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 as to be effectively cancelled, in the L2 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. LPP and BPP 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), univariate 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

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

2. Letting \( AH_i(z) \) be the error function for \( H_i(z) \), caused by rounding off its coefficients, the coefficients are optimized so that the error spectrum \( |AH_i(e^{j\omega})| \) is cancelled by \( H_i(z) \) in the L2 norm sense, where \( T \) is taken as unity, and

\[ H_i(z) = \prod_{i=1}^{T} H_i(z) \]

The discrete optimization can be formulated as

\[ E_3 = \int_{-\pi}^{\pi} |AH_i(e^{j\omega})H_i(e^{j\omega})|^2 d\omega \]

where \( E_3 \) is minimized for all \( H_i(z) \).

Filter Response Improvement

Let \( H(z) \) be expressed as

\[ H(z) = H_1(z)H_2(z) \]

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 2 bit variable wordlengths from the least significant bit (LSB). The following relations result

\[ |AH_{1\omega}(e^{j\omega})H_2(e^{j\omega})| = c_1 \]

\[ |H_1(e^{j\omega}) AH_{2\omega}(e^{j\omega})| = c_2 \]

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

\[ \frac{1}{2\pi} |AH_{1\omega}(e^{j\omega})|^2 d\omega = \frac{\Delta \omega}{3} N_1, \quad i = 1, 2 \]

where \( N_1 \) is the number of taps for \( H_i(z) \). From Eqs. (5) and (6),

\[ \frac{2\pi}{2\pi} \int_{-\pi/2}^{\pi/2} |H_{1\omega}(e^{j\omega})|^2 d\omega = \frac{\Delta \omega}{3} N_1, \quad i = 1, 2, j = 2, 1 \]

From this relation, \( c_1^2 \) can be obtained as

\[ c_1^2 = \frac{\Delta \omega}{3} N_1 |H_1(e^{j\omega})|^2, \quad i = 1, 2, j = 2, 1 \]

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

\[ |AH_{1\omega}(e^{j\omega})|^2 = c_1^2 + c_2^2 \]

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

\[ c_1^2 = \frac{\Delta \omega}{12} N_1 |H_1(e^{j\omega})|^2, \quad i = 1, 2, j = 2, 1 \]

484
where the $H_{1}(e^{j\omega})$ and $H_{2}(e^{j\omega})$ coefficients are also assumed to be uniformly distributed in the region $[-\Delta_{Q}/2, \Delta_{Q}/2]$. The $H(z)$ error spectrum is expressed

$$|\Delta H_{Q}(e^{j\omega})|^{2} = c_{1}^{*}e^{j\omega}c_{2}^{*}.$$  \hspace{1cm} (11)

If the following relation is satisfied

$$|\Delta H_{Q}(e^{j\omega})| < |\Delta H_{Q}(e^{j\omega})|$$  \hspace{1cm} (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 $L_{\infty}$ norms for $H_{1}(z)$, respectively. The frequency regions satisfying Eq. (12) usually appear in frequency selective filters, because an $L_{2}$ norm is well reduced from an $L_{\infty}$ norm.

Example

For simplicity, the amplitudes of $H_{1}(z)$ and $H_{2}(z)$ are assumed to be approximated by

$$|H_{1}(e^{j\omega})| = \frac{1}{1 + \omega^{2}}, \quad -\pi \leq \omega \leq \pi$$ \hspace{1cm} (13a)

$$|H_{2}(e^{j\omega})| = \frac{1}{1 + \omega^{4}}, \quad -\pi \leq \omega \leq \pi.$$ \hspace{1cm} (13b)

From Eqs. (8) and (9),

$$|\Delta H_{Q}(e^{j\omega})|^{2} = \frac{1}{12} \left( \frac{N_{1}}{1 + e^{j\omega}} \right)^{2} + \frac{N_{2}}{1 + e^{j2\omega}}.$$ \hspace{1cm} (14)

The $L_{2}$ norms for $H_{1}(z)$ and $H_{2}(z)$ are calculated as

$$\|H_{1}(e^{j\omega})\|^{2}_{2} = 153.8$$ \hspace{1cm} (15a)

$$\|H_{2}(e^{j\omega})\|^{2}_{2} = 1094.2.$$ \hspace{1cm} (15b)

Letting $\alpha$ be 2 bits, $\Delta_{q}$ relates to $\Delta_{Q}$ as

$$\Delta_{q} = 3.5\Delta_{Q}.$$ \hspace{1cm} (16)

Furthermore, by setting $N_{1}$ and $N_{2}$ to 100,

$$|\Delta B_{Q}(e^{j\omega})|^{2} = 3.03\Delta_{Q}^{2}.$$ \hspace{1cm} (17)

On the other hand,

$$|\Delta H_{Q}(e^{j\omega})|^{2}$$ \hspace{1cm} (18)

From Eq. (13),

$$|\Delta H_{Q}(e^{j\omega})|^{2} = \frac{100}{12} \left( \frac{1}{1 + \omega^{2}} \right)^{2} + \frac{1}{1 + \omega^{4}}.$$ \hspace{1cm} (19)

$$|\Delta B_{Q}(e^{j\omega})|^{2}$$ \hspace{1cm} (20)

where $W(z)$ is a weighting function for $P(z)$. 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 $\delta_{E}(z)$ so to be effectively cancelled by $W(z)$. Let $\delta_{I}$ and $\delta_{Q}$ be the impulse response for $\delta_{E}(z)$ and $W^{*}(z)$, respectively. The $\Delta H(z)$ impulse response can be expressed as

$$\Delta h_{n} = \sum_{m=0}^{N-1} w^{*} \delta_{I}^{*} \delta_{Q}^{*}, K = \min[n, M]$$ \hspace{1cm} (21)

where $M$ is degree of $W(z)$. From the Parseval relation, the error $E_{q}$ by Eq. (5) can be evaluated in a time domain as

$$E = \sum_{k=0}^{N-1} |\Delta h_{n}|^{2}$$ \hspace{1cm} (22)

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

$$E = \sum_{k=0}^{N-1} \sum_{m=0}^{n} w^{*} \delta_{I}^{*} \delta_{Q}^{*}, n = \min[n, M]$$ \hspace{1cm} (23)

Assume that $n$ satisfies

$$M \leq n$$ \hspace{1cm} (24)

then, $\Delta h_{n}$ and $\Delta h_{n+M+1}$ are expressed as

$$\Delta h_{n} = \sum_{m=0}^{N-1} w^{*} \delta_{I}^{*} \delta_{Q}^{*}$$ \hspace{1cm} (25a)

$$\Delta h_{n+M+1} = \sum_{m=0}^{N-1} w^{*} \delta_{I}^{*} \delta_{Q}^{*}.$$ \hspace{1cm} (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 $\Delta h_{K-k+1}$ to $\Delta h_{K}$ be $E(k)$

$$E(k) = \sum_{l=0}^{K-1} |\Delta h_{l}|^{2}$$ \hspace{1cm} (26)

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

$$k = \min\{K-k', K' \leq K \}$$ \hspace{1cm} (27)

where $L(K-k')$ and $E\{L+1\}(K-k')$ contain $(K'+M-1)$ common parameters. In the $E\{L\}(K-k')$ minimization, the parameters in the set $\{\delta_{I}^{*}L(K-k')\}$ are minimized at the $E(k)$, $(k < L(K-k'))$ minimization step, and the $L^{*}$ parameters in the set $\{L^{*}L(K-k')\}$ are optimized. After minimizing $E(k)$, the $L^{*}$ parameters included in the set $\{\delta_{I}^{*}L(K-k')\}$ are fixed at this step and the remaining $L^{*}$ parameters, included in the set, $\{\delta_{I}^{*}L(K-k')\}$ are used in minimizing.
$\varepsilon((L+1)(K+1))$ 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 $\varepsilon((L+1)(K+1))$ in the ($L+1$)'th power of $P$, where $P$ is a number of discrete value steps for each parameter. Therefore, a small ($L+1$)'th value means drastically saving computing time.

**Initial Guess:** The quantization error, caused by rounding off the approximate coefficients with infinite wordlengths, is taken as the initial guess for $A_1$.

**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. $P(z)$ is approximated by the Remez-exchange method (113 using $W(z)$ as a fixed weighting function. Coefficient wordlengths ($\*\$) do not include a sign bit, and the $P(z)$ coefficient values are not normalized. In the LPF case, zeros of $W(z)$ are all concentrated at $\pi$ radian on the $z$-plane. However, the parameters $A_1$ take only discrete values in the restricted range, and $[P(z)e^{j\theta}]$ is not strictly proportional to $[W(z)e^{j\theta}]^{-1}$. For this reason, zeros of $W(z)$ are set on $2\pi/3$ and $\pi$ 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, $W(z)P_{Q_0}(z)$, is almost the same as that of the original response $W(z)P_{Q_0}(z)$. The stopband attenuation for $W(z)P_{Q_0}(z)$ is somewhat improved, from that of $H_0(z)$, whose coefficients are rounded off only, because the $P_0(z)$ error spectrum is cancelled by $H_0(z)$ in the stopband. However, in the frequency range where the condition $|W(z)e^{j\theta}| < 1$ is not satisfied, that is, in the passband and the lower side in the stopband, the improvement rate becomes lower than $W(z)P_{Q_0}(z)$.

**Frequency Response Improvement Rate**

The maximum passband ripple (peak to peak) and the minimum stopband attenuation in $H_0(z)$, $R_0(z)$, $W(z)P_{Q_0}(z)$, and $W(z)P_{Q_0}(z)$ for the LPF and the BPF are shown in Fig. 3. The solid and dashed lines for $W(z)P_{Q_0}(z)$ indicate the variable wordlengths are 2 and 4 bits, respectively.

**Passband Ripple:** The improvement rates by $W(z)P_{Q_0}(z)$ are always very remarkable, compared with others.

**Stopband Attenuation:** In the case of LPF, $W(z)P_{Q_0}(z)$ is superior to $W(z)P_{Q_0}(z)$. They are, however, almost the same for the BPF case. However, the efficiency of $P_0(z)W(z)$ 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 $R_0(z)$ and $W(z)P_{Q_0}(z)$.

### 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


Fig. 1. Example of error spectrum shaping.

Table 1. Filter specifications and discrete optimization parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>LPF</th>
<th>BPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>H(z)</td>
<td>200 taps</td>
<td>200 taps</td>
</tr>
<tr>
<td>Sampling freq.</td>
<td>400 Hz</td>
<td>400 Hz</td>
</tr>
<tr>
<td>Passband</td>
<td>0-50 Hz</td>
<td>57-142 Hz</td>
</tr>
<tr>
<td>Ripple (Ap)</td>
<td>±0.085 dB</td>
<td>±0.035 dB</td>
</tr>
<tr>
<td>Stopband</td>
<td>56-200 Hz</td>
<td>0-50, 150-200 Hz</td>
</tr>
<tr>
<td>Attenuation</td>
<td>72.5 dB</td>
<td>80.0 dB</td>
</tr>
<tr>
<td>W(z)</td>
<td>(1+2z⁻¹z⁻²)</td>
<td>(1-z⁻²)²</td>
</tr>
<tr>
<td>W*(z)</td>
<td>(1+2z⁻¹z⁻²)²</td>
<td>(1-z⁻²)²</td>
</tr>
<tr>
<td>F(z)</td>
<td>196 taps</td>
<td>196 taps</td>
</tr>
<tr>
<td>Coefficient</td>
<td>8, 10, 12, 14* bits</td>
<td>8, 10, 12, 14* bits</td>
</tr>
<tr>
<td>wordlengths</td>
<td>bits</td>
<td>bits</td>
</tr>
<tr>
<td>Search region</td>
<td>1, 2 bits</td>
<td>1, 2 bits</td>
</tr>
<tr>
<td>K</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>K'</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>L</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>L'</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Fig. 2. Optimized frequency responses. (a) Solid line: W(z)P₀(z) with infinite precision coefficients, dashed line: H₀(z) with rounded off coefficients. (b) Solid line: W(z)P₀(z) with optimized coefficients, dashed line: W(z)P₀(z) with rounded off coefficients.

Fig. 3. Frequency response improvements. (a) LPF (b) BPF. Symbols A₀, A₉ and O correspond to H₀(z), W(z)P₀(z) and W(z)P₀(z), respectively. W(z)P₀(z) is shown by dashed and dotted line.