MINIMIZATION OF FINITE WORDLENGTH ERROR IN 2-D FIR DIGITAL FILTERS IN THE FREQUENCY DOMAIN

Mitsuhiko YAGYU *Akinori NISHIHARA* *Nobuo FUJII*

Department of Physical Electronics, Tokyo Institute of Technology
Tel: +81–3–5734–2560, Fax: +81–3–5734–2909, E-mail: mitsuhik@ss.titech.ac.jp

ABSTRACT

This paper presents a method to minimize the finite wordlength error in output signals of linear phase 2-D FIR filters. The finite wordlength errors can easily be analyzed in the frequency domain when the input signal statistics are known. In the case of white input signals, impulse responses corresponding to all levels of input impulses are optimized so as to minimize the errors. A new ROM-based filter structure is proposed in which the optimized impulse responses are stored in the ROM. The output signals are generated by superposing the impulse responses corresponding to the input levels. Many results of simulations confirm that the output signals of the proposed filters have far less errors than those of conventional filters.

1. INTRODUCTION

Two dimensional FIR digital filters are used in many applications such as image signal processing. In the practical implementations, the filters must have finite wordlength of both coefficients and internal signals. Rounding off the internal signals naturally causes errors in the output signal. Many methods have been proposed to design optimal 2-D FIR digital filters in the Chebyshev sense, even if the filter length and wordlength of coefficients are given as finite numbers[3]-[5]. Those methods, however, ignore finite wordlength effects of internal signals in the design procedure. If the internal signals of the filters have the finite wordlength, the products in the filters are rounded off and thus nonlinear operations are introduced. Then the roundoff errors are most properly analyzed by their impulse responses rather than frequency responses. The deteriorations in the impulse responses are dependent on the levels of the input impulses. In this paper, we propose a method to minimize the errors of the impulse responses in the frequency domain by using MILP for every possible input level. To realize the minimal error filters, the optimized responses of every input level are stored in a ROM and the output signals are generated by superposing the stored responses which correspond to the input signal. Many examples show the proposed filters are superior to the conventional ones.

2. ROM-BASED FILTER STRUCTURE

Generally, a response corresponding to an input impulse of a unit level is called as an *impulse response*. Hereafter, we also call a response corresponding to an input impulse of any level as an impulse response.

In this paper, we deal with linear phase 2-D FIR digital filters implemented as the direct form for image signal processing. It is assumed that all internal signals are expressed as fixed point binary numbers of a specified wordlength and that any samples of the impulse responses also have the same wordlength. Conventional filters have several multipliers and the products are rounded off to the specified wordlength.

We propose a filter structure having a ROM as shown in Fig. 1. Figure 1 shows a 1-D filter structure, but 2-D structures can be implemented in the same way. Superpositions of the shifted impulse responses can make the output signals, as any input signals are trains of impulses of different levels. The impulse responses to be stored in the ROM are optimized so that the filter has the minimal error. Depending on the input signal level, the impulse responses are successively referred and superposed to generate the output signal. Accordingly the roundoff operations are avoided unlike the conventional filters.

3. MINIMIZATION OF FINITE WORDLENGTH ERROR

3.1. Output signal estimation sequence (OSES)

Now we define the ideal zero phase 2-D FIR filter as F_d, and its frequency response as $H_d(\omega_1, \omega_2)$. The amplitude of $H_d(\omega_1, \omega_2)$ is 1.0 in whole passband, and 0.0 in whole stopband. Consider a given input image signal which is represented as binary numbers of a specified wordlength. Let the input signal be $x(n_1, n_2)$ and its size be $N_1 \times N_2$. Namely, $x(n_1, n_2)$ satisfies $x(n_1, n_2) = 0$, $n_1 < 0$, $n_1 \geq N_1$, $n_2 < 0$ or $n_2 \geq N_2$. The region where an image signal is defined is referred to as Φ. We define the vector $\boldsymbol{n}$ as $[n_1, n_2]^T$ and $\boldsymbol{\omega}$ as $[\omega_1, \omega_2]^T$. Then $x(n_1, n_2)$ and $H_d(\omega_1, \omega_2)$ can be written as $x(\boldsymbol{n})$ and $H_d(\boldsymbol{\omega})$, respectively. The discrete-time Fourier transform (DTFT) of the ideal output of F_d can be written as

$$Y_d(e^{j\omega_1}, e^{j\omega_2}) = \sum_{\boldsymbol{n}\in\Phi} x(\boldsymbol{n}) H_d(\boldsymbol{\omega}) e^{-j\boldsymbol{n}^T\boldsymbol{\omega}}. \quad (1)$$

Now let the tap size of the proposed filter be $T_1 \times T_2$. The analysis will be done for odd T_1 and T_2 as an example. Consider the input impulse at the point $\boldsymbol{n}$ and thus having the value $x(\boldsymbol{n})$. The corresponding impulse response is expressed as $h(x(\boldsymbol{n}), \boldsymbol{m})$, $n_1 - (T_1 - 1)/2 \leq m_1 \leq n_1 + (T_1 - 1)/2$, $n_2 - (T_2 - 1)/2 \leq m_2 \leq n_2 + (T_2 - 1)/2$. Now we define $H(x(\boldsymbol{n}), \boldsymbol{\omega}) e^{-j\boldsymbol{n}^T\boldsymbol{\omega}}$ as the DTFT of the impulse response $h(x(\boldsymbol{n}), \boldsymbol{m})$. The output of the proposed filter can be written as

$$Y(e^{j\omega_1}, e^{j\omega_2}) = \sum_{\boldsymbol{n}\in\Phi} H(x(\boldsymbol{n}), \boldsymbol{\omega}) e^{-j\boldsymbol{n}^T\boldsymbol{\omega}}. \quad (2)$$

The error in the output signal is given as

$$R(e^{j\omega_1}, e^{j\omega_2}) = Y(e^{j\omega_1}, e^{j\omega_2}) - Y_d(e^{j\omega_1}, e^{j\omega_2}). \quad (3)$$

Then we define the error which is included in $H(x(\boldsymbol{n}), \boldsymbol{\omega})$ as $R(x(\boldsymbol{n}), \boldsymbol{\omega})$, and the error can be written as

$$R(x(\boldsymbol{n}), \boldsymbol{\omega}) = H(x(\boldsymbol{n}), \boldsymbol{\omega}) - x(\boldsymbol{n}) H_d(\boldsymbol{\omega}). \quad (4)$$

By using (1), (2), (3) and (4), the error $R(e^{j\omega_1}, e^{j\omega_2})$ can be written as

$$R(e^{j\omega_1}, e^{j\omega_2}) = \sum_{\boldsymbol{n}\in\Phi} R(x(\boldsymbol{n}), \boldsymbol{\omega}) e^{-j\boldsymbol{n}^T\boldsymbol{\omega}}. \quad (5)$$

$R(x(\boldsymbol{n}), \boldsymbol{\omega})$ is a real function of $x(\boldsymbol{n})$ and $\boldsymbol{\omega}$. So if $\boldsymbol{\omega}$ is fixed, $R(x(\boldsymbol{n}), \boldsymbol{\omega})$, $n_1 = 0, 1, \cdots$, $n_2 = 0, 1, \cdots$ is a real sequence. We call this real sequence as an output signal estimation sequence (OSES). Equation (5) is the DTFT of the OSES. Then an input signal $x(\boldsymbol{n})$ determines an OSES $R(x(\boldsymbol{n}), \boldsymbol{\omega})$, $n_1 = 0, 1, \cdots$, $n_2 = 0, 1, \cdots$ corresponding to each frequency. The output error at $\boldsymbol{\omega}_a$ can be referred to as the value of the DTFT at $\boldsymbol{\omega}_a$ of the OSES corresponding to the frequency $\boldsymbol{\omega}_a$.

3.2. Mean squared output error spectrum (MSOES)

In this section, we show that the MSOES can be analyzed, even if the given input signals are not deterministic. Let the wordlength of the input signals be l and all the levels of input impulses be x_i, $i = 0, \cdots, L-1$ where $L = 2^l$. The OSES corresponding to a frequency $\boldsymbol{\omega}_a$ is $R(x(\boldsymbol{n}), \boldsymbol{\omega}_a)$ corresponding to an input signal $x(\boldsymbol{n})$ and then it is trains of elements in a set $\Psi_{\boldsymbol{\omega}_a}$ given by

$$\Psi_{\boldsymbol{\omega}_a} = \{R(x_i, \boldsymbol{\omega}_a) \mid i = 0, \cdots, L-1\}. \quad (6)$$

Now we assume that the stochastic input signals are stationary independent process and that their probability density function is defined as $p(x_i)$, $i = 0, \cdots, L-1$. Then the OSESs become stationary independent and thus their probability density function $q(R(x_i, \boldsymbol{\omega}_a))$ is obtained as

$$q(R(x_i, \boldsymbol{\omega}_a)) = p(x_i), \ i = 0, \cdots, L-1. \quad (7)$$

The mean power spectral density function of the OSESs corresponding to the frequency $\boldsymbol{\omega}_a$ is defined as $E[\mathcal{S}(e^{j\boldsymbol{\omega}})]$ and given in (23). Then the MSOES at frequency $\boldsymbol{\omega}_a$ is obtained as $E[\mathcal{S}(e^{j\boldsymbol{\omega}_a})]$. The second term in the right hand side in (23) becomes equivalent to the delta function and very large at DC, when N_1 and N_2 are infinite. So in this paper, the MSOES $E[\mathcal{P}(\boldsymbol{\omega})]$ is written as

$$E[\mathcal{P}(\boldsymbol{\omega})] \approx \begin{cases} R_{mse}(\mathbf{o}) + (N_1N_2 - 1)R_{me}^2(\mathbf{o}), & \boldsymbol{\omega}=\mathbf{o} \\ R_{mse}(\boldsymbol{\omega}) - R_{me}^2(\boldsymbol{\omega}) & \text{otherwise} \end{cases} \quad (8)$$

where

$$R_{me}(\boldsymbol{\omega}) = \sum_{i=0}^{L-1} R(x_i, \boldsymbol{\omega}) p(x_i), \quad (9)$$

$$R_{mse}(\boldsymbol{\omega}) = \sum_{i=0}^{L-1} R^2(x_i, \boldsymbol{\omega}) p(x_i). \quad (10)$$

If $R_{me}(\mathbf{o}) \neq 0$, the MSOES has very large error at DC. In other frequencies, the MSOES is given as the variance of the error responses $R(x_i, \boldsymbol{\omega}_a)$, $i = 0, \cdots, L-1$.

3.3. Optimization of All the Responses

From (8), the MSOES is given as the variance of the error responses. Then the MSOES at frequencies but DC can be written as

$$E[\mathcal{P}(\boldsymbol{\omega})] = \sum_{i=0}^{L-1} p(x_i) \{R(x_i, \boldsymbol{\omega}) - R_{me}(\boldsymbol{\omega})\}^2. \quad (11)$$

Only by optimizing all the error responses simultaneously, an optimum solution can be obtained. It is, however, difficult to carry out that optimization due to its enormous computing cost. Accordingly, we propose a method that $||R(x_i, \boldsymbol{\omega}) - R_{me}(\boldsymbol{\omega})||$, $i = 0, \cdots, L-1$ are minimized under the condition $R(x_i, \mathbf{o}) = 0$, iteratively, while the mean error response $R_{me}(\boldsymbol{\omega})$ is updated. By using that method, the MSOES shown in (8) can be minimized.

In the spatial domain, the responses $h(x_i, \boldsymbol{n})$ of linear phase filters have the symmetry. Then $H(x_i, \boldsymbol{\omega})$ can be written as

$$H(x_i, \boldsymbol{\omega}) = A(\boldsymbol{\omega}) \boldsymbol{h}_{x_i} \quad (12)$$

where $\boldsymbol{A}(\boldsymbol{\omega})$ is a vector whose elements are trigonometric functions and $\boldsymbol{h}_{x_i}$ is a vector whose elements are the independent coefficients of $h(x_i, \boldsymbol{n})$[2]. We propose the following algorithm to obtain the responses corresponding to all the input levels.

1. Let elements in a set Λ be probability densities $p(x_i)$, $i = 0, \cdots, L-1$, $R_{me}^*(\boldsymbol{\omega}) := 0$ and $s := 0$.
2. If the set Λ is empty, then stop.
3. Choose $p(x_k)$ which is the largest value in all elements in the set Λ. Exclude the element $p(x_k)$ from the set Λ. Let $H_d(\boldsymbol{\omega}) := x_k H_d(\boldsymbol{\omega})$.
4. The following mixed integer linear programing (MILP) problem is solved.

   Minimize $||\boldsymbol{A}(\boldsymbol{\omega})\boldsymbol{h}_{x_k} - H_d(\boldsymbol{\omega}) - R_{me}^*(\boldsymbol{\omega})||_\infty$,
   subject to $\boldsymbol{A}(\mathbf{o})\boldsymbol{h}_{x_k} = H_d(\mathbf{o})$ and each element of $\boldsymbol{h}_{x_k}$ has wordlength l.

5. By using the obtained $\boldsymbol{h}_{x_k}$, $R_{me}^*(\boldsymbol{\omega})$ is updated by

$$R_{me}^*(\boldsymbol{\omega}) := \frac{sR_{me}^*(\boldsymbol{\omega}) + p(x_k)\left\{\boldsymbol{A}(\boldsymbol{\omega})\boldsymbol{h}_{x_k} - H_d(\boldsymbol{\omega})\right\}}{s + p(x_k)}. \quad (13)$$

6. Let $s := s + p(x_k)$ and go to Step 2.

In the practical image signal processing, the probability density functions $p(x_i)$ of image signals are not known *a priori*. To prepare for various input signals, the probability density function used in the algorithm is given as a uniform distribution. Then the probability density function $p(x_i)$ can be written as

$$p(x_i) = \frac{1}{L}, \ i = 0, \cdots, L-1. \quad (14)$$

Let x_{me} be a mean value of input signals $x(\boldsymbol{n})$. In Step 3, the largest values of the probability densities are successively chosen. If the probability density function $p(x_i)$ is the uniform distribution, those largest values can not be determined uniquely. Therefore in that case, we modify Step 3 as follows.

3. Choose x_k which is the nearest value to x_{me} in all elements in the set Λ. Exclude the element x_k from the set Λ. Let $H_d(\boldsymbol{\omega}) := x_k H_d(\boldsymbol{\omega})$.

If a probability density function of input signals is known *a priori*, the original procedure is used as Step 3. Then better solutions can be obtained.

The MILP problem can be solved by using the branch and bound algorithm. $R_{me}^*(\boldsymbol{\omega})$ is the provisional mean error response during the optimization of the error responses. In Step 4, the difference between the error response $H(x_r, \boldsymbol{\omega}) - x_r H_d(\boldsymbol{\omega})$ and $R_{me}^*(\boldsymbol{\omega})$ is minimized. The algorithm minimizes not the error responses but the MSOES. The responses $H(x_i, \boldsymbol{\omega})$ are optimized so as to have similar error responses. Accordingly, $H(x_k, \boldsymbol{\omega})$ optimized by using the algorithm may have large deviations from $x_k H_d(\boldsymbol{\omega})$.

4. DESIGN EXAMPLES

The proposed filters obtained by using the above algorithm are compared with the conventional filters where the products are rounded off. Filter specifications are as follows.

	tap size	:	9×9
	wordlength	:	$6, 8$
Type I	passband	:	$\|\omega 1\| + \|\omega 2\| \leq 0.2\pi$
	stopband	:	$\|\omega 1\| + \|\omega 2\| \geq 0.6\pi$
Type II	passband	:	$\|\omega 1\| + \|\omega 2\| \leq 0.04\pi$
	stopband	:	$\|\omega 1\| + \|\omega 2\| \geq 0.4\pi$

The coefficients of the conventional filters are designed by using MILP. Those coefficients are designed under the condition that the errors at DC in the frequency responses of the conventional filters are strictly zero, because the power spectrums of image signals are usually very high at DC. To calculate PSNRs of output images of the conventional and the proposed filters, we need a reference which is regarded as the output images of the ideal filters. For this purpose, a 25×25 tap FIR filter is designed by using linear programing[1], which has -67.7dB Chebyshev error under the DC response condition.

Many standard images are quantized to 6 and 8 bits. Those quantized images are filtered by using the proposed, the conventional and the reference filters. The binary arithmetic is carried out in the proposed and the conventional filters, but the real arithmetic in the ideal filter. Then PSNRs of all the output images of the proposed and the conventional filters are calculated and shown in Tables 1(a), (b), (c) and (d). Figure 2 shows the quantized image of Lenna with wordlength 6. Figure 3 shows the output image by using the reference filter. Figures 4 and 5 show the output images of the proposed and the conventional filters which are designed so as to meet the specification wordlength 6 and Type I. From Figs.4 and 5, although the output image by the conventional filter has false contours, they are not observed in the image by the proposed filter. Tables 1 (a), (b), (c) and (d) indicate that the proposed design method is especially effective when the wordlength is short.

5. CONCLUSION

This paper proposes a method to minimize the finite wordlength error in the 2-D linear phase FIR digital filters. In the proposed method, the impulse responses corresponding to input impulses of all possible levels are optimized. The proposed filters have ROM where the optimized impulse responses are stored. The output signals are generated by superposing the impulse responses corresponding to the input impulses. In many design examples, we confirmed the superiority of the proposed filters to the conventional filters, where the roundoff operations are carried out.

REFERENCES

[1] J. V. Hu and L. R. Rabiner, "Design techniques for two-dimensional digital filters," IEEE Trans. Audio Electroacoust., vol. AU-20, pp. 259-257, Oct. 1972.

[2] C. Charalambous, "The performance of an algorithm for minimax design of two-dimensional linear phase FIR digital filters," IEEE Trans. on Circuits & Syst., vol. CAS-32, no. 10, pp. 1016-1028, Oct. 1985.

[3] P. Siohan and A. Benslimane, "Finite precision design of optimal linear phase 2-D FIR digital filters," IEEE Trans. on Circuits & Syst., vol. 36, no. 1, pp. 11-22, Jan. 1989.

[4] N. Benvenuto, M. Marchesi, and A. Uncini, "Applications of simulated annealing for the design of special digital filters," IEEE Trans. on Signal Process., vol. 40, no. 2, pp. 323-332, Feb. 1992.

[5] M. Park and W. Song, "A new design method of 2-d linear phase FIR filters with finite-precision coefficients," IEEE Trans. on Circuits & Syst., vol. 41, no. 7, pp. 478-482, Jul. 1994.

A THE MEAN POWER SPECTRAL DENSITY FUNCTION OF OSES

In this section, the OSES $R(x(\boldsymbol{n}), \boldsymbol{\omega}_a)$ is called $v(\boldsymbol{n})$, briefly. The stochastic OSESs $R(x(\boldsymbol{n}), \boldsymbol{\omega}_a)$ are stationary independent process and have the probability density function $q(R(x_i, \boldsymbol{\omega}_a))$ $i = 0, \cdots, L-1$. By using (6) and (7), the mean and the variance of OSESs can be obtained as

$$E[v(\boldsymbol{n})] = \sum_{i=0}^{L-1} R(x_i, \boldsymbol{\omega}_a) p(x_i) \tag{15}$$

and

$$V[v(\boldsymbol{n})] = \sum_{i=0}^{L-1} R^2(x_i, \boldsymbol{\omega}_a) p(x_i) - E^2[v(\boldsymbol{n})]. \tag{16}$$

Now we define another sequence $\hat{v}(\boldsymbol{n})$ satisfying

$$\hat{v}(\boldsymbol{n}) \stackrel{\triangle}{=} v(\boldsymbol{n}) - E[v(\boldsymbol{n})]. \tag{17}$$

Then $\hat{v}(\boldsymbol{n})$ also satisfies

$$E[\hat{v}(\boldsymbol{n})] = 0, \tag{18}$$

$$E[\hat{v}^2(\boldsymbol{n})] = E[v^2(\boldsymbol{n})] - E^2[v(\boldsymbol{n})] = V[v(\boldsymbol{n})] \tag{19}$$

and

$$E[\hat{v}(\boldsymbol{n})\hat{v}(\boldsymbol{n} - \boldsymbol{m})] = 0 \tag{20}$$

where $\boldsymbol{m} \neq \mathbf{0}$. Now let $E[\mathcal{S}(e^{j\boldsymbol{\omega}})]$ be the mean power spectral density function of $v(\boldsymbol{n})$ and $E[\hat{\mathcal{S}}(e^{j\boldsymbol{\omega}})]$ be that of $\hat{v}(\boldsymbol{n})$, respectively. By using (20), $E[\hat{\mathcal{S}}(e^{j\boldsymbol{\omega}})]$ is obtained as

$$E[\hat{\mathcal{S}}(e^{j\boldsymbol{\omega}})] = E[\hat{v}^2(\boldsymbol{n})] \tag{21}$$

By using $\hat{v}(\boldsymbol{n})$, $E[\mathcal{S}(e^{j\boldsymbol{\omega}})]$ can be written as

$$E[\mathcal{S}(e^{j\boldsymbol{\omega}})] = E[\hat{\mathcal{S}}(e^{j\boldsymbol{\omega}})] + \frac{E^2[v(\boldsymbol{n})]}{N_1 N_2} \left| \sum_{\boldsymbol{n} \in \Phi} e^{-j\boldsymbol{n}^T \boldsymbol{\omega}} \right|^2 + \frac{2E[v(\boldsymbol{n})]}{N_1 N_2} \sum_{\boldsymbol{n} \in \Phi} E[\hat{v}(\boldsymbol{n})] \sum_{\boldsymbol{m} \in \Phi} \cos[\boldsymbol{n} - \boldsymbol{m}]^T \boldsymbol{\omega}. \tag{22}$$

By substituting Eqs.(18), (19) and (21), (22) can be rewritten as

$$E[\mathcal{S}(e^{j\boldsymbol{\omega}})] = V[v(\boldsymbol{n})] + \frac{E^2[v(\boldsymbol{n})]}{N_1 N_2} \left| \sum_{\boldsymbol{n} \in \Phi} e^{-j\boldsymbol{n}^T \boldsymbol{\omega}} \right|^2. \tag{23}$$

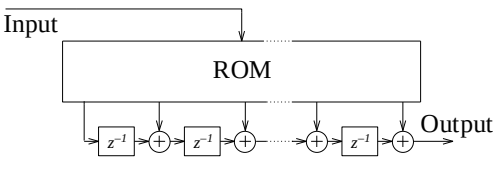

Figure 1. Rom-based filter structure

Table 1. PSNRs of the output images

(a) Wordlength 6 and Type I

	proposed	conventional
Lenna	26.0 [dB]	15.2 [dB]
Swiss mountain	22.9 [dB]	17.4 [dB]
Girl	21.0 [dB]	14.2 [dB]
Moon	26.1 [dB]	15.6 [dB]
Title	26.1 [dB]	9.94 [dB]

(b) Wordlength 8 and Type I

	proposed	conventional
Lenna	33.0 [dB]	30.5 [dB]
Swiss mountain	30.7 [dB]	30.5 [dB]
Girl	28.6 [dB]	27.2 [dB]
Moon	33.4 [dB]	30.2 [dB]
Title	30.4 [dB]	28.9 [dB]

(c) Wordlength 6 and Type II

	proposed	conventional
Lenna	24.5 [dB]	5.3 [dB]
Swiss mountain	21.5 [dB]	8.5 [dB]
Girl	19.6 [dB]	4.0 [dB]
Moon	24.6 [dB]	5.1 [dB]
Title	22.8 [dB]	7.1 [dB]

(d) Wordlength 8 and Type II

	proposed	conventional
Lenna	32.1 [dB]	25.5 [dB]
Swiss mountain	30.3 [dB]	27.4 [dB]
Girl	27.3 [dB]	23.1 [dB]
Moon	32.3 [dB]	25.6 [dB]
Title	29.9 [dB]	21.9 [dB]

Figure 2. Input image (lenna 64 bit/pixel)

Figure 3. Ideal output image

Figure 4. Output image by the proposed filter

Figure 5. Output image by the conventional filter