Electronics and Communications in Japan, Part 3, Vol. 89, No. 1, 2006
Translated from Denshi Joho Tsushin Gakkai Ronbunshi, Vol. J88-A, No. 1, January 2005, pp. 39–53

SNR-Maximizing Interpolation Filters for Band-Limited Signals with Quantization

Yoshinori Takei,[1] Kouichi Mogi,[2] Toshinori Yoshikawa,[1] and Xi Zhang[3]

[1]Department of Electrical Engineering, Nagaoka University of Technology, Nagaoka, 940-2188 Japan

[2]Nippon Seiki Co. Ltd., Nagaoka, 940-8580 Japan

[3]Department of Information and Communication Engineering, University of Electro-Communications, Chofu, 182-8585 Japan

SUMMARY

As an interpolation filter for sampling rate transformation, a half-band filter with a reduced amount of computation is often used. Due to the restrictions on its amplitude characteristic, it is not possible to sufficiently reduce the quantization noise of the transition region of the filter. In this paper, the output SNR of the L-interpolator filter is analyzed with a quantized band-limited signal as the input. A design method is proposed for the linear phase FIR filter maximizing the output SNR. For the design, both the SNR maximizing design within the Type I FIR filters and one supplemented with the restriction of the L-th band filter are presented. Each design is reduced to derivation of the solution of a system of linear equations with a coefficient matrix represented analytically. The filter based on the proposed method can attain an SNR not attainable under the sacrifice of the order in the conventional filter taking into account only the passband and the stopband. © 2005 Wiley Periodicals, Inc. Electron Comm Jpn Pt 3, 89(1): 31–46, 2006; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/ecjc.20178

Key words: interpolation filter; least square design; quantization noise; SNR improvement.

1. Introduction

Increasing the sampling rate by the interpolation method is one of the basic techniques for multirate signal processing. In general, when the sampling rate is increased, an interpolation filter is used in which null samples are inserted between the samples by the up-sampler and then the imaging components are eliminated by a low-pass filter [1]. Figure 1 shows an interpolator that increases the sampling rate by a factor of L, a positive integer, by a cascade connection of an L up-sampler and an interpolation filter $H(z)$. It is usual to use a linear phase FIR filter with a passband gain of L as an interpolation filter. In particular, when the sampling rate is doubled, FIR half-band filters, which are advantageous with regard to the amount of computation because about half of the coefficient values become zero, are often used [3, 7, 9]. In this case, a filter design specification to control the amplitude characteristics in the passband and the stopband is used, such as an equal-ripple design specification minimizing the maximum approximation error in the passband and the stopband. On

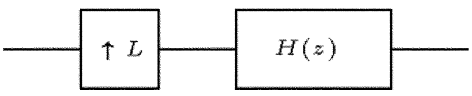


Fig. 1. An L-interpolator.

the other hand, with such specifications, the amplitude characteristics in the transition region cannot be controlled directly. Also, the amplitude characteristic is forced to be odd symmetric with regard to the normalized angular frequency $\pi/2$.

In a real digital signal processing system, the input signal to the interpolation filter is band-limited and quantized. For instance, in the case of audio CDs, the bandwidth is up to 20 kHz at the maximum but the signal is sampled at a sampling rate of 44.1 kHz and is quantized into 16 bits in a fixed point representation [4]. When the sampling rate of such a digital signal is doubled by using the configuration in Fig. 1, the input signal component reaching the interpolation filter $H(z)$, excluding the image and the quantization noise, is band-limited to the lower frequency side of $\pi/2 \cdot (2 \cdot 20/44.1)$, which is even lower than $\pi/2$. More generally, it can be stated that the signal is band-limited to the frequency range of $[0, \alpha\pi/2)$, where α is the band-limiting coefficient with $0 < \alpha < 1$. On the other hand, quantization noise is considered to exist at the input of the interpolation filter in the entire range of the frequency domain. This situation is modeled in Fig. 2, where the stippling indicates the quantization noise, the solid trapezoid the signal components to be passed, and the dashed trapezoid the imaging component.

Let us assume that an FIR half-band filter is used as an interpolation filter for the quantized band-limited signal. Of the quantization noise, the component located in the stopband of the interpolation filter is sufficiently suppressed if the stopband attenuation of the filter is sufficient. Hence, an increase of the filter order to make the amplitude characteristic in the stopband approach the ideal one is effective for suppression of quantization noise in the stopband. On the other hand, of the quantization noise, the component in the transition region (with the angular frequency ω near $\pi/2$) is not necessarily suppressed by increasing the order of the filter [6, 10]. This can be understood from the fact that the amplitude characteristics (2 for $\omega \in [0, \pi/2)$ and 0 for $\omega \in (\pi/2, \pi]$) that divide the frequency domain ideally into two let the component located in the frequency range of $\omega \in [\alpha\pi/2, \pi/2)$ pass without suppression. Hence, a half-band filter, even an ideal one, is not optimum for the objective of maximizing the signal-to-noise ratio (SNR) after interpolation of the quantized band-limited signal.

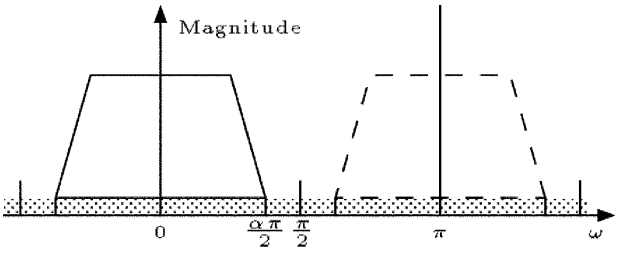


Fig. 2. Frequency spectrum characteristics of the input signal of the interpolation filter.

This paper treats the problem of maximizing the SNR at the output by optimization of the interpolation filter in the case where the quantized band-limited signal is L-interpolated. A design method for the interpolation filter is proposed that provides the best SNR in the output by an optimization of the interpolation filter with a given order. In Section 2, the method of evaluation of the noise at the output of the interpolation filter is described. The definition of the output SNR in the time domain is established. In Section 3, based on the analysis of the multirate random signals according to Sathe and Vaidyanathan [5] and Tuqan and Vaidyanathan [8], the output SNR defined in the time domain is transformed to a theoretical SNR equation in the frequency domain described by the frequency characteristics of the interpolation filter. The validity of the theoretical SNR equation derived is verified in Section 4 by comparison with a simulation based on the SNR definition in the time domain. In Section 5, the design method of the linear phase FIR filter which maximizes the theoretical SNR equation derived is proposed. The design problem is reduced to the least squares problem and the filter coefficients can be obtained by solving a system of linear equations. The design method is presented for the case in which the condition of the L-th band filter is imposed and not imposed. Comparisons of the output SNR performance between the proposed method, the conventional method, and the filter design example are presented in Section 6.

2. Definition of Output SNR in the Time Domain

In this section, an evaluation method is presented for the output error of the L-interpolation filter for the quantized band-limited signal and the output SNR is defined in the time domain.

2.1. Output error of the interpolator

Let us first consider output error evaluation for the interpolation filter in the absence of the quantization error. In the L-interpolator in Fig. 1, the number of samples at the output is L times the number of samples at the input. Hence, in order to evaluate the output error of this interpolator, the configuration in Fig. 3 is used. The original signal $y[m]$ is L-down-sampled to $x[n]$, which is input into the interpolator to be evaluated. By means of the L-up-sampler of the

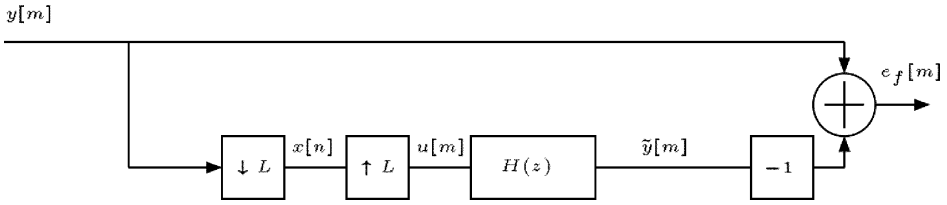


Fig. 3. Output error assessment of an interpolator in the absence of quantization noise.

interpolator, the signal $u[m]$ with the same number of samples as the original signal is obtained. This becomes the input to the interpolation filter $H(z)$. The following quantity, obtained by subtracting the output $\tilde{y}[m]$ of the interpolation filter from the original signal,

$$e_f[m] := y[m] - \tilde{y}[m] \tag{1}$$

is the result of the deviation of the characteristic of the interpolation filter from the ideal one and let us call it the filtering noise. Its square $e_f[m]^2$ can be used to evaluate the magnitude of the output error. This depends on the original signal $y[m]$ and the time m. They must be defined appropriately for evaluation of the interpolation filter $H(z)$.

2.2. Reference input signal

The original signal $y[m]$ must be one that is band-limited in the frequency range of $[0, \alpha\pi/2)$ and can evaluate the output error of $H(z)$ in a uniform manner. Hence, the real-valued random signal $y[m]$ obtained by ideally band-limiting the white noise $w[m]$ to the lower frequency side of $\alpha\pi/L$ is defined as the original signal. This is called the reference input signal. This $y[m]$ is a weakly stationary WSS (Wide Sense Stationary) process. Hence, the expectation value of the autocorrelation

$$R_{yy}[k] := \mathsf{E}\Big[y[m]y[m-k]\Big]$$

is defined independently of m and the spectral power density function

$$S_{yy}(e^{j\omega}) := \sum_{k\in\mathbb{Z}} R_{yy}[k]e^{-jk\omega} \tag{2}$$

is defined. The power spectral density of the white noise $w[m]$ prior to band limitation is expressed in terms of the variance σ_w^2 as

$$S_{ww}(e^{j\omega}) = \sigma_w^2 \qquad (\forall\omega \in \mathbb{R}) \tag{3}$$

Then, the power spectral density function $S_{yy}(e^{j\omega})$ of $y[m]$ obtained by ideally band-limiting the above is

$$S_{yy}(e^{j\omega}) = \begin{cases} \sigma_w^2 & \left(\begin{array}{c} \exists\, s \in \mathbb{Z}, \quad \omega \in \\ [2\pi s - \frac{\alpha\pi}{L}, 2\pi s + \frac{\alpha\pi}{L}] \end{array}\right) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

2.3. Quantization noise

The quantization noise caused by rounding real-valued $y[m]$ to a fixed point number whose fractional part has a bit length of b is modeled as additive noise $q[m]$. Figure 4 shows the method of evaluation of the output error from the interpolator in the presence of quantization noise. In the figure, $\tilde{q}[m]$ is the response to $q[m]$ of the cascade connection of the down-sampler, up-sampler, and interpolation filter $H(z)$, and

$$\widehat{y}[m] := \tilde{y}[m] + \tilde{q}[m] \tag{5}$$

is the output of the interpolation filter $H(z)$. The output error considering both the filtering noise and the quantization noise is

$$e[m] := y[m] - \widehat{y}[m] = y[m] - (\tilde{y}[m] + \tilde{q}[m]) \tag{6}$$

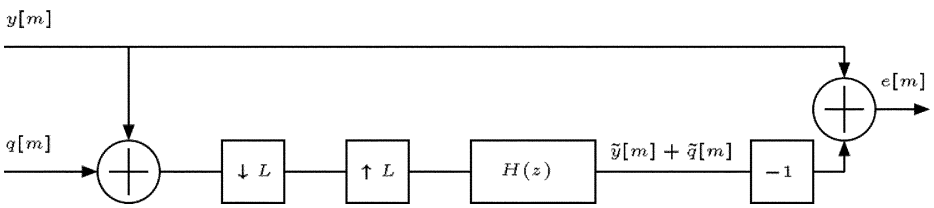


Fig. 4. Output error assessment of an interpolator in the presence of quantization noise.

With regard to the quantization error $q[m]$, the following assumption is used: $q[m]$ is a white noise independent of $w[m]$ and its power spectral density is

$$S_{qq}(e^{j\omega}) = \sigma_q^2 \qquad (\forall\, \omega \in \mathbb{R}) \tag{7}$$

Also, it is assumed that $y[m]$ and $q[m]$ are jointly WSS processes. The definition of the jointly WSS characteristic will be given in the next section.

If $w[m]$ takes a real value in (–1, 1) and $y[m]$ is rounded by quantization so that its fractional part is of length b bits, then the following holds:

$$\frac{\sigma_q^2}{\sigma_w^2} = 2^{-2(b+1)} \tag{8}$$

2.4. Definition of the output SNR in the time domain

With the above setting, it appears appropriate to define the output SNR of the interpolation filter by taking the energy ratio $y[m]^2/e[m]^2$ of the output error signal $e[m]$ in Eq. (6) and the reference input signal $y[m]$ and considering its expectation. However, due to the presence of the L-up-sampler, the WSS characteristic of $e[m]$ is not guaranteed even if the reference input signal $y[m]$ and the quantization error $q[m]$ have WSS characteristics. The expectation value of the problem cannot be determined independently of m. As seen in the next section, $e[m]$ is guaranteed to be an L-cycle wise sensor stationary (CWSS) process somewhat weaker than WSS. The expectation value $\mathsf{E}[L^{-1}\Sigma_{i=0}^{L-1} e[nL-i]^2]$ of the average of L samples of the squared error is determined independently of n. Hence, the SNR of the output signal of the interpolation filter is defined by

$$\mathrm{SNR} := \frac{\mathsf{E}\left[\frac{1}{L}\sum_{i=0}^{L-1} y[i]^2\right]}{\mathsf{E}\left[\frac{1}{L}\sum_{i=0}^{L-1} e[i]^2\right]} \tag{9}$$

2.5. Note about evaluation of the output SNR

The SNR defined in this section considers only the quantization error at the input of the interpolation filter. It does not include noise due to the interpolation filter operation with a finite word length and requantization at the output. It is of course desirable to carry out more precise analysis taking account of the effect of the quantization on the output side. Nevertheless, the model in this section is considered to be accurate in practice if the bit lengths of the output and the interpolation filter operation below the decimal point are sufficiently longer than the bit length b of the fractional part of the input of the interpolation filter.

3. Derivation of Theoretical Equation for SNR in the Frequency Domain

In this section, the SNR defined in the time domain in the previous section is transformed to the theoretical equation for the SNR in the frequency domain described by the interpolation ratio L, the band-limiting coefficient α, and the number of bits b of the fractional part, and the frequency characteristics of the interpolation filter.

The evaluation system of the output error of the interpolator in Fig. 4 in the previous section is not time-invariant due to the presence of the L-up-sampler. This makes the analysis of the output error somewhat more complicated. Hence, as shown in Refs. 5 and 8, an analysis is performed by using the vector-valued signal combining the adjacent L points of the signal.

For the reference input signal $y[m]$, the L-dimensional vector signal is defined as

$$\mathbf{y}[n] := \begin{bmatrix} y[nL] \\ \vdots \\ y[nL-i] \\ \vdots \\ y[nL-(L-1)] \end{bmatrix} \tag{10}$$

(see Fig. 5). The z transform of the signal $y[m]$ and its polyphase decomposition

$$Y(z) := \sum_{m\in\mathbb{Z}} y[m]\, z^{-m} = \sum_{i=0}^{L-1} z^i Y_i(z^L) \tag{11}$$

$$Y_i(z^L) := \sum_{n\in\mathbb{Z}} y[nL-i]\, z^{-nL} \tag{12}$$

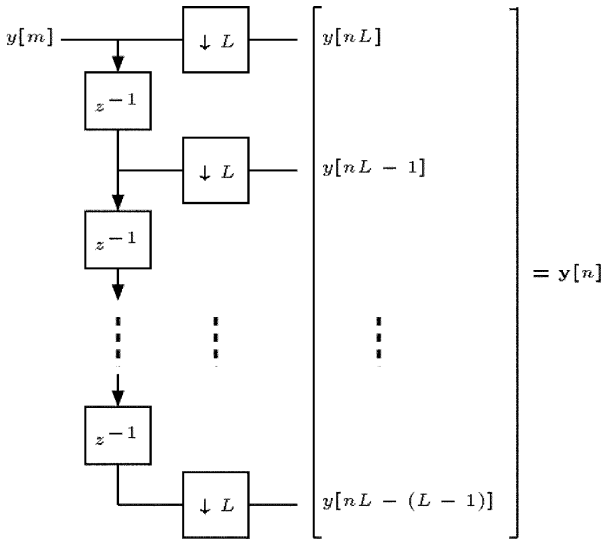


Fig. 5. Vector-valued signal $\mathbf{y}[n]$.

are related to the z transform of $\mathbf{y}[n]$

$$\mathbf{Y}(z^L) := \sum_{n\in\mathbb{Z}} \mathbf{y}[n]\, z^{-nL} \tag{13}$$

via the following relationship:

$$\mathbf{Y}(z^L) = \begin{bmatrix} Y_0(z^L) \\ \vdots \\ Y_i(z^L) \\ \vdots \\ Y_{L-1}(z^L) \end{bmatrix} \tag{14}$$

Here, by letting

$$\mathbf{\Phi}(z) := \begin{bmatrix} 1 & \dots & z^i & \dots & z^{L-1} \end{bmatrix} \tag{15}$$

Eq. (14) can be rewritten as

$$Y(z) = \mathbf{\Phi}(z)\mathbf{Y}(z^L) \tag{16}$$

In this section, it is assumed that $\mathbf{x}[n]$, $X(z)$, $X_i(z^L)$, and $\mathbf{X}(z^L)$ are defined in the same way as in Eqs. (10), (11), (12), and (13) for an arbitrary scalar signal $x[m]$. Also, for conciseness of the expression, let us define

$$\mathsf{LT}(s,t) := \begin{cases} 1 & (s < t) \\ 0 & \text{otherwise} \end{cases}$$

for $s, t \in \mathbb{R}$.

First, let us consider an equivalent rewriting of the input/output relationship in Fig. 3 without considering the quantization by means of the L-dimensional vector signal. For the L-dimensional vector $\mathbf{u}[n]$ of the output of the L-up-sampler, it is possible to write

$$\mathbf{U}(z^L) = \begin{bmatrix} 1 & \mathbf{O}_{1,L-1} \\ \mathbf{O}_{1,L-1} & \mathbf{O}_{L-1} \end{bmatrix} \mathbf{Y}(z^L) \tag{17}$$

Here, $\mathbf{O}_{1,L-1}$, $\mathbf{O}_{L-1,1}$, and $\mathbf{O}_{L-1}$ are 0 blocks of $1\times(L-1)$, $(L-1)\times 1$, and $(L-1)\times(L-1)$. Also, if the transfer function $H(z)$ of the interpolation filter is represented in polyphase as

$$H(z) = \sum_{\ell=0}^{L-1} z^{-\ell} H_\ell(z^L) \tag{18}$$

then the following is obtained:

$$\begin{aligned} \widetilde{Y}(z) &= H(z)U(z) \\ &= \sum_{\lambda=0}^{L-1} z^{\lambda} \sum_{\substack{0\le i<L \\ 0\le \ell<L \\ i-\ell\equiv\lambda \pmod L}} z^{i-\ell-\lambda} H_\ell(z^L) U_i(z^L) \end{aligned} \tag{19}$$

Since $l = i - \lambda + L\mathsf{LT}(i,\ \lambda)$ under the conditions $i - l \equiv \lambda \pmod L$, $0 \le i < L$, and $0 \le l < L$, we have

$$\widetilde{Y}(z) = \sum_{\lambda=0}^{L-1} z^{\lambda} \sum_{i=0}^{L-1} z^{-L\,\mathsf{LT}(i,\lambda)} H_{i-\lambda+L\,\mathsf{LT}(i,\lambda)}(z^L) U_i(z^L) \tag{20}$$

If the matrix

$$\mathbf{H}(z^L) = \begin{bmatrix} H^{[\lambda,i]}(z^L) \end{bmatrix}_{\substack{0\le\lambda<L \\ 0\le i<L}}$$

is defined by

$$H^{[\lambda,i]}(z^L) := z^{-L\,\mathsf{LT}(i,\lambda)} H_{i-\lambda+L\,\mathsf{LT}(i,\lambda)}(z^L) \tag{21}$$

then $\mathbf{H}(z^L)$ takes the form

$$\begin{bmatrix} H_0(z^L) & H_1(z^L) & \dots & H_{L-1}(z^L) \\ \frac{H_{L-1}(z^L)}{z^L} & H_0(z^L) & \dots & H_{L-2}(z^L) \\ \vdots & & \ddots & \vdots \\ \frac{H_1(z^L)}{z^L} & \frac{H_2(z^L)}{z^L} & \dots & H_0(z^L) \end{bmatrix}$$

Therefore, Eq. (20) can be written as $\widetilde{Y}(z) = \mathbf{\Phi}(z)\mathbf{H}(z^L)\mathbf{U}(z^L)$. Comparing this with $\widetilde{Y}(z) = \mathbf{\Phi}(z)\widetilde{\mathbf{Y}}(z)$, we obtain

$$\widetilde{\mathbf{Y}}(z^L) = \mathbf{H}(z^L)\mathbf{U}(z^L) \tag{22}$$

Combined with Eq. (17), we obtain

$$\widetilde{\mathbf{Y}}(z^L) = \mathbf{G}(z^L)\mathbf{Y}(z^L) \tag{23}$$

where

$$\mathbf{G}(z^L) := \mathbf{H}(z^L) \begin{bmatrix} 1 & \mathbf{O}_{1,L-1} \\ \mathbf{O}_{1,L-1} & \mathbf{O}_{L-1} \end{bmatrix} \tag{24}$$

From the above, the input and output relationships in Fig. 3 can be replaced by those in Fig. 6.

Next, let us consider the L-dimensional expression under the additive quantization noise. For the L-dimensional expression of $q[m]$ and $\tilde{q}[m]$, the relationship $\widetilde{\mathbf{Q}}(z^L) = \mathbf{G}(z^L)\mathbf{Q}(z^L)$ holds and Fig. 4 is replaced by Fig. 7. Here, let the reference input signal and the quantization noise be combined and written as the $2L$-dimensional $[\mathbf{y}[n]\ \mathbf{q}[n]]^T$. Then, by using $\mathbf{P}(z^L) = [\mathbf{Y}(z^L)\ \mathbf{Q}(z^L)]^T$, $\mathbf{E}(z^L)$ can be written as

$$\begin{aligned} \mathbf{E}(z^L) &= \mathbf{Y}(z^L) - \widetilde{\mathbf{Y}}(z^L) - \widetilde{\mathbf{Q}}(z^L) \\ &= \begin{bmatrix} \mathbf{I}_L - \mathbf{G}(z^L) & -\mathbf{G}(z^L) \end{bmatrix} \mathbf{P}(z^L) \end{aligned} \tag{25}$$

On the other hand, the output $\tilde{y}[m]$ of the interpolator for the signal $y[m]$ cannot be guaranteed to be WSS due to the effect of the up-sampler, but is an L-Cyclo Wide Sense

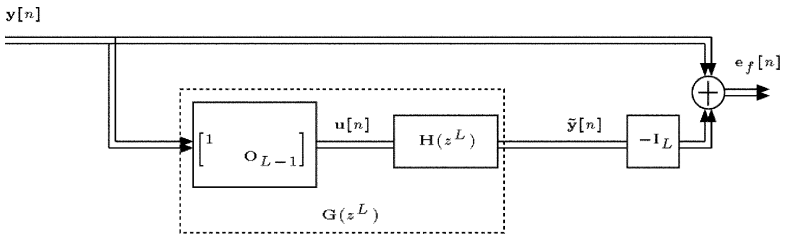


Fig. 6. An equivalent presentation of Fig. 3 using L-dimension signal.

Stationary $(\mathrm{CWSS})_L$ process. Therefore, $\tilde{\mathbf{y}}[n]$ obtained by combining L successive samples of $\tilde{y}[m]$ as one vector is an L-dimensional WSS process and the autocorrelation expectation value matrix $\mathbf{R}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}[k] := \mathsf{E}\,[\tilde{\mathbf{y}}[n]\tilde{\mathbf{y}}^{\dagger}[n-k]]$ is determined independently of n. In this case, the power spectrum density is defined as

$$\mathbf{S}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}(e^{jL\omega}) := \sum_{k\in\mathbb{Z}} \mathbf{R}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}[k]e^{-jkL\omega}$$

The output $\tilde{q}[m]$ of the filter for the quantization noise is also a $(\mathrm{CWSS})_L$ process and $\mathbf{R}_{\tilde{\mathbf{q}}\tilde{\mathbf{q}}}[k]$ and $\mathbf{S}_{\tilde{\mathbf{q}}\tilde{\mathbf{q}}}(e^{j\omega L})$ can be similarly defined.

In Section 2.3, it is assumed that $y[m]$ and $q[m]$ are jointly WSS processes. The definition is that the two-dimensional signal $[y[m]\ q[m]]^T$ is a WSS process. Hence, $\mathbf{y}[n]$ and $\mathbf{q}[n]$ are jointly WSS processes and thus the $2L$-dimensional vector

$$\mathbf{p}[n] = \begin{bmatrix}\mathbf{y}[n]\\ \mathbf{q}[n]\end{bmatrix}$$

is a WSS process. Also, from the assumption of independence of $w[m]$ and $q[m]$ in Section 2.3, $\mathbf{y}[n]$ and $\mathbf{q}[n]$ become uncorrelated and hence

$$\mathbf{S}_{\mathbf{pp}}(e^{jL\omega}) = \begin{bmatrix}\mathbf{S}_{\mathbf{yy}}(e^{jL\omega}) & \mathbf{O}_L\\ \mathbf{O}_L & \mathbf{S}_{\mathbf{qq}}(e^{jL\omega})\end{bmatrix} \tag{26}$$

Then, from the above jointly WSS characteristic and Eq. (25), it is found that $\mathbf{e}[n]$ is also a WSS process. So,

$$\mathbf{R}_{\mathbf{ee}}[k] := \mathsf{E}\Big[\mathbf{e}[n]\mathbf{e}^{\dagger}[n-k]\Big] \tag{27}$$

and

$$\mathbf{S}_{\mathbf{ee}}(e^{jL\omega}) := \sum_{k\in\mathbb{Z}} \mathbf{R}_{\mathbf{ee}}[k]\, e^{-jkL\omega} \tag{28}$$

are determined independently of n. In particular,

$$\begin{aligned}\mathcal{E} &:= \frac{1}{L}\mathrm{trace}\,(\mathbf{R}_{\mathbf{ee}}[0]) = \frac{1}{L}\mathsf{E}\Big[\mathbf{e}^{\dagger}[n]\mathbf{e}[n]\Big]\\ &= \mathsf{E}\Big[\frac{1}{L}\sum_{i=0}^{L-1}(e[nL-i])^2\Big]\end{aligned} \tag{29}$$

can be defined independently of n. From the last expression (29), it is found that $\mathcal{E}$ indicates an average energy of the output error per sample point under the quantization noise $q[m]$.

In order to analyze this $\mathcal{E}$ in the frequency domain, the following lemma is presented.

[Lemma 3.1]

$$\mathcal{E} = \frac{1}{2\pi L}\int_{-\pi}^{\pi} \mathbf{\Phi}(e^{j\omega})\mathbf{S}_{\mathbf{ee}}(e^{jL\omega})\mathbf{\Phi}^{\dagger}(e^{j\omega})d\omega$$

(Proof) From Eqs. (27) and (28) and the relationship between $\mathbf{e}[n]$ and $e[m]$, we have

$$\begin{aligned}&\frac{1}{L}\mathbf{\Phi}(e^{j\omega})\mathbf{S}_{\mathbf{ee}}(e^{jL\omega})\mathbf{\Phi}^{\dagger}(e^{j\omega})\\ &\quad = \frac{1}{L}\sum_{\mu=0}^{L-1}\sum_{k\in\mathbb{Z}}\sum_{\nu=0}^{L-1}\end{aligned}$$

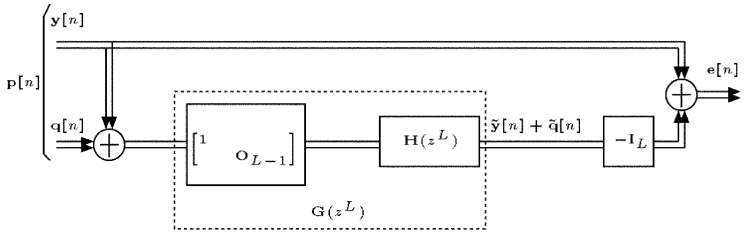


Fig. 7. An equivalent presentation of Fig. 4 using L-dimension signal.

$$\mathsf{E}\Big[e[nL-L+\mu]e[(n-k)L-\nu]\Big] \cdot e^{-j((k-1)L+\mu+\nu)\omega} \tag{30}$$

Because of the fact $\mathbf{e}[n]$ is a WSS process, this relationship holds regardless of $n \in \mathbb{Z}$. Letting $n = 1$ on the right-hand side, we have

$$\frac{1}{L}\sum_{\mu=0}^{L-1}\sum_{k\in\mathbb{Z}}\sum_{\nu=0}^{L-1} \mathsf{E}\Big[e[\mu]e[\mu-((k-1)L+\mu+\nu)]\Big] \cdot e^{-j((k-1)L+\mu+\nu)\omega} \tag{31}$$

Let us consider each term of the sum with respect to μ of this equation. For an arbitrarily fixed integer μ ($0 \le \mu < L$), let

$$s := k + \mathsf{LT}(L, \mu+\nu) - 1$$
$$t := \mu + \nu - L\,\mathsf{LT}(L, \mu+\nu)$$

Then, when k moves only once over all integers $\mathbb{Z}$ and ν moves over the integers $0 \le \nu < L$ once independently,

$$sL + t = (k-1)L + \nu + \mu$$

moves once on every member of all integers $\mathbb{Z}$, and t satisfies $0 \le t < L$. Hence, Eq. (31) can be rewritten as

$$= \frac{1}{L}\sum_{\mu=0}^{L-1}\sum_{t=0}^{L-1}\sum_{s\in\mathbb{Z}} \mathsf{E}\Big[e[\mu]e[\mu-(sL+t)]\Big]e^{-j(sL+t)\omega}$$
$$= \frac{1}{L}\sum_{\mu=0}^{L-1}\sum_{\ell\in\mathbb{Z}} \mathsf{E}\Big[e[\mu]e[\mu-\ell]\Big]e^{-j\ell\omega} \tag{32}$$

Extracting the term for $\ell = 0$ by the inverse Fourier transform, we obtain

$$\frac{1}{2\pi L}\int_{-\pi}^{\pi} \mathbf{\Phi}(e^{j\omega})\mathbf{S_{ee}}(e^{jL\omega})\mathbf{\Phi}^{\dagger}(e^{j\omega})d\omega = \frac{1}{L}\sum_{\mu=0}^{L-1}\mathsf{E}\Big[e[\mu]e[\mu]\Big]$$

This coincides with Eq. (29) with $n = 1$. Hence, the lemma is proved. □

Next, let us consider the relationship between the power spectral density matrix $\mathbf{S_{ee}}(e^{jL\omega})$ and the power spectral densities of the input and the quantization noise. Corresponding to Eq. (25), it is found from the general theory that the relationship

$$\mathbf{S_{ee}}(e^{jL\omega}) = \begin{bmatrix}\mathbf{I}_L - \mathbf{G}(e^{jL\omega}) & -\mathbf{G}(e^{jL\omega})\end{bmatrix} \mathbf{S_{pp}}(e^{jL\omega}) \begin{bmatrix}\mathbf{I}_L - \mathbf{G}^{\dagger}(e^{jL\omega}) \\ -\mathbf{G}^{\dagger}(e^{jL\omega})\end{bmatrix} \tag{33}$$

exists between $\mathbf{S_{ee}}(e^{jL\omega})$ and $\mathbf{S_{pp}}(e^{jL\omega})$. Further, from Eq. (26),

$$\mathbf{S_{ee}}(e^{jL\omega}) = (\mathbf{I}_L - \mathbf{G}(e^{jL\omega}))\mathbf{S_{yy}}(e^{jL\omega}) (\mathbf{I}_L - \mathbf{G}^{\dagger}(e^{jL\omega})) + \mathbf{G}(e^{jL\omega})\mathbf{S_{qq}}(e^{jL\omega})\mathbf{G}^{\dagger}(e^{jL\omega}) \tag{34}$$

The first term corresponds to the filtering noise $e_f[m]$ in Eq. (1). The second term corresponds to the response $\tilde{q}[m]$ corresponding to the quantization noise. By using this equation, the integrand of the right-hand side of the equality in Lemma 3.1 can be expanded as follows:

$$\mathbf{\Phi}(e^{j\omega})\mathbf{S_{ee}}(e^{jL\omega})\mathbf{\Phi}^{\dagger}(e^{j\omega})$$
$$= \mathbf{\Phi}(e^{j\omega})\mathbf{S_{yy}}(e^{jL\omega})\mathbf{\Phi}^{\dagger}(e^{j\omega}) \tag{35}$$
$$+\mathbf{\Phi}(e^{j\omega})\mathbf{G}(e^{jL\omega})\mathbf{S_{yy}}(e^{jL\omega}) \mathbf{G}^{\dagger}(e^{jL\omega})\mathbf{\Phi}^{\dagger}(e^{j\omega}) \tag{36}$$
$$-\mathbf{\Phi}(e^{j\omega})\mathbf{G}(e^{jL\omega})\mathbf{S_{yy}}(e^{jL\omega})\mathbf{\Phi}^{\dagger}(e^{j\omega}) \tag{37}$$
$$-\mathbf{\Phi}(e^{j\omega})\mathbf{S_{yy}}(e^{jL\omega})\mathbf{G}^{\dagger}(e^{jL\omega})\mathbf{\Phi}^{\dagger}(e^{j\omega}) \tag{38}$$
$$+\mathbf{\Phi}(e^{j\omega})\mathbf{G}(e^{jL\omega})\mathbf{S_{qq}}(e^{jL\omega}) \mathbf{G}^{\dagger}(e^{jL\omega})\mathbf{\Phi}^{\dagger}(e^{j\omega}) \tag{39}$$

Below, each term is calculated.

Equation (35): With a derivation similar to Eq. (32), it is found that

$$(35) = \sum_{\mu=0}^{L-1}\sum_{\ell\in\mathbb{Z}} \mathsf{E}\Big[y[\mu]y[\mu-\ell]\Big]e^{-j\ell\omega} \tag{40}$$

Since $y[m]$ is a WSS process, each term in the sum over μ is independent of μ and is equal to $S_{yy}(e^{j\omega})$. Hence,

$$(35) = LS_{yy}(e^{j\omega}) \tag{41}$$

Equation (36): After some calculations using Eqs. (21) and (24),

$$\mathbf{\Phi}(e^{j\omega})\mathbf{G}(e^{jL\omega}) = \begin{bmatrix}H(e^{j\omega}) & \mathbf{O}_{1,L-1}\end{bmatrix} \tag{42}$$

is obtained. Thus, the product of the (0, 0) component of $\mathbf{S_{yy}}(e^{j\omega})$ with $|H(e^{j\omega})|^2$ is calculated. We obtain

$$(36) = \frac{|H(e^{j\omega})|^2}{L}\sum_{k\in\mathbb{Z}} \mathsf{E}\Big[y[nL]y[nL-k]\Big]$$

$$\left(\sum_{r=0}^{L-1} e^{\frac{2\pi jkr}{L}}\right) e^{-jk\omega} \tag{43}$$

(Note that the interior of the last parentheses becomes L for $k \in L\mathbb{Z}$ and 0 otherwise. Hence,

$$(36) = \frac{|H(e^{j\omega})|^2}{L} \sum_{r=0}^{L-1} S_{yy}(e^{j\left(\omega - \frac{2\pi jr}{L}\right)}) \tag{44}$$

Equations (37) and (38): From (42), Eq. (38) becomes

$$\begin{aligned}(38) &= -\mathbf{\Phi}(e^{j\omega})\mathbf{S_{yy}}(e^{jL\omega}) \begin{bmatrix} H^*(e^{j\omega}) \\ \mathbf{O}_{L-1,1} \end{bmatrix} \\ &= -H^*(e^{j\omega}) \sum_{k\in\mathbb{Z}} \sum_{\mu=0}^{L-1} \\ &\quad \mathsf{E}\Big[y[nL-\mu]y[nL-\mu-(kL-\mu)]\Big] \\ &\quad e^{-j(kL-\mu)\omega} \\ &= -H^*(e^{j\omega})S_{yy}(e^{j\omega}) \end{aligned} \tag{45}$$

Note that Eq. (37) is its complex conjugate

$$(37) = -H(e^{j\omega})S_{yy}(e^{j\omega}) \tag{46}$$

Equation (39): Similarly to Eq. (36),

$$(39) = \frac{|H(e^{j\omega})|^2}{L} \sum_{r=0}^{L-1} S_{qq}(e^{j\left(\omega - \frac{2\pi jr}{L}\right)}) \tag{47}$$

The results of Eqs. (41), (44), (46), (45), and (47) are substituted and integrated,

$$\mathcal{E} = \mathcal{E}_p + \mathcal{E}_s + \mathcal{E}_q \tag{48}$$

where

$$\mathcal{E}_p := \int_{-\pi}^{\pi} |L - H(e^{j\omega})|^2 S_{yy}(e^{j\omega}) \frac{d\omega}{2\pi L^2}$$

$$\mathcal{E}_s := \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 \sum_{r=1}^{L-1} S_{yy}(e^{j\left(\omega - \frac{2\pi jr}{L}\right)}) \frac{d\omega}{2\pi L^2}$$

$$\mathcal{E}_q := \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 \sum_{r=0}^{L-1} S_{qq}(e^{j\left(\omega - \frac{2\pi jr}{L}\right)}) \frac{d\omega}{2\pi L^2}$$

Further, if Eq. (4) is used, we obtain

$$\mathcal{E}_p = \frac{\sigma_w^2}{2\pi L^2} \int_{-\frac{\alpha\pi}{L}}^{\frac{\alpha\pi}{L}} |L - H(e^{j\omega})|^2 d\omega \tag{49}$$

and

$$\begin{aligned}\mathcal{E}_s &= \sum_{r=1}^{L-1} \int_0^{2\pi} S_{yy}(e^{j\left(\omega - \frac{2\pi jr}{L}\right)}) \frac{d\omega}{2\pi L^2} \\ &= \frac{\sigma_w^2}{2\pi L^2} \sum_{r=1}^{L-1} \int_{2\pi\frac{r-\alpha/2}{L}}^{2\pi\frac{r+\alpha/2}{L}} |H(e^{j\omega})|^2 d\omega \end{aligned} \tag{50}$$

In the calculation of $\mathcal{E}_s$, the 2π periodicity of $|H(e^{j\omega})|$ and the exclusiveness of the support of $S_{yy}(e^{j(\omega - 2\pi jr/L)}$ $(1 \le r \le L-1)$ are used. These $\mathcal{E}_p$ and $\mathcal{E}_s$ denote the filtering noises in the passband and the stopband. Similarly, $\mathcal{E}_q$ is found from Eq. (7):

$$\mathcal{E}_q = \frac{\sigma_q^2}{2\pi L} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 d\omega \tag{51}$$

The energy of the reference input signal $y[m]$ per sample is

$$\begin{aligned}\mathcal{S} := \mathsf{E}\Big[y[m]^2\Big] &= \int_{-\pi}^{\pi} S_{yy}(e^{j\omega}) \frac{d\omega}{2\pi} \\ &= \frac{\sigma_w^2}{2\pi} \int_{-\frac{\alpha\pi}{L}}^{\frac{\alpha\pi}{L}} d\omega = \frac{\alpha\sigma_w^2}{L} \end{aligned}$$

Combining the above, the representation of the SNR in the frequency domain can be written as

$$\mathrm{SNR} = \frac{\mathcal{S}}{\mathcal{E}} = \frac{\mathcal{S}}{\mathcal{E}_p + \mathcal{E}_s + \mathcal{E}_q}$$

By normalization

$$\begin{bmatrix} S \\ E_p \\ E_s \\ E_q \end{bmatrix} := \frac{\pi L^2}{\sigma_w^2} \begin{bmatrix} \mathcal{S} \\ \mathcal{E}_p \\ \mathcal{E}_s \\ \mathcal{E}_q \end{bmatrix}$$

to $\mathcal{S}$, $\mathcal{E}_p$, $\mathcal{E}_s$, and $\mathcal{E}_q$ simpler expressions can be obtained.

The theoretical SNR equation after normalization is

$$\mathrm{SNR} = \frac{S}{E_p + E_s + E_q} \tag{52}$$

Here,

$$S = \alpha\pi L \tag{53}$$

$$E_p = \int_0^{\frac{\alpha\pi}{L}} |L - H(e^{j\omega})|^2 d\omega \tag{54}$$

$$E_s = \frac{1}{2} \sum_{r=1}^{L-1} \int_{2\pi\frac{r-\alpha/2}{L}}^{2\pi\frac{r+\alpha/2}{L}} |H(e^{j\omega})|^2 d\omega \tag{55}$$

$$E_q = L\frac{\sigma_q^2}{\sigma_w^2} \int_0^{\pi} |H(e^{j\omega})|^2 d\omega \tag{56}$$

Further, if it is assumed that the white noise $w[m]$ as the basis for the reference input signal $y[m]$ takes values in the range of (–1, 1), then by Eq. (8) E_q can be written specifically with bit length b of the fractional part as

$$E_q = L2^{-2b-2} \int_0^{\pi} |H(e^{j\omega})|^2 d\omega \tag{57}$$

It is assumed that this assumption is valid up to the end of this section.

Let us consider an ideal filter maximizing the theoretical SNR equation derived above and the resultant upper limit of the SNR. When the band-limiting coefficient α and the interpolation ratio L are given, the maximization of the SNR is reduced to the minimization of the denominator of Eq. (52):

$$E := E_p + E_s + E_q \tag{58}$$

The integration range of Eq. (57) is divided into $[0, \alpha\pi/L)$ and $[\alpha\pi/L, \pi)$. Considering it together with the integration range in Eq. (55), it is immediately found that the best choice of $H(e^{j\omega})$ in the range $[\alpha\pi, 2\pi - \alpha\pi/L)$ is to be identically zero. Then, E is

$$\int_0^{\frac{\alpha\pi}{L}} |L - H(e^{j\omega})|^2 + L2^{-2b-2}|H(e^{j\omega})|^2 d\omega \tag{59}$$

By some calculations, it is found that the best choice is obtained if $H(e^{j\omega})$ is to be identically $2^{2b+2}L/(2^{2b+2} + L)$ in this integration range. In total, the frequency characteristic of the ideal filter in the range of $[0, \pi)$ is

$$H_{\mathrm{ideal}}(e^{j\omega}) = \begin{cases} \frac{2^{2b+2}L}{2^{2b+2}+L} & (\omega \in [0, \alpha\pi/L)) \\ 0 & (\omega \in [\alpha\pi/L, \pi)) \end{cases} \tag{60}$$

Hence, the ideal filter passes the range within the input band limitation at a gain of $2^{2b+2}L/(2^{2b+2} + L)$ and completely suppresses the image and the quantization noise outside this bandwidth. Due to the band limiting coefficient α, this characteristic is different from that of the ideal L-th band filter in the sense of equal division of the band. This ideal filter H_{ideal} provides the upper limit of the SNR:

$$SNR_{\max} = 1 + L^{-1}2^{2b+2} \tag{61}$$

4. Verification of the Theoretical SNR Equation by Simulation

To verify validity of the theoretical SNR equation in the frequency domain derived in the previous section, it is compared with the numerical simulation of the SNR definition in the time domain defined by Eq. (9) in Section 2.

The simulation method is as follows. In the following real value operations, the floating point representation is used.

(i) Simulation of the reference input signal

By using a pseudorandom number generator MT19937 [2], a pseudo white noise sequence $w'[m]$ ($0 \le m < M$) that takes values in the range of (–1, 1) is generated. The sequence length is chosen as $M = 2^{14} = 16{,}384$ so that it is sufficiently large. By means of the M-point DFT, $w'[m]$ is band-limited to $[0, \alpha\pi/L)$ and is used as the pseudo reference input $y'[m]$.

(ii) Quantization

By discarding 0s and including 1s, $y'[m]$ is rounded so that its fractional part has b bits.

(iii) Down-sampling and interpolation process

For the output of (ii), the down-sampling and the up-sampling are applied according to the definition. Then filtering is carried out with $H(z)$ realized by floating point numbers, to generate the interpolation output $\hat{y}'[m]$ ($0 \le m < M$).

(iv) Calculation of square sums of signal and error

According to

$$\mathcal{S}' := \sum_{m=0}^{M-1} \hat{y}'[m]^2, \quad \mathcal{E}' := \sum_{m=0}^{M-1} \left(\hat{y}'[m] - y'[m]\right)^2$$

the square sum of the signals and the square sum of the errors are calculated for this sequence.

(v) Calculation of SNR

In regard to different 10,000 $w'[m]$ sequences, the sums of $\mathcal{S}'$ and $\mathcal{E}'$ in (iv) are taken. Let the results be $\mathcal{S}''$ and

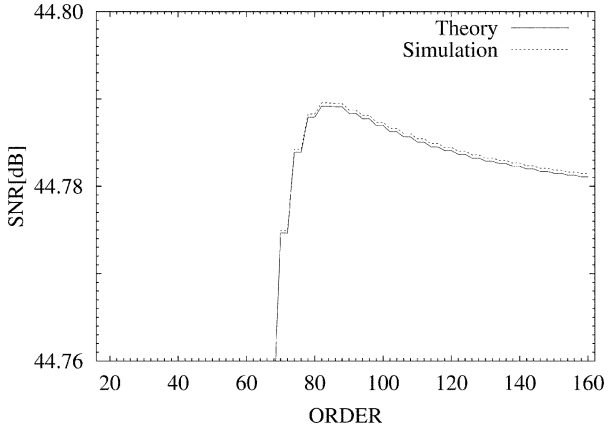


Fig. 8. SNR: theory versus simulation.

$\mathcal{E}''$, respectively. Then let the ratio $\mathcal{S}''/\mathcal{E}''$ be the SNR simulation value $\mathrm{SNR}_{\mathrm{sim}}$.

For an interpolation ratio $L = 2$, a band limiting coefficient $\alpha = 0.9$, and a number of quantization bits $b = 7$, Fig. 8 shows the variations of the theoretical SNR values and the simulated SNR values versus the order of the $H(z)$ of the interpolation filter. In the figure, *Theory* indicates the decibel representation of the theoretical value of the SNR, $10\log_{10}$ SNR. *Simulation* indicates the decibel expression of the simulation results, $10\log_{10}\mathrm{SNR}_{\mathrm{sim}}$. Both are in good agreement. Here an equal-ripple FIR half-band filter is used as $H(z)$. The theoretical SNR values are calculated by the theoretical SNR equation (52) from the amplitude characteristics.

5. Design of Interpolation SNR Maximizing FIR Filter

In this section, based on the theoretical SNR equation presented in Section 3, a design method for a linear phase FIR filter (Type I) is proposed for which the interpolation output SNR is the maximum. First, in Section 5.1, a design method is described for maximizing the SNR without the restriction of the L-th band filter, and then, in Section 5.2, the SNR maximizing design under the constraint of the L-th band filter. It is natural that the former filter is superior to the latter in SNR for the same filter order. On the other hand, the latter has the advantage that the number of multipliers is smaller (about one-half when $L = 2$).

5.1. Least square design of maximum SNR interpolation filter

In Section 3, Eq. (52) is derived, by which the SNR is obtained from the interpolation ratio L, the quantization bit number b of the input signal, the band limitation coefficient α, and the amplitude characteristics of the filter. As described toward the end of Section 3, the SNR maximization problem is reduced to that of minimization of Eq. (58):

$$E = E_p + E_s + E_q$$

when L and α are given. Here, E_p is the passband filtering noise, E_s is the stopband filtering noise, and E_q is the energy for the response of the interpolation filter for a quantization noise. They are given by Eqs. (54), (55), and (56). All of E_p, E_s, and E_q are in the form of squared amplitude errors. The error function E is their sum.

In this paper, E is used as the objective function and a design method is proposed for derivation of the filter minimizing E or maximizing the SNR by the method of least squares. It is assumed that the interpolation filter $H(z)$ is a linear phase Type I FIR filter. In order to express its frequency characteristics, let us define the vector $\boldsymbol{c}$ and the coefficient vector $\boldsymbol{a}$ as follows:

$$\boldsymbol{a} = \begin{bmatrix} a_0 & a_1 & \cdots & a_N \end{bmatrix}^T \tag{62}$$

$$\boldsymbol{c} = \begin{bmatrix} 1 & \cos\omega & \cdots & \cos N\omega \end{bmatrix}^T \tag{63}$$

Here,

$$a_m = \begin{cases} h[N] & (m = 0) \\ 2h[N+m] & (m = 1, 2, \cdots, N) \end{cases} \tag{64}$$

Using $\boldsymbol{a}$ and $\boldsymbol{c}$, the frequency characteristic $H_0(e^{j\omega})$ of the zero-phased version $H_0(z)$ of the interpolation filter $H(z)$ is given by

$$\begin{aligned} H_0(e^{j\omega}) &= a_0 + \sum_{m=1}^{N} a_m \cos\omega m \\ &= \boldsymbol{a}^T\boldsymbol{c} = \boldsymbol{c}^T\boldsymbol{a} \end{aligned} \tag{65}$$

$$\left|H_0(e^{j\omega})\right|^2 = \boldsymbol{a}^T\boldsymbol{c}\boldsymbol{c}^T\boldsymbol{a} \tag{66}$$

When these are substituted into $H(e^{j\omega})$ and $|H(e^{j\omega})|^2$ of E_p, E_s, and E_q in Eqs. (54), (55), and (56),

$$\begin{aligned} E_p &= \int_0^{\frac{\pi}{L}\alpha} \left|L - H_0(e^{j\omega})\right|^2 d\omega \\ &= L^2\int_0^{\frac{\pi}{L}\alpha} d\omega - 2\left(L\int_0^{\frac{\pi}{L}\alpha} \boldsymbol{c}^T d\omega\right)\boldsymbol{a} \\ &\quad + \boldsymbol{a}^T\left(\int_0^{\frac{\pi}{L}\alpha} \boldsymbol{c}\boldsymbol{c}^T d\omega\right)\boldsymbol{a} \end{aligned} \tag{67}$$

$$\begin{aligned} E_s &= \frac{1}{2}\sum_{r=1}^{L-1}\int_{\frac{\pi}{L}(2r-\alpha)}^{\frac{\pi}{L}(2r+\alpha)} \left|H_0(e^{j\omega})\right|^2 d\omega \\ &= \boldsymbol{a}^T\left(\frac{1}{2}\sum_{r=1}^{L-1}\int_{\frac{\pi}{L}(2r-\alpha)}^{\frac{\pi}{L}(2r+\alpha)} \boldsymbol{c}\boldsymbol{c}^T d\omega\right)\boldsymbol{a} \end{aligned} \tag{68}$$

$$\begin{aligned} E_q &= L\frac{\sigma_q^2}{\sigma_w^2}\int_0^{\pi} \left|H_0(e^{j\omega})\right|^2 d\omega \\ &= \boldsymbol{a}^T\left(L\frac{\sigma_q^2}{\sigma_w^2}\int_0^{\pi} \boldsymbol{c}\boldsymbol{c}^T d\omega\right)\boldsymbol{a} \end{aligned} \tag{69}$$

The reason for using the zero phase $H_0(z)$ in place of $H(z)$ is for correction of the N group delay in the output of the actual interpolation filter $H(z)$ in comparison to the reference input signal $y[m]$ in the error evaluation system in Fig. 4.

From Eqs. (67), (68), and (69), it is found that Eq. (58) for the error function can be written in the form

$$E = \boldsymbol{a}^T\boldsymbol{F}\boldsymbol{a} - 2\boldsymbol{g}^T\boldsymbol{a} \tag{70}$$

Here $\boldsymbol{F}$ and $\boldsymbol{g}$ are as follows:

$$\boldsymbol{F} = \int_0^{\frac{\pi}{L}\alpha} \boldsymbol{c}\boldsymbol{c}^T d\omega + \frac{1}{2}\sum_{r=1}^{L-1} \int_{\frac{\pi}{L}(2r-\alpha)}^{\frac{\pi}{L}(2r+\alpha)} \boldsymbol{c}\boldsymbol{c}^T d\omega + L\frac{\sigma_q^2}{\sigma_w^2}\int_0^{\pi} \boldsymbol{c}\boldsymbol{c}^T d\omega \quad (71)$$

$$\boldsymbol{g} = L\int_0^{\frac{\pi}{L}\alpha} \boldsymbol{c}\, d\omega \quad (72)$$

The k-th column l-th row (k, l = 0, 1, . . . , N) component $F_{k,l}$ of the matrix $\boldsymbol{F}$ and the k-th column (k = 0, 1, . . . , N) component g_k of the vector $\boldsymbol{g}$ can be calculated analytically as shown in Table 1. Note that under the assumption that the white noise $w[m]$ as the source of forming the reference input signal $y[m]$ takes a real value of (–1, 1) the following specific form can be found from Eq. (8):

$$F_{0,0} = \pi\left(\alpha + \frac{L}{2^{2b+2}}\right) \quad (73)$$

The matrix $\boldsymbol{F}$ is a symmetric positive-definite matrix. Therefore, when the error function E is the minimum, the following relationship exists for $\boldsymbol{F}$, $\boldsymbol{g}$, and $\boldsymbol{a}$ [11]:

$$\boldsymbol{F}\boldsymbol{a} = \boldsymbol{g} \quad (74)$$

By solving linear equation (74), the filter coefficient vector $\boldsymbol{a}$ minimizing the error function E can be derived.

In summary, the design specifications in the proposed method are

- the interpolation ratio L
- the input signal band limiting coefficient α
- the quantization bit number below the decimal point of the input signal b
- the filter order $2N$

and plugging them into Eqs. (71) and (72) and then solving the linear Eq. (74), the filter coefficients maximizing the SNR can be obtained.

5.2. SNR maximizing design under the L-th band restriction

Since about half the coefficients are 0 in an FIR half-band filter, this filter can be realized with about half as many multipliers as the general FIR filter. Here, a modified design method is presented in which the restriction of the L-th band filter is added to the design method in the previous section. The impulse response of the L-th band filter $h[m]$ is restricted as follows:

$$h[N + rL] = 0 \qquad (r = \pm 1, \pm 2, \cdots) \quad (75)$$

With regard to the center value $h[N]$ of the impulse response, no restriction is imposed for SNR optimization.

From restriction (75), the rL-row component $a_{(rL)}$ of the coefficient vector $\boldsymbol{a}$ given in Eq. (62) is restricted as

$$a_{(rL)} = 0 \qquad (r = 1, 2, 3, \cdots) \quad (76)$$

On the other hand, for the rL-row component of $\boldsymbol{g}$ in Eq. (72),

$$g_{(rL)} = 0 \qquad (r = 1, 2, 3, \cdots) \quad (77)$$

holds automatically. Hence, for the vector $\boldsymbol{a}$ in Eq. (62), let $\boldsymbol{a}'$ be the vector obtained by eliminating the rL-column:

Table 1. Elements of $\boldsymbol{F}$ and $\boldsymbol{g}$

$$F_{k,l} = \begin{cases} \frac{1}{2}\left[\frac{1}{k+l}\left\{1 + \sum_{r=1}^{L-1}\cos\frac{2(k+l)r\pi}{L}\right\}\left\{\sin\frac{(k+l)\alpha\pi}{L}\right\} + \frac{1}{k-l}\left\{1 + \sum_{r=1}^{L-1}\cos\frac{2(k-l)r\pi}{L}\right\}\left\{\sin\frac{(k-l)\alpha\pi}{L}\right\}\right] & (k \neq l) \\ \frac{1}{2}\left[\frac{1}{2k}\left\{1 + \sum_{r=1}^{L-1}\cos\frac{4kr\pi}{L}\right\}\left\{\sin\frac{2k\alpha\pi}{L}\right\} + F_{0,0}\right] & (k = l, k \neq 0, l \neq 0) \\ \pi\left(\alpha + L\frac{\sigma_q^2}{\sigma_w^2}\right) & (k = 0, l = 0) \end{cases}$$

$$g_k = \begin{cases} \frac{L}{k}\sin\frac{k\alpha\pi}{L} & (k \neq 0) \\ \alpha\pi & (k = 0) \end{cases}$$

$$\begin{bmatrix} a_0 & a_1 & \cdots & a_{(rL-1)} & a_{(rL+1)} & \cdots & a_N \end{bmatrix}^T$$

For $\boldsymbol{g}$, let us similarly define $\boldsymbol{g}'$. Also, let the matrix $\boldsymbol{F}'$ be the matrix obtained by eliminating the rL-th column and the rL-th row from $\boldsymbol{F}$ in Eq. (71). By solving the linear equation

$$\boldsymbol{F}'\boldsymbol{a}' = \boldsymbol{g}' \tag{78}$$

the coefficient vector $\boldsymbol{a}'$ of the L-th band filter maximizing the SNR can be derived.

6. SNR Performance Comparison and Design Examples

In this section, the SNR characteristics are compared between the SNR maximized interpolation filter proposed in the previous section and the interpolation filter by the conventional method. Several examples of design by the proposed method are also presented.

6.1. SNR performance comparison

Under the conditions of an interpolation ratio $L = 2$, a band limiting coefficient $\alpha = 0.9$, and a number of quantization bits $b = 7$, the SNR maximized interpolation filter designed by the method in Section 5.1, the SNR maximized half-band interpolation filter designed by the method in Section 5.2, and that designed by the conventional method are compared for each filter order. The filter designed by the conventional method is designed by seeking the passband edge frequency of the equal ripple filter that maximizes the SNR for each order. (The design needs to be repeated by changing the passband edge frequency. This is one of the shortcomings of the conventional design specifying the passband and stopband.) The SNR performance comparison is shown in Fig. 9. The SNR is computed from the theoretical SNR equation derived in Section 3 by using the amplitude characteristics of the interpolation filter. By applying 10 $\log_{10}$ operation, the decibel expression is obtained. The horizontal axis indicates the filter order. Proposed, Proposed (halfband), and Conventional indicate the SNR maximized filter in Section 5.1, the SNR maximized half-band filter in Section 5.2, and the conventional filter. Also, $SNR_{\max}$ is the upper limit of the SNR given by Eq. (61). It is first clear that SNR is not necessarily improved in the case of Conventional as the order is increased. When the order becomes about 80th or higher, the SNR is stationary or even decreases slightly. On the other hand, the SNR of Proposed is found to approach $SNR_{\max}$ as the order is increased. It is noted that the SNR of Proposed at the order of 80 is better than that of Conventional at 160. Since the increase in the order does not necessarily lead to improvement of the SNR in Conventional, even if comparisons between Proposed and Conventional are made with the same number of multipliers, their difference increases with an increasing number of multipliers. The SNR of Proposed (halfband) is slightly better than that of Conventional up to an order of about 100. The SNR is improved at higher orders. However, the improvement is slower than Proposed due to the half-band limitation.

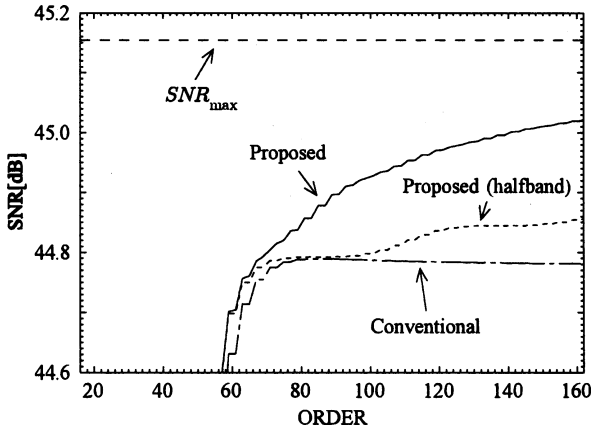


Fig. 9. SNR comparison.

The output noise for each order is divided into the quantization noise E_q, the passband filtering noise E_p, and the stopband filtering noise E_s, which are then compared in Figs. 10, 11, and 12. First, it should be noted that E_q is overwhelmingly predominant over E_p and E_s in terms of magnitude at orders larger than 40.

In the case of Conventional, the ripples in the passband and the stopband decrease as the order becomes higher, so that the filtering noises E_p and E_s are improved up to an order of about 100. On the other hand, the quanti-

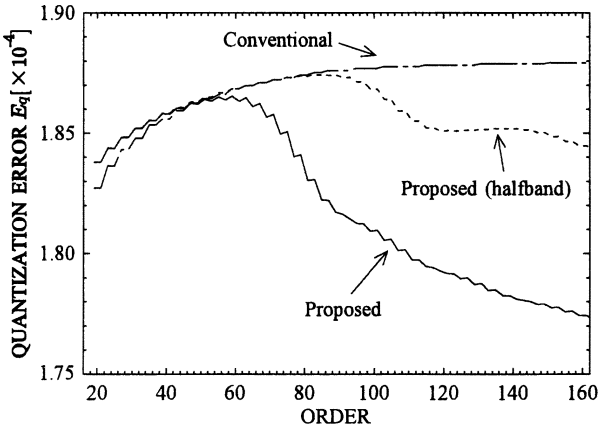


Fig. 10. Order versus quantization noise E_q.

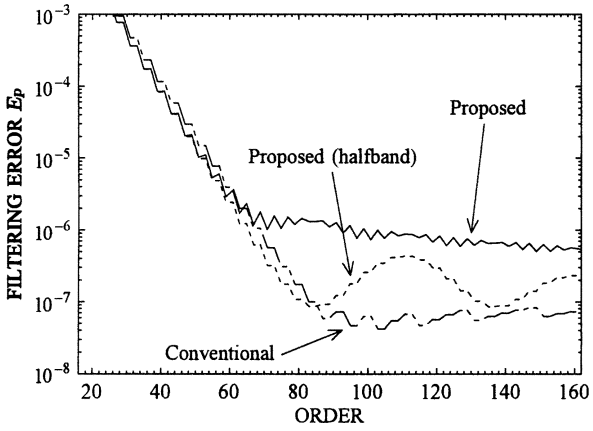


Fig. 11. Order versus passband-filtering noise E_p.

zation noise E_q tends to increase with the filter order. The amplitude characteristics of the filter designed by the conventional method in the transition region become sharper as the order becomes higher. This is not a suitable characteristic to reduce the quantization noise in $[\alpha\pi/2, \pi/2)$.

In Proposed, the improvement of E_q, E_p, and E_s is similar to that in Conventional for low orders (up to about 60). At higher orders, the quantization noise E_q is decreased more than the filtering errors E_p and E_s. The passband filtering error E_p of Proposed is worse than that in the conventional method at orders of more than 60. However, the quantization error E_q, which has a larger absolute value, is more significantly reduced.

In Proposed (halfband), with a long-period undulation as the order is increased, a characteristic that can be considered to be intermediate between that of Proposed and Conventional, is observed.

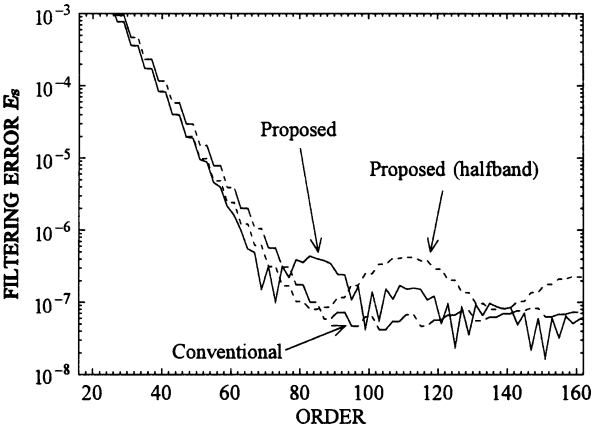


Fig. 12. Order versus stopband-filtering noise E_s.

6.2. Design example with an interpolation ratio of $L = 2$

The characteristics of the amplitude response of the interpolation filter designed by the proposed method are presented through a design example. The conditions of an interpolation ratio $L = 2$, a band limiting coefficient $\alpha = 0.9$, and a number of quantization bits $b = 7$ are identical to those in Section 6.1.

Figure 13 shows a comparison of the amplitude characteristics of the SNR maximizing filter (Proposed) in Section 5.1, the SNR maximizing half-band filter [Proposed (halfband)] in Section 5.2, and the conventional method (Conventional) when the filter order is 160. The horizontal axis is the normalized frequency $\omega/2\pi$. The two edges of the bandwidth $[\alpha/4, 1/2 - \alpha/4)$ in which only the quantization noise exists are indicated by notches on both sides of 1/4. This bandwidth is the transition region for Conventional without control of the amplitude. However, it is found in Proposed that the gain approaches zero while ripples exist. In Proposed (halfband), the gain in this bandwidth is kept small under the half-band restriction. The amplitude characteristics of Proposed and Proposed (halfband) for filter orders of 40, 80, and 160 are presented in Figs. 14 and 15. From Fig. 14, it is seen that the amplitude of Proposed becomes sharper near $\omega = \alpha\pi/2$ as the order is increased and exhibits a behavior approaching the ideal amplitude characteristic (60) in Section 3 (in some sense). Also, in Fig. 15, the amplitude characteristic becomes sharper with increasing order near $\omega = \alpha\pi/2$. However, due

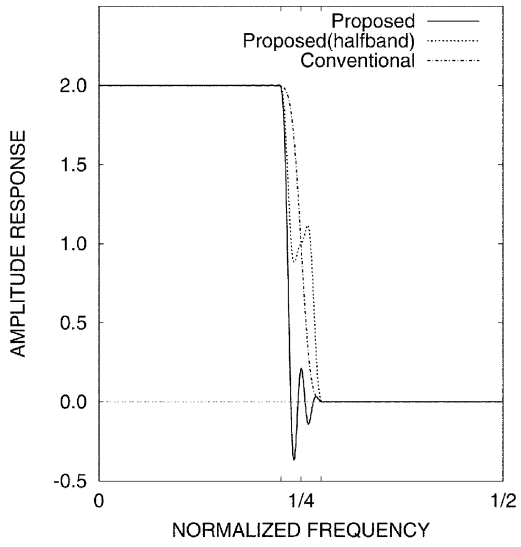


Fig. 13. Comparison of amplitude responses at order 160.

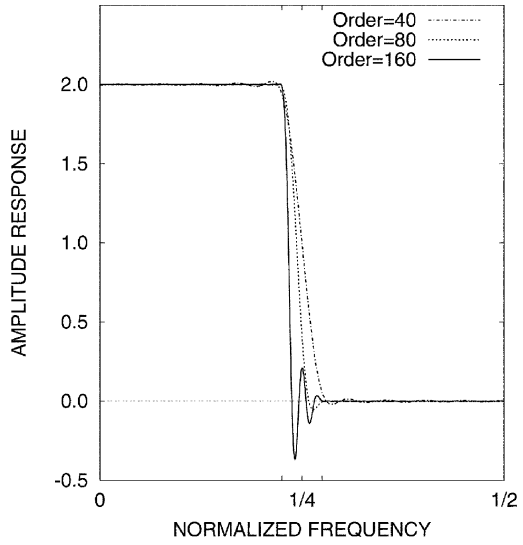


Fig. 14. Amplitude responses for order 40, 80, 160: Proposed.

to the restriction of the half band, the amplitude undulates around 1 in the frequency region in which only quantization noise exists.

6.3. Design example with interpolation ratio $L = 5$

The proposed method can be applied at an arbitrary value of L. However, in this section, only one example is presented, under conditions of an interpolation ratio $L = 5$, a band limiting coefficient $\alpha = 0.8$, and a number of quantization bits $b = 7$. Figure 16 shows the amplitude

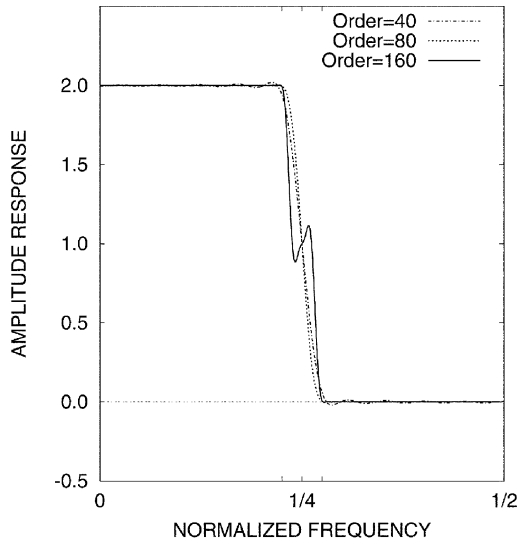


Fig. 15. Amplitude responses for order 40, 80, 160: Proposed (halfband).

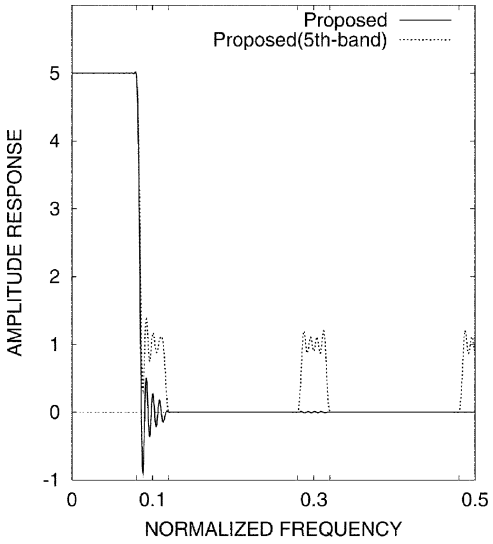


Fig. 16. Amplitude responses of 5-interpolation filters.

characteristics of the SNR maximized filter (Proposed) and the SNR maximized 5th-band filter [Proposed (5th-band)] with an order of 298. The frequency regions in which only quantization noise exists are [0.08, 0.12), [0.28, 0.32), and [0.48, 0.5) in terms of the normalized frequency $\omega/2\pi$. In Proposed, the amplitude is close to 0, but there are some ripples in these bands. In Proposed (5th-band), the amplitude undulates around 1 in these bands.

7. Conclusions

The output SNR of an L-interpolation filter for a quantized band-limited signal is analyzed. A theoretical equation for the SNR described in the frequency characteristic of the interpolation filter is derived. Its validity is confirmed by simulations. Further, a design method is proposed for an interpolation filter maximizing the SNR. The design methods are for SNR maximization within the Type I FIR filter and for SNR maximization with restriction of the L-th band filter. Each problem is reduced to solving a system of linear equations with the coefficient matrix expressed analytically. Hence, design for an arbitrary interpolation ratio L is easily carried out. In the proposed filter, an SNR not attainable by the conventional filter considering only the passband and the stopband, even with sacrifice of the order, is now attained by reducing the gain in the regions where only quantization noise exists.

REFERENCES

1. Kiya H. Multi-rate signal processing. Shoko-do; 1995.

2. Matsumoto M, Nishimura T. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans Model Comput Simul 1998;8:3–30.
3. McClellan JH, Parks TW, Rabiner LR. A computer program for designing optimum FIR linear phase digital filters. IEEE Trans Audio Electroacoust 1973;21:506–526.
4. Sagara I. Introduction to AD/DA converters. Nikkan-Kogyo-Shimbun; 1991.
5. Sathe VP, Vaidyanathan PP. Effects of multirate systems on the statistical properties of random signals. IEEE Trans Signal Process 1993;41:131–146.
6. Sato H, Yoshikawa T. Proposal and evaluation of an FIR half band filter suitable for quantized inputs. Tech Rep IEICE 1994;CAS94-59.
7. Takebe K. Design of digital filters. Tokai University Press; 1988.
8. Tuqan J, Vaidyanathan PP. Oversampling PCM techniques and optimum noise shapers for quantizing a class of nonbandlimited signals. IEEE Trans Signal Process 1999;47:389–407.
9. Vaidyanathan PP, Nguyen TQ. A "trick" for the design of FIR half-band filters. IEEE Trans Circuits Syst 1987;34:297–300.
10. Watanabe Y, Yoshikawa T. A proposal on the configuration method of an interpolation filter suitable for band limited signals. Tech Rep IEICE 1992;CAS92-72.
11. Zhang X, Iwakura H. Design of a linear phase FIR filter with a least square approximation with constraints. Bulletin of the University of Electro-Communications 1991;4:225–233.

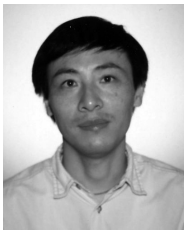
AUTHORS (from left to right)



Yoshinori Takei (member) graduated from the Department of Mathematics, Tokyo Institute of Technology, in 1990 and completed the M.S. program in 1992, receiving an M.S. degree in mathematics. In 2000, he completed the doctoral program in information processing, receiving a D.Eng. degree. From 1992 through 1995, he was with Kawasaki Steel R&D Inc. From 1999 to 2000, he was an assistant professor at Tokyo Institute of Technology. He moved to Nagaoka University of Technology in 2000, and is now an associate professor. He has been engaged in research on theoretical computer science and digital signal processing. He is a member of LA, SIAM, ACM, AMS, and IEEE.

Kouichi Mogi (member) graduated from the Department of Electrical and Electronic Systems Engineering, Nagaoka University of Technology, in 2001 and completed the M.S. program in 2003, receiving an M.S. degree in engineering. His student research dealt with digital signal processing. Since 2003, he has been engaged in system design of construction machines at Nippon Seiki.

Toshinori Yoshikawa (member) graduated from the Department of Electronic Engineering, Tokyo Institute of Technology, in 1971 and completed the doctoral program in 1976, receiving a D.Eng. degree. He became a research associate on the Faculty of Engineering, Saitama University. After serving as a lecturer, he became an associate professor at Nagaoka University of Technology in 1983. He is now a professor. His research concerns software applications for computers. He is a member of IEEE.



Xi Zhang (member) graduated from the Department of Electronic Engineering, Nanjing University of Aeronautics and Astronautics (China) in 1984. In 1993, he completed the doctoral program in information and communication engineering at the University of Electro-Communications, receiving a D.Eng. degree. In 1984, he became an assistant professor at Nanjing University of Aeronautics and Astronautics. He moved to the University of Electro-Communications in 1993. He became an associate professor at Nagaoka University of Technology in 1996. He is now an associate professor at the University of Electro-Communications. He was a visiting scholar supported by the Ministry of Education of Japan at MIT in 2000–2001. He received the third prize of the Science and Technology Progress Award of China in 1987, and the challenge prize of the 4th LSI IP Design Award of Japan in 2002. He served as an Associate Editor for *IEEE Signal Processing Letters* from 2002 to 2004. He has been engaged in research on digital signal processing, image processing, filter design theory, approximation theory, and wavelet and image compression. He is a senior member of IEEE.