Word Length Condition for DC Lossless DWT

Masahiro Iwahashi* and Hitoshi Kiya†
*Nagaoka University of Technology, Nagaoka 940-2188 Niigata Japan
E-mail: iwahashi@vos.nagaokaut.ac.jp Tel: +81-258-46-6000
†Tokyo Metropolitan University, Hino 191-0065 Tokyo Japan
E-mail: kiya@eei.metro-u.ac.jp Tel: +81-42-585-8603

***Abstract*— This report theoretically analyzes the condition on word length of coefficients and signals such that the discrete wavelet transform (DWT) becomes DC lossless. The DWT discussed here is irreversible for an arbitrary input signal. However, it becomes lossless for a constant valued (DC) input signal under the condition. In conventional approaches, error due to shortening of word length of signals (signal error) and that of coefficients (coefficient error) are treated as additive and multiplicative, respectively. In this report, we introduce a new model which shifts the coefficient error to the signal error in order to treat them as additive. Furthermore, utilizing the fact that the accumulated error inside the circuit is nullified by the rounding at its output, we derive the condition for the DC lossless DWT. Theoretical bound of the word length is derived and the minimum word length is found to be 14 [bit] for 8 [bit] input signals.**

I. INTRODUCTION

Recently the JPEG 2000 based on the discrete wavelet transform (DWT) was adopted as an international standard for digital cinema [1,2]. In the DWT circuit, all of signal values and coefficient values are expressed with finite word length. It contributes to high speed and low power implementation to shorten the word length [3,4]. However, the DWT is designed under the assumption that the word length is infinite. Therefore, it is inevitable to have loss due to shortening the word length in output signals of the DWT circuit.

The lifting structure has been widely developed since it can cancel the loss between the forward transform and the backward transform [5,6]. In the JPEG 2000, the reversible 5-3 DWT and the irreversible 9-7 DWT are utilized for lossless coding and lossy coding respectively [1]. The 9-7 DWT can achieve high performance lossy coding. However, it can't be lossless because of scaling for adjustment of signal gain [7].

Constructing the scaling with the lifting structure, a reversible 9-7 DWT is proposed [8]. However, its performance in lossless coding and in lossy coding is inferior to that of the reversible 5-3 DWT and the irreversible 9-7 DWT respectively.

In this report, we theoretically analyze the condition on word length of coefficient values and signal values such that the irreversible 9-7 DWT becomes DC lossless. Under this condition, output signals of the DWT contain no loss for a constant valued (DC) input signal. This DC lossless property is considered to be effective for the white balancing of a video system [9,10].

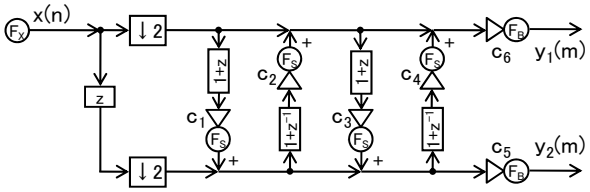
In conventional approaches, the error due to shortening the word length of signals (signal error) is described as additive to the signal [7]. It is treated as an independent and uniformly distributed white signal. On the other hand, the error of coefficients (coefficient error) is described as multiplicative to the signal and evaluated by the sensitivity [11]. However, both of them have been treated independently and their mutual effect has not been well studied.

In this report, we introduce a new model which shifts the coefficient error to the signal error in order to treat them as additive. As a result, their mutual effect is taken into account. Furthermore, utilizing the fact that the accumulated error inside the circuit is nullified by the rounding at its output, we derive the condition for the DC lossless DWT. As a result, theoretical bound is derived as a function of the word length of signals and coefficients. Defining a cost function, we also find the minimum word length under the condition.

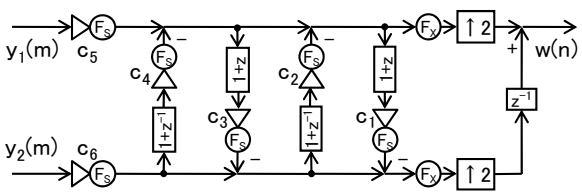
II. DC LOSSLESS DWT AND ITS WORD LENGTH

A. *Irreversible 9-7 DWT*

Fig.1 illustrates the irreversible 9-7 DWT in the JPEG 2000 [1]. The input signal $x(n)$, $n=\{1,2,\cdots,N\}$ is transformed to the band signals $y_1(m)$ and $y_2(m)$, $m=\{1,2,\cdots,N/2\}$. These are backward transformed to reconstruct the signal $w(n)$. In the figure, z^{-1} and $\downarrow 2$ indicate the delay and the down sampler respectively. The coefficients c_i, $i \in \{1,2,\cdots,6\}$ of multipliers are designed as real numbers. In implementation, their word lengths are shortened. The fraction part of a signal value is also truncated to F_S, F_B or F_X [bit] by the rounding operation illustrated as a circle in the figure.



(a) Forward transform



(b) Backward transform

Fig.1 Irreversible 9-7 DWT.

B. Word Length and Rounding Error

In this report, we use the rounding operation defined by

$$\begin{cases} R_0[s] = \lfloor s' \rfloor = s' - (s' \mod 1) \in \mathbf{Z}, \ s' = s + 2^{-1} \\ R_{F_S}[s] = R_0[s2^{F_S}]2^{-F_S}, \ 0 \le F_S \in \mathbf{Z} \end{cases} \quad (1)$$

as an example. It shortens the fraction part of the word length of a signal value s into F_S [bit]. It also generates the error:

$$\Delta_{F_S}[s] = s - R_{F_S}[s]. \quad (2)$$

Denoting the integer part as I_S [bit], the word length W_S [bit] of a signal s is defined by

$$W_S = I_S + F_S + 1 \ \text{[bit]} \quad (3)$$

including 1 [bit] for the sign part. Similarly, the word length W_C [bit] of a coefficient c is defined by

$$W_C = I_C + F_C + 1 \ \text{[bit]}. \quad (4)$$

Especially, in this report, we utilize the property [12]:

$$\begin{cases} R_{F_S}[s]2^{F_S} = p \in \mathbf{Z} \ , \ \left|\Delta_{F_S}[s]\right| \le 2^{-1-F_S} \\ \left|R_{F_S}[s]2^{F_S}\right| \le p \Leftrightarrow |s| < (p + 2^{-1})2^{-F_S} \\ s2^{F_S} \in \mathbf{Z} \Rightarrow R_{F_S}[s+t] = s + R_{F_S}[t] \end{cases} \quad (5)$$

to analyze the condition on the word length for DC lossless.

C. DC Lossless DWT

In a video system, an input signal is processed through a camera, a pair of an encoder and a decoder, and a display. When the camera and the display are adjusted, a white balancing signal which is a constant valued signal (DC signal) is commonly used [10]. In this case, it is desirable that the encoder and the decoder do not generate any loss.

In this report, we define the loss as the difference between the output signal of the DWT with infinite word length and that of the DWT with shortened word length. For DC input, when the output of the backward transform (reconstructed signal) $w(n)$ becomes lossless, we call it DC lossless in wide sense (DCL-W). When the output of the forward transform (band signal) $y_1(m)$ and $y_2(m)$ become lossless, we call it DC lossless in narrow sense (DCL-N). When all the outputs become lossless, we call it DC lossless.

III. Analysis on Word Length Condition

A. Shifted Error Model for Analysis

Fig.2 (a) illustrates a multiplier in the DWT circuit. A coefficient value c designed as a real number is rounded to a rational number c^* in the circuit. The fraction part of both of the input signal s and the output signal s' is rounded to F_S [bit]. These are denoted by

$$\begin{cases} s' = R_{F_S}[c^* s] \ , \ s2^{F_S} \in \mathbf{Z} \\ c^* = c - \Delta c \ , \ c^* = R_{F_C}[c] \ , \ \Delta c = \Delta_{F_C}[c] \end{cases} . \quad (6)$$

A conventional model for error analysis is illustrated in Fig.2 (b). It describes the coefficient error $-\Delta c \cdot s$ as multiplicative to the signal s [11], and the signal error e' as additive [7]. These are treated independently and approximately by

$$s' = cs - \Delta c \cdot s + e' \ , \ s2^{F_S} \in \mathbf{Z} \ , \ |e'| \le 2^{-1-F_S}. \quad (7)$$

On the contrary, as illustrated in Fig.2 (c), we describe the coefficient error e'' as additive by

$$\begin{cases} s' = R_{F_S}[cs] + e'' \ , \ s2^{F_S} \in \mathbf{Z} \\ e'' = R_{F_S}\left[\Delta_{F_S}[cs] - \Delta_{F_C}[c]s\right] \end{cases} . \quad (8)$$

It utilizes the fact that the coefficient error e'' is observed as a value 2^{-F_S} multiplier by an integer when both of coefficients and signals are rounded. It should be noticed that e'' in Eq.(8) is not an approximation but a strictly described value. This model can be denoted by

$$\begin{cases} s' = cs + e \ , \ s2^{F_S} \in \mathbf{Z} \\ e = R_{F_S}\left[-\Delta_{F_C}[c]s + \Delta_{F_S}[cs]\right] - \Delta_{F_S}[cs] \end{cases} \quad (9)$$

as illustrated in Fig.2 (d). Applying the properties in Eq.(5), under the assumption:

$$\left|\Delta_{F_S}[cs] - \Delta_{F_C}[c]s\right| < (p + 2^{-1})2^{-F_S} \ , \ 0 \le p \in \mathbf{Z}, \quad (10)$$

the maximum of the errors are given by

$$|e| \le (p + 2^{-1})2^{-F_S} \ , \ |e''| \le p2^{-F_S}. \quad (11)$$

As a result, the coefficient error e'' is shifted to the signal error e (shifted error) and it becomes possible to derive the word length condition considering mutual effect of the coefficient error and the signal error.

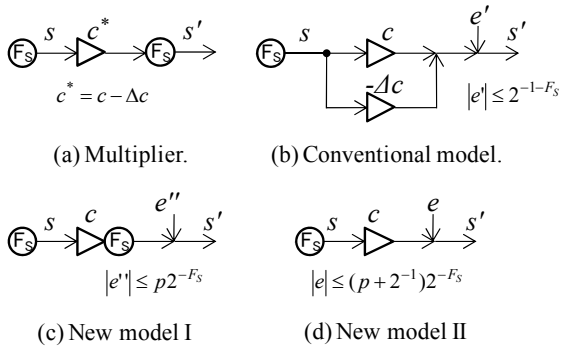


Fig. 2 A multiplier in the DWT and its models for analysis.

B. DC Equivalent Circuit

When the input is restricted to DC signals, $x(n)$ can be described as a scalar x independent of n. The delay z^{-1} can be

treated as 1 and $(1+z^{-1})$ can be replaced by 2. Therefore, instead of the circuits in Fig.1, we use the equivalent circuits for DC signals in Fig.3 to derive the condition.

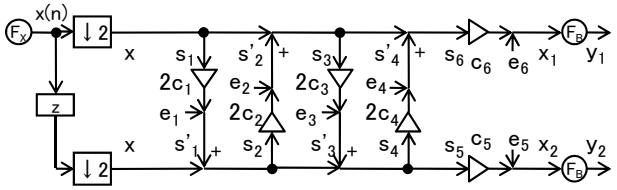
In Fig.3 (a), a scalar x with F_X [bit] fraction part is multiplied by the rational numbers c_i and rounded to F_S [bit]. Finally, the signals are rounded to F_B [bit] at its output to produce the band signals $[y_1\ y_2]$. The shifted errors inside the circuits are described by

$$e_i = e_i' + e_i'' \ , \quad i \in \{1,2,3,4,5,6\} \tag{12}$$

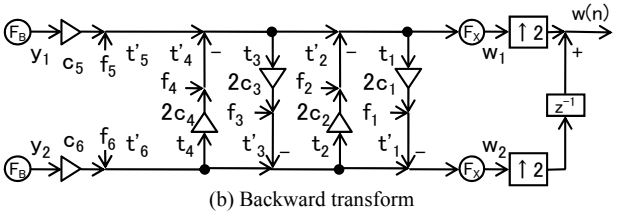
where

$$e_i' = -\Delta_{F_S}[c_i s_i] \ , \ e_i'' = R_{F_S}\left[\Delta_{F_S}[c_i s_i] - \Delta_{F_C}[c_i]s_i\right],$$
$$\begin{bmatrix} s_1 & s_3 & s_5 \\ s_2 & s_4 & s_6 \end{bmatrix} = \begin{bmatrix} 2x & 2(x+s'_2) & s_4 \\ 2(x+s'_1) & 2(s_2+s'_3) & s_3+s'_4 \end{bmatrix},$$
$$s'_i = c_i s_i + e_i = R_{F_S}[c_i s_i] \, .$$

For the backward transform in Fig.3 (b), signals and errors are similarly described. Since the reconstructed signal $w(n)$ is $[w_1\ w_2\ w_1\ w_2\ \cdots]$, it doesn't become DC for $w_1 \neq w_2$ and it impedes the white balancing of a video system.



(a) Forward transform



(b) Backward transform

Fig. 3 Equivalent circuits of the DWT for DC input signals.

C. Nullification of Accumulated Error

In Fig.3(a), the shifted errors in Eq.(12) are accumulated in the circuit. When the accumulated errors are nullified by the rounding at output of the transform, the DWT becomes DC lossless. The output $\mathbf{Y}_{12}=[y_1\ y_2]^T$ is described by

$$\begin{aligned}\mathbf{Y}_{12} = R_{F_B}\big[&\mathbf{I}_U e_6 + \mathbf{I}_L e_5 + \mathbf{K}\big(\mathbf{I}_U e_4 + \mathbf{H}_4\big(\mathbf{I}_L e_3 \\ &+ \mathbf{H}_3\big(\mathbf{I}_U e_2 + \mathbf{H}_2\big(\mathbf{I}_L e_1 + \mathbf{H}_1 \mathbf{I}_{UL} x\big)\big)\big)\big)\big)\big]\end{aligned} \tag{13}$$

where

$$\mathbf{I}_U = [1 \quad 0]^T, \ \mathbf{I}_L = [0 \quad 1]^T, \ \mathbf{I}_{UL} = \mathbf{I}_U + \mathbf{I}_L$$
$$\mathbf{H}_{i\in\{1,3\}} = \begin{bmatrix} 1 & 0 \\ 2c_i & 1 \end{bmatrix}, \ \mathbf{H}_{j\in\{2,4\}} = \begin{bmatrix} 1 & 2c_j \\ 0 & 1 \end{bmatrix}, \ \mathbf{K} = \begin{bmatrix} c_6 & 0 \\ 0 & c_5 \end{bmatrix}$$

and simplified as

$$\mathbf{Y}_{12} = R_{F_B}\left[(\mathbf{H}_{e1}\mathbf{E}_1 + \mathbf{H}_{e2}\mathbf{E}_2) + \mathbf{K}\mathbf{H}_{4321} x\right] \tag{14}$$

where

$$\begin{cases} \mathbf{H}_{e1} = \left[\mathbf{I}_U \quad \mathbf{K}\mathbf{I}_U \quad \mathbf{K}\mathbf{H}_{43}\mathbf{I}_U\right] \\ \mathbf{H}_{e2} = \left[\mathbf{I}_L \quad \mathbf{K}\mathbf{H}_4\mathbf{I}_L \quad \mathbf{K}\mathbf{H}_{432}\mathbf{I}_L\right] \\ \mathbf{H}_{43\ldots} = \mathbf{H}_4\mathbf{H}_3\cdots \end{cases}, \ \begin{cases} \mathbf{E}_1 = [e_6 \quad e_4 \quad e_2]^T \\ \mathbf{E}_2 = [e_5 \quad e_3 \quad e_1]^T \end{cases}.$$

It is described by the shifted errors $\mathbf{E}_1$ and $\mathbf{E}_2$. When the 9-7 DWT is DC lossless, its output becomes

$$\hat{\mathbf{Y}}_{12} = \mathbf{K}\mathbf{H}_{4321} x = \mathbf{I}_U x \in \mathbf{Z} \ . \tag{15}$$

Therefore, from Eq.(5), the loss in the band signal is

$$\mathbf{E}_{y12} = \mathbf{Y}_{12} - \hat{\mathbf{Y}}_{12} = R_{F_B}\left[(\mathbf{H}_{e1}\mathbf{E}_1 + \mathbf{H}_{e2}\mathbf{E}_2)\right] \tag{16}$$

and when the inequality:

$$\left|\mathbf{H}_{e1}\mathbf{E}_1 + \mathbf{H}_{e2}\mathbf{E}_2\right| < \mathbf{I}_{UL}\, 2^{-1-F_B} \tag{17}$$

holds, the DWT becomes DCL-N. Applying similar discussion to the backward transform in Fig.3 (b), the condition for DCL-W is also derived.

D. Condition on Word Length for DC Lossless DWT

According to Eq.(5), when Eq.(10) holds for any value of c and any value of s, the inequality:

$$2^{-F_C+F_S+I_S-1} < p \tag{18}$$

should be satisfied. In this case, the maximum of the shifted error is given by Eq.(11). Therefore, Eq.(17) becomes

$$\left(\left\|\mathbf{H}_{e1}\right\|_{L^1} + \left\|\mathbf{H}_{e2}\right\|_{L^1}\right)\cdot 2^{-F_S}(p + 2^{-1}) < \mathbf{I}_{UL}\, 2^{-1-F_B} \tag{19}$$

as the worst case, where $\left\|\mathbf{H}\right\|_{L^1}$ denotes a row vector composed of sum of absolute values in each column of $\mathbf{H}$. Substituting $F_X=F_B=0$ and the coefficient values of the 9-7 DWT, and including results on the backward transform, Eq.(19) implies

$$p < 2^{-1+F_S-G_E} - 2^{-1} \ , \ G_E = 2.66 \ , \tag{20}$$

where G_E means degree of accumulation of the shifted error inside the circuit. Compiling Eq.(18) and Eq.(20), we can finally derive the word length condition for DC lossless as

$$2^{-\Delta W_C} + 2^{-\Delta W_S} < 2^{-G_E} \tag{21}$$

where

$$\left[\Delta W_C \quad \Delta W_S\right] = \left[F_C - I_S \quad F_S\right].$$

The result means that the fraction part of signals and coefficients should be increased to attain DC lossless, and they can be traded each other.

The integer part of the word length is set to avoid overflow. Namely, $I_C = 1 > \log_2|c_1|$ and $I_S = I_X + 1 > \log_2|s_6|$ by the maximums $c_1 = -1.586$ and $s_6 = 1.230\max|x|$, where I_X is an integer part of the input signal x.

IV. Simulation Results

A. Verification of Condition on Word Length

Fig.4 illustrates a pair of (F_S, F_C) at which the DWT becomes DC lossless for any integer x with W_X=8 [bit]. The bold line indicates the theoretical lower bound derived from Eq.(21). It means the sufficient condition. "x" indicates experimentally measured points with the practically implemented DWT circuit. All of them satisfy the sufficient condition. Therefore, it can be concluded that the theory in Eq.(21) is verified.

B. Optimization of Word Length

Utilizing the result of our analysis, we calculate the optimum word length under the condition in Eq.(21). The cost function $J=2^{-1}(F_C+F_S)$ is minimized for the three examples. Ex.1 trades the word length between F_C and F_S, namely $F_C=F_0+T$ and $F_S=F_0-T$ where T is optimized. Ex.2 and Ex.3 are $F_C=F_S$ and $W_C=W_S$, respectively. Results are summarized in table 1 and table 2. Ex.1 has the minimum cost and it requires 14 [bit] and 13 [bit] for coefficients and signals respectively. Ex.2 requires 11 [bit] for fraction part. It was found that the optimum word length is 14 [bit] for both of coefficients and signals for Ex.3.

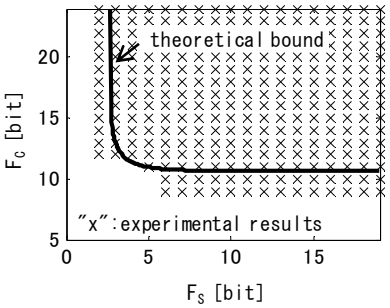


Fig. 4 Word length (F_S, F_C) which guarantees the DC lossless.

TABLE I THEORETICALLY DERIVED OPTIMUM WORD LENGTH.

	Ex.1	Ex.2	Ex.3
F_C F_S	G_E+1+I_S G_E+1	G_E+I_S*	G_E+I_C*+I_S-I_C G_E+I_C*
W_C W_S	G_E+2+I_S+I_C G_E+2+I_S	G_E+I_S*+1+I_C G_E+I_S*+1+I_S	G_E+I_C*+I_S+1
J	G_E+1+I_S/2	G_E+I_S*	G_E+I_C*+(I_S-I_C) /2

$I_S^*=\log_2(2^{I_S}+1)$, $I_C^*=\log_2(2^{I_C}+1)$

TABLE II THEORETICALLY DERIVED OPTIMUM WORD LENGTH FOR THE 9-7 DWT AT W_X=8 [BIT].

	Ex.1	Ex.2	Ex.3
F_C F_S	11.66 3.66	10.67	11.25 4.25
W_C W_S	13.66 12.66	12.67 19.67	13.25
J	7.66	10.67	7.75

V. Conclusions

In this report, we derived the condition on the word length of signals and coefficients such that the DWT becomes DC lossless. In our theoretical analysis, we introduced a new model which shifts the coefficient error to the signal error in order to consider their mutual effect. We also utilized the fact that the accumulated error inside the circuit is nullified by the rounding at its output. As a result, we derived the condition for the DC lossless DWT. Theoretical bound of the word length is derived and the minimum word length was found to be 14 [bit] for 8 [bit] input signals.

Discussion in this report should be extended to multi stage octave decomposition in the near future.

References

[1] ISO/IEC FCD15444-1, "JPEG2000 Image Coding System," March 2000.

[2] A. Descampe, F. O. Devaux, G. Rouvroy, J. D. Legat, J. J. Quisquater, B. Macq, "A Flexible Hardware JPEG 2000 Decoder for Digital Cinema," IEEE Trans. CAS for Video Technology, vol. 16, issue 11, pp.1397 - 1410, Nov. 2006

[3] M. Grangetto, E. Magli, M. Martina, G. Olmo, "Optimization and Implementation of the Integer Wavelet Transform for Image Coding," IEEE Trans. Image Processing, Vol. 11, Issue 6, pp. 596-604, June 2002.

[4] Y. Tonomura, S. Chokchaitam, M. Iwahashi, "Minimum Hardware Implementation of Multipliers of the Lifting Wavelet Transform," IEEE International Conference on Image Processing, WA-L4, pp.2499-2502, Oct. 2004.

[5] H. Kiya, M. Yae, M. Iwahashi, "Linear Phase Two Channel Filter Bank allowing Perfect Reconstruction," IEEE Proc. International Symposium on Circuits and Systems, no.2, pp.951-954, May 1992.

[6] W. Sweldens, "The lifting scheme: A Custom-design Construction of Biorthogonal Wavelets," Technical Report 1994:7, Industrial Mathematics Initiative, Department of Mathematics, University of South Carolina, 1994.

[7] A. M. Reza, Lian Zhu, "Analysis of error in the fixed-point implementation of two-dimensional discrete wavelet transforms," IEEE Trans. CAS, Fundamental Theory and Applications, vol.52, issue 3, pp.641-655, March 2005.

[8] I. Daubechies, W. Sweldens, "Factoring Wavelet Transforms into Lifting Steps," Journal of Fourier Analysis and Applications, Vol. 4, Nr. 3, 1998.

[9] H. Kiya , M. Iwahashi, O. Watanabe, "A New Class of Lifting Wavelet Transform for Guaranteeing Losslessness of Specific Signals," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.3273-3276, March 2008.

[10] K. Hirakawa, T. W. Parks, "Chromatic Adaptation and White Balancing Problem," IEEE Proc. International Conference Image Processing, vol.III, pp.984-987, Nov. 2005.

[11] C. Xiao; P. Agathoklis, D. J. Hill, "Coefficient Sensitivity and Structure Optimization of Multidimensional State-Space Digital Filters," IEEE Trans. Circuits, Systems I, Vol. 45, Issue 9, pp.993-998 Sept. 1998.

[12] M. Iwahashi, H. Kiya, "Finite Word Length Error Analysis based on Basic Formula of Rounding Operation," The International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), no.86, pp.49-52, Dec. 2008.