World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:1, No:2, 2007

Parallel Discrete Fourier Transform for Fast FIR Filtering Based on Overlapped-save Block Structure

Ying-Wen Bai and Ju-Maw Chen

***Abstract*—**To successfully provide a fast FIR filter with FTT algorithms, overlapped-save algorithms can be used to lower the computational complexity and achieve the desired real-time processing. As the length of the input block increases in order to improve the efficiency, a larger volume of zero padding will greatly increase the computation length of the FFT. In this paper, we use the overlapped block digital filtering to construct a parallel structure. As long as the down-sampling (or up-sampling) factor is an exact multiple lengths of the impulse response of a FIR filter, we can process the input block by using a parallel structure and thus achieve a low-complex fast FIR filter with overlapped-save algorithms. With a long filter length, the performance and the throughput of the digital filtering system will also be greatly enhanced.

***Keywords*—**FIR Filter, Overlapped-save Algorithm, Parallel Structure

I. Introduction

A parallel processing system is referred to as block processing which reduces the computational complexity of digital filtering systems [1], [2]. We can also use multirate digital signal processing to achieve the block digital filter [3] shown in Fig. 1, where L represents the input block size, N represents the output block size, and M is the down-sampling (up-sampling) factor. This system is shift-variant because of the down-samplers and up-samplers. If $L = M = N$ and the overall system is shift-invariant, then the system becomes non-overlapped block processing and $\mathbf{P}(z)$ becomes pseudocirculant matrix [4]. If $L \geq M$ and $N \geq M$, then the system becomes overlapped block processing [5], [6]. If $L = M$, the input blocks are not overlapped, but if $L > M$, the input blocks are overlapped. When $N = M$, the output blocks are not overlapped, but if $N > M$, the output blocks are overlapped.

The use of efficient FFT algorithms to implement FIR filtering has been known for quite some time. The conventional overlap-add and the overlap-save algorithms, when implemented using FFT methods, greatly reduce the computational complexity of FIR filtering [7]. We can make great use of the overlapped block digital filtering to achieve a DFT based fast FIR filter structure [5], [6].

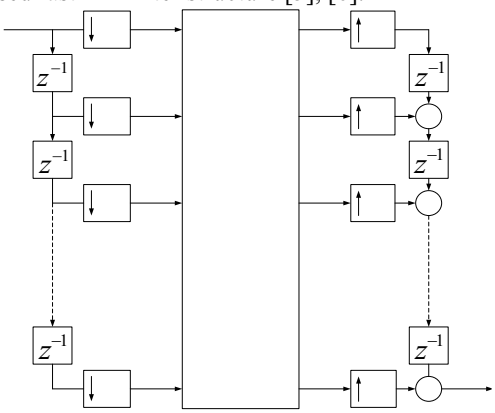

Fig. 1 Multirate representation of an overlapped block digital filter

However, if the filter length is prolonged or the input block length is increased, this structure will compute even longer discrete Fourier transforms for the zero padding. Hence, the length of DFT is the key factor that affects the performance.

We use the overlapped block digital filtering as the base to construct a parallel structure. It is, therefore, necessary to compute the original discrete Fourier transform to achieve high performance, even when the filter length is truly long.

This paper is organized as follows. In Section 2, the overlapped block filter structure is presented. In Section 3, the parallel DFT based on the fast FIR filter structure is described. In Section 4, the experimental performance and results are provided. In the final Section, our conclusions are presented.

II. Overlapped Block Filter Structure

Lin and Mitra have developed a parallel structure [5], [6] with an overlapped block digital filtering to achieve a FIR filter. The term, overlapped block digital filtering refers to give the up/down-sampling factor M. There are many different choices of the input block length L, the output block length N and the block transfer matrix $\mathbf{P}(z)$ but all lead to the same pseudocirculant matrix $\mathbf{Q}(z)$ [4] and the same transfer function $H(z)$ as in (1). The correlation is shown in (2). What satisfies the correlation-block transfer matrix $\mathbf{P}(z)$ is referred to as an extended pseudocirculant matrix. With respect to $\mathbf{R}(z)$ and

Ying-Wen Bai is with the Department of Electronic Engineering, Fu Jen Catholic University, Taipei, Taiwan, 242, R.O.C. (e-mail: bai@ee.fju.edu.tw)
Ju-Maw Chen is a graduate student of Fu Jen Catholic University, Taipei, Taiwan, 242, R.O.C. (e-mail: a9150604@st2.fju.edu.tw)

$\mathbf{S}(z)$, these two matrixes have the overlapped length of the output block and input block.

$$H(z)=H_0(z^M)+z^{-1}H_1(z^M)+\cdots+z^{-(M-1)}H_{M-1}(z^M) \quad (1)$$

$$\mathbf{Q}(z)=\mathbf{R}(z)\mathbf{P}(z)\mathbf{S}(z) \quad (2)$$

One of the $\mathbf{Q}(z)$ is a $M\times M$ matrix

$$\mathbf{Q}(z)=\begin{bmatrix} H_0(z) & H_1(z) & \cdots & H_{M-1}(z) \\ z^{-1}H_{M-1}(z) & H_0(z) & \cdots & H_{M-2}(z) \\ \vdots & \vdots & \ddots & \vdots \\ z^{-1}H_1(z) & z^{-1}H_2(z) & \cdots & H_0(z) \end{bmatrix} \quad (3)$$

$\mathbf{R}(z)$ is a $M\times N$ matrix of the following form

$$\mathbf{R}(z)=\begin{bmatrix} \cdots & 0 & 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \cdots & z^{-1} & 0 & 0 & 0 & \cdots & 1 & 0 & 0 \\ \cdots & 0 & z^{-1} & 0 & 0 & \cdots & 0 & 1 & 0 \\ \cdots & 0 & 0 & z^{-1} & 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \quad (4)$$

and $\mathbf{S}(z)$ is a $L\times M$ matrix of the form

$$\mathbf{S}(z)=\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ z^{-1} & 0 & 0 & \cdots & 0 \\ 0 & z^{-1} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \end{bmatrix} \quad (5)$$

When we choose an input block size of $L=2M-1$, and an output block size of $N=M$, we will achieve the block transfer function matrix $\mathbf{P}_2(z)$.

$$\mathbf{P}_2(z)=\begin{bmatrix} H_0 & H_1 & \cdots & H_{M-1} & 0 & \cdots & 0 \\ 0 & H_0 & \cdots & H_{M-2} & H_{M-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & H_0 & H_1 & \cdots & H_{M-1} \end{bmatrix} \quad (6)$$

$\mathbf{P}_2(z)$ is referred to as a Type S extended pseudocirculant matrix. In the structure of $\mathbf{P}_2(z)$, the input blocks are overlapped, while the output blocks are not overlapped. This structure also has the same nature as overlap-save algorithms for linear convolution.

III. Parallel DFT Based Fast FIR Filter Structure

With respect to the method shown in (6), if we further advocate the block transfer matrix $\mathbf{P}_2(z)$, let us then consider the transfer function.

$$H(z)=H_0(z^M)+z^{-1}H_1(z^M)+\cdots+z^{-(K-1)}H_{K-1}(z^M)+z^{-K}H_K(z^M)+\cdots+z^{-(M-1)}H_{M-1}(z^M)$$

which $K\le M$, and $H_K(z^M),\cdots H_{M-2}(z^M),H_{M-1}(z^M)$ are both equal to zero. Thus we can now decompose the block transfer matrix $\mathbf{P}_2(z)$ into $\mathbf{P}'_2(z)$ and $\mathbf{T}(z)$. These two matrix multiples are:

$$\mathbf{P}_2(z)=\mathbf{P}'_2(z)\mathbf{T}(z) \quad (7)$$

in which the

$$\mathbf{P}'_2(z)=\begin{bmatrix} H_0 & H_1 & \cdots & H_{K-1} & 0 & \cdots\cdots & 0 & 0 & \cdots & 0 \\ 0 & H_0 & \cdots & H_{K-2} & H_{K-1} & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 & H_0 & H_1 & \cdots & H_{K-1} \end{bmatrix} \quad (8)$$

$\mathbf{P}'_2(z)$ is a matrix by $M\times(K+M-1)$.

$$\mathbf{T}(z)=\begin{bmatrix}\mathbf{I}_{K+M-1} & \mathbf{0}_{K+M-1,M-K}\end{bmatrix} \quad (9)$$

$\mathbf{I}_{K+M-1}$ is a $(K+M-1)\times(K+M-1)$ identity matrix, $\mathbf{0}_{K+M-1,M-K}$ is a null matrix of the order $(K+M-1)\times(M-K)$.

According to the associative law of matrix multiplication:

$$\mathbf{Q}(z)=(\mathbf{P}'_2(z)\mathbf{T}(z))\mathbf{S}(z) \quad (10)$$

$$\mathbf{Q}(z)=\mathbf{P}'_2(z)(\mathbf{T}(z)\mathbf{S}(z)) \quad (11)$$

by multiplying $\mathbf{T}(z)$ by $\mathbf{S}(z)$, the results will be $\mathbf{S}'(z)$, which indicates that the input block size has been transformed from $2M-1$ to $K+M-1$.

$$\mathbf{S}'(z)=\mathbf{T}(z)\mathbf{S}(z) \quad (12)$$

$\mathbf{S}'(z)$ is a matrix of $(K+M-1)\times M$.

$$\mathbf{Q}(z)=\mathbf{P}'_2(z)\mathbf{S}'(z) \quad (13)$$

The correlation between $\mathbf{P}'_2(z)$ and $(K+M-1)\times(K+M-1)$ circulant matrix $\mathbf{C}_M$ is as shown in (14) and (15).

$$\mathbf{C}_M=\begin{bmatrix} H_0 & H_1 & \cdots & H_{K-1} & 0 & \cdots & 0 \\ 0 & H_0 & \cdots & H_{K-2} & H_{K-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & H_0 & H_1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & H_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & H_{K-1} \\ H_{K-1} & 0 & \cdots & 0 & 0 & \cdots & H_{K-2} \\ H_{K-2} & H_{K-1} & \cdots & 0 & 0 & \cdots & H_{K-3} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ H_1 & H_2 & \cdots & 0 & 0 & \cdots & H_0 \end{bmatrix} \quad (14)$$

$$\mathbf{P}'_2(z)=\begin{bmatrix}\mathbf{I}_M & \mathbf{0}_{M,K-1}\end{bmatrix}\mathbf{C}_M \quad (15)$$

where $\mathbf{I}_M$ is a $M\times M$ identity matrix, and $\mathbf{0}_{M,K-1}$ is a null matrix of order $M\times(K-1)$. The matrix $\mathbf{C}_M$ is a circulant matrix and therefore can be diagonalized by the DFT matrix. That is

$$\mathbf{C}_M=\hat{\mathbf{A}}\begin{bmatrix} G_0 & 0 & 0 & \cdots & 0 \\ 0 & G_1 & 0 & \cdots & 0 \\ 0 & 0 & G_2 & \ddots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & G_{K+M-2} \end{bmatrix}\hat{\mathbf{B}} \quad (16)$$

where $\hat{\mathbf{A}}$ is a $(K+M-1)$ -point IDFT matrix, $\hat{\mathbf{B}}$ is a $(K+M-1)$ -point DFT matrix and

$$\begin{bmatrix} G_0 \\ \vdots \\ G_{K-1} \\ G_K \\ \vdots \\ G_{K+M-2} \end{bmatrix}=(K+M-1)\hat{\mathbf{A}}\begin{bmatrix} H_0 \\ \vdots \\ H_{K-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (17)$$

Let us take $M=3$, $K=3$, or $H(z)=H_0(z^3)+z^{-1}H_1(z^3)+z^{-2}H_2(z^3)$ as an example. By using the method from (15) and (16), we get Fig. 2, which is the original structure that Lin and Mitra have suggested [5], and is a special case when $K=M$.

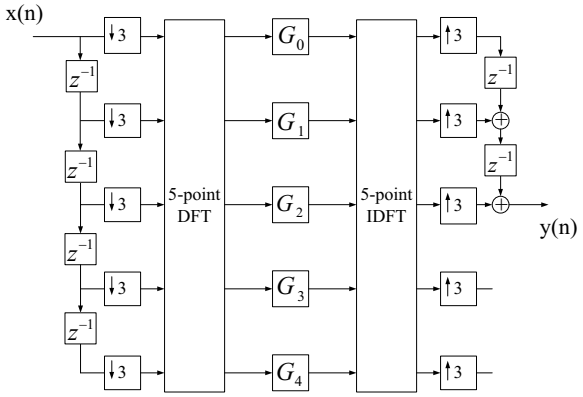

Fig. 2 Lin and Mitra have suggested the original structure: a DFT based algorithm using the Type S overlapped block structure [5]

When the up/down-sampling factor is M, the length K of the transfer function $H(z)$ and the input block length $L=K+M-1$, we have to calculate the DFT length of $K+M-1$. Thus, if the length of the input block increases, in order to increase the throughput, the larger volume of zero padding will greatly increase the computation length of the DFT. To solve this problem, if we consider $M=PK$, in which $P=1,2,3\ldots$, then the block transfer matrix $\mathbf{P}'_2(z)$ is derived by using the method from (8) shown below in parallel.

$$\mathbf{P}'_2(z)=\mathbf{VU} \tag{18}$$

in which

$$\mathbf{U}=\begin{bmatrix} \mathbf{X} & \mathbf{Y} & \mathbf{Z} & \mathbf{O}_2 & \mathbf{O}_1 & \mathbf{O}_2 & \mathbf{O}_1 & \cdots & \mathbf{O}_1 & \mathbf{O}_2 & \mathbf{O}_1 \\ \mathbf{O}_1 & \mathbf{O}_2 & \mathbf{X} & \mathbf{Y} & \mathbf{Z} & \mathbf{O}_2 & \mathbf{O}_1 & \cdots & \mathbf{O}_1 & \mathbf{O}_2 & \mathbf{O}_1 \\ \mathbf{O}_1 & \mathbf{O}_2 & \mathbf{O}_1 & \mathbf{O}_2 & \mathbf{X} & \mathbf{Y} & \mathbf{Z} & \cdots & \mathbf{O}_1 & \mathbf{O}_2 & \mathbf{O}_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{O}_1 & \mathbf{O}_2 & \mathbf{O}_1 & \mathbf{O}_2 & \mathbf{O}_1 & \mathbf{O}_2 & \mathbf{O}_1 & \cdots & \mathbf{X} & \mathbf{Y} & \mathbf{Z} \end{bmatrix} \tag{19}$$

$\mathbf{U}$ is a block matrix by $P\times(2P+1)$

in which

$$\begin{bmatrix}\mathbf{X} & \mathbf{Y} & \mathbf{Z}\end{bmatrix}=\mathbf{I}_{2K-1} \tag{20}$$

$\mathbf{I}_{2K-1}$ is a $(2K-1)\times(2K-1)$ identity matrix

$\mathbf{O}_1$ is a null matrix of order $(2K-1)\times(K-1)$, $\mathbf{O}_2$ is a null matrix of order $(2K-1)\times 1$

$$\mathbf{V}=\begin{bmatrix} \mathbf{P}_K(z) & \mathbf{O}_3 & \mathbf{O}_3 & \cdots & \mathbf{O}_3 \\ \mathbf{O}_3 & \mathbf{P}_K(z) & \mathbf{O}_3 & \cdots & \mathbf{O}_3 \\ \mathbf{O}_3 & \mathbf{O}_3 & \mathbf{P}_K(z) & \cdots & \mathbf{O}_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{O}_3 & \mathbf{O}_3 & \mathbf{O}_3 & \cdots & \mathbf{P}_K(z) \end{bmatrix} \tag{21}$$

$\mathbf{V}$ is a block matrix by $P\times P$. $\mathbf{O}_3$ is a null matrix of the order $K\times(2K-1)$.

Also, when $i=j=1,2,\cdots,P$

$$\left[\mathbf{V}\right]_{ij}=\mathbf{P}_K(z)=\begin{bmatrix} H_0 & H_1 & \cdots & H_{K-1} & 0 & \cdots & 0 \\ 0 & H_0 & \cdots & H_{K-2} & H_{K-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & H_0 & H_1 & \cdots & H_{K-1} \end{bmatrix} \tag{22}$$

in which $\mathbf{P}_K(z)$ is the $K\times(2K-1)$ block transfer matrix of the transfer function $H(z)$ of the original length K with the input block size of $2K-1$, which is the correlation between the block transfer matrix $\mathbf{P}_K(z)$ and $(2K-1)\times(2K-1)$ circulant matrix $\mathbf{C}_K$ is as shown in (14) and (23).

$$\mathbf{P}_K(z)=\begin{bmatrix}\mathbf{I}_K & \mathbf{0}_{K,K-1}\end{bmatrix}\mathbf{C}_K \tag{23}$$

$\mathbf{I}_K$ is a $K\times K$ identity matrix, $\mathbf{0}_{K,K-1}$ is a null matrix of the order $K\times(K-1)$.

The matrix $\mathbf{C}_K$ is a circulant matrix and therefore can be diagonalized by the DFT matrix with (16) and (23). $\mathbf{V}$ can be decomposed as four block matrix multiplications.

$$\mathbf{P}'_2(z)=\mathbf{FADBU} \tag{24}$$

$\mathbf{F},\mathbf{A},\mathbf{D},\mathbf{B}$ are block matrixes by $P\times P$, in which when $i=j=1,2,\cdots,P$, $\left[\mathbf{A}\right]_{ij}$ is a $(2K-1)$-point IDFT matrix, $\left[\mathbf{B}\right]_{ij}$ is a $(2K-1)$-point DFT matrix and

$$\left[\mathbf{D}\right]_{ij}=\begin{bmatrix} G_0 & 0 & 0 & \cdots & 0 \\ 0 & G_1 & 0 & \cdots & 0 \\ 0 & 0 & G_2 & \ddots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & G_{2K-2} \end{bmatrix} \tag{25}$$

$$\left[\mathbf{F}\right]_{ij}=\begin{bmatrix}\mathbf{I}_K & \mathbf{0}_{K,K-1}\end{bmatrix} \tag{26}$$

$\mathbf{I}_K$ is a $K\times K$ identity matrix, $\mathbf{0}_{K,K-1}$ is a null matrix of the order $K\times(K-1)$, if all $i\neq j$, $\left[\mathbf{F}\right]_{ij}$, $\left[\mathbf{A}\right]_{ij}$, $\left[\mathbf{D}\right]_{ij}$, $\left[\mathbf{B}\right]_{ij}$ are null matrixes.

According to the method used in (24), we can achieve a parallel DFT based algorithm with the discrete transformation length of $2K-1$ by using the Type S overlapped block structure. With $M=6$, $K=3$, $P=2$, or $H(z)=H_0(z^3)+z^{-1}H_1(z^3)+z^{-2}H_2(z^3)$ as an example, by following the method shown in (24), the result shown in Fig. 3 is formed.

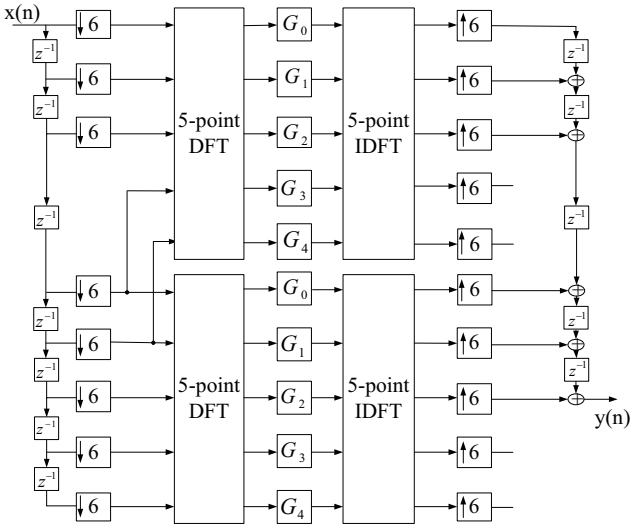

Fig. 3 2-Parallel 5-point DFT for fast FIR filtering based on overlapped-save block structure

By using this method, we will construct a very high order parallel structure according to (24) as long as the computation DFT have a length of $2K-1$ at the most. Due to the DFT fast algorithm, or FFT, we can use FFT to reduce the overall computational complexity by a wide margin. Therefore, as the block size of $2K-1$ equals an integer power of 2, we can realize DFT efficiently by using radix-2 FFT algorithm.

IV. Experimental Performance And Simulation Results

The results are derived by using the method shown in (24), with the formula programmed by Matlab to simulate the process in both the non-parallel and the parallel structure, with providing input 50400 data, to count the computation time through a different impulse response length. Fig. 4 shows the simulation results, in which K represents the filter length and M the down-sampling (or up-sampling) factor, we will discover that if the filter length is not too long and within a certain length, the total time required to finish the computation is otherwise longer. This is all due to the fact that the DFT computation time is not too big and the increase is only a limited one when the filter length is increased. Thus the total number of input blocks determines the computation time since the computation time increases with the number of input blocks. As regards the same total number of the input data, the total number of input blocks becomes smaller when the input block length gets longer. The influence is smaller when the curve has passed the lowest point. If the DFT length is used to reach the overall computation time, the master computation time will become larger due to the fact that the difference of the DFT computation time becomes larger when the filter length and the input block length is longer. It's not hard to see that the 2-parallel structure and 3-parallel structure computation time not only is shorter than the non-parallel structure but also is free of affection by the total number of input blocks and the input block length.

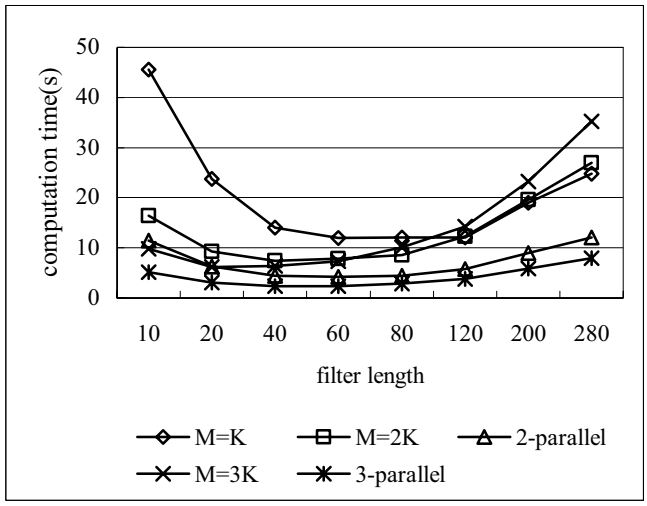

Fig. 4 Computation time with respect to different filter lengths

If we further divide the computation time of the original structure by the computation time of some other structure, we obtain the average gain factor. As shown in Fig. 5, we see that the total number of input blocks affects the gain factor if the filter length is not long enough. The greater the filter length is, the closer the gain factor is to the degree of the parallel structure. The gain factor of the non-parallel structure will also become smaller than one.

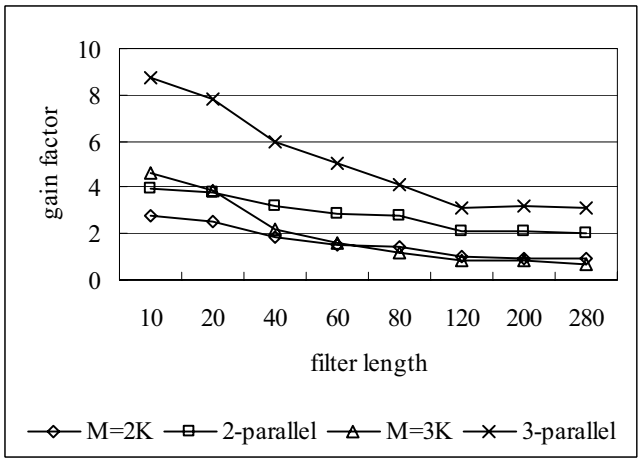

Fig. 5 Gain factor versus different filter lengths under the comparison of the proposed structures and original structure

V. Conclusion

We have proposed the ratiocination and the simulation between the conventional structure and the parallel structure with a parallel DFT based on a fast FIR filter structure. If we compare the performance of these two structures, we find that the 2-parallel and 3-parallel structure computation performance is almost two and three times than that of the original structure when the filter length lengthens. The FIR filter has more factors that might influence the computation performance. In addition to the structure design, the VLSI design of the digital adder and digital multipliers as well as the wave filtering computation all have a great impact on the performance. When all the other

factors are fixed, an appropriate structure can greatly improve the performance. Our parallel processing method results in two major advantages. One is avoiding a longer DFT computation time while the input blocks length is increase. The other advantage is, as the system impulse response length is increased, the major part of the computation time is still long enough to decrease the efficiency even if an FFT algorithm is used to provide the DFT. On the other hand, the parallel structure processing as presented will solve this problem.

REFERENCES

[1] M. Aziz, S. Boussakta, D.C.McLernon, "Parallelisation of the 1-D block filter algorithm to run on multiple DSPs", 9th International Conference on Electronics, Circuits and Systems, Vol. 3, 15-18 Sept. 2002, pp. 943-946.
[2] I. D. Moldovan, "Parallel processing from applications to systems", Morgan Kaufmann Publishers Inc., 1993, ch. 1.
[3] P. P.Vaidyanathan, "Multirate Systems and Filter Banks", Englewood Cliffs, NJ, Prentice-Hall, 1993, ch. 10.
[4] P. P. Vaidyanathan and S. K. Mitra, "Polyphase networks, block digital filtering, LPTV systems, and alias-free QMF banks: a unified approach based on pseudocirculants", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 36, Issue: 3, March 1988, pp. 381-391.
[5] Ing-Song Lin and S. K. Mitra, "Overlapped block digital filtering", IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, Vol. 43, Issue: 8, Aug. 1996, pp. 586-596.
[6] Ing-Song Lin and S. K. Mitra, "Fast FIR Filtering Algorithms Based On Overlapped Block Structure", 1993 IEEE International Symposium on Circuits and Systems, 3-6 May 1993, pp. 363-366.
[7] Sanjit K. Mitra, "Digital Signal Processing: A Computer-Based Approach", McGraw-Hall, 2001, ch. 3.