FLOATING POINT ERROR BOUND IN THE PRIME FACTOR FFT (U)
APR 80 D C MUNSON, B LIU
AFOSR-TR-80-1129
AFOSR-TR-80-3083
Floating Point Error Bound in the Prime Factor FFT.

David C. Munson Jr. and Bode Liu

Dept. of Electrical Eng. & Computer Science
Princeton University
Princeton, New Jersey 08544

Air Force Office of Scientific Research/NM
Bolling AFB, Washington, DC 20332

Approved for public release; distribution unlimited.

IEEE International Conference on Acoustics, Speech and Signal Processing,
Denver Colo., April 9-11, 1980

FFT, floating point roundoff error, prime factor FFT algorithm and error band

The prime factor FFT (PR FFT), developed by Kolba and Parks, makes use of recent computational complexity results by Winograd to compute the DFT with a fewer number of multiplications than that required by the FFT. Patterson and McClellan have derived an expression for the MSE in the PR FFT assuming finite precision fixed point arithmetic. In this paper we derive a bound on the MSE in the PR FFT assuming floating point arithmetic. In the course of the derivation an expression for the actual MSE is also presented, but is seen to be too complicated to be of practical use.
FLOATING POINT ERROR BOUND IN THE PRIME FACTOR FFT

David C. Munson, Jr.
Coordinated Science Laboratory
and Department of Electrical Engineering
University of Illinois
Urbana, IL 61801

ABSTRACT
The prime factor FFT (PF FFT), developed by Kolba and Parks, makes use of recent computational complexity results by Winograd to compute the DFT with a fewer number of multiplications than that required by the FFT. Patterson and McClellan have derived an expression for the MSE in the PF FFT assuming finite precision fixed point arithmetic. In this paper we derive a bound on the MSE in the PF FFT assuming floating point arithmetic. The resulting expression is quite cumbersome, but assuming floating point arithmetic with a fewer number of multiplications than that required by the algorithm. As an example, the 3 point DFT may be written as

\[ Y_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} Y_1 \]

In this particular example A and C are square, but this is not generally the case. An N-dimensional PF FFT is a cascade of M modules, each consisting of several of the computations given by (1). The PF FFT may be written as

\[ Y = [C_0 D_0 A_0] \cdots [C_{M-1} D_{M-1}] E \]

where "\( \cdot \)" denotes a Kronecker product, and \( x \) and \( \tilde{x} \) are the DFT input and output, respectively. It is assumed that the dimension of \( C_i D_i A_i \) is \( N \) and that \( m_0 \cdots m_{M-1} \) are relatively prime. Thus to compute the DFT according to (3), only \( N \) point DFT's are required.

Fig. I gives a pictorial representation of (3) for \( N \)-point modules. Each plane in the figure is a computation as given by (1). To date, algorithms have been published for only four mutually prime sequence lengths \( N_i \). Hence, \( N \) is the maximum number of modules.

B. Characterization of Floating Point Errors

We shall be concerned with binary machines using floating point arithmetic with a double precision accumulator. Thus, each machine number may be expressed as \( (\text{sign}) \cdot a \cdot b \) where the mantissa 'a' is a fraction between \( \frac{1}{2} \) and 1 and the exponent 'b' is an integer. It shall be assumed that 8 bits are used for the mantissa and that enough bits are allotted to the exponent to prevent overflow.

If we let \( f(x) \) denote the machine number result from a real floating point operation, then it is well known that \( f(x) \) for most machines

\[ 80 \ 1 \ 0 \ 6 \ 1 \ 0 \ 8 \]
\[
\begin{align*}
\sin(x+y) &= (\sin y)(1+\delta_1) \\
\sin(x-y) &= (x-y)(1+\delta_2)
\end{align*}
\]

where the errors \(\delta_1\) and \(\delta_2\) satisfy \(-z^{-3/2} \leq \delta_1, \delta_2 \leq z^{-3/2}\) for rounding and \(-z^{-3/2} \leq \delta_1, \delta_2 \leq 0\) for truncation.

The errors \(\delta_1\) and \(\delta_2\) are typically modelled as random variables, independent of \(x\) and \(y\) and uniformly distributed on their range [7]. The bound to be derived in the next section will require knowledge of only the mean squared values of the \(\delta_i\), which can be easily computed. For example, assuming rounding we have for real addition and multiplication

\[
E(\delta_i^2) = \frac{2}{12}
\]

For the remainder of this paper we shall assume rounding arithmetic.

The PF FFT is composed of many short one dimensional DFT's, each implemented as in (1). In the next section we shall require an expression for the error vector at the output of a single one dimensional DFT. Let \(x\) and \(y\) be the respective input and desired output of a single \(N_m\) point DFT. Then

\[
y = CDAx + JCDAx^T
\]

where \(x^R\) and \(y^T\) are the real and imaginary parts of \(y\) respectively. Denote the actual machine output by \(\hat{y}\). Then the error vector at the DFT output, \(\gamma(m) = \hat{y} - y\), may be written as

\[
\gamma(m) = Q(m)x^R + jQ(m)x^T + [G(m)x^R + jG(m)x^T]
\]

where \(Q(m)\) is an error matrix associated with a single \(N_m\) point DFT of a real input array. The error matrix \(Q(m)\) may be obtained as follows. The actual output \(\hat{y}\), resulting from a real input array, is computed by substituting \(f(\cdot)\) for each multiplication and addition occurring in (1). The ith substitution is made with an error source \(\delta_i\) where it is assumed the \(\{\delta_i\}\) are uncorrelated. If desired, \(d_{nk}(1+\delta_i)\) may be substituted for the diagonal elements \(d_{nk}\) in \(D\), so that the error due to storage of the multiplier coefficients is also accounted for. All second order effects involving terms of the form \(\delta_i\delta_j\) are dropped from (1). The output \(\hat{y}\) is then subtracted, giving \(\gamma(m)\) to first order. Each element of \(\gamma(m)\) is a linear combination of the \(y(m)\) with coefficients (the entries in \(Q(m)\)) given by linear combinations of the \(\delta_i\). The matrix \(Q(m)\) is the error matrix associated with an \(N_m\) point DFT of a purely imaginary array. \(Q(m)\) has the same form as \(Q(m)\), however the elements in these two matrices are assumed to be uncorrelated, since they arise from different multiplications. The last term in (6) is due to adding the DFT of \(x^R\) to the DFT of \(jx^T\) in (5). Both \(G(m)\) and \(G(m)\) are diagonal matrices with each element having variance \(z^{-3/2}\). A fact which will be neglected is that the first element in both \(G(m)\) and \(G(m)\) is actually zero. This follows since the component of \(Y_1\) due to \(Y_1^R\) is purely real and the component of \(Y_1\) due to \(jY_1^T\) is purely imaginary so that no error occurs in adding these two components.

III. FLOATING POINT ERROR BOUND

It is desired to bound the average MSE (over all outputs) at the PF FFT output. This error results from computational errors which originate within each module in (3) and then propagate to the output. It is expedient to consider the effects of the errors from each module separately.

It is somewhat difficult to picture the PF FFT for \(N > 2\). However, each portion of the computation, from the nth module to the output, may still be represented in a manner similar to Fig. 1. Fig. 2 illustrates a particular stage \(S_{k}(m)\), which is the transformation from a portion of the nth module input to the PF FFT output. The PF FFT contains \(N_1 \cdots N_m - 1\) of these stages in parallel. Hence, the index \(k\) runs from 1 to \(N_1 \cdots N_m - 1\) and each stage computes only part of the PF FFT output.

Let the error at the jth output of this DFT, due to errors originating within this DFT, be defined as

\[
\epsilon_{j} = (x^{R})_{j} + j(x^{T})_{j}, \quad j = 1, \ldots, N_m
\]

Let \(G(m)\) and \(G(m)\) be the respective matrices of \(G(m)\) and \(G(m)\) for the nth module. Then let

\[
\begin{align*}
\epsilon_{k}(m) &= \sum_{j=1}^{N_m} \epsilon_{j}^{(m)} \\
&= (x^{R})_{k} + j(x^{T})_{k}
\end{align*}
\]

be the rth output of the \(N_1 \cdots N_m - 1\) point DFT in \(S_{k}(m)\) due to the \(\epsilon_{j}^{(m)}\). We shall first compute

\[
\frac{1}{N_m} \sum_{k=1}^{N_m} \sum_{r=1}^{N_m} |\epsilon_{r}^{(m)}|^2
\]

which is the average MSE at the PF FFT output due to errors originating within the nth module. For fixed \(k\) and \(s\) it follows from Parseval that

\[
\sum_{k=1}^{N_m} |\epsilon_{r}^{(m)}|^2 = (N_1 \cdots N_m) \sum_{r=1}^{N_m} |\epsilon_{r}^{(m)}|^2.
\]

Therefore

\[
\frac{1}{N_m} \sum_{k=1}^{N_m} \sum_{r=1}^{N_m} |\epsilon_{r}^{(m)}|^2 = (N_1 \cdots N_m) \sigma^2(m) \quad (7)
\]

where

\[
\sigma^2(m) = \frac{1}{N_m} \sum_{k=1}^{N_m} \sum_{r=1}^{N_m} |\epsilon_{r}^{(m)}|^2
\]

(8)
is the average MSE at the output of the mth module due to errors originating within that module.

An expression for $\sigma^2(m)$ is now required in terms of the algorithm parameters. To investigate the errors occurring within the mth module, (3) is rewritten using a standard Kronecker product identity giving

$$\Delta = \sum_{i=0}^{M-1} (C_i D_i A_i)^{m-1} x_i^{m-1} \cdots x_1^1 \otimes \cdots \otimes (C_{M-1} D_{M-1} A_{M-1})^1 x_1^1.$$  

(9)

Similarly $e$ may be written as a product of matrices times $x$ where the $i$th matrix represents the transformation performed by the $i$th module. In (9) we have broken the computation around the mth module. The output $y(m)$ of this module may be written as

$$y(m) = \sum_{i=1}^{M} (C_i D_i A_i)^{m-1} x_i^{m-1} \cdots x_1^1.$$  

(10)

where $y(m)$ is the output to the mth module given by $y(m) = \sum_{i=1}^{M} (C_i D_i A_i)^{m-1} x_i^{m-1} \cdots x_1^1$.

It is apparent from (6) and (10) that the error $\epsilon(m)$ in $y(m)$ may be expressed as

$$\epsilon(m) = P(m) x(m-1) + \bar{P}(m) \bar{x}(m-1) + F(m) s(m)$$  

(11)

where

$$P(m) = [I_{N} \cdots I_{m-1} \otimes Q(m) \cdots I_1^1 x_1^1]^T,$$  

$$\bar{P}(m) = [I_{N} \cdots I_{m-1} \otimes G(m) \cdots I_1^1 x_1^1]^T.$$  

(12)

Here $^{T\otimes}$ indicates that each time $Q(m)$ or $G(m)$ is repeated in the Kronecker product, its elements are superscripted relating them to their respective $C_i D_i A_i$ implementation. Without this additional superscript, the $\delta$ arising from one implementation of $C_i D_i A_i$ would be assumed to be identical to those arising from another implementation. However, since the inputs to the various implementations are different, we shall in fact assume these errors to be uncorrelated.

From the definition of $\epsilon(m)$, the quantity $\sigma^2(m)$, given by (8) may be written as

$$\sigma^2(m) = \frac{1}{N} E [ \epsilon^T(m) \epsilon(m) ]$$

(13)

where $^{T\otimes}$ denotes complex conjugate transpose. Substituting from (11) gives

$$\sigma^2(m) = \frac{1}{N} E [ x(m-1) P(m) x(m) R(m-1) ] + \frac{1}{N} E [ x(m-1) P(m) x(m) \bar{R}(m) ]^T + \frac{1}{N} E [ x(m-1) P(m) x(m) ] R(m-1)$$

$$+ \frac{1}{N} E [ x(m) P(m) \bar{x}(m) \bar{R}(m) ]^T + \frac{1}{N} E [ x(m) P(m) \bar{x}(m) ] \bar{R}(m)$$

$$- \frac{1}{N} E [ x(m) P(m) ] R(m-1).$$

(14)

may be further simplified since

$$\epsilon^T(m) F(m) P(m) \epsilon(m) = \epsilon^T(m) R(m) E [ F(m) P(m) ] R(m)^T$$

$$E [ x(m) R(m) ] = \epsilon^T(0) \epsilon^T(m)$$

$$\frac{1}{N} E [ \epsilon^T(m) \epsilon(m) ] = \epsilon^T(0) \epsilon^T(m)$$

The second equality follows because $F(m)$ is a diagonal matrix with each element having variance $2/12$. The last term in (13) is equal to a similar expression so that the sum of the last two terms is given by

$$\frac{1}{N} E [ \epsilon^T(m) \epsilon(m) ] = \frac{1}{N} E [ \epsilon^T(0) \epsilon^T(m) ] = 2/12.$$

So far, only errors originating within the mth module have been considered. We shall now formulate an expression for the total average MSE at the PF FFT output due to all modules. It may be seen from (11) and (12) that the error vector at the output of the mth module depends on $Q(m)$ and $G(m)$. However, $Q(m)$ and $G(m)$ are uncorrelated with $Q(i)$ and $G(i)$ for $i \neq m$ and hence $\epsilon(m)$ is uncorrelated with $\epsilon(i)$. The total average MSE at the PF FFT output is therefore the sum of the average MSE's due to each module. From (7) the total average MSE $\sigma^2$ is given by

$$\sigma^2 = \frac{1}{N} \sum_{m=1}^{M-1} \sum_{l=0}^{N-1} \sum_{s=0}^{N-1} E [ \epsilon^T(m) \epsilon(m) ]$$

$$= \frac{1}{N} \sum_{m=1}^{M-1} \sum_{l=0}^{N-1} \sum_{s=0}^{N-1} \epsilon^T(m) \epsilon(m)$$

$$+ \frac{1}{N} \sum_{m=1}^{M-1} \sum_{l=0}^{N-1} \sum_{s=0}^{N-1} \epsilon^T(m) \epsilon(m)$$

(15)

This may be further simplified since

$$E [ x(m) R(m) ] = \epsilon^T(0) \epsilon^T(m)$$

where $\delta(m)$ is the total average MSE at the PF FFT output due to all modules. It may be seen from (11) and (12) that the error vector at the output of the mth module depends on $Q(m)$ and $G(m)$. However, $Q(m)$ and $G(m)$ are uncorrelated with $Q(i)$ and $G(i)$ for $i \neq m$ and hence $\epsilon(m)$ is uncorrelated with $\epsilon(i)$. The total average MSE at the PF FFT output is therefore the sum of the average MSE's due to each module. From (7) the total average MSE $\sigma^2$ is given by

$$\sigma^2 = \frac{1}{N} \sum_{m=1}^{M-1} \sum_{l=0}^{N-1} \sum_{s=0}^{N-1} E [ \epsilon^T(m) \epsilon(m) ]$$

$$= \frac{1}{N} \sum_{m=1}^{M-1} \sum_{l=0}^{N-1} \sum_{s=0}^{N-1} \epsilon^T(m) \epsilon(m)$$

$$+ \frac{1}{N} \sum_{m=1}^{M-1} \sum_{l=0}^{N-1} \sum_{s=0}^{N-1} \epsilon^T(m) \epsilon(m)$$

(15)

This may be further simplified since

$$E [ x(m) R(m) ] = \epsilon^T(0) \epsilon^T(m)$$

where $\delta(m)$ is the total average MSE at the PF FFT output due to all modules. It may be seen from (11) and (12) that the error vector at the output of the mth module depends on $Q(m)$ and $G(m)$. However, $Q(m)$ and $G(m)$ are uncorrelated with $Q(i)$ and $G(i)$ for $i \neq m$ and hence $\epsilon(m)$ is uncorrelated with $\epsilon(i)$. The total average MSE at the PF FFT output is therefore the sum of the average MSE's due to each module. From (7) the total average MSE $\sigma^2$ is given by

$$\sigma^2 = \frac{1}{N} \sum_{m=1}^{M-1} \sum_{l=0}^{N-1} \sum_{s=0}^{N-1} E [ \epsilon^T(m) \epsilon(m) ]$$

$$= \frac{1}{N} \sum_{m=1}^{M-1} \sum_{l=0}^{N-1} \sum_{s=0}^{N-1} \epsilon^T(m) \epsilon(m)$$

$$+ \frac{1}{N} \sum_{m=1}^{M-1} \sum_{l=0}^{N-1} \sum_{s=0}^{N-1} \epsilon^T(m) \epsilon(m)$$

(15)
REFERENCES