A Bound on the Output of a Circular Convolution with Application to Digital Filtering

ALAN V. OPPENHEIM, Member, IEEE
CLIFFORD WEINSTEIN, Student Member, IEEE
Lincoln Laboratory
Massachusetts Institute of Technology
Lexington, Mass. 02173

Abstract

When implementing a digital filter, it is important to utilize in the design a bound or estimate of the largest output value which will be obtained. Such a bound is particularly useful when fixed point arithmetic is to be used since it assists in determining register lengths necessary to prevent overflow. In this paper we consider the class of digital filters which have an impulse response of finite duration and are implemented by means of circular convolutions performed using the discrete Fourier transform. A least upper bound is obtained for the maximum possible output of a circular convolution for the general case of complex input sequences. For the case of real input sequences, a lower bound on the least upper bound is obtained. The use of these results in the implementation of this class of digital filters is discussed.

Manuscript received November 26, 1968.
This work was sponsored by the U. S. Air Force.

I. Introduction

When implementing a digital filter, either in hardware or on a computer, it is important to utilize in the design a bound or estimate of the largest output value which will be obtained. Such a bound is particularly useful when fixed point arithmetic is to be used, since it assists in determining register lengths necessary to prevent overflow. This paper considers the class of digital filters which have an impulse response of finite duration and are implemented by means of convolution sums performed using the discrete Fourier transform (DFT). The output samples of such a filter are obtained from the results of N-point circular convolutions of the filter impulse response (kernel) with sections of the input. These circular convolutions are obtained by computing the DFT of the input section, multiplying by the DFT of the impulse response, and inverse transforming the result. Stockham [1] has discussed procedures for utilizing the results of these circular convolutions to perform linear convolutions, rationales for choosing the transform length N, and speed advantages to be gained by using the fast Fourier transform (FFT) to implement the DFT. We concern ourselves here only with bounding the output of the N-point circular convolutions.

II. Problem Statement

According to the above discussion, we would like to determine an upper bound on the maximum modulus of an output value that can result from an N-point circular convolution. With $\{x_n\}$ denoting the input sequence, $\{h_n\}$ denoting the kernel, and $\{y_n\}$ denoting the output sequence, we have

$$y_n = \sum_{k=0}^{N-1} x_k h_{(n-k) \bmod N} \quad n = 0, 1, \cdots, N-1 \tag{1}$$

where it is understood that, in general, each of the three sequences may be complex. The circular convolution is accomplished by forming the product

$$Y_k = H_k X_k \tag{2}$$

where

$$X_k = \frac{1}{N} \sum_{n=0}^{N-1} x_n W^{nk} \quad k = 0, 1, \cdots, N-1 \tag{3}$$

$$Y_k = \frac{1}{N} \sum_{n=0}^{N-1} y_n W^{nk} \quad k = 0, 1, \cdots, N-1 \tag{4}$$

$$H_k = \sum_{n=0}^{N-1} h_n W^{nk} \quad k = 0, 1, \cdots, N-1 \tag{5}$$

with W defined as $W = \exp[j2\pi/N]$.

For convenience in notation, we imagine the computation to be carried out on fixed point fractions. Thus we bound the input values so that

$$|x_n| \leq 1. \tag{6}$$

By virtue of (3) we are then assured that

$$|X_k| \leq 1$$

so that the values of X_k do not overflow in the fixed point word.

In the typical cases, the sequence h_n is known and, consequently, so is the sequence H_k. Therefore it is not necessary to continually evaluate (5); that is, the sequence H_k is computed, normalized, and stored in advance. Thus it is reasonable to only apply a normalization to H_k and not to h_n, so that we require[1]

$$|H_k| \leq 1. \tag{7}$$

A normalization of the transform of the kernel so that the maximum modulus is unity allows maximum energy transfer through the filter, consistent with the requirement that Y_k does not overflow the register length.

Our objective is to obtain an upper bound on $|y_n|$ for all sequences $\{x_n\}$ and $\{H_k\}$ consistent with (6) and (7). This bound will specify, for example, the scaling factor to be applied in computing the inverse of (4) to guarantee that no value of y_n overflows the fixed point word. The following results will be obtained.

Result A: With the above constraints, the result of the N-point circular convolution of (1) is bounded by

$$|y_n| \leq \sqrt{N}.$$

Result B: In the general case where $\{x_n\}$ and $\{h_n\}$ are allowed to be complex, the bound in Result A is a least upper bound. This will be shown by demonstrating a sequence that can achieve the bound.

Result C: If we restrict $\{x_n\}$ and/or $\{h_n\}$ to be real valued, the bound of Result A is no longer a least upper bound for every N. However, the least upper bound $\beta(N)$ is itself bounded by

$$\frac{\sqrt{N}}{2} \leq \beta(N) \leq \sqrt{N}.$$

III. Derivation of Results

Proof of Result A

Parseval's relation requires that

$$\sum_{n=0}^{N-1} |y_n|^2 = N \sum_{k=0}^{N-1} |Y_k|^2 \tag{8}$$

and

$$\sum_{n=0}^{N-1} |x_n|^2 = N \sum_{k=0}^{N-1} |X_k|^2. \tag{9}$$

Substituting (2) into (8) and using (7),

$$\sum_{n=0}^{N-1} |y_n|^2 \leq N \sum_{k=0}^{N-1} |X_k|^2, \tag{10}$$

or, using (9),

$$\sum_{n=0}^{N-1} |y_n|^2 \leq \sum_{n=0}^{N-1} |x_n|^2 \tag{11}$$

with equality if and only if $|H_k| = 1$. However, (6) requires that

$$\sum_{n=0}^{N-1} |x_n|^2 \leq N \tag{12}$$

with equality if and only if $|x_n| = 1$. Combining (11) and (12),

$$\sum_{n=0}^{N-1} |y_n|^2 \leq N. \tag{13}$$

But

$$|y_n|^2 \leq \sum_{n=0}^{N-1} |y_n|^2 \tag{14}$$

and therefore

$$|y_n| \leq \sqrt{N}. \tag{15}$$

Proof of Result B

To show that $\sqrt{N}$ is a least upper bound on $|y_n|$, we review the conditions for equality in the inequalities used above. We observe that for equality to be satisfied in (15), it must be satisfied in (11), (12), and (14), requiring that

1) $|H_k| = 1$
2) $|x_n| = 1$
3) Any output sequence $\{y_n\}$ which has a point whose modulus is equal to $\sqrt{N}$ can contain only one nonzero point.

The third requirement can be rephrased as a requirement on the input sequence and on the sequence H_k. Specifically, if the output sequence contains only one nonzero point then Y_k for this sequence must be of the form

$$Y_k = AW^{n_0 k} = |A| \exp\left[j\left[\frac{2\pi}{N} n_0 k + \rho\right]\right]$$

where ρ is a real constant and n_0 is an integer so that, from (2),

$$H_k X_k = |A| \exp\left[j\left[\frac{2\pi}{N} n_0 k + \rho\right]\right]. \tag{16}$$

We can express H_k and X_k as

$$H_k = e^{j\eta_k}$$

and

$$X_k = |X_k| e^{j\theta_k}$$

where we have used the fact that $|H_k| = 1$. For (16) to be satisfied, then

$$|X_k| = |A| \tag{17}$$

and

$$\eta_k = -\theta_k + \frac{2\pi}{N} n_0 k + \rho. \tag{18}$$

Therefore, requirement 3) can be replaced by the requirement that:

3′) $|X_k|$ = constant and the phase of H_k be chosen to satisfy (18).

[1] The restrictions of (6) and (7) do not impose any loss of generality, and are introduced only for convenience. The bounds to be derived on max $|y_n|$ can be interpreted in a more general sense as bounds on the ratio max $|y_n| / \{\max|x_n| \max|H_k|\}$.

As an additional observation, we note that for *any* input sequence $\{x_n\}$,

$$|y_n| \leq \sum_{k=0}^{N-1} |H_k|\,|X_k|$$

with equality for some value of n if and only if $|H_k| = 1$ and the phase of H_k is chosen on the basis of (18). Therefore, for any $\{x_n\}$ the output modulus is maximized when H_k is chosen in this manner. This maximum value will only equal $\sqrt{N}$, however, if, in addition, $|x_n| = 1$ and $|X_k| =$ constant.

For N even, a sequence having the property that $|x_n| = 1$ and $|X_k| =$ constant is (see Appendix) the sequence

$$x_n = \exp\left[j\frac{\pi n^2}{N}\right] = W^{n^2/2}. \tag{19}$$

For N odd, a sequence with $|x_n| = 1$ and $|X_k| =$ constant is[2] (see Appendix)

$$x_n = \exp\left[j\frac{2\pi n^2}{N}\right] = W^{n^2}. \tag{20}$$

Using one of these sequences as the input, and choosing $H_k = e^{j\eta_k}$, with η_k given by (18), equality in (15) can be achieved for any N. Thus the bound given in Result A is a least upper bound.

Proof of Result C

Consider first the case where $\{x_n\}$ is restricted to be real. It can be verified by consideration of all possibilities that for $N = 2$ and $N = 3$, no real sequence exists for which both the sequence and its transform have constant modulus. Therefore, for these values of N at least, (15) does not provide a least upper bound, since requirements 2) and 3′) cannot be satisfied simultaneously. Note, however, that for $N = 4$ the real sequence $\{x_n\} = \{1, 1, -1, 1\}$ satisfies 2) and 3′), and therefore for $N = 4$, (15) is a least upper bound.

If $\{h_n\}$ is required to be real (with no such restriction on $\{x_n\}$), then one can verify for $N = 2$ that if $\{x_n\}$ is chosen to satisfy 2) and 3′), then the phase of H_k cannot be chosen to satisfy 3′), and thus (15) is not a least upper bound for this case.

To show that $\beta(N) \geq \sqrt{N}/2$ for $\{x_n\}$ and/or $\{h_n\}$ restricted to be real valued, it suffices to show that $\beta(N) \geq \sqrt{N}/2$ for both $\{x_n\}$ and $\{h_n\}$ real valued. This we will demonstrate only for the case where N is even, since the argument for N odd is identical.

Consider the complex sequence

$$f_n = \exp\left[j\pi n^2/N\right] = \cos\frac{\pi n^2}{N} + j\sin\frac{\pi n^2}{N}$$

with DFT denoted by

$$F_k = R_k + jI_k$$

where R_k and I_k are real valued and (see Appendix)

$$R_k^2 + I_k^2 = \frac{1}{N}. \tag{21}$$

Since $\exp\left[j\pi n^2/N\right]$ is an even function of n, i.e.,

$$\exp\left[j\pi n^2/N\right] = \exp\left[j\pi(N-n)^2/N\right],$$

R_k is the DFT of $\cos(\pi n^2/N)$ and I_k is the DFT of $\sin(\pi n^2/N)$. Now, if we choose

$$x_n = \cos\left(\frac{\pi n^2}{N}\right)$$

and

$$H_k = \begin{cases} 1 & R_k > 0 \\ -1 & R_k \leq 0, \end{cases}$$

then

$$y_0 = \sum_{k=0}^{N-1} |R_k|. \tag{22}$$

Similarly, if we choose $x_n' = \sin(\pi n^2/N)$, then we can choose $\{H_k\}$ in such a way that

$$y_0' = \sum_{k=0}^{N-1} |I_k|. \tag{23}$$

We note that since $\{x_n\}$ and $\{x_n'\}$ are both real, the values y_0 and y_0' will be obtained with $\{H_k\}$ having even magnitude and odd phase, corresponding to real $\{h_n\}$. Now, if β is the least upper bound for $|y_n|$, then

$$\beta \geq y_0 \tag{24a}$$

$$\beta \geq y_0' \tag{24b}$$

and, from (21),

$$|R_k| \leq \frac{1}{\sqrt{N}}$$

$$|I_k| \leq \frac{1}{\sqrt{N}}$$

and, hence,

$$|R_k| \geq \sqrt{N}\,|R_k|^2 \tag{25a}$$

$$|I_k| \geq \sqrt{N}\,|I_k|^2. \tag{25b}$$

Combining (22), (23), (24), and (25),

$$\beta \geq \sum_{k=0}^{N-1} |R_k| \geq \sqrt{N} \sum_{k=0}^{N-1} |R_k|^2 \tag{26a}$$

$$\beta \geq \sum_{k=0}^{N-1} |I_k| \geq \sqrt{N} \sum_{k=0}^{N-1} |I_k|^2. \tag{26b}$$

Adding (26a) and (26b) and using (21),

$$2\beta \geq \sqrt{N}$$

or

$$\beta \geq \frac{\sqrt{N}}{2}. \tag{27}$$

Since we argued previously that $\beta \leq \sqrt{N}$, Result C is proved.

[2] The sequences (19) and (20) were suggested to the authors by C. M. Rader of the M.I.T. Lincoln Laboratory.

IV. Discussion

The bound obtained in the previous sections can be utilized in several ways. If the DFT computation is carried out using a block floating-point strategy so that arrays are rescaled only when overflows occur, then a final rescaling must be carried out after each section is processed so that it is compatible with the results from previous sections. For general input and filter characteristics, the final rescaling can be chosen based on the bounds given here to insure that the output will not exceed the available register length.

The use of block floating-point computation requires the incorporation of an overflow test. In some cases we may wish instead to incorporate scaling in the computation in such a way that we are guaranteed never to overflow. For example, when we realize the DFT with a power of two algorithm, overflows in the FFT computation of $\{X_k\}$ will be prevented by including a scaling of $\frac{1}{2}$ at each stage, since the maximum modulus of an array in the computation is nondecreasing and increases by at most a factor of two as we proceed from one stage to the next [2]. With this scaling, the bound derived in this paper guarantees that with a power of two computation, scaling is not required in more than half the arrays in the inverse FFT computation. Therefore, including a scaling of $\frac{1}{2}$ in the first half of the stages in the inverse FFT will guarantee that there are no overflows in the remainder of the computation. The fact that $\beta \geq \sqrt{N}/2$ indicates that if we restrict ourselves to only real input data, at most one rescaling could be eliminated for some values of N.

The bounds derived and method of scaling mentioned above apply to the general case; that is, except for the normalization of (7), they do not depend on the filter characteristics. This is useful when we wish to fix the scaling strategy without reference to any particular filter. For specific filter characteristics, the bound can be reduced. Specifically, it can be verified from (1) and (6) that in terms of $\{h_n\}$

$$|y_n| \leq \sum_{l=0}^{M-1} |h_l| \tag{28}$$

where M denotes the length of the impulse response. This is a least upper bound since a sequence $\{x_n\}$ can be selected which will result in this value in the output. This will be significantly lower than the bound represented in (15) if, for example, the filter is very narrow band, or if the kernel has many points with zero value.

Appendix

We wish to demonstrate that for N even, the sequence

$$x_n = \exp\left[j\frac{\pi n^2}{N}\right] \qquad \begin{matrix} n = 0, 1, \cdots, N-1 \\ N \text{ even} \end{matrix} \tag{29}$$

has a discrete Fourier transform with constant modulus and that for N odd, the sequence

$$x_n = \exp\left[j\frac{2\pi n^2}{N}\right] \qquad \begin{matrix} n = 0, 1, \cdots, N-1 \\ N \text{ odd} \end{matrix} \tag{30}$$

has a discrete Fourier transform with constant modulus. We consider first the case of (29). Letting X_k denote the DFT of x_n,

$$X_k = \frac{1}{N}\sum_{n=0}^{N-1} \exp\left[j\frac{\pi n^2}{N}\right] \exp\left[j\frac{2\pi nk}{N}\right]$$

or

$$X_k = \frac{1}{N}\exp\left[-j\frac{\pi k^2}{N}\right] \sum_{n=0}^{N-1} \exp\left[j\pi(n+k)^2\right]. \tag{31}$$

We wish to show first that

$$\sum_{n=0}^{N-1} \exp\left[j\frac{\pi}{N}(n+k)^2\right]$$

is a constant. It is easily verified by a substitution of variables that

$$\sum_{n=0}^{2N-1} \exp\left[j\pi(n+k)^2/N\right] = \text{constant} \triangleq B. \tag{32}$$

But

$$\begin{aligned}
&\sum_{n=0}^{2N-1} \exp\left[j\pi(n+k)^2/N\right] \\
&= \sum_{n=0}^{N-1} \exp\left[j\pi(n+k)^2/N\right] + \sum_{n=N}^{2N-1} \exp\left[j\pi(n+k)^2/N\right] \\
&= \sum_{n=0}^{N-1} \exp\left[j\pi(n+k)^2/N\right] \\
&\quad + \sum_{n=0}^{N-1} \exp\left[j\pi(n+k)^2/N\right] \exp\left[j\pi N\right]
\end{aligned}$$

or, since N is even,

$$\begin{aligned}
&\sum_{n=0}^{2N-1} \exp\left[j\pi(n+k)^2/N\right] \\
&\qquad = 2\sum_{n=0}^{N-1} \exp\left[j\pi(n+k)^2/N\right].
\end{aligned} \tag{33}$$

Combining (31), (32), and (33),

$$X_k = \frac{1}{N}\cdot B \cdot \exp\left[-j\pi k^2/N\right].$$

To determine the modulus of B, Parseval's relation requires that

$$\sum_{n=0}^{N-1} |x_n|^2 = N\sum_{n=0}^{N-1} |X_k|^2$$

or

$$N = |B|^2.$$

Therefore

$$|B| = \sqrt{N}$$

or

$$|X_k| = \frac{1}{\sqrt{N}}.$$

It can be verified by example (try $N=3$) that the sequence of (29) does not have a DFT with constant modulus if N is odd.

Consider next the sequence of (30). We will show that X_k has constant modulus by showing that the circular autocorrelation of x_n, which we denote by c_n, is nonzero only at $n=0$. Specifically, consider

$$c_n = \sum_{r=0}^{N-1} x_r x^*_{(n+r) \bmod N}$$

$$= \sum_{r=0}^{N-1} \exp\left[j\frac{2\pi r^2}{N}\right] \exp\left[-j\frac{2\pi[(n+r)^2] \bmod N}{N}\right].$$

Now,

$$\exp\left[-j\frac{2\pi[(n+r)^2] \bmod N}{N}\right] = \exp\left[-j\frac{2\pi(n+r)^2}{N}\right].$$

Therefore,

$$c_n = \sum_{r=0}^{N-1} \exp\left[j\frac{2\pi r^2}{N}\right] \exp\left[-j\frac{2\pi(r+n)^2}{N}\right]$$

$$= \sum_{r=0}^{N-1} \exp\left[-j\frac{2\pi n^2}{N}\right] \exp\left[-j\frac{4\pi rn}{N}\right]$$

$$= \exp\left[-j\frac{2\pi n^2}{N}\right] \sum_{r=0}^{N-1} \exp\left[-j\frac{4\pi rn}{N}\right].$$

But

$$\sum_{r=0}^{N-1} \exp\left[-j\frac{4\pi rn}{N}\right]$$

$$= \begin{cases} 1 & n = 0 \\ 0 & n \neq 0, \quad N \text{ odd} \\ 1 & n = \dfrac{N}{2}, \quad N \text{ even} \\ 0 & n \neq 0,\ n \neq \dfrac{N}{2}, \quad N \text{ even.} \end{cases}$$

Since we are considering the case of N odd,

$$c_n = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0. \end{cases}$$

Since $|X_k|$ is constant, we may again use Parseval's theorem to show that $|X_k| = 1/\sqrt{N}$.

References

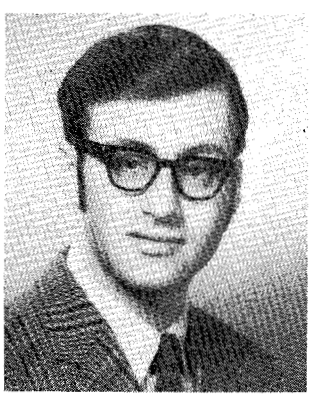
[1] T. G. Stockham, "High speed convolution and correlation," *1966 Spring Joint Computer Conf., AFIPS Proc.*, vol. 28. Washington, D. C.: Spartan, 1966, pp. 229–233.

[2] P. D. Welch, "A fixed-point fast Fourier transform error analysis," this issue, pp. 151–157.

Alan V. Oppenheim (S'57–M'65) was born in New York, N. Y., on November 11, 1937 He received the S.B. and S.M. degrees in 1961 and the Sc.D. degree in 1964 in electrical engineering from the Massachusetts Institute of Technology, Cambridge.

From 1961 to 1964 he was a member of the M.I.T. Research Laboratory of Electronics and an Instructor in the Department of Electrical Engineering, where he received a departmental teaching award in 1963. During this period his primary activities centered on system and communication theory. His doctoral research involved the application of modern algebra to the characterization of nonlinear systems. Since 1964 he has been an Assistant Professor in the M.I.T. Department of Electrical Engineering and a Staff Member of the Research Laboratory of Electronics. His present research activities are in the areas of speech communication and digital waveform processing. He is presently on a leave of absence at the M.I.T. Lincoln Laboratory, Lexington, Mass.

Dr. Oppenheim is a member of Tau Beta Pi, Eta Kappa Nu, Sigma Xi, and the Acoustical Society of America.



Clifford Weinstein (S'66) was born in New York, N. Y., on April 7, 1944. He received the B.S.E.E. and M.S.E.E. degrees in 1965 and 1967, respectively, from the Massachusetts Institute of Technology, Cambridge, where he is currently working toward the Ph.D. degree in electrical engineering.

He is also a member of the M.I.T. Lincoln Laboratory, Lexington, Mass., where he is doing doctoral research in the area of digital signal processing, involving a study of quantization effects in digital filters.

Mr. Weinstein is a member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.