Design of Optimal Finite Wordlength FIR Digital Filters Using Integer Programming Techniques

DUŠAN M. KODEK, MEMBER, IEEE

***Abstract*— The application of a general-purpose integer-programming computer program to the design of optimal finite wordlength FIR digital filters is described. Examples of two optimal low-pass FIR finite wordlength filters are given andthe results are compared with the results obtained by rounding the infinite wordlength coefficients. An analysis of the approach based on the results of more than 50 design cases is presented and the problem of optimal wordlength choice is discussed.**

I. INTRODUCTION

WHEN digital filters are implemented on a computer or with special-purpose hardware, each filter coefficient has to be represented by a finite number of bits. The simplest and most widely used approach to the problem is the rounding of the optimal[1] infinite precision coefficients to its b-bit representation. Several authors [1], [2] analyzed the effect of coefficient quantization on the frequency response of FIR filters. Statistical bounds on the error thus incurred were developed and verified by experimental data. It is now readily possible using these bounds to design FIR filters with finite wordlength coefficients.

However, the filters so obtained are not optimal anymore and in most cases there exists another set of finite wordlength coefficients which gives the best Chebyshev approximation to the desired frequency response. To find this set of coefficients it is necessary to include the finite wordlength restriction into the filter design procedure. The original problem of continuous optimal Chebyshev approximation becomes a much more complex discrete optimization problem. The standard methods of optimal FIR digital filter design, namely the Remez algorithm [3] and linear programming [4], do not work when the finite wordlength restriction is imposed, and one is forced to search for other methods.

Several authors [5]–[8], [15]–[19] have investigated the possibility of using suboptimal algorithms which systematically improve the coefficients obtained by the rounding of optimal infinite precision coefficients. They show that with these algorithms it is possible to considerably improve the rounded solution. Still, there are two important limitations of all these methods:

1) they are successful only for filters with a small number of coefficients (typically less than 10) and are therefore not well suited for the design of FIR digital filters; and

2) the solution obtained is suboptimal in most cases; even if it is optimal there is no proof of optimality and the designer is not aware that the optimal solution has been reached.

At present, the only known general way of producing the optimal finite wordlength FIR filter coefficients is by using the methods of mixed integer programming. These methods are well described in the literature (see, for example, [9],[10]) and there are several general-purpose computer programs which include integer programming algorithms. These programs are usually extensions of linear programming programs and can be used in much the same way as standard linear programming packages.

Application of general-purpose integer-programming computer programs to the design of optimal finite wordlength FIR digital filters is in many respects similar to application of linear programming techniques to the design of "infinite precision" FIR digital filters. It is therefore surprising that there are practically no papers[2] which would describe the use of integer programming programs for finite wordlength FIR design. These programs are by no means perfect for this purpose. Nevertheless, they produce the desired results and give insight into the somewhat obscure nature of optimal finite wordlength FIR filters.

II. STATEMENT OF THE PROBLEM

Using the standard notation [11], we write the frequency response $H(f)$ of an N length finite wordlength FIR linear phase digital filter as

$$H(f) = \sum_{k=0}^{N-1} h(k) e^{-j2\pi kf} \tag{1}$$

where $h(k)$, $k = 0, 1, \cdots, N-1$, are b-bit (sign bit included) filter coefficients. It can be shown [3], [11] that $H(f)$ can always be written as

Manuscript received December 22, 1978; revised September 18, 1979.

The author is with the Department of Electrical Engineering and Computer Science, University of Ljubljana, Ljubljana, Yugoslavia, on leave at the Department of Electrical Engineering and Computer Science, Princeton University, Princeton, NJ 08540.

[1]Throughout this paper the word optimal is used in a Chebyshev, or minimax, sense.

[2]The author has recently become aware of a paper [14] in which optimal CCD transversal filters were designed using mixed-integer programming techniques.

$$H(f) = G(f)\exp\left[j\frac{L\pi}{2} - \frac{(N-1)}{2}2\pi f\right] \quad (2)$$

where $G(f)$ is a real valued function and $L = 0$ or 1. There are exactly four cases of function $G(f)$ which can be expressed in the terms of b-bit coefficients $h(k)$ as

Case 1: N odd, $L = 0$

$$G(f) = h(n) + \sum_{k=1}^{n} h(n-k)2g(kf), \quad n = (N-1)/2 \quad (3)$$

$$g(k,f) = \cos(2\pi kf).$$

Cases 2, 3, and 4: N even, $L=0$; N odd, $L=1$; N even, $L=1$

$$G(f) = \sum_{k=1}^{n} h(n-k)2g(kf), \quad \begin{array}{ll} n = N/2 & N \text{ even} \\ n = (N-1)/2 & N \text{ odd} \end{array}$$

$$g(k,f) = \cos(2\pi(k-\frac{1}{2})f), \quad N \text{ even}, \quad L = 0$$

$$g(k,f) = \cos(2\pi kf), \quad N \text{ odd}, \quad L = 1$$

$$g(k,f) = \cos(2\pi(k-\frac{1}{2})f), \quad N \text{ even}, \quad L = 1. \quad (4)$$

The optimal b-bit wordlength linear-phase FIR filter design problem can now be stated as follows: given the number of bits b, the desired frequency response $D(f)$, and a positive weight function $W(f)$, both continuous on a compact subset $F \subset [0, \frac{1}{2}]$, and one of the four forms of $G(f)$, find the set of b-bit coefficients $h(n)$ which minimizes the maximum-weighted absolute error defined as

$$\| E(f) \| = \max_{f \in F} W(f) \mid D(f) - G(f) \mid . \quad (5)$$

It is instructive to compare this design problem with an infinite precision one. The inclusion of the b-bit restriction into the design problem formulation may not seem very severe at first. And indeed, one may even think that it might simplify the procedure since it is now obviously possible to enumerate all possible filters. It is, however, easy to see that such an approach fails even for a very small number of coefficients. It is also easy to see that it now becomes impossible to apply the Remez algorithm to the minimization of (5) since the alternation theorem does not hold anymore. Linear programming cannot be applied either since it does not allow the inclusion of b-bit coefficient constraints.

It is not the purpose of this paper to go into the intricate mathematical details of finding solutions to the above problem. Instead we shall use a general-purpose integer-programming package, which requires the following equivalent formulation of the problem [4]:

minimize E subject to constraints

$$h(n) + \sum_{k=1}^{n} h(n-k)2g(kf) - E/W(f) \le D(f)$$
$$-h(n) - \sum_{k=1}^{n} h(n-k)2g(kf) - E/W(f) \le -D(f) \quad f \in F \quad (6)$$

and

$h(k), \quad k = 0, 1, \cdots, n$ b-bit numbers (sign included).

We used form (3) of $G(f)$ to illustrate the problem. Each of the b-bit coefficients $h(k)$ is a b-bit binary number which can occupy one of the 2^b different values linearly distributed between some lower and upper bound. But since it is known that $\mid h(k) \mid \le 1, \ k = 0, 1, \cdots, n$, for all nonamplifying (output power lower or equal to input power) digital filters, we can without loss of generality express $h(k)$ as multiples of $2^{-(b-1)}$ bounded by

$$0 \le \mid h(k) \mid \le (2^{b-1} - 1)2^{-(b-1)}, \qquad k = 0, 1, \cdots, n. \quad (7)$$

Most of the integer-programming computer programs require that the discrete variables are nonnegative bounded integers. We shall therefore introduce substitution

$$h^*(k) \in 2^{b-1}(h(k) + 1), \qquad k = 0, 1, \cdots, n \quad (8)$$

and replace the formulation (6) by

minimize E subject to constraints

$$h^*(n) + \sum_{k=1}^{n} h^*(k-n)2g(kf) - E2^{b-1}/W(f) \le 2^{b-1}(D(f) + 1 + \sum_{k=1}^{n} 2g(kf)) \qquad f \in F$$

$$-h^*(n) - \sum_{k=1}^{n} h^*(k-n)2g(kf) - E2^{b-1}/W(f) \le -2^{b-1}(D(f) + 1 + \sum_{k=1}^{n} 2g(kf)) \quad (9)$$

and

$$h^*(k) \in (1, 2, 3, \cdots, 2^b - 1), \quad k = 0, 1, \cdots, n.$$

It is easy to see the equivalence of formulations (5), (6), and (9). The integers $h^*(k)$ which minimize (9) produce through (8) rationale $h(k)$ which minimize (6) and (5), and vice versa. We could also use the form (4) instead of (3) and show that the development applies to all four cases of FIR filters.

III. Description of Computer Implementation

The MPOS (multipurpose optimization system) program on CDC Cyber 72 was used for solving the problem (9). We also used the program of McClellan, Parks, and Rabiner [11] for generating the constraints in (9), using the same frequency grid F, and for comparison purposes. In fact, we added two additional subroutines into this program. The first subroutine generates the data file for MPOS. The data file is actually the formulation (9) written in the format required by MPOS. The second subroutine generates the program file for MPOS. The program file defines the type of the problem (mixed integer, minimize), chooses one of three available algorithms, and specifies the data file.

The same set of input data as in [11], with the addition of a number of bits b as the sixth parameter on the first card, was used. Parameter JPUNCH was used to choose one or none of the three MPOS algorithms. A typical run would produce a printout of infinite precision coefficients, b-bit rounded coeffi-

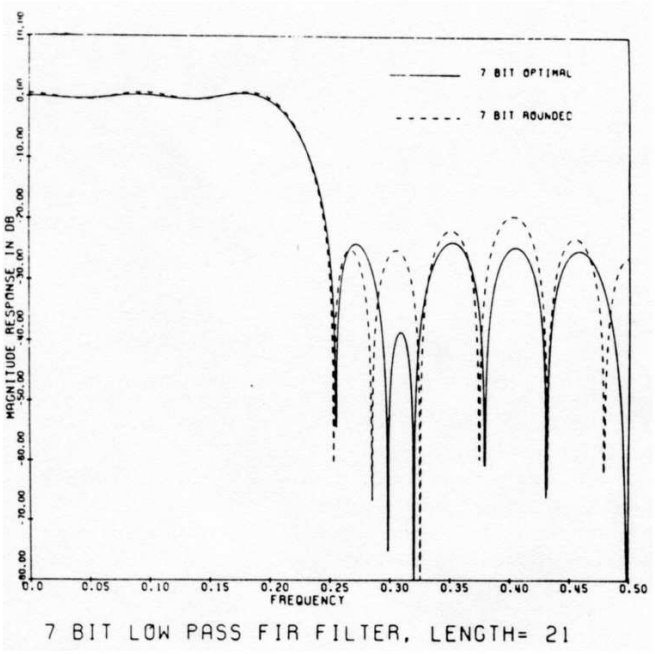


Fig. 1. Magnitude response and coefficients for $N = 21$, $b = 7$ low-pass filter.

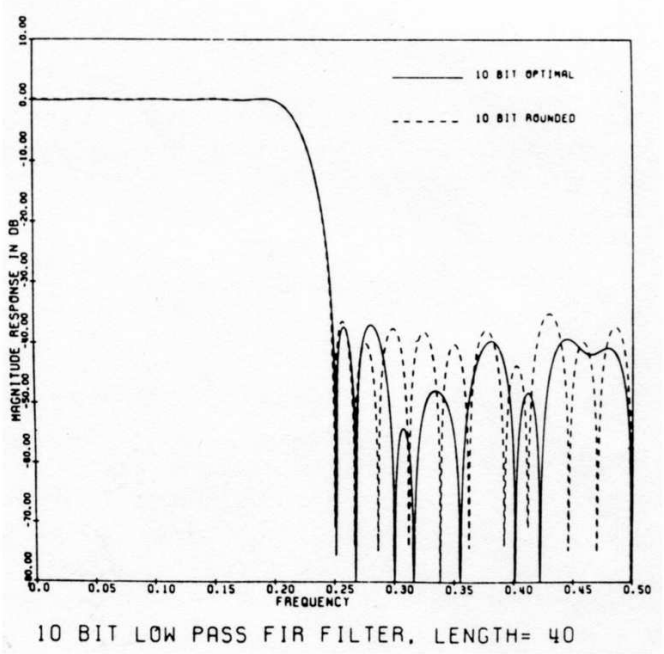


Fig. 2. Magnitude response and coefficients for $N = 40$, $b = 10$ low-pass filter.

Optimal 7 bit coefficients multiplied by $2^6 = 64$:

$h(0) = 2 = h(20)$
$h(1) = 0 = h(19)$
$h(2) = -2 = h(18)$
$h(3) = -1 = h(17)$
$h(4) = 2 = h(16)$
$h(5) = 3 = h(15)$
$h(6) = -3 = h(14)$
$h(7) = -6 = h(13)$
$h(8) = 3 = h(12)$
$h(9) = 20 = h(11)$
$h(10) = 28$

Rounded 7 bit coefficients multiplied by $2^6 = 64$:

$h(0) = 3 = h(20)$
$h(1) = 0 = h(19)$
$h(2) = -2 = h(18)$
$h(3) = -1 = h(17)$
$h(4) = 2 = h(16)$
$h(5) = 3 = h(15)$
$h(6) = -3 = h(14)$
$h(7) = -6 = h(13)$
$h(8) = 3 = h(12)$
$h(9) = 20 = h(11)$
$h(10) = 29$

Optimal 10 bit coefficients multiplied by $2^9 = 512$:

$h(0) = 2 = h(39)$
$h(1) = 2 = h(38)$
$h(2) = -1 = h(37)$
$h(3) = -4 = h(36)$
$h(4) = 0 = h(35)$
$h(5) = 6 = h(34)$
$h(6) = 1 = h(33)$
$h(7) = -8 = h(32)$
$h(8) = -5 = h(31)$
$h(9) = 8 = h(30)$
$h(10) = 10 = h(29)$
$h(11) = -9 = h(28)$
$h(12) = -17 = h(27)$
$h(13) = 5 = h(26)$
$h(14) = 27 = h(25)$
$h(15) = 2 = h(24)$
$h(16) = -43 = h(23)$
$h(17) = -25 = h(22)$
$h(18) = 92 = h(21)$
$h(19) = 211 = h(20)$

Rounded 10 bit coefficients multiplied by $2^9 = 512$:

$h(0) = 1 = h(39)$
$h(1) = 4 = h(38)$
$h(2) = -2 = h(37)$
$h(3) = -4 = h(36)$
$h(4) = 0 = h(35)$
$h(5) = 6 = h(34)$
$h(6) = 2 = h(33)$
$h(7) = -7 = h(32)$
$h(8) = -5 = h(31)$
$h(9) = 8 = h(30)$
$h(10) = 10 = h(29)$
$h(11) = -8 = h(28)$
$h(12) = -17 = h(27)$
$h(13) = 5 = h(26)$
$h(14) = 27 = h(25)$
$h(15) = 3 = h(24)$
$h(16) = -44 = h(23)$
$h(17) = -24 = h(22)$
$h(18) = 92 = h(21)$
$h(19) = 212 = h(20)$

cients, and b-bit optimal coefficients. The latter one was extracted from the MPOS output by a special program.

IV. Results

More than 50 optimal b-bit wordlength FIR digital filters were synthesized. As expected, the run time was very unpredictable and in some cases up to several hundred times longer than the run time for an equivalent infinite precision or rounded case.

Among three algorithms provided by MPOS, the branch and bound algorithm was by far the most successful one.[3] The other two algorithms, Direct Search 0-1 and Gomory, failed in most cases and were abandoned after initial attempts.

Due to the memory constraints on our particular machine it was possible to design only filters up to length 40 (with grid density LGRID = 8). This limitation was not a major problem in this work. It was much more difficult to provide the huge amounts of computer time that were necessary for this project.

Most of the design cases (all mentioned in this paper) were done for the low-pass filters defined as

$$\begin{aligned} &\text{passband} \quad D(f) = 1,\ W(f) = 1, \qquad 0 \le f \le 0.20 \\ &\text{stopband} \quad D(f) = 0,\ W(f) = 1, \quad 0.25 \le f \le 0.5. \end{aligned} \tag{10}$$

[3]Given enough computer time, the branch and bound algorithm was successful in all design cases.

Figs. 1 and 2 show specific examples of the use of the design program for $N = 21$, $b = 7$ and $N = 40$, $b = 10$ filters. These particular examples were reported in [12] and the comparison with the filters with rounded coefficients reveals a considerable improvement.[4]

It is also interesting to observe the closeness between the optimal and rounded coefficients. This observation was typical for all design cases and the difference between the optimal and rounded coefficients multiplied by 2^{b-1} was never greater than 4 for all N between 6 and 40 and all b between 3 and 15.

V. Optimal Wordlength Choice

During the implementation of a digital filter, one is often interested in the following question: given the desired frequency response, it is better to implement the filter with a small number of bits b and a large number of coefficients N, or vice versa?

[4]A comparison with suboptimal search algorithms is given in [20].

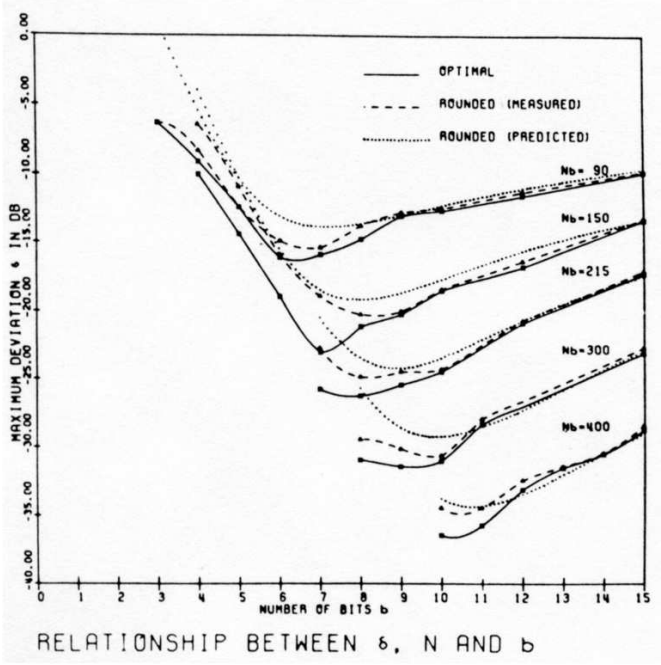


Fig. 3. Relationship between δ, N, and b for 36 low-pass filters specified by (10).

The answer will not only depend on the particular type of implementation, but also on the relation between b and N. We shall try to given some insight into this relation.

Let us first look at the case of filters with rounded coefficients. Using the approximate design formula for optimal infinite-precision low-pass FIR filters [13] and a statistical upper bound for error caused by the rounding [1], it is possible to get an approximate relationship between deviation $\delta = \max \mid D(f) - G(f) \mid$, number of coefficients N, and coefficients wordlength b. For the filters defined in (10) the transition band ΔF equals 0.05 and we would have

$$\delta \approx 2^{-(b-1)} \sqrt{\frac{2N-1}{3}} + \delta_\infty \tag{11}$$

where δ_∞ is the infinite precision deviation which can be calculated from the formula

$$\begin{aligned}(N-1)\Delta F \approx{}& 0.005309(\log_{10}\delta_\infty)^3 + (0.07114 - 0.00266) \\ & \cdot (\log_{10}\delta_\infty)^2 + (-0.4761 - 0.5941)\log_{10}\delta_\infty \\ & -0.4278 - 11.012(\Delta F)^2. \end{aligned} \tag{12}$$

The formula (12) becomes inaccurate for $\delta_\infty > 0.1$ and the correction suggested in [13] can be used in these cases.

No such relationship is known for optimal finite wordlength FIR filters. Since we know very little about the properties of these filters and since the attractive idea of a very short wordlength implementation, which is obviously impossible in the rounded case, seems feasible in the optimal case, we decided to study the relationship experimentally.

Fig. 3 shows the experimentally obtained relationship between δ , N, and b for 36 optimal finite wordlength low-pass FIR filters defined in (10), and for the equivalent filters with rounded coefficients. The values of N and b were chosen for 5 different values of product Nb. One could of course argue against the choice of Nb as a parameter since it is not the best criterion of hardware complexity. However, it was chosen because of the following.

1) The product Nb gives the number of bits that describe the filter and thus serves as a measure of information capacity of the filter.

2) It is convenient for computation purposes and allows an easy comparison with other criteria.

The results in Fig. 3 reveal that there is an optimal number of bits b for each value of Nb. It is interesting to compare these results with the results for equivalent rounded filters. The comparison shows that the optima move approximately one bit to the right in the rounded case and that the deviation predicted by (11) and (12) agrees closely with the measured values.

The results also indicate that it is probably not possible, in general, to substitute a small number of bits in coefficients by a higher number of coefficients and produce the same deviation. Several separate experiments confirmed this result but at the moment it is still too early to give a definite conclusion.

VI. Concluding Remarks

The results presented show that it is possible, in general, to design optimal finite wordlength FIR filters using general-purpose integer-programming techniques. The branch and bound algorithm proved to be successful in solving the approximation problem. It is, however, also very costly in terms of computer time and there is no doubt that better, more efficient methods will have to be found in order to make optimal finite wordlength FIR filters attractive to designers.

Acknowledgment

The author would like to express his thanks to Prof. K. Steiglitz, without whose encouragement this paper would not have been written, and to S. Krkic for his cooperation on the programming.

References

[1] D. S. K. Chan and L. R. Rabiner, "Analysis of quantization errors in the direct form for finite impulse response digital filters," *IEEE Trans. Audio Electroacoust.,* vol. AU-21, pp. 354–366, Aug. 1973.

[2] R. E. Crochiere, "A new statistical approach to the coefficient wordlength problem for digital filters," *IEEE Trans. Circuits Syst.,* vol. CAS-22, pp. 190–196, Mar. 1975.

[3] T. W. Parks and J. H. McClellan, "Chebyshev approximation for nonrecursive digital filters with linear phase," *IEEE Trans. Circuit Theory,* vol. CT-19, pp. 189–194, May. 1972.

[4] L. R. Rabiner, "Linear programming design of finite impulse response (FIR) digital filters," *IEEE Trans. Audio Electroacoust.,* vol. AU-20, pp. 280–288, Oct. 1972.

[5] E. Avenhaus, "On the design of digital filters with coefficients of limited wordlength," *IEEE Trans. Audio Electroacoust.,* vol. AU-20, pp. 206–212, Aug. 1972.

[6] K. Steiglitz, "Designing short-word recursive digital filters," in *Proc. 9th Annu. Allerton Conf. Circuits Syst. Theory,* Oct. 6-8, 1971, pp. 778–788.

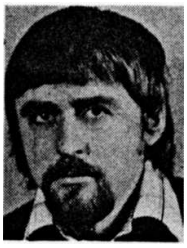
[7] C. Charalambos and M. J. Best, "Optimization of recursive digital filters with finite wordlength," *IEEE Trans. Acoust., Speech, SignalProcessing,* vol. ASSP-22, pp. 424–431, Dec. 1974.

[8] M. Suk and S. K. Mitra, "Computer-aided design of digital filters with finite wordlength," *IEEE Trans. Audio Electroacoust.,* vol. AU-20, pp. 356–363, Dec. 1972.

[9] H. M. Salkin, *Integer Programming.* Reading, MA: Addison-Wesley, 1975.

[10] A. Kaufman and A. Henry-Labordere, *Integer and Mixed Programming Theory and Applications.* New York: Academic, 1977.

[11] J. H. McClellan, T. W. Parks, and L. R. Rabiner, "A computer program for designing optimum FIR linear phase digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 506–526, Dec. 1973.
[12] D. Kodek and S. Krkic, "Synthesis of optimal quantized nonrecursive digital filters (in Slovene)," in *Proc. Informatica Int. Symp.*, Bled, Yugoslavia, Oct. 1976, pp. 3-119.
[13] O. Herrmann, L. R. Rabiner, and D. S. K. Chan, "Practical design rules for optimum finite impulse response lowpass digital filters," *Bell Syst. Tech. J.*, vol. 52, pp. 769–799, July-Aug. 1973.
[14] Y. Chen, S. M. Kang, and T. G. Marshall, "The optimal design of CCD transversal filters using mixed-integer programming techniques," in *Proc. 1978 IEEE Int. Symp. Circuits Syst.*, New York, May 17-19, 1978, pp. 748–751.
[15] L. Corgnier and R. Rocci, "Sintesi di filtei numerici di tipo trasversale con coefficienti quantisati e con riposta in frequenza prescritta mediante maschera," Centro Studie Laboratori Teleconrmunicazioni, Torino, Italy, Rap. Tec., vol. II (1974), pp. 27–35.
[16] F. Grenez, "Design of FIR linear phase digital filters to minimize the statistical wordlength of the coefficients," *IEE J. Electron. Circuits Syst.*, vol. 1, pp. 181–185, 1977.
[17] P. Chambon and A. Desblache, "Integer coefficients optimization of digital filters," in *Proc. IEEE Int. Symp. Circuits Syst.*, Munich, Germany, 1976, pp. 461–464.
[18] U. Heute, "Recent developments in digital FIR filtering," in *Proc. 5th Summer Symp. Circuit Theory*, Kladno (CSSR), 1977, pp. 14–27.
[19] W. H. Storzbach, "On the design of recursive digital filters with minimum coefficient length," in *Proc. Int. Symp. Circuit Theory*, Apr. 1972, pp. 279–282.
[20] D. Kodek and K. Steiglitz, "Comparison of optimal and local search methods for designing finite word length FIR digital filters," in *Proc. 13th Annu. Johns Hopkins Conf Inform Sci. Syst.*, Baltimore, MD, Mar. 28–30, 1979.



Dušan M. Kodek (M'79) was born in Ljubljana, Yugoslavia, on February 27, 1946. He received the B.S.E.E., M.S.E.E., and Sc.D. degrees from the University of Ljubljana, Ljubljana, Yugoslavia, in 1970, 1973, and 1975, respectively.

Since 1971 he has been an Assistant Professor with the Department of Electrical Engineering and Computer Science, University of Ljubljana. His current research interests are in digital signal processing, process control, and microprocessor applications.