INTERNAL REPORTS IN

SIMULATION, OPTIMIZATION AND CONTROL

No. SOC-36

DESIGN OF RECURSIVE DIGITAL FILTERS WITH OPTIMIZED WORD LENGTH COEFFICIENTS

J.W. Bandler, B.L. Bardakjian and J.H.K. Chen

April 1974

(Revised March 1975)

FACULTY OF ENGINEERING
McMASTER UNIVERSITY
HAMILTON, ONTARIO, CANADA

Design of recursive digital filters with optimized word length coefficients

J. W. BANDLER and B. L. BARDAKJIAN
(Group on Simulation, Optimization and Control and Department of Electrical Engineering, McMaster University, Hamilton, Canada)

J. H. K. CHEN
(Bell-Northern Research, Ottawa, Canada)

The problem of designing recursive digital filters with optimized word length coefficients to meet arbitrary, prescribed magnitude characteristics in the frequency domain is numerically investigated. The continuous nonlinear programming problem is formulated as an unconstrained minimax problem using the Bandler–Charalambous approach, and Dakin's branch-and-bound technique is used in conjunction with Fletcher's unconstrained minimization program to discretize the continuous solution. The objective function to be minimized is directly concerned with the word lengths of the coefficients, which are also introduced as variables.†

(Received in revised form on March 17th 1975)

INTRODUCTION

The problem of designing recursive digital filters with *a priori* specified finite word length for the representation of the coefficients, can be formulated as a nonlinear discrete optimization problem. Many approaches using random search optimization algorithms were proposed to solve this problem[1–3]. Störzbach[4] proposed a zero-one programming approach. Recently, Charalambous and Best[5] proposed an approach using a branch-and-bound technique in conjunction with an optimization algorithm for linearly constrained problems. Another approach is to formulate the continuous nonlinear programming problem as an unconstrained minimax problem[6], and use Dakin's branch-and-bound technique[7] in conjunction with Fletcher's unconstrained minimization program[8], to discretize the solution. An example using this approach is given. The main features reported in[6–8] have been implemented in a general computer program package called DISOPT[9].

Because the cost of a digital filter, if implemented as a special-purpose computer, depends heavily on the word length of the coefficients, it should be reduced as much as possible[1]. On the other hand, when the coefficients of a digital filter, initially specified with unlimited accuracy, are quantized by rounding or truncation, then coefficient quantization error occurs which affects the digital filter's response[10]. Therefore, it is desirable to incorporate the word lengths as additional parameters of the approximation problem in recursive digital filter design.

The problem of designing recursive digital filters with optimized word length coefficients to meet arbitrary, prescribed magnitude characteristics in the frequency domain, is formulated as a nonlinear integer programming problem, where the parameter vector consists essentially of the word lengths of the coefficients and the multipliers of the quantization step sizes. A function of the word lengths is minimized, subject to the prescribed constraints on the magnitude characteristics, while constraining all the constituents of the parameter vector to be integers.

DESCRIPTION OF THE PROBLEM

Suppose that the magnitude characteristics of a recursive digital filter, whose coefficients can be represented exactly using finite word lengths, are specified to lie within given upper and/or lower bounds at a prescribed discrete set of frequencies $f_1, f_2, \ldots, f_m$, corresponding to a discrete set of values of the variable z evaluated on the unit circle in the z domain:

$$z_i = e^{j\psi_i \pi} \quad i = 1, 2, \ldots, m \tag{1}$$

where

$$\psi_i = \frac{2f_i}{f_s} \qquad i = 1, 2, \ldots, m \tag{2}$$

and f_s is the sampling frequency.

One approach to this problem is to specify the word lengths required to represent the coefficients, and optimize the magnitude characteristics of the filter. Another approach is to optimize the word lengths required to represent the coefficients subject to the constraints that the magnitude characteristics lie within the specified upper and/or lower bounds.

We consider the transfer function to be of the cascade form with K second order sections, namely,

$$H(z) = A \sum_{k=1}^{K} \frac{1 + a_k z^{-1} + b_k z^{-2}}{1 + c_k z^{-1} + d_k z^{-2}} \tag{3}$$

If necessary, stability constraints may be dealt with either by the pole inversion approach[11,12] or by imposing the appropriate set of linear inequalities on c_k and d_k [2,5].

† This paper was presented at the Eighth Annual Princeton Conference on Information Sciences and Systems, Princeton, N.J., March 28–29, 1974.

PROBLEM FORMULATION

We let a coefficient, which is to be represented exactly using a finite word length, be

$$r_i 2^{-q_i}, \; i = 1, \ldots, n,$$

where r_i is an integer, q_i is a non-negative integer, 2^{-q_i} is the coefficient quantization step-size and $q_i + 1$ is the word length.

Case 1: a priori specified word lengths

Let

$$\underline{\phi}' = \begin{bmatrix} a_1 \\ b_1 \\ c_1 \\ d_1 \\ a_2 \\ b_2 \\ c_2 \\ d_2 \\ \cdot \\ \cdot \\ \cdot \\ A \end{bmatrix} \qquad (4)$$

which contains the n coefficients of the recursive digital filter. We want to find the coefficients given the coefficient quantization step sizes, to minimize an appropriately chosen objective function comprising the deviations of the response from its prescribed upper and lower bounds $S_u(\psi)$ and $S_l(\psi)$, respectively. Thus, the optimization of response is subject to

$$\phi_i'/2^{-q_i} \in I, \quad i = 1, 2, \ldots, n \qquad (5)$$

where $n = 4K + 1$ and I is the set of integers.

Case 2: optimum word lengths

Find an optimum n-dimensional grid having at least one element which also belongs to a specified region in the n-dimensional coefficient space, i.e., find

$$\underline{\phi} = \begin{bmatrix} q_1 \\ q_2 \\ \cdot \\ \cdot \\ \cdot \\ q_n \\ r_1 \\ r_2 \\ \cdot \\ \cdot \\ \cdot \\ r_n \end{bmatrix} \qquad (6)$$

to minimize the objective function

$$U(q_1, q_2, \ldots, q_n)$$

subject to

$$S_l(\psi) \leqslant |H(\underline{\phi}, \psi)| \leqslant S_u(\psi) \qquad (7)$$

and

$$q_i, r_i \in I, \quad i = 1, 2, \ldots, n \qquad (8)$$

where n, $S_u(\psi)$, $S_l(\psi)$ and I are as in Case 1.

THE PROGRAM DISOPT

DISOPT is a user-oriented computer program in FORTRAN 4 for solving continuous or discrete, constrained or unconstrained general optimization problems. Many recently proposed algorithms and techniques for nonlinear programming which have been reported to be efficient have been incorporated. This allows the user to fully employ some of the latest developments.

Two approaches are available in DISOPT to transform a constrained problem into an equivalent unconstrained objective. The first approach is the minimax approach proposed by Bandler and Charalambous[6]. This can be implemented by the following least pth approximation algorithms:

1. A least pth optimization with a large value of p [13].
2. A sequence of least pth optimizations with increasing values of p [13].
3. A sequence of least pth optimizations with geometrically increasing values of p in conjunction with an extrapolation technique[14].
4. A sequence of least pth optimizations with finite values of p [15].

The second approach or Algorithm 5 is a modification of an existing nonparametric exterior-point method described by Lootsma[16]. The quasi-Newton algorithm due to Fletcher[8] is then employed to perform the minimization.

The solution of a discrete problem follows the logic of the Dakin tree-search algorithm[7]. The discrete variables are forced to assume discrete values by automatically introducing additional variable constraints after the continuous solution is obtained.

Some of the options available in DISOPT to enhance the efficiency of the program are:

1. In the search for the optimum discrete solution, the new variable constraint added at each node always excludes the preceding optimum point from the current solution space. The constraint is therefore active if the function is locally unimodal. Thus, the value of the variable under the new constraint may be optionally fixed on the constraint boundary. Hence, a problem with one less parameter must be solved and the computational effort would be reduced.
2. To obtain an initial upper bound on the objective function for a discrete problem in order to avoid the search-

ing of some unlikely subtrees, DISOPT may be asked to check the nearest set of discrete solutions about the continuous optimum and store the best feasible solution.

3. If the constraints, including an upper bound on the objective, cannot be satisfied at the optimum of the least pth objective with any value of p greater than unity, then no feasible solution is attainable for all permissible values of p. Therefore, the user may request DISOPT to check the existence of a feasible solution before doing the actual minimization. This is particularly advantageous in the case of a discrete problem where the additional variable constraints may conflict with some of the original constraints on the continuous problem.
4. In case of multiple optimum discrete solutions, the user has the option of requesting only one solution to reduce the necessary computation time.
5. DISOPT may be optionally asked to check the derivatives of the objective function and the constraint functions at the starting point by numerical perturbation.

DISOPT can handle discrete problems of uniform as well as non-uniform quantization step sizes. The amount of programming effort required of a user has been reduced to a minimum. The user is responsible only for supplying the values and/or proper dimensioning of the parameters in the argument list and writing two service subroutines to define the objective function, the constraints and their respective partial derivatives. A documented listing of DISOPT is available from the first author at nominal charge[9].

EXAMPLES

Example 1: low-pass 7-bit filter

We consider, with f_s = 10 kHz, the following amplitude specifications:

f = 0,900 (100)	$S(f) = 1$
f = 1000	$S(f) = 1/\sqrt{2}$
f = 1200	$S(f) = 0$
f = 1500, 5000 (500)	$S(f) = 0$

TABLE 1. Results for example 1

Parameters	Suk and Mitra	DISOPT early solution	DISOPT final solution
a_1	−0.25	−0.296875	−0.328125
b_1	1.3125	1.015625	1.015625
c_1	−1.4375	−1.4375	−1.453125
d_1	0.65625	0.640625	0.65625
A	0.09375	0.109375	0.109375
Objective function	0.31535	0.29138	0.29059
Maximum error	0.41345	–	0.36685
Number of function evaluations	139	306	574 (terminated at 1030)

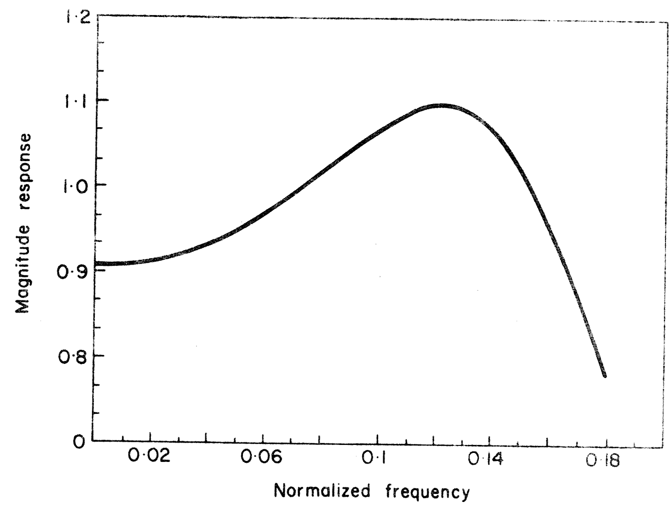

FIGURE 1. Passband response for example 1

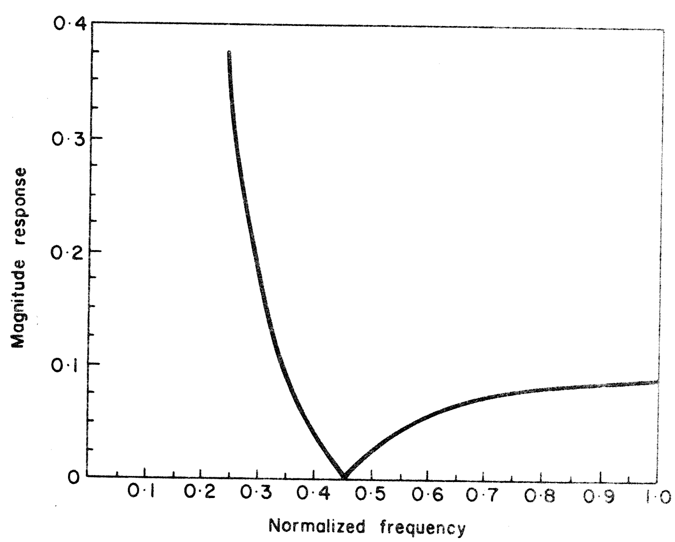

FIGURE 2. Stopband response for example 1

Using one section and the starting point of Suk and Mitra[3], namely, $[0\ 1\ -1\ 0.5\ 0.1]^T$ with the Case 1 formulation taking

$$U(\underline{\phi}') = \sum_{i=1}^{20} (|H(\underline{\phi}', \psi_i)| - S(\psi_i))^2 \qquad (9)$$

DISOPT gave the results shown in Table 1 and Figures 1 and 2 in less than 60 s on the CDC 6400 computer, using Algorithm 1 with $p = 10^7$ and options 1–5. Good solutions are obtained relatively soon in the optimization process as the table shows.

Example 2

Find an optimum grid having at least one element which also belongs to

$$R \triangleq \{ x_1, x_2 | 0.2 \leqslant x_1 \leqslant 0.4,\ 0.2 \leqslant x_2 \leqslant 0.8 \}.$$

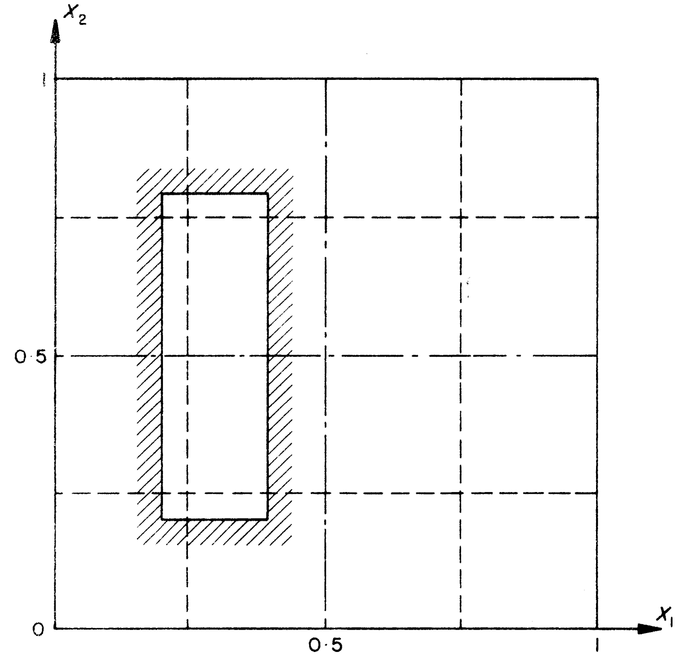

FIGURE 3. A representation of example 2

TABLE 2. Results for example 2

Node number	Objective function	Solution $\{q_1, q_2, r_1, r_2\}$	Description
0	0	{0, 0, 0.2, 0.781}	continuous
1	–	–	nonfeasible
2	1.322	{1.322, 0, 1, 0.358}	feasible
3	–	–	nonfeasible
4	2	{2, 0, 1.201, 0.51}	feasible
5	2	{2, 0, 1, 0.51}	feasible
6	–	–	nonfeasible
7	2.322	{2, 0.322, 1, 1}	feasible
8	–	–	nonfeasible
9	3	{2, 1, 1, 1.554}	feasible
10	3	{2, 1, 1, 1}	discrete
11	⩾3	–	abandoned
12	2.322	{2.322, 0, 2, 0.28}	feasible
13	–	–	nonfeasible
14	⩾3	–	abandoned

See Figure 3 for an illustration.

Let

$$x_1 = r_1 2^{-q_1}$$

$$x_2 = r_2 2^{-q_2}$$

Starting with $[1\ 1\ 1\ 1]^T$ using the Case 2 formulation taking

$$U(\underline{\phi}) = q_1 + q_2$$

DISOPT produced the results shown in Table 2 and Figure 4 in approximately 20 s on the CDC 6400, using Algorithm 3 with a third order extrapolation, initial value of $p = 4$, multiplying factor for p of 4 and options 2–5. In Figure 5 we have plotted contours of the minimax function[6] as incorporated into DISOPT for $r_2 = q_2 = 1$. The objective function value of 3 noted in Table 2, corresponding to $r_1 = 1$ and $q_1 = 2$, is clearly seen.

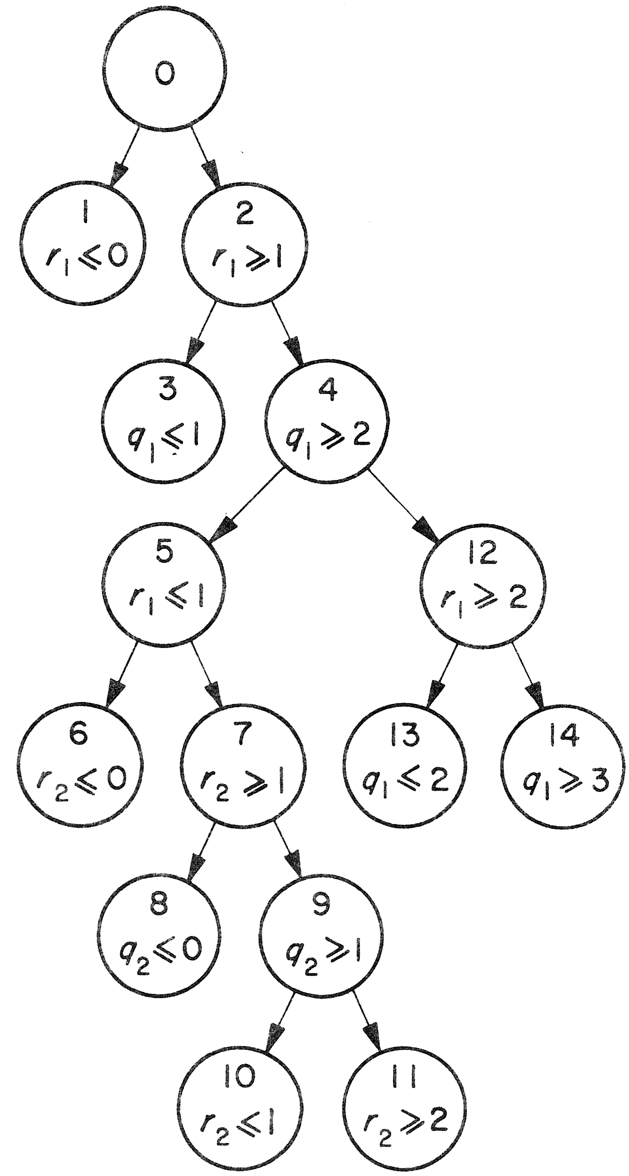

FIGURE 4. Tree structure for example 2

Example 3: low-pass optimized word length filter

Consider the following amplitude specifications:

$$\psi = 0, 0.18\ (0.02),\ S_u(\psi_i) = 1.3,\ S_l(\psi_i) = 0.7,$$

$$i = 1, 2, \ldots, 10$$

$$\psi = 0.24,\ S_u(\psi_i) = 0.3, \quad i = 11$$

$$\psi = 0.3, 1\ (0.1),\ S_u(\psi_i) = 0.3, \quad i = 12, 13, \ldots, 19.$$

Using one section and the starting point

$$[q\ r_1\ r_2\ r_3\ r_4\ r_5]^T = [1\ 0\ 2\ -2\ 1\ 0.2]^T$$

with the Case 2 formulation, where

$$q_1 = q_2 = \ldots = q_5 = q = U,$$

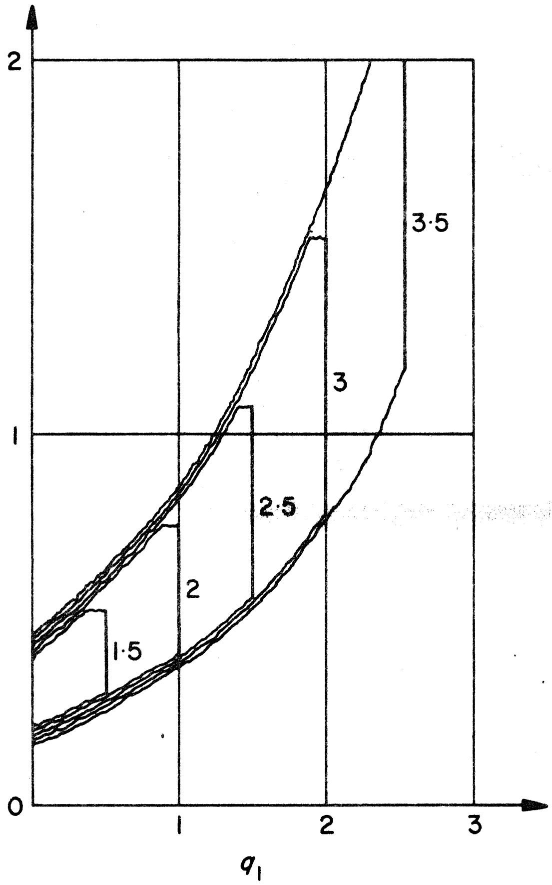

FIGURE 5. Contours of the unconstrained minimax objective function for example 2, with $r_2 = q_2 = 1$

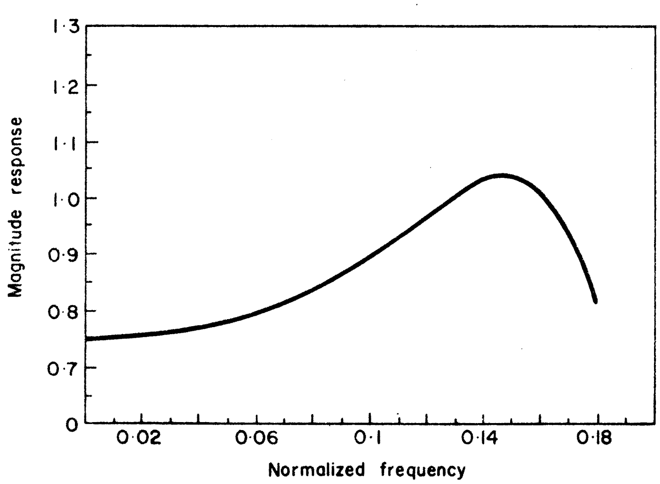

FIGURE 6. Passband response for example 3

DISOPT gave the solutions $[2\ -6\ 5\ -6\ 3\ 1]^T$, $[2\ -4\ 3\ -6\ 3\ 1]^T$ and $[2\ -5\ 4\ -6\ 3\ 1]^T$ using Algorithm 2 with the sequence of p values $\{2, 10, 10^2, 10^3, 10^4\}$ and options 2, 3 and 5. The corresponding coefficient sets are $\{-1.5, 1.25, -1.5, 0.75, 0.25\}$, $\{-1, 0.75, -1.5, 0.75, 0.25\}$ and $\{-1.25, 1, -1.5, 0.75, 0.25\}$, respectively. Figures 6 and 7 show the response for the last set. About 3 min computation time was required.

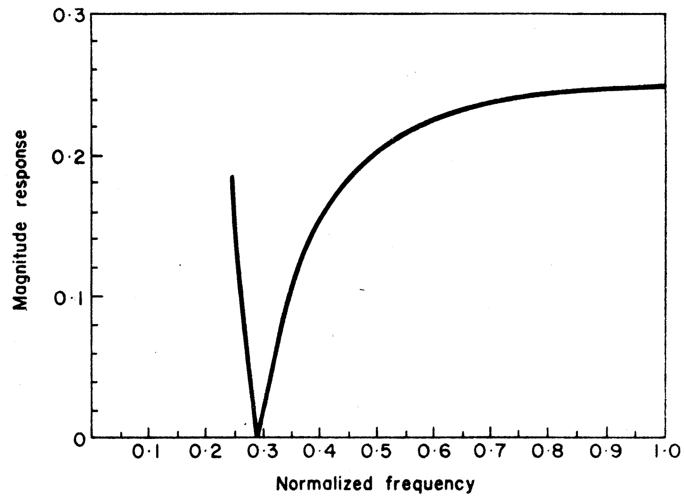

FIGURE 7. Stopband response for example 3

Example 4: low-pass optimized word length filter

This is the same as the last example except that all q_i's can vary and

$$U(\underline{\phi}) = q_1 + q_2 + q_3 + q_4 + q_5.$$

Starting with $[q_1 q_2 q_3 q_4 q_5 r_1 r_2 r_3 r_4 r_5]^T =$ $[2\ 0\ 1\ 2\ 2\ 0\ 1\ -2\ 2\ 0.4]^T$, DISOPT gave a solution $[2\ 0\ 1\ 2\ 2\ -5\ 1\ -3\ 3\ 1]^T$, which corresponds to the last coefficient set in Example 3 using the same algorithm and options as in Example 2. The solution was found in about 1 min but the program terminated after about 5 min.

CONCLUSIONS

The present approach to recursive digital filter design may be summarized as follows. First, a continuous feasible solution should be sought to determine the minimum necessary order of the filter. If the word lengths are specified, the best corresponding response would be sought using the Case 1 formulation. If the word lengths are to be optimized the Case 2 formulation may be used. Initially, a uniform, variable word length may be optimized. All feasible discrete solutions can be generated (the optimum word lengths solution being an element of this set), or we can stop after one discrete solution is found, allow the word lengths to differ and minimize a suitable function of these word lengths. Finally, if desired, the response corresponding to the optimum word lengths solution could be optimized using the Case 1 formulation.

ACKNOWLEDGEMENT

The authors thank Dr C. Charalambous, Dept. of Combinatorics and Optimization, University of Waterloo, Waterloo, Canada, for useful suggestions on optimization. The work was supported by the National Research Council of Canada under Grant A7239 and by a Scholarship to J. H. K. Chen.

REFERENCES

1 Avenhaus, E. 'On the design of digital filters with coefficients of limited word length', *IEEE Trans. Audio Electroacoust.*, Vol AU-20, (August 1972) pp 206–212

2 Steiglitz, K. *Designing short-word recursive digital filters*, Proc. 9th Allerton Conf. Circuit and System Theory, Urbana, Illinois, USA, October 1971, pp 778–788

3 Suk, M. and Mitra, S. K. 'Computer-aided design of digital filters with finite word lengths', *IEEE Trans. Audio Electroacoust.*, Vol AU-20, (December 1972) pp 356–363

4 Störzbach, W. H. *On the design of recursive digital filters with minimum coefficient word length*, Proc. Int. Symp. on Circuit Theory, North Hollywood, California, April 1972, pp 279–282

5 Charalambous, C. and Best, M. J. 'Optimization of recursive digital filters with finite word lengths', *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol ASSP-22, (December 1974) pp 424–431

6 Bandler, J. W. and Charalambous, C. 'Nonlinear programming using minimax techniques', *J. Optimiz. Theory and Appl.*, Vol 13, (June 1974) pp 607–619

7 Dakin, R. J. 'A tree-search algorithm for mixed integer programming problems', *Comput. J.*, Vol 8, (1966) pp 250–255

8 Fletcher, R. 'FORTRAN subroutines for minimization by quasi-Newton methods', Atomic Energy Research Establishment, Harwell, Berkshire, England, Report AERE-R7125, (1972)

9 Bandler, J. W. and Chen, J. H. K. 'DISOPT – a general program for continuous and discrete nonlinear programming problems', *Int. J. Syst Sci*, to be published. Also McMaster University, Hamilton, Canada, Internal Report in Simulation, Optimization and Control, No.SOC-29, March 1974 (full report by J. H. K. Chen)

10 Rabiner, L. R. *et al.*, 'Terminology in digital signal processing', *IEEE Trans. Audio Electroacoust.*, Vol AU-20, (December 1972) pp 322–337

11 Steiglitz, K. 'Computer-aided design of recursive digital filters', *IEEE Trans. Audio Electroacoust.*, Vol AU-18, (June 1970) pp 123–129

12 Bandler, J. W. and Bardakjian, B. L. 'Least *p*th optimization of recursive digital filters', *IEEE Trans. Audio Electroacoust.*, Vol AU-21, (October 1973) pp 460–470

13 Bandler, J. W. and Charalambous, C. 'Practical least *p*th optimization of networks', *IEEE Trans. Microwave Theory Tech.*, Vol MTT-20, (December 1972) pp 834–840

14 Chu, W. Y. 'Extrapolation in least *p*th approximation and nonlinear programming', McMaster University, Hamilton, Canada, Internal Report in Simulation, Optimization and Control, No.SOC-71, (December 1974)

15 Charalambous, C. and Bandler, J. W. 'New algorithms for network optimization', *IEEE Trans. Microwave Theory Tech.*, Vol MTT-21, (December 1973) pp 815–818

16 Lootsma, F. A. 'A survey of methods for solving constrained minimization problems via unconstrained minimization', *Numerical Methods for Non-Linear Optimization*, F. A. Lootsma, Ed. New York: Academic Press, (1972)

SOC-36

DESIGN OF RECURSIVE DIGITAL FILTERS WITH OPTIMIZED WORD LENGTH COEFFICIENTS

J.W. Bandler, B.L. Bardakjian and J.H.K. Chen

April 1974, No. of Pages: 6

Revised: March 1975

Key Words: Digital filters, discrete optimization, branch and bound methods, gradient minimization methods

Abstract: The problem of designing recursive digital filters with optimum word length coefficients to meet arbitrary, prescribed magnitude characteristics in the frequency domain is numerically investigated. The continuous nonlinear programming problem is formulated as an unconstrained minimax problem, and Dakin's branch and bound technique is used in conjunction with Fletcher's unconstrained minimization program to discretize the continuous solution. The objective function to be minimized is directly concerned with the word lengths of the coefficients, which are also introduced as variables.

Description: Reprint. Presented at the Eighth Annual Princeton Conference on Information Sciences and Systems (Princeton, N.J., March 1974). Published in Computer Aided Design, vol. 7, July 1975, pp. 151-156.

Related Work: SOC-29, SOC-93, SOC-113.

Price: $ 2.00.