Computer-Aided Design of Recursive Digital Filters

KENNETH STEIGLITZ, Member, IEEE
Department of Electrical Engineering
Princeton University
Princeton, N. J. 08540

Abstract

A practical method is described for designing recursive digital filters with arbitrary, prescribed magnitude characteristics. The method uses the Fletcher–Powell optimization algorithm to minimize a square-error criterion in the frequency domain. A strategy is described whereby stability and minimum-phase constraints are observed, while still using the unconstrained optimization algorithm. The cascade canonic form is used, so that the resultant filters can be realized accurately and simply. Design examples are given of low-pass, wide-band differentiator, linear discriminator, and vowel formant filters.

I. Introduction

While the problem of choosing the coefficients of a nonrecursive digital filter to approximate a specified magnitude characteristic has been thoroughly explored, the corresponding problem for recursive digital filters remains open [1], [2]. Design procedures for recursive filters generally deal only with the piecewise constant case, and involve transformations of well known continuous-time filter designs, such as the Butterworth or Chebyshev. The purpose of this paper is to describe a practical method for choosing the coefficients of a recursive digital filter to meet arbitrary specifications of the magnitude characteristic.

The proposed method uses the optimization algorithm described by Fletcher and Powell [3] to minimize a square-error criterion in the frequency domain. This technique has been used to design continuous-time filters [4]. In order to deal with the realization problem in the continuous-time case, a network topology is usually fixed, and the optimization method must incorporate the constraints that the element values be nonnegative. These restrictions are not present for digital filters, since any coefficients can be used for realization. The resulting digital filter must be stable, however, and this imposes the constraint that the poles lie inside the unit circle in the z-plane. It will be shown how this constraint, and an additional minimum-phase constraint, can be observed, while still using the unconstrained minimization method of Fletcher and Powell.

II. Choice of Canonic Form

The first important question to be resolved is the choice of the canonic form of the digital filter. A general recursive filter can be assumed to have the transfer function

$$Y(z) = \frac{\sum_{k=1}^{K} a_k z^{-(k-1)}}{1 + \sum_{k=1}^{N} b_k z^{-k}} \cdot \qquad (1)$$

This so-called direct form suffers from the following difficulties. First, if control is to be exercised over the pole locations, the denominator must be factored at certain stages in the optimization process. Second, the pole locations may be extremely sensitive functions of the coefficients b_k for high-order filters [1]. This means that the b_k must be found to very high precision, and that the error surface may be badly skewed. The cascade form

$$Y(z) = A \prod_{k=1}^{K} \frac{1 + a_k z^{-1} + b_k z^{-2}}{1 + c_k z^{-1} + d_k z^{-2}} \qquad (2)$$

avoids these difficulties, and has the additional advantage of yielding the realization shown in Fig. 1, which is known to be practical for high-order filters. This form also has the advantage of making the zeros easy to find, a feature not shared by a third possibility, the parallel form. For

Manuscript received September 10, 1969.

This work was supported by the U.S. Army Research Office-Durham, under Contract DA HC04 69 C 0012.

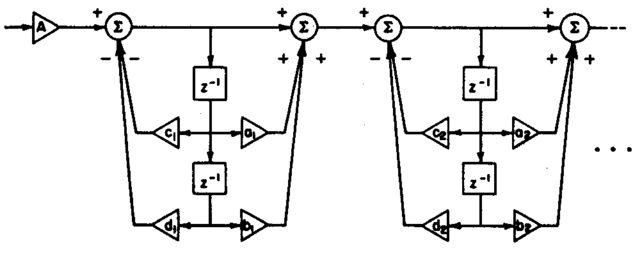


Fig. 1. Cascade realization corresponding to the canonic form.

these reaons the cascade canonic form will be assumed in what follows.

III. Statement of the Problem

Suppose now that the desired magnitude characteristic is prescribed at the discrete set of frequencies $W_1, \cdots, W_M$ where W_i is given in fractions of the Nyquist rate. These correspond to values of the variable z

$$z_i = e^{jW_i\pi} \qquad i = 1, \cdots, M. \tag{3}$$

Call the desired magnitude at these frequencies Y_i^d. Then the square-error in the frequency domain is

$$Q(\theta) = \sum_{i=1}^{M} (\mid Y(z_i) \mid - Y_i^d)^2 \tag{4}$$

where θ is the $(4K+1)$-vector of unknown coefficients

$$\theta = (a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2, \cdots, A)'. \tag{5}$$

The problem is to find a value of θ, say θ^*, such that for all θ

$$Q(\theta^*) \leq Q(\theta). \tag{6}$$

This square-error is a nonlinear function of the parameter vector θ, and an iterative method must be used to accomplish its minimization. Such numerical methods as are available seek a relative (local) minimum from a given starting point, and cannot in general be relied upon to find the global solution. Computational experience, gained by using different starting points for the same problem, often gives some indication of the likelihood that a given local solution is in fact global. In addition, a suboptimal value of θ can often provide a useful design.

IV. Elimination of A and Calculation of the Gradient

The method of Fletcher and Powell appears to be the most efficient and powerful nonlinear optimization method now available [3], [4]. It need not be described here, except to say that it performs a one-dimensional minimization at each cycle, along a direction determined by the gradient and an updated estimate of the Hessian. The double precision FORTRAN IV program DFMFP, supplied by IBM in the scientific subroutine package [5], was used without change. The Fletcher–Powell method requires the computation of the gradient of Q with respect to the parameter vector. This computation was performed using double precision complex arithmetic in FORTRAN IV.

The error function Q can be minimized analytically with respect to A for fixed a_i, b_i, c_i, d_i; and A need not be considered an unknown parameter. To eliminate A from Q, define the $4K$-dimensional parameter vector

$$\phi = (a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2, \cdots, d_K)' \tag{7}$$

and write

$$\begin{aligned} Y(z, A, \phi) &= A \prod_{k=1}^{K} \frac{1 + a_k z^{-1} + b_k z^{-2}}{1 + c_k z^{-1} + d_k z^{-2}} \\ &=: AH(z, \phi). \end{aligned} \tag{8}$$

Then

$$Q(A, \phi) = \sum_{i=1}^{M} (\mid AH(z_i, \phi) \mid - Y_i^d)^2. \tag{9}$$

Differentiating with respect to $|A|$ and setting the result to zero yields the following optimum value of $|A|$, say $|A^*|$:

$$\mid A^* \mid = \frac{\sum_{i=1}^{M} \mid H(z_i, \phi) \mid Y_i^d}{\sum_{i=1}^{M} \mid H(z_i, \phi) \mid^2} \cdot \tag{10}$$

The Fletcher–Powell method is then used to minimize the new error criterion

$$\hat{Q}(\phi) = Q(A^*, \phi). \tag{11}$$

Notice that the sign of A^* is immaterial, since it does not affect the magnitude characteristic. It will be taken positive. The gradient of $\hat{Q}$ with respect to ϕ can be computed as follows:

$$\frac{\partial \hat{Q}}{\partial \phi_n} = \frac{\partial Q(A^*, \phi)}{\partial \phi_n} + \frac{\partial Q(A^*, \phi)}{\partial A^*} \frac{\partial A^*}{\partial \phi_n} \qquad n = 1, \cdots, 4K. \tag{12}$$

The second term is zero, since A^* is chosen to minimize Q. Hence by (9),

$$\frac{\partial \hat{Q}}{\partial \phi_n} = 2A^* \sum_{i=1}^{M} (A^* \mid H(z_i, \phi) \mid - Y_i^d) \frac{\partial \mid H(z_i, \phi) \mid}{\partial \phi_n} \cdot \tag{13}$$

Writing

$$\mid H(z_i, \phi) \mid = [H(z_i, \phi) \quad \overline{H(z_i, \phi)}]^{1/2} \tag{14}$$

we have

$$\frac{\partial\,|H(z_i,\phi)|}{\partial\phi_n} = \frac{1}{|H(z_i,\phi)|} \cdot \mathrm{Re}\left\{\overline{H(z_i,\phi)}\;\frac{\partial H(z_i,\phi)}{\partial\phi_n}\right\} \tag{15}$$

which can be computed directly from (8) using complex arithmetic.

The subroutine which calculates $\hat{Q}(\phi)$ and grad $\hat{Q}(\phi)$, given ϕ, is summarized below.

1) Calculate

$$H_i = \prod_{k=1}^{K} \frac{1 + a_k z_i^{-1} + b_k z_i^{-2}}{1 + c_k z_i^{-1} + d_k z_i^{-2}}, \quad i = 1, \cdots, M. \tag{16}$$

2) Calculate

$$A^* = \frac{\sum_{i=1}^{M} |H_i|\, Y_i^d}{\sum_{i=1}^{M} |H_i|^2}. \tag{17}$$

3) Calculate

$$E_i = A^* |H_i| - Y_i^d, \quad i = 1, \cdots, M. \tag{18}$$

4) Calculate

$$Q = \sum_{i=1}^{M} E_i^2. \tag{19}$$

5) Calculate

$$\begin{aligned}\frac{\partial |H_i|}{\partial a_k} &= \frac{1}{|H_i|}\,\mathrm{Re}\left\{\overline{H}_i\,\frac{\partial H_i}{\partial a_k}\right\}\\ &= \frac{1}{|H_i|}\,\mathrm{Re}\left\{\overline{H}_i\,H_i\,\frac{z_i^{-1}}{1 + a_k z_i^{-1} + b_k z_i^{-2}}\right\}\\ &= |H_i|\;\mathrm{Re}\left\{\frac{z_i^{-1}}{1 + a_k z_i^{-1} + b_k z_i^{-2}}\right\}\\ &\qquad k = 1, \cdots, K;\quad i = 1, \cdots, M\end{aligned} \tag{20}$$

and similarly,

$$\frac{\partial |H_i|}{\partial b_k} = |H_i|\;\mathrm{Re}\left\{\frac{z_i^{-2}}{1 + a_k z_i^{-1} + b_k z_i^{-2}}\right\} \tag{21}$$

$$\frac{\partial |H_i|}{\partial c_k} = -\;|H_i|\;\mathrm{Re}\left\{\frac{z_i^{-1}}{1 + c_k z_i^{-1} + d_k z_i^{-2}}\right\} \tag{22}$$

$$\frac{\partial |H_i|}{\partial d_k} = -\;|H_i|\;\mathrm{Re}\left\{\frac{z_i^{-2}}{1 + c_k z_i^{-1} + d_k z_i^{-2}}\right\}. \tag{23}$$

6) Calculate

$$\frac{\partial \hat{Q}}{\partial \phi_n} = 2A^* \sum_{i=1}^{M} E_i\,\frac{\partial |H_i|}{\partial \phi_n}, \qquad n = 1, \cdots, 4K. \tag{24}$$

The elimination of A as an unknown parameter reduces by one the dimensionality of the search performed by the optimization program. An additional savings in computation time is achieved by computing the z_i once at the beginning of execution and storing them for later use.

V. Stability and Minimum-Phase Constraints

Suppose $Y(z)$ has a real pole at $z=\alpha$. Replacing this by a pole at $z=1/\alpha$ is equivalent to multiplying by the function

$$\frac{z-\alpha}{z-1/\alpha} \tag{25}$$

which has magnitude $|\alpha|$ when z is on the unit circle. Hence the inversion of a real pole with respect to the unit circle does not affect the shape of the magnitude characteristic. Since the gain constant A is chosen optimally, $\hat{Q}(\phi)$ is not affected at all by such inversion. Similarly, $\hat{Q}(\phi)$ is unaffected by inversions of complex pairs of poles, or real or complex pairs of zeros. At convergence of the optimization program, poles and zeros will appear randomly inside or outside the unit circle, depending on the starting point of the iteration, and upon the course of the iteration itself. It will be taken as a design criterion that all the poles and zeros lie within the unit circle. The poles must do so in order that the filter be stable. The zeros lying inside the unit circle ensure that there is no excess phase.

VI. Final Strategy and Example 1: A Low-Pass Filter

At first thought, it would appear that the following procedure would yield an optimum transfer function with all its poles and zeros inside the unit circle.

1) Use of the optimization program to minimize $\hat{Q}(\phi)$ without constraining pole or zero locations.
2) At convergence, invert all poles or zeros outside the unit circle.

If the optimization program is started anew from the result of step 2, however, it is found that further reduction in $\hat{Q}(\phi)$ is sometimes possible. The following example will show how this can happen. Consider the specification of an ideal low-pass filter with cutoff frequency at one-tenth the Nyquist frequency:

$$\begin{array}{ll} W = 0.00,\ 0.09\ (0.01); & Y^d = 1.0\\ W = 0.10; & Y^d = 0.5\\ W = 0.11,\ 0.20\ (0.01); & Y^d = 0.0\\ W = 0.2,\quad 1.0\quad (0.1); & Y^d = 0.0.\end{array} \tag{26}$$

This specification weights frequencies below $W=0.2$ more heavily than those above. If the optimization for a one-section filter ($K=1$) is started at

$$\phi = (0.,\ 0.,\ 0.,\ -0.25)' \tag{27}$$

convergence is obtained after 93 function evaluations. The resulting zeros and poles are given below and are plotted in Fig. 2(A):

$$\begin{aligned} \text{zeros: } & 0.67834430 \pm j\ 0.73474418 \\ \text{poles: } & 0.75677793, \qquad 1.3213916. \end{aligned} \tag{28}$$

The corresponding value of the error criterion is

$$\hat{Q} = 1.2611. \tag{29}$$

Notice that the two poles are very nearly inverses of each other. After inversion, 62 more function evaluations are required to produce convergence[1] to the following parameters (see Fig. 2(B)):

$$\begin{aligned} \text{zeros: } & 0.82191163 \ \pm j\ 0.56961501 \\ \text{poles: } & 0.89676390 \ \pm j\ 0.19181084 \\ A &= 0.11733978 \\ \hat{Q} &= 0.56731. \end{aligned} \tag{30}$$

The introduction of a complex pole pair is prevented in the first sequence of iterations by the fact that one pole is inside the unit circle and one is outside. After inversion, the two poles can split and become a complex pair, leading to the final minimum. The magnitude characteristic of the final filter is shown in Fig. 3.

The following algorithm was used to allow such convergence to take place.

1) Use the Fletcher–Powell optimization program until convergence takes place, or for a maximum of 25 cycles, and go to 2.
2) If any poles or zeros are outside the unit circle, invert them and go to 1. Otherwise, go to 3.
3) If convergence has not taken place, go to 1. Otherwise, go to 4.
4) Print out the final parameters and stop.

Fig. 3 also shows the resultant filter characteristic for $K=2$, corresponding to the following parameters at convergence:

$$\begin{aligned} \text{zeros: } & 0.92538461 \ \pm j\ 0.37902945 \\ & 0.61137175 \ \pm j\ 0.79134343 \\ \text{poles: } & 0.93121838 \ \pm j\ 0.27718988 \\ & 0.86454727 \ \pm j\ 0.13353860 \\ A &= 0.024867372 \\ Q &= 0.033959. \end{aligned} \tag{31}$$

The final parameters obtained from the $K=1$ design were used as a starting point for the $K=2$ optimization, and 376 further function evaluations were required for convergence.

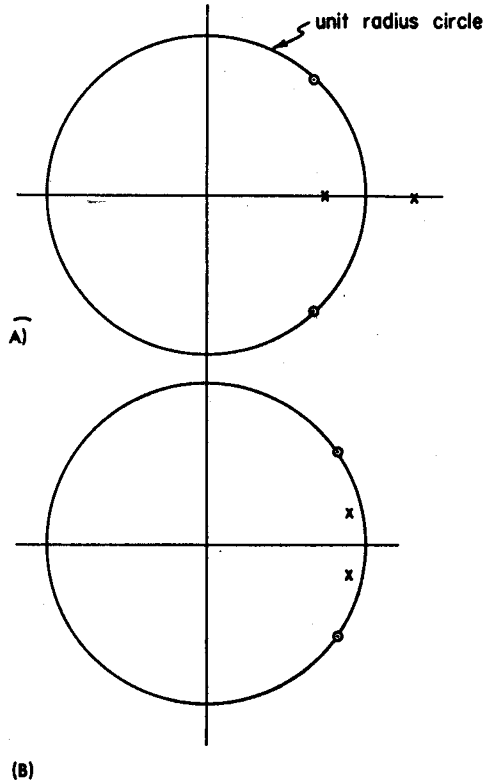
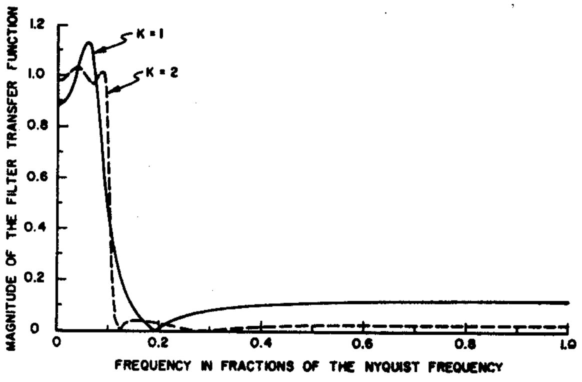


Fig. 2. Pole-zero configurations for the low-pass filter designs. (A) Intermediate local minimum. (B) Final minimum.

Fig. 3. Magnitude characteristic of the one- and two-section low-pass filters of Example 1.



VII. Examples

Example 2: A Wide-Band Differentiator

Consider the following specification,

$$W = 0.0,\ 1.0\ (0.05); \qquad Y^d = W, \tag{32}$$

which represents a linear amplitude characteristic, and hence an ideal differentiating filter, ignoring consideration of the phase for the moment. The one-section design

[1] The convergence criterion for this and all succeeding examples corresponds to the parameter EPS$=10^{-5}$ in DFMFP.

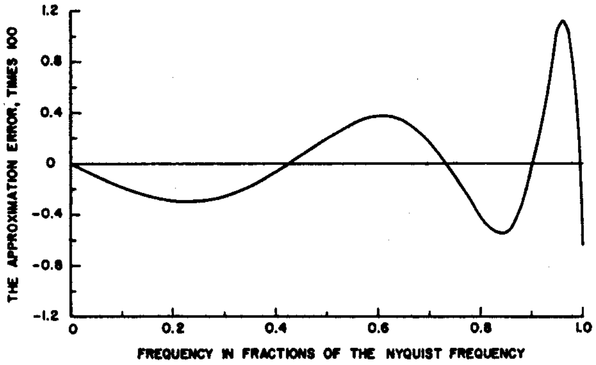
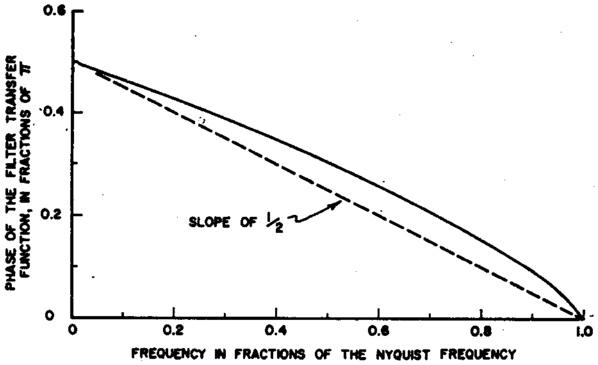


Fig. 4. The approximation error for the one-section wide-band differentiator filter.

Fig. 5. Phase characteristic of the one-section wide-band differentiator filter.



converged after 96 function evaluations to the following design:

$$\begin{aligned} \text{zeros: } & 1.0000000, \quad -0.67082621 \\ \text{poles: } & -0.14240300, \quad -0.71698670 \\ & A = 0.36637364 \\ & \hat{Q} = 2.7480 \times 10^{-4}. \end{aligned} \tag{33}$$

Fig. 4 shows the approximation error over the entire band of frequencies from zero to the Nyquist frequency. Of particular interest is the fact that the approximation is within about 1 percent of maximum over this entire range, in contrast with designs based on guard-band filters, which usually are accurate only up to about 80 percent of the Nyquist frequency (see [1]). Fig. 5 shows the phase characteristic, which approximates the phase of an ideal differentiator with an additional lag of one-half sampling period. Thus, this design introduces significantly less lag than designs reported in [1].

Starting with the one-section design above, 500 more function evaluations produced convergence to a more accurate two-section approximation, with $\hat{Q}=6.15\times 10^{-7}$. This two-section filter exhibited a similar characteristic ripple near the Nyquist frequency, and had almost the same phase characteristic. This points out the desirability of extending the method to include specifications on the phase characteristic.

Example 3: A Linear Discriminator

For the next example, consider the specification

$$W = 0.0,\ 1.0\ (0.05); \qquad Y^d = |1 - 2W| \tag{34}$$

which represents a linear discriminator with a zero at one-half the Nyquist frequency. After 40 function evaluations, the following one-section design was produced:

$$\begin{aligned} \text{zeros: } & 0.00000000 \pm j\, 1.00000000 \\ \text{poles: } & \pm 0.49614741 \\ & A = 0.35765018 \\ & \hat{Q} = 1.2299 \times 10^{-2}. \end{aligned} \tag{35}$$

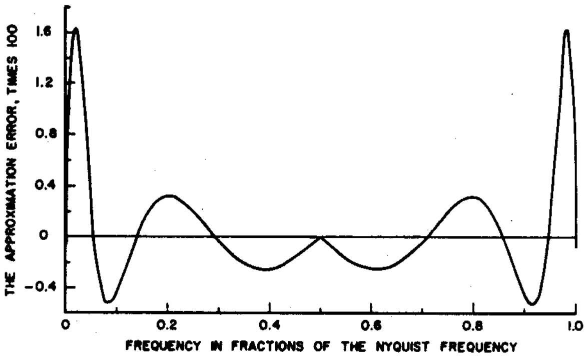


Fig. 6. The approximation error for the two-section linear discriminator filter.

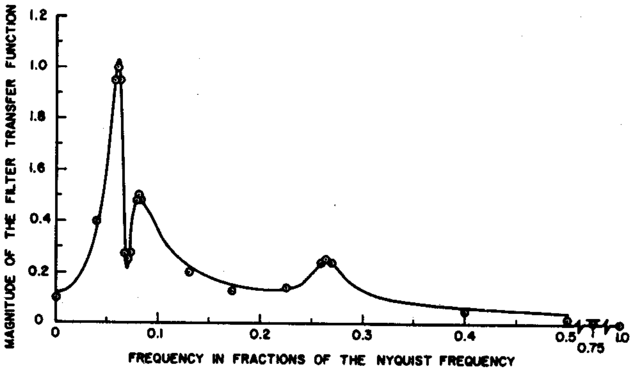
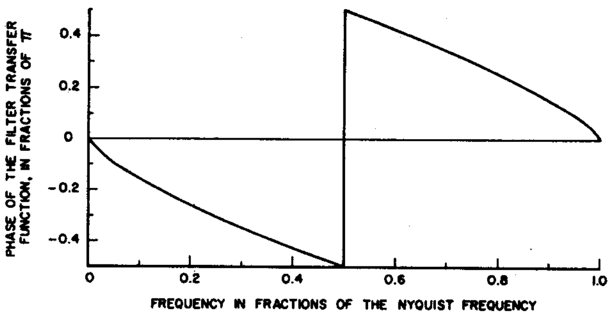


Fig. 8. Magnitude characteristic of the three-section vowel formant filter Circles indicate specification points.

Fig. 7. Phase characteristic of the two-section linear discriminator filter.



One-hundred and thirty more function evaluations produced convergence to the following two-section design:

$$
\begin{aligned}
\text{zeros: } & 0.00000004 \ \pm j\ 0.99999931 \\
& 0.81492900, \ -\ 0.81492888 \\
\text{poles: } & 0.84492845, \ -\ 0.84492830 \\
& 0.37204922, \ -\ 0.37204934 \\
& A = 0.36676649 \\
& \hat{Q} = 1.0807 \times 10^{-4}.
\end{aligned}
\tag{36}
$$

As might be expected, the resulting pole-zero patterns are symmetric with respect to the imaginary axis, within the precision allowed by the convergence criterion. Figs. 6 and 7 show the approximation error and phase characteristic of the two-section filter.

Example 4: A Vowel Formant Filter

Fig. 8 shows the specification of a filter which is to have a magnitude characteristic corresponding to the formant for the vowel ⊃ (as in "law") [6]. The principal requirements are taken to be that peaks occur at $W=0.06$, 0.08, and 0.26; with values of 1.0, 0.5, and 0.25, respectively; and that troughs occur midway between these peaks, with values one-half the lower peak. This design problem is considerably more difficult than the previous ones, since it involves approximating a rather arbitrary and complex shape. An acceptably good design required three sections, 1809 function evaluations and 2 minutes 34 seconds of computation time on the IBM 360/65. The final design is also shown in Fig. 8 and corresponds to the following parameters:

$$
\begin{aligned}
\text{zeros: } & 0.93470084, \ -\ 0.99966051 \\
& 0.62177384 \ \pm j\ 0.62464737 \\
& 0.97101465 \ \pm j\ 0.21383044 \\
\text{poles: } & 0.64343825 \ \pm j\ 0.70167849 \\
& 0.96361229 \ \pm j\ 0.19280318 \\
& 0.93515982 \ \pm j\ 0.20909432 \\
& A = 0.041075206 \\
& \hat{Q} = 9.4712 \times 10^{-3}.
\end{aligned}
\tag{37}
$$

VIII. Conclusions

A practical method has been described for designing recursive digital filters with arbitrary, prescribed magnitude characteristics. Examples have been given of such designs with 1, 2, and 3 cascade sections, corresponding to 5, 9, and 13 parameters. The most difficult of these designs takes about 2.5 minutes of computation time on the IBM 360/65 computer.

The important considerations in the development of this method have been: 1) a strategy for ensuring that the resulting filters are stable and minimum phase, 2) elimination of the gain factor A as an unknown parameter, 3) choice of the canonic form of the filter, and 4) choice of the unconstrained optimization algorithm.

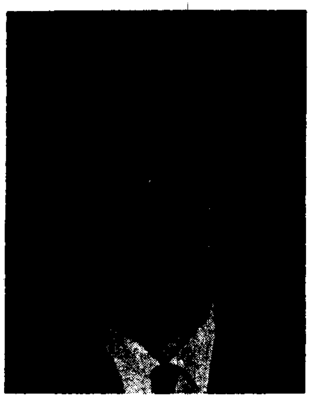
Further work along these lines might take into account specifications on the phase characteristic, and arbitrary weighting of the errors at different specification points.

Acknowledgment

The author is indebted to the following for valuable discussions about this problem: Dr. G. C. Winham, Department of Music, Princeton University; C. M. Rader, Lincoln Laboratories; and Dr. J. F. Kaiser, Bell Telephone Laboratories.

References

[1] J. F. Kaiser, "Digital filters," in *System Analysis by Digital Filter*, F. F. Kuo and J. F. Kaiser, Eds. New York: Wiley, 1966, ch. 7.
[2] C. M. Rader and B. Gold, "Digital filter design techniques in the frequency domain," *Proc. IEEE*, vol. 55, pp. 149–171, February 1967.
[3] R. Fletcher and M. J. D. Powell, "A rapidly convergent descent method for minimization," *Computer J.*, vol. 6, no. 2, pp. 163–168, 1963.
[4] G. C. Temes and D. A. Calahan, "Computer-aided network design—the state of the art," *Proc. IEEE*, vol. 55, pp. 1832–1863, November 1967.
[5] "System/360 scientific subroutine package (360A-CM-03X), version III, programmer's manual," IBM Data Processing Division, White Plains, N. Y., Document H20-0205-3, 1968.
[6] R. K. Potter and J. C. Steinberg, "Toward the specification of speech," *J. Acoust. Soc. Am.*, vol. 22, pp. 807–820, November 1950.



Kenneth Steiglitz (S'57–M'64) was born in Weehawken, N. J., on January 30, 1939. He received the B.E.E., M.E.E., and Eng. Sc.D. degrees from New York University, New York, N. Y., in 1959, 1960, and 1963, respectively.

Since 1963 he has been with the Department of Electrical Engineering, Princeton University, Princeton, N. J., where he is now an Associate Professor.

Dr. Steiglitz is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

Reprinted from IEEE TRANSACTIONS
ON *AUDIO AND ELECTROACOUSTICS*
Volume AU-18, Number 2, June, 1970
pp. 123-129