TWO METHODS FOR THE REDUCTION OF QUANTIZATION EFFECTS IN RECURSIVE DIGITAL FILTERS

Zaman Motamedi

A Thesis
in
The Faculty
of
Engineering and Computer Science

Presented in Partial Fulfillment of the Requirements
for the degree of Master of Engineering
Concordia University
Montreal, Quebec
Canada

July 1982

LIST OF TABLES

LIST OF FIGURES

LIST OF PRINCIPAL SYMBOLS

SYMBOL	DESCRIPTION
$x(nT)$	Discrete-time function used as filter input
$y(nT)$	Descrete-time function used as filter output
z	Used as an operator for z-transformation
$X(z), Y(z)$	The input and output functions in the z-domain
$H(z)$	Filter transfer function defined as $Y(z)/X(z)$
H_0	Filter multiplier constant
ω	Frequency in rad/s
$C(\phi)$	Constraint equation
$p(\phi,s,t)$	Shifted exterior penalty function
λ	Section scaling factor
PSD	Power spectral density function
NP	Noise power function
A	Denotes 2-section elliptic filters
B	Denotes 3-section elliptic filters
C	Denotes 3-section minimized-maximum-pole filters
$M(\alpha_i, r_i, \beta_i, \theta)$	Magnitude of the filter transfer function

LIST OF CONTENTS

ABSTRACT

Two Methods for the Reduction of Quantization Effects in Recursive Digital Filters

Zaman Motamedi

Two distinct methods for the reduction of coefficient and product quantization effects in recursive digital filters are described. A degree of freedom is introduced in the design by increasing the approximation order above the minimum. This is then used to increase the allowable margin for coefficient quantization error or to reduce the sensitivity to coefficient quantization. The two methods have been used to design a diverse range of lowpass filters, including some narrowband as well as high-selectivity filters. Experimental results reveal that the two methods lead to significant reductions in the required wordlength and in the inband noise power.

ACKNOWLEDGEMENTS

The author wishes to thank both Dr. C. Charalambous and Dr. A. Antoniou for their supervision, advice, and understanding and Marie Berryman for typing and preparation of the text of this thesis.

CHAPTER 1

PRINCIPAL CONCEPTS

1.0 Introduction

Emergence of digital technology began in the mid 1960's when high speed digital computers became widely available for serious research and development work. Many concepts that form the theoretical basis of digital signals and systems, such as the Z-transform and the Fourier analysis had been familiar to engineers for a long time. In the ensuing years the field of digital filters has matured considerably and its development is intimately tied with advances in the computer fiel[illegible] past decade has been marked with phenomenal progress in computer technology. With each stride forward, computers became more accessible and more affordable to an ever-increasing user community. The users discovered more new applications, generating new demands for even more sophisticated technology. These developments have had a profound impact on almost all scientific disciplines and the field of digital filters benefitted greatly from them.

In digital filters, we deal with signals and systems that are discrete-time counterparts of the more familiar continous-time systems. Digital filters can perform the same functions that analog filters do. The analog approach in some cases, may be difficult or unfeasible to implement practically. The use of digital filters offers important engineering advantages, such as perfect reproductibility and guaranteed level of performance, increased ease in changing the filter coefficients characterized by small physical size, and possibility of time sharing the same hardware system among a multiplicity of filtering functions. There is a further important advantage: the possibility of modularized hardware for customized large scale integration. These advantages alone would, in many cases, make digital filtering an attractive alternative to analog filters.

Digital systems, in general, are classified as linear or non-linear, causal or non-causal, time-dependent or time-invariant. Furthermore, one can divide the family of digital filters into two subfamilies; recursive and non-recursive. The z-transform, its properties and applications, as well as the stability criteria are essential ingredients upon which the complete theory of digital systems is based. Definitions, description of each class and subfamily as well as rigorous mathematical analysis of the above can be found in [1] of the reference section. This thesis is concerned with the design of linear, causal, time-invariant recursive digital filters.

The quantization of coefficients in digital filters introduces an error in the amplitude response whereas the quantization of products tends to introduce quantization noise. If prescribed filter specifications are to be achieved, the magnitude of coefficient quantization error must not exceed specified bounds. On the other hand, if the signal-to-noise ratio is to be maximized, the level of quantization noise must be as low as possible.

Reduction in coefficient quantization errors and quantization noise can be achieved in several ways, as follows:

1. By using low-sensitivity low-noise digital-filter structures [2]-[7].

2. By optimizing the amplitude response over a discrete-parameter space [7]-[12].

3. By choosing the filter approximation such that a specific sensitivity or noise measure is minimized [13]-[15].

In this thesis two alternative approaches for the reduction of quantization effects are examined. In both approaches the minimum-order elliptic approximation satisfying the desired specifications is deduced and the approximation order is increased in order to introduce a degree of freedom in the design. In the first approach the degree of freedom gained is used to minimize the passband ripple so as to maximize the allowable margin for coefficient quantization error. In the second approach the degree of freedom gained is used to maximize the minimum pole distance from the unit circle so as to reduce the sensitivity to coefficient quantization. This design is accomplished by using an appropriate optimization technique.

The two methods are used to design several 6th order lowpass filters. The designs obtained are then compared with the corresponding minimum-order elliptic designs with respect to the effects of coefficient and product quantization. The results show that both methods lead to a reduced wordlength for the implementation and also to reduced output quantization noise.

1.1 Basic Concepts

The filter structure assumed in this thesis is the cascade canonic realization of Fig. 1.1. Each building block in this structure is a second-order section of the form shown in Fig. 1.2. As can be seen each section is composed of three basic elements, namely

1. Unit-delays
2. Adders
3. Multipliers

The overall transfer function of each section as well as the partial transfer function from the outputs of the multipliers to the outputs of the sections can be determined by using the theorems of the z-transform. Using these theorems charactrization for the three basic elements shown in Fig. 1.3 can be obtained as:

Unit delay

$$Y(z) = z^{-1}X(z)$$

Adder

$$Y(z) = \sum_{n=1}^{K} X_i(z)$$

Multiplier

$$Y(z) = aX(z)$$

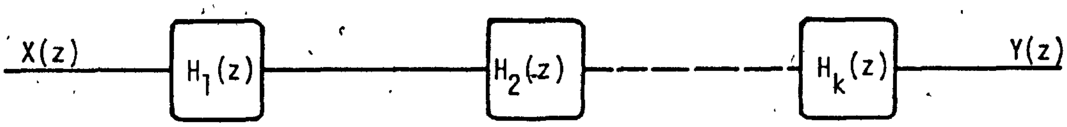


Figure 1.1 Cascade arrangement of filter sections

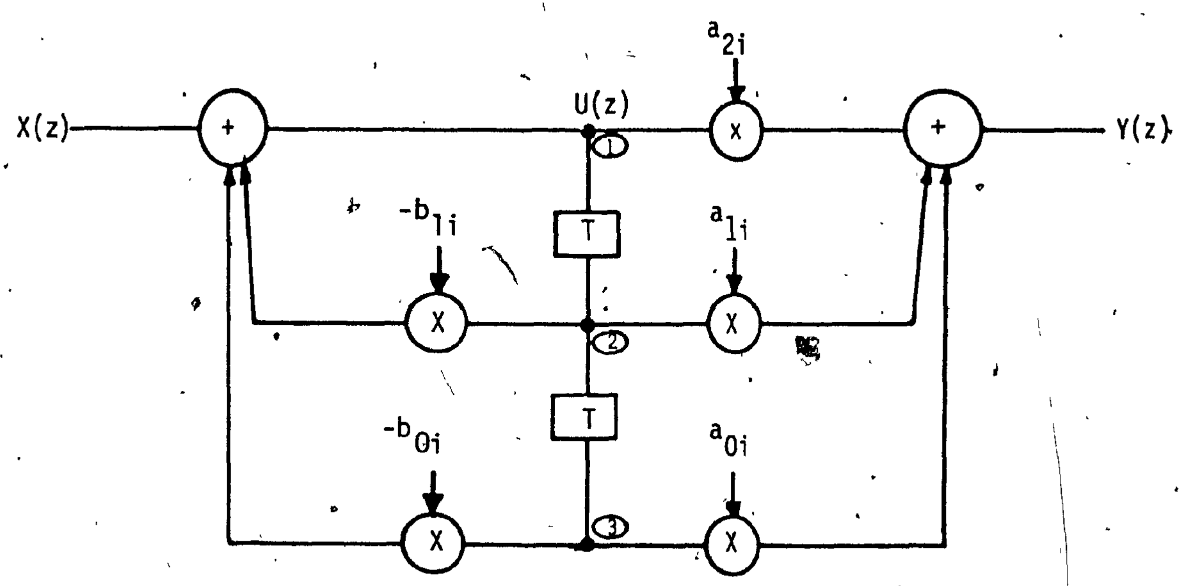


Figure 1.2 Canonic Realization of a second-order Filter section

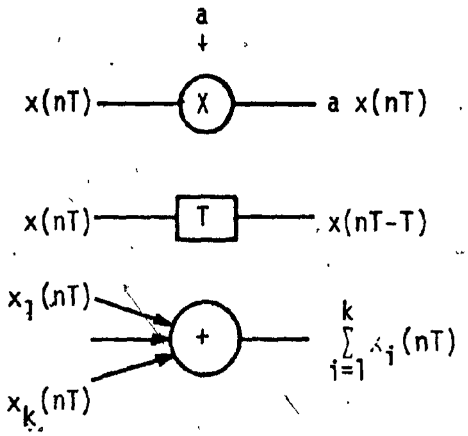


Figure 1.3 Digital filter elements

1.2 Transfer Function

From Fig. 1.2

$$U(z) = X(z) - b_{1i}z^{-1}U(z) - b_{0i}z^{-2}U(z) \qquad \ldots \; 1.1$$

and

$$Y(z) = z^{-1}U(z)a_{1i} + z^{-2}U(z)a_{0i} + a_{2i}U(z) \qquad \ldots \; 1.2$$

hence from Eqns. 1.1 and 1.2 we have

$$U(z) = \frac{X(z)}{1 + b_{1i}z^{-1} + b_{0i}z^{-2}} \qquad \ldots \; 1.3$$

Now from Eqns. 1.2 and 1.3

$$Y(z) = \frac{X(z)[a_{2i}+a_{1i}z^{-1}+a_{0i}z^{-2}]}{1 + b_{1i}z^{-1} + b_{0i}z^{-2}}$$

Therefore

$$H(z) = \frac{Y(z)}{X(z)} = \frac{a_{2i} + a_{1i}z^{-1} + a_{0i}z^{-2}}{1 + b_{1i}z^{-1} + b_{0i}z^{-2}}$$

or

$$H(z) = \frac{a_{2i}z^{2} + a_{1i}z + a_{0i}}{z^{2} + b_{1i}z + b_{0i}} \qquad \ldots \; 1.4$$

Similarly, the partial transfer function from input to node 1 in Fig. 1.2 is obtained as

$$H'(z) = \frac{U(z)}{X(z)} = \frac{1}{1+b_{1i}z^{-1}+b_{0i}z^{-2}} = \frac{z^2}{z^2+b_{1i}z+b_{0i}} \qquad \ldots\ 1.5$$

If K sections are connected in cascade as in Fig. 1.1, the overall transfer function

$$H(z) = H_o \prod_{i=1}^{K} \frac{a_{2i}z^2+a_{1i}z+a_{0i}}{z^2+b_{1i}z+b_{0i}} \qquad \ldots\ 1.6$$

is obtained where H_o is a multiplier constant.

1.3 Elliptic Approximation

Given a set of lowpass filter specifications

$$\underset{\sim}{S} = \{\omega_p, \omega_a, \omega_s, A_p, A_a\}$$

where

ω_p = passband frequency in rad/s

ω_a = stopband frequency in rad/s

ω_s = sampling frequency in rad/s

A_p = maximum passband ripple in dB

A_a = minimum stopband attenuation in dB a minimum-order elliptic approximation satisfying $\underset{\sim}{S}$ can be generated by using the approach in [1]. The zeros of elliptic filters are located on the unit circle while the poles are located within the unit circle of the z plane; hence for elliptic filters the transfer function of Eqn. 1.6 assumes the form

$$H(z) = H_0 \prod_{i=1}^{K} \frac{z^2+a_{1i}z+1}{z^2+b_{1i}z+b_{0i}} \qquad \ldots\ 1.7$$

that is $a_{0i} = a_{2i} = 1$.

In elliptic filters, the passband attenuation oscillates between zero and a prescribed maximum A_p, and the stopband attenuation oscillates between infinity and a prescribed minimum A_a, that is, the passband and stopband errors are equiripple. It can be shown that for a given set of specification $\underset{\sim}{S}$, the minimum-order elliptic approximation is the unique, lowest-order approximation that will satisfy $\underset{\sim}{S}$.

1.4 Amplitude Response

By writing the poles and zeros in terms of polar coordinates the transfer function of Eqn. 1.7 can be expressed as

$$H(z) = H_0 \prod_{i=1}^{K} \frac{(z-e^{j\alpha_i})(z-e^{-j\alpha_i})}{(z-r_ie^{j\beta_i})(z-r_ie^{-j\beta_i})} \qquad \ldots\ 1.8$$

where

$$a_{1i} = -(e^{j\alpha_i} + e^{-j\alpha_i}) = -2\cos\alpha_i$$

$$b_{1i} = -r_i(e^{j\beta_i} + e^{-j\beta_i}) = -2r_i \cos\beta_i$$

$$b_{0i} = r_i^2$$

If we let

$$z = e^{j\omega T} = e^{j2\pi\omega/\omega_s} = e^{j\theta} \qquad \ldots \quad 1:9$$

where $\theta = 2\pi\omega/\omega_s$.

We obtain

$$H(r_i,\alpha_i,\beta_i,\theta) = H_0 \prod_{i=1}^{K} \frac{(e^{j\theta}-e^{j\alpha_i})(e^{j\theta}-e^{-j\alpha_i})}{(e^{j\theta}-r_i e^{j\beta_i})(e^{j\theta}-r_i e^{-j\beta_i})}$$

$$= H_0 \prod_{i=1}^{K} \frac{(1-e^{j(\alpha_i-\theta)})(1-e^{-j(\alpha_i+\theta)})}{(1-r_i e^{j(\beta_i-\theta)})(1-r_i e^{-j(\beta_i+\theta)})}$$

$$= H_0 \prod_{i=1}^{K} \frac{N_i}{D_i} \qquad \ldots \quad 1.10$$

Where

$$D_i = [1-r_i e^{j(\beta_i-\theta)}][1-r_i e^{-j(\beta_i+\theta)}]$$

$$= 1-r_i[\cos(\theta+\beta_i)-j\sin(\theta+\beta_i)+\cos(\beta_i-\theta)+j\sin(\beta_i-\theta)]+r_i^2 e^{-2j\theta}$$

$$= [1-2r_i\cos\beta_i\cos\theta+2r_i^2\cos^2\theta-r_i^2]+j[2r_i\sin\theta][\cos\beta_i-r_i\cos\theta]$$

and

$$N_i = [1-e^{j(\alpha_i-\theta)}][1-e^{-j(\alpha_i+\theta)}]$$

$$= 1-e^{j(\alpha_i-\theta)}-e^{-j(\alpha_i+\theta)}+e^{-2j\theta}$$

$$= 1-[\cos(\theta+\alpha_i)-j\sin(\theta+\alpha_i)+\cos(\alpha_i-\theta)+j\sin(\alpha_i-\theta)]+e^{-2j\theta}$$

$$= (-2\cos\alpha_i\cos\theta+2\cos^2\theta)+j(2\sin\theta)(\cos\alpha_i-\cos\theta)$$

The amplitude response of the digital filter is given by

$$M(\alpha_i,r_i,\beta_i,\theta) = |H(\alpha_i,r_i,\beta_i,\theta)| \qquad \ldots \quad 1.11$$

From Eqns. 1.10 and 1.11

$$M(\alpha_i,r_i,\beta_i,\theta) = H_0 \prod_{i=1}^{K} \frac{|N_i|}{|D_i|}$$

where

$$|N_i| = [-8\cos\theta\cos\alpha_i+4\cos^2\theta+4\cos^2\alpha_i]^{1/2}$$

$$= 2\ |\cos\theta-\cos\alpha_i|$$

and

$$|D_i| = [(1-r_i^2)^2 -4\cos\theta\cos\beta_i\{r_i+r_i^3\} + 4r_i^2\{\cos^2\theta+\cos^2\beta_i\}]^{1/2}$$

Therefore

$$M(\underset{\sim}{x},\theta) = H_o \prod_{i=1}^{K} \frac{2|\cos\theta-\cos\alpha_i|}{[(1-r_i^2)^2 -4\cos\theta\cos\beta_i(r_i+r_i^3)+4r_i^2(\cos^2\theta+\cos^2\beta_i)]^{1/2}} \quad \ldots 1.12$$

where

$\underset{\sim}{x} = [\alpha_1,r_1,\beta_1,\ldots,\alpha_i,r_i,\beta_i,\ldots,H_o]^T$ is a 3K+1 dimensional column vector containing all the filter coefficients. In Eqn. 1.12

(i) zeros lie on the unit circle at angles of α_i

(ii) poles lie within the unit circle at angles of β_i

CHAPTER II

APPROXIMATION WITH OPTIMIZED POLE POSITION

2.0 Introduction

Coefficient quantization error is most objectionable in the passband where filter specifications are usually very tiaht. In order to render the effect of coefficient quantization insignificant, one may use the minimum-order elliptic approximation together with a sufficiently large wordlength in the implementation [1]. A second alternative is to increase the approximation order above the minimum and use the degree of freedom gained to minimize the passband ripple. In this way, the allowable margin for coefficient quantization error is increased and consequently a reduced wordlength can be employed in the implementation. A third possibility is to increase the approximation order as above and then use the degree of freedom gained to minimize the sensitivity to coefficient quantization. In this case the allowable margin for coefficient quantization error will remain the same as in minimum-order approximation. However, the magnitude of coefficient-quantization error will be reduced and presumably the required wordlength will also be reduced. Minimization of the passband ripple is a method which can easily be applied using well known techniques.

In this Chapter our efforts will be concentrated on the reduction of the coefficient sensitivity by minimization of the maximum pole radius of the filter. To do so, for a given set of specifications $\underset{\sim}{S}$, constraint equation for both the passband and the stopband must be derived.

Based on these equations, the problem of approximating the minimized-maximum-pole radius will be properly formulated. Finally, an optimization algorithm which can effectively solve the problem will be examined.

2.1 Error Function

The amplitude response $M(x,\theta)$ is constrained by the set of specification $\underset{\sim}{S}$ provided in Chapter I. Let the idealized passband response be unity and assume that

ε_p = Magnitude of passband tolerance

ε_a = Magnitude of stopband tolerance

With the set of specifications $\underset{\sim}{S}$ in mind, define

$$S_u = \text{Maximum allowable upper bound in the passband region} = 1+\varepsilon_p \quad \ldots\ 2.1$$

$$S_\ell = \text{Minimum allowable lower bound in the passband region} = 1-\varepsilon_p \quad \ldots\ 2.2$$

$$S_u' = \text{Maximum allowable upper bound in the stopband region} = \varepsilon_a \quad \ldots\ 2.3$$

Eqns. 2.1-2.3 provide the limits within which the filter specifications will be satified and beyond which the specifications will be violated. Thus if

(i) $e_u = M(\underset{\sim}{x},\theta)-S_u>0$ The upper limit in the passband is violated

(ii) $e_\ell = M(\underset{\sim}{x},\theta)-S_\ell<0$ The lower limit in the passband is violated

(iii) $e'_u = M(\underset{\sim}{x},\theta)-S'_u>0$ The stopband specification is violated

where e_u, e_ℓ, e'_u are the error functions in the passband and stopband regions. On the other hand if $\underset{\sim}{S}$ is to be satisfied then

$$\left.\begin{aligned} -e_u &= S_u-M(\underset{\sim}{x},\theta)\geq 0 \\ e_\ell &= M(\underset{\sim}{x},\theta)-S_\ell\geq 0 \end{aligned}\right\} \quad 0\leq\omega\leq\omega_p \qquad \ldots\ 2.4$$

$$-e'_u = S'_u-M(\underset{\sim}{x},\theta)\geq 0 \qquad \omega_a\leq\omega\leq\omega_{s/2}$$

Figs. 2.1(a) and (b) illustrate graphically the two conditions just discussed.

In the passband region, inequalities of 2.4 introduce two constraints (upper and lower limits) at every frequency point while in the stopband, for the same inequalities, one constraint must be satisfied at any given frequency point (upper limit).

If θ_i, for $i=1,2,\ldots,n_p$ are discrete frequency points located in the passband and θ_i, $i=n_p+1,\ldots, n_p+n_a$, are discrete frequency points located in the stopband as shown in Fig. 2.2, inequalities of 2.4 will be satisfied if and only if in the passband

$$S_\ell \leq M(\underset{\sim}{x},\ \theta_i) \leq S_u \qquad i=1,2,\ldots,n_p \qquad \ldots\ 2.5$$

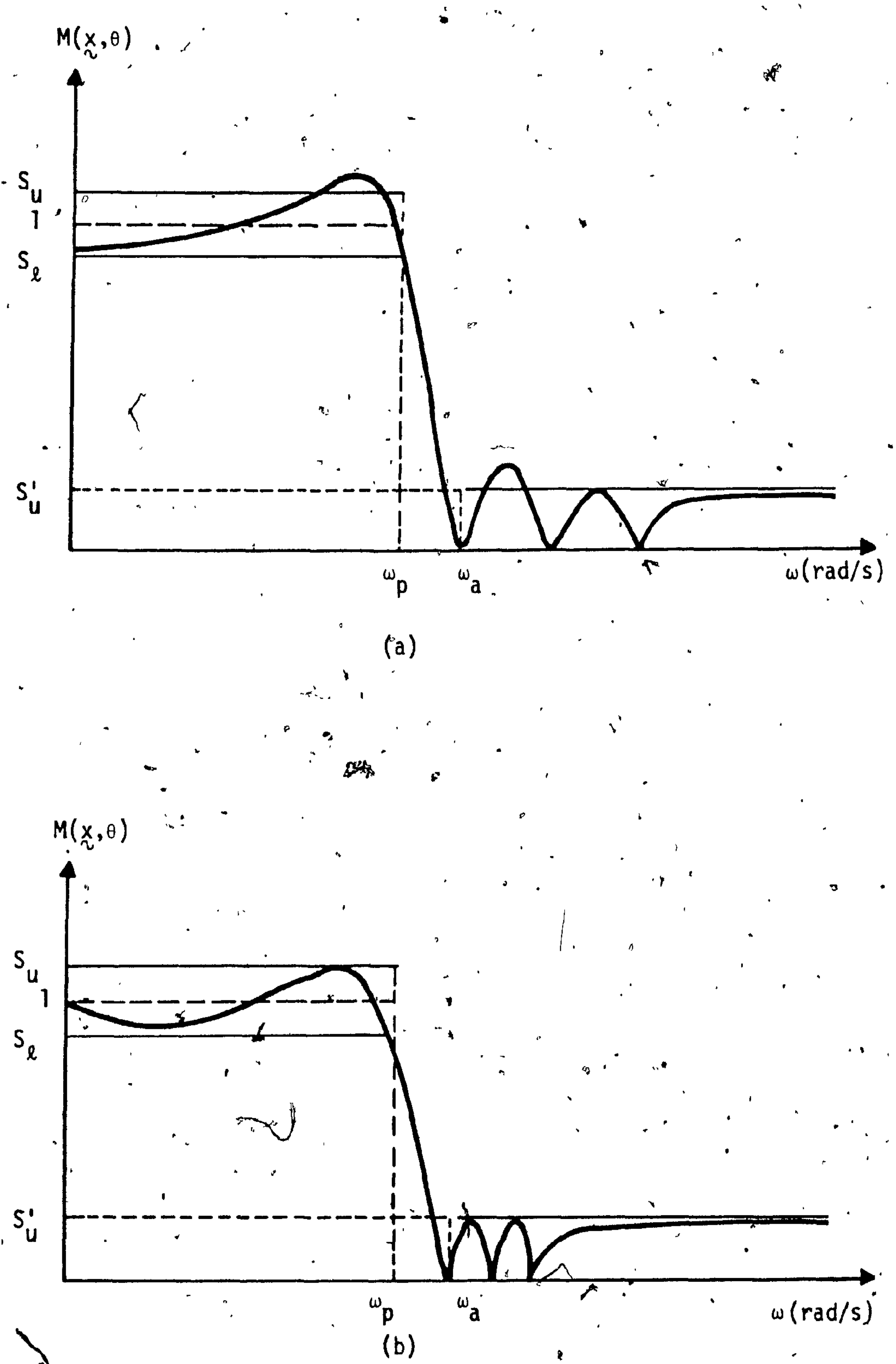


Figure 2.1(a) Amplitude response violating specification
(b) Amplitude response satisfying specifications

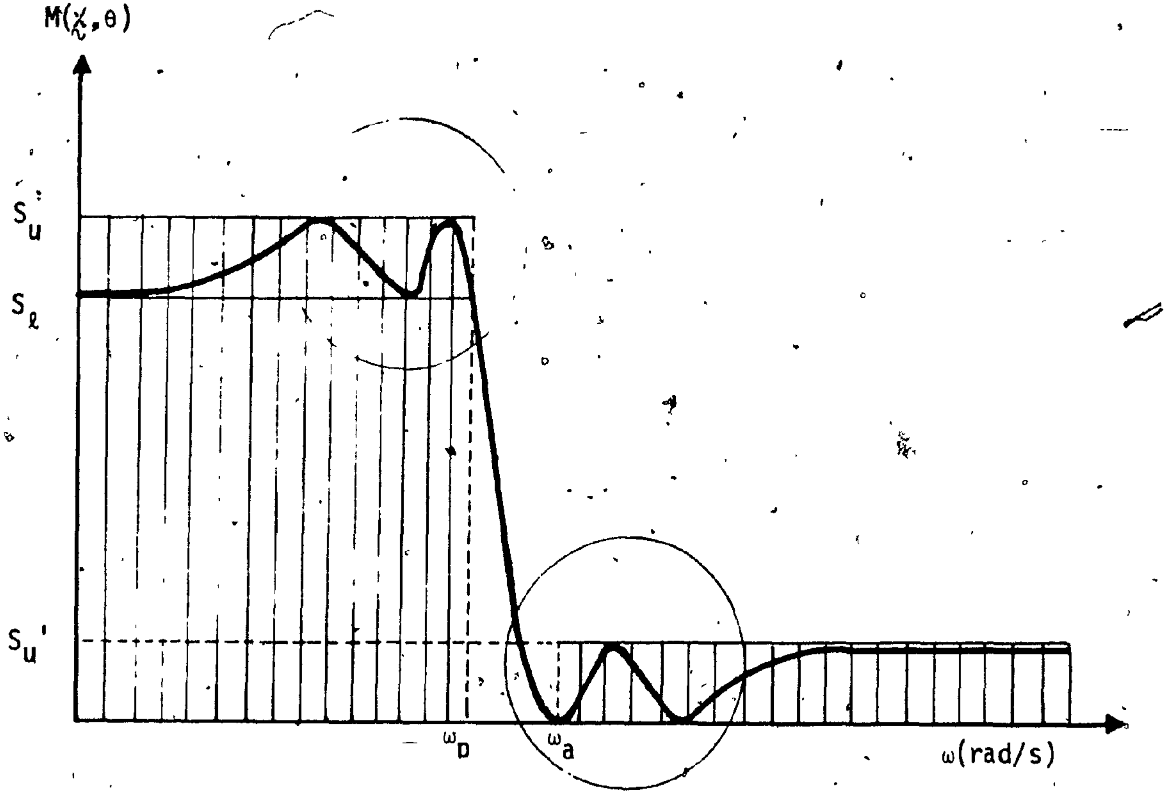


Figure 2.2 Discretization of frequency axis

In the stopband

$$0 \leq M(\underset{\sim}{x},\theta_i) \leq S'_u \qquad i=n_p+1,\ldots,n_p+n_a \qquad \ldots\ 2.6$$

Inequalities of 2.5 and 2.6 provide the following constraints:

In the passband

$$\left.\begin{array}{l} C_i(x) = M(\underset{\sim}{x},\theta_i)-S_\ell \geq 0 \\ C_{i+n_p}(\underset{\sim}{x}) = -M(\underset{\sim}{x},\theta_i)+S_u \geq 0 \end{array}\right\} \quad i=1,2,\ldots,n_p \qquad \ldots\ 2.7$$

and in the stopband

$$C_i(\underset{\sim}{x}) = -M(\underset{\sim}{x},\theta_{i-n_p})+S'_u \geq 0 \qquad i=1+2n_p,\ldots,2n_p+n_a \qquad \ldots\ 2.8$$

Hence, for any given set of specifications, if constraints 2.7 and 2.8 are satisfied, the filter specifications will be met and one may proceed towards evaluation of the filter cofficients.

The number of frequency points chosen in the passband and in the stopband regions depends on many factors such as the selectivity of the digital filter and the memory capacity of the computer being used. Naturally, the higher the number of frequency points the less the probability of violating the filter specifications at intervals between frequency points. However, the amount of computation will be increased. If the selectivity factor of the filter is high, the transition between passband and stopband regions will be sharper and coefficient quantization will yield a larger variation in the amplitude response of the digital filter near the passband and stopband edges. In such a case a larger number of constraints is required if the filter specifications are to be satisfied. There is no formal rule by which one may determine the 'adequate' number of constraints for realization of a given filter specification. However, experimental results have shown that the number of constraints should be 3 to 4 times the number of independent variables.

2.2 Description of the Problem

As stated previously, in order to render the effect of coefficient quantization insignificant one could

(i) Use a minimum-order elliptic approximation with sufficiently large wordlength in the implementation.

(ii) Introduce a degree of freedom by increasing the approximation order and use this degree of freedom to minimize the passband ripple.

(iii) Introduce a degree of freedom by increasing the approximation order and use this to minimize the coefficient sensitivity of the amplitude response .

In this chapter we explore the third possibility.

Variation in the location of the poles of the filter transfer function due to the use of finite wordlength registers affects the value of the amplitude response at any given frequency. Depending on the filter, small enough values of the wordlength could cause the poles to fall on or outside the unit circle of the z-plane, in which case filter instability will arise. In particular, sensitivity of the filter to coefficient variations is extremely high when the poles lie close to the unit circle (this is the case with highly selective filters). In order to minimize the coefficient sensitivity the degree of freedom introduced is used to maximize the minimum pole distance from the unit circle. This is done in such a way that constraints of 2.7-2.8 are satified at every discrete frequency point and, consequently, the filter specifications will be met.

2.3 Formulation of the Problem

Assuming that the feasible region defined by the constraints 2.7-2.8 is not empty, the objective of minimizing the maximum pole radius without violating the constraints reduces to the following optimization problem:

$$\underset{\underset{\sim}{x}}{\text{minimize}} \quad \{\max_{1\leq i\leq k} r_i\}$$

subject to the constraints

$$C_i(\underset{\sim}{x}) \geq 0 \qquad i=1,2,\ldots,n_p,n_{p+1},\ldots,2n_p+n_a$$

$$r_i \geq 0 \qquad i=1,2,\ldots,K$$

where K is the total number of filter sections and r_i is the radius of the ith pole.

By assuming that variable r_o is equal to or greater than the largest pole radius, the above optimization problem can be put in the following alternative but equivalent form:

$$\underset{\underset{\sim}{\phi}}{\text{minimize}}\ \{F(\underset{\sim}{\phi})\}$$

where $F(\underset{\sim}{\phi})=r_o$, and $\underset{\sim}{\phi}=[\underset{\sim}{x},r_o]^T$, subject to the constraints

$$C_i(\underset{\sim}{\phi})\geq 0 \qquad i=1,2,\ldots,2n_p+n_a \qquad 2.9$$

$$C_i(\underset{\sim}{\phi})=r_o-r_{i-2n_p-n_a}\geq 0 \quad i=2n_p+n_a+1,\ldots,2n_p+n_a+K$$

2.4 Minimization Algorithm

The constrained optimization problem of Eqn. 2.9 can be solved using a minimization algorithm proposed by C. Charalambous [16]. In this method the set of constraint functions is introduced into the objective function to form a new differentiable function called the penalty function. This new function, which is free of constraints, is then minimized by using any standard unconstrained optimization algorithm. According to Charalambous, if the problem of Eqn. 2.9 is to have local minima, the following theorems must hold true:

Theorem (1): (First-order necessary condition for optimality)

If $\underset{\sim}{\phi}^*$ is a local minimizer and the first-order constraint qualification holds, then there exist multipliers λ_i^*, i=1,2,...,m such that

$$\text{(i)}\qquad \nabla F(\underset{\sim}{\phi}^*) = \sum_{i=1}^{m} \lambda_i^* \nabla C_i(\underset{\sim}{\phi}^*)$$

$$\text{(ii)}\qquad \lambda_i^* C_i(\underset{\sim}{\phi}^*)=0 \qquad i=1,2,\ldots,m \qquad \ldots \quad 2.10$$

where $\lambda_i^* \geq 0$

constants λ_i^* are called Lagrangian multipliers and m is the number of inequality constraints.

If the gradients of the active constraints (i.e. the constraints which are equal to zero at ϕ^*) are linearly independent then the constraint qualifications hold.

Theorem (2): (Second-order sufficiency condition)

A sufficient condition for a feasible point ϕ^* to be a local minimizer is that there exist a vector λ^* such that

(i) Theorem (1) holds

(ii) $\underset{\sim}{d}^T \nabla_\phi^2 L(\underset{\sim}{\phi}^*, \underset{\sim}{\lambda}^*)\ \underset{\sim}{d} > 0 \quad \forall \underset{\sim}{d} \in D(\underset{\sim}{\phi}^*)$

where

$$L(\underset{\sim}{\phi}, \underset{\sim}{\lambda}) = \text{Lagrangian function} \equiv F(\underset{\sim}{\phi}) - \sum_{i=1}^{m} \lambda_i C_i(\underset{\sim}{\phi})$$

and

$$D(\underset{\sim}{\phi}^*) = \{\underset{\sim}{d} | \nabla C_i(\underset{\sim}{\phi}^*)^T \underset{\sim}{d} \geq 0 \ \ \forall i | C_i(\underset{\sim}{\phi}^*) = 0\}$$

The penalty function can be expressed as

$$p(\phi,s) = F(\phi) + \frac{1}{2} \sum_{i=1}^{m} s_i \{\max(0,-C_i(\underset{\sim}{\phi}))\}^2$$

$$s_i \geq 0 \quad i=1,2,\ldots,m$$

The function $p(\underset{\sim}{\phi},\underset{\sim}{s})$, called the exterior penalty function, is composed of two parts. The first part is the objective function to be minimized and the second part is the penalty part. Note that the penalty part is zero in the feasible region R set up by the constraints, while it has a positive contribution when the point $\underset{\sim}{\phi}$ is outside region R (see Figure 2.3). If the constrained minimium is the same as the unconstrained minimum, then $\underset{\sim}{\phi}^*$ is a minimum point of $p(\underset{\sim}{\phi},\underset{\sim}{s})$ for the value of $s_i \geq 0$; otherwise the optimum solution of $p(\underset{\sim}{\phi},\underset{\sim}{s})$ will be outside the feasible region (Figure 2.3). As $s_i \to \infty$, the tendency will be to draw the unconstrained minimum of $p(\underset{\sim}{\phi},\underset{\sim}{s})$ towards the boundary of the feasible region and under mild condition it can be proven that

$$\lim_{\underset{\sim}{s} \to \infty} \underset{\sim}{\phi}^*(\underset{\sim}{s}) = \underset{\sim}{\phi}^*,$$

The disadvantage of this method is that when s_i becomes large we introduce steep valleys in the unconstrained problem and even with the most efficient unconstrained algorithm difficulty will be experienced in minimizing $p(\underset{\sim}{\phi},\underset{\sim}{s})$ (ill-conditioning problem), i.e., although the exterior penalty function method is well supported by theory and requires

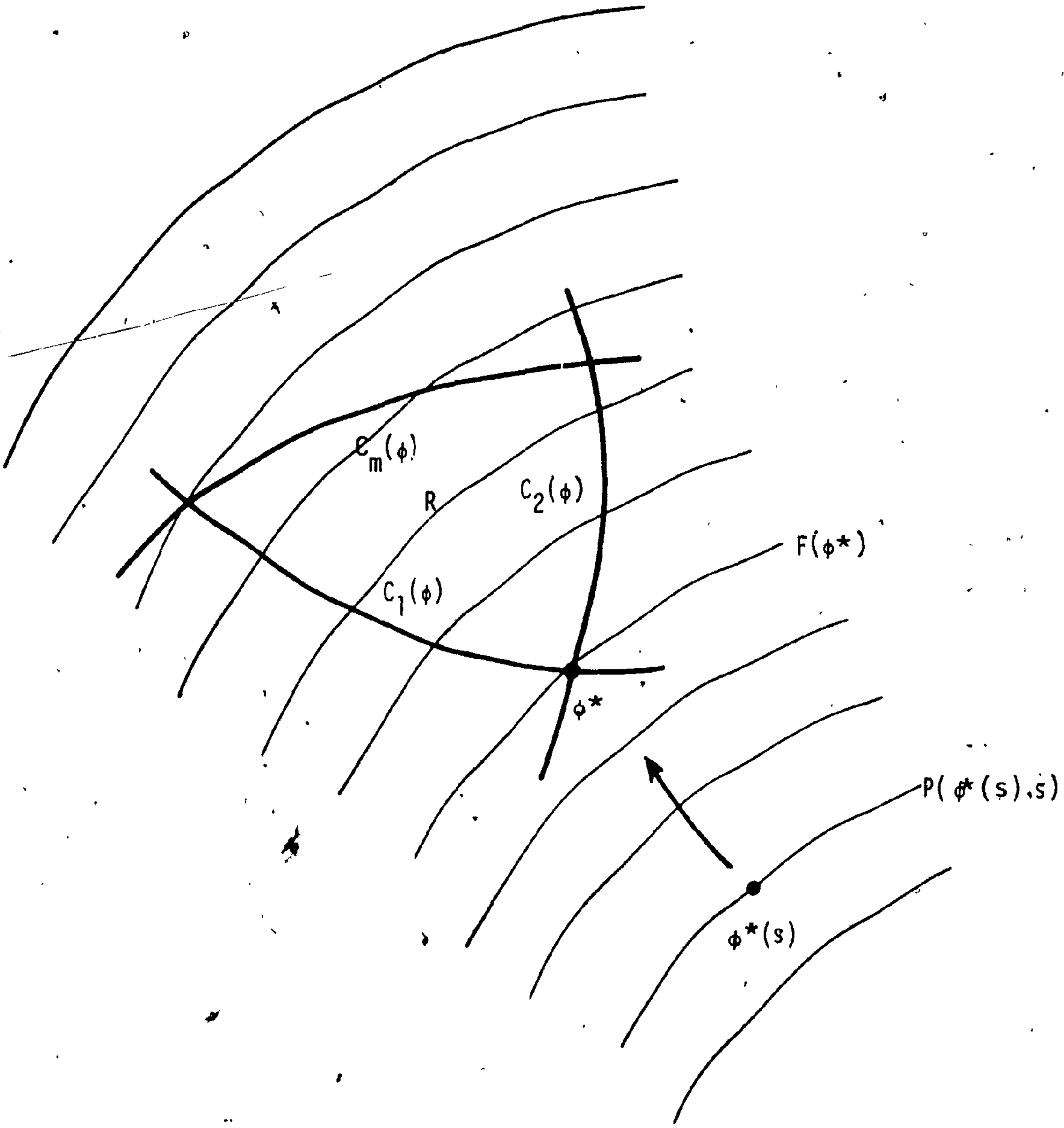


Figure 2.3 Relative location of the constrained and unconstrained minima $\phi^*,\phi^*(s)$ about the region defined by the constrains $C_1, C_2, \ldots, C_m$

mild assumptions to ensure theoretical convergence, it suffers from serious computational weaknesses which become more critical as the values of s_i tend to infinity. To avoid the ill-conditioning problem, modification to the exterior penalty function has been proposed so that the optimum point $\underset{\sim}{\phi}^*$ can be reached without having to force $\underset{\sim}{s}_i$ to very large values. An important property of the penalty function $p(\underset{\sim}{\phi},\underset{\sim}{s})$ is that at its optimum point $\phi^*(\underset{\sim}{s})$

$$F(\underset{\sim}{\phi}^*) \geq F[\underset{\sim}{\phi}^*(\underset{\sim}{s})] \qquad \ldots \quad 2.11$$

If $\nabla F(\underset{\sim}{\phi}^*)\neq 0$, we will get to $\underset{\sim}{\phi}^*$ only in the limit as $\underset{\sim}{s}_i\to\infty$. Suppose that each constraint $C_i(\underset{\sim}{\phi})$, $i=1,2,\ldots,m$ is perturbed inwards by a positive amount t_i as shown in Figure 2.4. The shifted constraints can be represented by the following set of equations:

$$\left.\begin{array}{l}\overline{C}_i(\underset{\sim}{\phi},t_i) = C_i(\underset{\sim}{\phi}) - t_i \\ \\ \qquad t_i>0\end{array}\right\} \qquad i=1,2,\ldots,m$$

With the perturbed constraints, the exterior penalty function $p(\underset{\sim}{\phi},\underset{\sim}{s})$ assumes the form

$$P(\underset{\sim}{\phi},\underset{\sim}{s},t) = F(\underset{\sim}{\phi}) + \frac{1}{2}\sum_{i=1}^{m} s_i[\max(0,-(C_i(\underset{\sim}{\phi})-t_i))]^2 \qquad \ldots \quad 2.12$$

Equation 2.12 is called the shifted exterior penalty function. Due to inequality 2.11 both the optimum solution of the shifted exterior penalty function $\underset{\sim}{\phi}^*(\underset{\sim}{s},t)$ and the optimum solution of the unconstrained function

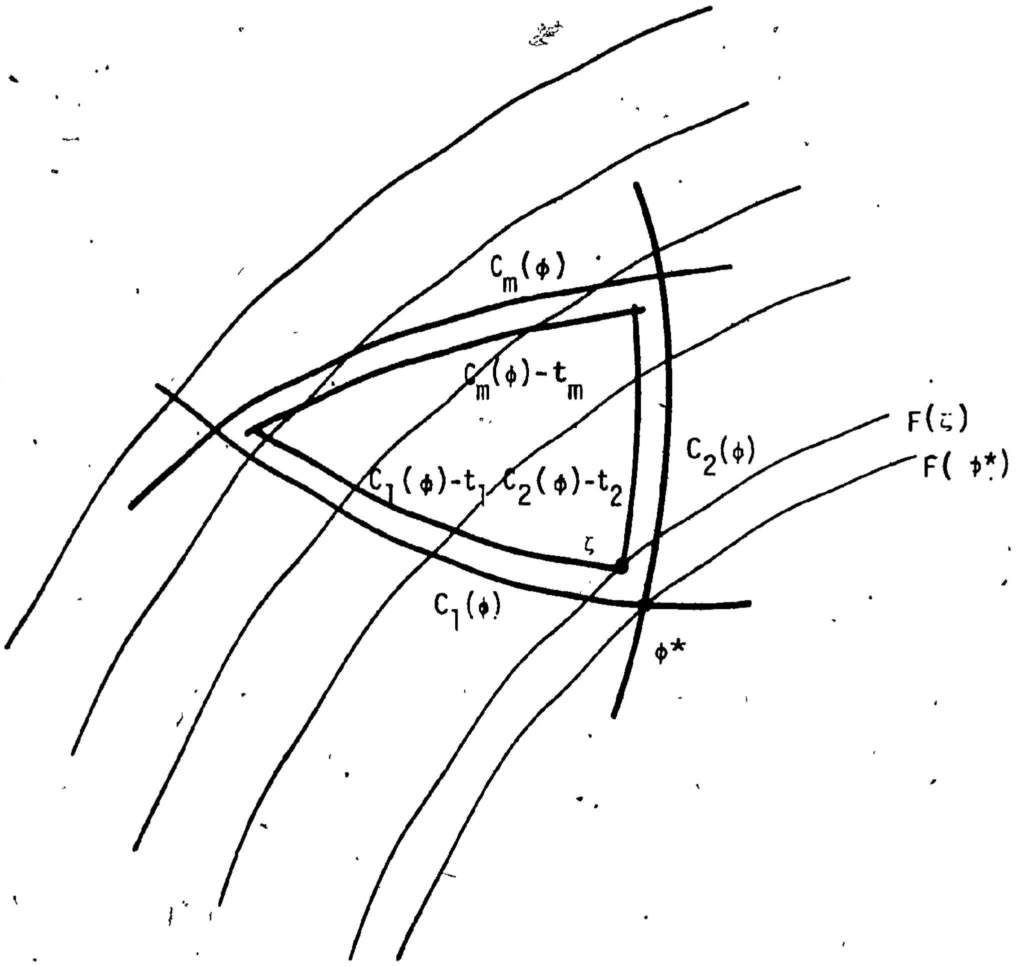


Figure 2.4 Perturbation of constraints

$F(\underset{\sim}{\phi})$, $\underset{\sim}{\phi}^*$, lie inside the region

$$R_t = \{\underset{\sim}{\phi} | F(\underset{\sim}{\phi}) \leq F(\underset{\sim}{\xi})\}$$

where $\underset{\sim}{\xi}$ is the solution of the problem:

$$\text{minimize } F(\underset{\sim}{\phi})$$
$$\text{subject to } C_i(\underset{\sim}{\phi}, t_i) \geq 0$$

According to Charalambous, the optimum solution of the shifted exterior penalty function $\underset{\sim}{\phi}^*(\underset{\sim}{s}, t)$ will be equal to $\underset{\sim}{\phi}^*$ for finite value of $\underset{\sim}{s}_i$ if

(i) the Kuhn-Tucker conditions for optimality of the non-linear programming problem are satisfied at $\underset{\sim}{\phi}^*$.

(ii) the parameters s_i and t_i are related as

$$\underset{\sim}{s}_i[\max\{0, -[C_i(\underset{\sim}{\phi}^*) - t_i]\}] = \underset{\sim}{\lambda}_i^* \qquad i=1,2,\ldots,m$$

$\underset{\sim}{\lambda}_i^*$ are the Lagrangian multipliers

and as long as $t_i \leq C_i(\phi^*)$ then the point $\underset{\sim}{\phi}^*$ will be a stationary point of $p(\underset{\sim}{\phi}, \underset{\sim}{s}, t)$. Defining the shifted parameters

$$t_i = \frac{\lambda_i}{s_i} \quad , \quad i=1,2,\ldots,m$$

the shifted exterior penalty function will become

$$p(\underset{\sim}{\phi},\underset{\sim}{s},\underset{\sim}{\lambda}) = F(\underset{\sim}{\phi}) + \frac{1}{2}\sum_{i=1}^{m} s_i[\max(0,-C_i(\underset{\sim}{\phi}) + \frac{\lambda_i}{s_i})]^2$$

and hence

$$\begin{aligned}\nabla p(\underset{\sim}{\phi}^*,\underset{\sim}{s}_i,\underset{\sim}{\lambda}^*) &= \nabla F(\underset{\sim}{\phi}^*) - \sum_{i=1}^{m} s_i[\max(0,-C_i(\underset{\sim}{\phi}^*) + \frac{\lambda_i^*}{s_i})]\nabla C_i(\underset{\sim}{\phi}^*)\\ &= \nabla F(\underset{\sim}{\phi}^*) - \sum_{i=1}^{m} \lambda_i^*\nabla C_i(\underset{\sim}{\phi}^*) = 0\end{aligned}$$

i.e. if $\underset{\sim}{\lambda}=\underset{\sim}{\lambda}^*$, $\underset{\sim}{\phi}^*$ is a stationary point of $p(\underset{\sim}{\phi},\underset{\sim}{s},\underset{\sim}{\lambda})$, and if $\underset{\sim}{\phi}^*$ satisfies the second-order sufficiency condition, then it can be proven that for s_i sufficiently large, $\underset{\sim}{\phi}^*$ is a strong local minimum of $p(\underset{\sim}{\phi},\underset{\sim}{s},\underset{\sim}{\lambda})$. The above argument is based on the following assumptions:

(i) The functions $F(\underset{\sim}{\phi})$ and $C_i(\underset{\sim}{\phi})$, $i=1,2,\ldots,m$ are all twice continuously differentiable.

(ii) Gradients $\nabla C_i(\underset{\sim}{\phi}^*)$ of active constraints are linearly independent.

(iii) The multipliers $\lambda_i^*>0$ when $C_i(\underset{\sim}{\phi}^*) = 0$.

(iv) The second-order sufficiency condition for a local constrained minimum of the problem holds at $\underset{\sim}{\phi}^*$.

The constrained optimization problem can be solved by using the optimization algorithm illustrated in Figure 2.5 [16]. In this algorithm a Quasi-Newton unconstrained optimization algorithm purposed by Fletcher has been used [17]. The Quasi-Newton method requires the gradient vector

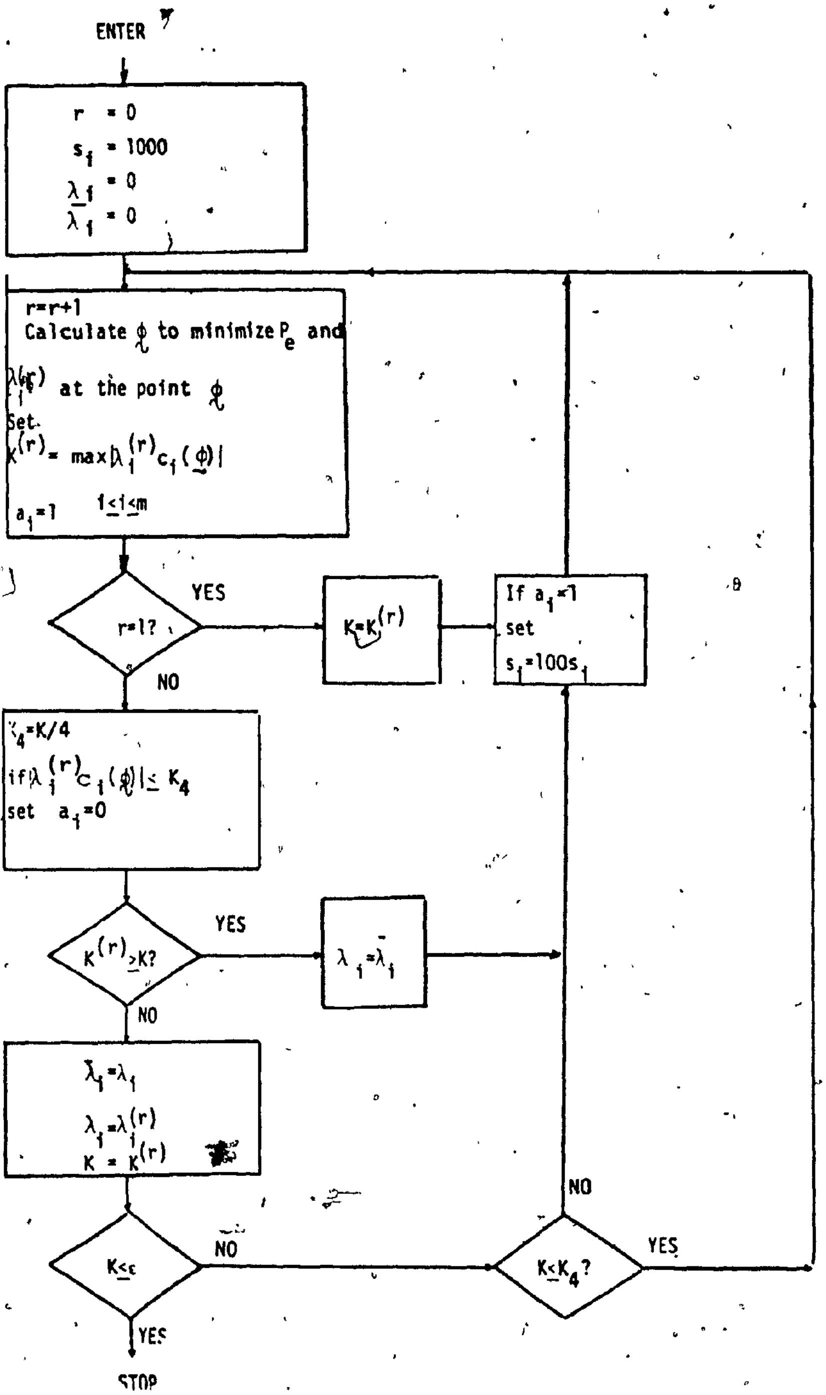


Figure 2.5 Flow-Chart of the optimization program

of the objective function in order to obtain the direction of descent which will eventually lead to the minimum point $\underset{\sim}{\phi}^*$. The basic steps in this algorithm are as follows. Let

$F(\underset{\sim}{\phi})$ = the unconstrained objective function to be minimized

H^0_{n*n} = any positive definite symmetric matrix

ε = convergence tolerance

$\underset{\sim}{\phi}^0$ = starting point

```
Step 1   set  $\underset{\sim}{\phi}=\underset{\sim}{\phi}^0$,  i=0
Step 2   if  $|\nabla F(\underset{\sim}{\phi}^i)|<\varepsilon$; stop
Step 3   set  $\underset{\sim}{d}^i = -[H^i]^T \nabla F(\underset{\sim}{\phi}^i)$
Step 4   perform exact line search to find the constant $\alpha_i^*$ which
         minimizes  $F(\underset{\sim}{\phi}^i+\alpha_i \underset{\sim}{d}^i)$; i.e.

             $\frac{\partial F}{\partial \alpha_i} = 0.$

Step 5   Set  $\underset{\sim}{\phi}^{i+1} = \underset{\sim}{\phi}^i + \underset{\sim}{\sigma}^i$  where  $\underset{\sim}{\sigma}^i = \alpha_i^* \underset{\sim}{d}^i$

Step 6   Set  $\underset{\sim}{\gamma}^i = \nabla F(\underset{\sim}{\phi}^{i+1}) - \nabla F(\underset{\sim}{\phi}^i)$
Step 7   Update  $H^i$, i.e.

             $H^{i+1} = H^i + \frac{\underset{\sim}{\sigma}^i \underset{\sim}{\sigma}^{i^T}}{\underset{\sim}{\sigma}^{i^T} \underset{\sim}{\gamma}^i} - \frac{H^i \underset{\sim}{\gamma}^i \underset{\sim}{\gamma}^{i^T} H^i}{\underset{\sim}{\gamma}^{i^T} H^i \underset{\sim}{\gamma}^i}$

Step 8   i=i+1; go to Step 2
```

It can be proven that the updating formula for H^i_{n*n} will, in the limit, converge to the inverse of the Hessian matrix and that if H^0_{n*n} is a positive definite symmetric matrix, then all H^i will be positive definite and symmetric. This will guarantee the descend direction. The Fletcher algorithm does not require the analytical evaluation of the second derivatives of the objective function (which constitute the elements of the Hessian matrix). The first derivatives of the objective function must, however, be supplied to the algorithm.

2.5 Gradient Functions

As stated earlier, the gradient of the objective function and the constraint equations must be supplied to the computer program carrying out the shifted exterior penalty function algorithm. The gradients for $i=1,2,\ldots n_p$, are as follows:

$$\text{(i)} \qquad \frac{\partial F(\phi)}{\partial \eta} = \begin{cases} 1 & \text{if} \quad \eta = r_0 \\ 0 & \text{if} \quad \eta \neq r_0 \end{cases}$$

$$\text{(ii)} \qquad \frac{\partial C_i}{\partial H_0} = \frac{\partial M(\phi,\theta_i)}{\partial H_0} = \frac{M(\phi,\theta_i)}{H_0}$$

$$\text{(iii)} \qquad \frac{\partial C_i}{\partial \alpha_s} = \frac{\partial M(\phi,\theta_i)}{\partial \alpha_s} = H_0 \left(\frac{\partial}{\partial \alpha_s} \prod_{i=1}^{k} \frac{N(\alpha_i)}{D(\beta_i,r_i)}\right)$$

$$= H_0 \, \frac{\prod\limits_{\substack{i=1 \\ i\neq s}}^{k} N(\alpha_i)}{\prod\limits_{i=1}^{k} D(\beta_i,r_i)} \cdot \frac{\partial}{\partial \alpha_s} (N(\alpha_s))$$

$$= M(\phi,\theta_i) \, \frac{\sin\alpha_s}{\cos\beta_i - \cos\alpha_s}$$

(iv) $$\frac{\partial C_i}{\partial \beta_s} = \frac{\partial M(\phi,\theta_i)}{\partial \beta_s} = H_o \frac{\partial}{\partial \beta_s} \frac{\prod_{i=1}^{K} N(\alpha_i)}{\prod_{\substack{i=1 \\ i \neq s}}^{k} D(r_i,\beta_i)} \cdot \frac{\partial}{\partial \beta_s} \left[\frac{1}{D(r_s,\beta_s)}\right]$$

$$= M(\phi,\theta_i) \frac{2r_s \sin\beta_s [2r_s \cos\beta_s - \cos\theta_i (1+r^2)]}{D^2(r_s,\beta_s)}$$

(v) $$\frac{\partial C_i}{\partial r_s} = H_o \frac{\partial}{\partial r_s} \frac{\prod_{i=1}^{k} N(\alpha_i)}{\prod_{i=1}^{k} D(r_i,\beta_i)}$$

$$= H_o \frac{\prod_{i=1}^{k} N(\alpha_i)}{\prod_{\substack{i=1 \\ i \neq s}}^{k} D(r_i,\beta_i)} \cdot \frac{\partial}{\partial r_s} \frac{1}{D(r_s,\beta_s)}$$

$$= M(\phi,\theta_i) \frac{-2r_s^3 + 2r_s[1 - 2\cos^2\theta_i - 2\cos^2\beta_s] + 2\cos\theta_i \cos\beta_s [1+3r_s^2]}{D^2(r_s,\beta_s)}$$

where $N(\alpha_i)$, $D(r_i,\beta_i)$ are magnitudes of the numerator and denominator functions respectively as defined in Chapter 1. For $i=1+n_p,\ldots,2n_p$, $1+2n_p,\ldots,2n_p+n_a$, the gradient functions would be the negative of the corresponding gradient functions in steps (i) through (v).

CHAPTER III

DESIGN EXAMPLES

3.4 Introduction

By solving the optimization problem of Section 2.3 for a given set of filter specifications $\underset{\sim}{S}$, the coefficients of a minimized-pole-distance filter can be obtained. Alternatively, passband ripple can be minimized and an equivalent elliptic filter satisfying the same set of specification may be realized. In either case, the starting point can be a minimum-order elliptic filter which itself satisfies $\underset{\sim}{S}$.

The task in this chapter is to design minimized-pole-distance filters and to compare them with the minimum-order and the minimized-ripple filters satisfying the same set of specifications. Several sets of specifications will be considered. There are 3 sets of specifications in a group and 3 groups will be considered. Particular attention will be placed on only one member of each group so that results can be verified and studied in detail. The sampling frequency is 5000, rad/s for all filters.

3.1 Specifications and Design

Table 3.1 represents the set of specifications considered. The approximations are based on the nominal values of A_p and A_a rather than the corresponding maximum values so as to allow a margin for the effects of coefficient quantization. The set of specifications are put into 3 groups, each group having 3 members. The individual sets of specifications were chosen such that designs satisfying those specifications would have

Table 3.1 Filter specifications

			A_p,dB		A_a,dB	
EXAMPLE	ω_p, rad/s	ω_a, rad/s	Nom	Max	Nom	Max
1	800	2100	.1	.105	70	66.50
2	250	975	.1	.105	70	66.50
3	80	350	.1	.105	70	66.50
4	800	1600	.5	.525	45	42.75
5	250	450	.5	.525	45	42.75
6	80	150	.5	.525	45	42.75
7	800	1100	1.	1.05	35	33.25
8	250	340	1.	1.05	35	33.25
9	80	110	1.	1.05	35	33.25

different selectivity factors. Hence, given the set of specifications in Table 3.1;

(i) the minimum-order elliptic approximations were obtained using the method in [1]. The order of approximation was 4 in each case and hence these filters can be implemented by connecting two second-order sections in cascade.

(ii) the approximation order was increased from 4 to 6 and the elliptic approximation leading to the minimum passband ripple without violating the remaining specifications was deduced by using File No. 6 in Appendix B of [1].

(iii) as in step (ii), the approximation order was increased from 4 to 6. The optimization technique of Chapter II was then used to minimize the maximum pole radius without violating the prescribed specifications. The initial starting point $\underset{\sim}{\phi}^{0}$ was computed by using the minimum-ripple transfer function obtained in step (ii) above. The initial value of r_0 was assumed to be the maximum pole radius. A suitable number of discrete frequencies for the optimization is 3 to 4 times the number of adjustable filter parameters.

The coefficients of the transfer functions obtained in steps (i), (ii), (iii) are given in Tables 3.2(a)-3.2(c) and the corresponding amplitude responses for example 3,5 and 7 are depicted in Figures 3.1(a) through 3.9(c).

Table 3.2(a) Coefficients of the 4th-order elliptic filters

EXAMPLE	a_{11}	b_{11}	b_{01}	a_{12}	b_{12}	b_{02}	H_o
1	1.961426	-.854963	.256905	1.7866161	-.7095954	.6604853	.025211
2	1.069546	-1.612782	.665201	-.537255	-1.735794	.855006	.0013868
3	-.99912606	-1.872388	.878487	-1.7795879	-1.937118	.9501383	$3.5583*10^{-4}$
4	1.7654032	-1.113335	.4023972	.969624	-.9039677	.7600289	.0208885
5	-.6702996	-1.717688	.7545492	-1.647928	-1.814799	.9141612	.007386
6	-1.7981827	-1.910299	.9143572	-1.959493	-1.960375	.9709445	.0049538
7	1.1466595	-1.178264	.4691026	-.2580147	-.9854304	.8374324	.0402949
8	-1.233856	-1.756025	.790902	-1.800438	-1.847536	.9427507	.019359
9	-1.9054172	-1.924145	.9279504	-1.9787137	-1.970923	.9809379	.0168689

$$H(z) = H_o \prod_{i=1}^{2} \frac{z^2+a_{1i}z+1}{z^2+b_{1i}z+b_{0i}}$$

Table 3.2(b) Coefficients of the 6th-order elliptic filters

EXAMPLE	a_{11}	b_{11}	b_{01}	a_{12}	b_{12}	b_{02}	a_{13}	b_{13}	b_{03}	H_o
1	1.98225	.60364	.11026	1.87174	.67455	.28678	1.76808	.85003	.67755	.14625
2	1.51203	-1.05706	.29323	-.02513	-1.19656	-.46944	-.61696	-1.44915	.78526	.00226
3	-.31477	-1.64812	.68195	-1.64084	-1.74120	.77754	-1.79701	-1.87999	.91884	.00040
4	1.88860	.22273	.0466616	1.31153	.173838	.28888	.90410	.13410	.70710	.09140
5	.07070	-1.36323	.47887	-1.46271	-1.5443	.68228	-1.67151	-1.7382	.89847	.00700
6	-1.58867	-1.76777	.78363	-1.93507	-1.86058	.8777	-1.96245	-1.94552	.963697	.00492
7	1.54919	-.36375	.0900178	.17833	-.51246	.44658	-.32455	-.66411	.822123	.06064
8	-.6630	-1.46021	.54960	-1.70637	-1.66430	.77411	-1.81194	-1.81419	.93939	.01678
9	-1.80465	-1.81143	.82287	-1.9679	-1.90589	.91860	-1.98000	-1.96595	.97940	.01560

$$H(z) = H_o \prod_{i=1}^{3} \frac{z^2+a_{1i}z+1}{z^2+b_{1i}z+b_{0i}}$$

Table 3.2(c) Coefficients of the 6th-order minimized-maximum-pole filters

EXAMPLE	a_{11}	b_{11}	b_{01}	a_{12}	b_{12}	b_{02}	a_{13}	b_{13}	b_{03}	H_o
1	1.801597	.317877	.219744	1.998705	.3342	.219716	1.999982	.333526	.219704	.060684
2	1.499	-1.24175	.555058	-.021474	-1.2546	.555186	-.618759	-1.09746	.337689	.0023523
3	-.368976	-1.65569	.690586	-1.62326	-1.78876	.8260952	-1.795378	-1.79059	.8261013	.0003658
4	2.0	-.058334	.2429383	2.0	-.0585763	.2429225	.9406022	-.053628	.242922	.035108
5	.05183051	-1.359988	.484097	-1.461139	-1.619555	.7688157	-1.671142	-1.61951	.768816	.007394
6	-1.50741	-1.78497	.800647	-1.93481	-1.89503	.912802	-1.962369	-1.89496	.912816	.004003
7	1.561064	-.4196122	.1097057	.1934951	-.659746	.6035348	-.3209151	-.659749	.6035349	.04431853
8	-.66778	-1.445262	.544173	-1.707233	-1.73248	.851716	-1.811467	-1.732323	.851561	.0180835
9	-1.78787	-1.815387	.827079	-1.967013	-1.93523	.948317	-1.97998	-1.93437	.948325	.013374

$$H(x) = H_o \prod_{i=1}^{3} \frac{z^2+a_{1i}z+1}{z^2+b_{1i}z+b_{0i}}$$

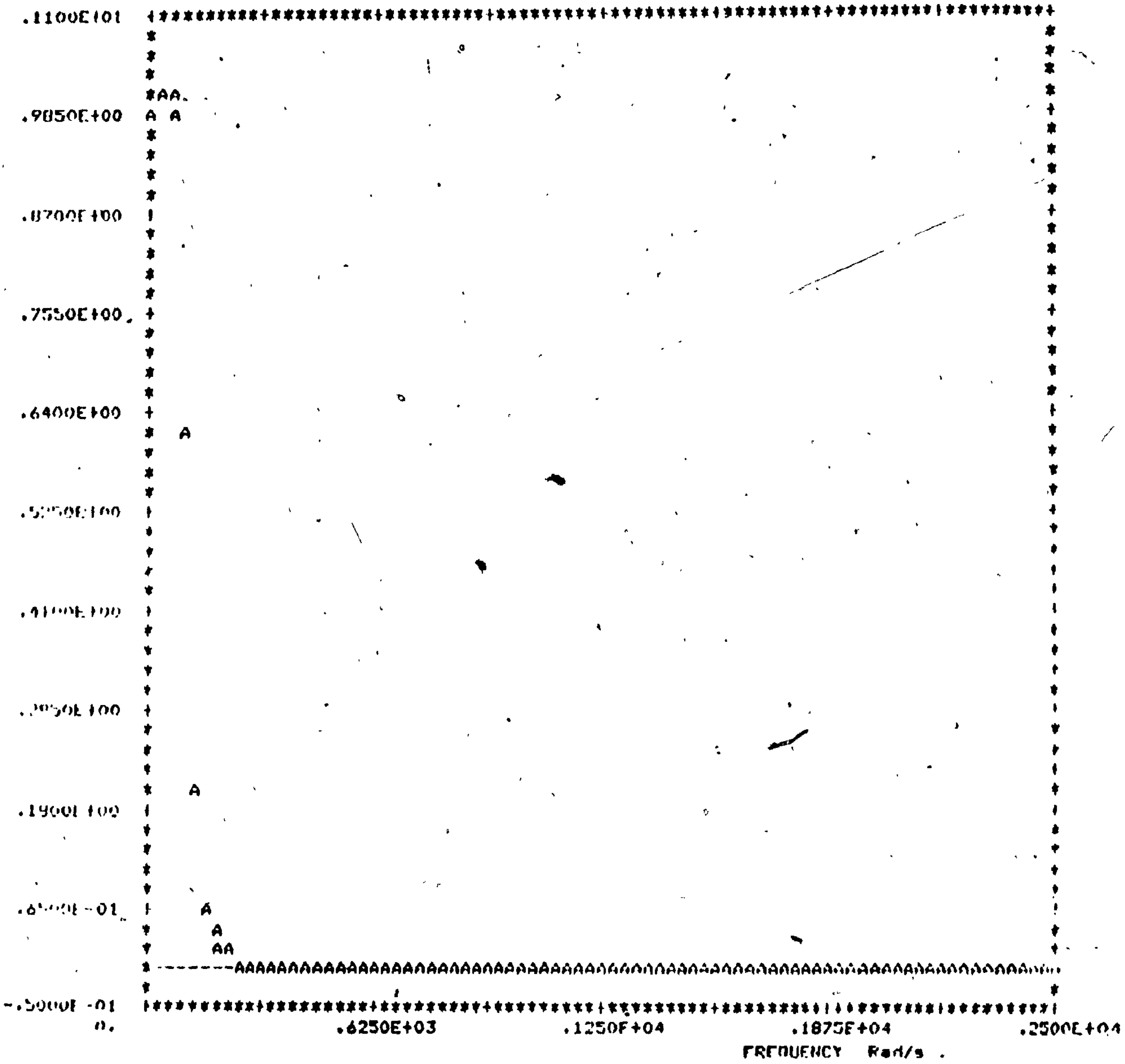


Figure 3.1(a) Overall amplitude response

Example: 3

Type: 2-section-elliptic

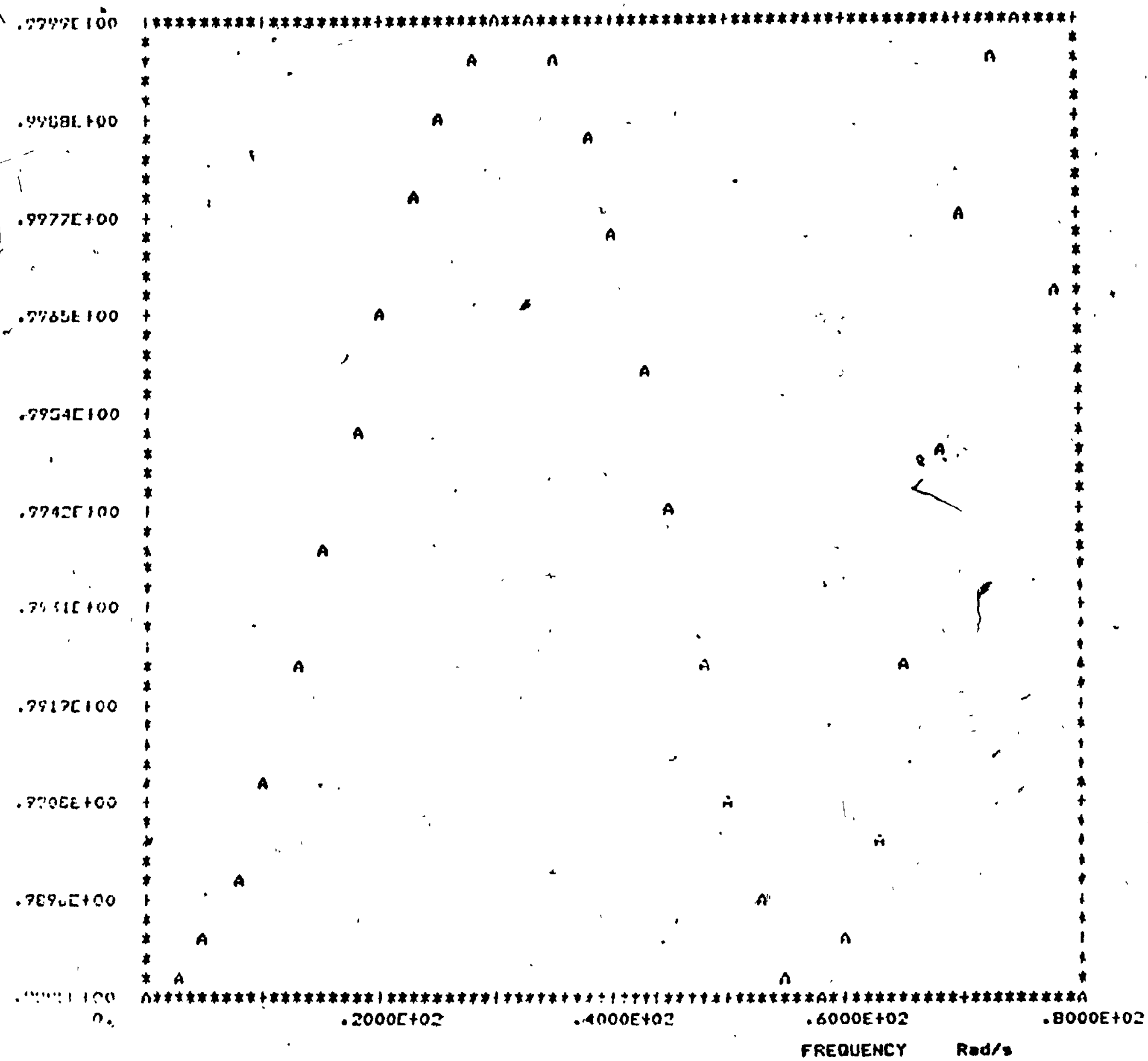


Figure 3.1(b) Amplitude response (passband)

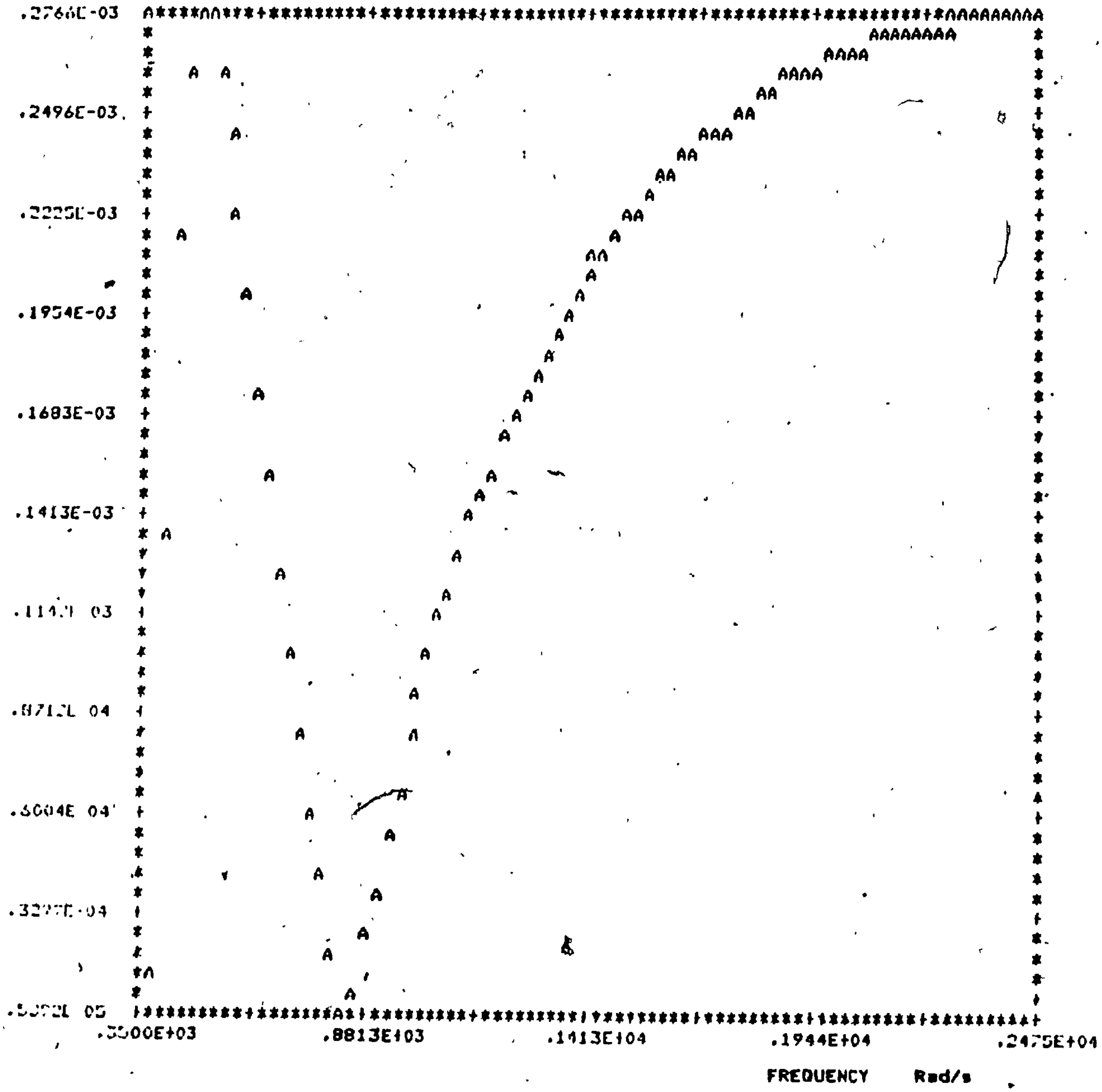


Figure 3.1(c) Amplitude response (stopband)

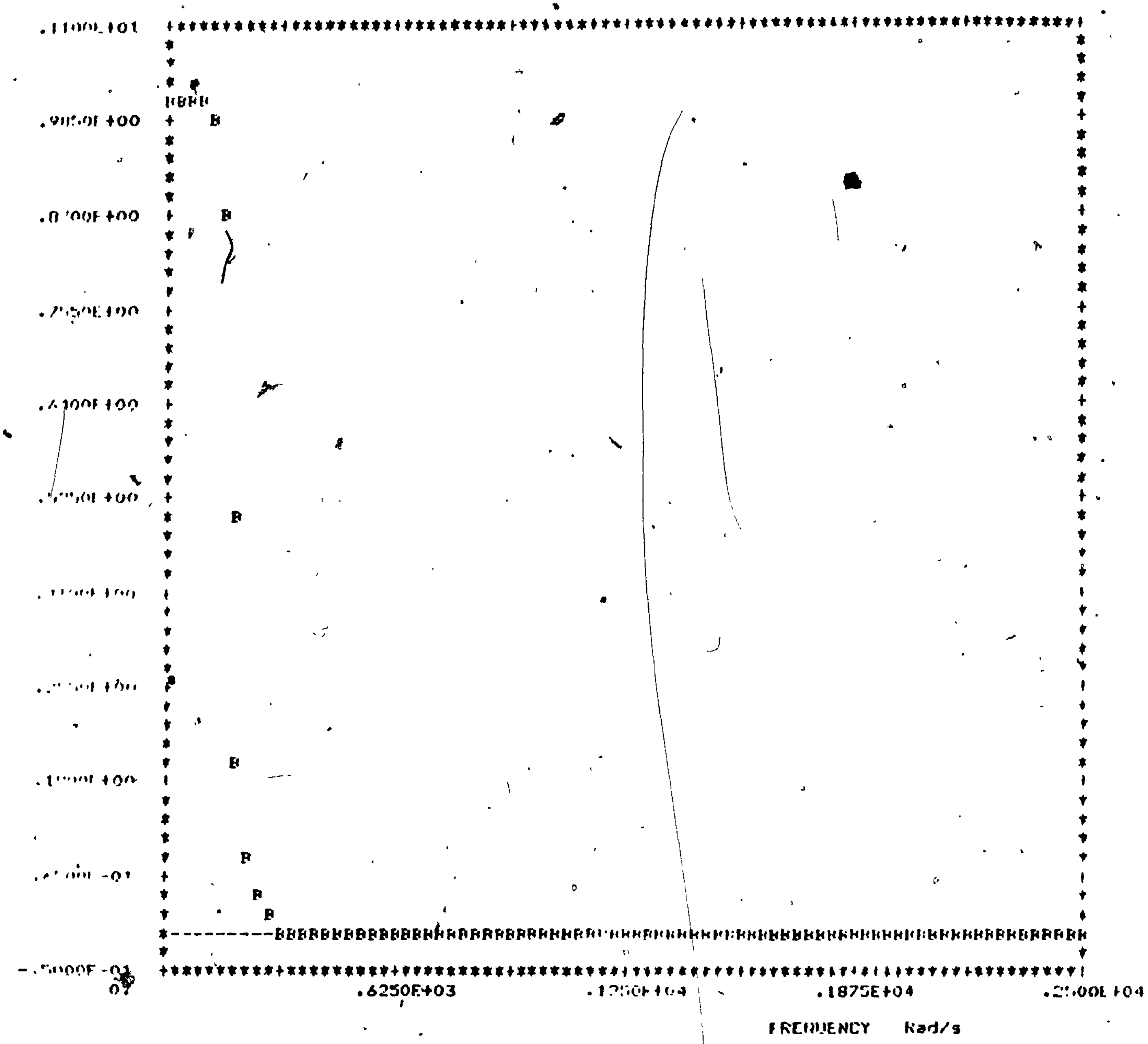


Figure 3.2(a) Overall amplitude response
Example: 3
Type: 3-section-elliptic

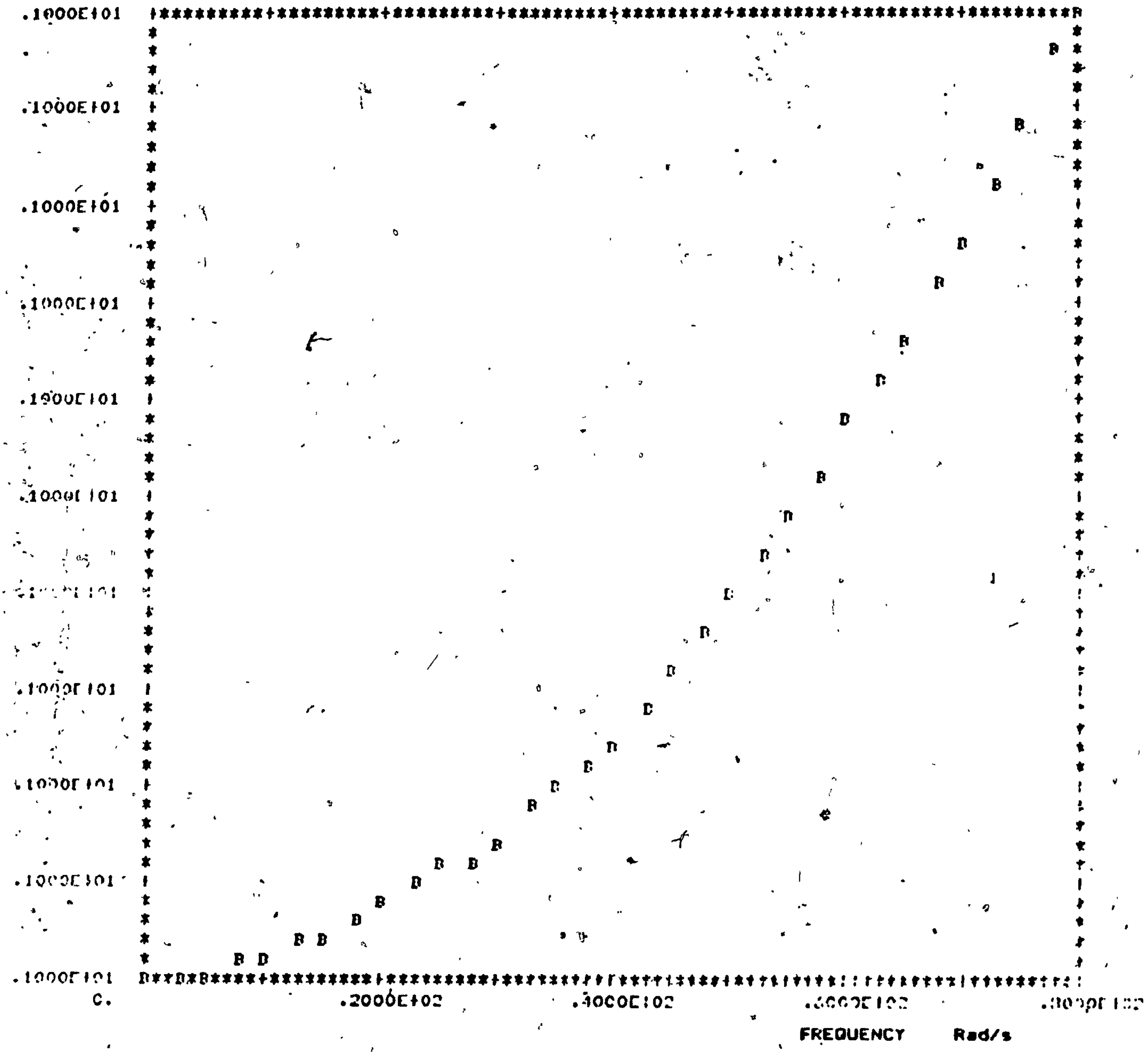


Figure 3.2(b) Amplitude response (passband)

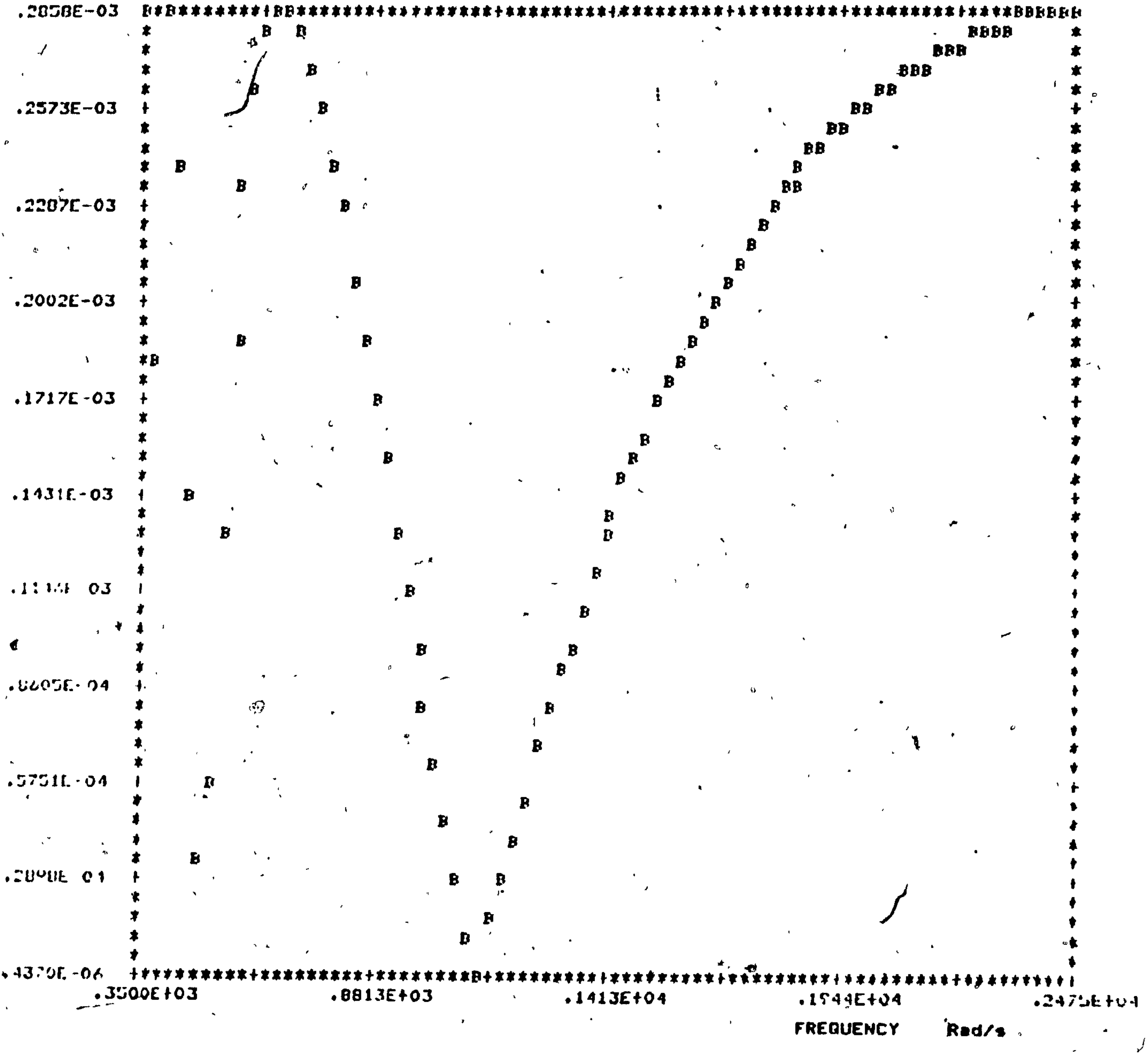


Figure 3.2(c) Amplitude response (stopband)

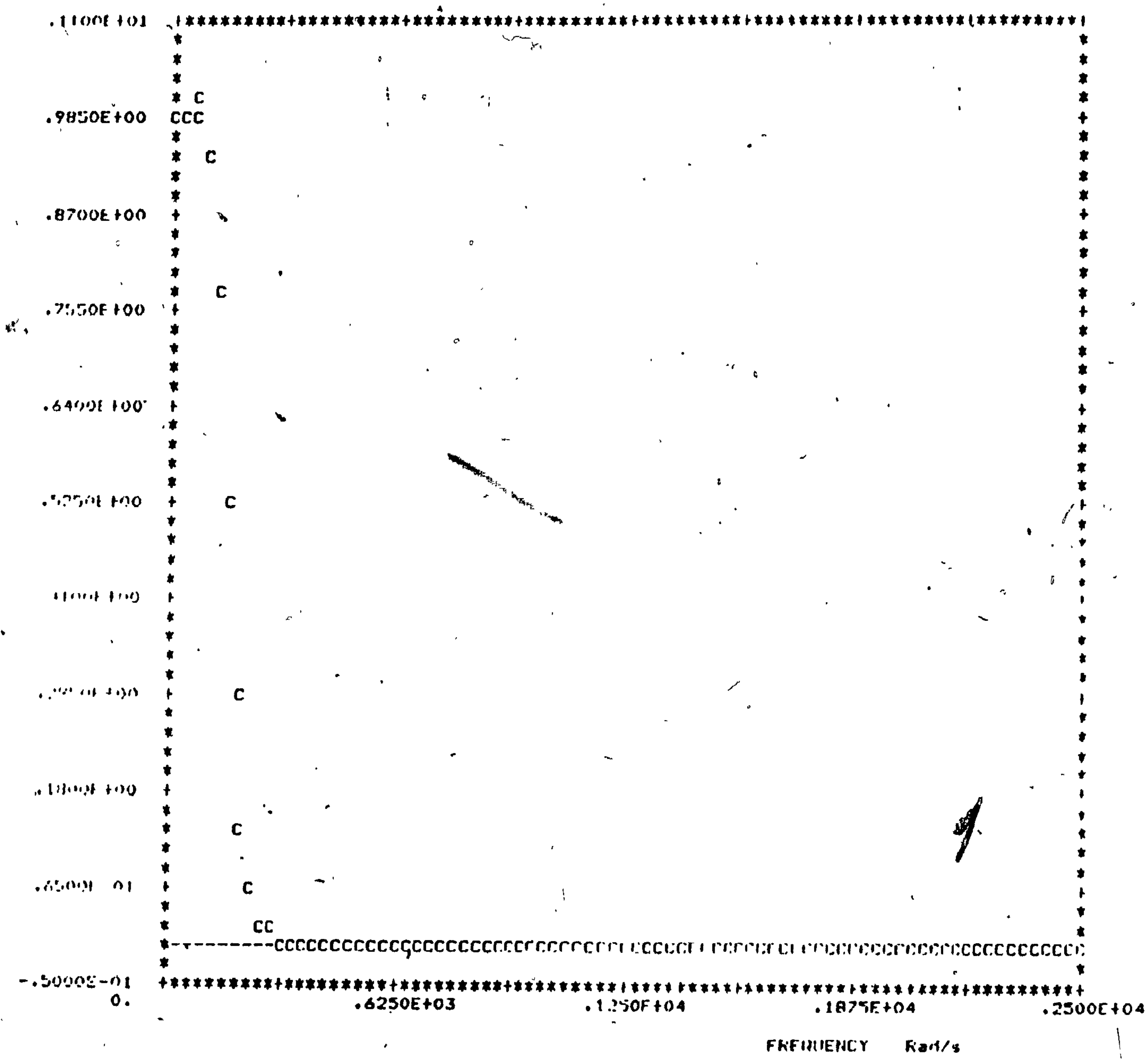


Figure 3.3(a) Overall amplitude response
Example: 3
Type: 3 section minimized-maximum-pole

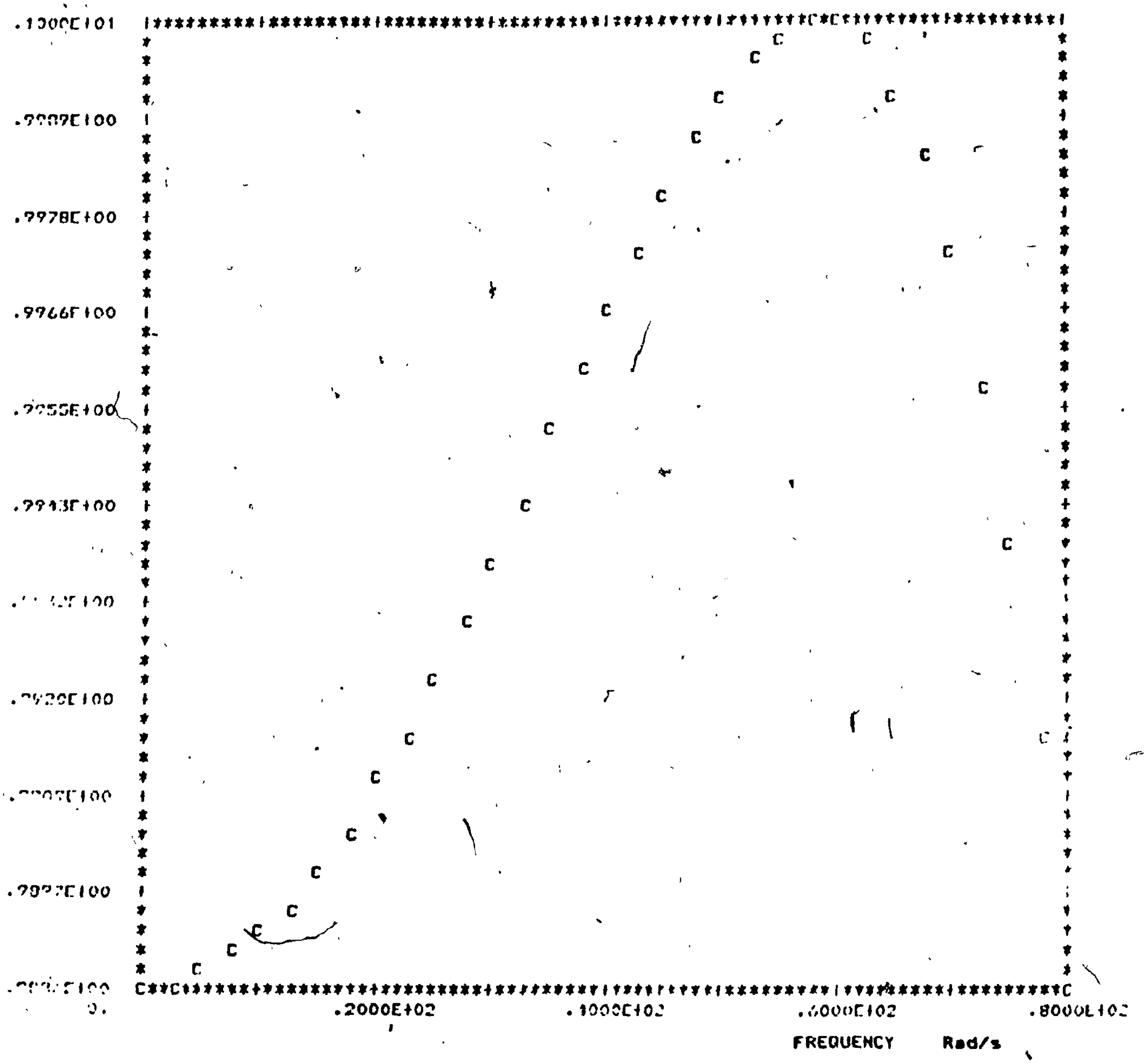


Figure 3.3(b) Amplitude response (passband)

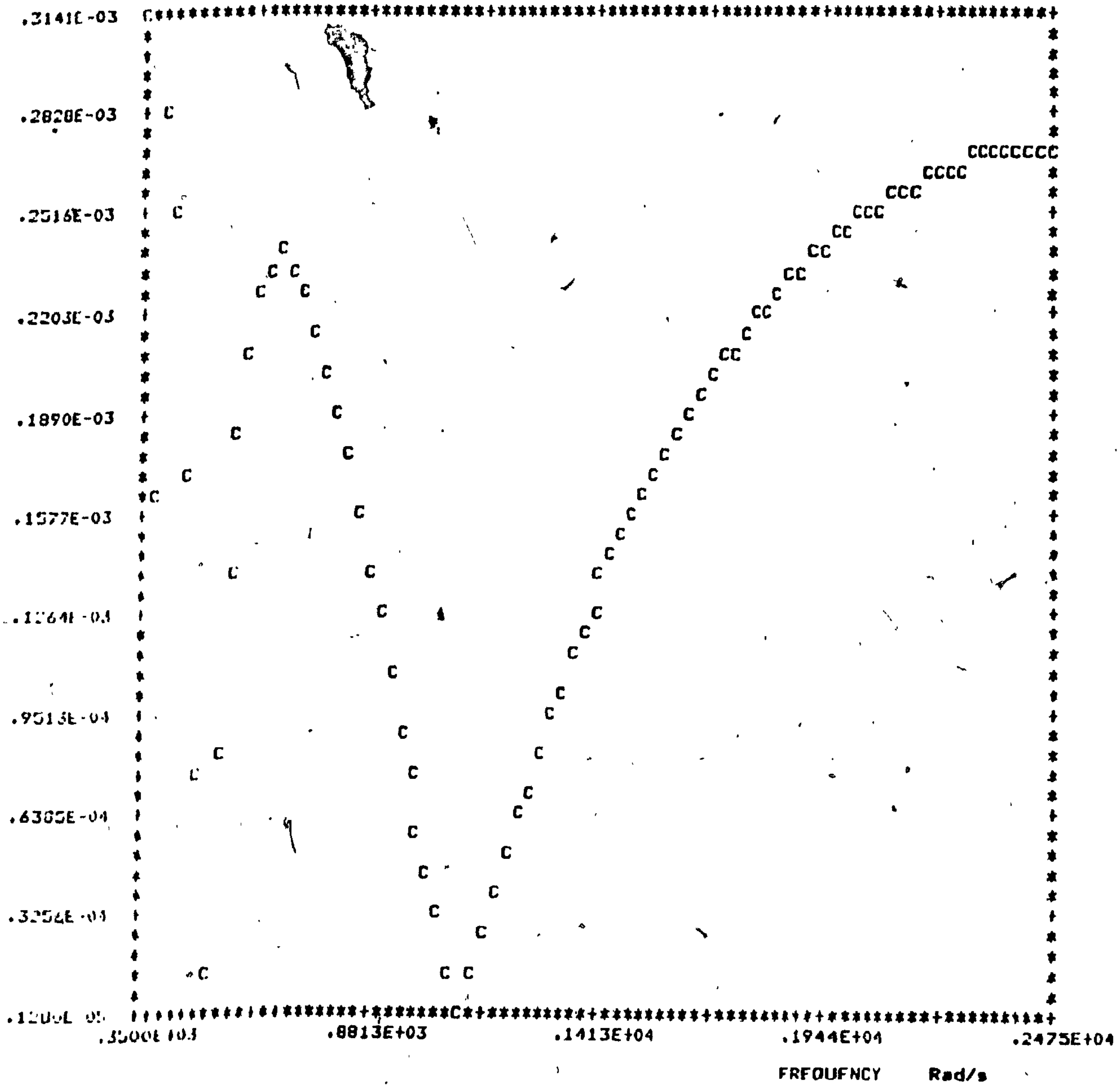


Figure 3.3(c) Amplitude response (stopband)

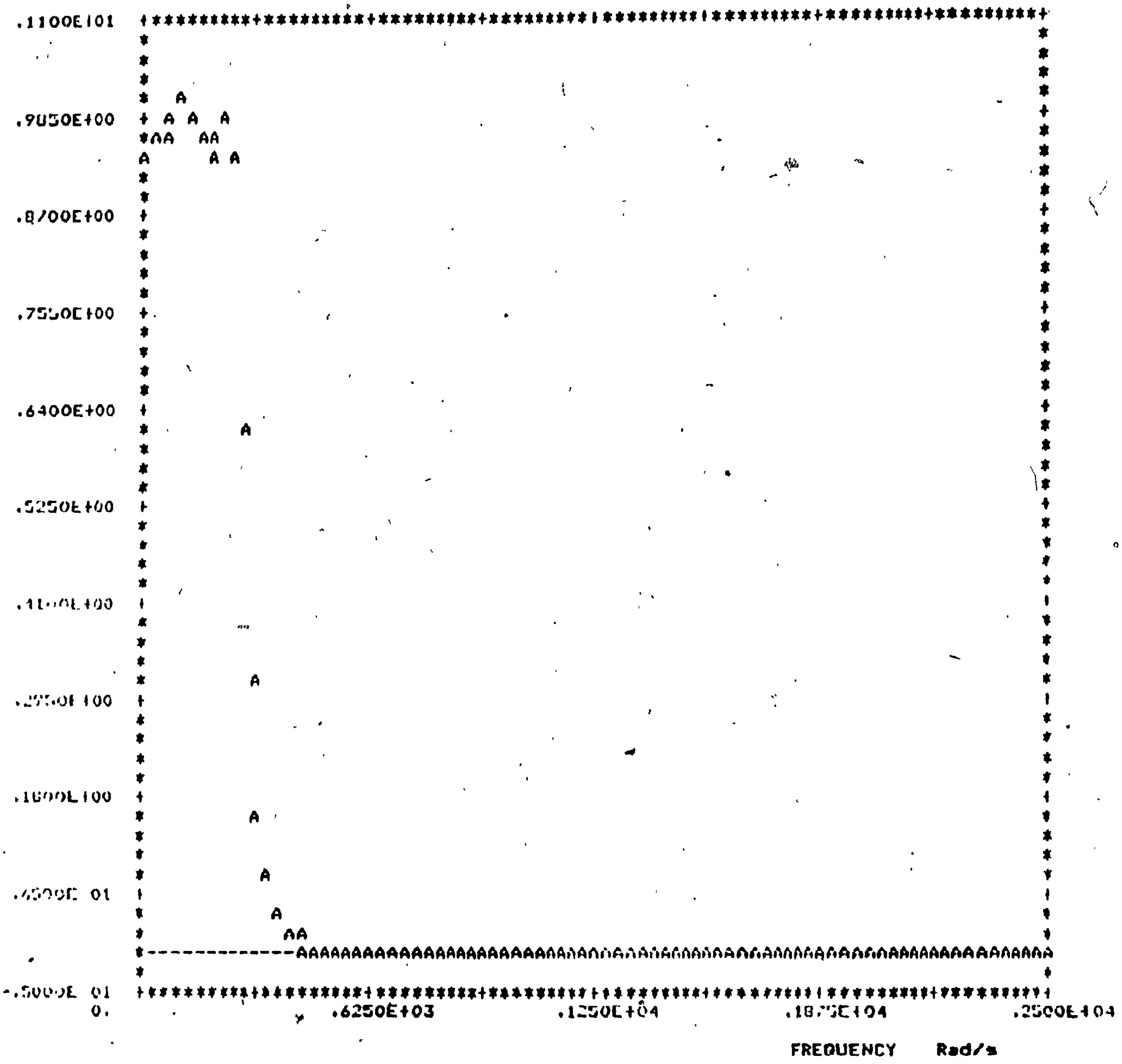


Figure 3.4(a) Overall amplitude response

Example: 5

Type: 2-section elliptic

FERIEURE

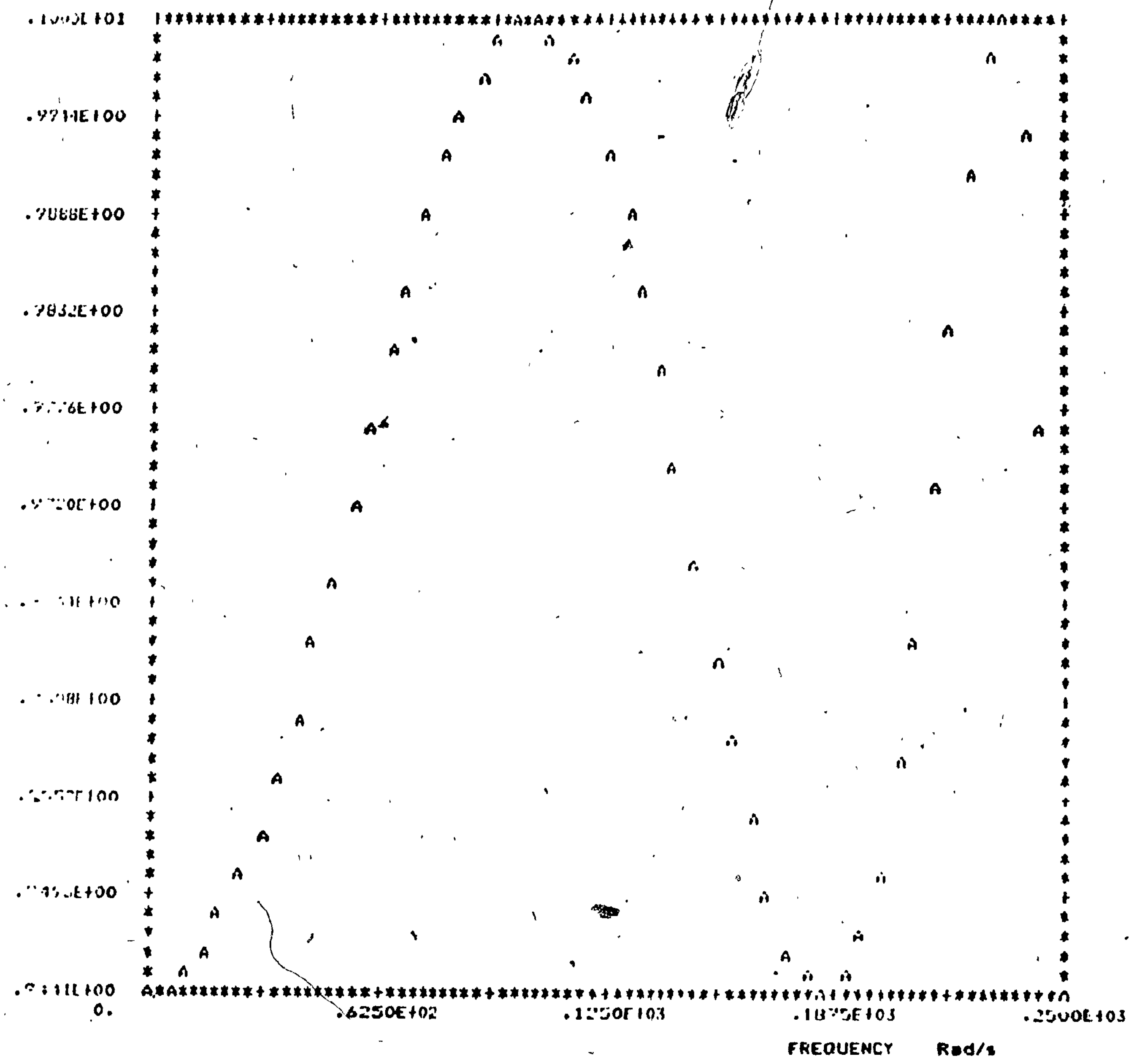


Figure 3.4(b) Amplitude response (passband)

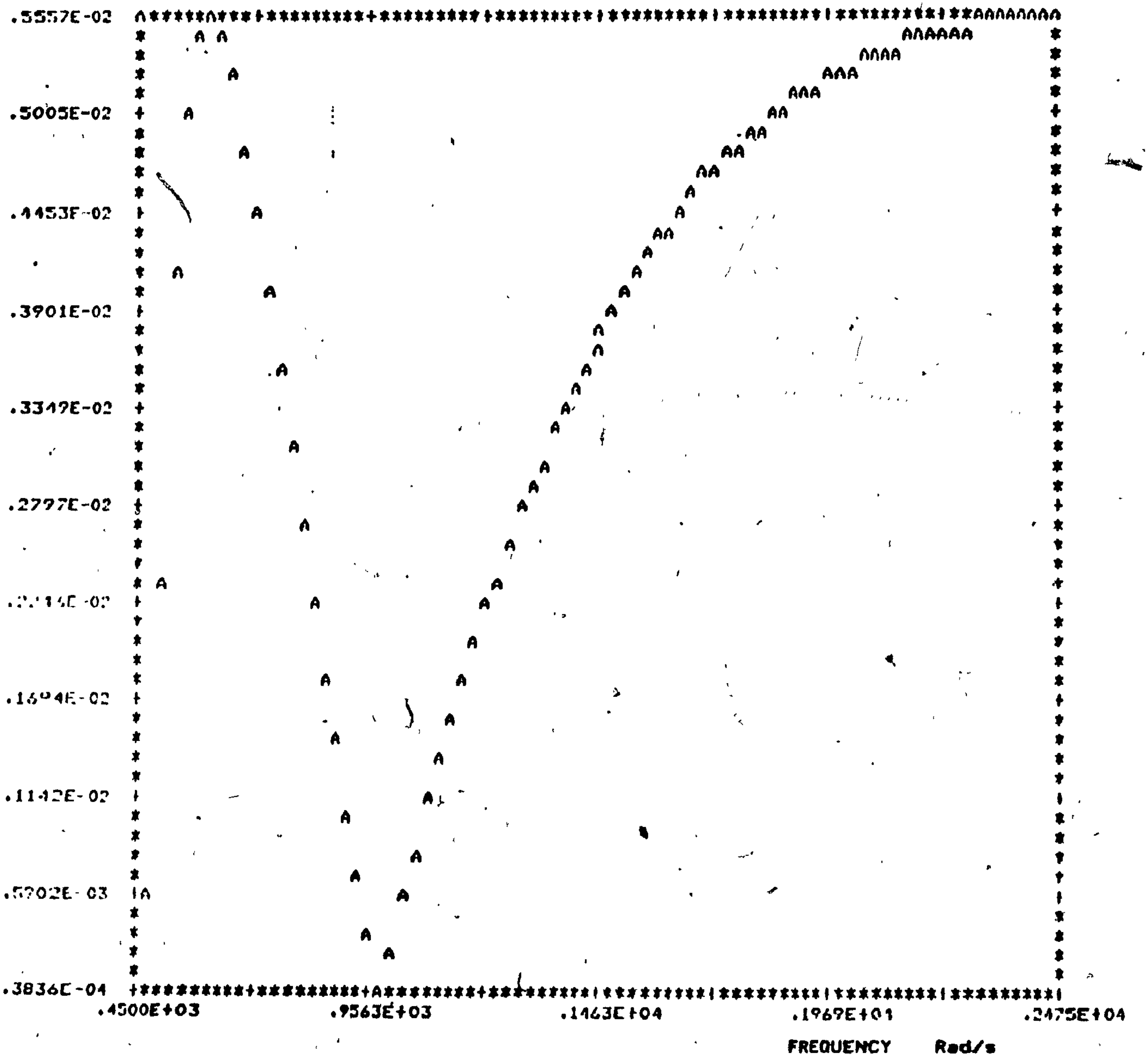


Figure 3.4(c) Amplitude response (stopband)

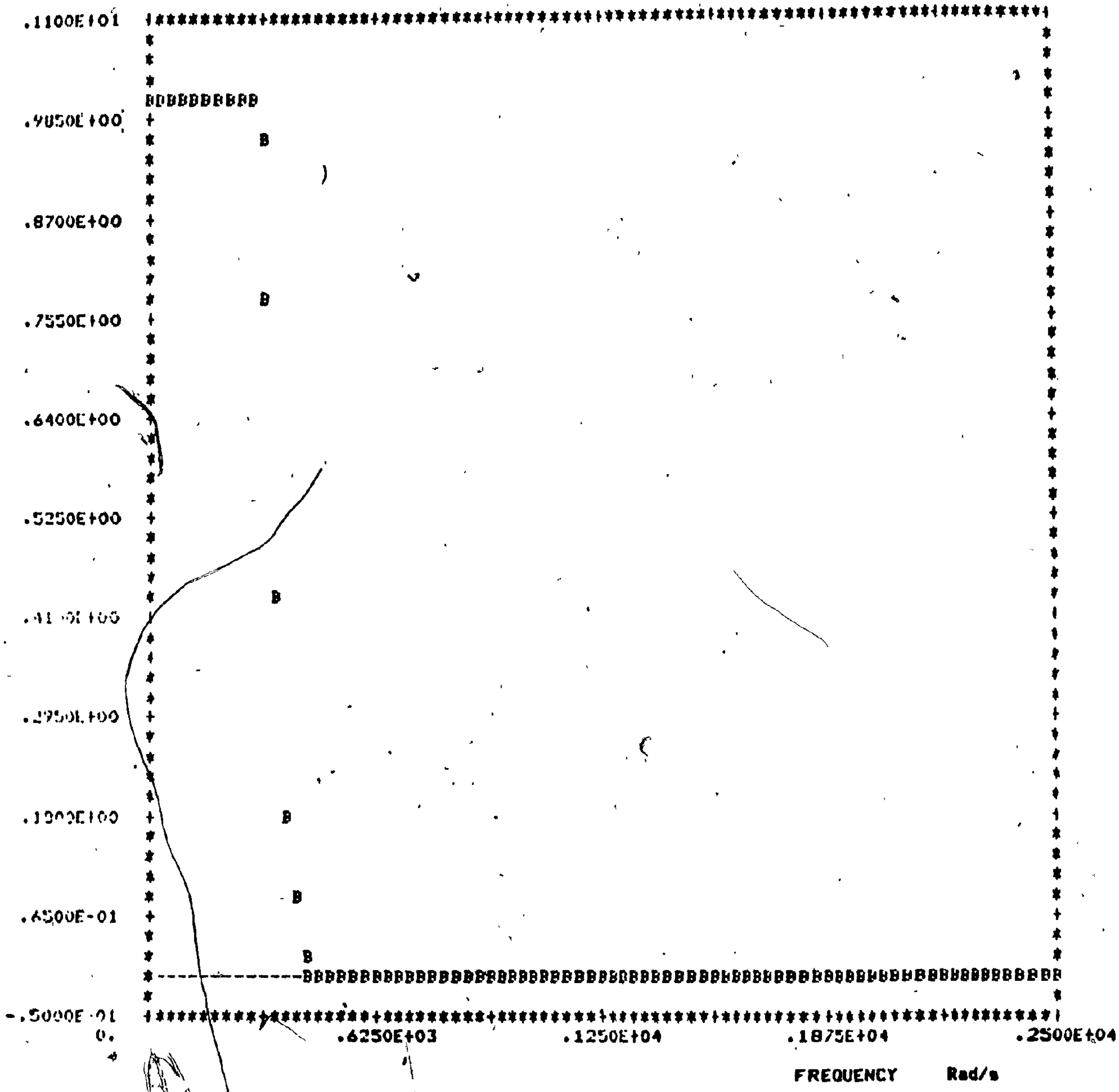


Figure 3.5(a) Overall amplitude response

Example: 5

Type: 3-section elliptic

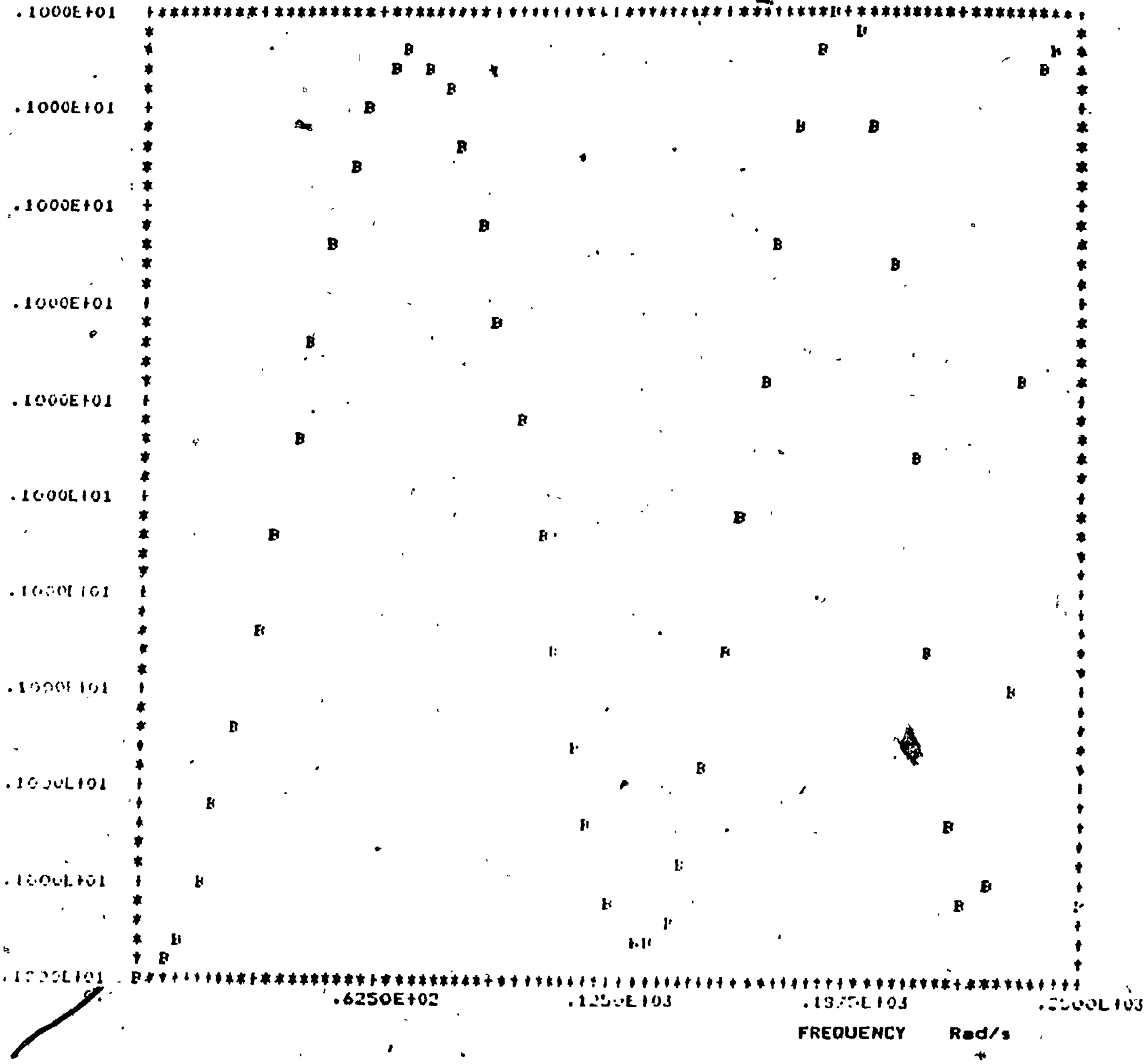


Figure 3.5(b) Amplitude response (passband)

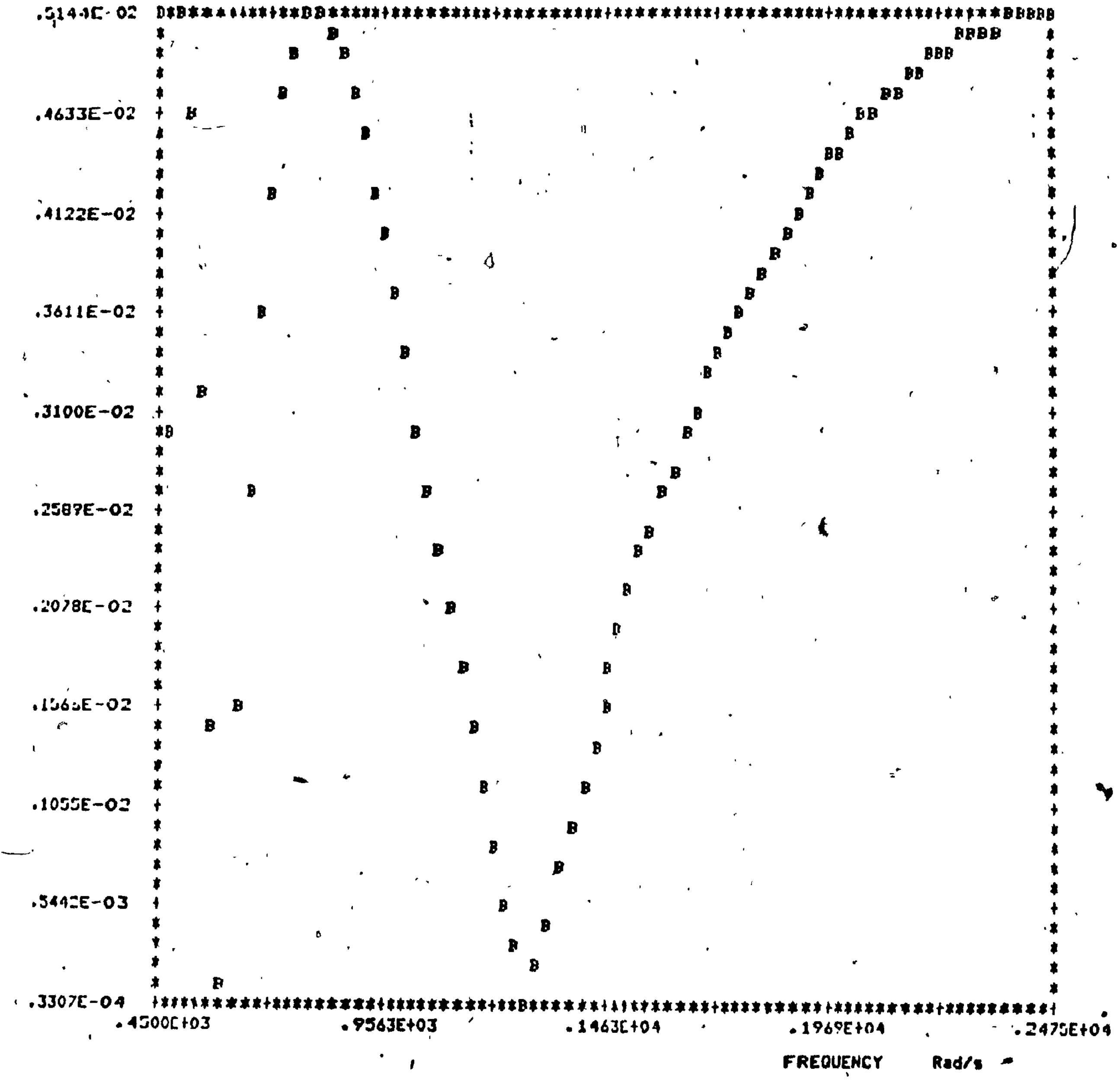


Figure 3.5(c) Amplitude response (stopband)

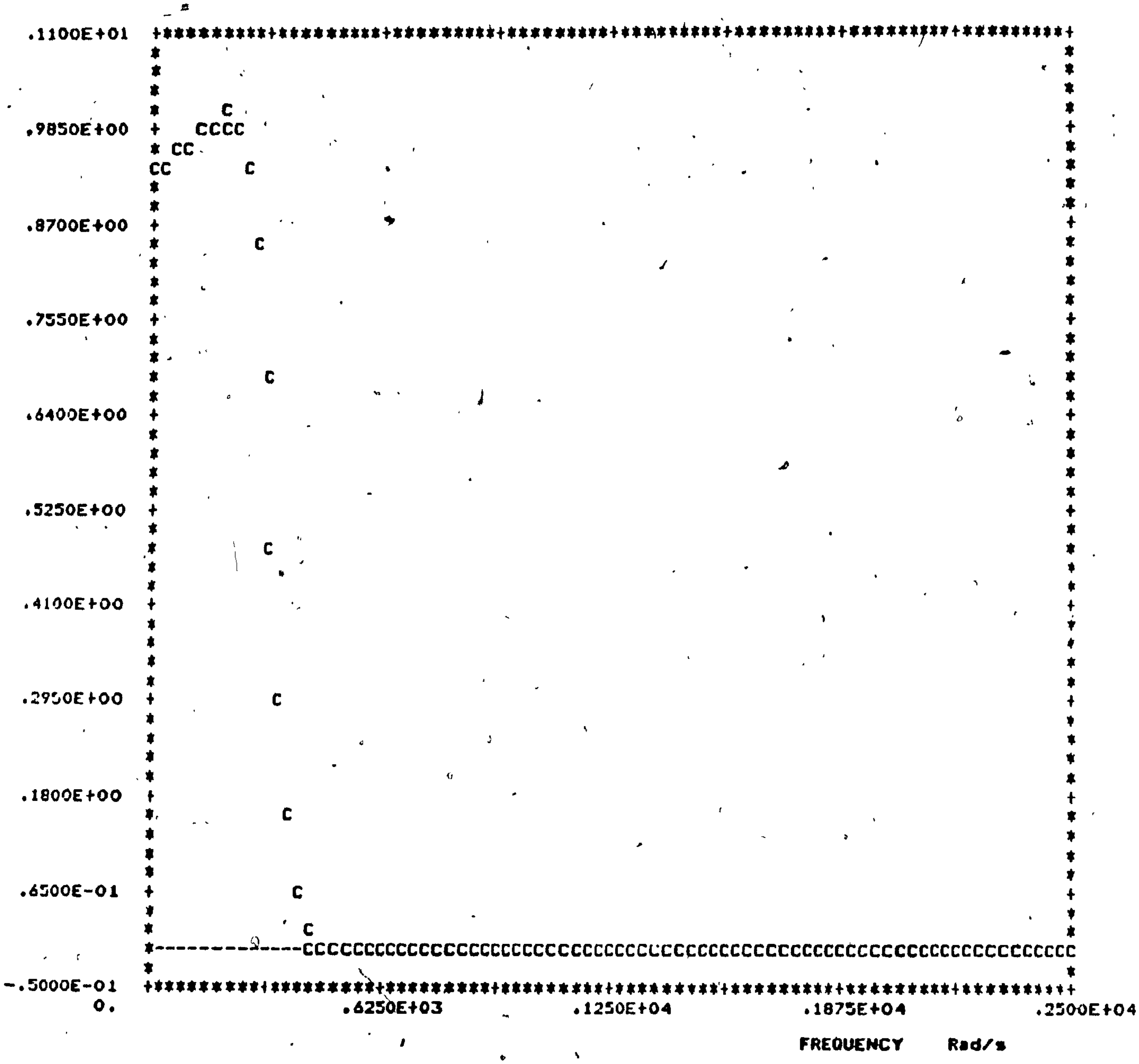


Figure 3.6(a) Overall amplitude response

Example: 5

Type: 3-section minimized-maximum pole

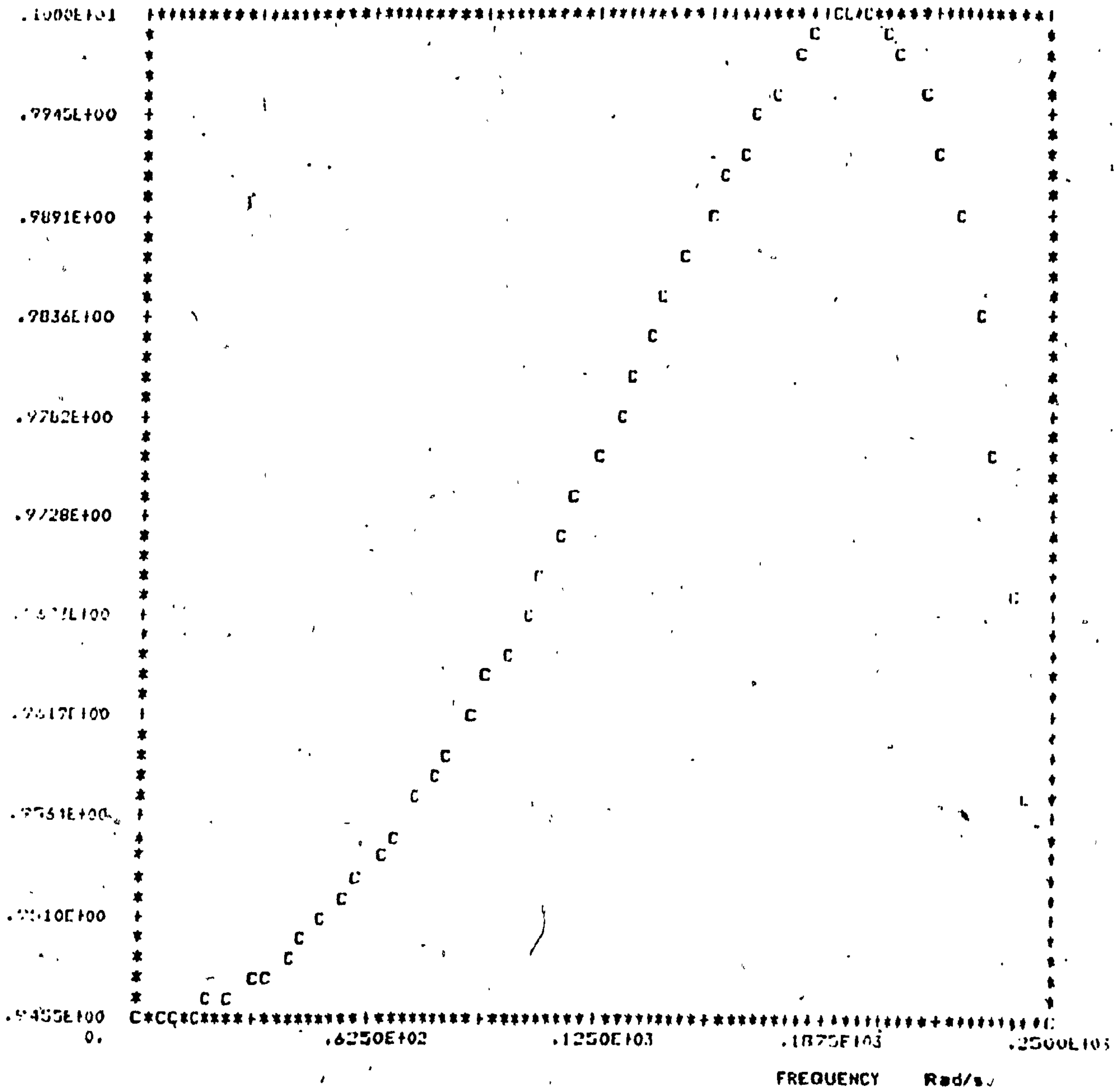


Figure 3.6(b) Amplitude response (passband)

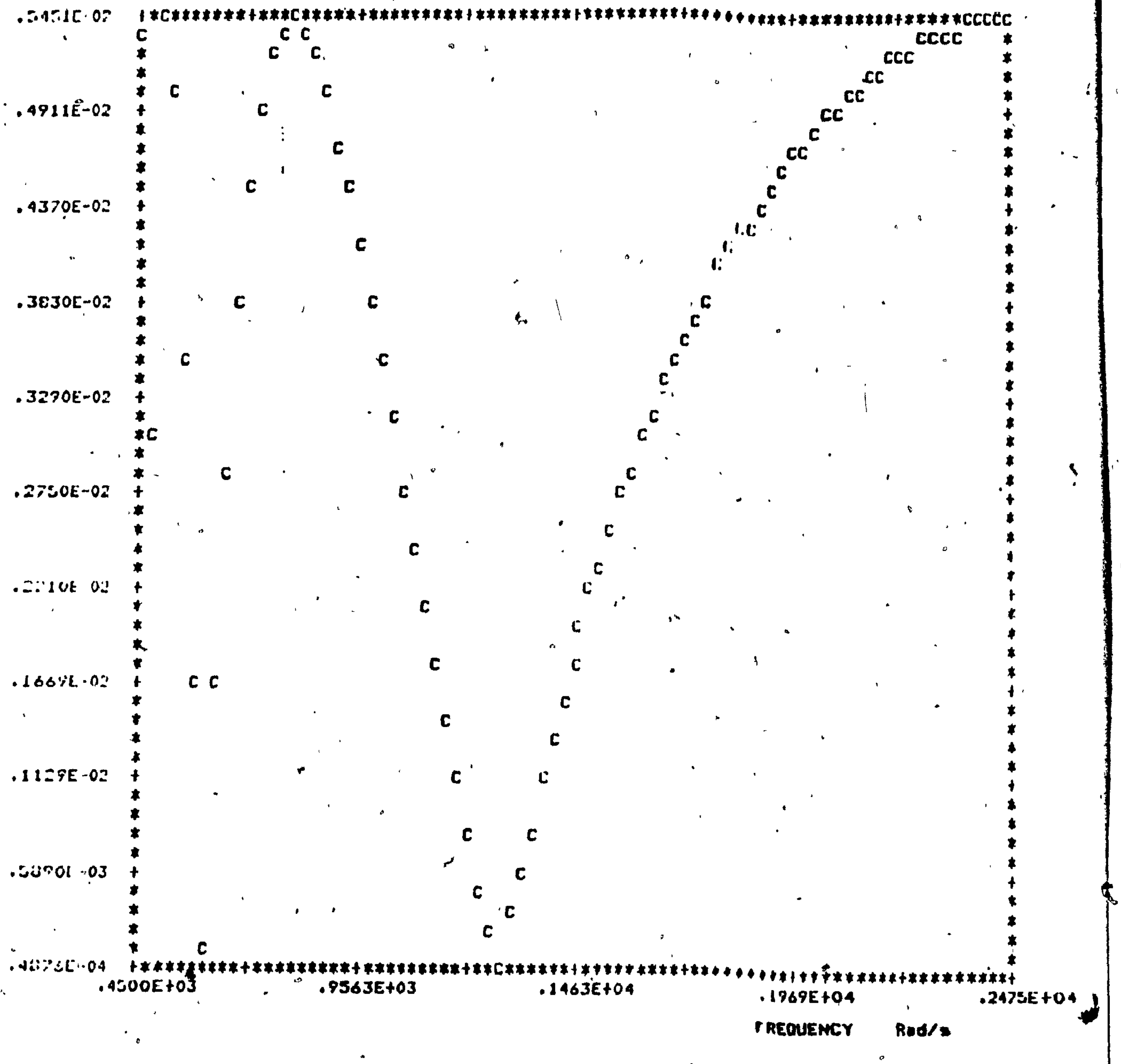


Figure 3.6(c) Amplitude response (stopband)

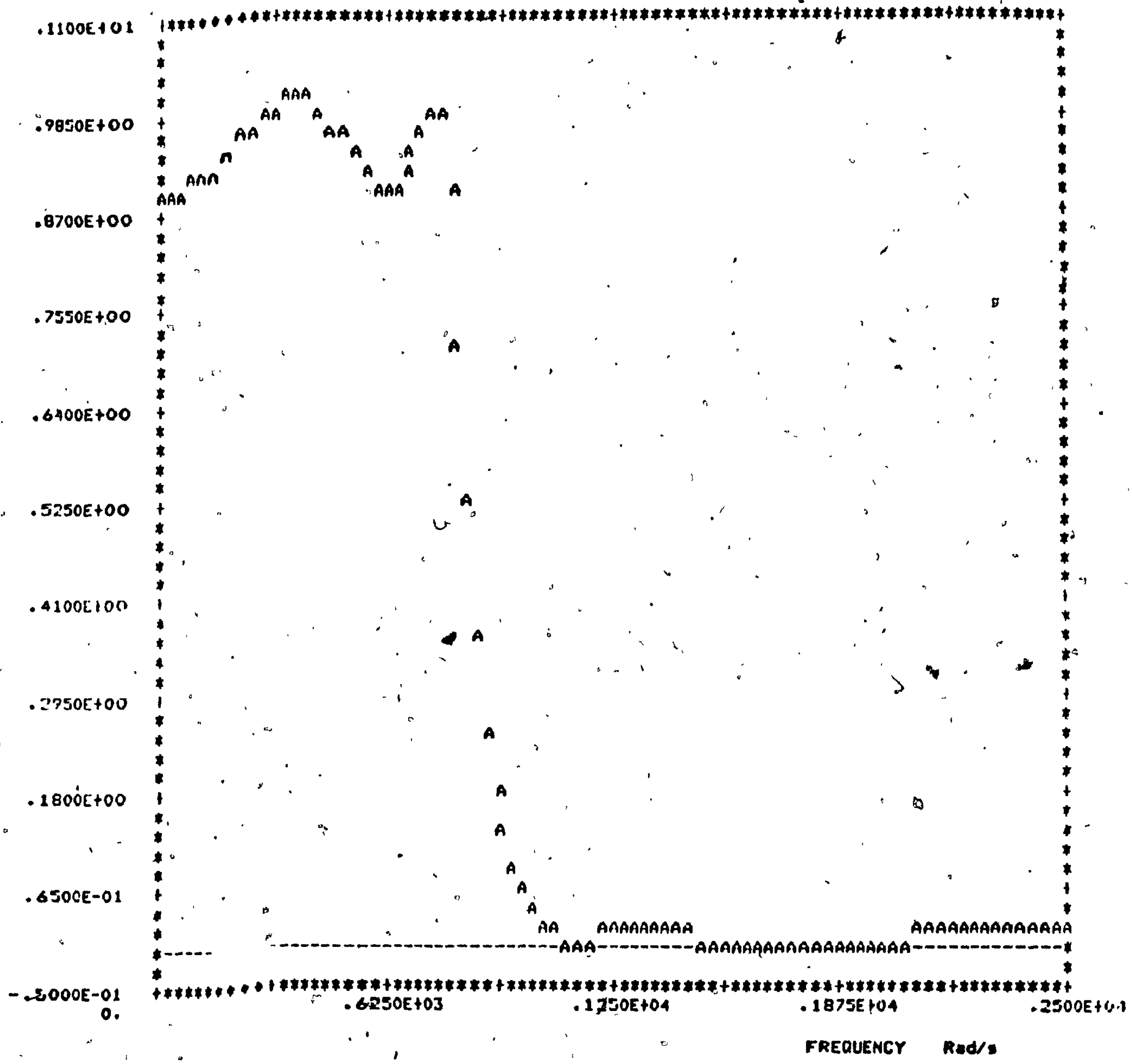


Figure 3.7(a) Overall amplitude response

Example: 7

Type: 2-section elliptic

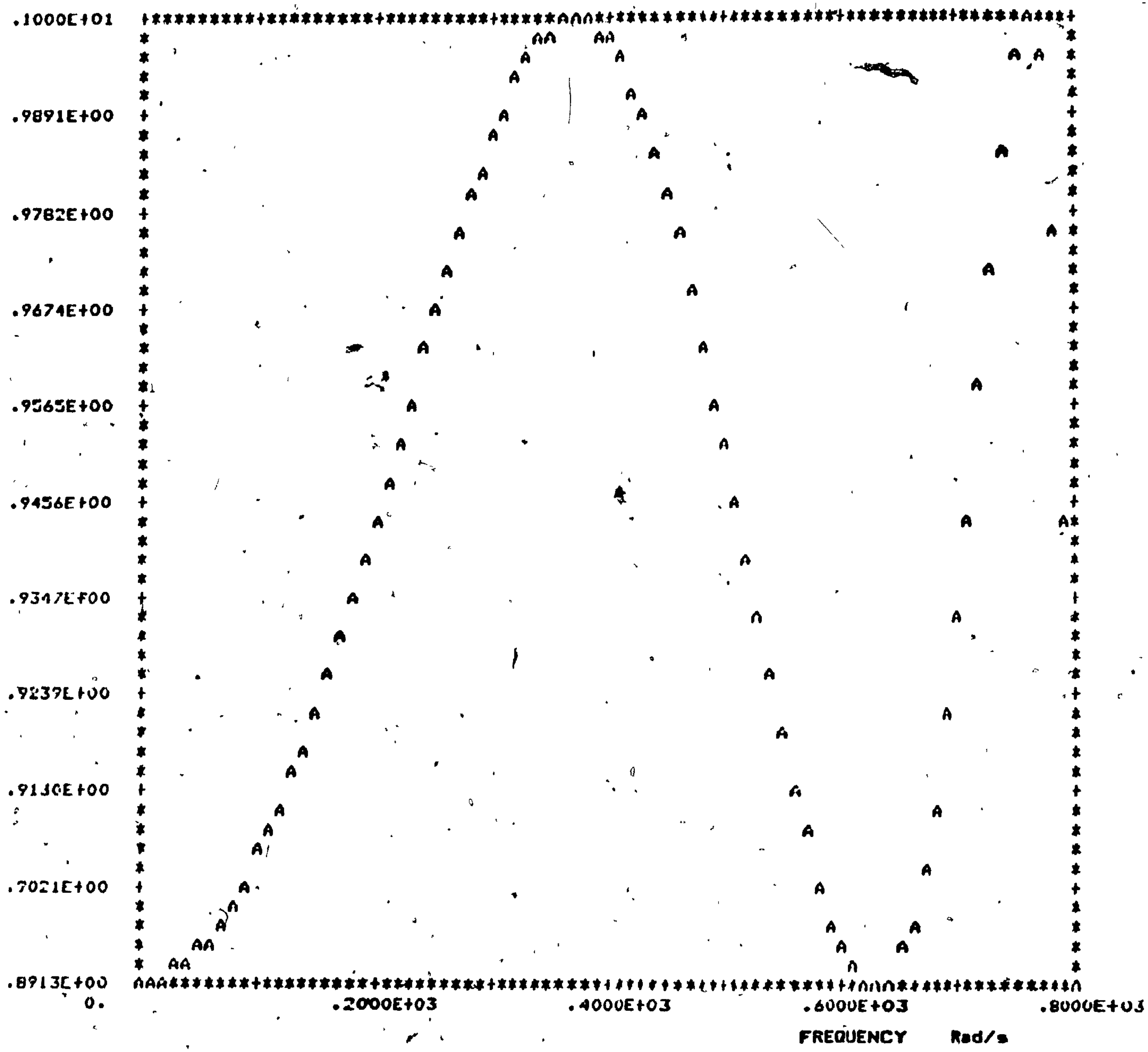


Figure 3.7(b) Amplitude response (passband)

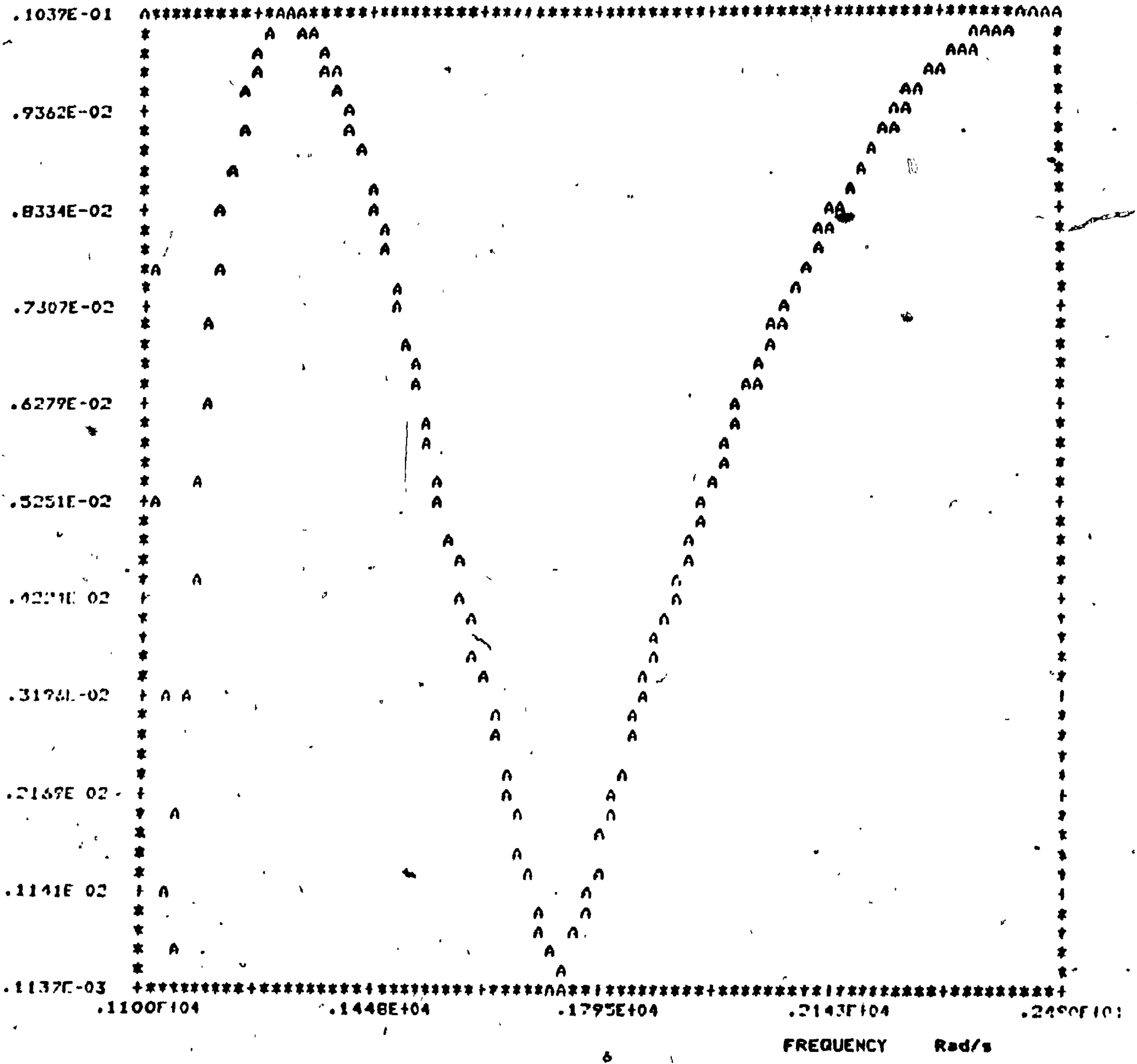


Figure 3.7(c) Amplitude response (stopband)

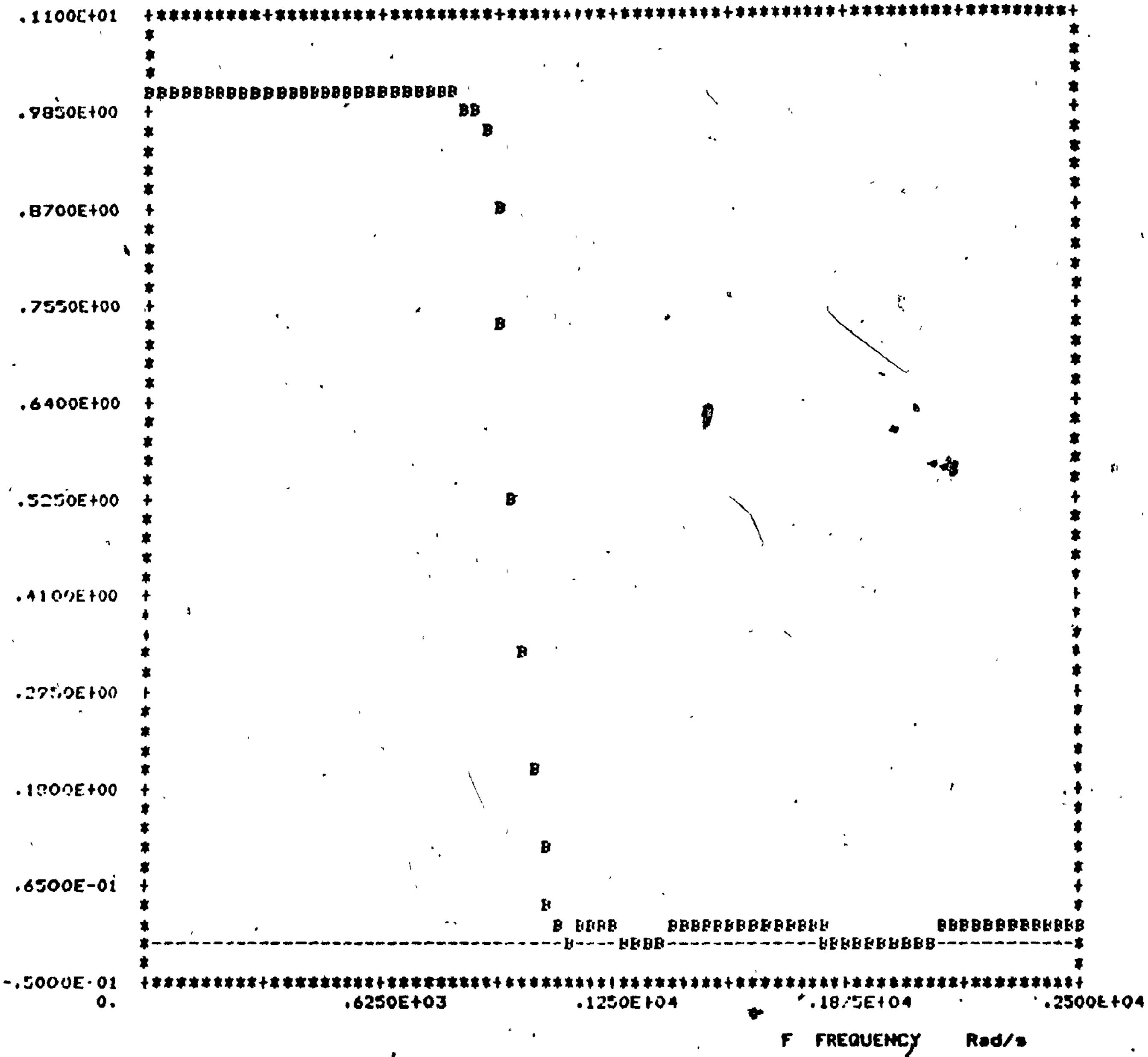


Figure 3.8(a) Overall amplitude response

Example: 7

Type: 3-section elliptic

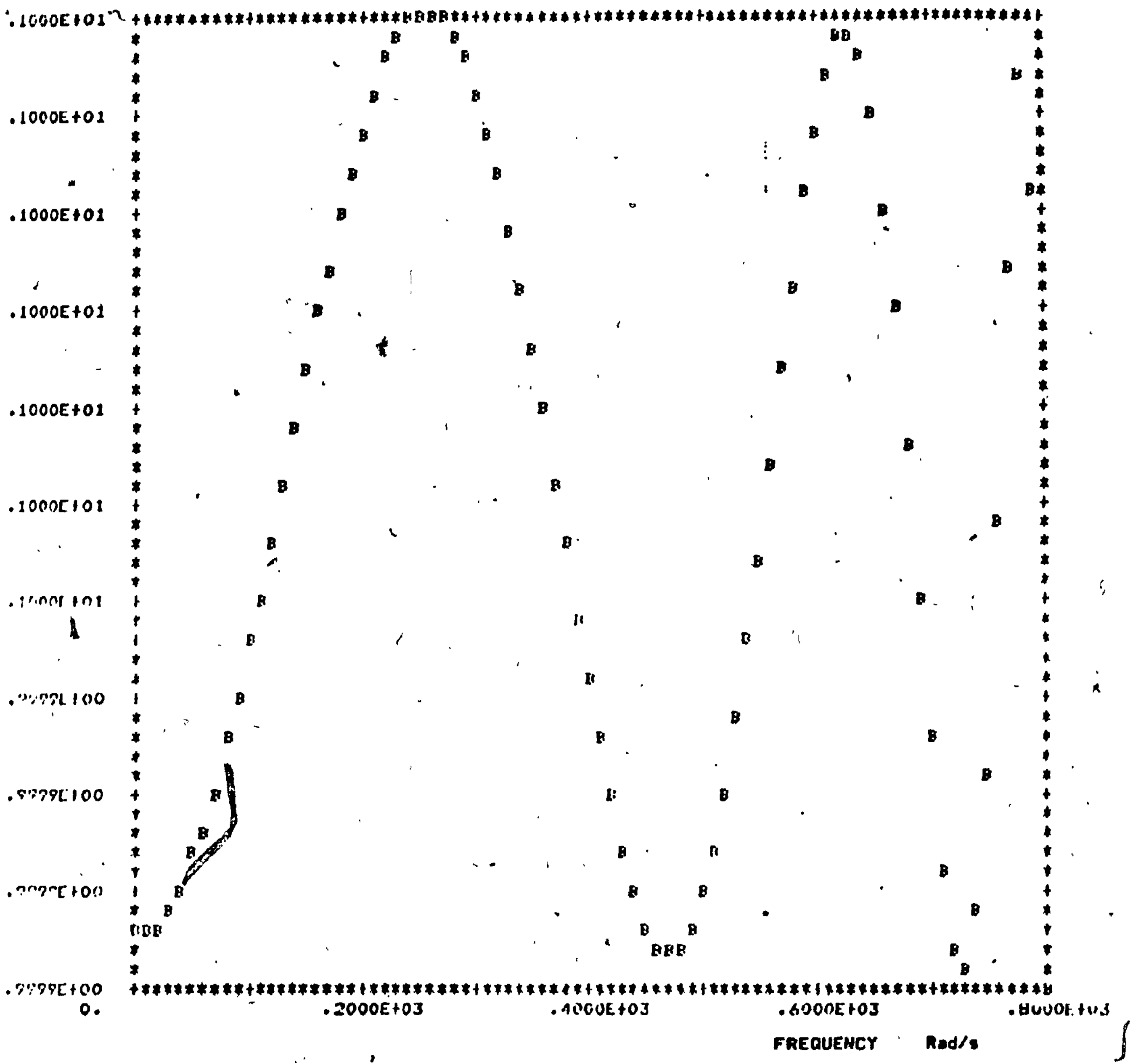


Figure 3.8(b) Amplitude response (passband)

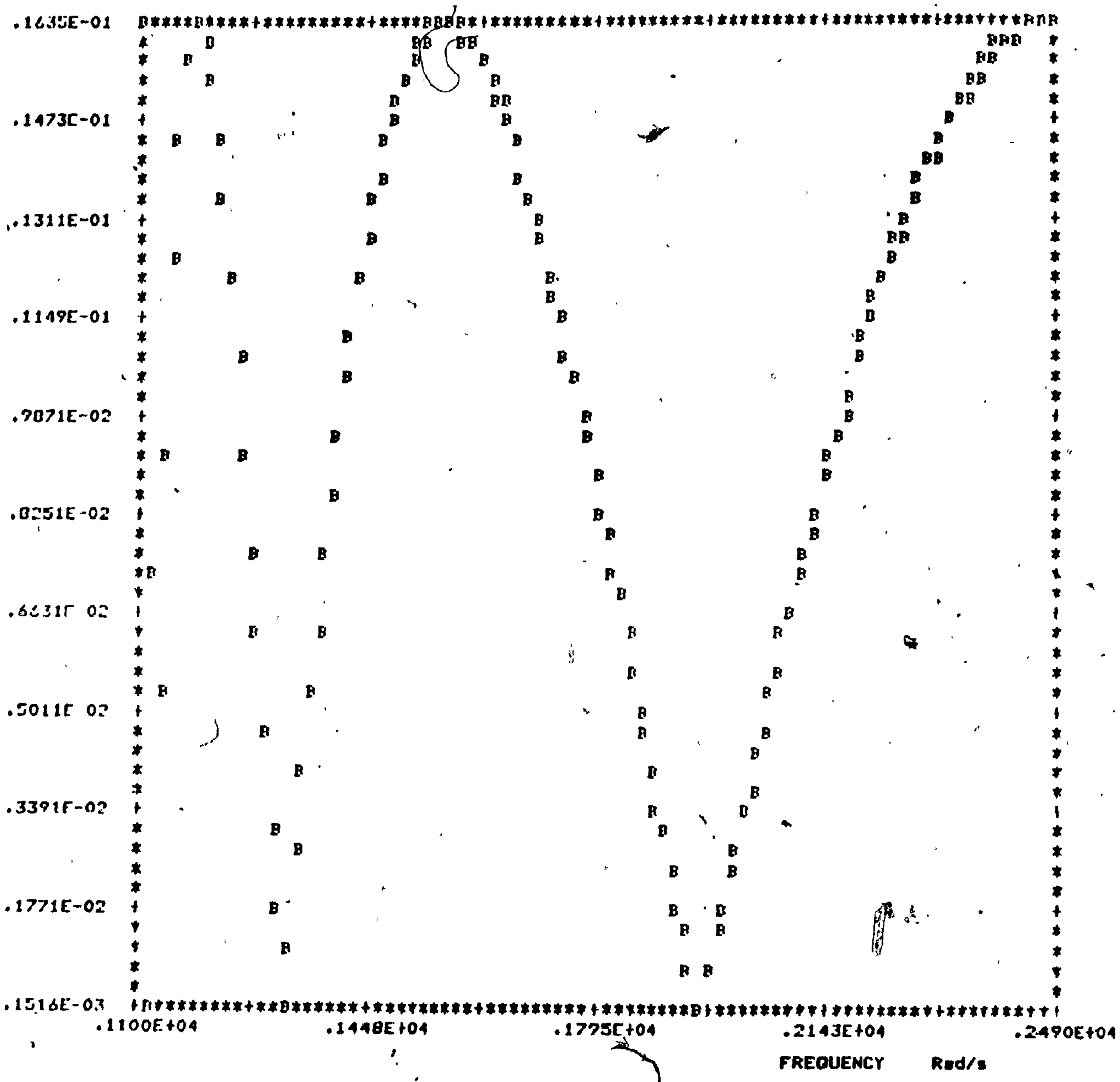


Figure 3.8(c) Amplitude response (stopband)

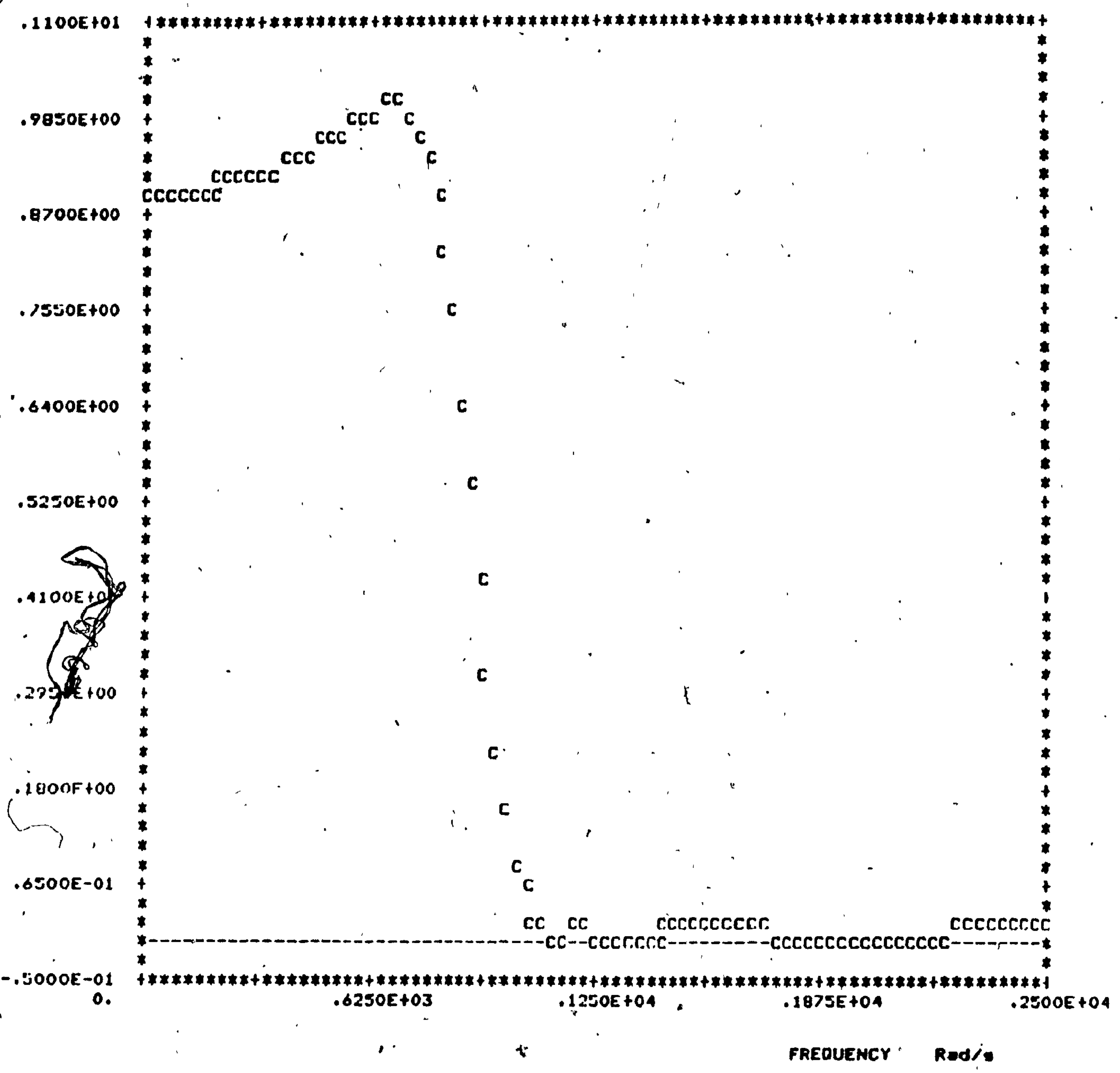


Figure 3.9(a) Overall amplitude response

Example: 7

Type: 3-section minimized-maximum pole

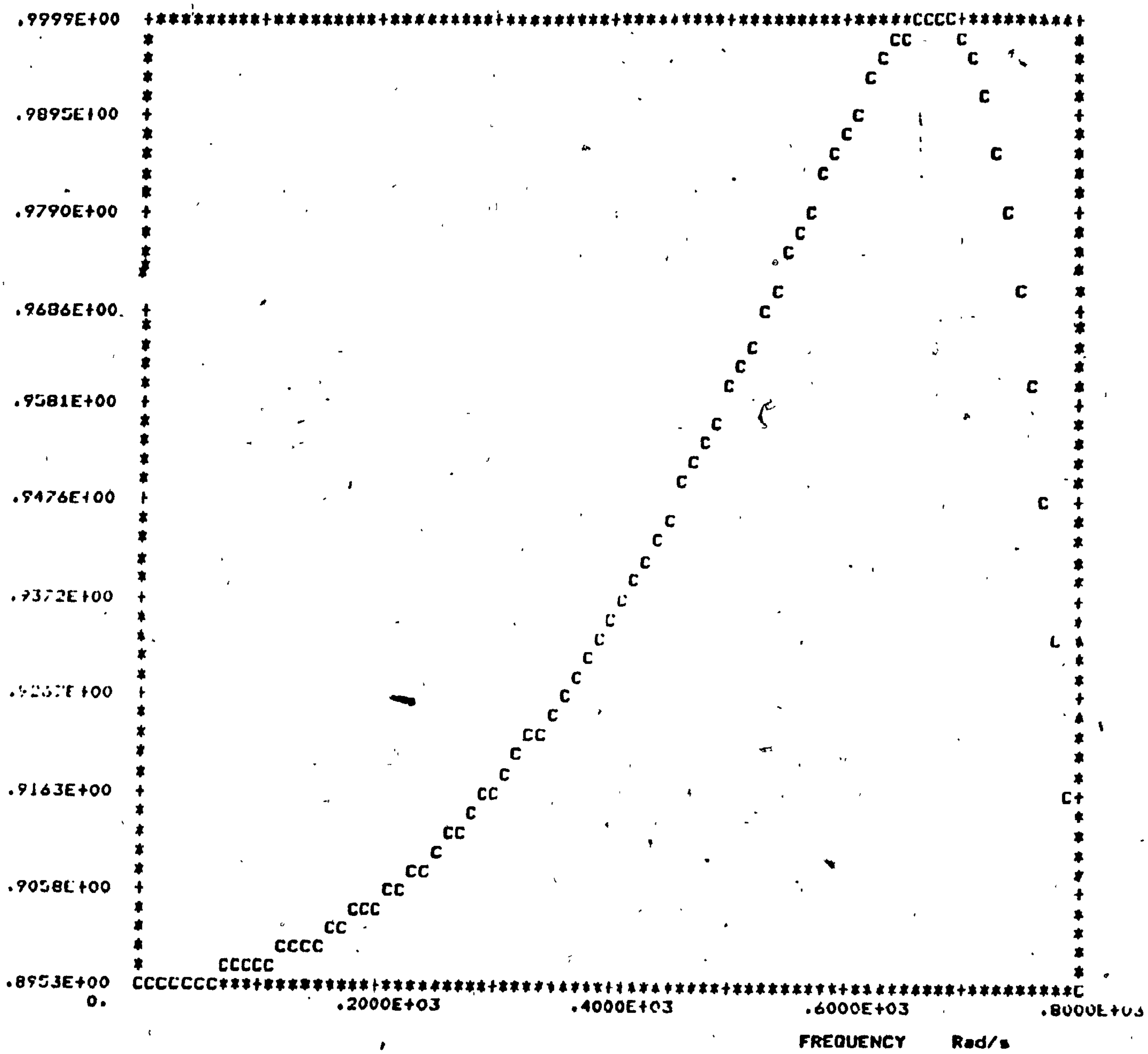


Figure 3.9(b) Amplitude response (passband)

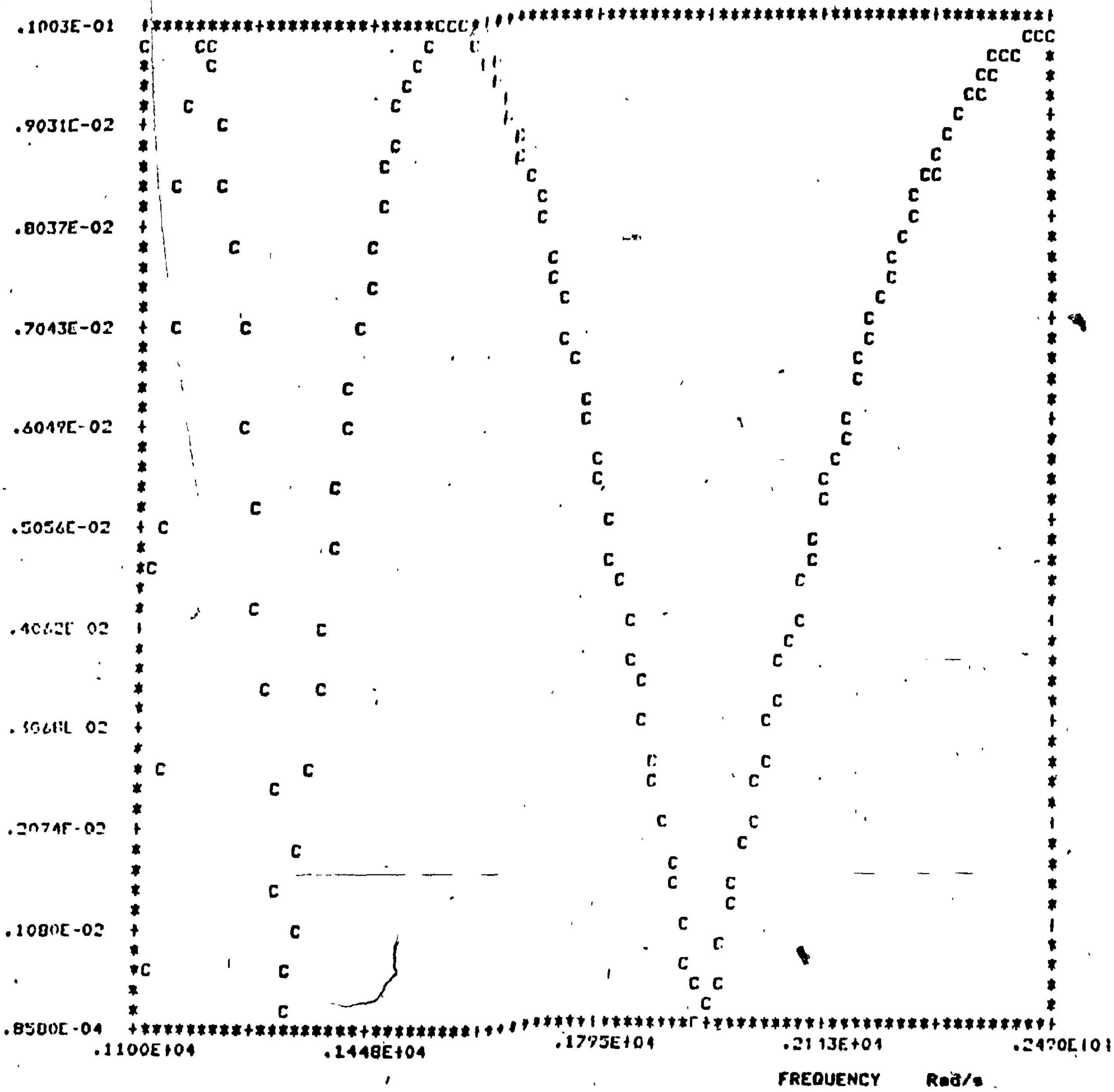


Figure 3.9(c) Amplitude response (stopband)

3.2 Location of Poles

The pole radii for each filter are tabulated in Table 3.3. The purpose of this Table is to illustrate the relative location of the poles and to compare the reduction in maximum-pole distance. From Table 3.3 the following conclusions may be drawn:

(i) in all 9 cases, reduction in the maximum-pole distance is achieved.

(ii) reduction in the maximum pole distance is inversely related to the selectivity of the filter.

(iii) differential pole distance is smallest in the case of minimized-pole-distance filters.

(iv) since all poles lie within a circle of radius r_o, one would expect the minimized-pole-distance filters to be less sensitive with respect to small variations in the transfer function coefficients.

Table 3.3 Pole radii for filters No. 1 to No. 9

Section No. 1	1	2	3
A	.50686	.8127	-
B	.33205	.52552	.823135
C	.46877	.46874	.46726

Section No. 2	1	2	3
A	.31585	.92466	-
B	.541507	.685157	.885149
C	.745022	.745108	.5811

Section No. 3	1	2	3
A	.9372	.9747	-
B	.825803	.881782	.959081
C	.831015	.908897	.908901

Section No. 4	1	2	3
A	.6343	.87179	-
B	.216013	.537293	.840894
C	.49289	.49287	.49287

Table 3.3 Pole radii for filters No. 1 to No. 9

Section No.5	1	2	3
A	.86865	.95611	-
B	.69201	.8260	.947874
C	.69577	.87682	.87682

Section No.6	1	2	3
A	.9562	.9853	-
B	.885229	.936857	.981681
C	.894789	.95541	.95541

Section No.7	1	2	3
A	.6849	.9152	-
B	.30003	.668269	.906711
C	.331218	.776875	.776875

Section No.8	1	2	3
A	.9060	.9709	-
B	.741408	.879836	.970274
C	.737681	.922885	.9228005

Table 3.3 Pole radii for filters No. 1 to No. 9

Section No.9	1	2	3
A	.9633	.9904	-
B	.907117	.958437	.988919
C	.90944	.973816	.97382

A = minimum-order elliptic

B = minimum-ripple ellptic

C = minimized-pole-distance filter

CHAPTER IV

COMPARISON OF FILTERS

4.0 Introduction

In this Chapter the effect of coefficient and product quantization in various designs will be explored and the three types of realizations will be compared. Cascade canonic implementation, fixed-point arithmetic, two's complement number representation, and quantization by rounding will be assumed. The approach used is as follows:

(i) The three different designs for each set of specifications were scaled using Jackson's signal scaling technique [6],[18] considering all possible section sequences in each case.

(ii) The signal scaling determined in (i) was incorporated in each design and the output noise power spectral density (PSD) was computed for all possible section sequences. The optimum section sequence, namely the one yielding the lowest inband-noise power, was chosen for each case.

(iii) The three optimum designs for each set of specifications, properly scaled and connected were subjected to coefficient quantization. The resulting passband ripple $\tilde{A}_p$ and a minimum stopband loss $\tilde{A}_a$ were then computed for wordlengths in the range 7 to 20 bits (excluding the sign bit). The minimum wordlength such that

$$\tilde{A}_p \leq \max(A_p) \qquad \tilde{A}_a \geq \min(A_a)$$

were determined for each case.

(iv) The analyses in step (iii) were repeated using floating-point arthmetic but without signal scaling.

4.1 Signal Scaling

If the amplitude of any internal signal in a fixed-point implementation is allowed to exceed the dynamic range, overflow will occur and the output signal will be severly distorted. On the other hand if signal amplitudes throughout the filter are unduly low, the filter will be operating inefficiently and the signal-to-noise ratio will be poor. Therefore, for optimum filter performance suitable signal scaling must be employed to adjust various signal levels.

A scaling technique to one's or two's complement implementations was developed by Jackson. In this technique a scaling multiplier λ_i is used at the input of each filter section H_i, so that amplitudes of multiplier inputs are bounded when the input of the filter is bounded. Under these circumstances, adder outputs are also bounded and overflow can not occur.

In order to obtain the values of the scaling multipliers, λ_i, one method would be to employ the L_p-norm notation. The L_p-norm of an arbitrary periodic function $A(\omega)$ with a period ω_0 is defined as

$$\|A\|_p = [\frac{1}{\omega_0} \int_0^{\omega_0} |A(\omega)|^p d\omega]^{1/p} \qquad \ldots \quad 4.1$$

If $A(\omega)$ is a continuous function, then one can show that

$$\lim_{p\to\infty} \|A\|_p = \|A\|_\infty = \max_{0\le\omega\le\omega_0} |A(\omega)| .$$

Using the properties of $A(\omega)$, (periodicity and continuity), and the L_∞-norm notation Jackson proved that for values of scaling constants

$$\lambda_i \le \frac{1}{\|H_i\|_\infty}$$

multiplier inputs would be bounded and overflow will not occur. In the case of the parallel or cascade realization efficient scaling can be accomplished using one multiplier per section. The arrangement of the scaling multipliers for the cascade realization is illustrated in Figure 4.1. The scaling multipliers are given by

$$\lambda_i = \frac{1}{Q_0[\max\{\|\prod_{n=1}^{i+1} H_n\|_\infty,\ H'_{i+1}Q_1\|_\infty\}]} \qquad \ldots\ 4.2$$

where

$$Q_0 = \begin{cases} 1 & \text{for } i=0 \\ \prod_{n=0}^{i-1} \lambda_n & \text{for } i\ge 1 \end{cases}$$

and

$$Q_1 = \begin{cases} 1 & \text{for } i=0 \\ \prod_{n=1}^{i} H_n & \text{for } i\ge 1 \end{cases}$$

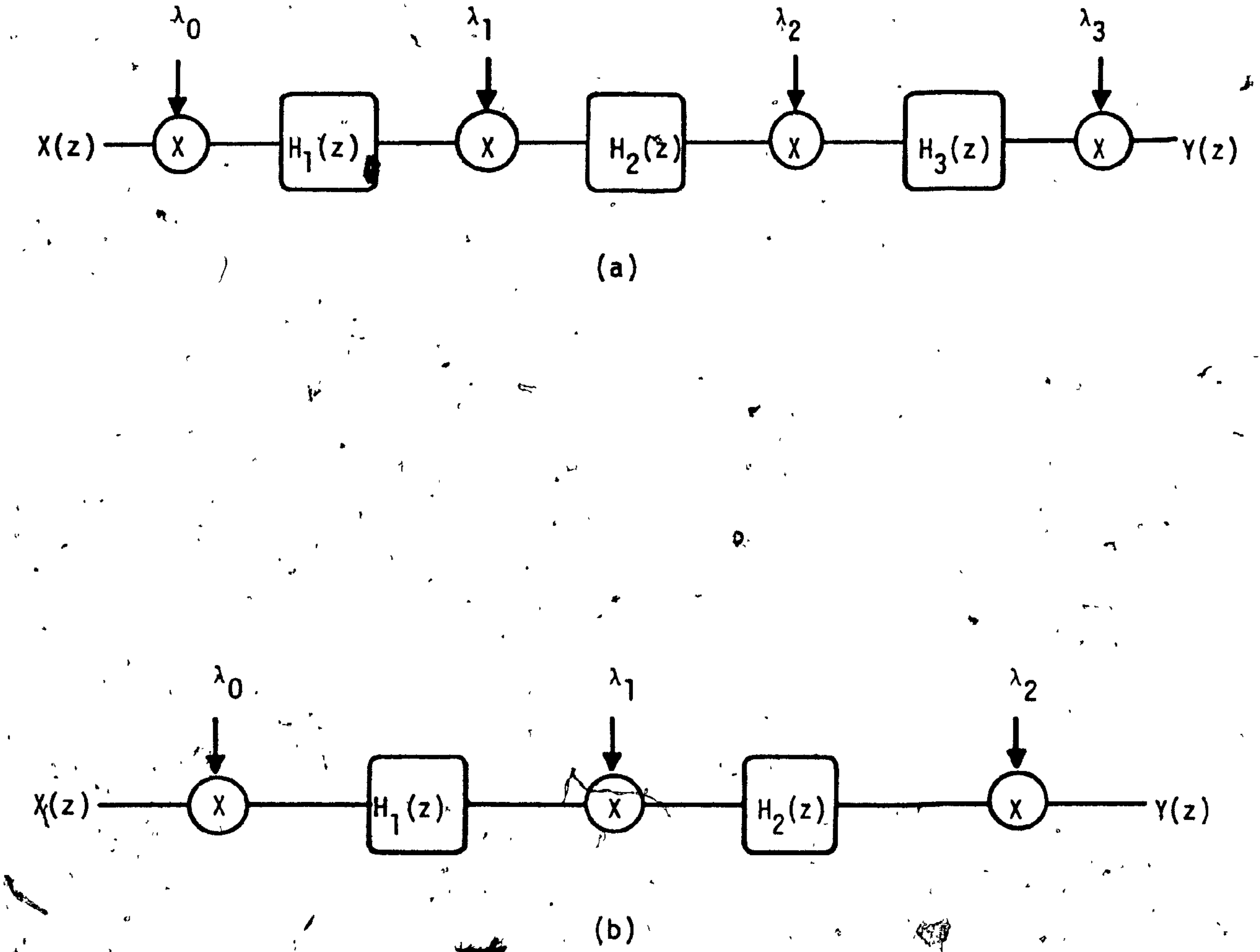


Figure 4.1 Scaling of cascade realization
(a) 3 sections
(b) 2 sections

H_n and H_n' represent the magnitudes of the overall and partial transfer functions of the nth section respectively.

According to Eqn. 4.2 the scaling constants for each section of a two-section cascade realization can be obtained as follows:

$$\lambda_0 = \frac{1}{\max[\|H_1\|_\infty, \|H_1'\|_\infty]}$$

$$\lambda_1 = \frac{1}{\lambda_0 \max[\|H_1H_2\|_\infty, \|H_1H_2'\|_\infty]}$$

$$\lambda_2 = \frac{1}{\lambda_0\lambda_1 \max[\|H_1H_2\|_\infty]}$$

For a three-section filter we would have

$$\lambda_0 = \frac{1}{\max[\|H_1\|_\infty, \|H_1'\|_\infty]}$$

$$\lambda_1 = \frac{1}{\lambda_0 \max[\|H_1H_2\|_\infty, \|H_1H_2'\|_\infty]}$$

$$\lambda_2 = \frac{1}{\lambda_0\lambda_1 \max[\|H_1H_2H_3\|_\infty, \|H_1H_2H_3'\|_\infty]}$$

$$\lambda_3 = \frac{1}{\lambda_0\lambda_1\lambda_2 \max[\|H_1H_2H_3\|_\infty]}$$

4.2 Product Quantization

The output of a finite-wordlength multiplier can be expressed as

$$Q = mx(n) + e(n)$$

where

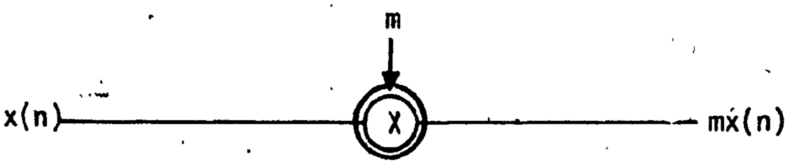
$x(n)$ = multiplier input

m = multiplier coefficient

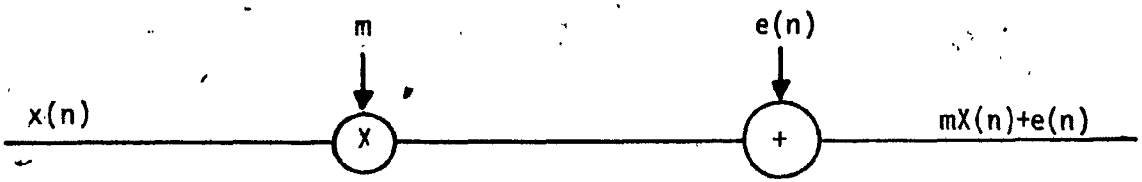
$e(n)$ = quantization noise

If quantization is by rounding, $e(n)$ is a random process with uniform probability density (white noise).

The nonideal multiplier of Fig. 4.2(a) can thus be modeled in terms of an ideal multplier in series with a noise source as shown in Fig. 4.2 (b). Under reasonable assumptions one can show that in a filter, the noise generated by individual multiplier elements are statistically independent and add up in linear fashion. Based on the model of Fig. 4.2(b), a canonic configuration can be represented as in Fig. 4.3, where each multiplier is accompanied by a noise source $e(n)$. Since the noise sources in the canonic section are statistically independent random processes, the power spectral density (PSD) of a sum of several processes is the sum of their respective power spectral densities. If H_i is the transfer function of the ith section of the filter, then the PSD of that section is given by



(a)



(b)

Figure 4.2 (a) A multiplier element

(b) Model assumed for the multiplier element in (a)

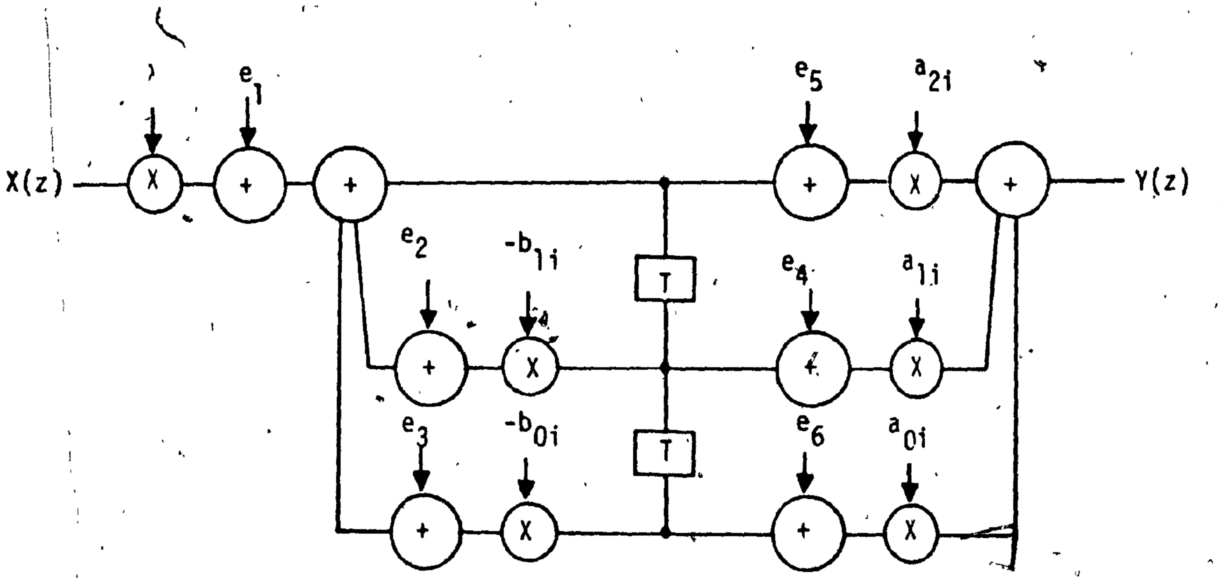


Figure 4.3 Realistic model of a canonical section including the noise sources

$$P_i = H_i(z)H_i(z^{-1}) \sum_{j=1}^{3} S_{ej}(z) + \sum_{j=4}^{6} S_{ej}(z)$$

where

$$S_{ej}(z) = \frac{2^{-L}}{12} = \frac{q^2}{12}$$

L = wordlength of register

q = quantization step

In canonic second-order sections, which constitute the building blocks of our realizations, $a_{0i}=a_{2i}=1$ and hence the expression for the power spectral density is changed to

$$P_i = H_i(z)H_i(z^{-1}) \sum_{j=1}^{3} S_{ej}(z) + S_{e4}(z)$$

For a K-section filter, the overall filter power spectral density is therefore obtained as

$$PSD = \sum_{i=1}^{K} P_i$$

Clearly, the power spectral density function is frequency dependent. If ω_p is the passband edge, then the passband noise power can be computed as:

$$NP = \frac{1}{\omega_p} \int_0^{\omega_p} PSD(\omega)\ d\omega$$

For each filter, the value of noise power will be different for different possible section sequences. The optimum section sequence will be the arrangement of filter sections H_i such that the noise power is minimum. In case of two-section filters there are two possible section arrangements of which one arrangement will have the lowest noise power value. In the case of three-section filters there are six possible section arrangements of which one arrangement would provide the optimum inband noise power value. Table 4.1 depicts all possible section configurations for two-section and three-section filters. Based on the optimum value of noise power, proper scaling constants λ_i can be evaluated. The set of scaling coefficients λ_i^* for which the noise power is minimum, is called the optimal scaling coefficient set. The optimum sets of λ_i for the various types of filters and specifications are given in Table 4.2. The corresponding values of noise power are presented in Table 4.3. As can be seen, the designs with minimum passband ripple yield similar results as the designs with minimized maximum pole radius, and both designs lead to significant reduction in the inband noise power relative to that in the minimum-order elliptic design. The reduction in the passband average of the PSD that can be achieved is in the range of 1.7 to 12.2 dB. Power spectral density curves were plotted for each optimum-sequence design and the PSD was found to be approximately constant in each design. Figs. 4.4-4.6 illustrate such curves for 3 different examples.

Table 4.1 Section arrangements

	2-Section filter	
Configuration	Section	Section
A	1	2
B	2	1

	3-Section Filter		
Configuration	Section	Section	Section
A	1	2	3
B	1	3	2
C	2	1	3
D	2	3	1
E	3	1	2
F	3	2	1

Table 4.2 Scaling constants for filters No. 1 to No. 9

Scaling constant Filter No.1	λ_0	λ_1	λ_2	λ_3
A	.1015565	.248244	.248244	-
B	.5060145	.4308629	.670787	1.0
C	.4042439	.386584	.385956	1.0

Scaling constant Filter No.2	λ_0	λ_1	λ_2	λ_3
A	.037145	.037336	-	-
B	.1371215	.1510497	.1090255	1.0
C	.1739195	.1519232	.088522	1.0

Scaling constant Filter No.3	λ_0	λ_1	λ_2	λ_3
A	.0060937	.012354	.4726855	-
B	.0353152	.0590116	.1908546	1.0
C	.0309583	.0843908	.139233	1.0

Scaling constant Filter No.4	λ_0	λ_1	λ_2	λ_3
A	.10109	.20663	1.0	-
B	.438743	.3286003	.633999	1.0
C	.402738	.293543	.2888894	1.0

Table 4.2 Scaling constants for filters No. 1 to No. 9

Scaling constant Filter No.5	λ_0	λ_1	λ_2	λ_3
A	.0277023	.070237	.379612	-
B	.112889	.1270203	.095355	5.11658
C	.0886661	.188963	.104672	4.10115

Scaling constant Filter No.6	λ_0	λ_1	λ_2	λ_3
A	.004053	.038666	31.61463	-
B	.015861	.041622	.173484	42.96605
C	.0156771	.0331266	.2183175	34.3472

Scaling constant Filter No.7	λ_0	λ_1	λ_2	λ_3
A	.136914	.29431	1.0	-
B	.403864	.217275	.3725154	1.85527
C	.308005	.270713	.420897	1.195687

Scaling constant Filter No.8	λ_0	λ_1	λ_2	λ_3
A	.03325	.063867	.911769	-
B	.066924	.1079435	.174967	13.275364
C	.074245	.084074	.27762	9.896252

Table 4.2 Scaling constants for filters No. 1 to No. 9

Scaling constant Filter No.9	λ_0	λ_1	λ_2	λ_3
A	.0036425	.057586	80.437655	-
B	.011431	.063968	.175635	121.4572
C	.011692	.045416	.264296	90.370699

A = minimum order elliptic filter

B = minimum ripple elliptic filter

C = minimized pole distance filter

Table 4.3 Optimum section configurations for filters No. 1 to No. 9

Filter type / Filter No.	Filter A		Filter B		Filter C	
	Seq.	N_p dB	Seq.	N_p dB	Seq.	N_p dB
1	A	31.16	C	19.02	A	23.15
2	B	48.09	D	40.63	F	39.76
3	A	61.38	D	54.76	D	54.12
4	D	32.49	C	21.43	E	25.81
5	A	41.52	C	37.67	C	37.86
6	A	51.58	A	46.14	A	46.87
7	B	28.96	A	23.9	C	25.49
8	A	36.47	A	34.75	A	33.74
9	A	46.61	A	42.94	A	43.15

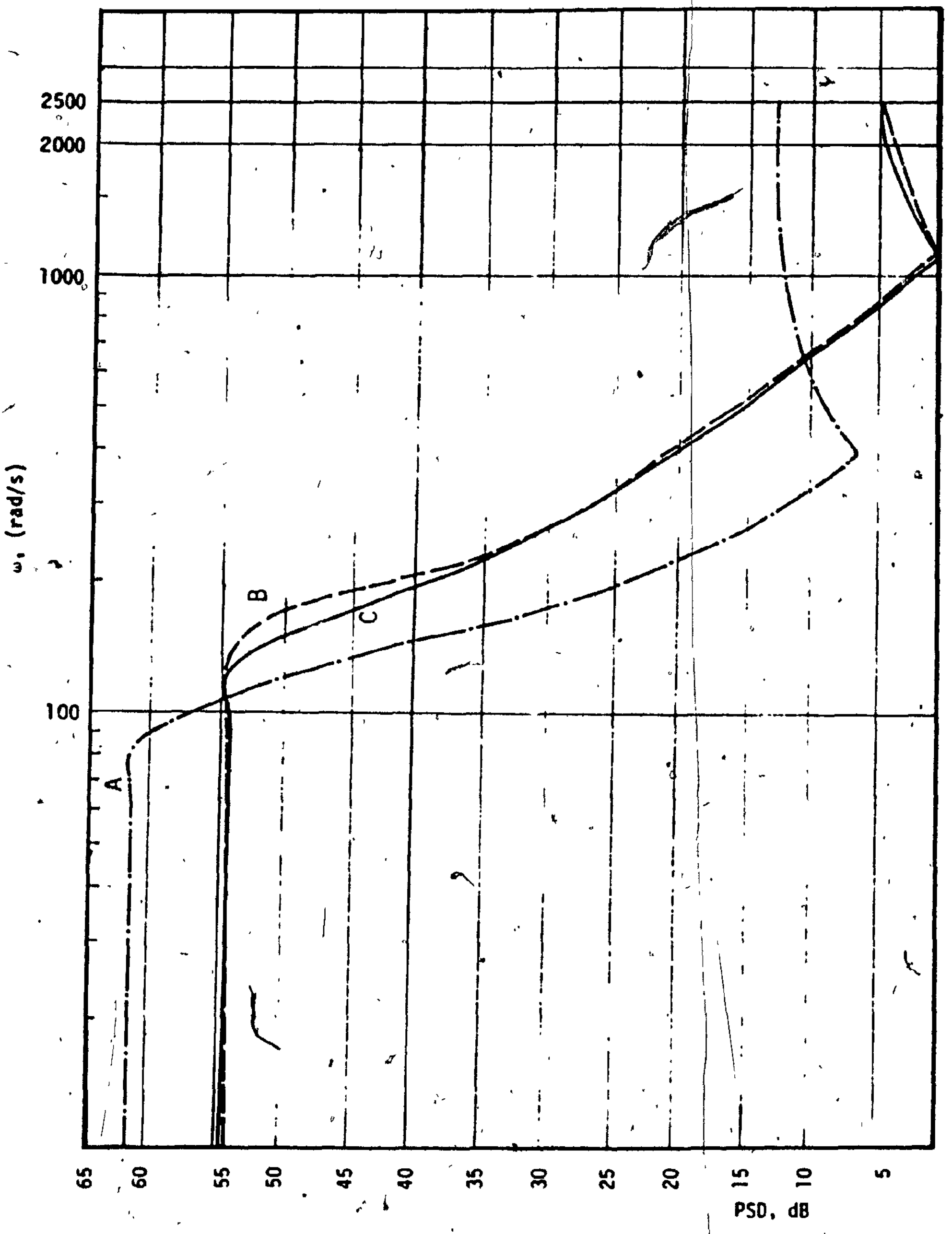


Figure 4.4 PSD versus frequency filter No. 3

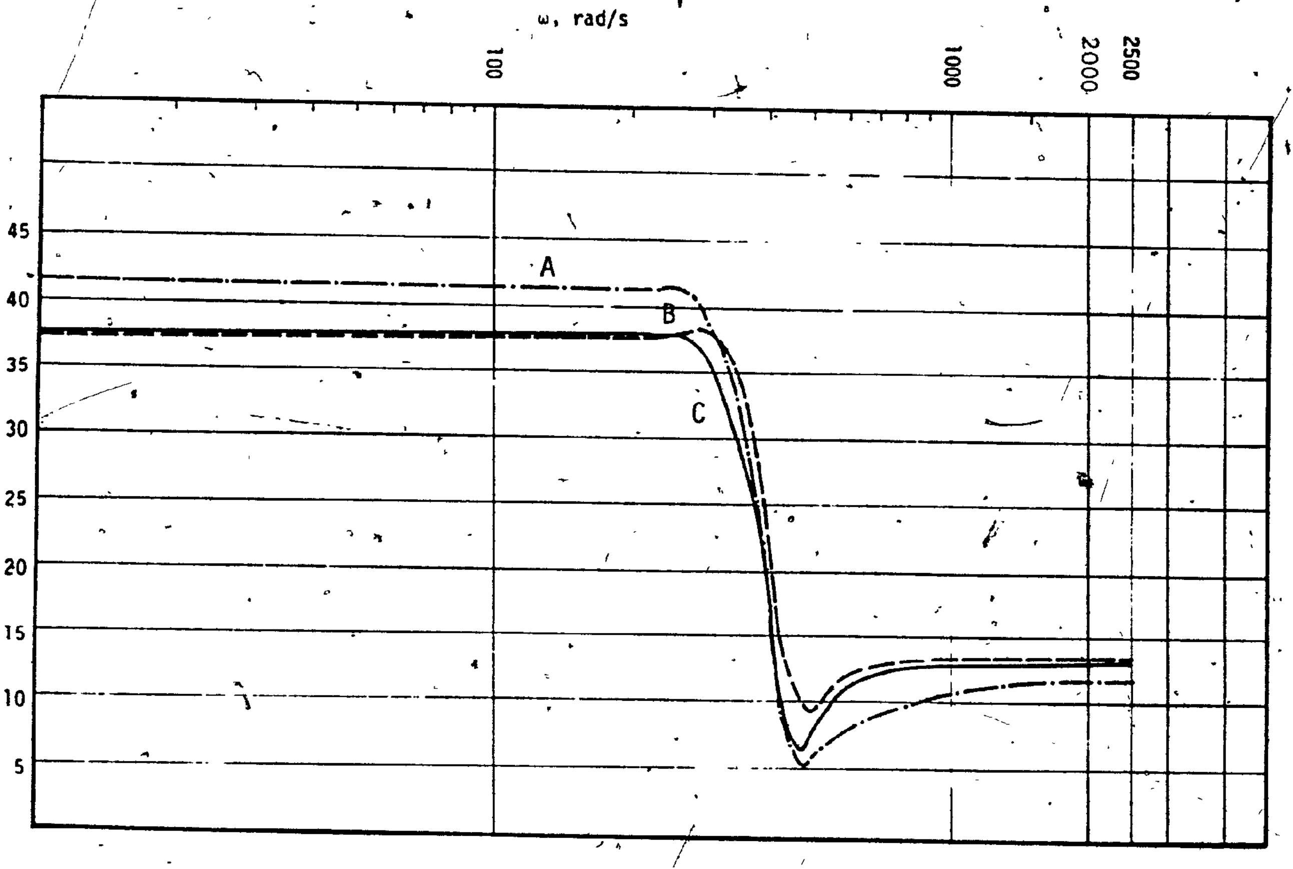


Figure 4.5 PSD versus frequency filter No. 5

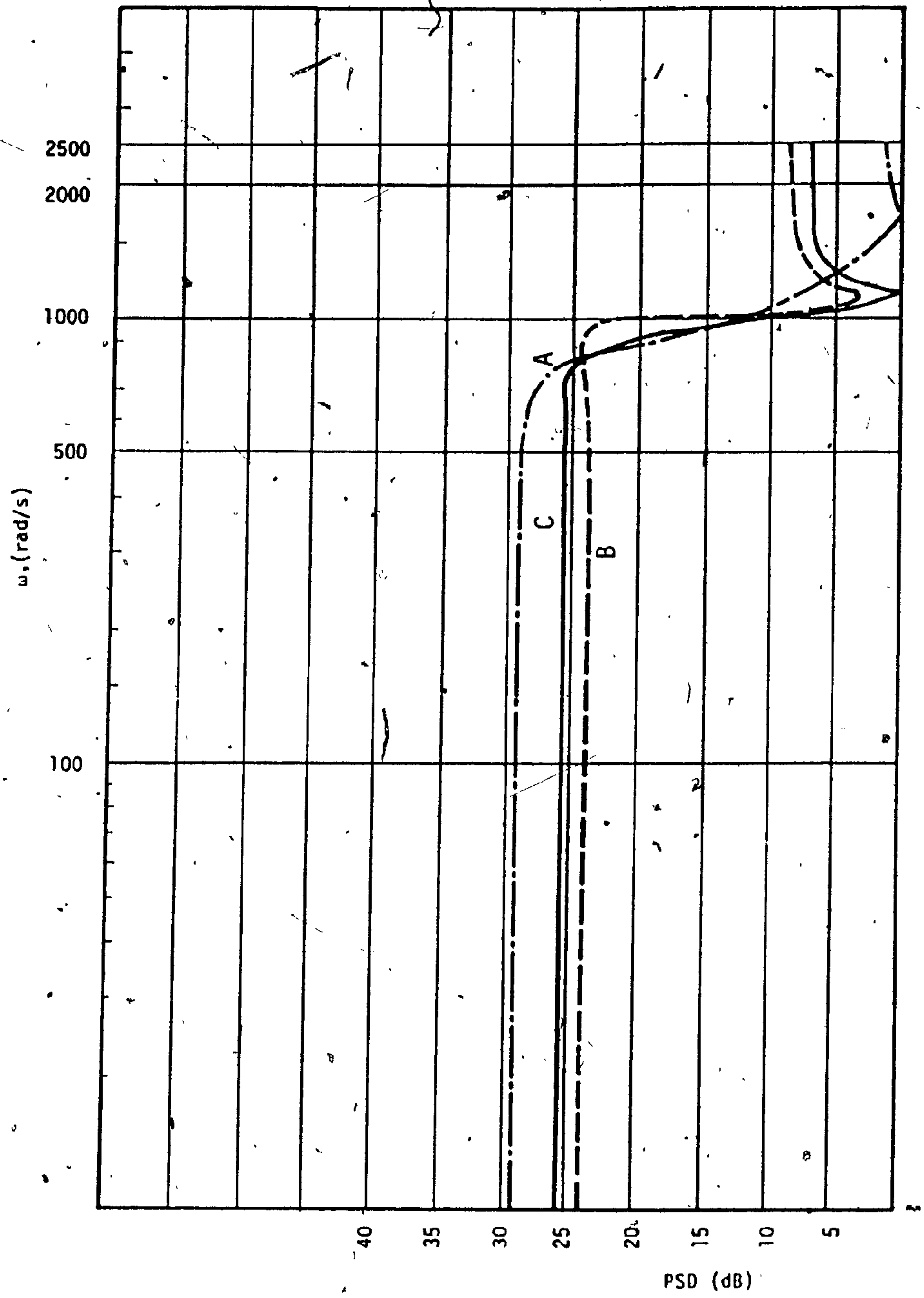


Figure 4.6 PSD versus frequency filter No. 7

4.3 Coefficient Quantization

In software as well as hardware digital-filter implementations, numbers are ultimately stored in finite-length registers. Consequently, coefficients and signal values must be quantized by rounding or truncation before they can be stored. Coefficients of the transfer function are normally evaluated to a high degree of accuracy during the approximation step. If they are quantized, errors will arise. Errors in coefficient quantization introduce perturbations in the zeros and poles of the transfer function which in turn manifest themselves as errors in the frequency response, that is the passband ripple A_p will increase and the stopband loss A_a will be reduced. Depending on the length of the register, errors in coefficient quantization can cause the filter to violate the desired set of specification. Sensitivity of the filter to ceofficient quantization is particularly important when the poles in the z-plane are closed to the unit circle. In such cases, coefficient quantization can move the poles on or outside the unit circle which in turn will lead to unstable filter.

The hardware implementation of digital filters, like the implementation of any other digital hardware, is based on the binary number representation which can be of the fixed-point or floating-point type. In the fixed-point arithmetic, as the name implies, the true binary point occupies a specific physical position in the register where as no specific physical position of the register is assigned to the true binary point in the case of the floating-point arithmetic.

In fixed-point arithmetic, numbers are usually assumed to be proper fractions and the binary point is usually set between the first and second bit positions of the register. The first position is reserved for the sign of the number. Depending on the representation of negative numbers, fixed-point arithmetic can assume three forms:

1. Signed magnitude
2. One's complement
3. Two's complement

Of the above three forms, two's complement arithmetic is widely used in fixed-point implementation of digital filters. In floating-point arithmetic, numbers are stored in registers which are subdivided into two segments, one for the signed mantissa and the other for the signed exponent.

Fixed-point and floating-point arithmetics have their own merits and demerits. In the case of fixed-point arithmetic the range of numbers that can be handled is small and the percentage error produced by truncation or rounding tends to increase as the magnitude of the number is decreased. Floating-point arithmetic alleviates these problems to a large extent, however, implementation of floating-point arithmetic increases the hardware cost and reduces the speed of processing.

In order to examine the effects of coefficient quantization on each design, coefficients of Tables 3.2 were used. Both fixed-point and floating-point arithmetic were considered and coefficient quantization was assumed to be by rounding. Variations in the values of the passband ripple A_p and the stopband loss A_a with respect to different wordlengths were studied. The minimum wordlength required to satisfy the prescribed specifications are given in Table 4.4 for the fixed-point arithmetic and in Table 4.5 for the floating-point arithmetic. In both

Table 4.4 Required wordlength (fixed-point arithmetic)

EXAMPLE	A	B	C
1	11	7	9
2	15	7	7
3	13	10	10
4	9	7	7
5	13	8	8
6	15	9	11
7	8	7	7
8	13	9	9
9	13	10	10

Table 4.5 Required wordlength (floating-point arithmetic)

EXAMPLE	A	B	C
1	11	7	6
2	12	6	7
3	14	9	9
4	9	5	5
5	11	7	7
6	12	9	10
7	13	8	8
8	8	6	6
9	14	11	10

cases, designs with minimized passband ripple yield similar results as the designs with minimized-maximum pole radius. Both designs lead to significant reduction in the wordlength relative to that in the minimum-order design. Reduction in the wordlength ranges from 1 to 8 bits for the case of fixed-point arithmetic and from 2 to 6 bits for the case of floating-point arithmetic depending on the design. Figs. 4.7-4.13 illustrate the variation of passband ripple A_p and stopband loss A_a with respect to the wordlength for examples 3, 5 and 7. The relative insensitivity of minimized-maximum-pole-distance filters is to be noted.

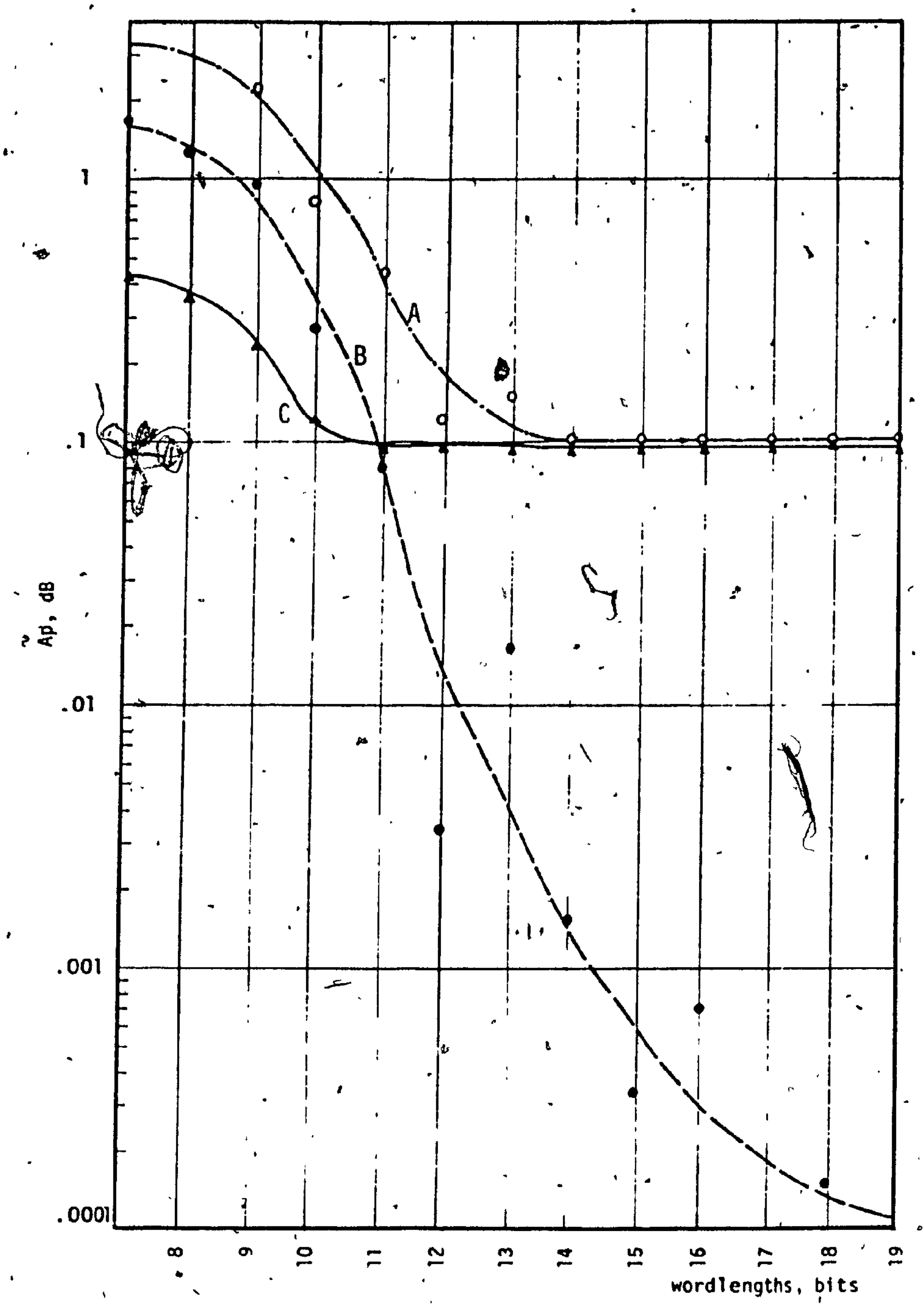


Figure 4.7 $\tilde{A}_p$ versus wordlenqth (fixed-point arithmetic) filter No. 3

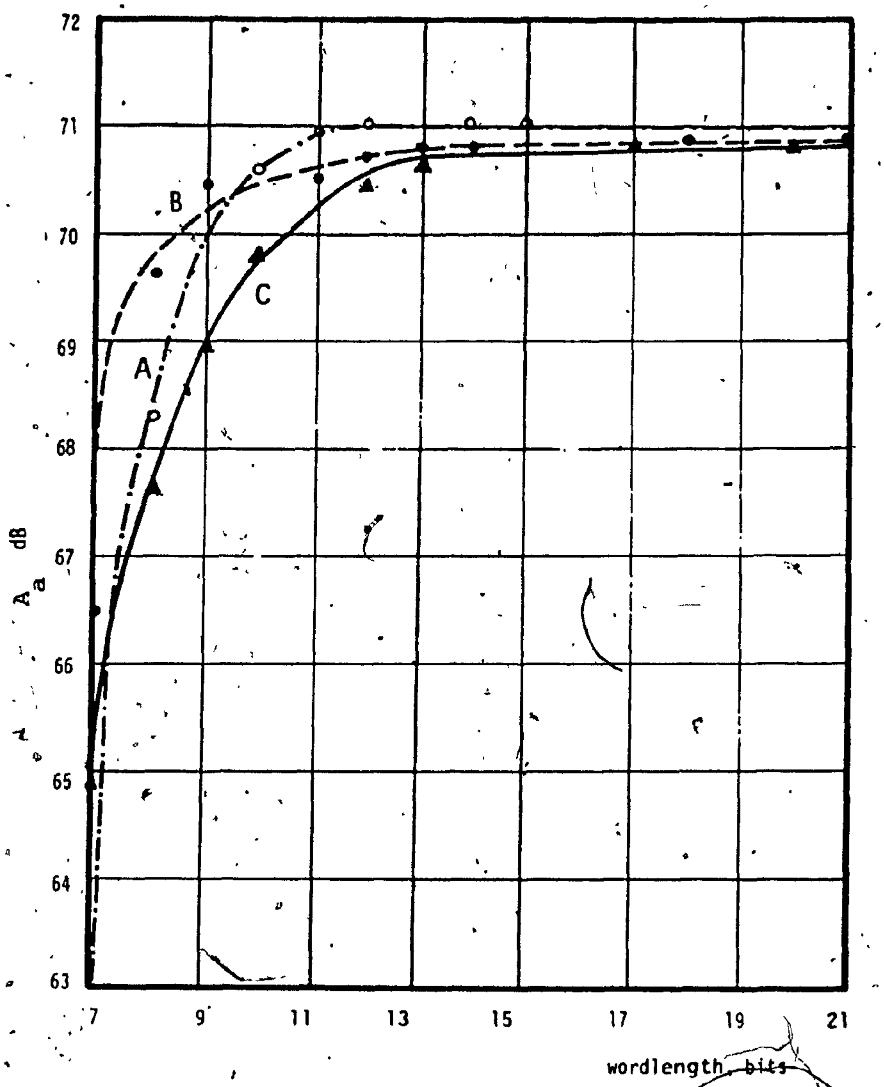


Figure 4.8 $\tilde{A}_a$ versus wordlength (fixed-point airthmetic) filter No. 3

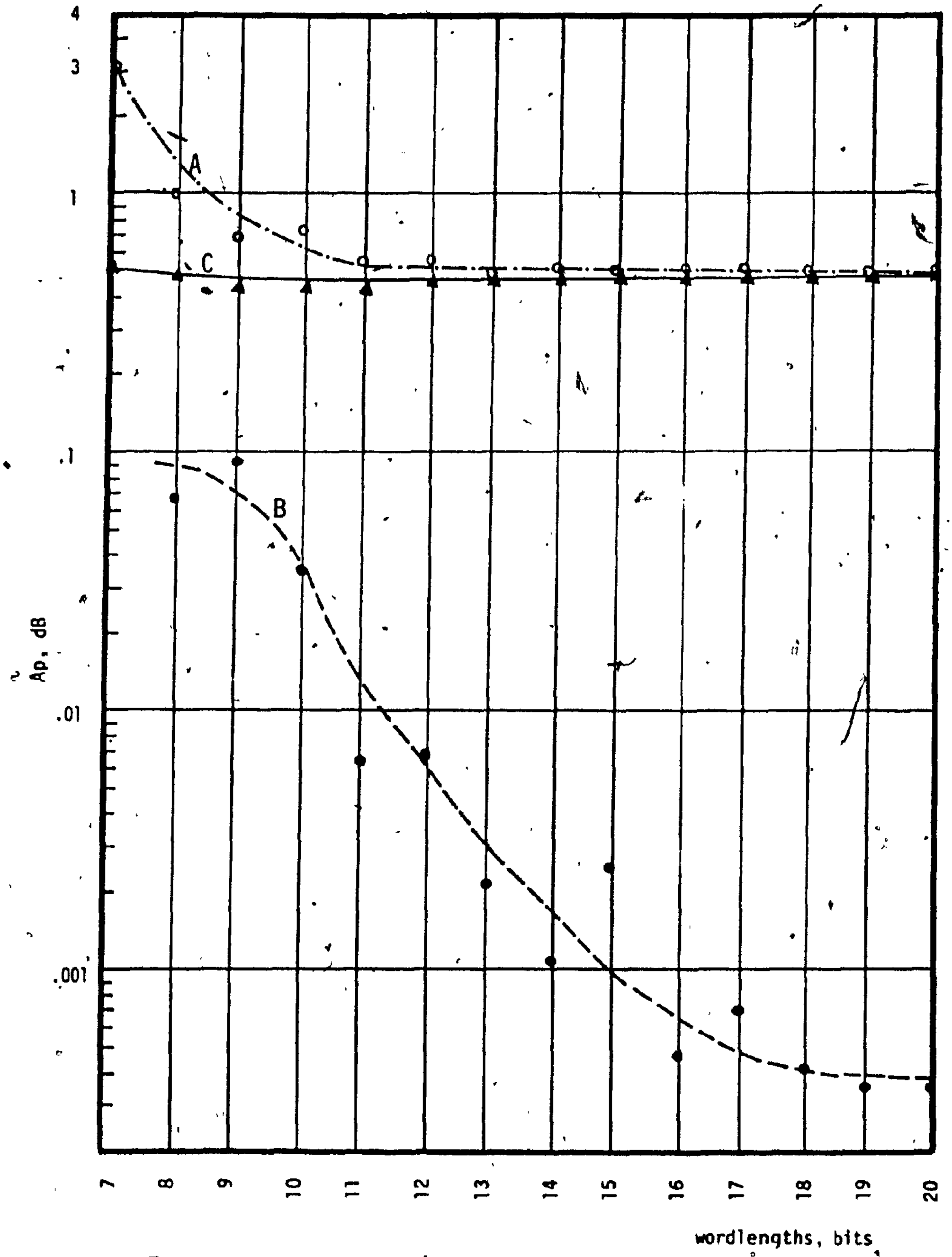


Figure 4.9 A_p versus wordlength (fixed-point arithmetic filter No. 5

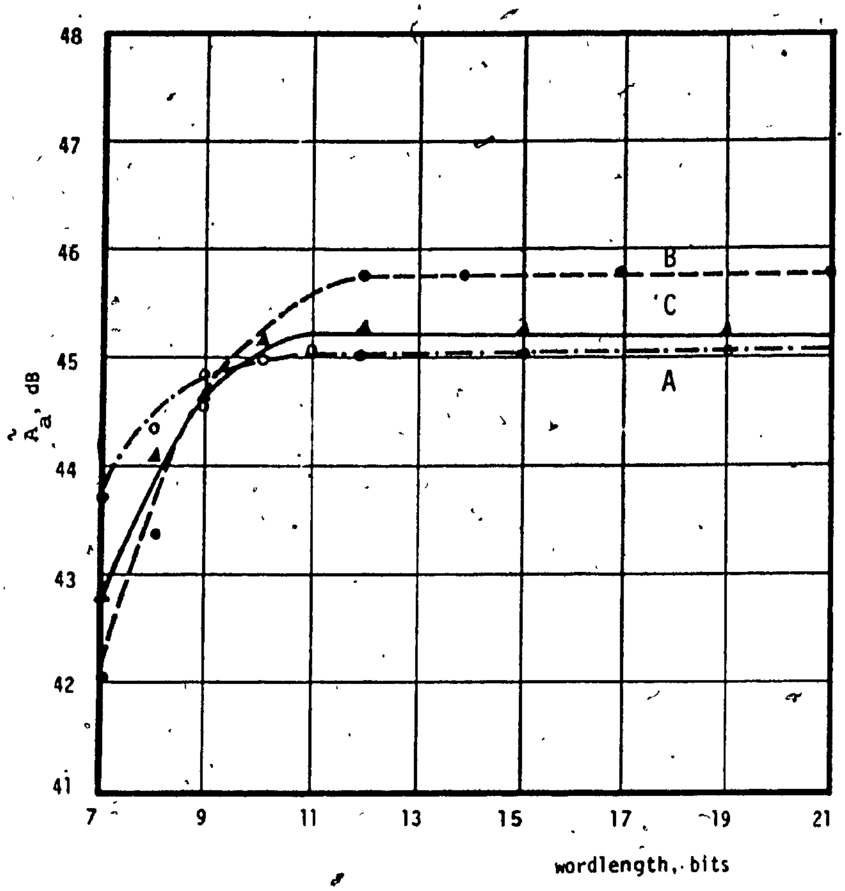


Figure 4.10 $\tilde{A}_a$ versus wordlength (fixed-point arithmetic) filter No. 5

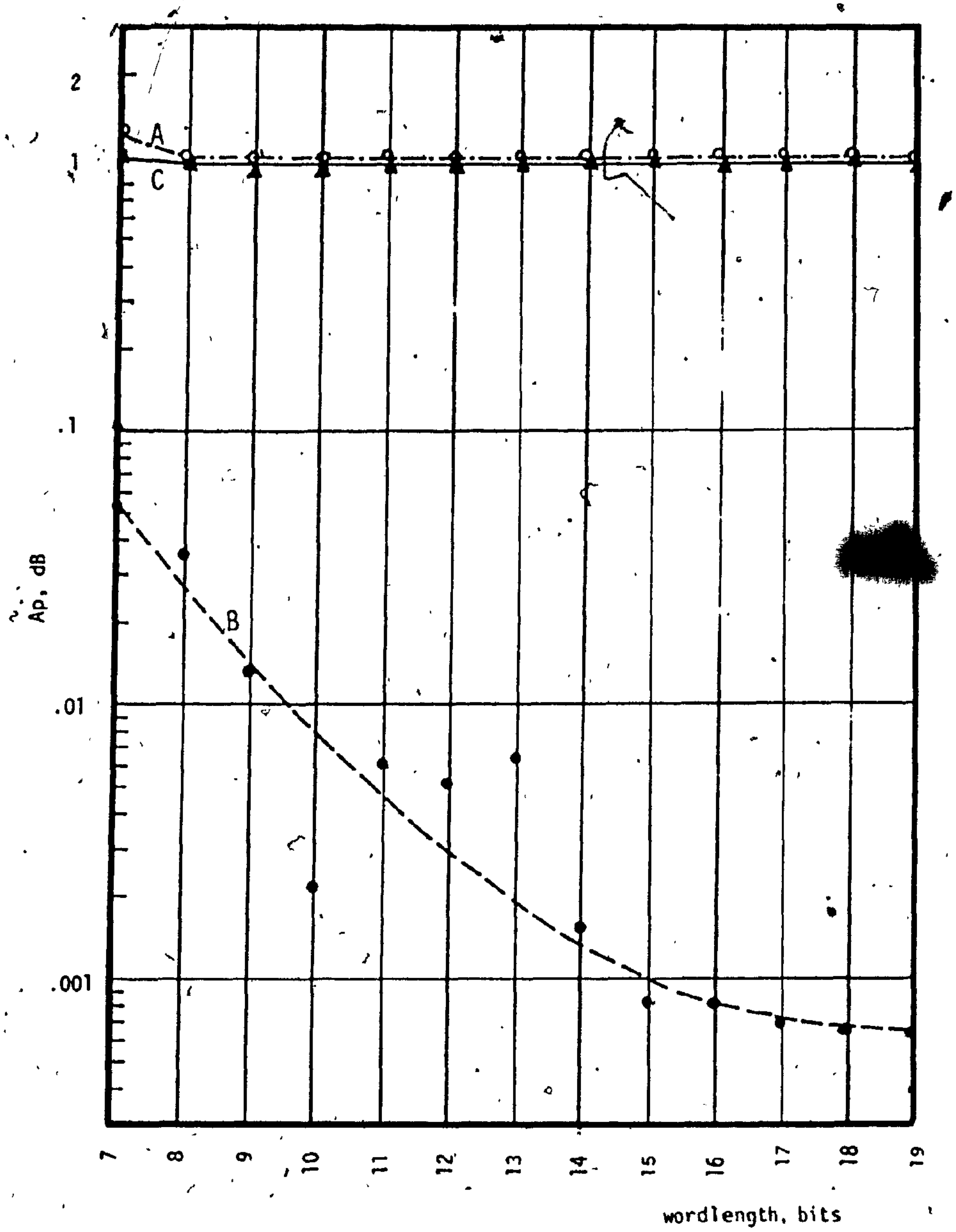


Figure 4.11 $\tilde{A}_p$ versus wordlength (fixed-point arithmetic) filter No. 7

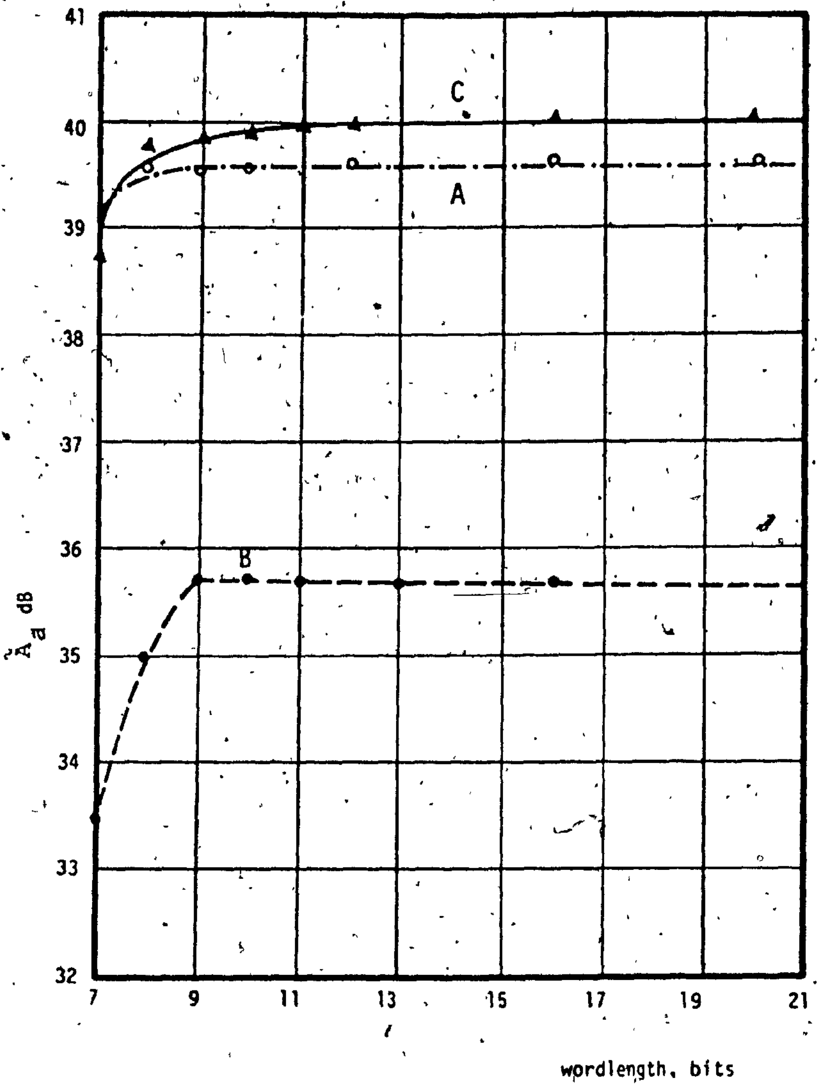


Figure 4.12 $\tilde{A}_a$ versus wordlength (fixed-point arithmetic) filter No. 7

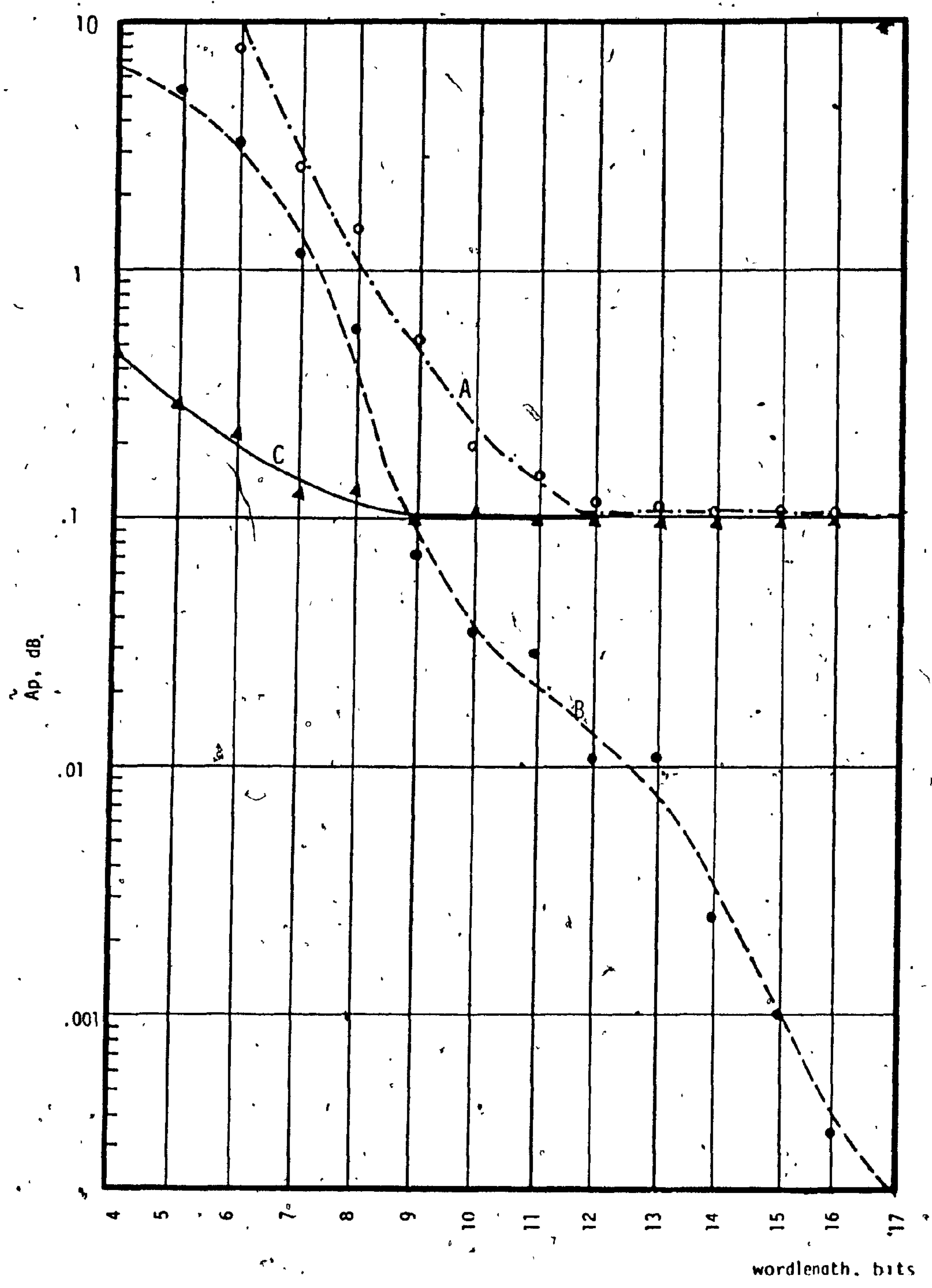


Figure 4.13 $\tilde{A}_p$ versus wordlength (floating-point arithmetic) filter No. 3

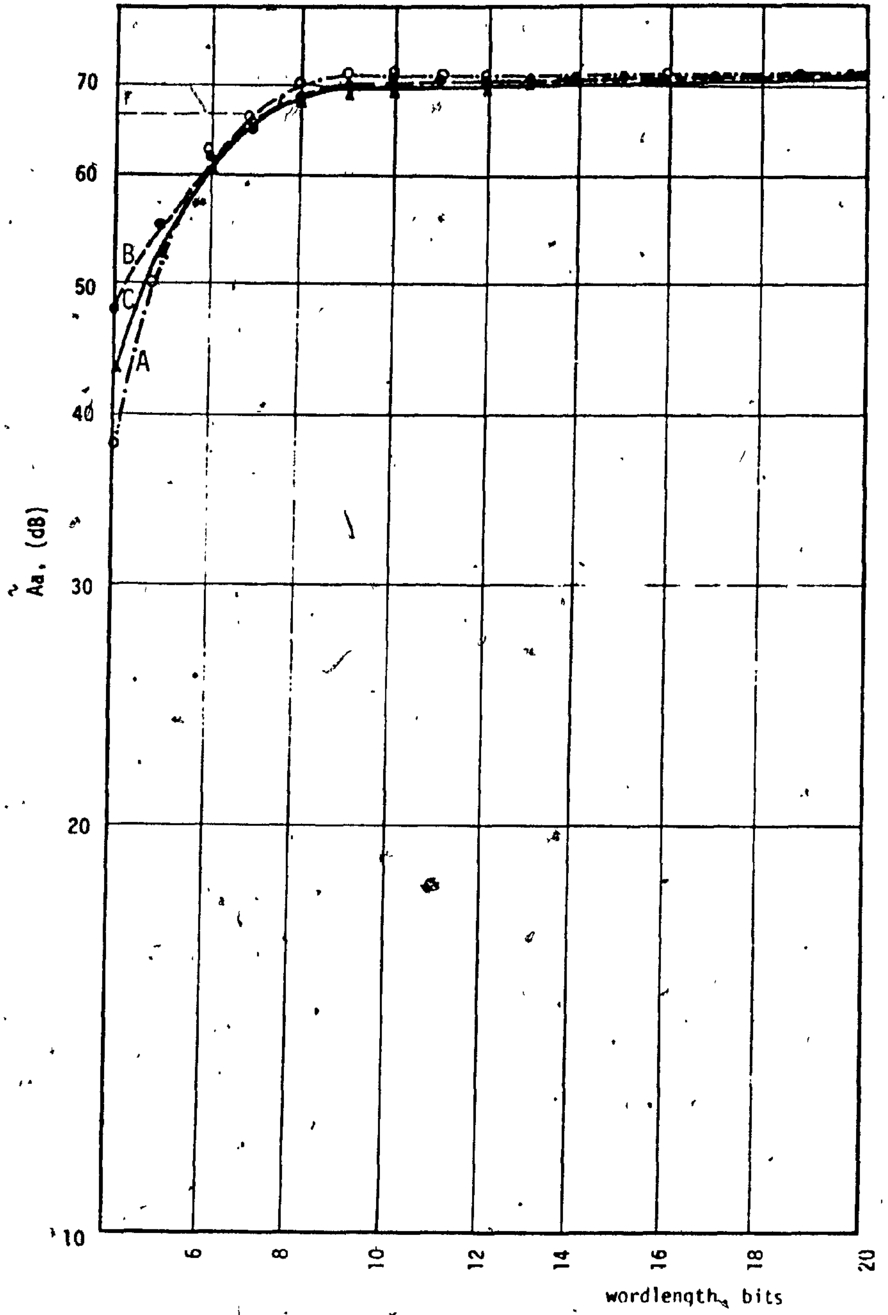


Figure 4.14 $\tilde{A}_a$ versus wordlength (floating-point arithmetic)

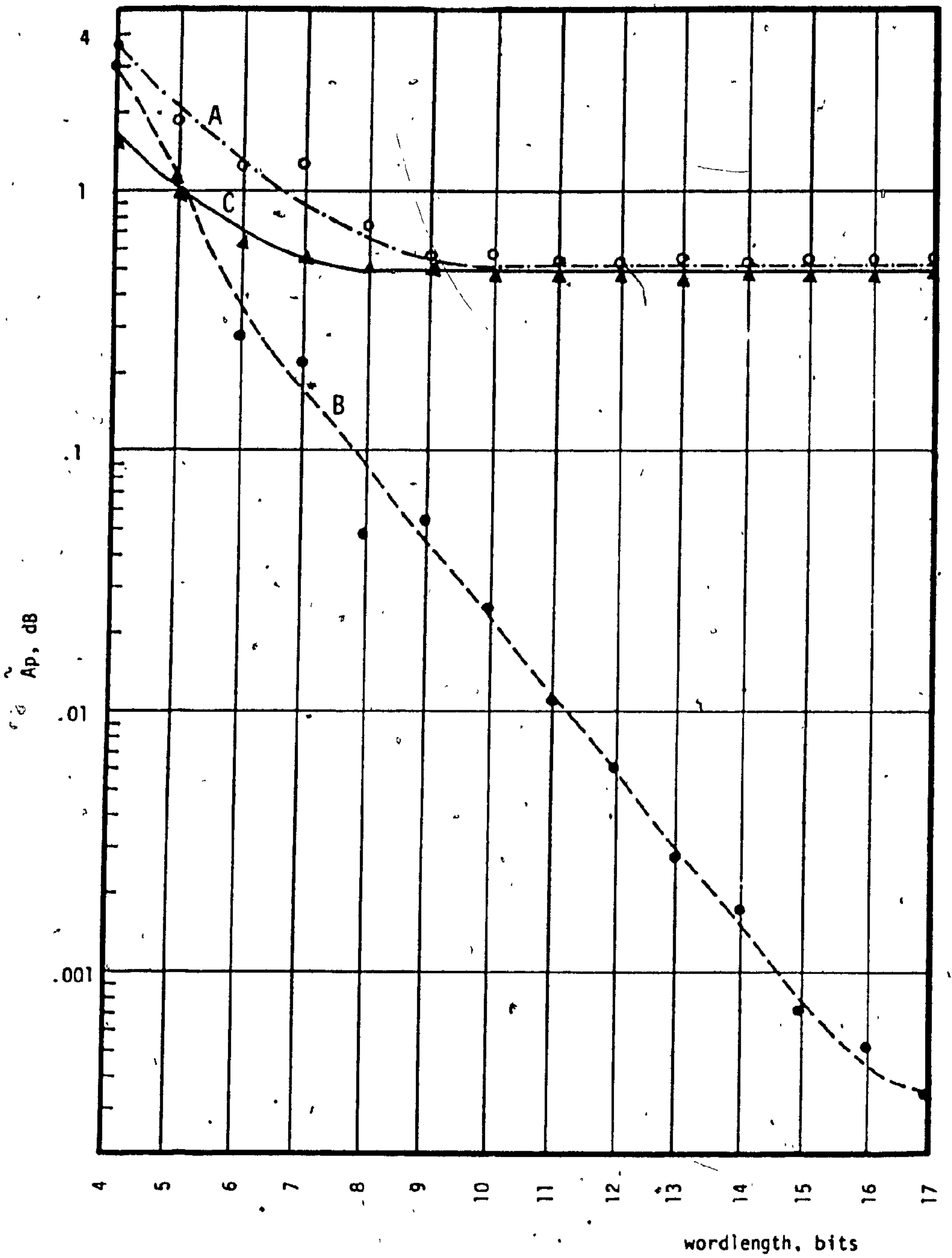


Figure 4.15 $\tilde{A}_p$ versus wordlength (floating-point arithmetic) filter No. 5

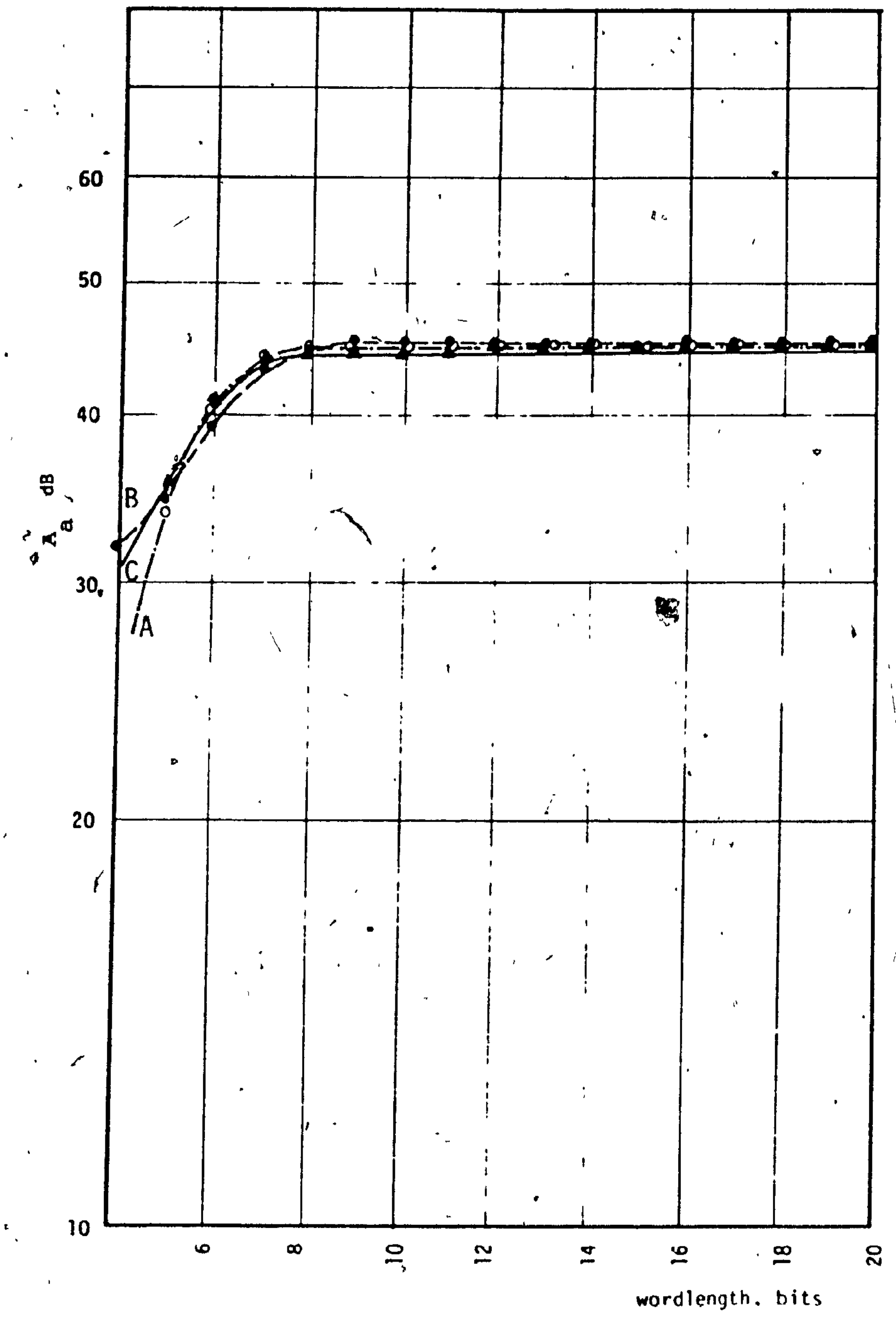


Figure 4.16 $\hat{\lambda}_a$ versus wordlength (floating-point arithmetic) filter No. 5

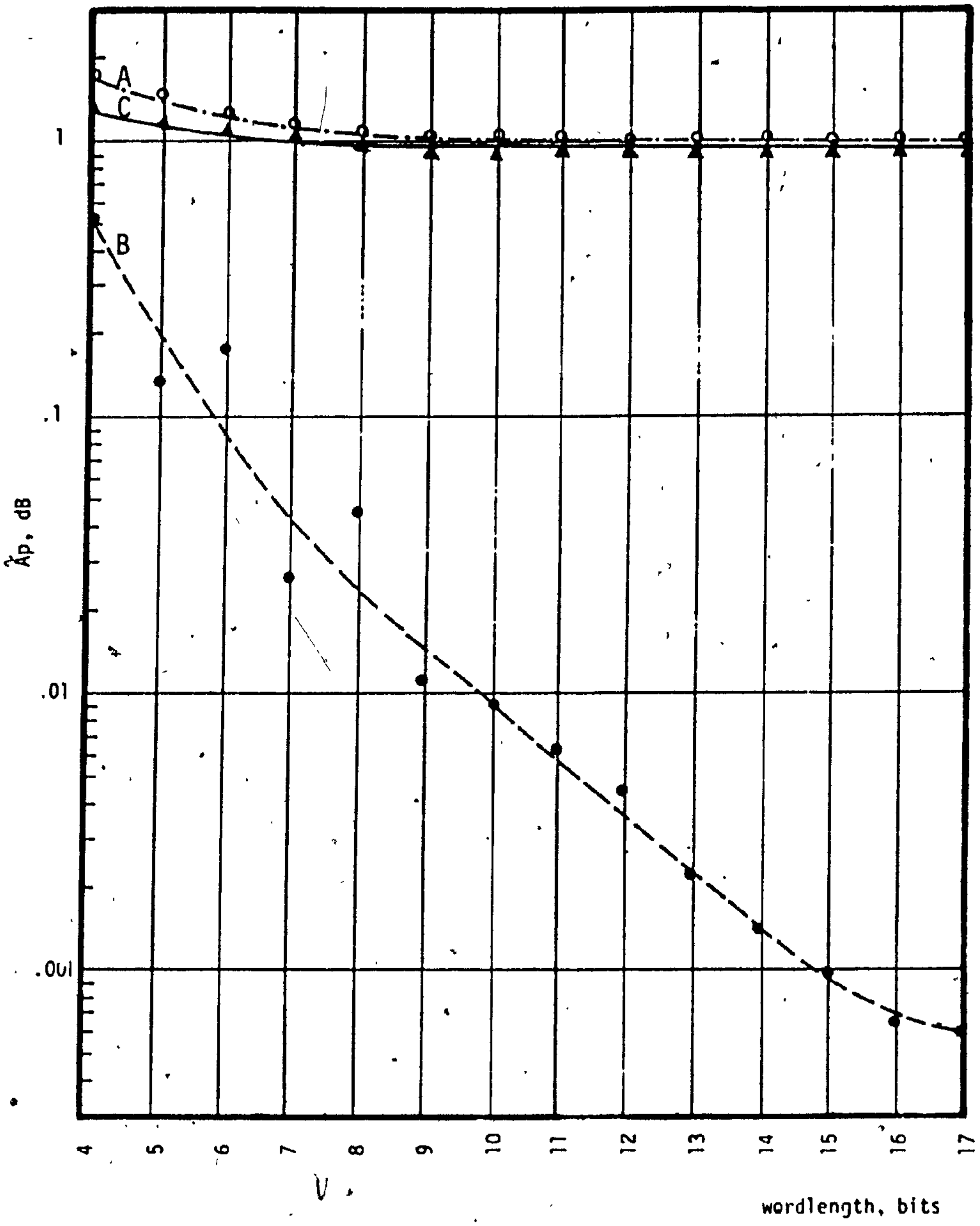


Figure 4.17 λ_p versus wordlength (floating-point arithmetic) filter No. 7

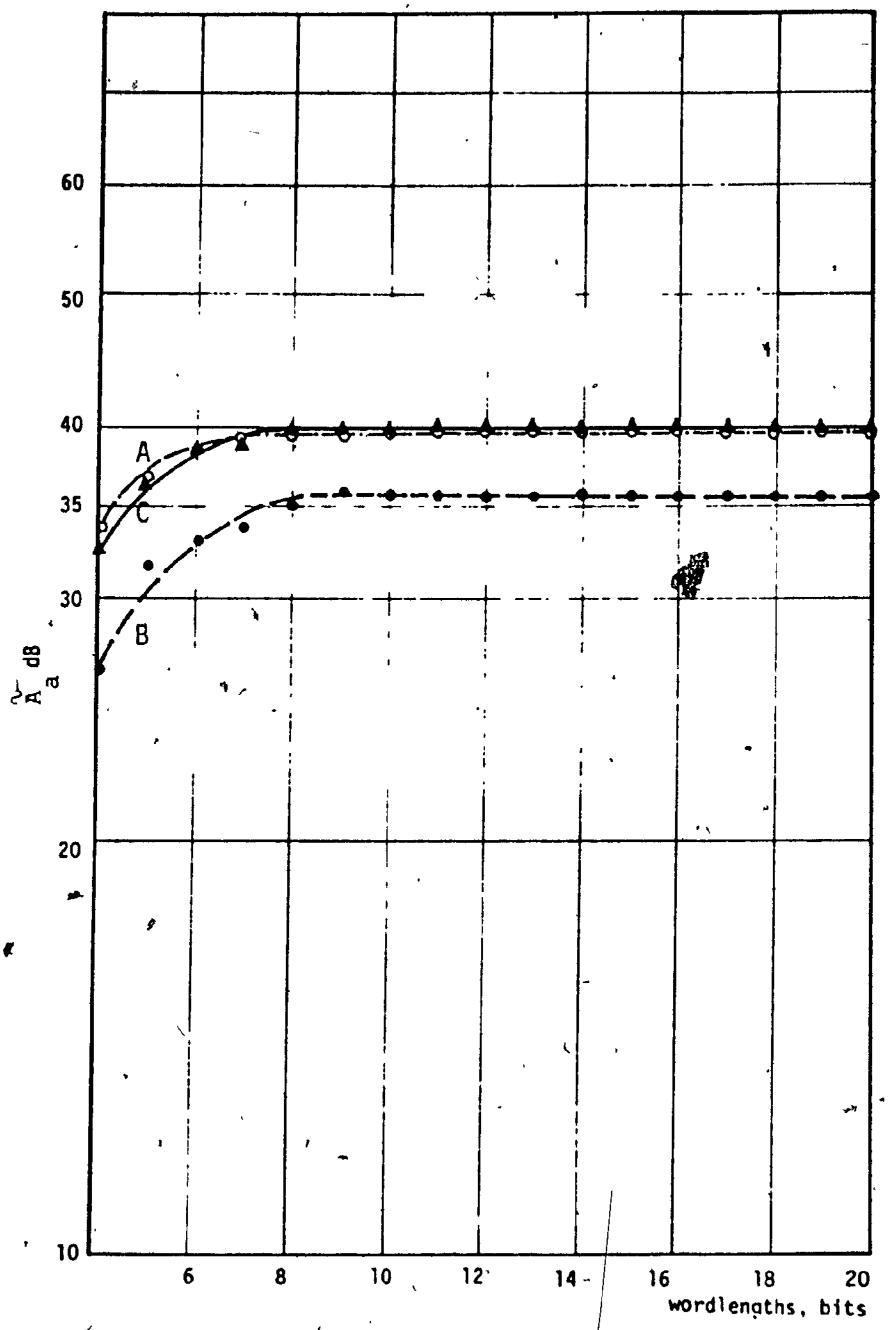


Figure 4.18 $\tilde{A}_a$ versus wordlength (floating-point arithmetic filter No. 7

CHAPTER V

CONCLUSIONS

Two distinct methods for the reduction of coefficient and product quantization effects have been investigated. In both methods a degree of freedom is introduced in the design by increasing the approximation order. The degree of freedom gained is used to minimize either the passband ripple or the maximum pole radius. In the first method the allowable margin for coefficient quantization error is increased while in the second method the sensitivity to coefficient quantization is indirectly reduced.

The two methods were used to design a diverse range of lowpass filters including some narrowband, high-selectivity, low-passband ripple, and high-stopband-attenuation filters. The results show that both methods lead to significant reductions in the required wordlength and the inband-noise power. The reduction in the wordlength that can be achieved ranges from 1 to 8 bits for fixed-point arithmetic and from 2 to 6 bits for floating-point arithmetic. The corresponding reduction in the passband average of the PSD ranges from 1.7 to 12.2 dB.

As can be seen in Tables 4.3 to 4.5, the two methods yield similar improvements. The reason is that both methods tend to reduce the pole radii relative to the pole radii in the minimum-order designs, as is evident in Tables 3.2(a),(b),(c) and in effect the sensitivity to coefficient quantization is reduced in both types of designs. Minimization of the maximum pole radius results in a lower sensitivity to coefficient quantization. However, it appears that this advantage is almost exactly counterbalanced by the associated increase in the allowable specifications margin in the case of the minimized passband-ripple design. The reduction in the output noise in both types of designs is closely linked to a known correlation between attenuation sensitivity and output noise [19].

The minimized passband-ripple method, unlike the minimized maximum-pole-radius method, entails a minimal amount of computation and is therefore to be preferred. In certain applications, however, the latter method might be preferable since the reduction in the maximum pole radius tends to reduce the amplitude of limit-cycle oscillations [1].

An increase in the approximation order corresponds to an increase in the number of arithmetic operations per sampling period. Consequently, the proposed methods introduce a trade-off whereby additional arithmetic operations are performed in order to reduce the required wordlength and output noise. This trade-off would be of value in applications where the wordlength of available VLSI chips is inadequate for the required filter specifications or in applications where noise specifications are very tight.

Only the cascade realization has been considered. However, since the improvements are brought about at the approximation stage, similar improvements are expected if other realizations (except Butterworth) are utilized [4],[6],[12].

REFERENCES

1. A. Antoniou, Digital Filters: Analysis and Design, McGraw-Hill, New York, 1979.

2. C.M. Rader and B. Gold, "Effect of parameter quantization on the poles of a digital filter", Proc. IEEE, Vol. 55, pp. 688-689, May 1967.

3. A. Fettweis, "Digital filter structures related to classical filter networks", Arch. Elektron. Uebertrag., Vol. 25, pp. 79-89, 1971.

4. A. Sedlmeyer and A. Fettweis, "Digital filters with true ladder configuration", Int. J. of Circuit Theory and Appl., Vol. 1, pp. 5-10, March 1973.

5. L.T. Bruton, "Low-sensitivity digital ladder filters", IEEE Trans. Circuits Syst., Vol. CAS-22, pp. 168-176, March 1975.

6. A. Antoniou and M.G. Rezk, "A comparison of cascade and wave fixed-point digital-filter structures", IEEE Trans. Circuits Syst., Vol. CAS-27, pp. 1184-1194, Dec. 1980.

7. E. Avenhaus, "On the design of digital filters with coefficients of limited word length", IEEE Trans. Audio Electroacoust., Vol. AU-20, pp. 206,212, Aug. 1972.

8. E. Avenhaus and W. Schussler, "On the approximation problem in the design of digital filters with limited wordlength", Arch. Elektron. Uebertrag., Vol. 24, pp. 571-572, 1970.

9. C. Charalambous and M.J. Best, "Optimization of recursive digital filters with finite word lengths", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-22, pp. 424-431, Dec. 1974.

10. F. Brglez, "Digital filter design with short word-length coefficients", IEEE Trans. Circuits Syst., Vol. CAS-25, pp. 1044-1050, Dec. 1978.

11. N.I. Smith, "A random-search method for designing finite-wordlength recursive digital filters", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27, pp. 40-46, Feb. 1979.

12. H. Kwan, "On the problem of designing IIR digital filters with short word lengths", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27, pp. 620-624, Dec. 1979.

13. W. Schussler, "On the approximation problem in the design of digital filters", Proc. 5th Annual Princeton Conf. Inf. Sc. and Syst., pp. 54-63, 1971.

14. G. Dehner, "On the design of digital Cauer filters with coefficients of limited wordlength", Arch. Elektron. Uebertrag., Vol. 29, pp. 165-168, April 1975.

15. R.E. Crochiere, "A new statistical approach to the coefficient word length problem for digital filters", IEEE Trans. Circuits Syst., Vol. CAS-22, pp. 190-196, March 1975.

16. C. Charalambous, "A method to overcome the ill-conditioning problem of differentiable penalty functions", Operations Research, Vol. 28, pp. 650-667, May-June 1980.

17. R. Fletcher, Fortran Subroutines for Minimization by Quasi-Newton Methods, Report AERE-R7125, Theoretical Physics Division, Atomic Energy Research Establishment, Harwell, Berkshire, 1972.

18. L.B. Jackson, "Roundoff-noise analysis for fixed-point digital filters realized in cascade or parallel form", IEEE Trans. Audio Electroacoust., Vol. AU-18, pp. 107-122, June 1970.

19. A. Fettweis, "Roundoff noise and attenuation sensitivity in digital filters with fixed-point arithmetic", IEEE Trans. Circuit Theory, Vol. CT-20, pp. 174-175, Mar. 1973.