Optimum Monotonic Step Response Filters

Peter H. Halpern

University of Central Florida
OPTIMUM MONOTONIC
STEP RESPONSE FILTERS

BY
PETER H. HALPERN
B. S. Physics, Rensselaer Polytechnic Institute 1956

THESIS
Submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Graduate Studies Program of the College of Engineering
at the University of Central Florida, Orlando, Florida

Spring Quarter
1980
OPTIMUM MONOTONIC TRANSIENT RESPONSE FILTERS

by

Peter H. Halpern

ABSTRACT

The problem of designing sharp cutoff filters with monotonic step responses is addressed. The impulse responses of the filters are expanded in terms of finite duration trigonometric polynomials. The coefficients of the trigonometric polynomials are obtained, for arbitrary frequency penalty functions, by solving a generalized eigenvalue problem. Once the trigonometric polynomial is specified the network can be synthesized with known techniques.

Two theorems which assist in the numerical solution are proven.
# Table Of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Formulation Of The Problem</td>
<td>3</td>
</tr>
<tr>
<td>Numerical Methods</td>
<td>10</td>
</tr>
<tr>
<td>Example Design</td>
<td>14</td>
</tr>
<tr>
<td>Results</td>
<td>17</td>
</tr>
<tr>
<td>References</td>
<td>19</td>
</tr>
</tbody>
</table>
Introduction

System designers often specify that filter designs introduce no overshoot and have monotone step response. At the same time there are often some requirements to deal with a variety of frequency dependent noise backgrounds. Consequently we are looking for filters which approximately band limit the input signal. Simultaneously we require the step response to be monotonely non-decreasing or equivalently, the impulse response to be non-negative. These combined time domain and frequency domain specifications lead to an interesting generalized eigenvalue problem.

Historically there have been a variety of low pass filters designed for flat amplitude or for linear phase response or transitional filters which are some compromise to both. When the transient response has been in question, the tendency has been to choose some sort of linear phase characteristic. In fact linear phase is neither necessary nor a sufficient condition for a (no overshoot) monotonic step response. To wit, a filter which has an impulse response

\[ h(t) = 1 + \sum_{k=1}^{n} a_k \cos 2k \pi t, \quad 0 \leq t \leq 1 \]

\[ h(t) = 0 \text{ elsewhere}, \]
will have linear phase for any choice of coefficients, $a_k$, for all frequencies; but at the same time the coefficients can be chosen so that there is overshoot. On the other hand a filter whose impulse response is

$$\sum a_k e^{-a_k t} \text{ with } a_k > 0 \text{ and } d_k > 0,$$

will have a monotonic step response whose final value is $\sum a_k$, but this filter does not have any flat amplitude or linear phase response. It is true however that the so-called Bessel filter derived by Thomson [1] and Storch [2] exhibit linear phase and virtually no overshoot.

In order to solve our problem using classical analysis we first convert the frequency domain specifications into the time domain. Then for ease in identifying our pulse responses with known lumped networks we expand our impulse responses in finite duration trigonometric pulses. According to [3] these responses can be synthesized as accurately as desired. In earlier work [4] a simple version of our problem was solved. The monotone step responses with maximum asymptotic cutoff led to impulse responses all of the form $(\sin \pi t)^n$ for $0 \leq t \leq 1$. In the following we will abandon the maximum asymptotic cutoff criterion and consider more general criteria.
Formulation Of The Problem

We want to construct an impulse response \( h(t) \) which is positive and of finite time duration, say \( 0 < t < 1 \). The function \( h(t) \) is normalized so that the step response has a final value equal to its input. This means

\[
\int_{0}^{1} h(t) \, dt = 1. \tag{3}
\]

There is a "badness" function or a "goodness" function associated with the magnitude squared of the Fourier transform of \( h(t) \). For our \( h(t) \) the Fourier transform is

\[
H(\omega) = \int_{0}^{1} e^{-j\omega t} h(t) \, dt. \tag{4}
\]

Let \( G(\omega) \) be a prescribed positive real and even function of \( \omega \) which we call our goodness function. Then the quantity to be maximized is

\[
J_1 = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega)H(\omega)H^*(\omega) \, d\omega. \tag{5}
\]

Alternately we might want to minimize the quantity,

\[
J_2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} B(\omega)H(\omega)H^*(\omega) \, d\omega. \tag{6}
\]
where \( B(\omega) \) is an appropriately defined "badness" function. These "goodness" or "badness" functions are generally derived from systems considerations of the background noise, of the spectral occupancy of the data and possibly the sampling rate if the system is also a sampled data system.

We will first treat the form of (5) and then maneuver to get (6) into the same form.

In our development we will further restrict the \( h(t) \). Although we could theoretically approximate any \( h(t) \) with arbitrary accuracy, the problem is simplified if we start with \( h(t) \) expressed as sums of known functions. That is, we want to finally express \( h(t) \) as the impulse response of lumped linear networks. Using the techniques of [3] we can accurately synthesize responses of the form

\[
h(t) = \frac{a_0}{\sqrt{2}} + \sum_{1}^{n} a_k \cos 2k\pi t , \quad 0 < t < 1 \quad (7)
\]

\( h(t) = 0 \) elsewhere.

Alternately we can accurately approximate these time functions with the form

\[
h(t) = \sum_{1}^{n} a_k \sin(2k-1)\pi t \quad 0 \leq t \leq 1 \quad , \quad (8)
\]

\( h(t) = 0 \) elsewhere.
The trigonometric polynomials in (8) can also be expressed using trigonometric identities as

\[ h(t) = \sin \pi t \sum_{0}^{n-1} b_k \cos 2k\pi t, \quad (9) \]

In general the complexity of our filter is determined by the number of terms \( n \).

Our optimization problem is then to find the set of coefficients \( \{ a_k \} \) or \( \{ b_k \} \) which maximizes \( J_1 \) of (5) and satisfies the normalization (3) and which results in a non-negative \( h(t) \).

\( J_1 \) of (5) is easily converted into a double integral in the time domain. To this end we use the fact that Fourier transforms preserve inner products (generalized Parsevals theorem; see any text on Fourier Integrals).

\[
\frac{1}{2\pi} \int_{-\infty}^{\infty} H_1(\omega)H_2^*(\omega) d\omega = \int_{-\infty}^{\infty} h_1(t)h_2(t)dt, \quad (10)
\]

where the capital letters are Fourier transforms of the small letters. Multiplication in the frequency domain becomes convolution in the time domain, thus (5) can be written

\[
J_1 = \int_{0}^{1} F^{-1} \left\{ G(\omega)H(\omega) \right\} h(t) dt
= \int_{0}^{1} \int_{0}^{1} h(x) g(t-x) h(t) dx dt, \quad (11)
\]
where \( g(x) \) is the Fourier transform of \( G(\omega) \), and \( F^{-1} \) denotes inverse Fourier transform operations. The finite limits are obtained from the restricted form of \( h(t) \) in (7) and (8). Since \( J_1 \) is positive for all non-zero \( h(t) \), the double integral in (11) is positive definite for all non-zero \( h(t) \). We will now prove Theorem I.

For the cosine (polynomial) expansion of (7) or of (9) to optimize \( J_1 \) of (5) it is necessary that all its roots occur in the closed interval \( 0 \leq t \leq 1 \) and the roots in the open interval must be of even multiplicity.

Proof: Suppose the trigonometric polynomial had a factor either linear or quadratic that did not have its roots in the interval \( 0 \leq t \leq 1 \), then we could vary that factor by \( \epsilon - C_1 \epsilon \cos 2\pi t \), for sufficiently small \( \epsilon \), since that factor must be positive in the interval. To maintain our normalization we can solve for \( C_1 \), we have

\[
\delta \int_0^1 h(t)dt = \int_0^1 Q(\epsilon - C_1 \epsilon \cos 2\pi t)dt = 0 ,
\]

(12)

where \( Q \) represents the rest of the factors* of \( h(t) \). Note

* If \( \int_0^1 Q \cos 2\pi t dt = 0 \) we use this same argument without the constant \( \epsilon \) term.
that $C_1$ is independent of $\epsilon$. There results a variation of the quantity $J_1$ of (11) which can be written

$$\delta J_1 = \int_0^1 \int (h(t) + \epsilon \hat{z}(t)) g(t-x) \left(h(x) + \epsilon \hat{z}(x)\right) \, dt \, dx$$

$$- \int_0^1 h(t) g(t-x) h(x) \, dt \, dx$$

$$= 2\epsilon \int_0^1 h(t) g(t-x) \hat{z}(x) \, dt \, dx$$

$$+ \epsilon^2 \int_0^1 \hat{z}(t) g(t-x) \hat{z}(x) \, dt \, dx \quad (13)$$

The first term on the R.H.S. of (13) must vanish for optimality, but there still remains the second term which is always positive contrary to the hypothesis of optimality. Therefore all factors must have real roots in the interval $0 \leq t \leq 1$, but for positiveness of $h(t)$ the interior roots must be of even multiplicity to avoid changes of sign.\[\star\]

In view of just the even multiplicity of the roots we can express $h(t)$ in the form

$$h(t) = \left( \frac{a_0}{\sqrt{2}} + \sum_{k=1}^{m} a_k \cos 2k\pi t \right)^2 \quad (14)$$

with the possible factor of $\sin \pi t$, if we use the form of (9).

\[\star\] denotes end of proof
Substituting in (11) (for even degree) and adjoining the constraint of (3) we arrive at the calculus of variation problem: Maximize J with respect to the coefficients $a_k$ where

$$J = \int_{0}^{1} \left( \frac{a_0}{\sqrt{2}} + \sum_{k=1}^{m} a_k \cos 2k \pi t \right)^2 g(t-x) \left( \frac{a_0}{\sqrt{2}} + \sum_{k=1}^{m} a_k \cos 2k \pi x \right)^2 dtdx$$

$$- \lambda \int_{0}^{1} \left( \frac{a_0}{2} + \sum_{k=1}^{m} a_k \cos 2k \pi t \right)^2 dt.$$  

A necessary condition is that $\frac{\partial J}{\partial a_k} = 0$. This results in a generalized eigenvalue problem:

$$\lambda a_q = \sum_{i=0}^{m} \sum_{j=0}^{m} \sum_{k=0}^{m} a_i a_j a_k \gamma_{kj} \gamma_{i} \gamma_{kq}, \quad (for \ q = 0, 1, \ldots, m)$$

where the $\gamma$ array entries are

$$\gamma = \int_{0}^{1} \int \cos 2i \pi t \cos 2j \pi t \ g(t-x) \cos 2k \pi x \ \cos 2q \pi x \ dtdx.$$  

Again there may be possible extra factors in the integrand of $\sin \pi t \sin \pi x$. It is understood that if any index is zero its corresponding cosine term in the integrand is replaced by the constant of $1/\sqrt{2}$.

For the normalized response of (3) $\lambda$ of (16) represents
$J_1$. This can be seen by multiplying (16) by $a_q$ and summing, which reproduces $J_1$ on the R.H.S. and $\lambda$ on the L.H.S. We therefore are looking for the largest $\lambda$. 
Numerical Methods

As an aid to finding the solution to our eigenvalue problem we can prove Theorem II:

The optimum solution of (16) is such that its vector of coefficients, $\bar{a}$, corresponds to the largest eigenvalue of the matrix $\bar{a} \cdot \bar{V} \cdot \bar{a}$.

Here we use the shorthand dot product* notation

$$\bar{a} \cdot \bar{V} \cdot \bar{a} = \sum_{i,k} a_i \gamma_{ij} k \bar{a}_k$$  \hspace{1cm} (18)

To see this assume that on the contrary $\bar{a}$ is expandable in normalized eigenvectors** as

$$\bar{a} = \sum C_i \bar{u}_i,$$  \hspace{1cm} (19)

where $\sum C_i^2 = 1$ and $\bar{u}_1$ is assumed to be the eigenvector corresponding to the maximum eigenvalue.

Now we vary $\bar{a}$ by adding $\epsilon \bar{u}_1$ and noting the variation in $J_1$. We start with

* More dots mean summation over more repeated indices.
** The matrix $\bar{a} \cdot \bar{V} \cdot \bar{a}$ is positive definite symmetric.
\[ J_1 = \bar{a} : \bar{\gamma} : \bar{a} = \sum c_i^2 \lambda_i \]  

(20)

The modified \( \bar{a} \) must be renormalized to satisfy (3)

\[ C(\bar{a} + \epsilon \bar{u}_1) = C(\sum c_i \bar{u}_1 + \epsilon \bar{u}_1) \]  

(21)

or \[ C^2 \approx \frac{1}{1+2\epsilon c_1} \]  

(22)

This results in a modified \( J_1 \) of approximately

\[ J_1^* \approx \frac{1}{(1+2\epsilon c_1)^2} (\bar{a} + \epsilon \bar{u}_1)^2 : \bar{\gamma} : (\bar{a} + \epsilon \bar{u}_1)^2 \]  

(23a)

\[ \approx \frac{1}{(1+2\epsilon c_1)^2} (\bar{a} : \bar{\gamma} : \bar{a} + 4\epsilon \bar{a} \bar{u}_1 : \bar{\gamma} : \bar{a}) \]  

(23b)

\[ = \frac{1}{(1+2\epsilon c_1)^2} (\sum c_i^2 \lambda_1 + 4\epsilon c_1 \lambda_1) \]  

(23c)

Choose \( \epsilon \) with the same sign as \( c_1 \) and we see that there results an increase in \( J_1^* \) since \( \lambda_1 \) is max.

If \( c_1 \) is zero then the vector \( \frac{\bar{a} + \bar{u}_1}{\sqrt{2}} \) will result in a better \( J_1^* \).
In this case

\[ J_1^* = \frac{1}{2} \bar{a}^2 : \overline{\nabla} : \bar{a}^2 + \bar{u}_1 \bar{a} : \nabla : \bar{a}u_1 + \frac{1}{2} \bar{u}_1^2 : \nabla : \bar{u}_1^2 \]

\[ = \frac{1}{2} \sum_{i=1}^{2} \lambda_i + \lambda_1 + \text{positive} \]

which is certainly bigger than \( \sum_{i=1}^{2} \lambda_i \).

In the above discussion we have used the fact that we have a positive definite tensor

\[ \bar{a} \bar{b} : \overline{\nabla} : \bar{a} \bar{b} > 0 \]

for non-zero \( \bar{a}, \bar{b} \).

For the "badness" criterion \( J_2 \) of (6) we are attempting to minimize a positive form or equivalently maximize a negative form. We can secure a positive form by adjoining a constraint of \( A ( \int h(t) \, dt )^2 \) with \( \int h(t) \, dt \). This will make \( \overline{\nabla} \) positive definite for sufficiently large \( A \) and not change the eigenvectors. Consequently the same mathematics applies to this version of the problem.

The value of Theorem II is that standard iterative procedures which are used to obtain the eigenvector corresponding to the largest eigenvalue (see Hildebrand [5]) can be used to solve our problem.

The procedure is as follows:

We take an initial guess \( \bar{a}_0 \), and form matrix \( \bar{a}_0 \cdot \overline{\nabla} \cdot \bar{a}_0 \).

We operate on \( \bar{a} \) a few iterations, say \( m \), renormalizing as necessary to get
\[ \bar{a}_1 = k(\bar{a}_0 \cdot \bar{Y} \cdot \bar{a}_0)^m \bar{a}_0 \]  

(26)

\( \bar{a}_1 \) is an approximation to the "largest eigenvector" of \((\bar{a}_0 \cdot \bar{Y} \cdot \bar{a}_0) \) according to Hildebrand [5]. Then we form the matrix \( \bar{a}_1 \cdot \bar{Y} \cdot \bar{a}_1 \) and repeat the above steps. According to Theorem II the optimum solution will be such that the best \( \bar{a} \) is the "largest eigenvector" of \( \bar{a} \cdot \bar{Y} \cdot \bar{a} \) and that we can continually improve our \( \bar{a} \) until this condition is obtained.
Example Design

In almost all cases we will need numerical methods to obtain our filter designs. The maximum asymptotic cutoff criterion of \([4]\) is a notable exception. We considered the following ideal filter "goodness" criterion as our first case:

\[
\begin{align*}
G(\omega) &= 1, \quad 0 \leq |\omega| \leq \omega_1 \\
G(\omega) &= 0, \quad \omega_1 < |\omega| 
\end{align*}
\]

(27)

This was done for several values of \(\omega_1\).

The corresponding kernel in the integrand of (11) is the inverse Fourier transform of (27) and is simply

\[
g(t-x) = \frac{\sin \omega_1 (t-x)}{\pi (t-x)}.
\]

(28)

One of the simplest non trivial filters is the 4th degree trigonometric polynomial

\[
h(t) = \left( \frac{a_0 + a_1 \cos 2\pi t}{\sqrt{2}} \right)^2
\]

(29)

The ratio of \(a_1/a_0\) ought to depend on the parameter \(\omega_1\) of (27). We know from Theorem I that

\[
\frac{a_1}{a_0} \leq -\frac{1}{\sqrt{2}}
\]

(30)
is necessary to ensure real roots of (29). The normalizing constraint of (3) requires

\[ a_1^2 + a_0^2 = 2. \]  

(31)

Observe that for our 4th degree trigonometric polynomial we need only solve for one parameter, namely the ratio of \( a_1/a_0 \).

This was solved using the method outlined in Theorem II, but for any value of \( \omega_1 \) we could have simply plotted the "goodness" function of \( a_1/a_0 \).

The \( \overline{\gamma} \) array has 16 entries but because of the symmetries:

\[
\begin{align*}
10 \gamma_{00} &= 01 \gamma_{00} = 00 \gamma_{10} = 00 \gamma_{01} \\
11 \gamma_{00} &= 00 \gamma_{11} \\
10 \gamma_{10} &= 10 \gamma_{01} = 01 \gamma_{10} = 01 \gamma_{01} \\
11 \gamma_{10} &= 11 \gamma_{01} = 10 \gamma_{11} = 01 \gamma_{11}.
\end{align*}
\]

(32)

Only six different integrals of the form of (17) were calculated for every \( \omega_1 \). This was done by a two dimensional Simpson's rule.

An initial guess of \( a_0 \sqrt{2} = 1, a_1 = 0 \) was used. After each iteration \( a_0 \sqrt{2} \) was renormalized to 1. The matrices \( \overline{a} \cdot \overline{\gamma} \cdot \overline{a} \) were reformed every 5th iteration of the vector \( \overline{a} \).

For a range of \( \omega_1 \) from 3 to 12 we obtained practi-
cally the same coefficients which yielded
\[ h(t) \approx \frac{2}{3} \left( 1 - \cos 2\pi t \right)^2 = \frac{8}{3} \sin^4 \pi t. \]  
(33)

This is the same result as that of [4].

If we had used a significantly different goodness function \( G(\omega) \) we could expect considerably different filters. A practical example which will be worked in the future is one where there is strong narrow band interference. This results in near impulses in the badness function and will create zeroes of transfer at the interfering frequencies in the resulting monotone step response filter.
Results

A theoretical development is given for finding optimum monotonic step response filters under a variety of criteria. The filters are developed in terms of trigonometric pulse forming networks. There results an interesting generalized eigenvalue problem. We solved this eigenvector problem numerically for only low degree cases with the "goodness" function of the form

\[ G(\omega) = \begin{cases} 1 & |\omega| \leq \omega_1 \\ 0 & |\omega| > \omega_1 \end{cases} \]  

For a wide range of \( \omega_1 \) we obtained answers which differed very little from the maximum asymptotic cutoff criterion of \([4]\). The impulse responses were all nearly

\[ h(t) \approx \frac{8}{3} (\sin \pi t)^4 \quad 0 \leq t \leq 1 \]  

We believe higher degree filters will be required to appreciate a significant difference with this new approach.

It was found emperically that the numerical method outlined in Theorem II yielded local maxima. We had to search for what we hoped was a global maximum. At this time we do not have a sufficient test for global maxima nor do we even
know how many distinct eigenvalues to expect. This phenomenon is peculiar to our non-linear eigenvalue problem. The iterative techniques for the linear eigenvalue problem will always converge to the extreme magnitude eigenvalues as long as they are unique.

Nonetheless using our method combined with appropriate simple search techniques we should be able to find the solutions for higher order filters and for any "goodness" or "badness" functions, and obtain the optimum filters.
References


