Abstract
Quantization of data and representation of the coefficients is studied in digital matched filters for weak-signal detection. An algorithm for optimum coefficients and equations for the optimum input quantizer are obtained for the known signal in additive noise problem. Some numerical performance results are given.

I. INTRODUCTION
A matched filter is often used as a detector for testing a hypothesis about an input signal; it is the optimal processor in this role for Gaussian input noise, and may also be considered as the optimum processing scheme in non-Gaussian noise in the weak-signal case. We will elaborate on this in the next section; the general theory and applications of matched filters may be found, for example, in [1,2].

In this paper we will be concerned with digital matched filters operating on discrete-time data, and we will examine the effects of, and optimization with respect to, finite-bit representation or quantization of the coefficients and analog inputs. Most previous investigations in this direction have been concerned either with data quantization only (e.g. [3]) or have assumed only simple one-bit coefficient representations [2].

In the next section we briefly consider the basic results on local, or weak-signal, detection of signals based on quantized data and finite-bit coefficients. In Section III implementation of optimum digital matched filtering is considered, and an algorithm for determining the best coefficient representation is discussed. In Section IV we give some numerical examples.

II. PERFORMANCE WITH QUANTIZED COEFFICIENTS AND INPUTS
Let us assume that an observation vector \( \mathbf{X} = (X_1, X_2, \ldots, X_n) \) is available, and is described by the equation
\[
X_i = \theta s_i + N_i, \quad i=1,2,\ldots,n, \ \theta > 0. \tag{1}
\]

Here the vector \( \mathbf{s} = (s_1, s_2, \ldots, s_n) \) is a known signal vector, \( \theta \) is the amplitude of the signal, and the vector \( \mathbf{N} = (N_1, N_2, \ldots, N_n) \) is a vector of independent identically distributed noise samples each with symmetric density \( f \).

If we consider as a detection statistic for testing \( H_0: \theta = 0 \) vs. \( H_1: \theta > 0 \) a
\[
T = \sum_{i=1}^{n} g_i(X_i), \tag{2}
\]
and use as a criterion of optimality the differential signal-to-noise ratio
\[
DSNR = \left( \frac{d}{d\theta} E(T) \bigg|_{\theta = 0} \right)^2 \tag{3}
\]
we find (from the Schwarz inequality) that the optimum \( T \) maximizing DSNR is
\[
T_{opt} = \sum_{i=1}^{n} s_i g_{opt}(X_i) \tag{4}
\]
where
\[
g_{opt}(X_i) = -f'(X_i)/f(X_i) \tag{5}
\]
the prime denoting the first derivative of the function. The criterion of (3) is a reasonable one for the weak-signal case, being a modified case of the usual SNR criterion [2]. It is also well-known, of course, that \( T_{opt} \) is the locally-optimum statistic maximizing the slope of the power function for testing \( H_0 \) vs. \( H_1 \) [4].

In the case of Gaussian noise \( T_{opt} \) is the output of a linear filter matched to the signal vector \( s \). In the general case, we may consider \( T_{opt} \) to be the output of a similar filter preceded by the instantaneous nonlinearity \( g_{opt} \). For the digital matched filter both the coefficients, the \( s_i \), and the data inputs, the \( g_{opt}(X_i) \), have to be replaced by suitably quantized versions \( r_i \) and \( q(X_i) \), respectively.
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)

NOTICE OF TRANSMITTAL TO DDC

This technical report has been reviewed and is approved for public release IAW AFR 190-12 (7b).

Distribution is unlimited.

A. D. BLOSE
Technical Information Officer
To optimize the performance of the digital matched filter, then, we have to optimize DSNR of (3) for the case where the detection statistic is
\[ T_d = \sum_{i=1}^{n} r_i q(x_i), \]
the optimization being with respect to a choice of the \( r_i \) and the input quantizer \( q \), given \( 2^b = 2k \) levels or \( b \) bits for coefficient representation and \( 2^t = 2m \) levels or \( t \) bits for the data quantization. Note that (6) represents a general digital matched filter operating on quantized data, for any kind of input noise density.

Applying (3) to \( T_d \), we find that DSNR for \( T_d \) with an odd-symmetric, even-state quantizer \( q \) is given by
\[
\text{DSNR} = \left( \frac{\sum_{i=1}^{n} s_i r_i}{\sum_{i=1}^{n} r_i^2} \right)^2.
\]
where \( F \) is the distribution function corresponding to \( f \). In (7) the \( \ell_j \) are the output levels for positive \( x \) of the symmetric quantizer \( q \), with \( q(x) \) being \( \ell_j \) whenever \( x \) is in the interval \((a_j, a_{j-1})\). The breakpoints \( a_j \) satisfy \( a_j < a_{j-1} \), and by definition \( a_0 = -\infty \) and \( a_m = \infty \).

The problem is to maximize the quantity
\[
J = \left( \frac{\sum_{i=1}^{n} s_i r_i}{\sum_{i=1}^{n} r_i^2} \right)^2,
\]
which is a factor in \( \text{DSNR} \). In addition, for a given set of levels \( \ell_j \) for the quantizer \( q \), the best set of breakpoints \( a_j \) may be obtained. Then the overall performance of the digital matched filter may be optimized with respect to allocation of \( b+t \) bits between coefficient and data quantization.

III. PERFORMANCE OPTIMIZATION

As discussed in the previous section, the objective is to maximize DSNR of (7). The two factors in (7) are decoupled, so that the coefficient representation and data quantization problems can be treated independently of each other. We first consider maximization of the coefficient factor \( J \) defined in (8). This quantity can be written as
\[
J = \frac{(r \cdot s)^2}{||z||^2},
\]
where \( s \) is the previously defined signal vector and \( r = (r_1, r_2, \ldots, r_n) \), the vector of filter coefficients by which the reference signal \( s \) is represented. If we let \( \phi \) be the angle between \( s \) and \( r \), so that
\[
\cos^2 \phi = \frac{(r \cdot s)^2}{||z||^2 ||s||^2},
\]
we can express \( J \) as
\[
J = ||s||^2 \cos^2 \phi.
\]
Thus, maximization of \( J \) for a given reference signal is simply maximization of \( \cos^2 \phi \), and we therefore have to pick the vector \( r \) closest to \( s \).

Since it was not easy to get an analytic solution for the coefficient vector \( r \) maximizing \( \cos^2 \phi \) for a given \( s \) (and with a constraint on the number of bits, \( b \), for coefficient representation), an efficient computer technique was developed which is described below.

It was assumed that the coefficient representation \( r \) is obtained through a quantizer with one of three possible ranges of \( 2^b = 2k \) levels:

(a) The levels 0, 1, ..., 2k-1
(b) The levels -k+1, -k+2, ..., k
(c) The levels -k, -k+1, ..., k-1

Range (a) is obviously to be used if \( s \) has positive components. A signal vector with all negative components may be complemented, and therefore be represented in this range also. Ranges (b) and (c) are more natural choices if positive as
well as negative components are present
in $s$. Note that we are considering the
class of quantizers with an even number
of levels, one of which is the 0 level.

The search algorithm for $r$ is de-
scribed by the following sequence of
steps. It is assumed that the signal
vector components have been ordered,
that is $s_1 \leq s_2 \leq \ldots \leq s_n$, and are non-zero.

0. Initialization: $j = 2k-1$, $C = 0$.
1. Check if $s_j > 0$. If not, invert the
sequence of components in $s$ so that
now $s = (-s_n, -s_{n-1}, \ldots, -s_1)$, and go
to step 3.
2. Check if $\sum_{i=1}^{n} s_i > 0$. If the sum is
negative, invert the sequence of
components in $s$, so that now
$s = (-s_n, -s_{n-1}, \ldots, -s_1)$. Note that
inverting the sequence in step 1
results in a sequence which fulfills
the condition in this step.
3. Form the vector $s(j) = (s_1(j),
\ldots, s_n(j))$ where $s_i(j) = s_i / s_n$.
Let $I_i(j)$ be the largest integer
less than, or equal to, $s_i(j)$. Thus
$I_n(j) = j$; the algorithm now looks
for the first $(n-1)$ components of
$r(j) = (r_1(j), r_2(j), \ldots, r_{n-1}(j), j)$
so that the cosine squared of the
angle between $s$ and $r(j)$ is maxi-
mized.
4. If $j = k$, go to step 6. If $j < k$, go
to step 7.
5. If $s_i(j) > 0$, consider the two possi-
bilities $I_i(j)$ and $U_i(j)$ for $r_i(j)$,
where $U_j(j)$ is the smallest integer
larger than, or equal to, $s_j(j)$.
If $s_i(j) < 0$, set $r_i(j) = 0$. Go to
step 8.
6. If $s_i(j) \geq -k+1$, consider the possi-
bilities $I_i(j)$ and $U_i(j)$ for
$r_i(j)$. If $s_i(j) < k+1$, set
$r_i(j) = -k+1$. Go to step 8.
7. If $s_i(j) \geq -k$, consider the possi-
bilities $I_i(j)$ and $U_i(j)$ for $r_i(j)$.
If $s_i(j) < -k$, set $r_i(j) = -k$.
8. Compute the square of the cosine,
c$^2(j)$, of the angle between $s$ and
$r(j)$ for each different possible
combination of components of $r(j)$;
let $r_m(j)$ be the vector yielding
a maximum value $c^2_m(j)$ for $c^2(j)$.
If $c^2_m > C$, assign the value $c^2_m(j)$
to $C$ and define $r$ to be $r_m(j)$.
9. If $j > 1$, set $j = j - 1$ and go to step 3.
10. Stop. The optimum coefficient vec-
tor is $r$, and $C$ is the square of
the cosine of the angle between $r$
and $s$, if the two conditions of
steps 1 and 2 had been fulfilled.
Otherwise the optimum coefficient
vector is in inverse order in $r$.

On an intuitive basis, one might
choose the best coefficient vector in
the following way for the simple case
where the $s_i$ are positive. Pick the
vector of coefficients $\bar{r}$, where the
components $\bar{r}_i$ are the integers closest to
the $s_i(2k-1)$. However, it is clear that
$\bar{r}$ need not be an optimum set maximizing
$\cos^2\theta$. This can be seen by considering
the case $b=1$, giving two-level represen-
tation, for $(n-1)$ identical, small $s_i$
and a large value for $s_n$. In the next
section this is further illustrated for
an example which is not as extreme.

We also need to consider optimiza-
tion with respect to the data quantizer.
For the symmetric even-state input quan-
tizer we are considering, specification
of the number of levels $2^{b-2}=m$ (with $t$
bits) fixes the levels at values
$\pm 1, \pm 2, \ldots, \pm m$. With $l_j = m-j+1, j=1,2,\ldots, m$, the
second factor $K$ in (9), defined by
\begin{equation}
K = \frac{1}{m} \sum_{j=1}^{m} \left( \frac{e^{(a_j)}}{e^{(a_{j-1})}} \right)^2, \quad (12)
\end{equation}
can be maximized with respect to the $a_j$.
The optimum set of $a_j$'s can easily be
obtained for specific densities $f$. Set-
ting the partial derivative of $K$ with
respect to $a_j$ equal to zero, we find a
necessary condition for the maximizing
values of $a_j$:
Equation (13) may be solved for specific densities. For the case of a Gaussian density with variance $\sigma^2$, the equations reduce to

$$a_j = \sigma^2 \left[ \frac{\ell_j + \ell_{j+1}}{2} \right] L, \quad j=1, \ldots, m$$

(14)

Substitution of (14) in (15) leads to a single equation for $L$, which may then be solved (numerically) and hence the optimum $a_j$ can be obtained. The next section contains some specific numerical results.

IV. RESULTS AND DISCUSSION

The considerations of the previous section were applied to several specific cases, two of which are presented here. A signal vector $s$ of 10 components and one of 15 components is shown in Table I, together with optimum representations for one, two and three bits. The maximum values $J_{\text{max}}$ of $J$ [Equation (11)] normalized by $\|s\|^2$ are also shown in Table I. The results indicate that even with only two bits for coefficient representation ($\cos^2 \phi$)$_{\text{max}}$ is very close to unity, which corresponds to the analog case. The exact numerical values depend on the coefficient vector and its length.

In the second part of Section III the optimization of the second factor $K$ in (7) was discussed. The optimum performance for the specific case of Gaussian noise with unit variance is shown in Table II. Again, in this case the value $K=1$ is obtained for unquantized data. Table III combines the results of Tables I and II, giving the overall performance for the different allocations of a fixed number of bits (six bits) between coefficient representation and data quantization. In general, the optimum allocation will depend on the length of the vector $s$, its components and the type of noise density. It is seen that one need consider only a small number of bits to achieve near-analog performance.

In Section III an approximation method of quantizing the coefficient sequence $s$ was also discussed. Table IV gives an example of this approximation method using the same sequence as in Table I(a) with $n=10$. Note that the discrepancy is larger for the lower-order quantizers, as expected. In fact, the result of the approximation for 3 bits is identical to that in Table I. It is seen that in this example, the approximation methods gives performance very close to optimum.

ACKNOWLEDGEMENT

This research is supported by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR-77-3154.

REFERENCES


Table I: Optimum Coefficient Representation

(a) n=10

<table>
<thead>
<tr>
<th>COEFFICIENTS</th>
<th>-1.82</th>
<th>-0.92</th>
<th>-0.08</th>
<th>0.77</th>
<th>1.22</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.58</td>
<td>1.98</td>
<td>2.30</td>
<td>2.41</td>
<td>3.10</td>
</tr>
<tr>
<td>ONE BIT REPRESENTATION</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>TWO BIT REPRESENTATION</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>THREE BIT REPRESENTATION</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

(b) n=15

<table>
<thead>
<tr>
<th>COEFFICIENTS</th>
<th>-4.3</th>
<th>-2.1</th>
<th>-1.8</th>
<th>-1.7</th>
<th>-1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.7</td>
<td>0.1</td>
<td>0.3</td>
<td>1.1</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>2.1</td>
<td>3.1</td>
<td>3.3</td>
<td>4.1</td>
<td>4.5</td>
</tr>
<tr>
<td>ONE BIT REPRESENTATION</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TWO BIT REPRESENTATION</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>THREE BIT REPRESENTATION</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

\[ \frac{J_{\text{max}}}{||s||^2} = (\cos^2 \phi)_{\text{max}} \]

Table II: Optimum Data Quantization (Gaussian Noise, Unit Variance)

<table>
<thead>
<tr>
<th>QUANTIZER BITS</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPTIMUM VALUE OF K</td>
<td>0.637</td>
<td>0.842</td>
<td>0.946</td>
<td>0.984</td>
</tr>
</tbody>
</table>
Table III: Bit Allocation Between Coefficient and Data Quantization (Gaussian Noise, Unit Variance)

(a) n=10

<table>
<thead>
<tr>
<th>Coefficient Bits</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Bits</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>DSNR</td>
<td>&lt;0.792</td>
<td>0.941</td>
<td>0.936</td>
<td>&lt;0.842</td>
<td>&lt;0.637</td>
</tr>
<tr>
<td>$</td>
<td></td>
<td>\bar{g}</td>
<td></td>
<td>^2$</td>
<td></td>
</tr>
</tbody>
</table>

(b) n=15

<table>
<thead>
<tr>
<th>Coefficient Bits</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Bits</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>DSNR</td>
<td>&lt;0.604</td>
<td>0.892</td>
<td>0.926</td>
<td>&lt;0.842</td>
<td>&lt;0.637</td>
</tr>
<tr>
<td>$</td>
<td></td>
<td>\bar{g}</td>
<td></td>
<td>^2$</td>
<td></td>
</tr>
</tbody>
</table>

Table IV: Suboptimum Coefficient Representation

<table>
<thead>
<tr>
<th>COEFFICIENTS</th>
<th>-1.82</th>
<th>-0.92</th>
<th>-0.08</th>
<th>0.77</th>
<th>1.22</th>
<th>$J_{\max} = (\cos^2\phi)_{\max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{|\bar{g}|^2}{|\bar{g}|^2}$</td>
<td>1.58</td>
<td>1.98</td>
<td>2.30</td>
<td>2.41</td>
<td>3.10</td>
<td></td>
</tr>
<tr>
<td>ONE BIT REPRESENTATION</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.775</td>
</tr>
<tr>
<td>TWO BIT REPRESENTATION</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.930</td>
</tr>
<tr>
<td>THREE BIT REPRESENTATION</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0.990</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>
Quantization of data and representation of the coefficients is studied in digital matched filters for weak-signal detection. An algorithm for optimum coefficients and equations for the optimum input quantizer are obtained for the known signal in additive noise problem. Some numerical performance results are given.
**Title**: Coefficient and Data Quantization in Matched Filters for Detection

**Authors**: Tong Leong Lim, Saleem A. Kassam

**Performing Organization**: University of Pennsylvania, Dept of Systems Engineering, Philadelphia, PA 19104

**Controlling Office**: Air Force Office of Scientific Research/NM, Bolling AFB, Washington, DC 20332

**Report Date**: 1977

**Number of Pages**: 6

**Distribution Statement**: Approved for public release; distribution unlimited.

**Abstract**: Quantization of data and representation of the coefficients is studied in digital matched filters for weak-signal detection. An algorithm for optimum coefficients and equations for the optimum input quantizer are obtained for the known signal in additive noise problem. Some numerical performance results are given.