SIGNAL ESTIMATION USING ADAPTED TREE-STRUCTURED BASES
AND THE MDL PRINCIPLE

Pierre Moulin

University of Illinois
Beckman Institute and ECE Dept.
405 N. Mathews Ave.
Urbana, IL 61801
Email: moulin@itf.uiuc.edu

ABSTRACT
We present a new signal denoising technique based on adapted tree-structured bases and the following paradigms:
(a) use of the MDL principle for estimating the signal in each branch of the tree and the best tree structure;
(b) use of signal-adapted filter banks.

1. MODEL FOR SIGNAL ESTIMATION
Consider the following model for signal estimation:
\[ y_i = f_i + \epsilon_i, \quad 0 \leq i < N, \]  
(1)
where \( \epsilon_i \) are iid \( N(0, \sigma^2) \) and \( \{f_i, 0 \leq i < N\} \) is an unknown, discrete-time signal. The orthogonal-series approach to estimation is based on the representation of the signal in a suitable orthonormal basis \( \{\phi_n\} \),
\[ f_i = \sum_{n=0}^{N-1} \theta_n \phi_n(i), \quad 0 \leq i < N, \]
in which the energy of the coefficients \( \theta_n \) is concentrated along a small number of coordinates. The problem (1) is mapped into the equivalent problem of estimating \( \theta \) from the rotated data
\[ \tau_n = \theta_n + w_n, \quad 0 \leq n < N, \]  
(2)
where \( \tau_n = \sum_{i=0}^{N-1} y_i \phi_n(i) \) and \( w_n = \sum_{i=0}^{N-1} \epsilon_i \phi_n(i) \) are iid \( N(0, \sigma^2) \) by orthonormality of the transform.

2. MDL AND THRESHOLDING
The popular thresholding technique consists in estimating each coefficient \( \theta_n \) individually as \( \hat{\theta}_n = T_\lambda(\tau_n) \), where
\[ T_\lambda(\tau) \triangleq \begin{cases} 
0 & |\tau| < \lambda \\
\tau & \text{else}
\end{cases} \]  
(3)
is the "hard threshold" function. The particular choice \( \lambda = \sigma \sqrt{2 \ln N} \), motivated by minimax considerations, has been studied in [1]. Other choices are possible. The papers [2, 3] point to the equivalence of the thresholding approach with the complexity-penalized maximum-likelihood approach
\[ \text{Maximize } \hat{\theta}_k \left( -\frac{1}{2\sigma^2} \sum_{i=0}^{N-1} |\tau_n - \hat{\theta}_n|^2 - \frac{(\lambda/\sigma)^2}{2}k \right) \]  
(4)
where \( k \) is the number of nonzero coefficients. In particular, a useful form of the Minimum Description Length (MDL) principle [4] arises with the choice
\[ \lambda = \sigma \sqrt{3 \ln N} \]  
(5)
in which case the negative maximum penalized likelihood criterion (4) scaled \( 1 \) by \( \frac{1}{\ln 2} \) is interpreted as a codelength for \( \tau \) (ignoring additive constants),
\[ L(\tau) = \sum_{n=0}^{N-1} \log_2 p(\tau_n|\hat{\theta}_n) + \frac{3}{2}k \log_2 N \]
\[ = \frac{1}{2(\ln 2)\sigma^2} \sum_{n=0}^{N-1} |\tau_n - \hat{\theta}_n|^2 + \frac{3}{2}k \log_2 N. \]  
(6)
The nonzero coefficients \( \hat{\theta}_n \) specify a particular model for describing the data \( \tau \). According to the MDL principle applied to estimation, the best model among a collection of models is the one that minimizes the total codelength for the data plus the model. The second term in (6) represents the codelength for the model \( \hat{\theta} \) [5] and improves on the formulation in [2, 7]. It accounts for the familiar \( \frac{1}{2}k \log_2 N \) bits for encoding the \( k 

1Thanks are due to Andrew Barron for indicating that this scaling factor was missing from the log-likelihood function in [5, 3].
real parameters $\theta_n$ and $k \log_2 N$ bits for encoding their
index (assuming a uniform prior on indices). An addi-
tional $\log_2 N$ bits is used for encoding the parameter $k$
but plays no role in the minimization problem.

3. MDL AND OPTIMAL TREE DESIGN

The second basic problem is to find an orthonormal
basis in which the energy of the coefficients $\theta_n$
is concentrated along a small number of coordinates. The use
of wavelet bases for this purpose has been the object
of considerable research [1]. The search for even better
bases has led to adaptive tree structures such as wavelet
packets [6, 5, 7, 8]. (Here each branch of the tree corre-
sponds to a subband channel.) This raises the funda-
mental problem of choosing a criterion that determines
the tree structure. Various choices have been consid-
ered in the literature; a popular one is based upon the
"entropy" cost function [6, 5, 7]

$$- \sum_n \gamma_n^2 \ln \gamma_n^2. \quad (7)$$

For instance, [7] compares two values of the cost func-
tion for each splitting decision, and makes the decision
based on a hypothesis test.

Although the cost function (7) has a reasonable
heuristic interpretation, we seek a cost function more
attuned with fundamental estimation principles. We
view the choice of a tree as a choice between competing
models, and choose the best model according to the
MDL principle. The MDL cost function is then an
extension of (6),

$$MDL = L(\text{Tree}) + \frac{1}{2(\ln 2)^2} \sum_{n=0}^{N-1} |\gamma_n - \hat{\theta}_n|^2 + \frac{3}{2} k \log_2 N \quad (8)$$

where the additional term $L(\text{Tree})$ accounts for the cost
of the tree.

The tree is encoded using the following standard
(and convenient) algorithm. Nodes are visited in a pre-
determined order (e.g., starting from the root node
and visiting left children first), and one bit is used to spec-
ify whether each node is a leaf or not, see example in
Fig. 1. $L(\text{Tree})$ is thus equal to the number of nodes
in the tree and is in the range $[1, 2N - 1]$. The wavelet
tree is a simple tree with only $2d + 1$ nodes, where
$d < \log_2 N$ is the depth of the tree. Highly complex
trees have up to $2N - 1$ nodes.

2 A similar approach has been proven beneficial in the closely
related signal coding problem, choosing the tree structure that
optimizes the rate–distortion performance of the encoder [9].

![Figure 1: Code for depth–2 wavelet tree is 00111.](image)

![Figure 2: Splitting Decision at node $A$.](image)

3.1. Optimization Algorithms

Our optimization algorithms consist in a sequence of
decisions at the nodes. At a given node we denote by $\tau_{\text{parent}}$
the signal in the parent branch, and by $\tau_{\text{Lchild}}$
and $\tau_{\text{Rchild}}$ the signals in the left and right children
branches (Fig. 2). The decision for or against splitting
the parent branch is the outcome of the following com-
parison between codelengths before and after splitting:

$$L(\tau_{\text{parent}}) < L(\tau_{\text{Lchild}}) + L(\tau_{\text{Rchild}}) + 2. \quad (9)$$

The constant 2 represents the increased complexity of
the tree (two new nodes) following the splitting deci-
sion. The shortest $L(\tau_{\text{parent}})$, $L(\tau_{\text{Lchild}})$, and
$L(\tau_{\text{Rchild}})$ are obtained as described in Section 2; in particular,
the threshold (5) is used in each branch. The optimal
splitting decision is the one which results in the smaller
codelength.

Two specific algorithms are suggested:

Dynamic Programming (DP) Algorithm. A global
solution to the optimization problem may be found
using DP. The key is the orthonormality of the
wavelet–packet basis family and the additivity of the
cost function (8) over leaves of the tree. Such
algorithms are studied in detail in [6, 9].

Greedy Algorithm. This is a straightforward, top-
down approach. Starting with the root node,
make each splitting decision according to the criterion (9). This algorithm progressively grows the tree, and the order in which nodes are visited is immaterial.

3.2. Relation to Other Work

In [5], MDL is used for finding the best of several wavelet-packet best bases. However, each wavelet-packet best basis is chosen so as to optimize the entropy criterion (7).

In the recent paper [8], the best basis is chosen according to the criterion

\[ \sum_n \min(\rho_n^2, \Lambda) \]

where \( \Lambda = \mu^2 \sigma^2 (1 + \sqrt{2\ln(N \log_2 N)})^2 \) and \( \mu \geq 8 \), and the coefficient estimates are obtained from (3) as

\[ \hat{\theta}_n = T_n(z_n). \]

This estimator maximizes the criterion

\[ -\frac{1}{2\sigma^2} \sum_{i=0}^{N-1} |r_n - \hat{\theta}_n|^2 - \frac{\Lambda}{2} k \]

and differs from (8) in that (a) complex trees are not penalized, and (b) the threshold \( \sqrt{\Lambda} \) is larger than (5) by at least a factor \( \sqrt{2}/3 \). The criterion (8) implies a significant penalty (up to \( 2N - 1 \) bits) for complex trees, but also a smaller penalty on nonzero coefficients (the threshold is the same as in MDL wavelet denoising).

4. SIGNAL–ADAPTED FILTER BANKS

A third point of interest concerns the choice of the filter bank. The traditional approach consists in choosing a particular filter bank, e.g., using Daubechies’ filters or coiflets, and leads to the wavelets and wavelet packets construction discussed above. Here we propose another approach, based on the use of filter banks adapted to the statistics of the signal in each channel. This method has shown great promise in coding applications [10, 11], and a fast algorithm for maximizing energy compaction has recently been introduced in [12, 13]. This technique may be applied to each node of the tree, so the optimal basis \( \{\phi_n\} \) is determined not only by the tree structure, but also by the filter bank at each node.

5. REFERENCES


