SIGNAL PROCESSING AND COMPRESSION WITH WAVELET PACKETS

Ronald R. Coifman
Yves Meyer*
Steven Quake†
M. Victor Wickerhauser‡

Numerical Algorithms Research Group
Department of Mathematics
Yale University
New Haven, Connecticut 06520

5 April 1990**

****Historical Note.** This article discusses results obtained in the Summer of 1989 and first presented at a short wavelets conference at CNRS/Marseille in November, 1989. It was originally intended for publication in the Proceedings of the Conference on Wavelets and Applications, Marseille, 1989 (edited by Y. Meyer; published by Masson, Paris, 1992). However, that volume is restricted to results presented at the big meeting in May, 1989, and the short November conference issued no proceedings.

Although it has been available in electronic form since 1990, there has been a continuing demand for paper copies of this article. The authors felt that it should be published, despite having been superceded by more recent work, because the original article has some historical and tutorial value.

0. Introduction

We describe some new algorithms for signal processing and data compression based on a collection of orthogonal functions which we shall call wavelet packets. Wavelet packets generalize the compactly supported wavelets of Daubechies and Meyer described in [D]. The algorithms we describe combine the projection of a sequence onto wavelet packet components, the selection of an optimal orthonormal basis subset, some linear or quasilinear processing of the coefficients, and then reconstruction of the transformed sequence.

The present algorithms were inspired by a subband expansion described by Nicolas. Quake obtained graphs of wavelet packets, some of which are included in the appendix to this paper. Coifman and Meyer [CM] obtained analytic formulas for generating wavelet packets. Wickerhauser employed wavelet packets to compress speech signals in [W1], pictures in [W2], and certain matrices in [W3]. Coifman and Wickerhauser described in [CW] the general relationship between a wavelet packet's index and its principal frequency.

The projection of a vector in R^N onto wavelet packet components has complexity $O(N \log N)$, much like the discrete fast Fourier transform (or FFT). Significant differences from FFT are that it is a real-valued algorithm, and that it produces a tree of $N \log N$ coefficients. These correspond to windowed spectral transforms at all dyadic window widths, using smooth windows. From them we may select more than 2^N orthogonal representations. It is also possible to conjugate the wavelet packet algorithm by FFT and obtain all the dyadic windowed Fourier transforms for a vector at once. These windows will have different shapes at each scale, which is known to be necessary for orthogonality.

We describe the notion of a best basis, introduced in [W1] and [CW]. We mention some measures of information, notably Shannon-Weaver entropy, which we can minimize over a collection of bases. Some of these describe quite generally the complexity of transmitting a sequence, or of numerical operations involving it. We can optimize the basis with respect to which we perform a particular operation. In some cases, this

Research supported in part by ONR Grant N00014-88-K0020

can drastically reduce the number of computations or transmitted bits needed for a given degree of accuracy. The search for a minimum will have complexity $0(N \log N)$.

1. Wavelet Packet Analysis

Roughly speaking, a *wavelet packet* ψ is a square integrable modulated wave form with mean 0, well localized in both position and frequency. It may be assigned three parameters: frequency, scale, and position. The first and third may be taken to be the centers of mass of $|\psi|^2$ and $|\hat{\psi}|^2$, where $\hat{\psi}$ is the Fourier transform of ψ. The second might be taken to be a characteristic width of $|\psi|^2$, or equivalently the uncertainty in the position. By Heisenberg's principle, it is also the reciprocal of the uncertainty in the frequency.

Examples of modulated waveforms are easy to construct. Let ϕ be any "sufficiently nice" function with mean 0, and define the modulation, dilation, and translation operators by $\mu_f\phi(t) = e^{ift}\phi(t)$, $\delta_s\phi(t) = s^{1/2}\phi(st)$, and $\tau_p\phi(t) = \phi(t-p)$, respectively. Then the collection of dilated, translated, and modulated ϕ's forms a family of wavelet packets with parameters f, s, p. These transformations conserve energy, so the waveforms can be normalized to be unit vectors in L^2. The component of a function x at f, s, p is the inner product of x with the modulated wave form whose parameters are f, s, p. If it is large, we may conclude that x has considerable energy near frequency f, position p, and scale s.

Definition of Wavelet Packets. We introduce a new class of orthonormal bases of $L^2(\mathbf{R}^n)$ by constructing a "library" of modulated wave forms out of which various bases can be extracted. In particular, the wavelet basis, the Walsh functions, and rapidly oscillating "wavelet packet" bases are obtained.

We'll use the notation and terminology of [D], whose results we shall assume.

We are given an exact quadrature mirror filter $h(n)$ satisfying the conditions of Theorem (3.6) in [D], p. 964, i.e.

$$\sum h(n-2k)h(n-2\ell) = \delta_{k,\ell}\ , \ \sum h(n) = \sqrt{2}.$$

We let $g_k = h_{k+1}(-1)^k$ and define the operations F_i on $\ell^2(\mathbf{Z})$ into "$\ell^2(2\mathbf{Z})$"

$$\begin{aligned} F_0\{s_k\}(i) &= 2\sum s_k h_{k-2i} \\ F_1\{s_k\}(i) &= 2\sum s_k g_{k-2i}. \end{aligned} \tag{1.0}$$

The map $\mathbf{F}(s_k) = F_0(s_k) \oplus F_1(s_k) \in \ell^2(2\mathbf{Z}) \oplus \ell^2(2\mathbf{Z})$ is orthogonal and

$$F_0^*F_0 + F_1^*F_1 = I \tag{1.1}$$

We now define the following sequence of functions.

$$\begin{cases} W_{2n}(x) = \sqrt{2}\sum h_k W_n(2x-k) \\ W_{2n+1}(x) = \sqrt{2}\sum g_k W_n(2x-k) \end{cases}. \tag{1.2}$$

Clearly the function $W_0(x)$ can be identified with the function φ in [D] and W_1 with the function ψ.

Let us define $m_0(\xi) = \frac{1}{\sqrt{2}}\sum h_k e^{-ik\xi}$ and

$$m_1(\xi) = -e^{i\xi}\bar{m}_0(\xi+\pi) = \frac{1}{\sqrt{2}}\sum g_k e^{ik\xi}$$

Remark. The quadrature mirror condition on the operation $\mathbf{F} = (F_0, F_1)$ is equivalent to the unitarity of the matrix

$$\mathcal{M} = \begin{bmatrix} m_0(\xi) & m_1(\xi) \\ m_0(\xi+\pi) & m_1(\xi+\pi) \end{bmatrix}$$

Taking Fourier transform of (1.2) when $n = 0$ we get

$$\hat{W}_0(\xi) = m_0(\xi/2)\hat{W}_0(\xi/2)$$

i.e.,

$$\hat{W}_0(\xi) = \prod_{j=1}^{\infty} m_0(\xi/2^j)$$

and

$$\hat{W}_1(\xi) = m_1(\xi/2)\hat{W}_0(\xi/2) = m_1(\xi/2)m_0(\xi/4)m_0(\xi/2^3)\cdots$$

More generally, the relations (1.2) are equivalent to

$$\hat{W}_n(\xi) = \prod_{j=1}^{\infty} m_{\varepsilon_j}(\xi/2^j) \tag{1.3}$$

and $n = \sum_{j=1}^{\infty} \varepsilon_j 2^{j-1}$ ($\varepsilon_j = 0$ or 1).

We can rewrite (1.1) as follows.

$$\begin{aligned} W_{2n}(x-\ell) &= \sqrt{2}\sum h_{j-2\ell} W_n(2x-j) = F_0\{W_n(2x-j)\}(\ell) \\ W_{2n+1}(x-\ell) &= \sqrt{2}\sum g_{j-2\ell} W_n(2x-j) = F_1\{W_n(2x-j)\}(\ell) \end{aligned} \tag{1.4}$$

where $W_n(2x-j)$ is viewed as a sequence in j for (x,n) fixed. Using (1.1) we find:

$$\boxed{W_n(x-j) = \sqrt{2}\sum_i h_{j-2i} W_{2n}\left(\frac{x}{2}-i\right) + g_{j-2i} W_{2n+1}\left(\frac{x}{2}-i\right).} \tag{1.5}$$

In the case $n = 0$ we obtain:

$$W_0(x-k) = \sqrt{2}\sum h_{k-2i} W_0\left(\frac{x}{2}-i\right) + g_{k-2i} W_1\left(\frac{x}{2}-i\right) \tag{1.6}$$

from which we deduce the usual decomposition of a function f in the space Ω_0 (V_0 in [D]) i.e., a function f of the form

$$\begin{aligned} f(x) &= \sum s_k^0 W_0(x-k) \\ &= \sqrt{2}\sum(\sum s_k^0 h_{k-2i}) W_0\left(\frac{x}{2}-i\right) + \sqrt{2}\sum(\sum s_k^0 g_{k-2i}) W_1\left(\frac{x}{2}-i\right) \\ &= \sum \frac{1}{\sqrt{2}} F_0(s_k^0)(i) W_0\left(\frac{x}{2}-i\right) + \sum \frac{1}{\sqrt{2}} F_1(s_k^0)(i) W_1\left(\frac{x}{2}-i\right) \end{aligned}$$

More generally, if we define

$$\Omega_n = \{f : f = \sum \omega_k W_n(x-k)\}. \tag{1.7}$$

We find

$$f(x) = \sum \frac{1}{\sqrt{2}} F_0(\omega_k)(i) W_{2n}\left(\frac{x}{2}-i\right) + \sum \frac{1}{\sqrt{2}} F_1(\omega_k)(i) W_{2n+1}\left(\frac{x}{2}-i\right) \tag{1.8}$$

or

$$\sqrt{2} f(2x) = h + g \qquad h \in \Omega_{2n} \; g \in \Omega_{2n+1}$$

We now prove

Theorem (1.1). *The functions $W_n(x-k)$ form an orthonormal basis of $L^2(\mathbf{R})$.*

Proof. We proceed by induction on n, assuming that $W_n(x-k)$ form an orthonormal set of functions and, proving that, $W_{2n}(x-k), W_{2n+1}(x-k)$ form an orthonormal set.

By assumption $\|\sqrt{2}f(2x)\|_2^2 = \sum \omega_k^2$ if $f \in \Omega_n$ from the quadrature mirror condition on (F_0, F_1) we get

$$\sum \omega_k^2 = \sum F_0(\omega_k)(i)^2 + F_1(\omega_k)(i)^2.$$

Since $F_0(\omega_k)(i) = \mu_i,\ F_0(\omega_k)(i) = \nu_i$ can be chosen as two arbitrary sequences of ℓ^2 (arising from $\omega = F_0^*\mu_i + F_1^*\nu_i$) it follows that

$$\int |\sum \mu_i W_{2n}(x-i) + \sum \nu_i W_{2n+1}(x-i)|^2 = \sum \mu_i^2 + \sum \nu_i^2$$

which is equivalent to $W_{2n}(x-i), W_{2n+1}(x-j)$ being an orthonormal set of functions. □

Let us now define $\delta f = \sqrt{2}f(2x)$. Formula (1.8) shows that $\delta\Omega_n = \Omega_{2n} \oplus \Omega_{2n+1}$ as an orthogonal sum or,

(1.9)
$$\begin{aligned}
\delta\Omega_0 - \Omega_0 &= \Omega_1 \\
\delta^2\Omega_0 - \delta\Omega_0 &= \delta\Omega_1 = \Omega_2 \oplus \Omega_3 \\
\delta^3\Omega_0 - \delta^2\Omega_0 &= \delta\Omega_2 \oplus \delta\Omega_3 = \Omega_4 \oplus \Omega_5 \oplus \Omega_6 \oplus \Omega_7 \text{ or} \\
\delta^k\Omega_0 - \delta^{k-1}\Omega_0 &= \Omega_{2^{k-1}} \oplus \Omega_{2^{k-1}+1} \cdots \oplus \Omega_{2^k-1}
\end{aligned}$$

and

$$\delta^k\Omega_0 = \Omega_0 \oplus \Omega_1 \oplus \cdots \oplus \Omega_{2^k-1}$$

More generally, we let $\mathcal{W}_k = \delta^{k+1}\Omega_0 - \delta^k\Omega_0 = \delta^k\Omega_1 = \delta^k\mathcal{W}_1$. Therefore we have

Proposition (1.1).

$$\mathcal{W}_k = \delta^k W_1 = \Omega_{2^k} \oplus \Omega_{2^k+1} \oplus \cdots \oplus \Omega_{2^{k+1}-1}.$$

Alternatively, the functions

$$W_n(x-j) \qquad j \in \mathbf{Z}\ 2^k \le n < 2^{k+1}$$

form an orthonormal basis of $\mathcal{W}_k$.

Since the spaces $\mathcal{W}_k$ are mutually orthogonal and span $L^2(\mathbf{R})$ see [D], it follows that $W_n(x-j)$ are complete.

Numerical expansion of functions in wavelet packet bases. Earlier [CW] we introduced an indexing notation for wavelet packets which we shall use here as well. Order the frequency, scale, and position parameters as (f, s, p), and set $w_{0,0,0}(t) = 2^{L/2}W_0(2^L t)$ for some fixed integer L. Define recursively $w_{2f,0,0}(t) = F_0 w_{f,0,0}(t)$ and $w_{2f+1,0,0}(t) = F_1 w_{f,0,0}(t)$, for $f = 0, 1, \ldots$. The integer f' is approximately the center of energy of $\hat{w}_{f,0,0}$, where f' is the Gray code image of f. The position is set to p by the equation $w_{f,0,p}(t) = \tau_p w_{f,0,0}(t) = w_{f,0,0}(t-p)$, and the scale to s by $w_{f,s,p}(t) = \delta_{2^{-s}} w_{f,0,p}(t) = 2^{-s/2} w_{f,0,0}(2^{-s}t - p)$.

We may approximate a uniformly continuous function $x \in L^2(\mathbf{R})$ in the uniform norm topology to arbitary accuracy by sums of orthogonal compactly supported bumps. We have

$$\sup_t \left| x(t) - 2^L \sum_p \langle x, w_{0,0,p}\rangle w_{0,0,p}(t) \right| \to 0 \qquad \text{as } L \to \infty.$$

Conversely, for sufficiently large L, the evaluations $x(p2^{-L})$ approximate well the l^2 sequence of inner products $x_p = 2^{L/2}\langle x, w_{0,0,p}\rangle$.

Suppose in addition that $x \in L^2(\mathbf{R}^n)$ has d uniformly continuous derivatives, for $d \geq 0$. If $\phi \in L^2(\mathbf{R})$ satisfies the conditions

$$\int_R \phi(t)\,dt = 1, \qquad \int_{\mathbf{R}} t^m \phi(t)\,dt = 0 \quad \text{if } 0 < m < d, \text{ and} \qquad \int_R t^d \phi(t)\,dt < \infty,$$

then Taylor's theorem implies that the discrete values $x(k2^{-\nu})$ are very good approximations to the inner products $\langle x, \phi_{-\nu,k}\rangle$ for $k \in \mathbf{Z}^n$ and $\nu \in \mathbf{N}$, where $\phi_{-\nu,k}(t) = 2^{-\nu}\phi(2^{-\nu}t - k)$. We obtain an estimate for the rate of convergence of evaluations to inner products:

$$\sup_{t \in I_{\nu,k}} |\langle x, \phi_{-\nu,k}\rangle - x(t)| < C2^{-\nu d},$$

where $I_{\nu,k} = \times \prod_{i=1}^{n} [2^{-\nu}k_i, 2^{-\nu}(k_i+1)[$, and $0 < C < \infty$ may be chosen independently of ν and k. For ϕ one may use w_{000}. The vanishing moment properties are obtained with appropriate quadrature mirror filters.

In numerical applications there is a limit to precision, say ϵ. This determines a minimum grid size dependent on the smoothness of x and the number of vanishing moments of ϕ. To this precision, the inner products $\langle x, w_{0,0,i}\rangle$ may be replaced by evaluations of x. From these the other wavelet packet coefficients $\langle x, w_{f,s,p}\rangle$ are computed recursively for $s > 0$, $f \geq 0$, and integer p. By transposition,

$$\langle x, w_{2f,s+1,p}\rangle = \sum_j h_j \langle x, w_{f,s,2p+j}\rangle$$
$$\langle x, w_{2f+1,s+1,p}\rangle = \sum_j g_j \langle x, w_{f,s,2p+j}\rangle$$

These recurrences also have periodized analogs. The parameter ranges then become $0 \leq f < 2^s$, and $0 \leq p < 2^{L-s}$, for $0 \leq s \leq L$.

Library of rapidly constructible functions. Wavelet packets form a library of functions. There are infinitely many of them in the continuum limit, but their approximations by vectors in $\mathbf{R}^N$ form a set of $N \log N$ vectors. The vectors are arranged in a homogeneous tree, with any two disjoint maximal subtrees spanning orthogonal subspaces. A useful picture of the tree of wavelet packet coefficients is that of a rectangle of coefficients. The row number within the rectangle indexes the scale of the wavelet packets listed therein. The column number indexes both the frequency and position parameters. We may choose to group the wavelet packets either by frequency or by position. The first method leads to more efficient implementations, but the second yields a more intuitive picture. Grouping by position fills each row of the rectangle with adjacent windowed spectral transforms, with the window size determined by the row number and the window position corresponding to the location of the group. The frequency parameter increases within the group.

We will describe an algorithm to produce a rectangle in which coefficients are grouped by frequency, since this is simpler and since the transformation to the other form is evident. For definiteness, consider a function defined at 8 points $\{x_1, \ldots, x_8\}$, i.e., a vector in $\mathbf{R}^8$. We may develop the (periodized) wavelet packet coefficients of this function by filling out the following rectangle:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
s_1	s_2	s_3	s_4	d_1	d_2	d_3	d_4
ss_1	ss_2	ds_1	ds_2	sd_1	sd_2	dd_1	dd_2
sss_1	dss_1	sds_1	dds_1	ssd_1	dsd_1	sdd_1	ddd_1

Figure 1. A rectangle of wavelet packet coefficients.

Each row is computed from the row above it by one application of either F_0 or F_1, which we think of as "summing" (s) or "differencing" (d) operations, respectively. Thus, for example the subblock $\{ss_1, ss_2\}$ is obtained by convolution-decimation of $\{s_1, s_2, s_3, s_4\}$ with F_0, while $\{ds_1, ds_2\}$ comes from similar convolution-decimation with F_1. In the simplest case, where we use the Haar filters $h = \{\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\}$ and $g = \{\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\}$, we have in particular $ss_1 = \frac{1}{\sqrt{2}}(s_1 + s_2)$, $ss_2 = \frac{1}{\sqrt{2}}(s_3 + s_4)$, $ds_1 = \frac{1}{\sqrt{2}}(s_1 - s_2)$, and $ds_2 = \frac{1}{\sqrt{2}}(s_3 - s_4)$. The two daughter s and d subblocks on the $n+1$st row are determined by their mutual parent on the nth row, which conversely is determined by them through the adjoint anticonvolution.

Reconstructing the nth row from the $n+1$st row consists of applying F_0^* to the left daughter and F_1^* to the right daughter, then summing the images into the parent. In this manner, we generated the graphs of the functions which are included in the appendix. We used a rectangle of size 1024×10 to obtain 1024-point approximations. We filled the rectangle with 0's except for a single 1, then applied the deconvolutions F_0^* and F_1^* up to 10 times in various orders, so as to generate a vector of length 1024. This vector approximates one of the 10240 wavelet packets in R^{1024}. The details of this reconstruction determine the frequency, scale, and location parameters.

From this rectangle, we may choose subsets of N coefficients which correspond to orthonormal bases for $\mathbf{R}^N$. For example, the subset corresponding to the labelled boxes in Figure 2 is the wavelet basis.

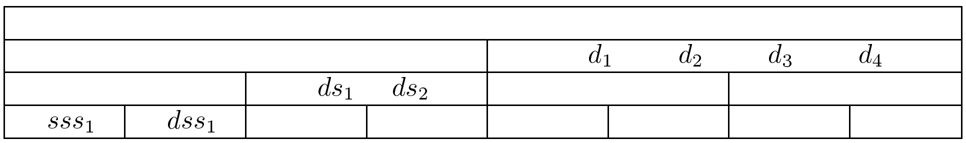

Figure 2. The wavelet basis.

Figures 3 and 4 give other orthonormal basis subsets.

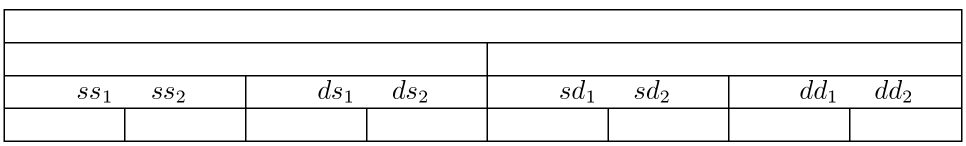

Figure 3. A subband basis.

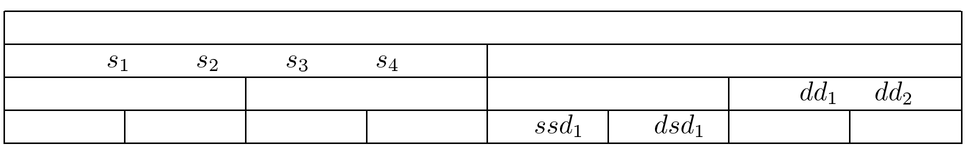

Figure 4. An orthonormal basis subset.

The boxes of coefficients in the rectangle have a natural binary tree structure. Each box is a direct sum of its two children. Call a subset of the rectangle a *graph* if it contains only whole boxes and each column of the rectangle has exactly one element. We have the following:

Proposition. *Every graph is an orthonormal basis subset.*

The number of graphs may be counted by induction. If $N = 2^L$, let A_L be the number of graphs in the coefficient rectangle of N columns and L rows. Then $A_0 = 1$ and we have the relation $A_{L+1} = 1 + A_L^2$, which implies that $A_{L+1} > 2^{2^L} = 2^N$.

Influence of the QMFs. Since wavelet packets are limits of repeated convolutions by perfect reconstruction quadrature mirror filters, the choice of filter influences their various properties, including smoothness and number of vanishing moments. Good filters exist with only a few coefficients, i.e., less than 25. Dual pairs of filters may be found in which the coefficients are dyadic rationals, making convolution on binary computers very fast. Longer filters have more degrees of freedom, and may be optimized for smoothness, number of vanishing moments, attenuation of certain frequencies, or other useful properties. The cost is computational complexity, which grows with filter length.

Orthogonal projection and reconstruction from partial or redundant coefficients. The maps $F_0^*F_0$ and $F_1^*F_1$ are orthogonal projections. Iteration of these projections divides the original Hilbert space into a tree of orthogonal subspaces generalizing multiresolution approximations. This is easiest to describe in the periodic case, where we can count dimensions. A vector in $\mathbf{R}^N$ injects into the space $\mathbf{R}^{N\log N}$ of wavelet packet components. From a basis subset of just N of these components, the vector with those components may be constructed by the adjoint anticonvolutions. If we reconstruct from only a part of the basis subset, say N' of them, we obtain an orthogonal projection onto an N' dimensional subspace of $\mathbf{R}^N$. Thus, given any *a priori* knowledge of the importance of certain wavelet packet coefficients in a signal, we have a least-squares projection algorithm from a given signal onto these coefficients. Counting operations shows that the complexity is $O(N\log N)$.

The reconstruction algorithm also provides a projection of $\mathbf{R}^{N\log N}$ onto a rank-N subspace. We consider $\mathbf{R}^{N\log N}$ to be an ordered stack of $\log N$ rows, each holding the wavelet packet coefficients at a particular scale. We reconstruct a vector by summing the reconstructions from each of the rows, then rescaling by the factor $(\log N)^{-1}$. Finally, we reexpand the vector into its $N\log N$ wavelet packet coefficients. It is easily seen that this is an orthogonal projection, and that it may be computed in $O(N\log N)$ operations.

Analytic interpretation. Although approximations by vectors in R^N will always be used for numerical algorithms, the analytic properties of wavelet packets derive from properties of the continuum limits of the filter convolutions. We shall use here the notation defined above for the wavelet packet $w_{f,s,p}(t)$.

For each f, the Fourier transform $\hat{w}_{f,0,0}$ is a smooth, rapidly decreasing function with a principal bump near f'. The relation between f and f' is explained below. In general, there will be at least $O(\log f)$ auxiliary bumps of lower energy near other frequencies. By orthogonality, the collection $\{|\hat{w}_{f,0,0}|^2 : f\in\mathbf{Z}\}$ forms a partition of unity in the frequency variable.

The center of energy of $\hat{w}_{f,0,0}$ is at f', where $f\mapsto f'$ is a permutation of the integers defined by the following recurrence: $f'=f$ if $f=0$ or $f=1$, and

$$(2f+k)' = \begin{cases} f'+k, & \text{if } f' \text{ is even,}\\ f'+(1-k), & \text{if } f' \text{ is odd.}\end{cases}$$

This permutation is the Gray code transformation, which may be described by $f_j = f'_{j+1} + f'_j \pmod 2$, where f_j is the jth binary digit of f. By interchanging filters p and q appropriately, successive convolutions will find inner products of a vector with $w_{f',0,0}$, which are arranged monotonically by "main" frequency.

Wavelet packets generalize discrete compactly supported wavelets. The wavelet packet $w_{1,0,0}$ may be used as a "mother wavelet" ψ. Dyadic dilations and integer translations of ψ form an orthonormal basis of L^2, which is also an unconditional basis of the common function spaces, as described in [M]. This basis is part of the rectangle or tree above. Explicitly, for a periodic vector of length 2^L, it is the set $\{w_{2^n,n-L,p} : 0\le n\le L, 0\le p<2^{n-L}\}\cup\{w_{0,0,0}\}$. Alternatively, it is the set of coefficients in the rectangle of Figure 1 which consist of a single leading d followed by only s's, together with the lower left-hand corner coefficient which consists of L s's.

The multiresolution or wavelet decomposition is a particular descending chain of maximal subtrees in our picture. Let V_s be $\langle w_{0,s,p}, p\in\mathbf{Z}\rangle$, namely the linear span of the integer translates of $w_{0,s,0}$. Then the recurrence relation among the w's gives that $V_s\subset V_{s-1}$, and the two-sided chain of subspaces is a multiresolution decomposition of L^2 based on the bump function ϕ. The quotient $W_s = V_s^\perp\cap V_{s-1}$ is the linear span $\langle w_{1,s,p}, p\in\mathbf{Z}\rangle$. The collection $\{W_s, s\in\mathbf{Z}\}$ is just a single basis subset. The other subspaces in the tree of wavelet packet coefficients constitute a refinement of this decomposition.

2. Algorithms For Compression

Let x be a vector in $\mathbf{R}^N$ with coordinates determined up to some fixed precision. We wish to represent x with fewer coefficients, as a linear combination of elements of our wavelet packet library. We will discard components with negligibly small amplitude by some criterion, and we will try to arrange that the resulting approximation differs minimally from the original.

This procedure is an orthogonal projection of x onto a lower dimensional subspace, and is therefore linear. The choice of subspace will contain some of the information lost by the projection. Any choice of library will result in some improvement in the efficiency of the representation. We can adjust our library, however, to take advantage of *a priori* knowledge about the signal, such as its bandwidth, or the relative importance of certain frequencies.

For a given library, we seek the most efficient representation of x by trying to minimize the information content of the representation. There are several ways to measure this, depending upon the application: we list a few below.

Measures of information. Define an *additive measure of information* on l^2 to be a functional M satisfying $M(x \times y) = M(x) + M(y)$, and $M(0) = 0$. Here $l^2 \times l^2 \cong l^2$; any fixed isomorphism will do. Any such measure may be minimized over a family of orthogonal bases of $\mathbf{R}^N \subset l^2$. Furthermore, since $\mathbf{R}^N$ factors into a cartesian product of N one-dimensional spaces, we see that evaluating an additive measure of information requires $O(N)$ operations. Three useful examples of M are listed below.

* *Entropy.* The Shannon-Weaver entropy of a sequence $x = \{x_j\}$ is $\mathcal{H}(x) = -\sum_j p_j \log p_j$, where $p_j = \frac{|x_j|^2}{\|x\|^2}$. This is not an additive measure of information. However, the $l^2 \log l^2$ "norm" $\lambda(x) = -\sum_j |x_j|^2 \log |x_j|^2$ is. The relation $\mathcal{H}(x) = \|x\|^{-2}\lambda(x) + \log \|x\|^2$ insures that minimizing the latter minimizes the former. For this entropy, $\exp \mathcal{H}(x)$ is related to the number of coefficients needed to represent the signal to a fixed accuracy.
* *Number above a threshold.* Set an arbitrary threshold ϵ and count the elements in the sequence x whose absolute value exceeds ϵ. This is an additive measure of information. It gives the number of coefficients needed to transmit the signal to accuracy ϵ.
* *Bit counts.* Choose an arbitrary $\epsilon > 0$ and count the (binary) digits in $\lfloor |x_j|/\epsilon \rfloor$. Summing over j gives an additive measure of information. It corresponds to the number of bits needed to transmit the signal to accuracy ϵ.

Choosing a basis. We search through the family of bases to find the one that minimizes the additive measure of information. Since the measure is additive across cartesian products of subspaces, we can examine orthogonal subspaces independently, minimize locally, then recombine the minimal pieces into a best basis for the whole space. In the wavelet packet case, these decompositions are organized as a homogeneous tree in which a node is the cartesian product of its children. There the search for a global minimum for any M is a sequence of comparisons beween a node and its children, followed by a depth-first search of the tree for the nodes which beat all basis subsets of their descendents. For any M, in a homogeneous tree with N nodes, minimizing takes $O(N \log N)$ operations.

* *Best level.* Define *level m* of the representation of x to be the collection of wavelet packet coefficients of x obtained by applying exactly m convolutions of p and q, in all possible orders. It is easy to see that a level is a basis set. For expansions down to L levels, choose that m for which M is minimal. This is a generalization of traditional subband coding, which would always choose the bottom or Lth level after deciding (in advance) what the optimal filter and number of levels should be. Since subband coding works so well, we expect good results even with this simple algorithm.
* *Restricted best basis.* Given the complete rectangle of wavelet packet coefficients down to some level, exclude certain coefficients for statistical or other reasons. For example, wavelet packets whose main frequencies are above the Nyquist frequency may be ignored. Search the remaining coefficients for the basis subset minimizing M on their span.
* *Best basis.* Search the entire collection of basis subsets for the one in which x has an M-minimal representation.
* *Best basis in the time domain.* We can conjugate the wavelet packet algorithm by the discrete FFT to obtain an order $N \log N$ time-domain algorithm. This is exactly analogous to finding the windowed Fourier transforms at all dyadic scales, with the windows being the scaled bump functions $w_{0,s,0}$ translated appropriately.
* *Best basis in both time and frequency.* Nothing prevents us from searching both the time and frequency domains for the M-minimizing representation. The transmission cost is one extra bit to distinguish between the two methods. The encoding time is tripled, and the decoding time is at most doubled.

Discarding negligible coefficients. Several methods exist for deciding which coefficients in an optimal basis are negligible. Of course, this decision is intimately related to the choice of information measure M.

* *Absolute cutoff.* Fix $\epsilon > 0$, and treat as negligible any coefficient c with $|c| < \epsilon$. The number of these will be maximized if we use "number of coefficients above a threshold" for the measure of information.
* *Relative energy.* Fix $0 < \epsilon \leq 1$, and discard any coefficient c for which $|c|^2 < \epsilon\|x\|^2$. One may also use weights, and local or windowed measures of energy.
* *Entropy criterion.* Since $\exp \mathcal{H}(x)$ is a measure of the number of coefficients needed to determine the signal, we may define the *average energy of a significant coefficient* to be $\|x\|^2 \exp -\mathcal{H}(x)$. This has the convenient form $\exp(-\lambda(x)/\|x\|^2)$ in terms of the $l^2 \log l^2$ norm λ. We may choose $0 < \epsilon \leq 1$, and declare negligible any coefficient c for which $|c|^2 < \epsilon \exp(-\lambda(x)/\|x\|^2)$. The appropriate entropy to use is the minimum achieved by a basis selection using $M = \lambda$. This will maximize the cutoff energy and therefore minimize the number of retained coefficients for each ϵ.
* *Decreasing rearrangements and fixed percentages.* If we are allowed to retain only a fixed fraction of the bits in the original signal, we may sort the coefficients of any optimal representation in decreasing order of absolute value and then keep only as many of the largest as we can afford, discarding the rest. An optimal basis for this method is one in which a decreasing rearrangement decreases at the maximal rate. Observe that Shannon-Weaver entropy, and every additive measure of information, is invariant under rearrangements.

3. Some Results

Various combinations of the above techniques have been applied to acoustic signals including speech and music, seismic data, fluid velocity measurements, and digitized pictures. In addition, some pseudodifferential operators written as matrices were "compressed" by finding their sparsest matrix representations.

Acoustic signals. Speech signals were recorded and compressed by various methods. As described in [W1], the sound quality degrades gracefully down to bit rates around 1 kbps.

* *Best level with many samples.* We start with speech sampled at 22050 8-bit linear samples per second, or 176.4 kbps. We construct the best-level representation from among 15 levels of a 32768 sample segment, using the threshold criterion. At 14 kbps, distortion is essentially undetectable.
* *Best basis with many samples.* Again starting with 176.4 kbps sampling, with phrases of 32768 8-bit samples, the best basis representation by the entropy method allows us to discard all but 4.5 kbps with good quality, although there is some distortion. No filtering or other spectrum modification was done.
* *Best basis with few samples.* Using 8012 samples per second and recording 8-bit μ-law samples with a standard CODEC gives an initial rate of 64.1 kbps. Using 256 sample windows and the best basis by the entropy criterion, this reduces to 14 kbps with very good quality and 4 kbps with reasonable quality.
* *Dependence upon filter length.* Longer filters give better compression with less distortion. This is particularly noticeable in the best level experiments.

Preliminary results using L^2 distortion estimates suggest that best-basis compression algorithms are competitive with the state of the art in signal processing. Further experiments, for example to quantify the subjective distortion, are needed to judge the practical value of the method.

Pictures. We prepared a digitized ray-traced image rich with textures and varying scales. The resolution was 256×256 pixels, with 8 bits of gray level per pixel. The picture was expanded in coefficients with respect to two-dimensional tensor-product wavelet packets.

* *Best basis, keep specified fraction.* After choosing the best basis by the entropy criterion, the coefficients were sorted in decreasing order of absolute value. The picture was reconstructed from a specified fraction of the largest coefficients. Distortion was unnoticeable above 1 bit per pixel, becoming objectionable around 0.4 bits per pixel.
* *Dependence upon filter length.* Longer filters give better results, particularly at features with sharp curved boundaries. The low resolution contributed to poor results at low bit rates.
* *Visibility criterion.* We used the luminance visibility table of the draft JPEG picture compression standard. Wavelet packet amplitudes with a given frequency were weighted by the visibility coefficients of this table, which is intended to be used with block discrete cosine transforms. This weighting ignores scale,

but even so it improves the subjective distortion at high compression ratios. The results suggest further experiments to determine the actual visibility of wavelet packets.

Flow velocity. One-dimensional data from a hot wire velocity probe was expressed in the best basis and reconstructed from a small number of the largest coefficients. Major features of the signal could be recognized at compression ratios of 50 to 700. Selecting wavelet packet coefficients by frequency resulted in incremental reconstructions showing features at different scales.

Turbulence pictures. Gray scale digitized photos of turbulent flow were treated as the pictures above. By reconstructing pictures from certain ranges of coefficients, particular features are isolated. These features may be isolated by scale, frequency, and position, the three indices of wavelet packet coefficients.

4. Comparison With Previous Results

Wavelet methods. Wavelet packet coefficients represent signals at least as efficiently as wavelet coefficients. The search for the best wavelet packet basis includes the wavelet basis. However, the wavelet packet algorithm has complexity $O(N \log N)$ versus $O(N)$ for wavelets. Certain operators, i.e., those with smooth but oscillatory kernels, will not compress in the wavelet representation. Likewise, smooth oscillatory signals like speech or music will compress significantly better in the wavelet packet basis.

Sub-band coding methods. With perfect reconstruction filters, this is a special case of wavelet packet methods, in which we always choose the bottom level. The best-level wavelet packet algorithm must be at least as good, although experiments with speech suggest that 2 or 3 of the levels will be chosen quite consistently.

References

R. R. Coifman and M. V. Wickerhauser, *Best-adapted wavelet packet bases*, Yale University preprint, February, 1990.

R. R. Coifman et Yves Meyer, *Nouvelles bases orthonormées de $L^2(\mathbf{R})$ ayant la structure du système de Walsh*, preprint, Yale University, 1989.

Ingrid Daubechies, *Orthonormal bases of compactly supported wavelets*, Communications on Pure and Applied Mathematics **XLI** (1988), 909–996.

Yves Meyer, *De la recherche petroliére à la geometrie des espaces de Banach en passant par les paraproduits*, Seminaire equations aux derivées partielles, 1985–1986, École Polytechnique, Palaiseaux.

M. V. Wickerhauser, *Acoustic signal compression with wavelet packets*, preprint, Yale University, August, 1989.

———, *Picture compression by best-basis wavelet packet coding*, preprint, Yale University, January, 1990.

———, *Nonstandard matrix multiplication*, preprint, Yale University, April, 1990.

Current addresses (1992): *CEREMADE, Université de Paris–Dauphine; †Department of Physics, Cambridge University; ‡Department of Mathematics, Washington University.