Proceedings of Symposia in Applied Mathematics
Volume **00**, 0000

Best-adapted wavelet packet bases

MLADEN VICTOR WICKERHAUSER

March 7, 1993

Abstract. This paper is a review of the construction of orthogonal wavelet packets, using the quadrature mirror filter algorithm slightly generalized to the case of $p \geq 2$ wavelets and scaling functions. It is part of the AMS short course on "Wavelets and Applications" held in San Antonio, 11-12 January 1993.

Introduction. We begin with a classical reproducing formula for functions $f \in H^2$, the Hardy space of square-integrable functions whose Fourier transforms vanish on the negative half-line:

$$\text{If } c = 2\pi \int_0^\infty \frac{|\hat{\psi}(\xi)|^2}{|\xi|}\, d\xi < \infty \quad \text{and} \quad T_f(a,b) = \int_{\mathbf{R}} f(x)\bar{\psi}(ax+b)\, dx,$$
$$\text{then } f(x) = \frac{1}{c} \iint_{\mathbf{R}\times\mathbf{R}^+} Wf(a,b)\psi(ax+b)\, da\, db$$

This formula was studied by Calderón in the 60's and revived by Grossmann and Morlet in their 1984 paper [**GM**]. A function ψ satisfying the *admissibility condition* $c < \infty$ is called a "wavelet," and the map $f \mapsto T_f$ is called the (continuous) wavelet transform. The discrete (dyadic) wavelet transform transform is the restriction $f \mapsto \{T_f(2^j, k),\ j, k \in \mathbf{Z}\}$. A well-established sampling theory [**FJW**], [**Me1**] exists which provides necessary and sufficient conditions on ψ for the discrete transform to be bijective. There are compactly-supported functions ψ with any given degree of smoothness for which the discrete wavelet transform is orthogonal [**D**],[**Ma**]; there is also an orthonormal basis of C^∞ wavelets with exponential decay [**Me2**].

AMS Subject Classification 33C45,42C10,94A11.
This paper is in final form and no version of it will be submitted elsewhere.
Research supported in part by AFOSR Grant F49620-92-J-0106.

The "fast" discrete wavelet transform computes $\{T_f(2^j,k),\ j,k\in\mathbf{Z}\}$ by iterating a pair of operators called quadrature mirror filters (QMFs). Suppose that H, G are bounded linear operators defined by

(QMF)
$$Hz_n=\sum_k h_k z_{2n-k},\quad Gz_n=\sum_k g_k z_{2n-k}\quad \text{from } \ell^2(\mathbf{Z}) \text{ to } \ell^2(\mathbf{Z});$$
$$Hf(x)=\sum_k h_k f(2x-k),\quad Gf(x)=\sum_k g_k z(2x-k)\quad \text{from } L^2(\mathbf{R}) \text{ to } L^2(\mathbf{R}).$$

They are QMFs if $HG^*=0$, $HH^*=GG^*=I$, and $H^*H\oplus G^*G=I$. It is conventional to assume that $H\mathbf{1}=\sqrt{2}\mathbf{1}$ (where $\mathbf{1}$ is the sequence of 1's) and $G\mathbf{1}=0$, and to call H and G the *low-pass* and *high-pass* filters, respectively. The fast discrete wavelet transform computes coefficients by the "pyramid scheme" depicted in the diagram below:

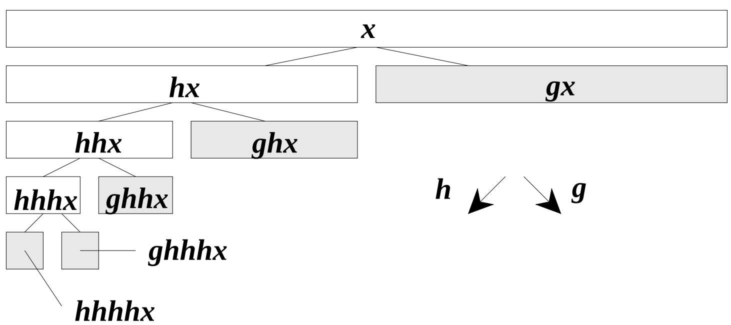

Figure 1.

Pyramid scheme for the "fast" discrete wavelet transform.

The functions underlying the expansion are called "wavelets" and "scaling" functions or "mother" and "father" functions respectively. If $\{h_k\}$ and $\{g_k\}$ are finite sequences, then we have:

Lemma 1. *There is a unique function $\phi\in L^2\cap L^1$ of compact support solving $H\phi=\phi$, $\int\phi=1$.*

Lemma 2. *The function $\psi=G\phi\in L^2\cap L^1$ is a wavelet with compact support.*

Meyer and Mallat introduced the *multiresolution analysis* of $L^2(\mathbf{R})$ (or MRA); it is a sequence of subspaces $\{V_j : j\in\mathbf{Z}\}$ based on a single function ϕ, with $V_j \overset{\text{def}}{=} \operatorname{span}\{\phi(2^jx-k) : k\in\mathbf{Z}\}$ satisfying $j<j' \implies V_j\subset V_{j'}$, $\overline{\bigcup_j V_j}=L^2$, and $\bigcap_j V_j=0$. Given an MRA, we may put $W_j=V_j^\perp\cap V_{j+1}$ to get $L^2=\overline{\bigoplus_j W_j}$. If ϕ is the function in Lemma 1, then $\{V_j\}$ is an MRA and $\psi=G\phi$ is a wavelet. Conversely, it is known [**L**] that if the function ψ has compact support and the

set $\{\psi(2^j x - k) : j, k \in \mathbf{Z}\}$ is an orthonormal basis for L^2, then it comes from an MRA with a compactly-supported ϕ.

There is a fast numerical functional calculus based on the compactly-supported orthonormal wavelet transform. It is implemented by defining orthogonal projections $P_j : f \twoheadrightarrow V_j$ and $Q_j : f \twoheadrightarrow W_j$; these are approximated either by sampling $P_j f(k) = \int f(x)\phi(2^j x - k)\,dx \approx 2^{-j} f(2^{-j}k)$ or by using some higher-order numerical quadrature formula. Thereafter, all computations within the MRA are performed by iterating the low complexity maps H and G on just those coefficients (from $\{P_j f(k) : k \in \mathbf{Z}\}$) which are larger than some threshold depending upon the numerical precision of the calculation.

Coifman, Meyer and Wickerhauser [**CMW**] have described a large "library" of orthonormal bases generalizing the wavelet basis. These bases consist of "wavelet packets" which are superpositions of wavelets and are described efficiently by short sequences of H and G. Wavelet packets come with independent frequency, position and duration parameters and can be used to build individually-adapted orthonormal bases for oscillatory functions and operator kernels.

By using extra filters, it is possible to introduce fast wavelet packet transformations which decimate by arbitrary numbers. Such transformations generalize algorithms which decimate by 2. The method produces new libraries of orthonormal basis vectors. We will describe the *best basis* algorithm for selecting a most efficient representation from this library. The extra generality of p filters is not expensive and in fact may clarify certain points. We will prove that the best-basis algorithm has complexity $O(N \log_p N)$ for a sequence of length N. We will also discuss some of the analytic properties and applications of such representations.

Aperiodic filters and bases in l^2. Consider first the construction of bases on l^2. Let p be a positive integer and introduce p absolutely summable sequences $f_0, \dots, f_{p-1}$ satisfying the properties:

(1) For some $\epsilon > 0$, $\sum_m |f_i(m)||m|^\epsilon < \infty$,
(2) $\sum_m f_0(pm + i) = 1/\sqrt{p}$, for $i = 0, 1, \dots, p-1$, and
(3) $\sum_m f_i(m) f_j(m + kp) = \delta_{i-j}\delta_k$, where δ is the Kronecker symbol.

To these sequences are associated p convolution operators $F_0, \dots, F_{p-1}$ and their adjoints $F_0^*, \dots, F_{p-1}^*$ defined by

$$F_i : l^2 \to l^2, \quad F_i v(k) = \sum_m f_i(m + pk) v(m),$$
$$F_i^* : l^2 \to l^2, \quad F_i^* v(m) = \sum_k f_i(m + pk) v(k).$$

These convolution operators will be called *filters* by analogy with quadrature mirror filters in the case $p = 2$. They have the following properties:

Lemma 2. *For $i, j = 0, 1, \dots, p-1$,*

(1) $F_i F_j^* = 0$, *if $i \neq j$,*
(2) $F_i F_i^* = I$,

(3) *$F_i^* F_i$ is an orthogonal projection of l^2, and for $i \neq j$ the ranges of $F_i^* F_i$ and $F_j^* F_j$ are orthogonal, and*
(4) $F_0^* F_0 + \cdots + F_{p-1}^* F_{p-1} = I$.

Proof. Properties (1) and (2) follow by interchanging the order of summation:

$$\begin{aligned} F_i F_j^* v(k') &= \sum_m \sum_k f_i(m + pk') f_j(m + pk) v(k) \\ &= \sum_k \left(\sum_{m'} f_i(m') f_j(m' + p[k - k']) \right) v(k) \\ &= \sum_k \delta_{i-j} \delta_{k-k'} v(k) = \begin{cases} v(k'), & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases} \end{aligned}$$

For property (3) we use (1) and (2): $F_i^* F_i F_i^* F_i = F_i^* F_i$, and $F_i^* F_i F_j^* F_j = 0$. Orthogonality is shown by transposition: $\langle F_i^* F_i x, F_j^* F_j y \rangle = \langle F_i x, F_i F_j^* F_j y \rangle = \langle F_i x, 0 \rangle = 0$.

To prove (4), let $m_j(\xi) = \sum_k f_j(k) e^{ik\xi}$ be the (bounded, Hölder continuous, periodic) function determined by the filter f_j, for $j = 0, \ldots, p-1$. Then $f_j(k) = \hat{m}_j(k)$ is a real number, and each $F_i^* F_i$ is unitarily equivalent to multiplication by $|m_i|^2$ on $L^2(-\pi, \pi)$.

Now Plancherel's theorem gives

$$\int_0^{2\pi} e^{ilp\xi} m_j(\xi) \bar{m}_{j'}(\xi)\, d\xi = \sum_k f_j(k) f_{j'}(k + lp) = \delta_{j-j'} \delta_l.$$

In particular, $|m_j|^2$ has integral 1, and the Fourier coefficient $(|m_j|^2)\hat{}(lp)$ vanishes if $l \neq 0$. This is equivalent to the average of $|m_j(\xi)|^2$ over $\{\xi, \xi + 2\pi/p, \ldots, \xi + 2\pi(p-1)/p\}$ being identically 1.

The same vanishing is true of the Fourier coefficients of the cross terms $m_j \bar{m}_{j'}$, and for those it also holds when $l = 0$. Thus, the average of $m_j(\xi)\bar{m}_{j'}(\xi)$ over $\{\xi, \xi + 2\pi/p, \ldots, \xi + 2\pi(p-1)/p\}$ vanishes identically. Hence, the conditions on the filters f_i are equivalent to the unitarity of the following matrix:

$$\begin{pmatrix} m_0(\xi) & m_0(\xi + \frac{2\pi}{p}) & \ldots & m_0(\xi + \frac{2\pi(p-1)}{p}) \\ \ldots & \ldots & \ldots & \ldots \\ m_{p-1}(\xi) & \ldots & \ldots & m_{p-1}(\xi + \frac{2\pi(p-1)}{p}) \end{pmatrix}$$

But then $\sum_{k=0}^{p-1} |m_k(\xi)|^2 = 1$ for all ξ. Thus $F_0^* F_0 + \cdots + F_{p-1}^* F_{p-1}$ is unitarily equivalent to multiplication by 1 in $L^2(-\pi, \pi)$, proving (4). □

With this lemma we can decompose l^2 into mutually orthogonal subspaces $W_0^1 \perp \cdots \perp W_{p-1}^1$, where $W_i^1 = F_i^* F_i(l^2)$ for $i = 0, \ldots, p-1$. The map F_i finds the coordinates of a vector with respect to an orthonormal basis of W_i^1. One level of this decomposition is displayed in the figure below:

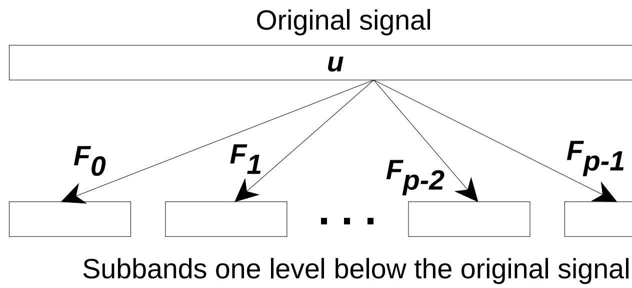

Figure 2.

One level of decomposition into p subspaces.

Since each $F_iW_i^1 = F_i(l^2)$ is another copy of l^2, there is nothing to prevent us from reapplying the filter convolutions recursively. At the mth stage, we obtain $l^2 = W_0^m \perp \cdots \perp W_{p^m-1}^m$, where $W_n^m = F_{n_1}^* \dots F_{n_m}^* F_{n_m} \dots F_{n_1}(l^2)$ and $n_m \dots n_1$ is the radix-p representation of n. The map $F_{n_m} \dots F_{n_1}$ transforms into standard coordinates in W_n^m. For convenience, we will introduce the notations $\mathbf{F}_n^m = F_{n_m} \dots F_{n_1}$, and $\mathbf{F}_n^{m*} = F_{n_1}^* \dots F_{n_m}^*$.

The subspaces W_n^m form a p-ary tree. Every node W_n^m is a parent with p daughters $W_{pn}^{m+1}, \dots, W_{pn+p-1}^{m+1}$. The root of the tree is the original space l^2, which we may label W_0^0 for consistency. Call the whole tree $\mathbf{W}$. This tree is a partially ordered set with minimal element W_0^0, which we call the *root*. We will say that $W' \in \mathbf{W}$ is greater than $W \in \mathbf{W}$ if the unique path between W' and W_0^0 contains W. The set $\{W' : W' \geq W\}$ will be called the *descendants* of W.

Now fix m and suppose w belongs to W_n^m, where $0 \leq n \leq p^m - 1$, and $\mathbf{F}_n^m w = e_k$ is the elementary sequence with 1 in the kth position and 0's elsewhere. The collection of all such w forms an orthonormal basis of l^2 with some remarkable properties. In particular, if $p = 2$ and the filters F_0 and F_1 are taken as low-pass and high-pass quadrature mirror filters, respectively, then the spaces $W_0^m, \dots, W_{2^m-1}^m$ are all the subbands at level m. These have been used for a long time in digital signal processing and compression. An earlier paper [**W**] described experiments with an algorithm for choosing m so as to reduce the bit rate of digitized acoustic signal transmission. This produced good signal quality at rather low bit rates.

The tree contains other orthogonal bases of W_0^0. In fact, it forms a library of bases which may be adapted to classes of functions. The tree structure allows the library to be searched efficiently for the extremum of certain cost functionals.

To every node in $\mathbf{W}$ we associate the subtree of all its descendants. Define a *graph* to be any finite subset of the nodes of $\mathbf{W}$ with the property that the union of the associated subtrees is disjoint and contains a complete level $W_0^m, \ldots, W_{p^m-1}^m$ for some m. For example, the singleton $\{W_0^0\}$ is a graph with $m = 0$. The following may be called the graph theorem.

THEOREM 3. *Every graph corresponds to a decomposition of l^2 into a finite direct sum of orthogonal subspaces.*

PROOF. Every graph is a finite set, of cardinality no more than p^m for the m in the definition. Fix a graph, and suppose that $W_{n_1}^{m_1}$ and $W_{n_2}^{m_2}$ are subspaces corresponding to two nodes. Without loss, suppose that $m_1 \le m_2$. Then $W_{n_2}^{m_2}$ is contained in a subspace $W_n^{m_1}$ for some $n \ne n_1$. Since the subspaces at a given level are orthogonal, we conclude that $W_{n_2}^{m_2} \perp W_{n_1}^{m_1}$.

To show that the decomposition is complete, observe that a node contains the sum of its immediate descendent nodes, or *children*. By induction, it contains the sum of all of the nodes in its subtree. Hence a graph contains the sum of all the subspaces at some level m. But this sum is all of l^2. $\square$

COROLLARY 4. *Graphs are in one-to-one correspondence with finite disjoint covers of* $[0, 1)$ *by p-adic intervals* $I_n^m = p^{-m}[n, n+1)$, $n = 0, 1, \ldots, p^m - 1$.

PROOF. The correspondence is evidently $W_n^m \leftrightarrow I_n^m$. The subtree associated to W_n^m corresponds to all p-adic subintervals of I_n^m. The details are left to the reader. $\square$

This correspondence induces a partial order on graphs. We will say that graph u is greater than or equal to graph v if the cover associated to u is a refinement of the cover associated to v. This partial order has a minimal element $\{W_0^0\}$. For each *maximum level* $L \ge 0$ it also has a maximal element $\{W_0^L, \ldots, W_{p^L-1}^L\}$. Some example graphs are depicted in the figures below:

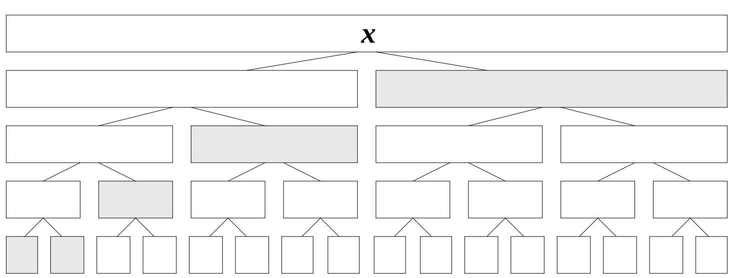

Figure 3.

The "wavelet" graph basis, $p = 2$.

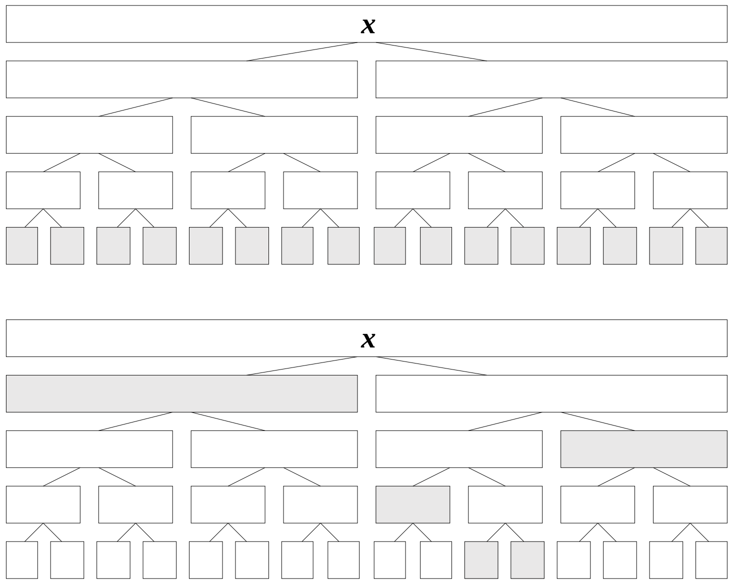

Figure 4.

Some other examples of graph bases, $p = 2$**.**

Analytic properties of graphs: continuous wavelet packets. Each filter F_j (and its adjoint F_j^*) maps the class of rapidly decreasing sequences to itself. Likewise, the projections $\mathbf{F}_n^{m*}\mathbf{F}_n^m$ preserve that class. In practice, we shall consider only finite sequences in l^2. For actual computations the filters must be finitely supported as well. Convolution with such filters preserves the property of finite support. Let the support width of the filters be r, and let z_m be the maximum width of any vector of the form $F_{j_1}^* \ldots F_{j_m}^*(e_k)$. Then $z_0 = 1$ and $z_{m+1} = pz_m + r - p$. By induction, we see that $z_m = p^m + (p^m - 1)(r - p)$.

In [**CMW**] we observed that the basis elements $\mathbf{F}_n^{m*} e_k$ form wave packets over $\mathbf{R}$. Because they are superpositions of Daubechies' compactly-supported *wavelets*, we will call these basis elements *wavelet packets*. A slightly generalized paraphrase of the construction follows. Many of the basic facts we use were proved by Daubechies in [**D**].

Let w be a function defined by $\hat{w}(\xi) = \prod_{j=1}^{\infty} m_0(\xi/p^j)$, where m_0 is the analytic function defined by F_0, as above. Then w has mass 1, decreases rapidly, and is Hölder continuous, as proved in [**D**]. If m_0 is a trigonometric polynomial of degree r, then w is supported in the interval $[-r, r]$. Arranging that w has r continuous derivatives requires m_0 with degree at most $O(r)$. See [**D**] for a discussion of the constant in this relation for $p = 2$. Put $w_0^0 = w$, and define the family of wavelet packets recursively by the formula $w_{pn+j}^{m+1}(t) = \sum_{-\infty}^{\infty} f_j(i) w_n^m(pt-i)$. This produces one function w_n^m for each pair (m,n), where $m = 0, 1, \ldots$ and $n = 0, 1, \ldots, p^m - 1$.

We can renormalize the wavelet packets to a fixed scale p^L. Write

$$w_{n,m,k}^L(t) = p^{(L-m)/2} w_n^m(p^{L-m}t - k).$$

Then $w_{0,0,k}^L$ is a collection of orthonormal functions of mass $p^{L/2}$, concentrated in intervals of size $O(p^{-L})$. This makes them suitable for sampling continuous functions. Let $x(t)$ be any continuous function, and put

$$s_0^0(k) = \langle x, p^{L/2} w_{0,0,k}^L \rangle = \int_{-\infty}^{\infty} x(t) p^L w_0^0(p^L t - k)\, dt.$$

We may use $s_0^0(k)$ as a representative value of $x(t)$ in the interval $I_k^L = p^{-L}[k, k+1)$. The closeness of the approximation to values of x depends, of course, on the smoothness of x. Suppose that x is Hölder continuous with exponent ϵ. Then if t_0 is any point in I_k^L, we have

$$|x(t_0) - s_0^0(k)| = |\int_{I_k^L} (x(t_0) - x(t))\, p^L w_0^0(p^L t - k)\, dt| = O(p^{-\epsilon L}).$$

We can also take advantage of differentiability of x if we construct w_0^0 with vanishing moments. Given d vanishing moments and d derivatives of x, the approximation improves to $|x(t_0) - s_0^0(k)| = O(p^{-dL})$.

The map $x \mapsto s_0^0$ sends $L^2(\mathbf{R})$ to l^2, and pulls back the orthonormal bases of l^2 constructed in the last section. To see this, define $s_n^m(k) = \langle x, w_{n,m,k}^L \rangle$. By interchanging the order of recurrence relation and inner product, we obtain the formula $s_n^m = \mathbf{F}_n^m s_0^0$. Thus, the coordinates $s_n^m(k)$ are coefficients with respect to an orthonormal basis of W_n^m.

The resulting subspaces of $L^2(\mathbf{R})$ form a finer type of multiresolution decomposition than that of Mallat [**Ma**]. The coordinates $s_n^m(k)$ are rapidly computable. As we shall see, they contain a mixture of location and frequency information about x.

Ordering the basis elements. The parameters n, m, k, L in $w_{n,m,k}^L$ have a natural interpretation as frequency, scale, position, and resolution, respectively. However, n is not monotonic with frequency, because our construction yields wavelet packets in the so-called Paley (natural, or p-adic) ordering. The following results show how to permute $n \mapsto n'$ into a frequency-based ordering.

THEOREM 5. *We can choose rapidly decreasing filters* $F_0, \ldots, F_p - 1$ *such that* $w_{n,m,k}^L$ *is concentrated near the interval* I_k^{L-m}, *and* $\hat{w}_{n,m,k}^L$ *is concentrated near the interval* $I_{n'}^m$, *where* $n \mapsto n'$ *is a permutation of the integers.*

PROOF. For the first part, we note that for any family of rapidly decreasing filters, w_0^0 decreases rapidly away from $[0,1)$. The dilate and translate $w_{0,m,k}^L$ of this function to the interval I_k^{L-m} similarly has rapid decrease. Likewise,

$w^L_{n,m,k}$ has the same concentration as $w^L_{0,m,k}$, since all the filters F_i are rapidly decreasing.

The second part may be deduced from the Fourier transform of the recurrence relation:

$$\hat{w}^{m+1}_{pn+j}(\xi) = \left(p^{-1}\sum_k f_j(k)e^{-ix\xi/p}\right)\hat{w}^m_n(\xi/p) = p^{-1}m_j(\xi/p)\hat{w}^m_n(\xi/p),$$

where m_j is the multiplier defined above. Recall that $\sum_{j=0}^{p-1}|m_j(\xi)|^2 \equiv 1$ and that $m_0(0) = 1$. Thus, the periodic functions $|m_j|^2$ form a partition of unity into p functions, with 0 being in the support of m_0 alone.

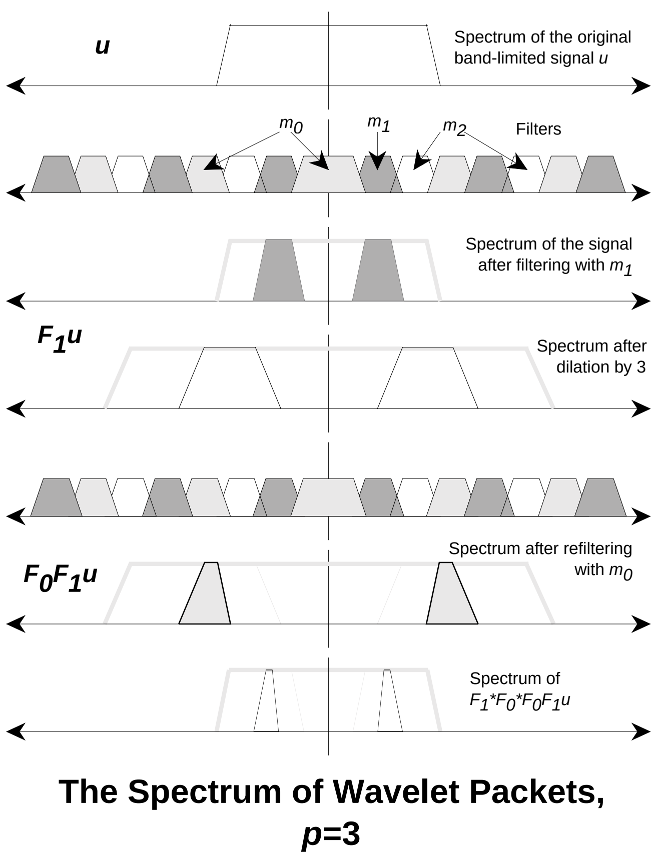

Figure 5.

Successive application of $p = 3$ convolution-decimations.

Now suppose for simplicity that we have chosen filters in such a way that

$|m_j(\xi)| = \sum_{k=-\infty}^{\infty} \chi_{\pm\frac{\pi}{p}[j,j+1)}(\xi - 2\pi k)$. Such m_j may be approximated in $L^2(-\pi,\pi)$ as closely as we like by multipliers arising from exponentially decreasing filters. In this simple case, it is immediate that $\hat{w}_0^0(\xi) = m_0(\xi/p)|_{(-\pi,\pi)}$ is the characteristic function of $(-\pi,\pi)$, so that $\hat{w}_{0,0,0}^L$ is the characteristic function of $(-\pi p^L, \pi p^L)$. Likewise, $\hat{w}_{j,1,0}^L$ is the characteristic function of $\pi p^{L-1}(-j-1,-j] \cup \pi p^{L-1}[j,j+1)$. From the recurrence relation, we see that $\hat{w}_{n,m,0}^L$ will be the characteristic function of the union of the intervals $\pm\pi p^{L-m}[n', n'+1)$, where $n \mapsto n'$ is a permutation. These intervals cover $p^L(-\pi,\pi)$ as $n = 0,\dots,p^m-1$. This arrangement of frequencies is depicted in Figure 5. The permutation $n \mapsto n'$ is given by the recurrence relation

$$n' = n, \quad \text{if } n = 0,\dots,p-1; \quad (np+j)' = \begin{cases} n'p + j, & \text{if } n' \text{ is even,} \\ n'p + (p-1) - j, & \text{if } n' \text{ is odd.} \end{cases}$$

Write n_j for the jth digit of n in radix p, numbering from the least significant. Set $n_m = 0$ if n has fewer than m digits. Then the recurrence relation implies that $n_j = \pi(n'_{j+1}, n'_j)$, where

$$\pi(x,y) = \begin{cases} y, & \text{if } x \text{ is even,} \\ p-1-y, & \text{if } x \text{ is odd..} \end{cases}$$

For each value of the first variable, π is a permutation of the set $\{0,\dots,p-1\}$ in the second variable. Thus the map $n' \mapsto n$ and its inverse $n \mapsto n'$ are permutations of the integers. It is not hard to see that these are permutations of order 2 if p happens to be odd. Otherwise they have infinite order, as may be seen by considering an increasing sequence of integers n' all of which have only odd digits in radix p. $\square$

Corollary 6. *With filters $F_0,\dots,F_p - 1$ chosen as above, we can modify the recurrence relation for $w_{n,m,k}^L$ such that $\hat{w}_{n,m,k}^L$ is concentrated near the interval I_n^m.*

Proof. Simply reorder the functions w_n^m by using the alternative recurrence relation:

$$w_{pn+j}^{m+1}(t) = \begin{cases} \sum_k f_j(k) w_n^m(pt-k), & \text{if } n \text{ is even,} \\ \sum_k f_{p-1-j}(k) w_n^m(pt-k), & \text{if } n \text{ is odd.} \end{cases}$$

Since we are enforcing $n = n'$ at each level m, we are composing with the permutation defined above. Of course, this algorithm has complexity identical to the original. $\square$

Periodic filters and bases for $\mathbf{R}^d$. A sampled periodic function may be represented as a vector in $\mathbf{R}^d$ for some d. In this case let p be any factor of d. Introduce as filters a family of p vectors $\{\tilde{f}_i \in \mathbf{R}^d : i = 0,\dots,p-1\}$. These are obviously summable. Suppose in addition that they are orthogonal as periodic discrete functions, i.e., that $\sum_{m=1}^{d} \tilde{f}_i(m)\tilde{f}_j(m+kp \mod d) = \delta_{i-j}\delta_k$.

Let the associated convolution operators be $\{\tilde{F}_0, \dots, \tilde{F}_{p-1}\}$, defined as above by

$$\tilde{F}_i : \mathbf{R}^d \to \mathbf{R}^{d/p}, \quad \tilde{F}_i v(k) = \sum_{m=1}^{d} \tilde{f}_i(m + pk \mod d) v(m), \quad \text{for } k = 1, 2, \dots, d/p,$$

$$\tilde{F}_i^* : \mathbf{R}^{d/p} \to \mathbf{R}^d, \quad \tilde{F}_i^* v(m) = \sum_{k=1}^{d/p} \tilde{f}_i(m + pk \mod d) v(k), \quad \text{for } m = 1, 2, \dots, d.$$

The reduction modulo d is intentionally emphasized. These operators satisfy conditions similar to those of aperiodic filters:

Lemma 7.

(1) $\tilde{F}_i \tilde{F}_j^* = 0, \qquad$ *if* $i \neq j$,

(2) $\tilde{F}_i \tilde{F}_i^* = I_{d/p}$

(3) $\tilde{F}_i^* \tilde{F}_i$ *is a rank* d/p *orthogonal projection on* $\mathbf{R}^d$*, and for* $i \neq j$ *the ranges of* $\tilde{F}_i^* \tilde{F}_i$ *and* $\tilde{F}_j^* \tilde{F}_j$ *are orthogonal,*

(4) $\tilde{F}_0^* \tilde{F}_0 + \cdots + \tilde{F}_{p-1}^* \tilde{F}_{p-1} = I_d$

where I_d *is the identity on* $\mathbf{R}^d$.

Proof. The proof is nearly identical with the one in the aperiodic case. □

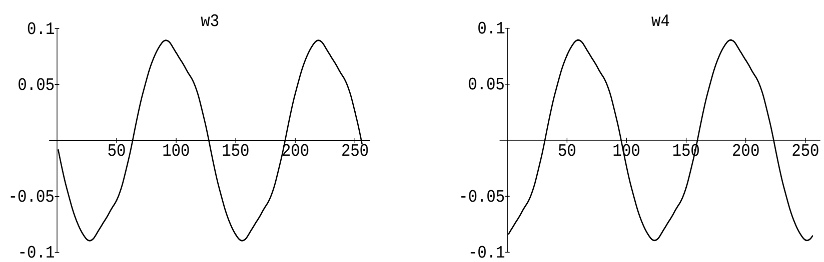

Figure 6.

Periodized wavelet packets, $p = 2$**, sequencies 3 and 4.**

The decomposition suggested by equation (4) may be recursively applied to the p subspaces $\mathbf{R}^{d/p}$ to generate *periodic wavelet packets*. We must extend the action of the filter family to $\mathbf{R}^{d/p}$ in the natural way. For $d = p_1 \dots p_L$ and $0 \leq n < d$, we have a unique representation $n = n_1 + n_2 p_1 + n_3 p_2 p_1 + \cdots + n_L p_{L-1} \dots p_1$, where $0 \leq n_i < p_i$. This defines a one-to-one correspondence between $\{0, \dots, d-1\}$ and an index set of L-tuples $I = \{ (n_1, \dots, n_L) : 0 \leq n_i < p_i \}$. We can construct a basis of $\mathbf{R}^d$ whose elements are indexed by I. For $n = (n_1, \dots, n_L) \in I$, define $\tilde{\mathbf{F}}_n^L = \tilde{F}_{n_L}^L \dots \tilde{F}_{n_1}^1$, where $\tilde{F}^i$ is a family of p_i periodic filters. Then $\tilde{\mathbf{F}}_n^{L*} \tilde{\mathbf{F}}_n^L$ is an orthogonal projection onto a 1-dimensional subspace of $\mathbf{R}^d$. This is shown by induction on the rank in (3). Now let W_n^L be the range of this projection. The collection $\{u_n = \tilde{\mathbf{F}}_n^{L*}1 : n \in I\}$ of standard basis vectors of W_n^L will be an orthonormal basis of $\mathbf{R}^d$, and the map $\tilde{\mathbf{F}}_n^L : \mathbf{R}^d \to \mathbf{R}$ gives the component in the u_n direction. Some examples of periodic wavelet packets are depicted in Figures 6 and 7.

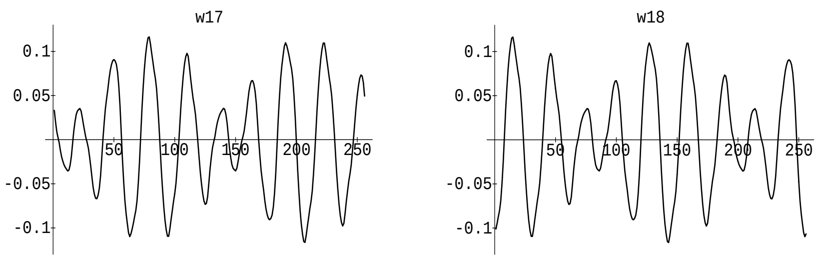

Figure 7.

Periodized wavelet packets, $p = 2$, sequencies 17 and 18.

These periodic wavelet packets are more useful in practice than the aperiodic wavelet packets we first considered, because the number of coefficients produced by the periodic wavelet packet transform is no more than the original number of signal samples. Of course, this advantage is balanced by the requirement that we treat the original signal as periodic. Notice that the periodized wavelet packets at sequencies $2n-1$ and $2n$ differ only by a shift. This shift, which is always close to $1/4$ period, bears a simple relationship to the binary expansion of n, but its main significance is that there are really only $N/2$ distinct frequencies in a collection of N periodized wavelet packets. This is a reflection of Nyquist's theorem, which states that if we sample a periodic function N times within a period, then we can only distinguish frequencies up to $N/2$.

As before, we are not limited to the basis defined by the index set I. Products of fewer than L filters form orthogonal projections onto a tree of subspaces of $\mathbf{R}^d$. A node arising from a product of m filters will correspond to the subspace $W_n^m = \tilde{\mathbf{F}}_n^{m*}\tilde{\mathbf{F}}_n^m\mathbf{R}^d$, where $n = n_1 + \cdots + n_m p_{m-1} \ldots p_1$ indexes a composition of m filters. The tree will be nonhomogeneous in general, although all nodes i levels from the root will have the same number p_i of children. Define a *nonhomogeneous graph* as a finite union of nodes whose associated subtrees form a disjoint cover of some level $m \leq L$. A graph theorem holds for this tree of subspaces as well. It and its corollary may be stated as follows:

THEOREM 8. *Every nonhomogeneous graph corresponds to an orthogonal decomposition of* $\mathbf{R}^d$. □

COROLLARY 9. *Graphs are in one-to-one correspondence with finite disjoint covers of* $[0,1)$ *by intervals of the form* $I_n^m = (p_1 \ldots p_m)^{-1}[n, n+1)$. □

This correspondence also induces a partial order on nonhomogeneous graphs. We say that u is greater than or equal to v if the cover associated to u is a refinement of the cover associated to v. This partial order has both a minimal element $\{W_0^0\}$ and a maximal element $\{W_0^L, \ldots, W_{p_1 \ldots p_L - 1}^L\}$. Any permutation of the factors of d gives a (possibly different) set of bases.

Concentration criteria for the best-adapted basis. Define an *additive information cost function* on l^2 to be any functional $M : l^2 \to \mathbf{R}^+$ satisfying

$$(i) \quad M(\{x_i\}) = \sum_i m(|x_i|) \text{for some function } m = m(t),$$

$$(ii) \quad M(\{0\}) = 0.$$

Some useful information cost functions are the threshold counting norm $\#\{k : |x_k| > \epsilon\}$ and the bit-length norm $\sum_k \log(1 + |x_k|/\epsilon)$.

Let U be a finite library of orthonormal bases of l^2. A vector x has coefficients x_u in the basis $u \in U$, where $x_u(k) = \langle x, u(k) \rangle$ for $u(k) \in u$. The information cost of a particular representation may be measured by $M(x_u)$. Also, the information contained in the the choice of u is $\log |U|$, where $|U|$ is the cardinality of the library. Define a *best-basis* for x as any element $u \in U$ for which $M(x_u)$ is minimal. The *best-basis information cost* of x in the library U is therefore $M(x_u) + \log |U|$.

Our goal is to find an efficient algorithm for reducing the information cost of vectors in a class. The library U, which depends on the class, should be very large but easy to search. It takes a naïve algorithm $O(|U|)$ operations to find the least-information representation of a fixed vector x. This procedure is inefficient because it requires a global reëvaluation of the information cost for each basis in the library.

We can evidently reduce the information $M(x \oplus y)$ by reducing $M(x)$ and $M(y)$ individually. Such a procedure is local in the following sense. Suppose that 2 orthonormal bases u and v in a library U partially coincide, and we write $x_u = x_{u \cap v} \times x_{u'}$ and $x_v = x_{u \cap v} \times x_{v'}$. Then $M(x_u) < M(x_v) \iff M(x_{u'}) < M(x_{v'})$. The decomposition and individual reduction of M may then be reapplied to the pieces $x_{u'}$ and $x_{v'}$.

Let $\mathbf{W}$ be a tree of subspaces of l^2. If $W \in \mathbf{W}$ is a node, then the associated coordinate map is $\mathbf{F}_W$, and the associated orthogonal projection can be denoted by $\mathbf{F}_W^* \mathbf{F}_W$. Every node W may be regarded as the Cartesian product of its daughters, or more generally of the elements of any graph of its descendants. Let U be the library of bases corresponding to graphs in $\mathbf{W}$. Then each element of U has a unique factorization into a Cartesian product of the standard bases of the subspaces in the associated graph. Namely, if x is a sequence and if u is a basis corresponding to some graph $G \subset \mathbf{W}$, then $x_u = \times \prod_{W \in G} \mathbf{F}_W x$. Consequently, $M(x_u) = \sum_{W \in G} M(\mathbf{F}_W x)$. To find the best basis, we must choose a graph of those subspaces W which contribute the least information. But this large choice may be factored into a sequence of small subchoices.

Recall now our definition of a partial ordering on graphs through the tree $\mathbf{W}$, which is inherited from our definition of the partial order on trees. We must keep track of the lowest achievable measure of information M as we progress down the tree to its root. This may be defined inductively. We must suppose now that the tree is finite with L levels, so that its set of graphs has a maximal element. Let G_L be this maximal graph in $\mathbf{W}$, and for $W \in G_L$ set $M^*_W(x) = M(\mathbf{F}_W x)$. Then for $W \in \mathbf{W}$ let $M^*_W(x) = \min\{M^*_V(x) : V \geq W\}$. This functional M^*_W records the minimum value of $M(\mathbf{F}_W x_u)$ achievable by bases of W coming from graphs through subtrees above node W. Evidently, if W_0 is the root (or minimal) node of $\mathbf{W}$, then $M^*_{W_0}(x) = M(x_{u_{\min}})$ is the best-basis measure of information for the vector x.

The search algorithm may now be described. Mark all maximal nodes $W \in G_L$ as "kept." Suppose now that node W has children $W_1, \ldots, W_n$. Mark W as "kept" if $M^*_W(x) \leq M^*_{W_1}(x) + \cdots + M^*_{W_n}(x)$; otherwise mark it as "not kept". Namely, keep W if including it reduces M^*. Observe that we can compute $M^*_W(x) = \min\{M(F_W x), M^*_{W_1}(x) + \cdots + M^*_{W_n}(x)\}$ without having to search the entire subtree above W. We may proceed down the tree to the root, at which point all the nodes in the tree have been marked either as "kept" or as "not kept." We claim the following:

Proposition 10. *The union of the minimal "kept" nodes is a graph corresponding to the best-basis representation of x.*

Proof. That it is a graph is clear by induction. Every minimal "kept" node W is the root of a subtree containing some of the maximal nodes of $\mathbf{W}$. This set of subtrees is disjoint, since if two subtrees intersect then one must contain the other and so their roots cannot both be minimal. The union of these disjoint subtrees covers all of the maximal nodes of $\mathbf{W}$, which form a complete level of the tree.

So call this graph G. Notice that the sum $\sum_{W \in G} M^*_W(x)$ over the minimal "kept" nodes $W \in G$ is equal to $M^*_{W_0}(x)$, where W_0 is the root of $\mathbf{W}$. By the remarks immediately above the proposition, this is the minimum achievable information cost. □

Operations required to find the best basis. We may count the operations in our search algorithm above as follows. Let $E(W)$ be the number of operations required to evaluate $M(F_W x)$, and let $D(W)$ be the number of children of the node $W \in \mathbf{W}$. Then it will require $\sum_{W \in G_L} E(W) + \sum_{W \notin G_L} [E(W) + D(W)]$ operations to construct the functional M^* on the tree and to mark the appropriate nodes as "kept." Finding the minimal "kept" nodes requires a depth-first search of $\mathbf{W}$, which takes at most $|\mathbf{W}|$ operations.

For definiteness, consider the example of a homogeneous p-adic tree generated by periodizing a family of p filters. Suppose we start with a vector x of $N = p^L$ components and develop the tree of its representations $\mathbf{W}$ as far as we can, namely L levels. We can label the nodes W^m_n as before. We observe that

$\mathbf{F}_n^m x$ has p^{L-m} components so that $E(W_n^m) = cp^{L-m}$, where c is the number of operations required per non-zero coefficient to compute M. This tree has p^m nodes at level m. $D(W) = p$ for all $W \in \mathbf{W}$, so it requires $\sum_{n=0}^{p^L-1} cp^0 + \sum_{m=0}^{L-1} \sum_{n=0}^{p^m-1} [cp^{L-m} + p] = cp^L + p(p^L - 1)/(p-1) = O(p^L) = O(N)$ operations to build M^*. Then $|\mathbf{W}| = \sum_{m=0}^{L} p^m = (p^{L+1} - 1)/(p-1) = O(p^L) = O(N)$, so that the entire search takes $O(N)$ operations for a periodic vector of length N.

The library of graphs through a tree grows rapidly with the number of levels in the tree. Let $|U_L|$ be the number of bases in a p-adic tree of L levels. Then $|U_L|$ satisfies the recurrence $|U_{L+1}| = 1 + |U_L|^p$, which is easily estimated as greater than $2^{p^{L-1}} = 2^{N/p}$. We list the first 7 values $|U_0|, \ldots, |U_6|$ for $p = 3$:

```
1, 2, 9, 730, 389017001, 58871587162270593034051002,
204040901322752673844230437877671861543858084850895762746141813554591014612009
```

By contrast, one may also list the number of operations required to find the best basis representation and information of a vector of length $N = 3^L$ in a tree of $L = 0, \ldots, 6$ levels. We shall suppose that evaluating M requires 3 operations per coefficient, so that the operation count is at most $\frac{3}{2}(3^{L+1} - 1)$.

$\{$ `3,12,39,120,363,1092,3279` $\}$

This is a good example of combinatorial explosion tamed by an efficient search algorithm.

The entropy criterion and estimates. Let $x \in l^2$ and denote by $\|x\|$ the usual norm: $\|x\|^2 = \sum_k |x_k|^2$. Then the sequence defined by $|x_k|^2/\|x\|^2$ gives a probability distribution of the energy of x. This distribution has a Shannon entropy which we shall denote by $\mathcal{H}(x) = -\sum_k (|x_k|^2/\|x\|^2) \log(|x_k|^2/\|x\|^2)$, where the summand is interpreted as 0 for any $x_k = 0$. This entropy is a well-known measure of the information of a distribution.

We may also use the $L^2 \log L^2$ norm rather than entropy. Denote this by $H(x) = -\sum_k |x_k|^2 \log |x_k|^2$ with the same convention for the case $x_k = 0$. Note that

$$\mathcal{H}(x) = H(x)\|x\|^{-2} + \log \|x\|^2$$

H is an additive measures of information. $\mathcal{H}$ is not, but the above relation guarantees that whenever $\|x\| = \|y\|$, we have $\mathcal{H}(x) < \mathcal{H}(y) \iff H(x) < H(y)$. Suppose that u is a best basis for x with respect to H. It is clear that this will also be the element for which the more classical $\mathcal{H}(x_u)$ is minimal, and also that for which $\exp \mathcal{H}(x_u)$ is minimal. The exponential of entropy has the following suggestive property:

LEMMA 11. *If $x \in l^2$ is a sequence of 0's and 1's, then* $\exp \mathcal{H}(x)$ *is the number of 1's in the sequence.* $\square$

Notice that $\exp \mathcal{H}(x) = \|x\|^2 \exp\big(H(x)\|x\|^{-2}\big)$.

Suppose now that $x \in l^2$ is any sequence, and we project it onto a sequence y_ϵ defined for $\epsilon \geq 0$ by

$$y_k = \begin{cases} 0, & \text{if } |x_k|^2 < \epsilon \exp\left(-H(x)\|x\|^{-2}\right), \\ x_k, & \text{otherwise.} \end{cases}$$

Then there will be at most $\exp\left(-H(x)\|x\|^{-2}\right)/\epsilon$ nonzero terms in y. The energy error $\|x-y\|^2$ will be the sum of the squares of the omitted terms.

Existence and construction of filters. We can construct finitely supported filters of any support length greater that p. Longer support lengths allow more degrees of freedom. Let M be a positive integer and consider the problem of finding filters of length pM, i.e., p trigonometric polynomials $m_0, \ldots, m_{p-1}$ of degree pM for which the above matrix of values of m_j is unitary. By a construction similar to Pollen's in [**P**], this is equivalent to finding an element of the group $SU(p, \mathbf{C}[z, 1/z])$ which is the product of M inverse factors.

Given any pair P, Q of (perfect reconstruction) quadrature mirror filters, we can build a family of $p = 2^q$ filters by taking all distinguishable compositions of P and Q of length q. Alternatively, we can take all distinguishable products of q filters. This method serves to build filters for q-dimensional signals. Given a signal $s = s(x) = s(x_1, \ldots, x_q)$, and $J = j_q \ldots j_1$ radix 2, we can define 2^q filters F_J by taking a one-dimensional filter for each dimension: $F_J s(x) = \sum_{k_1,\ldots,k_q} f_{j_1}(k_1 + px_1) \ldots f_{j_q}(k_q + px_q)\, s(k_1, \ldots, k_q)$. Such filters are useful for image processing and matrix multiplication.

Gopinath and Burrus [**GB**] have given a construction of "multiplicity p" wavelets similar to the one in this paper. Their scheme for generating filter families is based on "cosine modulation," and they provided examples of filter families with Hölder regularity. In practice it is desirable to have smooth basis elements, since a certain degree of smoothness (one derivative in L^2) is needed to have finite variance in frequency. We can define a smoothness property for filter sequences:

Definition. *A summable sequence f is a smooth filter (of degree $d \leq \infty$) if there is a nonzero solution ϕ in $L^1(\mathbf{R}) \cap L^2(\mathbf{R}) \cap C^d(\mathbf{R})$ to the functional equation*

$$\phi(x) = p^{1/2} \sum_m f(m) \phi(px + m).$$

Daubechies has shown in [**D**] that finitely supported filters of any degree of smoothness may be constructed in the case $p = 2$. An obvious consequence is that smooth filters exist in the case $p = 2^q$. For arbitrary p, Lundberg and Welland [**LW**] give a construction of p-families of filters whose wavelet packets are m-differentiable, where p and m are arbitrary.

References

[**CMW**]. Ronald R. Coifman, Yves Meyer, and M. Victor Wickerhauser, *Wavelet analysis and signal processing*, Wavelets and Their Applications, ed. B.

Ruskai et al., ISBN 0-86720-225-4, Jones and Bartlett, Boston, 1992, pp. 153–178..

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

[**FJW**]. Michael Frazier, Bjørn Jawerth and Guido Weiss, *Littlewood–Paley Theory and the Study of Function Spaces*, CBMS Regional Conference Series, American Mathematical Society, Providence, 1990.

[**GB**]. R. A. Gopinath and C. S. Burrus, *Wavelet transforms and filter banks*, Wavelets–A Tutorial in Theory and Applications, C. K. Chui (ed.) ISBN 0-12-174590-2, Academic Press, Boston, 1992, pp. 603–654.

[**GM**]. A. Grossman and J. Morlet, *Decomposition of Hardy functions into square-integrable wavelets of constant shape*, SIAM J. Math. Anal. **8** (1985), 4.

[**L**]. Pierre Gilles Lemarié, *Fonctions a support compact dans les analyses multi-résolutions*, Revista Matemática Iberoamericana **7** (1991), 157–182.

[**LW**]. M. Lundberg and G. Welland, *Construction of compact p-wavelets*, to appear, J. Constructive Approximation (1992).

[**Ma**]. Stephane G. Mallat, *A Theory for Multiresolution Signal Decomposition: The Wavelet Decomposition*, IEEE Transactions on Pattern Analysis and Machine Intelligence **11** (1989), 674–693.

[**Me1**]. Yves Meyer, *Ondelettes et Opérateurs I,II*, Hermann, Paris, 1990.

[**Me2**]. Yves Meyer, *De la recherche pétrolière à la géometrie des espaces de Banach en passant par les paraproduits*, preprint, École Polytechnique, Palaiseau, Seminaire équations aux dérivées partielles (1985–1986).

[**P**]. David Pollen, *Parametrization of Compactly Supported Wavelets*, preprint AD890503.1.4, Aware, Inc., Cambridge, Mass. (1989).

[**W**]. M. Victor Wickerhauser, *Acoustic signal compression with wavelet packets*, Wavelets–A Tutorial in Theory and Applications, C. K. Chui (ed.) ISBN 0-12-174590-2, Academic Press, Boston, 1992, pp. 679–700.