ENTROPY-BASED ALGORITHMS FOR BEST BASIS SELECTION

Ronald R. Coifman, Mladen Victor Wickerhauser

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

We would like to describe a method permitting efficient compression of a variety of signals such as sound and images. While similar in goals to vector quantization, the new method uses a codebook or library of predefined modulated waveforms with some remarkable orthogonality properties. We can apply the method to two particularly useful libraries of recent vintage, orthogonal wavelet-packets [CM1],[CW] and localized trigonometric functions [CM2], for which the time-frequency localization properties of the waveforms are reasonably well controlled. The idea is to build out of the library functions an orthonormal basis relative to which the given signal or collection of signals has the lowest information cost. We may define several useful cost functionals; one of the most attractive is Shannon entropy, which has a geometric interpretation in this context.

Practicality is built into the foundation of this orthogonal best-basis methods. All bases from each library of waveforms described below come equipped with fast $O(N \log N)$ transformation algorithms, and each library has a natural dyadic tree structure which provides $O(N \log N)$ search algorithms for obtaining the best basis. The libraries are rapidly constructible, and never have to be stored either for analysis or synthesis. It is never necessary to construct a waveform from a library in order to compute its correlation with the signal. The waveforms are indexed by three parameters with natural interpretations (position, frequency, and scale), and we have experimented with feature-extraction methods that use best-basis compression for front-end complexity reduction.

The method relies heavily on the remarkable orthogonality properties of the new libraries. It is obviously a nonlinear transformation to represent a signal in its own best basis, but since the transformation is orthogonal once the basis is chosen, compression via the best-basis method is not drastically affected by noise: the noise energy in the transform values cannot exceed the noise energy in the original signal. Furthermore, we can use information cost functionals defined for signals with normalized energy, since all expansions in a given library will conserve energy. Since two expansions will have the same energy globally, it is not necessary to normalize expansions to compare their costs. This feature greatly enlarges the class of functionals usable by the method, speeds the best-basis search, and provides a geometric interpretation in certain cases.

Definitions of Two Modulated Waveform Libraries. We now introduce the concept of a "library of orthonormal bases." For the sake of exposition we restrict our attention to two classes of numerically useful waveforms, introduced recently by Y. Meyer and the authors.

We start with trigonometric waveform libraries. These are localized sine transforms associated to a covering by intervals of $\mathbf{R}$ or, more generally, of a manifold.

We consider a strictly increasing sequence $\{a_i\} \subset \mathbf{R}$, and build an orthogonal decomposition of $L^2(\mathbf{R})$. Let b_i be a continuous real-valued function on the interval $[a_{i-1}, a_i]$ satisfying:

$$b_i(a_{i-1}) = 0; \qquad b_i(a_i) = 1; \qquad b_i^2(t) + b_i^2(2a_i - t) = 1 \quad \text{for } a_{i-1} < t < a_i.$$

Then the function which we may define by $\tilde{b}_i(t) = b_i(2a_i - t)$ is the reflection of b_i about the midpoint of $[a_{i-1}, a_i]$, and we have $b_i^2(t) + \tilde{b}_i^2(t) = 1$. Now define

$$p_i(t) = \begin{cases} b_i, & \text{if } a_{i-1} \le t < a_i, \\ \tilde{b}_{i+1}, & \text{if } a_i \le t \le a_{i+1}, \\ 0, & \text{if } t < a_{i-1} \text{ or } t > a_{i+1}. \end{cases}$$

Each p_i is supported on the interval $[a_{i-1}, a_{i+1}]$, and we have

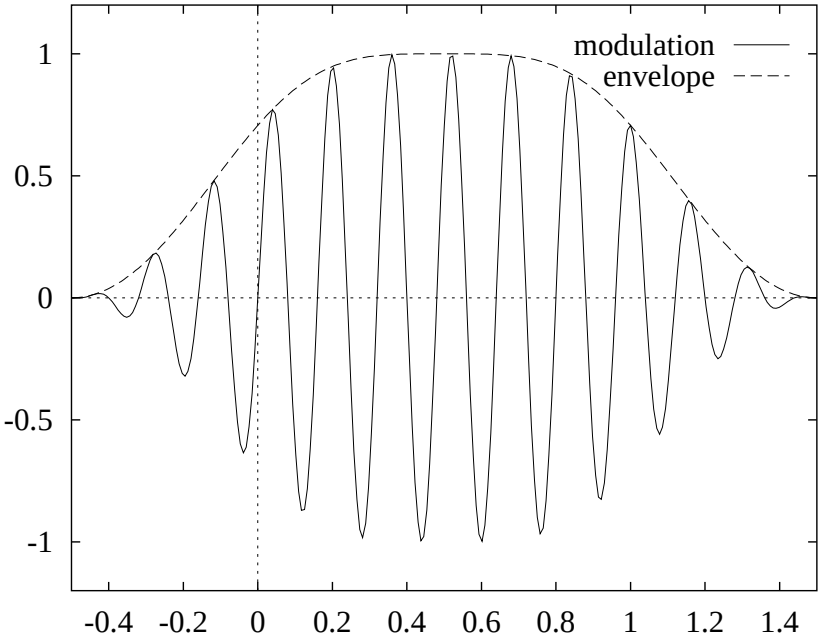
$$\sum_i p_i^2 = 1.$$

The middle of the bump function p_i lies over the interval $I_i = [c_i, c_{i+1})$, where $c_i = (a_i + a_{i-1})/2$; these intervals form a disjoint partition of $\mathbf{R}$, and we can show that the following functions form an orthonormal basis for $L^2(\mathbf{R})$ localized to this partition:

$$S_{i,k}(t) = \sqrt{\frac{2}{|I_i|}}\, p_i(t) \sin\left[\pi(k+\frac{1}{2})\frac{t-c_i}{|I_i|}\right].$$

This is what we shall call a *local sine basis*. Certain modifications are possible, for example sine can be replaced by cosine, so we shall refer to it also as a local trigonometric basis.

Below is a plot of one such function, localized to the interval $[0, 1]$:



Example of a localized sine function.

The indices of each function $S_{i,k}$ have a natural interpretation as "position" and "frequency." The collection $\{S_{i,k} : k \in \mathbf{N}\}$ forms an oscillatory orthonormal basis for a subspace of $L^2(\mathbf{R})$ consisting of continuous functions supported in $[a_{i-1}, a_{i+1}]$. If we denote this subspace by H_{I_i}, then $H_{I_i} + H_{I_{i+1}}$ is spanned by the functions

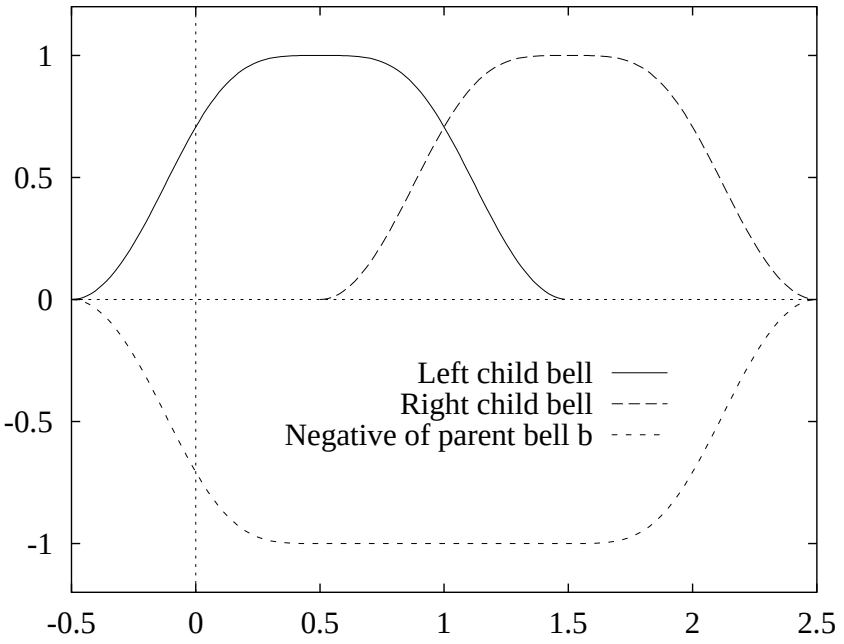
$$\sqrt{\frac{2}{|I_i| + |I_{i+1}|}}\, P(t) \sin\left[(k+\frac{1}{2})\pi\frac{t-c_i}{|I_i| + |I_{i+1}|}\right]$$

where

$$P^2(t) = p_i^2(t) + p_{i+1}^2(t)$$

is a "window" function whose middle lies over the interval $I_i \cup I_{i+1}$.

The relationship between the larger interval and its 2 "children" is illustrated by the following figure:



The larger subspace is the direct sum of the 2 smaller subspaces.

It can now be seen how to construct such an orthonormal basis for any partition of $\mathbf{R}$ which has $\{a_i\}$ as a refinement. For each disjoint cover $\mathbf{R} = \cup_n J_n$, where the J_n are unions of contiguous I_i, we have $L^2(\mathbf{R}) = \oplus_n H_{J_n}$. The local trigonometric bases associated to all such partitions may be said to form a *library of orthonormal bases*. There is a partial ordering of such partitions by refinement; the graph of the partial order can be made into a tree, and the tree can be efficiently searched for a "best basis" as will be described below.

A second new library of orthonormal bases, called the wavelet packet library, can also be constructed. This collection of modulated wave forms corresponds roughly to a covering of "frequency" space. This library contains the wavelet basis, Walsh functions, and smooth versions of Walsh functions called wavelet packets.

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.

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

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

$$F_0\{s_k\}(2i) = \sum_k s_k h_{k-2i} \tag{1.0}$$

$$F_1\{s_k\}(2i) = \sum_k s_k g_{k-2i}.$$

The map $\mathbf{F} : \ell^2(\mathbf{Z}) \to \ell^2(2\mathbf{Z}) \oplus \ell^2(2\mathbf{Z})$ defined by $\mathbf{F} = F_0 \oplus F_1$ is orthogonal. We also have $F_0 F_0^* = F_1 F_1^* = I$, $F_1 F_0^* = F_0 F_1^* = 0$, and

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

We now define the sequence of functions $\{W_k\}_{k=0}^{\infty}$ from a given function W_0 as follows:

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

Notice that W_0 is determined up to a normalizing constant by the fixed-point problem obtained when $n = 0$. The function $W_0(x)$ can be identified with the scaling function φ in [D] and W_1 with the basic wavelet 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

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

Taking the 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

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

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

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

We define our *library of wavelet packet bases* to be the collection of orthonormal bases composed of functions of the form $W_n(2^\ell x - k)$, where $\ell, k \in \mathbf{Z}$, $n \in \mathbf{N}$. Here, each element of the library is determined by a subset of the indices: a scaling parameter ℓ, a localization parameter k and an oscillation parameter n. These are natural parameters, for the function $W_n(2^\ell x - k)$ is roughly centered at $2^{-\ell} k$, has support of size $\approx 2^{-\ell}$, and oscillates $\approx n$ times. We have the following simple description of the orthonormal bases in the library:

Proposition. *Any collection of indices $(\ell, n, k) \subset \mathbf{N} \times \mathbf{N} \times \mathbf{Z}$ such that the intervals $[2^\ell n, 2^\ell (n + 1))$ form a disjoint cover*[1] *of $[0, \infty)$, and k ranges over all the integers, corresponds to an orthonormal basis of $L^2(\mathbf{R})$.*

If we use Haar filters, there will be elements of the library which do not correspond to disjoint dyadic covers. For the sake of generality, we will not consider such other bases.

This collection of disjoint covers forms a partially ordered set. Just like the local trigonometric basis library, the wavelet packet basis library organizes itself into a tree, which may be efficiently searched for a "best basis."

Entropy of a Vector. We now define a real-valued cost functional $\mathcal{M}$ on sequences and search for its minimum over all bases in a library. Such a functional should, for practical reasons, describe "concentration" or the number of coefficients required to accurately describe the sequence. By this we mean that $\mathcal{M}$ should be large when the coefficients are roughly the same size and small when all but a few coefficients are negligible. In particular, any averaging process should increase the information cost, suggesting that we consider convex functionals. This property should also hold on the unit sphere in ℓ^2, since we will be measuring coefficient sequences in various orthogonal bases. Finally, we will restrict our attention to those functionals which split nicely across cartesian products, so that the search is a fast divide-and-conquer.

[1] We can think of this as an even covering of frequency space by windows roughly localized over the corresponding intervals.

Definition. *A map $\mathcal{M}$ from sequences $\{x_i\}$ to $\mathbf{R}$ is called an* additive information cost function *if $\mathcal{M}(0) = 0$ and $\mathcal{M}(\{x_i\}) = \sum_i \mathcal{M}(x_i)$.*

If we fix a vector $x \in \mathbf{R}^N$, we can make an additive information cost function into a functional on the manifold of orthonormal bases, i.e., the orthogonal group $\mathbf{O}(N)$. Let $B \in \mathbf{O}(N)$ be an orthonormal basis, written as a matrix of row vectors. Then Bx is the vector of coefficients of x in the orthonormal basis B, and $\mathcal{M}(Bx)$ is the information cost of x in the basis B.

Since $\mathbf{O}(N)$ is compact, there is a global minimum for every continuous information cost. Unfortunately, this minimum will not be a rapidly computable basis in general, nor will the search for a minimum be of low complexity. Therefore, we will restrict our attention to a library $\mathcal{B} \subset \mathbf{O}(N)$ of orthonormal bases each of which has an associated fast transform (of order $O(N \log N)$ or better) and for which the search for a constrained minimum of $\mathcal{M}$ converges in $O(N)$ operations.

Definition. *The* best basis *relative to $\mathcal{M}$ for a vector x in a library $\mathcal{B}$ of bases is that $B \in \mathcal{B}$ for which $\mathcal{M}(Bx)$ is minimal.*

Motivated by ideas from signal processing and communication theory we were led to measure the "distance" between a basis and a function in terms of the Shannon entropy of the expansion. More generally, let H be a Hilbert space. Let $v \in H$, $\|v\| = 1$ and assume that H is an orthogonal direct sum:

$$H = \oplus \sum H_i$$

We write $v = \oplus \sum_i v_i$ for the decomposition of v into its H_i-components, and define

$$\varepsilon^2(v; \{H_i\}) = -\sum \|v_i\|^2 \ell n \|v_i\|^2$$

as a measure of distance between v and the orthogonal decomposition. ε^2 is characterized by the Shannon equation which is a version of Pythagoras' theorem.

Let

$$\begin{aligned} H &= \oplus(\sum H^i) \oplus (\sum H_j) \\ &= H_+ \oplus H_- \end{aligned}$$

Thus, H^i and H_j give orthogonal decompositions $H_+ = \sum H^i, H_- = \sum H_j$. Then

$$\varepsilon^2(v; \{H^i, H_j\}) = \varepsilon^2(v; \{H_+, H_-\} + \|v_+\|^2 \varepsilon^2 \left(\frac{v_+}{\|v_+\|}; \{H^i\} \right) + \|v_-\|^2 \varepsilon^2 \left(\frac{v_-}{\|v_-\|}; \{H_j\} \right)$$

This is Shannon's equation for entropy if we interpret $\|P_{H_+} v\|^2$ to be, as in quantum mechanics, the "probability" of v being in the subspace H_+. This equation enables us to search for a smallest-entropy spatial decomposition of a given vector.

Remark. The Karhunen-Loéve basis is the minimum-entropy orthonormal basis for an ensemble of vectors. The best basis as defined above is useful even for a single vector, where the Karhunen-Loéve method trivializes. The constraint to a library $\mathcal{B}$ can keep us within the class of "fast" orthonormal expansions.

Suppose that $\{x_n\}$ belongs to both L^2 and $L^2 \log L$. If $x_n = 0$ for all sufficiently large n, then in fact the signal is finite dimensional. Generalizing this notion, we can compare sequences by their rate of decay, i.e., the rate at which their elements become negligible if they are rearranged in decreasing order. This allows us to introduce a notion of the dimension of a signal.

Definition. *The* theoretical dimension *of $\{x_n\}$ is*

$$d = \exp(-\sum_n p_n \log p_n)$$

where $p_n = |x_n|^2 \|x\|^{-2}$.

Our nomenclature is supported by the following ideas, which are proved in most information theory texts:

Proposition. *If $x_n = 0$ for all but finitely many (say N) values of n, then $1 \le d \le N$.*

Proposition. *If $\{x_n\}$ and $\{x'_n\}$ are rearranged so that both $\{p_n\}$ and $\{p'_n\}$ are monotone decreasing, and if we have $\sum_{0<n<m} p_n \ge \sum_{0<n<m} p'_n$ for all m, then $d \le d'$.*

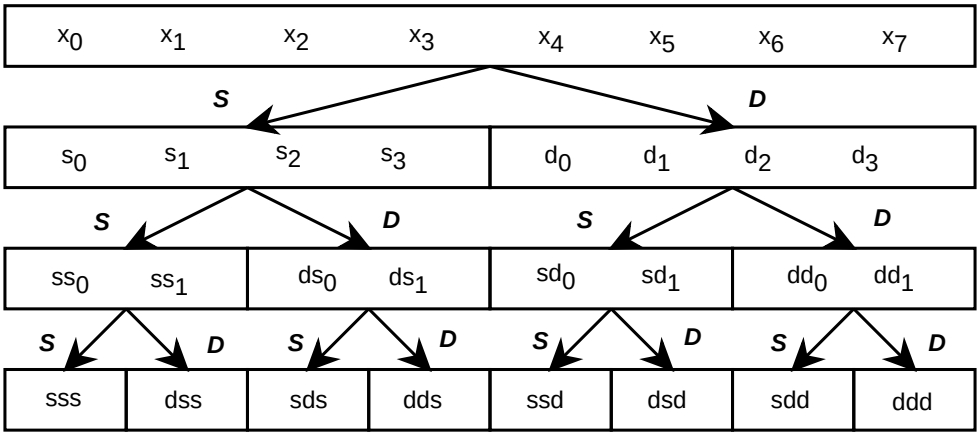
Of course, while entropy is a good measure of concentration or efficiency of an expansion, various other information cost functions are possible, permitting discrimination and choice between special function expansions.

Selecting the Best Basis. For the local trigonometric basis library example, we can build a minimum-entropy basis from the most refined partition upwards. We start by calculating the entropy of an expansion relative to intervals of length one, then we compare the entropy of each adjacent pair of intervals to the entropy of an expansion on their union. We pick the expansion of lesser entropy and continue up to some maximum interval size. This uncovers the minimum entropy expansion for that range of interval sizes. This rough idea can be made precise as well as generalized to all libraries with a tree structure:

Definition. *A library of orthonormal bases is a* (binary) tree *if it satisfies:*

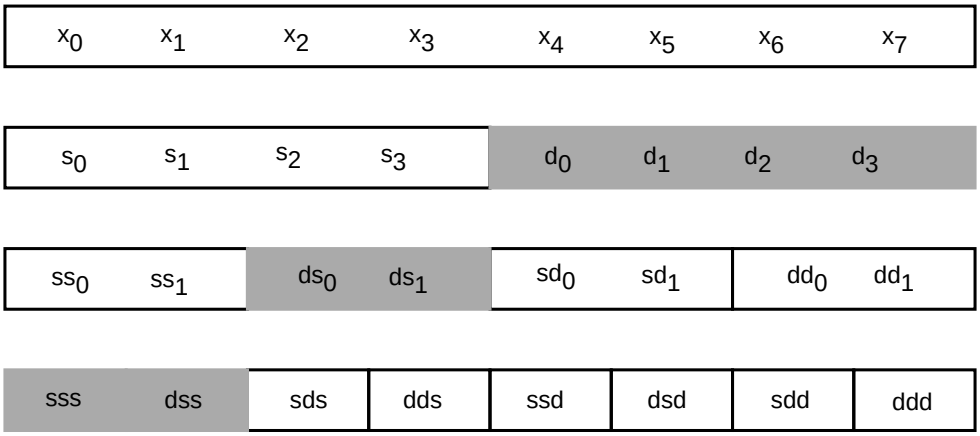
(1) *Subsets of basis vectors can be identified with intervals of* $\mathbf{N}$ *of the form* $I_{nk} = [2^k n, 2^k(n+1)[$, *for* $k, n \ge 0$.
(2) *Each basis in the library corresponds to a disjoint cover of* $\mathbf{N}$ *by intervals* I_{nk}.
(3) *If* V_{nk} *is the subspace identified with* I_{nk}, *then* $V_{n,k+1} = V_{2n,k} \oplus V_{2n+1,k}$.

The two example libraries above satisfy this definition. The library of wavelet packet bases is naturally organized as subsets of a binary tree. The tree structure is depicted in the figures below:

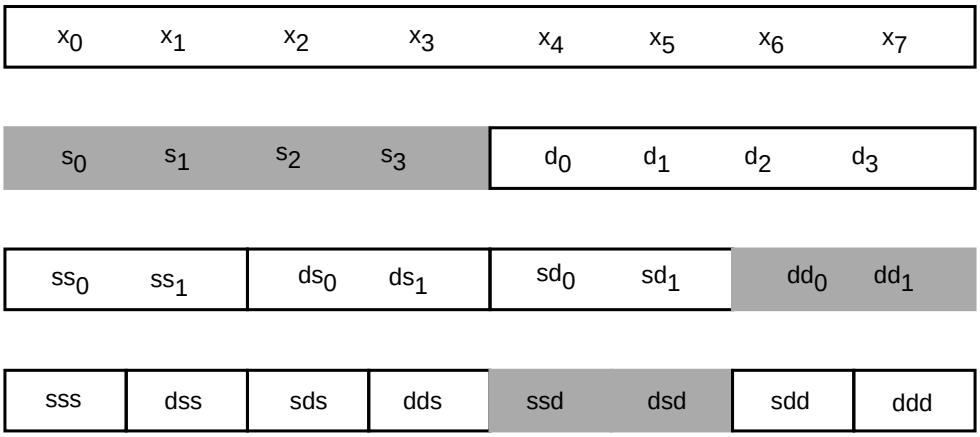


Wavelet packets organized as a binary tree.

Each node represents a subspace of the original signal. Each subspace is the orthogonal direct sum of its two children nodes. The leaves of every connected subtree give an orthonormal basis. Two example bases from this library are depicted in the figures below:

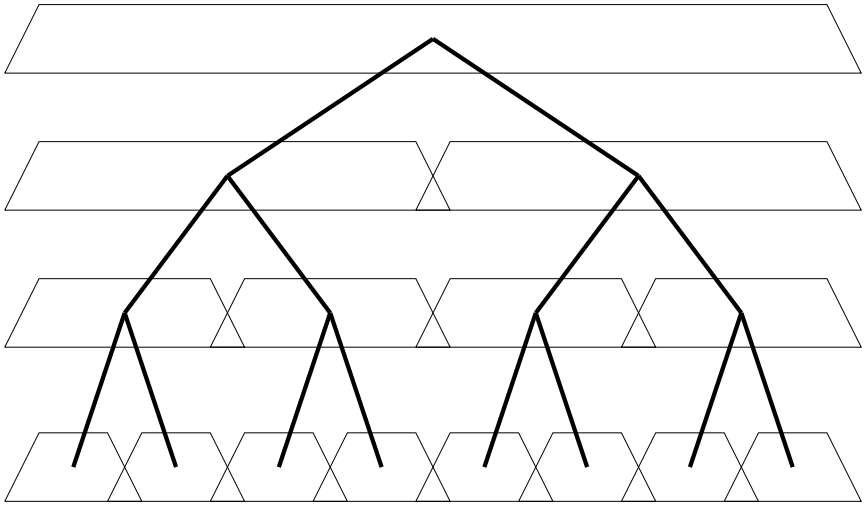


Part of the wavelet packet basis library: the wavelet basis.



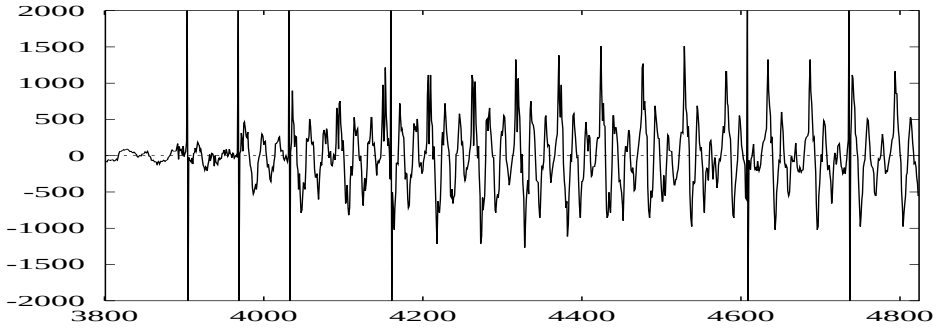
Part of the wavelet packet basis library: some unnamed basis.

The library of local trigonometric bases over a compact interval U may be organized as a binary tree by taking partitions localized to a dyadic decomposition of U. Then I_{00} will correspond to the sine basis on U, and I_{nk} will correspond to the local sine basis over interval n of the 2^k intervals at level k of the tree. This organization is depicted schematically in the figure below:



Organization of localization intervals into a binary tree.

This procedure permits the segmentation of acoustic signals into those dyadic windows best adapted to the local frequency content. An example is the segmentation of part of the word "armadillo," in the figure below:



Automatic lowest-entropy segmentation of part of a word.

REMARK. In multidimensions, we must extend this notion to libraries which can be organized as more general trees. This can be done by replacing condition (3) with the condition that for each $k, n \geq 0$, we have an integer $b > 1$ such that $V_{n,k+1} = V_{bn,k} \oplus \cdots \oplus V_{bn+b-1,k}$. It will not change the subsequent argument.

If the library is a tree, then we can find the best basis by induction on k. Denote by B_{nk} the basis of vectors corresponding to I_{nk}, and by A_{nk} the best basis for x restricted to the span of B_{nk}. For $k = 0$, there is a single basis available, namely the one corresponding to $I_{n,0}$, which is therefore the best basis: $A_{n,0} = B_{n,0}$ for all $n \geq 0$. We construct $A_{n,k+1}$ for all $n \geq 0$ as follows:

$$A_{n,k+1} = \begin{cases} B_{n,k+1}, & \text{if } \mathcal{M}(B_{n,k+1}x) < \mathcal{M}(A_{2n,k}x) + \mathcal{M}(A_{2n+1,k}x), \\ A_{2n,k} \oplus A_{2n+1,k}, & \text{otherwise.} \end{cases} \tag{1.4}$$

Fix $K \geq 0$ and let V be the span of I_{0K}. We have the following:

Proposition. *The algorithm in Eq.* (1.4) *yields the best basis for x relative to $\mathcal{M}$.*

Proof. This can be shown by induction on K. For $K = 0$, there is only one basis for V. If A' is any basis for $V_{0,K+1}$, then either $A' = B_{0,K+1}$ or $A' = A'_0 \oplus A'_1$ is a direct sum of bases for $V_{0,K}$ and $V_{1,K}$. Let A_0 and A_1 denote the best bases in these subspaces. By the inductive hypothesis, $\mathcal{M}(A_i x) \leq \mathcal{M}(A'_i x)$ for $i = 0, 1$, and by Eq. (1.4) $\mathcal{M}(Ax) \leq \min\{\mathcal{M}(B_{0,K+1}x), \mathcal{M}(A_0 x) + \mathcal{M}(A_1 x)\} \leq \mathcal{M}(A'x)$.

Comparisons are always made between two adjacent generations of the binary tree. Therefore, the complexity of the search is proportional to the number of nodes in the tree, which for a vector in $\mathbf{R}^N$ is just $O(N)$. This complexity is dominated by the cost of calculating all coefficients for all bases in the library. This takes $O(N \log N)$ for the wavelet packet library, and $O(N[\log N]^2)$ for the local trigonometric library. In practice, the coefficients are small: approximately 20 for wavelet packets, and approximately 1 for localized sines.

The number of bases in a binary tree library may be calculated recursively. Let A_L be the number of bases in a binary tree of $1 + L$ levels, i.e., L levels below the root or standard basis. We can combine two such trees, plus a new root, into a new tree of $2 + L$ levels. The two subtrees are independent, so we obtain the recursive formula

$$A_{L+1} = 1 + A_L^2$$

from which we can estimate $A_{L+1} \geq 2^{2^L}$. Thus a signal of $N = 2^L$ points can be expanded in 2^N different orthogonal bases in $O(N \log N)$ operations, and the best basis from the entire collection may be obtained in an additional $O(N)$ operations.

For voice signals and images this procedure leads to remarkable compression algorithms; see the references [W2] and [W3] below. The best basis method may be applied to ensembles of vectors, more like classical Karrhunen-Loéve analysis. The so-called "energy compaction function" may be used as an information cost to compute the joint best basis over a set of random vectors. The idea is to concentrate most of the variance of the sample into a few new coordinates, to reduce the dimension of the problem and make factor analysis tractable. The algorithm and an application to recognizing faces is described in [W1].

Some other libraries are known and should be mentioned. The space of frequencies can be decomposed into pairs of symmetric windows around the origin, on which a smooth partition of unity is built. This and other constructions were obtained by one of our students E. Laeng [L]. Higher dimensional libraries can also be easily constructed, and there are generalizations of local trigonometric bases for certain manifolds.

References

anonymous InterNet ftp site at Yale University, `ceres.math.yale.edu[130.132.23.22]`.

R. R. Coifman and M. V. Wickerhauser, *Best-adapted wavelet packet bases*, preprint, Yale University, (February, 1990), available from [ceres] in `/pub/wavelets/baseb.tex`.

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

R. R. Coifman et Yves Meyer, *Remarques sur l'analyse de Fourier à fenêtre*, série I, C. R. Acad. Sci. Paris **312** (1991), 259–261.

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

Enrico Laeng, *Nouvelles bases orthogonales de* L^2, C. R. acad. sci. Paris (1989).

M. V. Wickerhauser, *Fast approximate Karhunen-Loéve expansions*, preprint, Yale University (May, 1990), available from [ceres] in `/pub/wavelets/fakle.tex`.

M. V. Wickerhauser, *Picture compression by best-basis sub-band coding*, preprint, Yale University (January, 1990), available from [ceres] in `/pub/wavelets/pic.tar`.

M. V. Wickerhauser, *Acoustic signal compression with wavelet packets*, preprint, Yale University (August, 1989), available from [ceres] in `/pub/wavelets/acoustic.tex`.