Instructions to the Reader

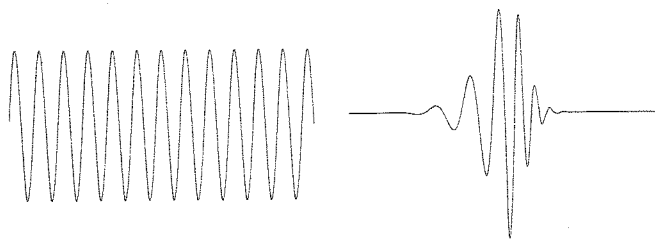
Although this book in arranged in a somewhat progressive order, starting with basic ideas and definitions, moving to a rather complete discussion of the basic wavelet system, and then on to generalizations, one should skip around when reading or studying from it. Depending on the background of the reader, he or she should skim over most of the book first, then go back and study parts in detail. The Introduction at the beginning and the Summary at the end should be continually consulted to gain or keep a perspective; similarly for the Table of Contents and Index. The MATLAB programs in the Appendix or the Wavelet Toolbox from Mathworks or other wavelet software should be used for continual experimentation. The list of references should be used to find proofs or detail not included here or to pursue research topics or applications. The theory and application of wavelets are still developing and in a state of rapid growth. We hope this book will help open the door to this fascinating new subject.

Chapter 1

Introduction to Wavelets

This chapter will provide an overview of the topics to be developed in the book. Its purpose is to present the ideas, goals, and outline of properties for an understanding of and ability to use wavelets and wavelet transforms. The details and more careful definitions are given later in the book.

A *wave* is usually defined as an oscillating function of time or space, such as a sinusoid. Fourier analysis is wave analysis. It expands signals or functions in terms of sinusoids (or, equivalently, complex exponentials) which has proven to be extremely valuable in mathematics, science, and engineering, especially for periodic, time-invariant, or stationary phenomena. A *wavelet* is a "small wave", which has its energy concentrated in time to give a tool for the analysis of transient, nonstationary, or time-varying phenomena. It still has the oscillating wave-like characteristic but also has the ability to allow simultaneous time and frequency analysis with a flexible mathematical foundation. This is illustrated in Figure 1.1 with the wave (sinusoid) oscillating with equal amplitude over $-\infty \leq t \leq \infty$ and, therefore, having infinite energy and with the wavelet having its finite energy concentrated around a point.



(a) A Sine Wave (b) Daubechies' Wavelet ψ_{D20}

Figure 1.1. A Wave and a Wavelet

We will take wavelets and use them in a series expansion of signals or functions much the same way a Fourier series uses the wave or sinusoid to represent a signal or function. The signals are functions of a continuous variable, which often represents time or distance. From this series expansion, we will develop a discrete-time version similar to the discrete Fourier transform where the signal is represented by a string of numbers where the numbers may be samples of a signal,

samples of another string of numbers, or inner products of a signal with some expansion set. Finally, we will briefly describe the continuous wavelet transform where both the signal and the transform are functions of continuous variables. This is analogous to the Fourier transform.

1.1 Wavelets and Wavelet Expansion Systems

Before delving into the details of wavelets and their properties, we need to get some idea of their general characteristics and what we are going to do with them [Swe96b].

What is a Wavelet Expansion or a Wavelet Transform?

A signal or function $f(t)$ can often be better analyzed, described, or processed if expressed as a linear decomposition by

$$f(t) = \sum_{\ell} a_\ell \, \psi_\ell(t) \tag{1.1}$$

where ℓ is an integer index for the finite or infinite sum, a_ℓ are the real-valued expansion coefficients, and $\psi_\ell(t)$ are a set of real-valued functions of t called the expansion set. If the expansion (1.1) is unique, the set is called a *basis* for the class of functions that can be so expressed. If the basis is orthogonal, meaning

$$\langle \psi_k(t), \psi_\ell(t) \rangle = \int \psi_k(t) \, \psi_\ell(t) \, dt = 0 \qquad k \neq \ell, \tag{1.2}$$

then the coefficients can be calculated by the *inner product*

$$a_k = \langle f(t), \psi_k(t) \rangle = \int f(t) \, \psi_k(t) \, dt. \tag{1.3}$$

One can see that substituting (1.1) into (1.3) and using (1.2) gives the single a_k coefficient. If the basis set is not orthogonal, then a dual basis set $\widetilde{\psi}_k(t)$ exists such that using (1.3) with the dual basis gives the desired coefficients. This will be developed in Chapter 2.

For a Fourier series, the orthogonal basis functions $\psi_k(t)$ are $\sin(k\omega_0 t)$ and $\cos(k\omega_0 t)$ with frequencies of $k\omega_0$. For a Taylor's series, the nonorthogonal basis functions are simple monomials t^k, and for many other expansions they are various polynomials. There are expansions that use splines and even fractals.

For the *wavelet expansion*, a two-parameter system is constructed such that (1.1) becomes

$$f(t) = \sum_{k} \sum_{j} a_{j,k} \, \psi_{j,k}(t) \tag{1.4}$$

where both j and k are integer indices and the $\psi_{j,k}(t)$ are the wavelet expansion functions that usually form an orthogonal basis.

The set of expansion coefficients $a_{j,k}$ are called the *discrete wavelet transform* (DWT) of $f(t)$ and (1.4) is the inverse transform.

What is a Wavelet System?

The wavelet expansion set is not unique. There are many different wavelets systems that can be used effectively, but all seem to have the following three general characteristics [Swe96b].

1. A wavelet system is a set of *building blocks* to construct or represent a signal or function. It is a two-dimensional expansion set (usually a basis) for some class of one- (or higher) dimensional signals. In other words, if the wavelet set is given by $\psi_{j,k}(t)$ for indices of $j, k = 1, 2, \cdots$, a linear expansion would be $f(t) = \sum_k \sum_j a_{j,k} \, \psi_{j,k}(t)$ for some set of coefficients $a_{j,k}$.

2. The wavelet expansion gives a time-frequency *localization* of the signal. This means most of the energy of the signal is well represented by a few expansion coefficients, $a_{j,k}$.

3. The calculation of the coefficients from the signal can be done *efficiently*. It turns out that many wavelet transforms (the set of expansion coefficients) can calculated with $O(N)$ operations. This means the number of floating-point multiplications and additions increase linearly with the length of the signal. More general wavelet transforms require $O(N \log(N))$ operations, the same as for the fast Fourier transform (FFT) [BP85].

Virtually all wavelet systems have these very general characteristics. Where the Fourier series maps a one-dimensional function of a continuous variable into a one-dimensional sequence of coefficients, the wavelet expansion maps it into a two-dimensional array of coefficients. We will see that it is this two-dimensional representation that allows localizing the signal in both time and frequency. A Fourier series expansion localizes in frequency in that if a Fourier series expansion of a signal has only one large coefficient, then the signal is essentially a single sinusoid at the frequency determined by the index of the coefficient. The simple time-domain representation of the signal itself gives the localization in time. If the signal is a simple pulse, the location of that pulse is the localization in time. A wavelet representation will give the location in both time and frequency simultaneously. Indeed, a wavelet representation is much like a musical score where the location of the notes tells when the tones occur and what their frequencies are.

More Specific Characteristics of Wavelet Systems

There are three additional characteristics [Swe96b, Dau92] that are more specific to wavelet expansions.

1. All so-called first-generation wavelet systems are generated from a single scaling function or wavelet by simple *scaling* and *translation*. The two-dimensional parameterization is achieved from the function (sometimes called the generating wavelet or mother wavelet) $\psi(t)$ by

$$\psi_{j,k}(t) = 2^{j/2} \, \psi(2^j t - k) \qquad j, k \in \mathbf{Z} \tag{1.5}$$

   where $\mathbf{Z}$ is the set of all integers and the factor $2^{j/2}$ maintains a constant norm independent of scale j. This parameterization of the time or space location by k and the frequency or scale (actually the logarithm of scale) by j turns out to be extraordinarily effective.

2. Almost all useful wavelet systems also satisfy the *multiresolution* conditions. This means that if a set of signals can be represented by a weighted sum of $\varphi(t-k)$, then a larger set (including the original) can be represented by a weighted sum of $\varphi(2t-k)$. In other words, if the basic expansion signals are made half as wide and translated in steps half as wide, they will represent a larger class of signals exactly or give a better approximation of any signal.

3. The lower resolution coefficients can be calculated from the higher resolution coefficients by a tree-structured algorithm called a *filter bank*. This allows a very efficient calculation of the expansion coefficients (also known as the discrete wavelet transform) and relates wavelet transforms to an older area in digital signal processing.

The operations of translation and scaling seem to be basic to many practical signals and signal-generating processes, and their use is one of the reasons that wavelets are efficient expansion functions. Figure 1.2 is a pictorial representation of the translation and scaling of a single mother wavelet described in (1.5). As the index k changes, the location of the wavelet moves along the horizontal axis. This allows the expansion to explicitly represent the location of events in time or space. As the index j changes, the shape of the wavelet changes in scale. This allows a representation of detail or resolution. Note that as the scale becomes finer (j larger), the steps in time become smaller. It is both the narrower wavelet and the smaller steps that allow representation of greater detail or higher resolution. For clarity, only every fourth term in the translation ($k = 1, 5, 9, 13, \cdots$) is shown, otherwise, the figure is a clutter. What is not illustrated here but is important is that the shape of the basic mother wavelet can also be changed. That is done during the design of the wavelet system and allows one set to well-represent a class of signals.

For the Fourier series and transform and for most signal expansion systems, the expansion functions (bases) are chosen, then the properties of the resulting transform are derived and

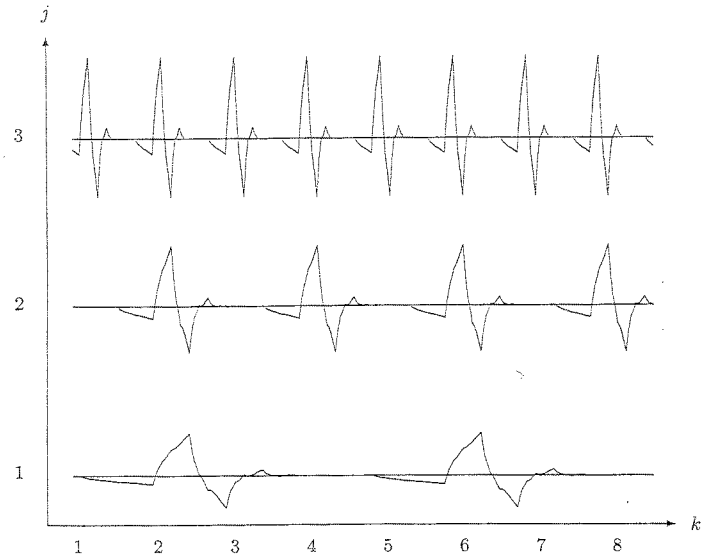


Figure 1.2. Translation (every fourth k) and Scaling of a Wavelet ψ_{D4}

analyzed. For the wavelet system, these properties or characteristics are mathematically required, then the resulting basis functions are derived. Because these constraints do not use all the degrees of freedom, other properties can be required to customize the wavelet system for a particular application. Once you decide on a Fourier series, the sinusoidal basis functions are completely set. That is not true for the wavelet. There are an infinity of very different wavelets that all satisfy the above properties. Indeed, the understanding and design of the wavelets is an important topic of this book.

Wavelet analysis is well-suited to transient signals. Fourier analysis is appropriate for periodic signals or for signals whose statistical characteristics do not change with time. It is the localizing property of wavelets that allow a wavelet expansion of a transient event to be modeled with a small number of coefficients. This turns out to be very useful in applications.

Haar Scaling Functions and Wavelets

The multiresolution formulation needs two closely related basic functions. In addition to the wavelet $\psi(t)$ that has been discussed (but not actually defined yet), we will need another basic function called the *scaling function* $\varphi(t)$. The reasons for needing this function and the details of the relations will be developed in the next chapter, but here we will simply use it in the wavelet expansion.

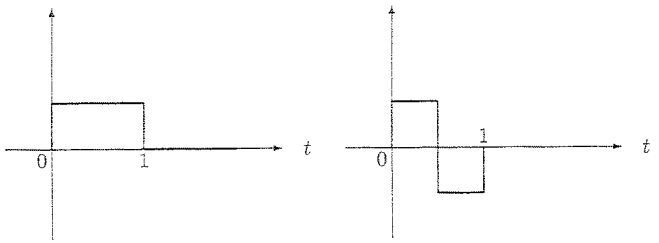
The simplest possible orthogonal wavelet system is generated from the Haar scaling function and wavelet. These are shown in Figure 1.3. Using a combination of these scaling functions and wavelets allows a large class of signals to be represented by

$$f(t) = \sum_{k=-\infty}^{\infty} c_k \, \varphi(t-k) + \sum_{k=-\infty}^{\infty} \sum_{j=0}^{\infty} d_{j,k} \, \psi(2^j t - k). \tag{1.6}$$

Haar [Haa10] showed this result in 1910, and we now know that wavelets are a generalization of his work. An example of a Haar system and expansion is given at the end of Chapter 2.

What do Wavelets Look Like?

All Fourier basis functions look alike. A high-frequency sine wave looks like a compressed low-frequency sine wave. A cosine wave is a sine wave translated by 90° or $\pi/2$ radians. It takes a



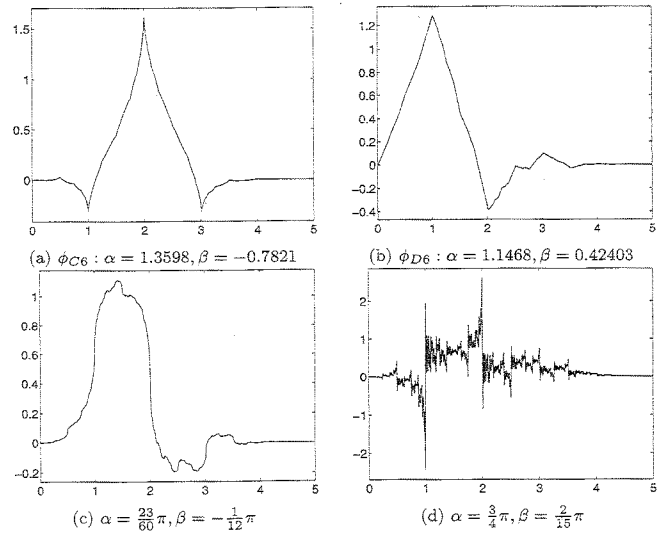
(a) $\phi(t)$ (b) $\psi(t)$

Figure 1.3. Haar Scaling Function and Wavelet

large number of Fourier components to represent a discontinuity or a sharp corner. In contrast, there are many different wavelets and some have sharp corners themselves.

To appreciate the special character of wavelets you should recognize that it was not until the late 1980's that some of the most useful basic wavelets were ever seen. Figure 1.4 illustrates four different scaling functions, each being zero outside of $0 < t < 6$ and each generating an orthogonal wavelet basis for all square integrable functions. This figure is also shown on the cover to this book.

Several more scaling functions and their associated wavelets are illustrated in later chapters, and the Haar wavelet is shown in Figure 1.3 and in detail at the end of Chapter 2.



(a) $\phi_{C6} : \alpha = 1.3598, \beta = -0.7821$ (b) $\phi_{D6} : \alpha = 1.1468, \beta = 0.42403$

(c) $\alpha = \frac{23}{60}\pi, \beta = -\frac{1}{12}\pi$ (d) $\alpha = \frac{3}{4}\pi, \beta = \frac{2}{15}\pi$

Figure 1.4. Example Scaling Functions (See Section 5.8 for the Meaning of α and β)

Why is Wavelet Analysis Effective?

Wavelet expansions and wavelet transforms have proven to be very efficient and effective in analyzing a very wide class of signals and phenomena. Why is this true? What are the properties that give this effectiveness?

1. The size of the wavelet expansion coefficients $a_{j,k}$ in (1.4) or $d_{j,k}$ in (1.6) drop off rapidly with j and k for a large class of signals. This property is called being an *unconditional basis* and it is why wavelets are so effective in signal and image compression, denoising, and detection. Donoho [Don93b, DJKP95b] showed that wavelets are near optimal for a wide class of signals for compression, denoising, and detection.

2. The wavelet expansion allows a more accurate local description and separation of signal characteristics. A Fourier coefficient represents a component that lasts for all time and, therefore, temporary events must be described by a phase characteristic that allows cancellation or reinforcement over large time periods. A wavelet expansion coefficient represents a component that is itself local and is easier to interpret. The wavelet expansion may allow a separation of components of a signal that overlap in both time and frequency.

3. Wavelets are adjustable and adaptable. Because there is not just one wavelet, they can be designed to fit individual applications. They are ideal for adaptive systems that adjust themselves to suit the signal.

4. The generation of wavelets and the calculation of the discrete wavelet transform is well matched to the digital computer. We will later see that the defining equation for a wavelet uses no calculus. There are no derivatives or integrals, just multiplications and additions—operations that are basic to a digital computer.

While some of these details may not be clear at this point, they should point to the issues that are important to both theory and application and give reasons for the detailed development that follows in this and other books.

1.2 The Discrete Wavelet Transform

This two-variable set of basis functions is used in a way similar to the short-time Fourier transform, the Gabor transform, or the Wigner distribution for time-frequency analysis [Coh89, Coh95, HB92]. Our goal is to generate a set of expansion functions such that any signal in $L^2(\mathbf{R})$ (the space of square integrable functions) can be represented by the series

$$f(t) = \sum_{j,k} a_{j,k}\, 2^{j/2}\, \psi(2^j t - k) \tag{1.7}$$

or, using (1.5), as

$$f(t) = \sum_{j,k} a_{j,k}\, \psi_{j,k}(t) \tag{1.8}$$

where the two-dimensional set of coefficients $a_{j,k}$ is called the *discrete wavelet transform* (DWT) of $f(t)$. A more specific form indicating how the $a_{j,k}$'s are calculated can be written using inner products as

$$f(t) = \sum_{j,k} \langle \psi_{j,k}(t), f(t) \rangle\, \psi_{j,k}(t) \tag{1.9}$$

if the $\psi_{j,k}(t)$ form an orthonormal basis[1] for the space of signals of interest [Dau92]. The inner product is usually defined as

$$\langle x(t), y(t) \rangle = \int x^*(t)\, y(t)\, dt. \tag{1.10}$$

[1] Bases and tight frames are defined in Chapter 4.

The goal of most expansions of a function or signal is to have the coefficients of the expansion $a_{j,k}$ give more useful information about the signal than is directly obvious from the signal itself. A second goal is to have most of the coefficients be zero or very small. This is what is called a *sparse* representation and is extremely important in applications for statistical estimation and detection, data compression, nonlinear noise reduction, and fast algorithms.

Although this expansion is called the discrete wavelet transform (DWT), it probably should be called a wavelet series since it is a series expansion which maps a function of a continuous variable into a sequence of coefficients much the same way the Fourier series does. However, that is not the convention.

This wavelet series expansion is in terms of two indices, the time translation k and the scaling index j. For the Fourier series, there are only two possible values of k, zero and $\pi/2$, which give the sine terms and the cosine terms. The values j give the frequency harmonics. In other words, the Fourier series is also a two-dimensional expansion, but that is not seen in the exponential form and generally not noticed in the trigonometric form.

The DWT of a signal is somewhat difficult to illustrate because it is a function of two variables or indices, but we will show the DWT of a simple pulse in Figure 1.5 to illustrate the localization of the transform. Other displays will be developed in the next chapter.

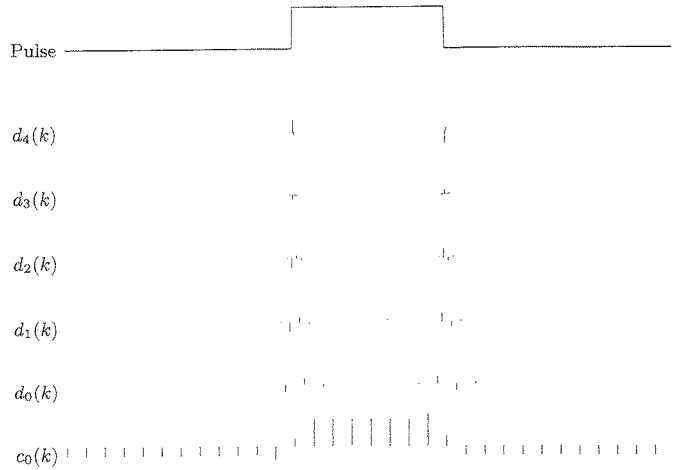


Figure 1.5. Discrete Wavelet Transform of a Pulse, using ψ_{D6} with a Gain of $\sqrt{2}$ for Each Higher Scale.

1.3 The Discrete-Time and Continuous Wavelet Transforms

If the signal is itself a sequence of numbers, perhaps samples of some function of a continuous variable or perhaps a set of inner products, the expansion of that signal is called a discrete-time wavelet transform (DTWT). It maps a sequence of numbers into a sequence of numbers much the same way the discrete Fourier transform (DFT) does. It does not, however, require the signal to be finite in duration or periodic as the DFT does. To be consistent with Fourier terminology, it probably should be called the discrete-time wavelet series, but this is not the convention. If the discrete-time signal is finite in length, the transform can be represented by a finite matrix. This formulation of a series expansion of a discrete-time signal is what filter bank methods accomplish [Vai92, VK95] and is developed in Chapter 8 of this book.

If the signal is a function of a continuous variable and a transform that is a function of two continuous variables is desired, the continuous wavelet transform (CWT) can be defined by

$$F(a,b) = \int f(t)\, w(\frac{t-a}{b})\, dt \tag{1.11}$$

with an inverse transform of

$$f(t) = \iint F(a,b)\, w(\frac{t-a}{b})\, da\, db \tag{1.12}$$

where $w(t)$ is the basic wavelet and $a, b \in \mathbf{R}$ are real continuous variables. Admissibility conditions for the wavelet $w(t)$ to support this invertible transform is discussed by Daubechies [Dau92], Heil and Walnut [HW89], and others and is briefly developed in Section 7.8 of this book. It is analogous to the Fourier transform or Fourier integral.

1.4 Exercises and Experiments

As the ideas about wavelets and wavelet transforms are developed in this book, it will be very helpful to experiment using the Matlab programs in the appendix of this book or in the MATLAB Toolbox [MMOP96]. An effort has been made to use the same notation in the programs in Appendix C as is used in the formulas in the book so that going over the programs can help in understanding the theory and vice versa.

1.5 This Chapter

This chapter has tried to set the stage for a careful introduction to both the theory and use of wavelets and wavelet transforms. We have presented the most basic characteristics of wavelets and tried to give a feeling of how and why they work in order to motivate and give direction and structure to the following material.

The next chapter will present the idea of multiresolution, out of which will develop the scaling function as well as the wavelet. This is followed by a discussion of how to calculate the wavelet expansion coefficients using filter banks from digital signal processing. Next, a more detailed development of the theory and properties of scaling functions, wavelets, and wavelet transforms is given followed by a chapter on the design of wavelet systems. Chapter 8 gives a detailed development of wavelet theory in terms of filter banks.

The earlier part of the book carefully develops the basic wavelet system and the later part develops several important generalizations, but in a less detailed form.

Chapter 2

A Multiresolution Formulation of Wavelet Systems

Both the mathematics and the practical interpretations of wavelets seem to be best served by using the concept of resolution [Mey93, Mal89b, Mal89c, Dau92] to define the effects of changing scale. To do this, we will start with a *scaling function* $\varphi(t)$ rather than directly with the wavelet $\psi(t)$. After the scaling function is defined from the concept of resolution, the wavelet functions will be derived from it. This chapter will give a rather intuitive development of these ideas, which will be followed by more rigorous arguments in Chapter 5.

This multiresolution formulation is obviously designed to represent signals where a single event is decomposed into finer and finer detail, but it turns out also to be valuable in representing signals where a time-frequency or time-scale description is desired even if no concept of resolution is needed. However, there are other cases where multiresolution is not appropriate, such as for the short-time Fourier transform or Gabor transform or for local sine or cosine bases or lapped orthogonal transforms, which are all discussed briefly later in this book.

2.1 Signal Spaces

In order to talk about the collection of functions or signals that can be represented by a sum of scaling functions and/or wavelets, we need some ideas and terminology from functional analysis. If these concepts are not familiar to you or the information in this section is not sufficient, you may want to skip ahead and read Chapter 5 or [VD95].

A *function space* is a linear vector space (finite or infinite dimensional) where the vectors are functions, the scalars are real numbers (sometime complex numbers), and scalar multiplication and vector addition are similar to that done in (1.1). The *inner product* is a scalar a obtained from two vectors, $f(t)$ and $g(t)$, by an integral. It is denoted

$$a = \langle f(t), g(t) \rangle = \int f^*(t)\, g(t)\, dt \tag{2.1}$$

with the range of integration depending on the signal class being considered. This inner product defines a *norm* or "length" of a vector which is denoted and defined by

$$\|f\| = \sqrt{|\langle f, f \rangle|} \tag{2.2}$$

which is a simple generalization of the geometric operations and definitions in three-dimensional Euclidean space. Two signals (vectors) with non-zero norms are called *orthogonal* if their inner product is zero. For example, with the Fourier series, we see that $\sin(t)$ is orthogonal to $\sin(2t)$.

A space that is particularly important in signal processing is call $L^2(\mathbf{R})$. This is the space of all functions $f(t)$ with a well defined integral of the square of the modulus of the function. The "L" signifies a Lebesque integral, the "2" denotes the integral of the square of the modulus of the function, and $\mathbf{R}$ states that the independent variable of integration t is a number over the whole real line. For a function $g(t)$ to be a member of that space is denoted: $g \in L^2(\mathbf{R})$ or simply $g \in L^2$.

Although most of the definitions and derivations are in terms of signals that are in L^2, many of the results hold for larger classes of signals. For example, polynomials are not in L^2 but can be expanded over any finite domain by most wavelet systems.

In order to develop the wavelet expansion described in (1.5), we will need the idea of an expansion set or a basis set. If we start with the vector space of signals, $\mathcal{S}$, then if any $f(t) \in \mathcal{S}$ can be expressed as $f(t) = \sum_k a_k \varphi_k(t)$, the set of functions $\varphi_k(t)$ are called an expansion set for the space $\mathcal{S}$. If the representation is unique, the set is a basis. Alternatively, one could start with the expansion set or basis set and define the space $\mathcal{S}$ as the set of all functions that can be expressed by $f(t) = \sum_k a_k \varphi_k(t)$. This is called the *span* of the basis set. In several cases, the signal spaces that we will need are actually the *closure* of the space spanned by the basis set. That means the space contains not only all signals that can be expressed by a linear combination of the basis functions, but also the signals which are the limit of these infinite expansions. The closure of a space is usually denoted by an over-line.

2.2 The Scaling Function

In order to use the idea of multiresolution, we will start by defining the scaling function and then define the wavelet in terms of it. As described for the wavelet in the previous chapter, we define a set of scaling functions in terms of integer translates of the basic scaling function by

$$\varphi_k(t) = \varphi(t-k) \qquad k \in \mathbf{Z} \qquad \varphi \in L^2. \tag{2.3}$$

The subspace of $L^2(\mathbf{R})$ spanned by these functions is defined as

$$\mathcal{V}_0 = \overline{\operatorname{Span}_k\{\varphi_k(t)\}} \tag{2.4}$$

for all integers k from minus infinity to infinity. The over-bar denotes closure. This means that

$$f(t) = \sum_k a_k \varphi_k(t) \quad \text{for any} \quad f(t) \in \mathcal{V}_0. \tag{2.5}$$

One can generally increase the size of the subspace spanned by changing the time scale of the scaling functions. A two-dimensional family of functions is generated from the basic scaling function by scaling and translation by

$$\varphi_{j,k}(t) = 2^{j/2} \varphi(2^j t - k) \tag{2.6}$$

↑ normalized

whose span over k is

$$\mathcal{V}_j = \overline{\underset{k}{\text{Span}}\{\varphi_k(2^j t)\}} = \overline{\underset{k}{\text{Span}}\{\varphi_{j,k}(t)\}} \tag{2.7}$$

for all integers $k \in \mathbf{Z}$. This means that if $f(t) \in \mathcal{V}_j$, then it can be expressed as

$$f(t) = \sum_k a_k \varphi(2^j t + k). \tag{2.8}$$

For $j > 0$, the span can be larger since $\varphi_{j,k}(t)$ is narrower and is translated in smaller steps. It, therefore, can represent finer detail. For $j < 0$, $\varphi_{j,k}(t)$ is wider and is translated in larger steps. So these wider scaling functions can represent only coarse information, and the space they span is smaller. Another way to think about the effects of a change of scale is in terms of resolution. If one talks about photographic or optical resolution, then this idea of scale is the same as resolving power.

Multiresolution Analysis

In order to agree with our intuitive ideas of scale or resolution, we formulate the basic requirement of multiresolution analysis (MRA) [Mal89c] by requiring a nesting of the spanned spaces as

$$\cdots \subset \mathcal{V}_{-2} \subset \mathcal{V}_{-1} \subset \mathcal{V}_0 \subset \mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \subset L^2 \tag{2.9}$$

or

$$\mathcal{V}_j \subset \mathcal{V}_{j+1} \quad \text{for all} \quad j \in \mathbf{Z} \tag{2.10}$$

with

$$\mathcal{V}_{-\infty} = \{0\}, \qquad \mathcal{V}_{\infty} = L^2. \tag{2.11}$$

The space that contains high resolution signals will contain those of lower resolution also.

Because of the definition of $\mathcal{V}_j$, the spaces have to satisfy a natural scaling condition

$$f(t) \in \mathcal{V}_j \quad \Leftrightarrow \quad f(2t) \in \mathcal{V}_{j+1} \tag{2.12}$$

which insures elements in a space are simply scaled versions of the elements in the next space. The relationship of the spanned spaces is illustrated in Figure 2.1.

The nesting of the spans of $\varphi(2^j t - k)$, denoted by $\mathcal{V}_j$ and shown in (2.9) and (2.12) and graphically illustrated in Figure 2.1, is achieved by requiring that $\varphi(t) \in \mathcal{V}_1$, which means that if $\varphi(t)$ is in $\mathcal{V}_0$, it is also in $\mathcal{V}_1$, the space spanned by $\varphi(2t)$. This means $\varphi(t)$ can be expressed in terms of a weighted sum of shifted $\varphi(2t)$ as

$$\boxed{\varphi(t) = \sum_n h(n)\sqrt{2}\,\varphi(2t - n), \quad n \in \mathbf{Z}} \tag{2.13}$$

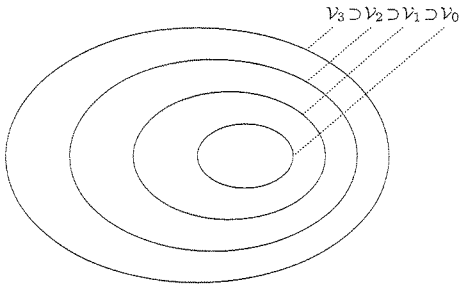


Figure 2.1. Nested Vector Spaces Spanned by the Scaling Functions

where the coefficients $h(n)$ are a sequence of real or perhaps complex numbers called the scaling function coefficients (or the scaling filter or the scaling vector) and the $\sqrt{2}$ maintains the norm of the scaling function with the scale of two.

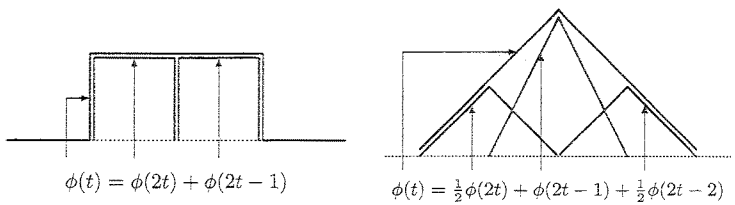
This recursive equation is fundamental to the theory of the scaling functions and is, in some ways, analogous to a differential equation with coefficients $h(n)$ and solution $\varphi(t)$ that may or may not exist or be unique. The equation is referred to by different names to describe different interpretations or points of view. It is called the refinement equation, the multiresolution analysis (MRA) equation, or the dilation equation.

The Haar scaling function is the simple unit-width, unit-height pulse function $\varphi(t)$ shown in Figure 2.2, and it is obvious that $\varphi(2t)$ can be used to construct $\varphi(t)$ by

$$\varphi(t) = \varphi(2t) + \varphi(2t - 1) \tag{2.14}$$

which means (2.13) is satisfied for coefficients $h(0) = 1/\sqrt{2}$; $h(1) = 1/\sqrt{2}$.

The triangle scaling function (also a first order spline) in Figure 2.2 satisfies (2.13) for $h(0) = \frac{1}{2\sqrt{2}}$, $h(1) = \frac{1}{\sqrt{2}}$, $h(2) = \frac{1}{\sqrt{2}}$, and the Daubechies scaling function shown in the first part of



(a) Haar (same as φ_{D2}) (b) Triangle (same as φ_{S1})

Figure 2.2. Haar and Triangle Scaling Functions

Figure 6.1 satisfies (2.13) for $h = \{0.483, 0.8365, 0.2241, -0.1294\}$ as do all scaling functions for their corresponding scaling coefficients. Indeed, the design of wavelet systems is the choosing of the coefficients $h(n)$ and that is developed later.

2.3 The Wavelet Functions

The important features of a signal can better be described or parameterized, not by using $\varphi_{j,k}(t)$ and increasing j to increase the size of the subspace spanned by the scaling functions, but by defining a slightly different set of functions $\psi_{j,k}(t)$ that span the *differences* between the spaces spanned by the various scales of the scaling function. These functions are the wavelets discussed in the introduction of this book.

There are several advantages to requiring that the scaling functions and wavelets be orthogonal. Orthogonal basis functions allow simple calculation of expansion coefficients and have a Parseval's theorem that allows a partitioning of the signal energy in the wavelet transform domain. The orthogonal complement of $\mathcal{V}_j$ in $\mathcal{V}_{j+1}$ is defined as $\mathcal{W}_j$. This means that all members of $\mathcal{V}_j$ are orthogonal to all members of $\mathcal{W}_j$. We require

$$\langle \varphi_{j,k}(t), \psi_{j,\ell}(t) \rangle = \int \varphi_{j,k}(t)\, \psi_{j,\ell}(t)\, dt = 0 \tag{2.15}$$

for all appropriate $j, k, \ell \in \mathbf{Z}$.

The relationship of the various subspaces can be seen from the following expressions. From (2.9) we see that we may start at any $\mathcal{V}_j$, say at $j = 0$, and write

$$\mathcal{V}_0 \subset \mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \subset L^2. \tag{2.16}$$

We now define the wavelet spanned subspace $\mathcal{W}_0$ such that

$$\mathcal{V}_1 = \mathcal{V}_0 \oplus \mathcal{W}_0 \tag{2.17}$$

which extends to

$$\mathcal{V}_2 = \mathcal{V}_0 \oplus \mathcal{W}_0 \oplus \mathcal{W}_1. \tag{2.18}$$

In general this gives

$$L^2 = \mathcal{V}_0 \oplus \mathcal{W}_0 \oplus \mathcal{W}_1 \oplus \cdots \tag{2.19}$$

when $\mathcal{V}_0$ is the initial space spanned by the scaling function $\varphi(t-k)$. Figure 2.3 pictorially shows the nesting of the scaling function spaces $\mathcal{V}_j$ for different scales j and how the wavelet spaces are the disjoint differences (except for the zero element) or, the orthogonal complements.

The scale of the initial space is arbitrary and could be chosen at a higher resolution of, say, $j = 10$ to give

$$L^2 = \mathcal{V}_{10} \oplus \mathcal{W}_{10} \oplus \mathcal{W}_{11} \oplus \cdots \tag{2.20}$$

or at a lower resolution such as $j = -5$ to give

$$L^2 = \mathcal{V}_{-5} \oplus \mathcal{W}_{-5} \oplus \mathcal{W}_{-4} \oplus \cdots \tag{2.21}$$

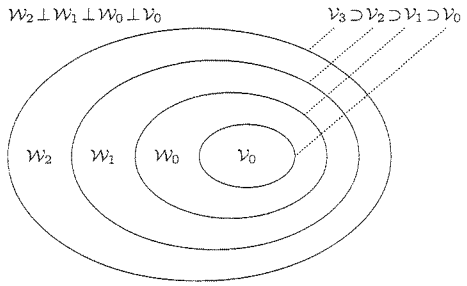


Figure 2.3. Scaling Function and Wavelet Vector Spaces

or at even $j = -\infty$ where (2.19) becomes

$$L^2 = \cdots \oplus \mathcal{W}_{-2} \oplus \mathcal{W}_{-1} \oplus \mathcal{W}_0 \oplus \mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \cdots \tag{2.22}$$

eliminating the scaling space altogether and allowing an expansion of the form in (1.9).

Another way to describe the relation of $\mathcal{V}_0$ to the wavelet spaces is noting

$$\mathcal{W}_{-\infty} \oplus \cdots \oplus \mathcal{W}_{-1} = \mathcal{V}_0 \tag{2.23}$$

which again shows that the scale of the scaling space can be chosen arbitrarily. In practice, it is usually chosen to represent the coarsest detail of interest in a signal.

Since these wavelets reside in the space spanned by the next narrower scaling function, $\mathcal{W}_0 \subset \mathcal{V}_1$, they can be represented by a weighted sum of shifted scaling function $\varphi(2t)$ defined in (2.13) by

$$\boxed{\psi(t) = \sum_n h_1(n)\, \sqrt{2}\, \varphi(2t - n), \quad n \in \mathbf{Z}} \tag{2.24}$$

for some set of coefficients $h_1(n)$. From the requirement that the wavelets span the "difference" or orthogonal complement spaces, and the orthogonality of integer translates of the wavelet (or scaling function), it is shown in the Appendix in (12.48) that the wavelet coefficients (modulo translations by integer multiples of two) are required by orthogonality to be related to the scaling function coefficients by

$$h_1(n) = (-1)^n h(1-n). \tag{2.25}$$

One example for a finite even length-N $h(n)$ could be

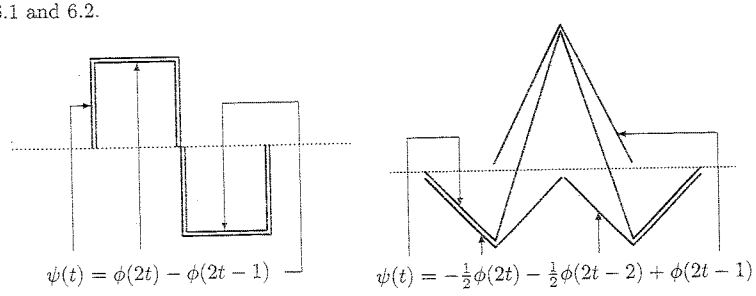
$$h_1(n) = (-1)^n h(N-1-n). \tag{2.26}$$

The function generated by (2.24) gives the prototype or mother wavelet $\psi(t)$ for a class of expansion functions of the form

$$\psi_{j,k}(t) = 2^{j/2}\, \psi(2^j t - k) \tag{2.27}$$

where 2^j is the scaling of t (j is the $\log_2$ of the scale), $2^{-j}k$ is the translation in t, and $2^{j/2}$ maintains the (perhaps unity) L^2 norm of the wavelet at different scales.

The Haar and triangle wavelets that are associated with the scaling functions in Figure 2.2 are shown in Figure 2.4. For the Haar wavelet, the coefficients in (2.24) are $h_1(0) = 1/\sqrt{2}$, $h_1(1) = -1/\sqrt{2}$ which satisfy (2.25). The Daubechies wavelets associated with the scaling functions in Figure 6.1 are shown in Figure 6.2 with corresponding coefficients given later in the book in Tables 6.1 and 6.2.



(a) Haar (same as ψ_{D2}) (b) Triangle (same as ψ_{S1})

Figure 2.4. Haar and Triangle Wavelets

We have now constructed a set of functions $\varphi_k(t)$ and $\psi_{j,k}(t)$ that could span all of $L^2(\mathbf{R})$. According to (2.19), any function $g(t) \in L^2(\mathbf{R})$ could be written

$$g(t) = \sum_{k=-\infty}^{\infty} c(k)\,\varphi_k(t) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty} d(j,k)\,\psi_{j,k}(t) \tag{2.28}$$

as a series expansion in terms of the scaling function and wavelets.

In this expansion, the first summation in (2.28) gives a function that is a low resolution or coarse approximation of $g(t)$. For each increasing index j in the second summation, a higher or finer resolution function is added, which adds increasing detail. This is somewhat analogous to a Fourier series where the higher frequency terms contain the detail of the signal.

Later in this book, we will develop the property of having these expansion functions form an orthonormal basis or a tight frame, which allows the coefficients to be calculated by inner products as

$$c(k) = c_0(k) = \langle g(t), \varphi_k(t)\rangle = \int g(t)\,\varphi_k(t)\,dt \tag{2.29}$$

and

$$d_j(k) = d(j,k) = \langle g(t), \psi_{j,k}(t)\rangle = \int g(t)\,\psi_{j,k}(t)\,dt. \tag{2.30}$$

The coefficient $d(j,k)$ is sometimes written as $d_j(k)$ to emphasize the difference between the time translation index k and the scale parameter j. The coefficient $c(k)$ is also sometimes written as $c_j(k)$ or $c(j,k)$ if a more general "starting scale" other than $j = 0$ for the lower limit on the sum in (2.28) is used.

It is important at this point to recognize the relationship of the scaling function part of the expansion (2.28) to the wavelet part of the expansion. From the representation of the nested spaces in (2.19) we see that the scaling function can be defined at any scale j. Equation (2.28) uses $j = 0$ to denote the family of scaling functions.

You may want to examine the Haar system example at the end of this chapter just now to see these features illustrated.

2.4 The Discrete Wavelet Transform

Since

$$L^2 = \mathcal{V}_{j_0} \oplus \mathcal{W}_{j_0} \oplus \mathcal{W}_{j_0+1} \oplus \cdots \tag{2.31}$$

using (2.6) and (2.27), a more general statement of the expansion (2.28) can be given by

$$g(t) = \sum_k c_{j_0}(k)\,2^{j_0/2}\,\varphi(2^{j_0}t - k) + \sum_k \sum_{j=j_0}^{\infty} d_j(k)\,2^{j/2}\,\psi(2^j t - k) \tag{2.32}$$

or

$$g(t) = \sum_k c_{j_0}(k)\,\varphi_{j_0,k}(t) + \sum_k \sum_{j=j_0}^{\infty} d_j(k)\,\psi_{j,k}(t) \tag{2.33}$$

where j_0 could be zero as in (2.19) and (2.28), it could be ten as in (2.20), or it could be negative infinity as in (1.8) and (2.22) where no scaling functions are used. The choice of j_0 sets the coarsest scale whose space is spanned by $\varphi_{j_0,k}(t)$. The rest of $L^2(\mathbf{R})$ is spanned by the wavelets which provide the high resolution details of the signal. In practice where one is given only the samples of a signal, not the signal itself, there is a highest resolution when the finest scale is the sample level.

The coefficients in this wavelet expansion are called the *discrete wavelet transform* (DWT) of the signal $g(t)$. If certain conditions described later are satisfied, these wavelet coefficients completely describe the original signal and can be used in a way similar to Fourier series coefficients for analysis, description, approximation, and filtering. If the wavelet system is orthogonal, these coefficients can be calculated by inner products

$$c_j(k) = \langle g(t), \varphi_{j,k}(t) = \int g(t)\,\varphi_{j,k}(t)\,dt \tag{2.34}$$

and

$$d_j(k) = \langle g(t), \psi_{j,k}(t) = \int g(t)\,\psi_{j,k}(t)\,dt. \tag{2.35}$$

If the scaling function is well-behaved, then at a high scale, the scaling is similar to a Dirac delta function and the inner product simply samples the function. In other words, at high enough resolution, samples of the signal are very close to the scaling coefficients. More is said about this later. It has been shown [Don93b] that wavelet systems form an unconditional basis for a large class of signals. That is discussed in Chapter 5 but means that even for the worst case signal in the class, the wavelet expansion coefficients drop off rapidly as j and k increase. This is why the DWT is efficient for signal and image compression.

The DWT is similar to a Fourier series but, in many ways, is much more flexible and informative. It can be made periodic like a Fourier series to represent periodic signals efficiently.

However, unlike a Fourier series, it can be used directly on non-periodic transient signals with excellent results. An example of the DWT of a pulse was illustrated in Figure 3.3. Other examples are illustrated just after the next section.

2.5 A Parseval's Theorem

If the scaling functions and wavelets form an orthonormal basis[1], there is a Parseval's theorem that relates the energy of the signal $g(t)$ to the energy in each of the components and their wavelet coefficients. That is one reason why orthonormality is important.

For the general wavelet expansion of (2.28) or (2.33), Parseval's theorem is

$$\int |g(t)|^2 \, dt = \sum_{l=-\infty}^{\infty} |c(l)|^2 + \sum_{j=0}^{\infty} \sum_{k=-\infty}^{\infty} |d_j(k)|^2 \tag{2.36}$$

with the energy in the expansion domain partitioned in time by k and in scale by j. Indeed, it is this partitioning of the time-scale parameter plane that describes the DWT. If the expansion system is a tight frame, there is a constant multiplier in (2.36) caused by the redundancy.

Daubechies [Dau88a, Dau92] showed that it is possible for the scaling function and the wavelets to have compact support (i.e., be nonzero only over a finite region) and to be orthonormal. This makes possible the time localization that we desire. We now have a framework for describing signals that has features of short-time Fourier analysis and of Gabor-based analysis but using a new variable, scale. For the short-time Fourier transform, orthogonality and good time-frequency resolution are incompatible according to the Balian-Low-Coifman-Semmes theorem [Dau90, Sie86]. More precisely, if the short-time Fourier transform is orthogonal, either the time or the frequency resolution is poor and the trade-off is inflexible. This is not the case for the wavelet transform. Also, note that there is a variety of scaling functions and wavelets that can be obtained by choosing different coefficients $h(n)$ in (2.13).

Donoho [Don93b] has noted that wavelets are an unconditional basis for a very wide class of signals. This means wavelet expansions of signals have coefficients that drop off rapidly and therefore the signal can be efficiently represented by a small number of them.

We have first developed the basic ideas of the discrete wavelet system using a scaling multiplier of 2 in the defining equation (2.13). This is called a *two-band wavelet system* because of the two channels or bands in the related filter banks discussed in Chapters 3 and 8. It is also possible to define a more general discrete wavelet system using $\varphi(t) = \sum_n h(n) \sqrt{M} \, \varphi(Mt - n)$ where M is an integer [SHGB93]. This is discussed in Section 7.2. The details of numerically calculating the DWT are discussed in Chapter 9 where special forms for periodic signals are used.

2.6 Display of the Discrete Wavelet Transform and the Wavelet Expansion

It is important to have an informative way of displaying or visualizing the wavelet expansion and transform. This is complicated in that the DWT is a real-valued function of two integer indices and, therefore, needs a two-dimensional display or plot. This problem is somewhat analogous to plotting the Fourier transform, which is a complex-valued function.

There seem to be five displays that show the various characteristics of the DWT well:

1. The most basic time-domain description of a signal is the signal itself (or, for most cases, samples of the signal) but it gives no frequency or scale information. A very interesting property of the DWT (and one different from the Fourier series) is for a high starting scale j_0 in (2.33), samples of the signal are the DWT at that scale. This is an extreme case, but it shows the flexibility of the DWT and will be explained later.

2. The most basic wavelet-domain description is a three-dimensional plot of the expansion coefficients or DWT values $c(k)$ and $d_j(k)$ over the j, k plane. This is difficult to do on a two-dimensional page or display screen, but we show a form of that in Figures 2.5 and 2.8.

3. A very informative picture of the effects of scale can be shown by generating time functions $f_j(t)$ at each scale by summing (2.28) over k so that

$$f(t) = f_{j_0} + \sum_j f_j(t) \tag{2.37}$$

where

$$f_{j_0} = \sum_k c(k) \, \varphi(t - k) \tag{2.38}$$

and

$$f_j(t) = \sum_k d_j(k) \, 2^{j/2} \, \psi(2^j t - k). \tag{2.39}$$

This illustrates the components of the signal at each scale and is shown in Figures 2.7 and 2.10.

4. Another illustration that shows the time localization of the wavelet expansion is obtained by generating time functions $f_k(t)$ at each translation by summing (2.28) over k so that

$$f(t) = \sum_k f_k(t) \tag{2.40}$$

where

$$f_k(t) = c(k) \, \varphi(t - k) + \sum_j d_j(k) \, 2^{j/2} \, \psi(2^j t - k). \tag{2.41}$$

This illustrates the components of the signal at each integer translation.

5. There is another rather different display based on a partitioning of the time-scale plane as if the time translation index and scale index were continuous variables. This display is called "tiling the time-frequency plane." Because it is a different type of display and is developed and illustrated in Chapter 9, it will not be illustrated here.

Experimentation with these displays can be very informative in terms of the properties and capabilities of the wavelet transform, the effects of particular wavelet systems, and the way a

2.7 Examples of Wavelet Expansions

In this section, we will try to show the way a wavelet expansion decomposes a signal and what the components look like at different scales. These expansions use what is called a length-8 Daubechies basic wavelet (developed in Chapter 6), but that is not the main point here. The local nature of the wavelet decomposition is the topic of this section.

These examples are rather standard ones, some taken from David Donoho's papers and web page. The first is a decomposition of a piecewise linear function to show how edges and constants are handled. A characteristic of Daubechies systems is that low order polynomials are completely contained in the scaling function spaces $\mathcal{V}_j$ and need no wavelets. This means that when a section of a signal is a section of a polynomial (such as a straight line), there are no wavelet expansion coefficients $d_j(k)$, but when the calculation of the expansion coefficients overlaps an edge, there is a wavelet component. This is illustrated well in Figure 2.6 where the high resolution scales gives a very accurate location of the edges and this spreads out over k at the lower scales. This gives a hint of how the DWT could be used for edge detection and how the large number of small or zero expansion coefficients could be used for compression.

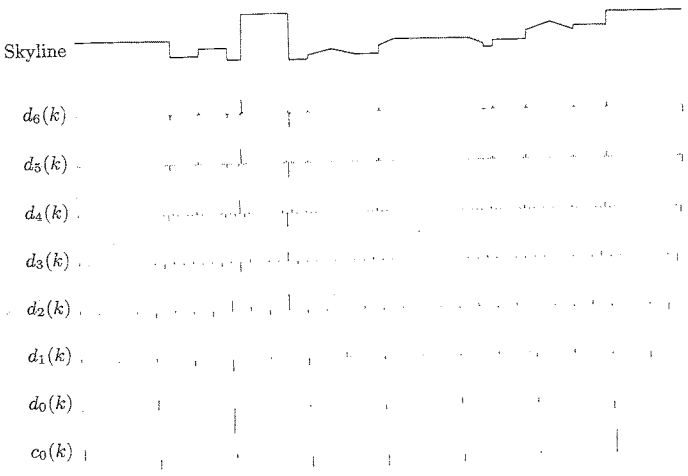


Figure 2.5. Discrete Wavelet Transform of the Houston Skyline, using $\psi_{D8'}$ with a Gain of $\sqrt{2}$ for Each Higher Scale

Figure 2.6 shows the approximations of the skyline signal in the various scaling function spaces $\mathcal{V}_j$. This illustrates just how the approximations progress, giving more and more resolution at higher scales. The fact that the higher scales give more detail is similar to Fourier methods, but

the components of the signal that exist in the wavelet spaces $\mathcal{W}_j$ at different scales j. This shows the same expansion as Figure 2.6, but with the wavelet components given separately rather than being cumulatively added to the scaling function. Notice how the large objects show up at the lower resolution. Groups of buildings and individual buildings are resolved according to their width. The edges, however, are located at the higher resolutions and are located very accurately.

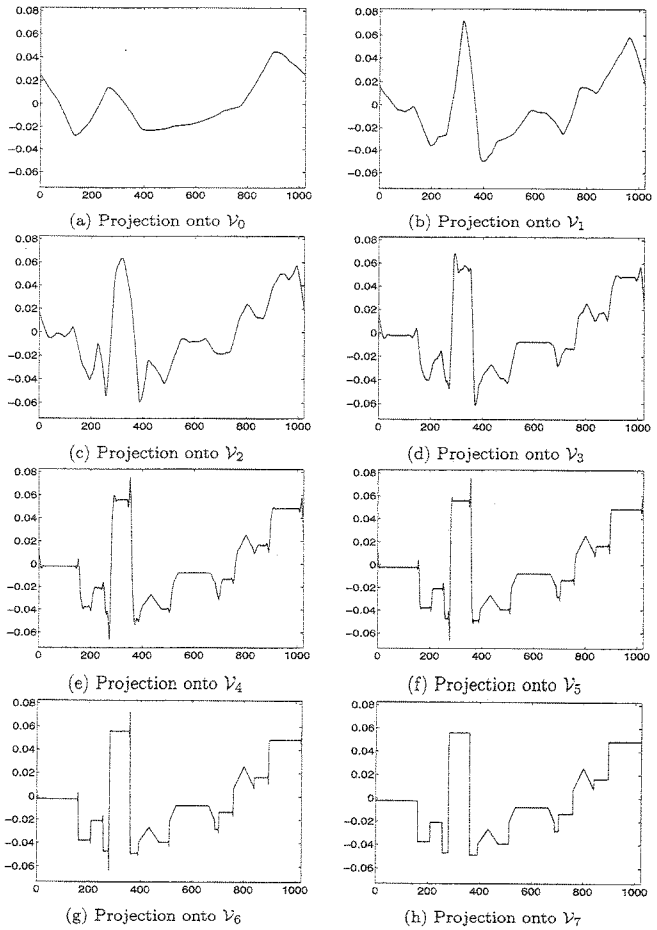


Figure 2.6. Projection of the Houston Skyline Signal onto $\mathcal{V}$ Spaces using $\phi_{D8'}$

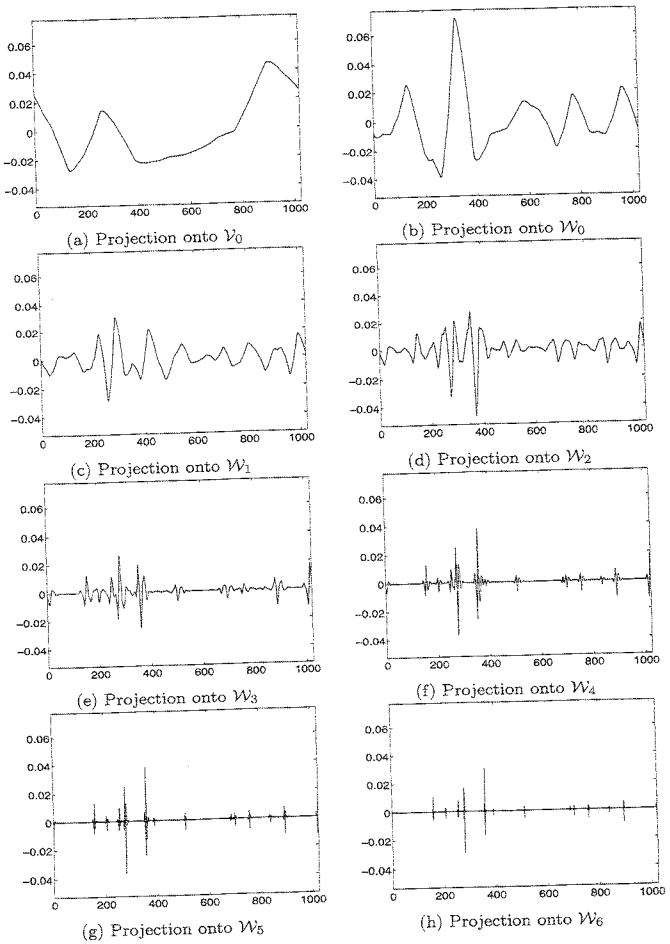


Figure 2.7. Projection of the Houston Skyline Signal onto $\mathcal{W}$ Spaces using $\psi_{D8'}$

The second example uses a chirp or doppler signal to illustrate how a time-varying frequency is described by the scale decomposition. Figure 2.8 gives the coefficients of the DWT directly as a function of j and k. Notice how the location in k tracks the frequencies in the signal in a way the Fourier transform cannot. Figures 2.9 and 2.10 show the scaling function approximations

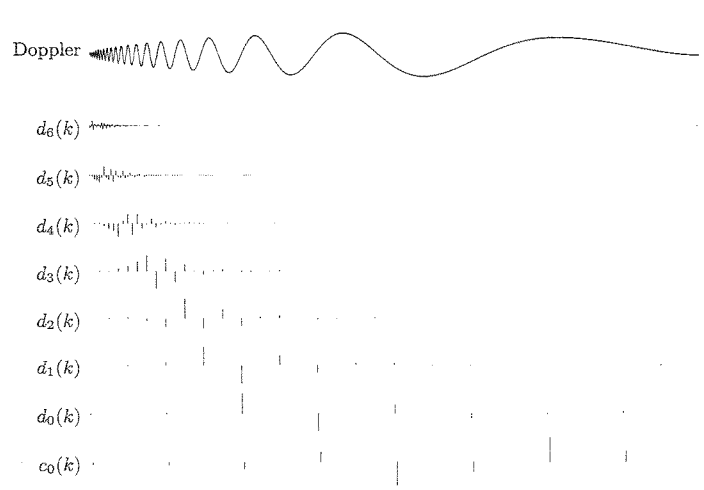


Figure 2.8. Discrete Wavelet Transform of a Doppler, using $\psi_{D8'}$ with a gain of $\sqrt{2}$ for each higher scale.

2.8 An Example of the Haar Wavelet System

In this section, we can illustrate our mathematical discussion with a more complete example. In 1910, Haar [Haa10] showed that certain square wave functions could be translated and scaled to create a basis set that spans L^2. This is illustrated in Figure 2.11. Years later, it was seen that Haar's system is a particular wavelet system.

If we choose our scaling function to have compact support over $0 \leq t \leq 1$, then a solution to (2.13) is a scaling function that is a simple rectangle function

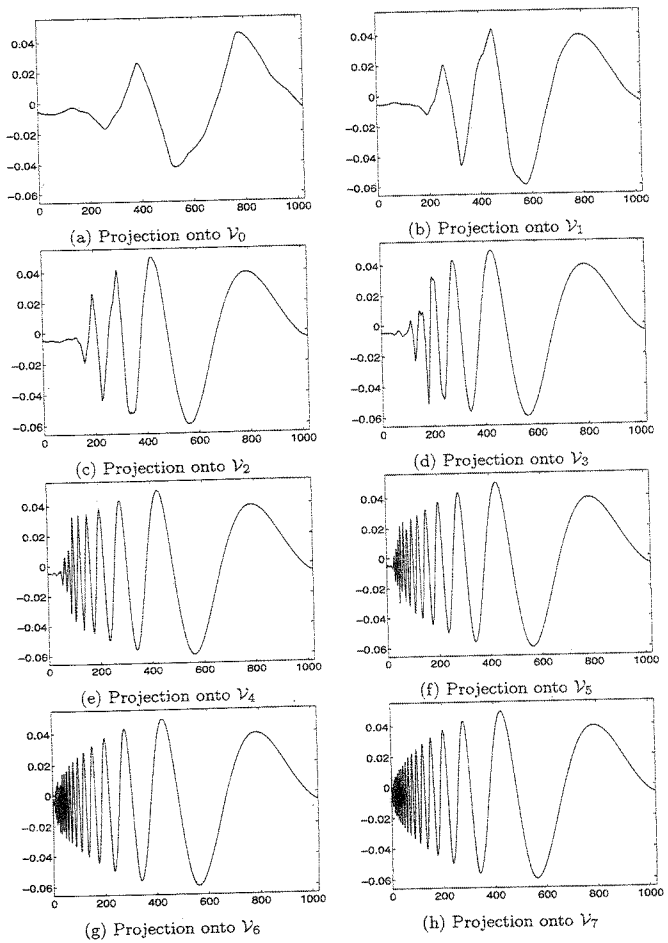
$$\varphi(t) = \begin{cases} 1 & \text{if } 0 < t < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.42}$$

with only two nonzero coefficients $h(0) = h(1) = 1/\sqrt{2}$ and (2.24) and (2.25) require the wavelet to be

$$\psi(t) = \begin{cases} 1 & \text{for } 0 < t < 0.5 \\ -1 & \text{for } 0.5 < t < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.43}$$

with only two nonzero coefficients $h_1(0) = 1/\sqrt{2}$ and $h_1(1) = -1/\sqrt{2}$.

$\mathcal{V}_0$ is the space spanned by $\varphi(t-k)$ which is the space of piecewise constant functions over integers, a rather limited space, but nontrivial. The next higher resolution space $\mathcal{V}_1$ is spanned by $\varphi(2t-k)$ which allows a somewhat more interesting class of signals which does include $\mathcal{V}_0$. As we consider higher values of scale j, the space $\mathcal{V}_j$ spanned by $\varphi(2^j t - k)$ becomes better able



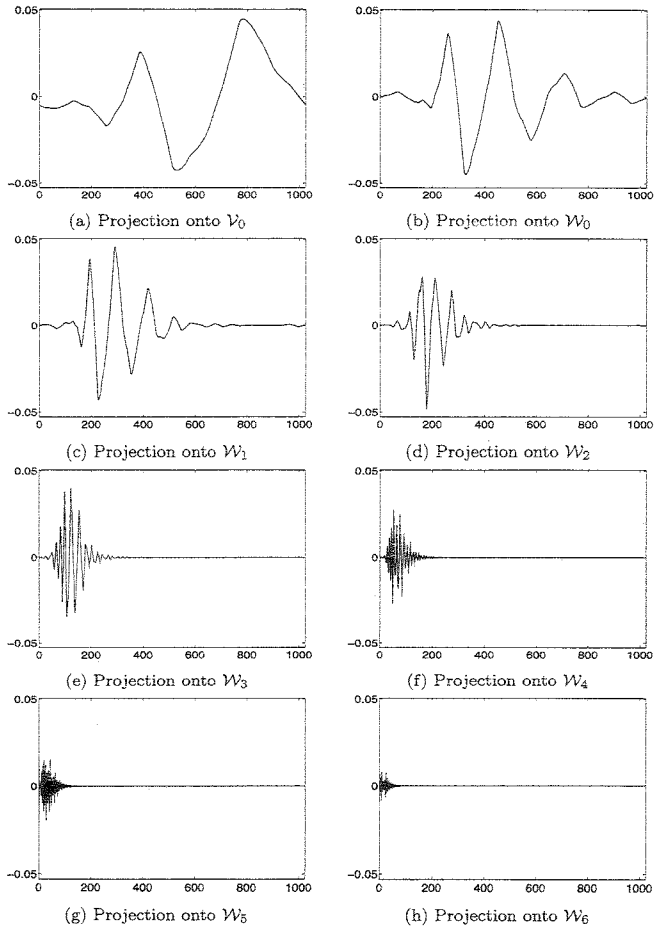
(a) Projection onto $\mathcal{V}_0$ (b) Projection onto $\mathcal{V}_1$

(c) Projection onto $\mathcal{V}_2$ (d) Projection onto $\mathcal{V}_3$

(e) Projection onto $\mathcal{V}_4$ (f) Projection onto $\mathcal{V}_5$

(g) Projection onto $\mathcal{V}_6$ (h) Projection onto $\mathcal{V}_7$

Figure 2.9. Projection of the Doppler Signal onto $\mathcal{V}$ Spaces using $\phi_{D8'}$



(a) Projection onto $\mathcal{V}_0$ (b) Projection onto $\mathcal{W}_0$

(c) Projection onto $\mathcal{W}_1$ (d) Projection onto $\mathcal{W}_2$

(e) Projection onto $\mathcal{W}_3$ (f) Projection onto $\mathcal{W}_4$

(g) Projection onto $\mathcal{W}_5$ (h) Projection onto $\mathcal{W}_6$

Figure 2.10. Projection of the Doppler Signal onto $\mathcal{W}$ Spaces using $\psi_{D8'}$

Haar showed that as $j \to \infty$, $\mathcal{V}_j \to L^2$. We have an approximation made up of step functions approaching any square integrable function.

The Haar functions are illustrated in Figure 2.11 where the first column contains the simple constant basis function that spans $\mathcal{V}_0$, the second column contains the unit pulse of width one half and the one translate necessary to span $\mathcal{V}_1$. The third column contains four translations of a pulse of width one fourth and the fourth contains eight translations of a pulse of width one eighth. This shows clearly how increasing the scale allows greater and greater detail to be realized. However, using only the scaling function does not allow the decomposition described in the introduction. For that we need the wavelet. Rather than use the scaling functions $\varphi(8t-k)$ in $\mathcal{V}_3$, we will use the orthogonal decomposition

$$\mathcal{V}_3 = \mathcal{V}_2 \oplus \mathcal{W}_2 \tag{2.44}$$

which is the same as

$$\overline{\underset{k}{\text{Span}}\{\varphi(8t-k)\}} = \overline{\underset{k}{\text{Span}}\{\varphi(4t-k)\}} \oplus \overline{\underset{k}{\text{Span}}\{\psi(4t-k)\}} \tag{2.45}$$

which means there are two sets of orthogonal basis functions that span $\mathcal{V}_3$, one in terms of $j=3$ scaling functions, and the other in terms of half as many coarser $j=2$ scaling functions plus the details contained in the $j=2$ wavelets. This is illustrated in Figure 2.12.

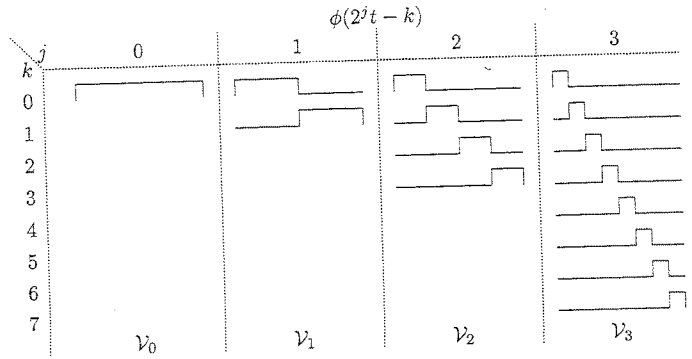


Figure 2.11. Haar Scaling Functions that Span $\mathcal{V}_j$

The $\mathcal{V}_2$ can be further decomposed into

$$\mathcal{V}_2 = \mathcal{V}_1 \oplus \mathcal{W}_1 \tag{2.46}$$

which is the same as

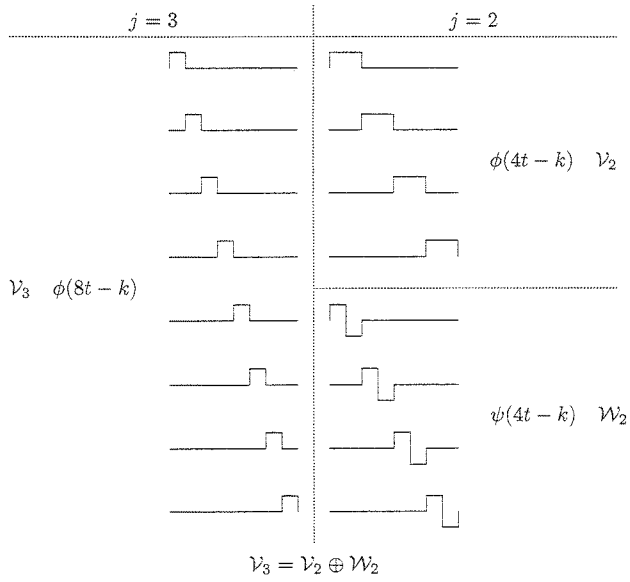
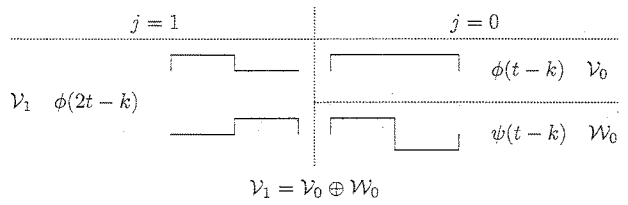


Figure 2.12. Haar Scaling Functions and Wavelets Decomposition of $\mathcal{V}_3$

and this is illustrated in Figure 2.14. This gives $\mathcal{V}_1$ also to be decomposed as

$$\mathcal{V}_1 = \mathcal{V}_0 \oplus \mathcal{W}_0 \tag{2.48}$$

which is shown in Figure 2.13. By continuing to decompose the space spanned by the scaling function until the space is one constant, the complete decomposition of $\mathcal{V}_3$ is obtained. This is symbolically shown in Figure 2.16.



Finally we look at an approximation to a smooth function constructed from the basis elements in $\mathcal{V}_3 = \mathcal{V}_0 \oplus \mathcal{W}_0 \oplus \mathcal{W}_1 \oplus \mathcal{W}_2$. Because the Haar functions form an orthogonal basis in each subspace, they can produce an optimal least squared error approximation to the smooth function. One can easily imagine the effects of adding a higher resolution "layer" of functions to $\mathcal{W}_3$ giving an approximation residing in $\mathcal{V}_4$. Notice that these functions satisfy all of the conditions that we have considered for scaling functions and wavelets. The basic wavelet is indeed an oscillating function which, in fact, has an average of zero and which will produce finer and finer detail as it is scaled and translated.

The multiresolution character of the scaling function and wavelet system is easily seen from Figure 2.12 where a signal residing in $\mathcal{V}_3$ can be expressed in terms of a sum of eight shifted scaling functions at scale $j = 3$ or a sum of four shifted scaling functions and four shifted wavelets at a scale of $j = 2$. In the second case, the sum of four scaling functions gives a low resolution approximation to the signal with the four wavelets giving the higher resolution "detail". The four shifted scaling functions could be further decomposed into coarser scaling functions and wavelets as illustrated in Figure 2.14 and still further decomposed as shown in Figure 2.13.

Figure 2.15 shows the Haar approximations of a test function in various resolutions. The signal is an example of a mixture of a pure sine wave which would have a perfectly localized Fourier domain representation and a two discontinuities which are completely localized in time domain. The component at the coarsest scale is simply the average of the signal. As we include more and more wavelet scales, the approximation becomes close to the original signal.

This chapter has skipped over some details in an attempt to communicate the general idea of the method. The conditions that can or must be satisfied and the resulting properties, together with examples, are discussed in the following chapters and/or in the references.

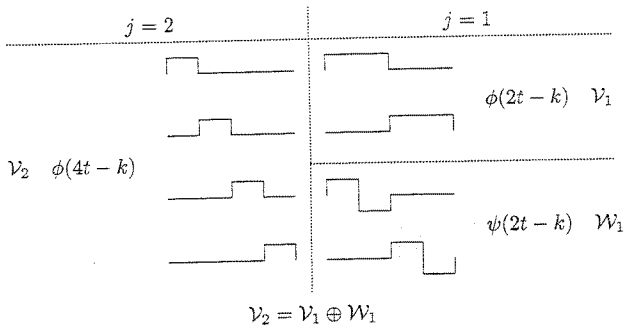


Figure 2.14. Haar Scaling Functions and Wavelets Decomposition of $\mathcal{V}_2$

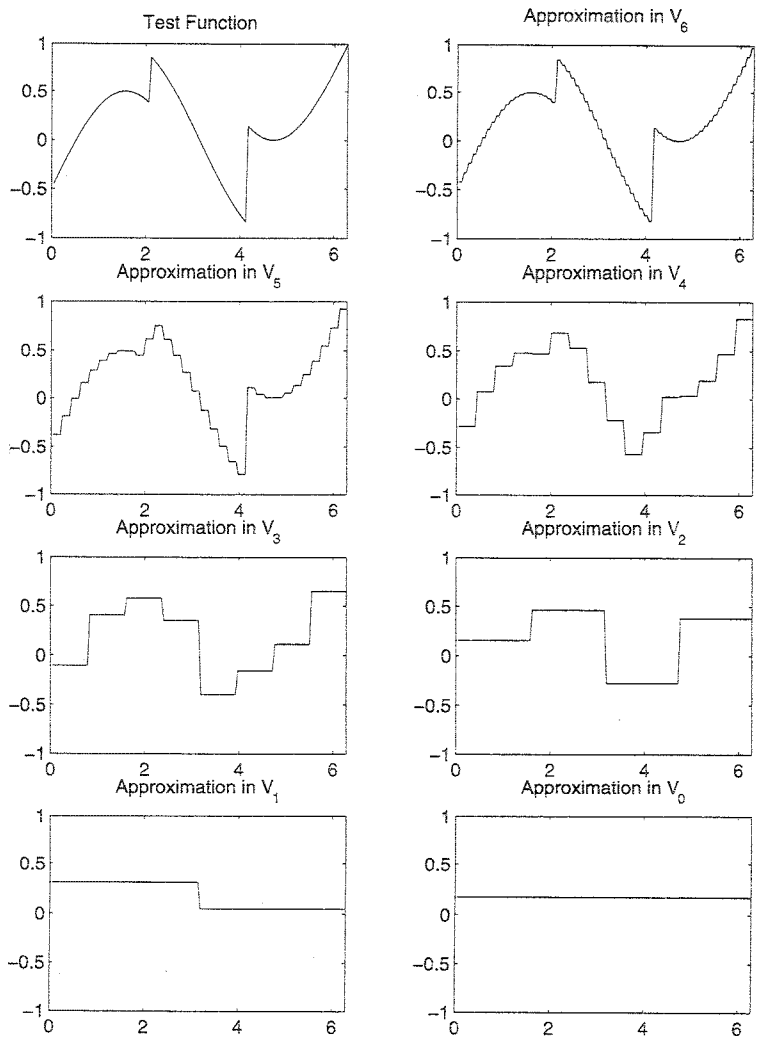


Figure 2.15. Haar Function Approximation in $\mathcal{V}_j$

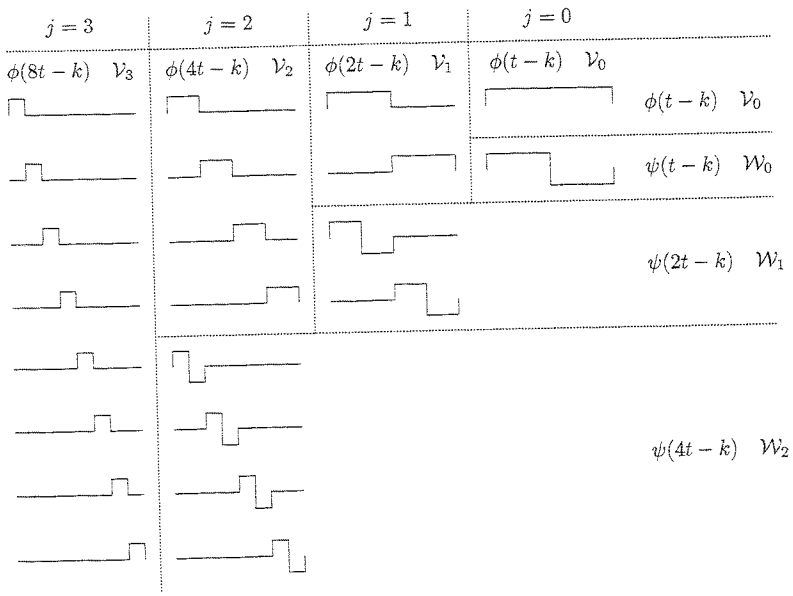


Figure 2.16. Haar Scaling Functions and Wavelets Decomposition of $\mathcal{V}_3$

Chapter 3

Filter Banks and the Discrete Wavelet Transform

In many applications, one never has to deal directly with the scaling functions or wavelets. Only the coefficients $h(n)$ and $h_1(n)$ in the defining equations (2.13) and (2.24) and $c(k)$ and $d_j(k)$ in the expansions (2.28), (2.29), and (2.30) need be considered, and they can be viewed as digital filters and digital signals respectively [GB92c, Vai92]. While it is possible to develop most of the results of wavelet theory using only filter banks, we feel that both the signal expansion point of view and the filter bank point of view are necessary for a real understanding of this new tool.

3.1 Analysis – From Fine Scale to Coarse Scale

In order to work directly with the wavelet transform coefficients, we will derive the relationship between the expansion coefficients at a lower scale level in terms of those at a higher scale. Starting with the basic recursion equation from (2.13)

$$\varphi(t) = \sum_n h(n)\, \sqrt{2}\, \varphi(2t-n) \tag{3.1}$$

and assuming a unique solution exists, we scale and translate the time variable to give

$$\varphi(2^j t-k) = \sum_n h(n)\, \sqrt{2}\, \varphi(2(2^j t-k)-n) = \sum_n h(n)\, \sqrt{2}\, \varphi(2^{j+1}t-2k-n) \tag{3.2}$$

which, after changing variables $m = 2k+n$, becomes

$$\varphi(2^j t-k) = \sum_m h(m-2k)\, \sqrt{2}\, \varphi(2^{j+1}t-m). \tag{3.3}$$

If we denote $\mathcal{V}_j$ as

$$\mathcal{V}_j = \operatorname*{Span}_k\{2^{j/2}\, \varphi(2^j t-k)\} \tag{3.4}$$

then

$$f(t) \in \mathcal{V}_{j+1} \quad \Rightarrow \quad f(t) = \sum_k c_{j+1}(k)\, 2^{(j+1)/2}\, \varphi(2^{j+1}t-k) \tag{3.5}$$

is expressible at a scale of $j+1$ with scaling functions only and no wavelets. At one scale lower resolution, wavelets are necessary for the "detail" not available at a scale of j. We have

$$f(t) = \sum_k c_j(k)\, 2^{j/2}\, \varphi(2^j t - k) + \sum_k d_j(k)\, 2^{j/2}\, \psi(2^j t - k) \tag{3.6}$$

where the $2^{j/2}$ terms maintain the unity norm of the basis functions at various scales. If $\varphi_{j,k}(t)$ and $\psi_{j,k}(t)$ are orthonormal or a tight frame, the j level scaling coefficients are found by taking the inner product

$$c_j(k) = \langle f(t),\, \varphi_{j,k}(t) \rangle = \int f(t)\, 2^{j/2}\, \varphi(2^j t - k)\, dt \tag{3.7}$$

which, by using (3.3) and interchanging the sum and integral, can be written as

$$c_j(k) = \sum_m h(m-2k) \int f(t)\, 2^{(j+1)/2}\, \varphi(2^{j+1} t - m)\, dt \tag{3.8}$$

but the integral is the inner product with the scaling function at a scale of $j+1$ giving

$$\boxed{c_j(k) = \sum_m h(m-2k)\, c_{j+1}(m).} \tag{3.9}$$

The corresponding relationship for the wavelet coefficients is

$$\boxed{d_j(k) = \sum_m h_1(m-2k)\, c_{j+1}(m).} \tag{3.10}$$

Filtering and Down-Sampling or Decimating

In the discipline of digital signal processing, the "filtering" of a sequence of numbers (the input signal) is achieved by convolving the sequence with another set of numbers called the filter coefficients, taps, weights, or impulse response. This makes intuitive sense if you think of a moving average with the coefficients being the weights. For an input sequence $x(n)$ and filter coefficients $h(n)$, the output sequence $y(n)$ is given by

$$y(n) = \sum_{k=0}^{N-1} h(k)\, x(n-k) \tag{3.11}$$

There is a large literature on digital filters and how to design them [PB87, OS89]. If the number of filter coefficients N is finite, the filter is called a Finite Impulse Response (FIR) filter. If the number is infinite, it is called an Infinite Impulse (IIR) filter. The design problem is the choice of the $h(n)$ to obtain some desired effect, often to remove noise or separate signals [OS89, PB87].

In multirate digital filters, there is an assumed relation between the integer index n in the signal $x(n)$ and time. Often the sequence of numbers are simply evenly spaced samples of a function of time. Two basic operations in multirate filters are the down-sampler and the up-sampler. The down-sampler (sometimes simply called a sampler or a decimator) takes a signal

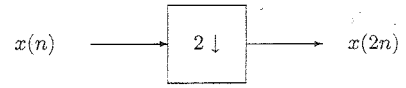


Figure 3.1. The Down Sampler or Decimator

$x(n)$ as an input and produces an output of $y(n) = x(2n)$. This is symbolically shown in Figure 3.1. In some cases, the down-sampling is by a factor other than two and in some cases, the output is the odd index terms $y(n) = x(2n+1)$, but this will be explicitly stated if it is important.

In down-sampling, there is clearly the possibility of losing information since half of the data is discarded. The effect in the frequency domain (Fourier transform) is called aliasing which states that the result of this loss of information is a mixing up of frequency components [PB87, OS89]. Only if the original signal is band-limited (half of the Fourier coefficients are zero) is there no loss of information caused by down-sampling.

We talk about digital filtering and down-sampling because that is exactly what (3.9) and (3.10) do. These equations show that the scaling and wavelet coefficients at different levels of scale can be obtained by convolving the expansion coefficients at scale j by the time-reversed recursion coefficients $h(-n)$ and $h_1(-n)$ then down-sampling or decimating (taking every other term, the even terms) to give the expansion coefficients at the next level of $j-1$. In other words, the scale-j coefficients are "filtered" by two FIR digital filters with coefficients $h(-n)$ and $h_1(-n)$ after which down-sampling gives the next coarser scaling and wavelet coefficients. These structures implement Mallat's algorithm [Mal89b, Mal89c] and have been developed in the engineering literature on filter banks, quadrature mirror filters (QMF), conjugate filters, and perfect reconstruction filter banks [SB86a, SB87, Vet86, VG89, Vet87, Vai87a, Vai92] and are expanded somewhat in Chapter 8 of this book. Mallat, Daubechies, and others showed the relation of wavelet coefficient calculation and filter banks. The impiementation of equations (3.9) and (3.10) is illustrated in Figure 3.2 where the down-pointing arrows denote a decimation or down-sampling by two and the other boxes denote FIR filtering or a convolution by $h(-n)$ or $h_1(-n)$. To ease notation, we use both $h(n)$ and $h_0(n)$ to denote the scaling function coefficients for the dilation equation (2.13).

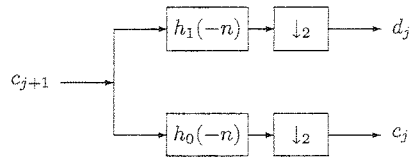


Figure 3.2. Two-Band Analysis Bank

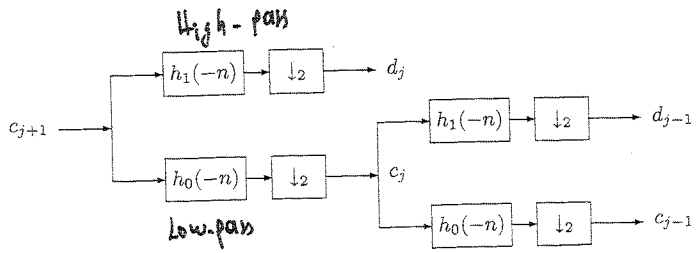


Figure 3.3. Two-Stage Two-Band Analysis Tree

As we will see in Chapter 5, the FIR filter implemented by $h(-n)$ is a lowpass filter, and the one implemented by $h_1(-n)$ is a highpass filter. Note the average number of data points out of this system is the same as the number in. The number is doubled by having two filters; then it is halved by the decimation back to the original number. This means there is the possibility that no information has been lost and it will be possible to completely recover the original signal. As we shall see, that is indeed the case. The aliasing occurring in the upper bank can be "undone" or cancelled by using the signal from the lower bank. This is the idea behind perfect reconstruction in filter bank theory [Vai92, Fli94].

This splitting, filtering, and decimation can be repeated on the scaling coefficients to give the two-scale structure in Figure 3.3. Repeating this on the scaling coefficients is called *iterating the filter bank*. Iterating the filter bank again gives us the three-scale structure in Figure 3.4.

The frequency response of a digital filter is the discrete-time Fourier transform of its impulse response (coefficients) $h(n)$. That is given by

$$H(\omega) = \sum_{n=-\infty}^{\infty} h(n)\, e^{i\omega n}. \tag{3.12}$$

The magnitude of this complex-valued function gives the ratio of the output to the input of the filter for a sampled sinusoid at a frequency of ω in radians per seconds. The angle of $H(\omega)$ is the phase shift between the output and input.

The first stage of two banks divides the spectrum of $c_{j+1}(k)$ into a lowpass and highpass band, resulting in the scaling coefficients and wavelet coefficients at lower scale $c_j(k)$ and $d_j(k)$. The second stage then divides that lowpass band into another lower lowpass band and a bandpass band. The first stage divides the spectrum into two equal parts. The second stage divides the lower half into quarters and so on. This results in a logarithmic set of bandwidths as illustrated in Figure 3.5. These are called "constant-Q" filters in filter bank language because the ratio of the band width to the center frequency of the band is constant. It is also interesting to note that a musical scale defines octaves in a similar way and that the ear responds to frequencies in a similar logarithmic fashion.

For any practical signal that is bandlimited, there will be an upper scale $j = J$, above which the wavelet coefficients, $d_j(k)$, are negligibly small [GOB94]. By starting with a high resolution description of a signal in terms of the scaling coefficients c_J, the analysis tree calculates the DWT

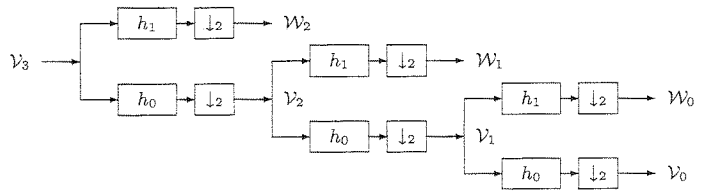


Figure 3.4. Three-Stage Two-Band Analysis Tree

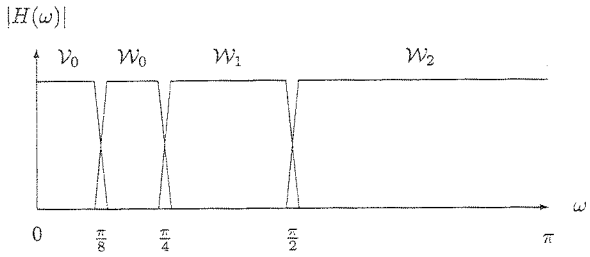


Figure 3.5. Frequency Bands for the Analysis Tree

down to as low a resolution, $j = j_0$, as desired by having $J - j_0$ stages. So, for $f(t) \in \mathcal{V}_J$, using (2.8) we have

$$f(t) = \sum_k c_J(k)\, \varphi_{J,k}(t) \tag{3.13}$$

$$= \sum_k c_{J-1}(k)\, \varphi_{J-1,k}(t) + \sum_k d_{J-1}(k)\, \psi_{J-1.k}(t) \tag{3.14}$$

$$f(t) = \sum_k c_{J-2}(k)\, \varphi_{J-2,k}(t) + \sum_k \sum_{j=J-2}^{J-1} d_j(k)\, \psi_{j,k}(t) \tag{3.15}$$

$$f(t) = \sum_k c_{j_0}(k)\, \varphi_{j_0,k}(t) + \sum_k \sum_{j=j_0}^{J-1} d_j(k)\, \psi_{j,k}(t) \tag{3.16}$$

which is a finite scale version of (2.33). We will discuss the choice of j_0 and J further in Chapter 9.

3.2 Synthesis – From Coarse Scale to Fine Scale

As one would expect, a reconstruction of the original fine scale coefficients of the signal can be made from a combination of the scaling function and wavelet coefficients at a coarse resolution. This is derived by considering a signal in the $j+1$ scaling function space $f(t) \in \mathcal{V}_{j+1}$. This function can be written in terms of the scaling function as

$$f(t) = \sum_k c_{j+1}(k)\, 2^{(j+1)/2}\, \varphi(2^{j+1}t - k) \tag{3.17}$$

or in terms of the next scale (which also requires wavelets) as

$$f(t) = \sum_k c_j(k)\, 2^{j/2}\, \varphi(2^j t - k) + \sum_k d_j(k)\, 2^{j/2}\, \psi(2^j t - k). \tag{3.18}$$

Substituting (3.1) and (2.24) into (3.18) gives

$$f(t) = \sum_k c_j(k) \sum_n h(n)\, 2^{(j+1)/2}\, \varphi(2^{j+1}t - 2k - n) + \sum_k d_j(k) \sum_n h_1(n)\, 2^{(j+1)/2}\, \varphi(2^{j+1}t - 2k - n). \tag{3.19}$$

Because all of these functions are orthonormal, multiplying (3.17) and (3.19) by $\varphi(2^{j+1}t - k')$ and integrating evaluates the coefficient as

$$\boxed{c_{j+1}(k) = \sum_m c_j(m)\, h(k - 2m) + \sum_m d_j(m)\, h_1(k - 2m).} \tag{3.20}$$

Filtering and Up-Sampling or Stretching

For synthesis in the filter bank we have a sequence of first up-sampling or stretching, then filtering. This means that the input to the filter has zeros inserted between each of the original terms. In other words,

$$y(2n) = x(n) \quad \text{and} \quad y(2n+1) = 0 \tag{3.21}$$

where the input signal is stretched to twice its original length and zeros are inserted. Clearly this up-sampling or stretching could be done with factors other than two, and the two equation above could have the $x(n)$ and 0 reversed. It is also clear that up-sampling does not lose any information. If you first up-sample then down-sample, you are back where you started. However, if you first down-sample then up-sample, you are not generally back where you started.

Our reason for discussing filtering and up-sampling here is that is exactly what the synthesis operation (3.20) does. This equation is evaluated by up-sampling the j scale coefficient sequence $c_j(k)$, which means double its length by inserting zeros between each term, then convolving it with the scaling coefficients $h(n)$. The same is done to the j level wavelet coefficient sequence and the results are added to give the $j+1$ level scaling function coefficients. This structure is illustrated in Figure 3.6 where $g_0(n) = h(n)$ and $g_1(n) = h_1(n)$. This combining process can be continued to any level by combining the appropriate scale wavelet coefficients. The resulting two-scale tree is shown in Figure 3.7.

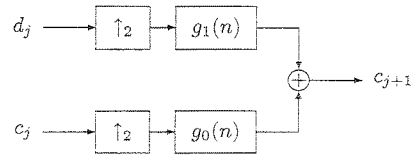


Figure 3.6. Two-Band Synthesis Bank

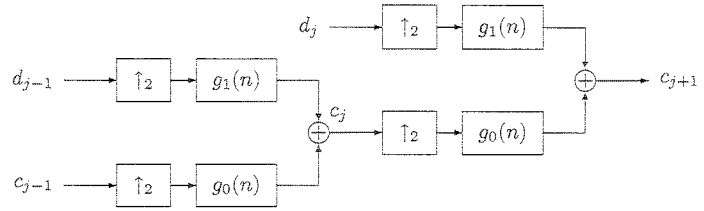


Figure 3.7. Two-Stage Two-Band Synthesis Tree

3.3 Input Coefficients

One might wonder how the input set of scaling coefficients c_{j+1} are obtained from the signal to use in the systems of Figures 3.2 and 3.3. For high enough scale, the scaling functions act as "delta functions" with the inner product to calculate the high scale coefficients as simply a sampling of $f(t)$ [GOB92, OGB92]. If the samples of $f(t)$ are above the Nyquist rate, they are good approximations to the scaling coefficients at that scale, meaning no wavelet coefficients are necessary at that scale. This approximation is particularly good if moments of the scaling function are zero or small. These ideas are further explained in Section 6.8 and Chapter 9.

An alternative approach is to "prefilter" the signal samples to make them a better approximation to the expansion coefficients. This is discussed in [Str86].

This set of analysis and synthesis operations is known as Mallat's algorithm [Mal89b, Mal89c]. The analysis filter bank efficiently calculates the DWT using banks of digital filters and down-samplers, and the synthesis filter bank calculates the inverse DWT to reconstruct the signal from the transform. Although presented here as a method of calculating the DWT, the filter bank description also gives insight into the transform itself and suggests modifications and generalizations that would be difficult to see directly from the wavelet expansion point of view. Filter banks will be used more extensively in the remainder of this book. A more general development of filter banks is presented in Section 7.2.

Although a pure wavelet expansion is possible as indicated in (1.7) and (2.22), properties of the wavelet are best developed and understood through the scaling function. This is certainly true if the scaling function has compact support because then the wavelet is composed of a finite sum of scaling functions given in (2.24).

In a practical situation where the wavelet expansion or transform is being used as a computational tool in signal processing or numerical analysis, the expansion can be made finite. If the basis functions have finite support, only a finite number of additions over k are necessary. If the scaling function is included as indicated in (2.28) or (3.6), the lower limit on the summation over j is finite. If the signal is essentially bandlimited, there is a scale above which there is little or no energy and the upper limit can be made finite. That is described in Chapter 9.

3.4 Lattices and Lifting

An alternative to using the basic two-band tree-structured filter bank is a lattice-structured filter bank. Because of the relationship between the scaling filter $h(n)$ and the wavelet filter $h_1(h)$ given in (2.25), some of the calculation can be done together with a significant savings in arithmetic. This is developed in Chapter 9 [Vai92].

Still another approach to the calculation of discrete wavelet transforms and to the calculations of the scaling functions and wavelets themselves is called "lifting." Although it is related to several other schemes [Mar92, Mar93, DM93, KS92], this idea was first explained by Wim Sweldens as a time-domain construction based on interpolation [Swe95]. Lifting does not use Fourier methods and can be applied to more general problems (e.g., nonuniform sampling) than the approach in this chapter. It was first applied to the biorthogonal system [Swe96a] and then extended to orthogonal systems [DS96a]. The application of lifting to biorthogonal is introduced in Section 7.4 later in this book. Implementations based on lifting also achieve the same improvement in arithmetic efficiency as the lattice structure do.

3.5 Different Points of View

Multiresolution versus Time-Frequency Analysis

The development of wavelet decomposition and the DWT has thus far been in terms of multiresolution where the higher scale wavelet components are considered the "detail" on a lower scale signal or image. This is indeed a powerful point of view and an accurate model for many signals and images, but there are other cases where the components of a composite signal at different scales and/or time are independent or, at least, not details of each other. If you think of a musical score as a wavelet decomposition, the higher frequency notes are not details on a lower frequency note; they are independent notes. This second point of view is more one of the time-frequency or time-scale analysis methods [Coh89, Coh95, HB92, Boa92, LP89] and may be better developed with wavelet packets (see Section 7.3), M-band wavelets (see Section 7.2), or a redundant representation (see Section 7.6), but would still be implemented by some sort of filter bank.

Periodic versus Nonperiodic Discrete Wavelet Transforms

Unlike the Fourier series, the DWT can be formulated as a periodic or a nonperiodic transform. Up until now, we have considered a nonperiodic series expansion (2.33) over $-\infty < t < \infty$ with

the calculations made by the filter banks being an on-going string of coefficients at each of th scales. If the input to the filter bank has a certain rate, the output at the next lower scale wi be two sequences, one of scaling function coefficients $c_{j-1,k-1}$ and one of wavelet coefficient $d_{j-1,k-1}$, each, after down-sampling, being at half the rate of the input. At the next lower scale the same process is done on the scaling coefficients to give a total output of three strings, on at half rate and two at quarter rate. In other words, the calculation of the wavelet transfor coefficients is a multirate filter bank producing sequences of coefficients at different rates but wit the average number at any stage being the same. This approach can be applied to any signa finite or infinite in length, periodic or nonperiodic. Note that while the average output rate i the same as the average input rate, the number of output coefficients is greater than the numbe of input coefficients because the length of the output of convolution is greater than the length c the input.

An alternative formulation that can be applied to finite duration signals or periodic signal (much as the Fourier series) is to make all of the filter bank filters cyclic or periodic convolutio which is defined by

$$y(n) = \sum_{\ell=0}^{N-1} h(\ell)\, x(n-\ell), \tag{3.22}$$

for $n, \ell = 0, 1, \cdots, N-1$ and all indices and arguments are evaluated modulo N. For a length N input at scale $j = J$, we have after one stage two length $N/2$ sequences, after two stages, on length $N/2$ and two length $N/4$ sequences, and so on. If $N = 2^J$, this can be repeated J time with the last stage being length one; one scaling function coefficient and one wavelet coefficient An example of how the periodic DWT of a length 8 can be seen Figure 3.8.

$c_j(k)$	$d_j(k)$	$d_{j+1}(k)$	$d_{j+1}(k+1)$	$d_{j+2}(k)$	$d_{j+2}(k+1)$	$d_{j+2}(k+2)$	$d_{j+2}(k+3)$

Figure 3.8. The length-8 DWT vector

The details of this periodic approach are developed in Chapter 9 showing the aliasing that takes place in this system because of the cyclic convolution (3.22). This formulation is particularly clean because there are the same number of terms in the transform as in the signal. It can be represented by a square matrix with a simple inverse that has interesting structure. It can be efficiently calculated by an FFT although that is not needed for most applications.

For most of the theoretical developments or for conceptual purposes, there is little difference in these two formulations. However, for actual calculations and in applications, you should make sure you know which one you want or which one your software package calculates. As for the Fourier case, you can use the periodic form to calculate the nonperiodic transform by padding the signal with zeros but that wastes some of the efficiency that the periodic formulation was set up to provide.

The Discrete Wavelet Transform versus the Discrete-Time Wavelet Transform

Two more points of view concern looking at the signal processing methods in this book as based on an expansion of a signal or on multirate digital filtering. One can look at Mallat's algorithm either as a way of calculating expansion coefficients at various scales or as a filter bank for processing

discrete-time signals. The first is analogous to use of the Fourier series (FS) where a continuous function is transformed into a discrete sequence of coefficients. The second is analogous to the discrete Fourier transform (DFT) where a discrete function is transformed into a discrete function. Indeed, the DFT (through the FFT) is often used to calculate the Fourier series coefficients, but care must be taken to avoid or minimize aliasing. The difference in these views comes partly from the background of the various researchers (i.e., whether they are "wavelet people" or "filter bank people"). However, there are subtle differences between using the series expansion of the signal (using the discrete wavelet transform (DWT)) and using a multirate digital filter bank on samples of the signal (using the discrete-time wavelet transform (DTWT)). Generally, using both views gives more insight into a problem than either achieves alone. The series expansion is the main approach of this book but filter banks and the DTWT are also developed in Chapters 7.8 and 8.

Numerical Complexity of the Discrete Wavelet Transform

Analysis of the number of mathematical operations (floating-point multiplications and additions) shows that calculating the DTWT of a length-N sequence of numbers using Mallat's algorithm with filter banks requires $O(N)$ operations. In other words, the number of operations is linear with the length of the signal. What is more, the constant of linearity is relatively small. This is in contrast to the FFT algorithm for calculating the DFT where the complexity is $O(N \log(N))$ or calculating a DFT directly requires $O(N^2)$ operations. It is often said that the FFT algorithm is based on a "divide and conquer" scheme, but that is misleading. The process is better described as a "organize and share" scheme. The efficiency (in fact, optimal efficiency) is based on organizing the calculations so that redundant operations can be shared. The cascaded filtering (convolution) and down-sampling of Mallat's algorithm do the same thing.

One should not make too much of this difference between the complexity of the FFT and DTWT. It comes from the DTWT having a logarithmic division of frequency bands and the FFT having a uniform division. This logarithmic scale is appropriate for many signals but if a uniform division is used for the wavelet system such as is done for wavelet packets (see Section 7.3) or the redundant DWT (see Chapter 7.6), the complexity of the wavelet system becomes $O(N \log(N))$.

If you are interested in more details of the discrete wavelet transform and the discrete-time wavelet transform, relations between them, methods of calculating them, further properties of them, or examples, see Section 7.8 and Chapter9.

Chapter 4

Bases, Orthogonal Bases, Biorthogonal Bases, Frames, Tight Frames, and Unconditional Bases

Most people with technical backgrounds are familiar with the ideas of expansion vectors or basis vectors and of orthogonality; however, the related concepts of biorthogonality or of frames and tight frames are less familiar but also important. In the study of wavelet systems, we find that frames and tight frames are needed and should be understood, at least at a superficial level. One can find details in [You80, Dau92, Dau90, HW89]. Another perhaps unfamiliar concept is that of an unconditional basis used by Donoho, Daubechies, and others [Don93b, Mey90, Dau92] to explain why wavelets are good for signal compression, detection, and denoising [GOL*94b, GOL*94a]. In this chapter, we will very briefly define and discuss these ideas. At this point, you may want to skip these sections and perhaps refer to them later when they are specifically needed.

4.1 Bases, Orthogonal Bases, and Biorthogonal Bases

A set of vectors or functions $f_k(t)$ *spans* a vector space $\mathcal{F}$ (or $\mathcal{F}$ is the *Span* of the set) if any element of that space can be expressed as a linear combination of members of that set, meaning: Given the finite or infinite set of functions $f_k(t)$, we define $\text{Span}_k\{f_k\} = \mathcal{F}$ as the vector space with all elements of the space of the form

$$g(t) = \sum_k a_k \, f_k(t) \tag{4.1}$$

with $k \in \mathbf{Z}$ and $t, a \in \mathbf{R}$. An *inner product* is usually defined for this space and is denoted $\langle f(t), g(t) \rangle$. A *norm* is defined and is denoted by $\|f\| = \sqrt{\langle f, f \rangle}$.

We say that the set $f_k(t)$ is a *basis set* or a *basis* for a given space $\mathcal{F}$ if the set of $\{a_k\}$ in (4.1) are unique for any particular $g(t) \in \mathcal{F}$. The set is called an *orthogonal basis* if $\langle f_k(t), f_\ell(t) \rangle = 0$ for all $k \neq \ell$. If we are in three dimensional Euclidean space, orthogonal basis vectors are coordinate vectors that are at right (90°) angles to each other. We say the set is an *orthonormal basis* if $\langle f_k(t), f_\ell(t) \rangle = \delta(k - \ell)$ i.e. if, in addition to being orthogonal, the basis vectors are normalized to unity norm: $\|f_k(t)\| = 1$ for all k.