Chapter 1

Haar Wavelets

The purpose of computing is insight, not numbers.

Richard W. Hamming

The purpose of computing is insight, not pictures.

Lloyd N. Trefethen[1]

A Haar wavelet is the simplest type of wavelet. In discrete form, Haar wavelets are related to a mathematical operation called the *Haar transform.* The Haar transform serves as a prototype for all other wavelet transforms. Studying the Haar transform in detail will provide a good foundation for understanding the more sophisticated wavelet transforms which we shall describe in the next chapter. In this chapter we shall describe how the Haar transform can be used for compressing audio signals and for removing noise. Our discussion of these applications will set the stage for the more powerful wavelet transforms to come and their applications to these same problems. One distinctive feature that the Haar transform enjoys is that it lends itself easily to simple hand calculations. We shall illustrate many concepts by both simple hand calculations and more involved computer computations.

1.1 The Haar transform

In this section we shall introduce the basic notions connected with the Haar transform, which we shall examine in more detail in later sections.

[1] Hamming's quote is from [HAM]. Trefethen's quote is from [TRE].

First, we need to define the type of signals that we shall be analyzing with the Haar transform.

Throughout this book we shall be working extensively with *discrete signals.* A discrete signal is a function of time with values occurring at discrete instants. Generally we shall express a discrete signal in the form $\mathbf{f} = (f_1, f_2, \ldots, f_N)$, where N is a positive even integer which we shall refer to as the *length* of $\mathbf{f}$. The *values* of $\mathbf{f}$ are the N real numbers $f_1, f_2, \ldots, f_N$. These values are typically measured values of an analog signal g, measured at the time values $t = t_1, t_2, \ldots, t_N$. That is, the values of $\mathbf{f}$ are

$$f_1 = g(t_1),\ f_2 = g(t_2),\ \ldots,\ f_N = g\left(t_N\right). \tag{1.1}$$

For simplicity, we shall assume that the increment of time that separates each pair of successive time values is always the same. We shall use the phrase *equally spaced sample values*, or just *sample values,* when the discrete signal has its values defined in this way. An important example of sample values is the set of data values stored in a computer audio file, such as a `.wav` file. Another example is the sound intensity values recorded on a compact disc. A non-audio example, where the analog signal g is not a sound signal, is a digitized electrocardiogram.

Like all wavelet transforms, the Haar transform decomposes a discrete signal into two subsignals of half its length. One subsignal is a running average or *trend;* the other subsignal is a running difference or *fluctuation.*

Let's begin by examining the trend subsignal. The *first trend* subsignal, $\mathbf{a}^1 = (a_1, a_2, \ldots, a_{N/2})$, for the signal $\mathbf{f}$ is computed by taking a running average in the following way. Its first value, a_1, is computed by taking the average of the first pair of values of $\mathbf{f}$: $(f_1 + f_2)/2$, and then multiplying it by $\sqrt{2}$. That is, $a_1 = (f_1 + f_2)/\sqrt{2}$. Similarly, its next value a_2 is computed by taking the average of the next pair of values of $\mathbf{f}$: $(f_3 + f_4)/2$, and then multiplying it by $\sqrt{2}$. That is, $a_2 = (f_3 + f_4)/\sqrt{2}$. Continuing in this way, all of the values of $\mathbf{a}^1$ are produced by taking averages of successive pairs of values of $\mathbf{f}$, and then multiplying these averages by $\sqrt{2}$. A precise formula for the values of $\mathbf{a}^1$ is

$$a_m = \frac{f_{2m-1} + f_{2m}}{\sqrt{2}}, \tag{1.2}$$

for $m = 1, 2, 3, \ldots, N/2$.

For example, suppose $\mathbf{f}$ is defined by eight values, say

$$\mathbf{f} = (4, 6, 10, 12, 8, 6, 5, 5);$$

then its first trend subsignal is $\mathbf{a}^1 = (5\sqrt{2}, 11\sqrt{2}, 7\sqrt{2}, 5\sqrt{2})$. This result can be obtained using Formula (1.2). Or it can be calculated as indicated in the following diagram:

$$\begin{tikzcd}
\mathbf{f}\colon & 4 \arrow[dr] & & 6 \arrow[dl] & 10 \arrow[dr] & & 12 \arrow[dl] & 8 \arrow[dr] & & 6 \arrow[dl] & 5 \arrow[dr] & & 5 \arrow[dl] \\
& & 5 \arrow[d] & & & 11 \arrow[d] & & & 7 \arrow[d] & & & 5 \arrow[d] & \\
\mathbf{a}^1\colon & & 5\sqrt{2} & & & 11\sqrt{2} & & & 7\sqrt{2} & & & 5\sqrt{2}\,. &
\end{tikzcd}$$

You might ask: Why perform the extra step of multiplying by $\sqrt{2}$? Why not just take averages? These questions will be answered in the next section, when we show that multiplication by $\sqrt{2}$ is needed in order to ensure that the Haar transform preserves the energy of a signal.

The other subsignal is called the *first fluctuation.* The first fluctuation of the signal $\mathbf{f}$, which is denoted by $\mathbf{d}^1 = (d_1, d_2, \ldots, d_{N/2})$, is computed by taking a running difference in the following way. Its first value, d_1, is calculated by taking half the difference of the first pair of values of $\mathbf{f}$: $(f_1 - f_2)/2$, and multiplying it by $\sqrt{2}$. That is, $d_1 = (f_1 - f_2)/\sqrt{2}$. Likewise, its next value d_2 is calculated by taking half the difference of the next pair of values of $\mathbf{f}$: $(f_3 - f_4)/2$, and multiplying it by $\sqrt{2}$. In other words, $d_2 = (f_3 - f_4)/\sqrt{2}$. Continuing in this way, all of the values of $\mathbf{d}^1$ are produced according to the following formula:

$$d_m = \frac{f_{2m-1} - f_{2m}}{\sqrt{2}}\,, \tag{1.3}$$

for $m = 1, 2, 3, \ldots, N/2$.

For example, for the signal $\mathbf{f} = (4, 6, 10, 12, 8, 6, 5, 5)$ considered above, its first fluctuation $\mathbf{d}^1$ is $(-\sqrt{2}, -\sqrt{2}, \sqrt{2}, 0)$. This result can be obtained using Formula (1.3), or it can be calculated as indicated in the following diagram:

$$\begin{tikzcd}
\mathbf{f}\colon & 4 \arrow[dr] & & 6 \arrow[dl] & 10 \arrow[dr] & & 12 \arrow[dl] & 8 \arrow[dr] & & 6 \arrow[dl] & 5 \arrow[dr] & & 5 \arrow[dl] \\
& & -1 \arrow[d] & & & -1 \arrow[d] & & & 1 \arrow[d] & & & 0 \arrow[d] & \\
\mathbf{d}^1\colon & & -\sqrt{2} & & & -\sqrt{2} & & & \sqrt{2} & & & 0. &
\end{tikzcd}$$

Haar transform, 1-level

The Haar transform is performed in several stages, or levels. The first level is the mapping $\mathbf{H}_1$ defined by

$$\mathbf{f} \overset{\mathbf{H}_1}{\longmapsto} (\mathbf{a}^1 \,|\, \mathbf{d}^1) \tag{1.4}$$

from a discrete signal $\mathbf{f}$ to its first trend $\mathbf{a}^1$ and first fluctuation $\mathbf{d}^1$. For example, we showed above that

$$(4, 6, 10, 12, 8, 6, 5, 5) \overset{\mathbf{H}_1}{\longmapsto} (5\sqrt{2}, 11\sqrt{2}, 7\sqrt{2}, 5\sqrt{2} \,|\, -\sqrt{2}, -\sqrt{2}, \sqrt{2}, 0). \tag{1.5}$$

The mapping $\mathbf{H}_1$ in (1.4) has an inverse. Its inverse maps the transform signal $(\mathbf{a}^1 \,|\, \mathbf{d}^1)$ back to the signal $\mathbf{f}$, via the following formula:

$$\mathbf{f} = \left(\frac{a_1 + d_1}{\sqrt{2}}, \frac{a_1 - d_1}{\sqrt{2}}, \dots, \frac{a_{N/2} + d_{N/2}}{\sqrt{2}}, \frac{a_{N/2} - d_{N/2}}{\sqrt{2}} \right). \tag{1.6}$$

In other words, $f_1 = (a_1 + d_1)/\sqrt{2}$, $f_2 = (a_1 - d_1)/\sqrt{2}$, $f_3 = (a_2 + d_2)/\sqrt{2}$, $f_4 = (a_2 - d_2)/\sqrt{2}$, and so on. For instance, the following diagram shows how to invert the transformation in (1.5):

$$\begin{tikzcd}
\mathbf{a}^1: & & 5\sqrt{2} & & & 11\sqrt{2} & & & 7\sqrt{2} & & & 5\sqrt{2} & \\
\mathbf{d}^1: & & -\sqrt{2} \arrow[dl] \arrow[dr] & & & -\sqrt{2} \arrow[dl] \arrow[dr] & & & \sqrt{2} \arrow[dl] \arrow[dr] & & & 0 \arrow[dl] \arrow[dr] & \\
\mathbf{f}: & 4 & & 6 & 10 & & 12 & 8 & & 6 & 5 & & 5.
\end{tikzcd}$$

Let's now consider what advantages accrue from performing the Haar transformation. These advantages will be described in more detail later in this chapter, but some basic notions can be introduced now. All of these advantages stem from the following cardinal feature of the Haar transform (a feature that will be even more prominent for the Daubechies transforms described in the next chapter):

Small Fluctuations Feature. *The magnitudes of the values of the fluctuation subsignal are often significantly smaller than the magnitudes of the values of the original signal.*

For instance, for the signal $\mathbf{f} = (4, 6, 10, 12, 8, 6, 5, 5)$ considered above, its eight values have an average magnitude of 7. On the other hand, for its first fluctuation $\mathbf{d}^1 = (-\sqrt{2}, -\sqrt{2}, \sqrt{2}, 0)$, the average of its four magnitudes is $0.75\sqrt{2}$. In this case, the magnitudes of the fluctuation's values are an average of 6.6 times smaller than the magnitudes of the original signal's values. For a second example, consider the signal shown in Figure 1.1(a). This signal was generated from 1024 sample values of the function

$$g(x) = 20x^2(1-x)^4 \cos 12\pi x$$

over the interval $[0, 1)$. In Figure 1.1(b) we show a graph of the 1-level Haar transform of this signal. The trend subsignal is graphed on the left half, over the interval $[0, 0.5)$, and the fluctuation subsignal is graphed on the right half, over the interval $[0.5, 1)$. It is clear that a large percentage of the fluctuation's values are close to 0 in magnitude, another instance of the Small Fluctuations Feature. Notice also that the trend subsignal looks like the original signal, although shrunk by half in length and expanded by a factor of $\sqrt{2}$ vertically.

The reason that the Small Fluctuations Feature is generally true is that typically we are dealing with signals whose values are samples of a continuous analog signal g with a very short time increment between the samples.

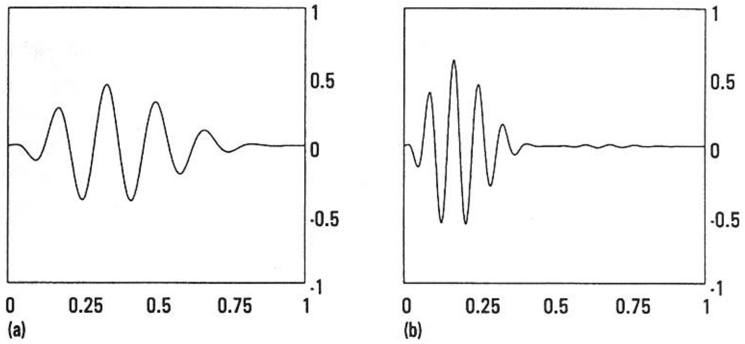

FIGURE 1.1
(a) Signal, (b) Haar transform, 1-level.

In other words, the equations in (1.1) hold with a small value of the time increment $h = t_{k+1} - t_k$ for each $k = 1, 2, \ldots, N-1$. If the time increment is small enough, then successive values $f_{2m-1} = g\left(t_{2m-1}\right)$ and $f_{2m} = g\left(t_{2m}\right)$ of the signal $\mathbf{f}$ will be close to each other due to the continuity of g. Consequently, the fluctuation values for the Haar transform satisfy

$$d_m = \frac{g\left(t_{2m-1}\right) - g\left(t_{2m}\right)}{\sqrt{2}} \approx 0.$$

This explains why the Small Fluctuations Feature is generally true for the Haar transform. A similar analysis shows why the trend subsignal has a graph that is similar in appearance to the first trend. If g is continuous and the time increment is very small, then $g(t_{2m-1})$ and $g(t_{2m})$ will be close to each other. Expressing this fact as an approximation, $g(t_{2m-1}) \approx g(t_{2m})$, we obtain the following approximation for each value a_m of the trend subsignal

$$a_m \approx \sqrt{2}\, g(t_{2m}).$$

This equation shows that $\mathbf{a}^1$ is approximately the same as sample values of $\sqrt{2}\, g(x)$ for $x = t_2, t_4, \ldots, t_N$. In other words, it shows that the graph of the first trend subsignal is similar in appearance to the graph of g, as we pointed out above in regard to the signal in Figure 1.1(a). We shall examine these points in more detail in the next chapter when we discuss other wavelet transforms.

One of the reasons that the Small Fluctuations Feature is important is that it has applications to *signal compression.* By compressing a signal we mean transmitting its values, or approximations of its values, by using a smaller number of bits. For example, if we were only to transmit the trend subsignal for the signal shown in Figure 1.1(a) and then perform Haar transform inversion (treating the fluctuation's values as all zeros), then we would obtain an approximation of the original signal. Since the length of the trend subsignal is half the length of the original signal, this would achieve 50% compression. We shall discuss compression in more detail in Section 1.5.

Once we have performed a 1-level Haar transform, then it is easy to repeat the process and perform multiple-level Haar transforms. We shall discuss this in the next section.

1.2 Conservation and compaction of energy

In the previous section we defined the 1-level Haar transform. In this section we shall discuss its two most important properties: (1) It conserves the energies of signals; (2) It performs a compaction of the energy of signals. We shall also complete our definition of the Haar transform by showing how to extend its definition to multiple levels.

Conservation of energy

An important property of the Haar transform is that it *conserves the energies of signals.* By the *energy* of a signal $\mathbf{f}$ we mean the sum of the squares of its values. That is, the energy $\mathcal{E}_{\mathbf{f}}$ of a signal $\mathbf{f}$ is defined by

$$\mathcal{E}_{\mathbf{f}} = f_1^2 + f_2^2 + \cdots + f_N^2. \tag{1.7}$$

We shall provide some explanation for why we give the name Energy to this quantity $\mathcal{E}_{\mathbf{f}}$ in a moment. First, however, let's look at an example of calculating energy. Suppose $\mathbf{f} = (4, 6, 10, 12, 8, 6, 5, 5)$ is the signal considered in Section 1.1. Then $\mathcal{E}_{\mathbf{f}}$ is calculated as follows:

$$\mathcal{E}_{\mathbf{f}} = 4^2 + 6^2 + \cdots + 5^2 = 446.$$

So the energy of $\mathbf{f}$ is 446. Furthermore, using the values for its 1-level Haar transform $(\mathbf{a}^1 \,|\, \mathbf{d}^1) = (5\sqrt{2}, 11\sqrt{2}, 7\sqrt{2}, 5\sqrt{2} \,|\, -\sqrt{2}, -\sqrt{2}, \sqrt{2}, 0)$, we find that

$$\mathcal{E}_{(\mathbf{a}^1 \,|\, \mathbf{d}^1)} = 25 \cdot 2 + 121 \cdot 2 + \cdots + 2 + 0 = 446$$

as well. Thus the 1-level Haar transform has kept the energy constant. In fact, this is true in general:

Conservation of Energy. *The 1-level Haar transform conserves energy, i.e.,* $\mathcal{E}_{(\mathbf{a}^1 \,|\, \mathbf{d}^1)} = \mathcal{E}_{\mathbf{f}}$ *for every signal* $\mathbf{f}$.

We will explain why this Conservation of Energy property is true for all signals at the end of this section.

Before we go any further, we should say something about why we have given the name Energy to the quantity $\mathcal{E}_{\mathbf{f}}$. The reason is that sums of squares frequently appear in physics when various types of energy are calculated. For instance, if a particle of mass m has a velocity of $\mathbf{v} = (v_1, v_2, v_3)$, then its kinetic energy is $(m/2)(v_1^2 + v_2^2 + v_3^2)$. Hence its kinetic energy is proportional to $v_1^2 + v_2^2 + v_3^2 = \mathcal{E}_{\mathbf{v}}$. Ignoring the constant of proportionality, $m/2$, we obtain the quantity $\mathcal{E}_{\mathbf{v}}$ which we call the energy of $\mathbf{v}$.

While Conservation of Energy is certainly an important property, it is even more important to consider how the Haar transform redistributes the energy in a signal by compressing most of the energy into the trend subsignal. For example, for the signal $\mathbf{f} = (4, 6, 10, 12, 8, 6, 5, 5)$ we found in Section 1.1 that its trend $\mathbf{a}^1$ equals $(5\sqrt{2}, 11\sqrt{2}, 7\sqrt{2}, 5\sqrt{2})$. Therefore, the energy of $\mathbf{a}^1$ is

$$\mathcal{E}_{\mathbf{a}^1} = 25 \cdot 2 + 121 \cdot 2 + 49 \cdot 2 + 25 \cdot 2 = 440.$$

On the other hand, the fluctuation $\mathbf{d}^1$ is $(-\sqrt{2}, -\sqrt{2}, \sqrt{2}, 0)$, which has energy

$$\mathcal{E}_{\mathbf{d}^1} = 2 + 2 + 2 + 0 = 6.$$

Thus the energy of the trend $\mathbf{a}^1$ accounts for $440/446 = 98.7\%$ of the total energy of the signal. In other words, the 1-level Haar transform has redistributed the energy of $\mathbf{f}$ so that over 98% is concentrated into the subsignal $\mathbf{a}^1$ which is half the length of $\mathbf{f}$. For obvious reasons, this is called *compaction of energy.* As another example, consider the signal $\mathbf{f}$ graphed in Figure 1.1(a) and its 1-level Haar transform shown in Figure 1.1(b). In this case, we find that the energy of the signal $\mathbf{f}$ is 127.308 while the energy of its first trend $\mathbf{a}^1$ is 127.305. Thus 99.998% of the total energy is compacted into the half-length subsignal $\mathbf{a}^1$. By examining the graph in Figure 1.1(b) it is easy to see why such a phenomenal energy compaction has occurred; the values of the fluctuation $\mathbf{d}^1$ are so small, relative to the much larger values of the trend $\mathbf{a}^1$, that its energy $\mathcal{E}_{\mathbf{d}^1}$ contributes only a small fraction of the total energy $\mathcal{E}_{\mathbf{a}^1} + \mathcal{E}_{\mathbf{d}^1}$.

These two examples illustrate the following general principle:

Compaction of Energy. *The energy of the trend subsignal* $\mathbf{a}^1$ *accounts for a large percentage of the energy of the transformed signal* $(\mathbf{a}^1 \,|\, \mathbf{d}^1)$.

Compaction of Energy will occur whenever the magnitudes of the fluctuation's values are significantly smaller than the trend's values (recall the Small Fluctuations Feature from the last section).

In Section 1.5, we shall describe how compaction of energy provides a framework for applying the Haar transform to compress signals. We now turn to a discussion of how the Haar transform can be extended to multiple levels, thereby increasing the energy compaction of signals.

Haar transform, multiple levels

Once we have performed a 1-level Haar transform, then it is easy to repeat the process and perform multiple level Haar transforms. After performing a 1-level Haar transform of a signal $\mathbf{f}$ we obtain a first trend $\mathbf{a}^1$ and a first fluctuation $\mathbf{d}^1$. The second level of a Haar transform is then performed by computing a second trend $\mathbf{a}^2$ and a second fluctuation $\mathbf{d}^2$ *for the first trend* $\mathbf{a}^1$ *only.*

For example, if $\mathbf{f} = (4, 6, 10, 12, 8, 6, 5, 5)$ is the signal considered above, then we found that its first trend is $\mathbf{a}^1 = (5\sqrt{2}, 11\sqrt{2}, 7\sqrt{2}, 5\sqrt{2})$. To get the second trend $\mathbf{a}^2$ we apply Formula (1.2) *to the values of* $\mathbf{a}^1$. That is, we add successive pairs of values of $\mathbf{a}^1$ and divide by $\sqrt{2}$ as indicated in the following diagram:

$$\begin{tikzcd}
\mathbf{a}^1\colon & 5\sqrt{2} \arrow[dr] & & 11\sqrt{2} \arrow[dl] & 7\sqrt{2} \arrow[dr] & & 5\sqrt{2} \arrow[dl] \\
\mathbf{a}^2\colon & & 16 & & & 12 &
\end{tikzcd}$$

And to get the second fluctuation $\mathbf{d}^2$ we subtract successive pairs of values of $\mathbf{a}^1$ and divide by $\sqrt{2}$ as indicated in this diagram:

$$\begin{tikzcd}
\mathbf{a}^1\colon & 5\sqrt{2} \arrow[dr] & & 11\sqrt{2} \arrow[dl] & 7\sqrt{2} \arrow[dr] & & 5\sqrt{2} \arrow[dl] \\
\mathbf{d}^2\colon & & -6 & & & 2 &
\end{tikzcd}$$

Thus the 2-level Haar transform of $\mathbf{f}$ is the signal

$$(\mathbf{a}^2 \,|\, \mathbf{d}^2 \,|\, \mathbf{d}^1) = (16, 12 \,|\, -6, 2 \,|\, -\sqrt{2}, -\sqrt{2}, \sqrt{2}, 0).$$

For this signal $\mathbf{f}$, a 3-level Haar transform can also be done, and the result is

$$(\mathbf{a}^3 \,|\, \mathbf{d}^3 \,|\, \mathbf{d}^2 \,|\, \mathbf{d}^1) = (14\sqrt{2} \,|\, 2\sqrt{2} \,|\, -6, 2 \,|\, -\sqrt{2}, -\sqrt{2}, \sqrt{2}, 0).$$

It is interesting to calculate the energy compaction that has occurred with the 2-level and 3-level Haar transforms that we just computed. First, we know that $\mathcal{E}_{(\mathbf{a}^2 \,|\, \mathbf{d}^2 \,|\, \mathbf{d}^1)} = 446$ because of Conservation of Energy. Second, we compute that $\mathcal{E}_{\mathbf{a}^2} = 400$. Thus the 2-level Haar transformed signal $(\mathbf{a}^2 \,|\, \mathbf{d}^2 \,|\, \mathbf{d}^1)$ has almost 90% of the total energy of $\mathbf{f}$ contained in the second trend $\mathbf{a}^2$ which is 1/4 of the length of $\mathbf{f}$. This is a further compaction, or *localization,* of the energy of $\mathbf{f}$. Furthermore, $\mathcal{E}_{\mathbf{a}^3} = 392$; thus $\mathbf{a}^3$ contains 87.89% of the total energy of $\mathbf{f}$. This is even further compaction; the 3-level Haar transform $(\mathbf{a}^3 \,|\, \mathbf{d}^3 \,|\, \mathbf{d}^2 \,|\, \mathbf{d}^1)$ has almost 88% of the total energy of $\mathbf{f}$ contained in the third trend $\mathbf{a}^3$ which is 1/8 the length of $\mathbf{f}$.

For those readers who are familiar with Quantum Theory, there is an interesting phenomenon here that is worth noting. By Heisenberg's Uncertainty Principle, it is impossible to localize a fixed amount of energy into an arbitrarily small time interval. This provides an explanation for why the energy percentage dropped from 98% to 90% when the second-level Haar transform was computed, and from 90% to 88% when the third-level Haar transform was computed. When we attempt to squeeze the energy into ever smaller time intervals, it is inevitable that some energy leaks out.

As another example of how the Haar transform redistributes and localizes the energy in a signal, consider the graphs shown in Figure 1.2. In Figure 1.2(a) we show a signal, and in Figure 1.2(b) we show the 2-level Haar transform of this signal. In Figures 1.2(c) and (d) we show the respective cumulative energy profiles of these two signals. By the *cumulative energy profile* of a signal $\mathbf{f}$ we mean the signal defined by

$$\left(\frac{f_1^2}{\mathcal{E}_{\mathbf{f}}}, \frac{f_1^2 + f_2^2}{\mathcal{E}_{\mathbf{f}}}, \frac{f_1^2 + f_2^2 + f_3^2}{\mathcal{E}_{\mathbf{f}}}, \dots, 1 \right).$$

The cumulative energy profile of $\mathbf{f}$ thus provides a summary of the accumulation of energy in the signal as time proceeds. As can be seen from comparing the two profiles in Figure 1.2, the 2-level Haar transform has redistributed and localized the energy of the original signal.

Justification of Energy Conservation

We close this section with a brief justification of the Conservation of Energy property of the Haar transform. First, we observe that the terms a_1^2 and d_1^2 in the formula $\mathcal{E}_{(\mathbf{a}^1 \,|\, \mathbf{d}^1)} = a_1^2 + \cdots + a_{N/2}^2 + d_1^2 + \cdots + d_{N/2}^2$ add up as follows:

$$\begin{aligned} a_1^2 + d_1^2 &= \left[\frac{f_1 + f_2}{\sqrt{2}} \right]^2 + \left[\frac{f_1 - f_2}{\sqrt{2}} \right]^2 \\ &= \frac{f_1^2 + 2f_1 f_2 + f_2^2}{2} + \frac{f_1^2 - 2f_1 f_2 + f_2^2}{2} \\ &= f_1^2 + f_2^2. \end{aligned}$$

Similarly, $a_m^2 + d_m^2 = f_{2m-1}^2 + f_{2m}^2$ for all other values of m. Therefore, by adding a_m^2 and d_m^2 successively for each m, we find that

$$a_1^2 + \cdots + a_{N/2}^2 + d_1^2 + \cdots + d_{N/2}^2 = f_1^2 + \cdots + f_N^2.$$

In other words, $\mathcal{E}_{(\mathbf{a}^1 \,|\, \mathbf{d}^1)} = \mathcal{E}_{\mathbf{f}}$, which justifies the Conservation of Energy property.

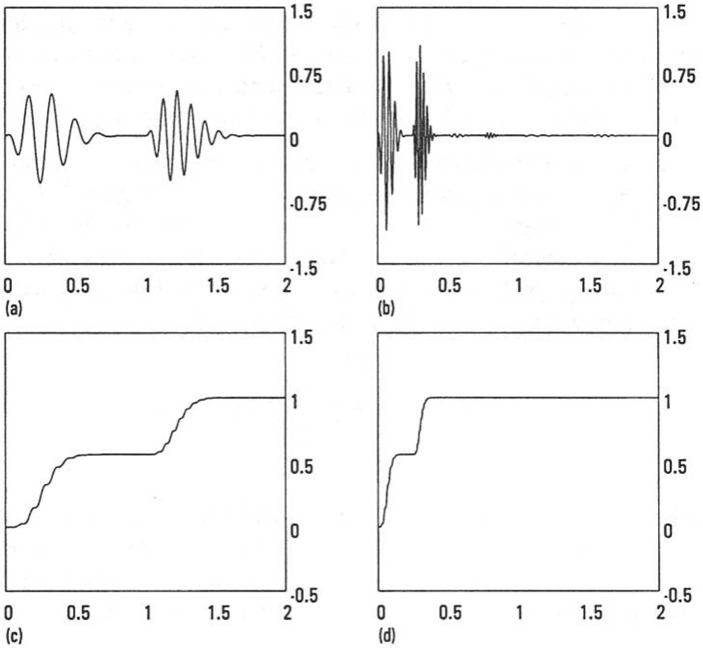

FIGURE 1.2
(a) Signal. (b) 2-level Haar transform of signal. (c) Cumulative energy profile of Signal. (d) Cumulative energy profile of 2-level Haar transform.

1.3 Haar wavelets

In this section we discuss the simplest wavelets, the Haar wavelets. This material will set the stage for the more sophisticated Daubechies wavelets described in the next chapter.

We begin by discussing the 1-level *Haar wavelets.* These wavelets are defined as

$$\begin{aligned}\mathbf{W}_1^1 &= \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0, \dots, 0\right)\\ \mathbf{W}_2^1 &= \left(0, 0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0, \dots, 0\right)\\ &\vdots\\ \mathbf{W}_{N/2}^1 &= \left(0, 0, \dots, 0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right). \end{aligned} \tag{1.8}$$

These 1-level Haar wavelets have a number of interesting properties. First, they each have an energy of 1. Second, they each consist of a rapid fluctuation between just two non-zero values, $\pm 1/\sqrt{2}$, with an average value of 0. Hence the name *wavelet.* Finally, they all are very similar to each other in that they are each a translation in time by an even number of time-units of the first Haar wavelet $\mathbf{W}_1^1$. The second Haar wavelet $\mathbf{W}_2^1$ is a translation forward in time by two units of $\mathbf{W}_1^1$, and $\mathbf{W}_3^1$ is a translation forward in time by four units of $\mathbf{W}_1^1$, and so on.

The reason for introducing the 1-level Haar wavelets is that we can express the 1-level fluctuation subsignal in a simpler form by making use of scalar products with these wavelets. The scalar product is a fundamental operation on two signals, and is defined as follows.

Scalar product: *The scalar product* $\mathbf{f} \cdot \mathbf{g}$ *of the signals* $\mathbf{f} = (f_1, f_2, \ldots, f_N)$ *and* $\mathbf{g} = (g_1, g_2, \ldots, g_N)$ *is defined by*

$$\mathbf{f} \cdot \mathbf{g} = f_1 g_1 + f_2 g_2 + \cdots + f_N g_N. \tag{1.9}$$

Using the 1-level Haar wavelets, we can express the values for the first fluctuation subsignal $\mathbf{d}^1$ as scalar products. For example,

$$d_1 = \frac{f_1 - f_2}{\sqrt{2}} = \mathbf{f} \cdot \mathbf{W}_1^1.$$

Similarly, $d_2 = \mathbf{f} \cdot \mathbf{W}_2^1$, and so on. We can summarize Formula (1.3) in terms of scalar products with the 1-level Haar wavelets:

$$d_m = \mathbf{f} \cdot \mathbf{W}_m^1 \tag{1.10}$$

for $m = 1, 2, \ldots, N/2$.

We can also use the idea of scalar products to restate the Small Fluctuations Feature from Section 1.1 in a more precise form. If we say that the *support* of each Haar wavelet is the set of two time-indices where the wavelet is non-zero, then we have the following more precise version of the Small Fluctuations Feature:

Property 1. *If a signal* $\mathbf{f}$ *is (approximately) constant over the support of a* 1*-level Haar wavelet* $\mathbf{W}_k^1$*, then the fluctuation value* $d_k = \mathbf{f} \cdot \mathbf{W}_k^1$ *is (approximately) zero.*

This property will be considerably strengthened in the next chapter.

Note: *From now on, we shall refer to the set of time-indices* m *where* $f_m \neq 0$ *as the* support *of a signal* $\mathbf{f}$*.*

We can also express the 1-level trend values as scalar products with certain elementary signals. These elementary signals are called 1-level *Haar* *scaling signals,* and they are defined as

$$\begin{aligned}
\mathbf{V}_1^1 &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0, \dots, 0\right) \\
\mathbf{V}_2^1 &= \left(0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0, \dots, 0\right) \\
&\vdots \\
\mathbf{V}_{N/2}^1 &= \left(0, 0, \dots, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).
\end{aligned} \tag{1.11}$$

Using these Haar scaling signals, the values $a_1, \dots, a_{N/2}$ for the first trend are expressed as scalar products:

$$a_m = \mathbf{f} \cdot \mathbf{V}_m^1 \tag{1.12}$$

for $m = 1, 2, \dots, N/2$.

The Haar scaling signals are quite similar to the Haar wavelets. They all have energy 1 and have a support consisting of just two consecutive time-indices. In fact, they are all translates by an even multiple of time-units of the first scaling signal $\mathbf{V}_1^1$. Unlike the Haar wavelets, however, the average values of the Haar scaling signals are not zero. In fact, they each have an average value of $1/\sqrt{2}$.

The ideas discussed above extend to every level. For simplicity, we restrict our discussion to the second level. The 2-level Haar scaling signals are defined by

$$\begin{aligned}
\mathbf{V}_1^2 &= \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, 0 \dots, 0\right) \\
\mathbf{V}_2^2 &= \left(0, 0, 0, 0, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, 0, \dots, 0\right) \\
&\vdots \\
\mathbf{V}_{N/4}^2 &= \left(0, 0, \dots, 0, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right).
\end{aligned} \tag{1.13}$$

These scaling signals are all translations by multiples of four time-units of the first scaling signal $\mathbf{V}_1^2$, and they all have energy 1 and average value 1/2. Furthermore, the values of the 2-level trend $\mathbf{a}^2$ are scalar products of these scaling signals with the signal $\mathbf{f}$. That is, $\mathbf{a}^2$ satisfies

$$\mathbf{a}^2 = \left(\mathbf{f} \cdot \mathbf{V}_1^2, \mathbf{f} \cdot \mathbf{V}_2^2, \dots, \mathbf{f} \cdot \mathbf{V}_{N/4}^2\right). \tag{1.14}$$

Likewise, the 2-level Haar wavelets are defined by

$$\mathbf{W}_1^2 = \left(\frac{1}{2}, \frac{1}{2}, \frac{-1}{2}, \frac{-1}{2}, 0, 0 \dots, 0\right)$$

$$\mathbf{W}_2^2 = \left(0,0,0,0,\frac{1}{2},\frac{1}{2},\frac{-1}{2},\frac{-1}{2},0,0,\dots,0\right)$$

$$\vdots$$

$$\mathbf{W}_{N/4}^2 = \left(0,0,\dots,0,\frac{1}{2},\frac{1}{2},\frac{-1}{2},\frac{-1}{2}\right). \tag{1.15}$$

These wavelets all have supports of length 4, since they are all translations by multiples of four time-units of the first wavelet $\mathbf{W}_1^2$. They also all have energy 1 and average value 0. Using scalar products, the 2-level fluctuation $\mathbf{d}^2$ satisfies

$$\mathbf{d}^2 = \left(\mathbf{f}\cdot\mathbf{W}_1^2,\, \mathbf{f}\cdot\mathbf{W}_2^2,\dots,\, \mathbf{f}\cdot\mathbf{W}_{N/4}^2\right). \tag{1.16}$$

1.4 Multiresolution analysis

In the previous section we discussed how the Haar transform can be described using scalar products with scaling signals and wavelets. In this section we discuss how the inverse Haar transform can also be described in terms of these same elementary signals. This discussion will show how discrete signals are synthesized by beginning with a very low resolution signal and successively adding on details to create higher resolution versions, ending with a complete synthesis of the signal at the finest resolution. This is known as *multiresolution analysis* (MRA). MRA is the heart of wavelet analysis.

In order to make these ideas precise, we must first discuss some elementary operations that can be performed on signals. Given two signals $\mathbf{f} = (f_1, f_2, \dots, f_N)$ and $\mathbf{g} = (g_1, g_2, \dots, g_N)$, we can perform the following elementary algebraic operations:

Addition and Subtraction: The *sum* $\mathbf{f}+\mathbf{g}$ of the signals $\mathbf{f}$ and $\mathbf{g}$ is defined by adding their values:

$$\mathbf{f}+\mathbf{g} = (f_1+g_1, f_2+g_2, \dots, f_N+g_N). \tag{1.17}$$

Their *difference* $\mathbf{f}-\mathbf{g}$ is defined by subtracting their values:

$$\mathbf{f}-\mathbf{g} = (f_1-g_1, f_2-g_2, \dots, f_N-g_N). \tag{1.18}$$

Constant multiple: A signal $\mathbf{f}$ is multiplied by a constant c by multiplying each of its values by c. That is,

$$c\,\mathbf{f} = (cf_1, cf_2, \dots, cf_N). \tag{1.19}$$

For example, by repeatedly applying the addition operation, we can express a signal $\mathbf{f} = (f_1, f_2, \ldots, f_N)$ as follows:

$$\mathbf{f} = (f_1, 0, 0, \ldots, 0) + (0, f_2, 0, 0, \ldots, 0) + \cdots + (0, 0, \ldots, 0, f_N).$$

Then, by applying the constant multiple operation to each of the signals on the right side of this last equation, we obtain

$$\mathbf{f} = f_1(1, 0, 0, \ldots, 0) + f_2(0, 1, 0, 0, \ldots, 0) + \cdots + f_N(0, 0, \ldots, 0, 1).$$

This formula is a very natural one; it amounts to expressing $\mathbf{f}$ as a sum of its individual values at each discrete instant of time.

If we define the elementary signals $\mathbf{V}_1^0, \mathbf{V}_2^0, \ldots, \mathbf{V}_N^0$ as

$$\begin{aligned} \mathbf{V}_1^0 &= (1, 0, 0, \ldots, 0) \\ \mathbf{V}_2^0 &= (0, 1, 0, 0, \ldots, 0) \\ &\vdots \\ \mathbf{V}_N^0 &= (0, 0, \ldots, 0, 1) \end{aligned} \tag{1.20}$$

then the last formula for $\mathbf{f}$ can be rewritten as

$$\mathbf{f} = f_1\mathbf{V}_1^0 + f_2\mathbf{V}_2^0 + \cdots + f_N\mathbf{V}_N^0. \tag{1.21}$$

Formula (1.21) is called the *natural expansion* of a signal $\mathbf{f}$ in terms of the *natural basis* of signals $\mathbf{V}_1^0, \mathbf{V}_2^0, \ldots, \mathbf{V}_N^0$. We shall now show that the Haar MRA involves expressing $\mathbf{f}$ as a sum of constant multiples of a different basis set of elementary signals, the Haar wavelets and scaling signals defined in the previous section.

In the previous section, we showed how to express the 1-level Haar transform in terms of wavelets and scaling signals. It is also possible to express the inverse of the 1-level Haar transform in terms of these same elementary signals. This leads to the first level of the Haar MRA. To define this first level Haar MRA we make use of (1.6) to express a signal $\mathbf{f}$ as

$$\begin{aligned} \mathbf{f} = &\left(\frac{a_1}{\sqrt{2}}, \frac{a_1}{\sqrt{2}}, \frac{a_2}{\sqrt{2}}, \frac{a_2}{\sqrt{2}}, \ldots, \frac{a_{N/2}}{\sqrt{2}}, \frac{a_{N/2}}{\sqrt{2}}\right) \\ &+ \left(\frac{d_1}{\sqrt{2}}, \frac{-d_1}{\sqrt{2}}, \frac{d_2}{\sqrt{2}}, \frac{-d_2}{\sqrt{2}}, \ldots, \frac{d_{N/2}}{\sqrt{2}}, \frac{-d_{N/2}}{\sqrt{2}}\right). \end{aligned}$$

This formula shows that the signal $\mathbf{f}$ can be expressed as the sum of two signals that we shall call the first averaged signal and the first detail signal. That is, we have

$$\mathbf{f} = \mathbf{A}^1 + \mathbf{D}^1 \tag{1.22}$$

where the signal $\mathbf{A}^1$ is called the *first averaged signal* and is defined by

$$\mathbf{A}^1 = \left(\frac{a_1}{\sqrt{2}}, \frac{a_1}{\sqrt{2}}, \frac{a_2}{\sqrt{2}}, \frac{a_2}{\sqrt{2}}, \dots, \frac{a_{N/2}}{\sqrt{2}}, \frac{a_{N/2}}{\sqrt{2}} \right) \tag{1.23}$$

and the signal $\mathbf{D}^1$ is called the *first detail signal* and is defined by

$$\mathbf{D}^1 = \left(\frac{d_1}{\sqrt{2}}, \frac{-d_1}{\sqrt{2}}, \frac{d_2}{\sqrt{2}}, \frac{-d_2}{\sqrt{2}}, \dots, \frac{d_{N/2}}{\sqrt{2}}, \frac{-d_{N/2}}{\sqrt{2}} \right). \tag{1.24}$$

Using Haar scaling signals and wavelets, and using the basic elementary algebraic operations with signals, the averaged and detail signals are expressible as

$$\mathbf{A}^1 = a_1 \mathbf{V}_1^1 + a_2 \mathbf{V}_2^1 + \cdots + a_{N/2} \mathbf{V}_{N/2}^1, \tag{1.25a}$$

$$\mathbf{D}^1 = d_1 \mathbf{W}_1^1 + d_2 \mathbf{W}_2^1 + \cdots + d_{N/2} \mathbf{W}_{N/2}^1. \tag{1.25b}$$

Applying the scalar product formulas for the coefficients in Equations (1.10) and (1.12), we can rewrite these last two formulas as follows

$$\mathbf{A}^1 = (\mathbf{f} \cdot \mathbf{V}_1^1)\mathbf{V}_1^1 + (\mathbf{f} \cdot \mathbf{V}_2^1)\mathbf{V}_2^1 + \cdots + (\mathbf{f} \cdot \mathbf{V}_{N/2}^1)\mathbf{V}_{N/2}^1,$$

$$\mathbf{D}^1 = (\mathbf{f} \cdot \mathbf{W}_1^1)\mathbf{W}_1^1 + (\mathbf{f} \cdot \mathbf{W}_2^1)\mathbf{W}_2^1 + \cdots + (\mathbf{f} \cdot \mathbf{W}_{N/2}^1)\mathbf{W}_{N/2}^1.$$

These formulas show that the averaged signal is a combination of Haar scaling signals, with the values of the first trend subsignal as coefficients; and that the detail signal is a combination of Haar wavelets, with the values of the first fluctuation subsignal as coefficients.

As an example of these ideas, consider the signal

$$\mathbf{f} = (4, 6, 10, 12, 8, 6, 5, 5).$$

In Section 1.1 we found that its first trend subsignal was

$$\mathbf{a}^1 = (5\sqrt{2}, 11\sqrt{2}, 7\sqrt{2}, 5\sqrt{2}).$$

Applying Formula (1.23), the averaged signal is

$$\mathbf{A}^1 = (5, 5, 11, 11, 7, 7, 5, 5). \tag{1.26}$$

Notice how the first averaged signal consists of the repeated average values $5, 5$, and $11, 11$, and $7, 7$, and $5, 5$ about which the values of $\mathbf{f}$ fluctuate. Using Formula (1.25a), the first averaged signal can also be expressed in terms of Haar scaling signals as

$$\mathbf{A}^1 = 5\sqrt{2}\, \mathbf{V}_1^1 + 11\sqrt{2}\, \mathbf{V}_2^1 + 7\sqrt{2}\, \mathbf{V}_3^1 + 5\sqrt{2}\, \mathbf{V}_4^1.$$

Comparing these last two equations we can see that *the positions of the repeated averages correspond precisely with the supports of the scaling signals.*

We also found in Section 1.1 that the first fluctuation signal for $\mathbf{f}$ was $\mathbf{d}^1 = (-\sqrt{2}, -\sqrt{2}, \sqrt{2}, 0)$. Formula (1.24) then yields

$$\mathbf{D}^1 = (-1, 1, -1, 1, 1, -1, 0, 0).$$

Thus, using the result for $\mathbf{A}^1$ computed above, we have

$$\mathbf{f} = (5, 5, 11, 11, 7, 7, 5, 5) + (-1, 1, -1, 1, 1, -1, 0, 0).$$

This equation illustrates the basic idea of MRA. The signal $\mathbf{f}$ is expressed as a sum of a lower resolution, or averaged, signal $(5, 5, 11, 11, 7, 7, 5, 5)$ added with a signal $(-1, 1, -1, 1, 1, -1, 0, 0)$ made up of fluctuations or details. These fluctuations provide the added details necessary to produce the full resolution signal $\mathbf{f}$.

For this example, using Formula (1.25b), the first detail signal can also be expressed in terms of Haar wavelets as

$$\mathbf{D}^1 = -\sqrt{2}\,\mathbf{W}^1_1 - \sqrt{2}\,\mathbf{W}^1_2 + \sqrt{2}\,\mathbf{W}^1_3 + 0\,\mathbf{W}^1_4.$$

This formula shows that the values of $\mathbf{D}^1$ occur in successive pairs of rapidly fluctuating values positioned at the supports of the Haar wavelets.

Multiresolution analysis, multiple levels

In the discussion above, we described the first level of the Haar MRA of a signal. This idea can be extended to further levels, as many levels as the number of times that the signal length can be divided by 2.

The second level of a MRA of a signal $\mathbf{f}$ involves expressing $\mathbf{f}$ as

$$\mathbf{f} = \mathbf{A}^2 + \mathbf{D}^2 + \mathbf{D}^1. \tag{1.27}$$

Here $\mathbf{A}^2$ is the second averaged signal and $\mathbf{D}^2$ is the second detail signal. Comparing Formulas (1.22) and (1.27) we see that

$$\mathbf{A}^1 = \mathbf{A}^2 + \mathbf{D}^2. \tag{1.28}$$

This formula expresses the fact that computing the second averaged signal $\mathbf{A}^2$ and second detail signal $\mathbf{D}^2$ simply consists of performing a first level MRA of the signal $\mathbf{A}^1$. Because of this, it follows that the second level averaged signal $\mathbf{A}^2$ satisfies

$$\mathbf{A}^2 = (\mathbf{f} \cdot \mathbf{V}^2_1)\mathbf{V}^2_1 + (\mathbf{f} \cdot \mathbf{V}^2_2)\mathbf{V}^2_2 + \cdots + (\mathbf{f} \cdot \mathbf{V}^2_{N/4})\mathbf{V}^2_{N/4}$$

and the second level detail signal $\mathbf{D}^2$ satisfies

$$\mathbf{D}^2 = (\mathbf{f} \cdot \mathbf{W}^2_1)\mathbf{W}^2_1 + (\mathbf{f} \cdot \mathbf{W}^2_2)\mathbf{W}^2_2 + \cdots + (\mathbf{f} \cdot \mathbf{W}^2_{N/4})\mathbf{W}^2_{N/4}.$$

For example, if $\mathbf{f} = (4, 6, 10, 12, 8, 6, 5, 5)$, then we found in Section 1.2 that $\mathbf{a}^2 = (16, 12)$. Therefore

$$\begin{aligned} \mathbf{A}^2 &= 16\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0\right) + 12\left(0, 0, 0, 0, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right) \\ &= (8, 8, 8, 8, 6, 6, 6, 6). \end{aligned} \tag{1.29}$$

It is interesting to compare the equations in (1.26) and (1.29). The second averaged signal $\mathbf{A}^2$ has values created from averages that involve twice as many values as the averages that created $\mathbf{A}^1$. Therefore, the second averaged signal reflects more long term trends than those reflected in the first averaged signal. Consequently, these averages are repeated for twice as many time-units.

We also found in Section 1.2 that this signal $\mathbf{f} = (4, 6, 10, 12, 8, 6, 5, 5)$ has the second fluctuation $\mathbf{d}^2 = (-6, 2)$. Consequently

$$\begin{aligned} \mathbf{D}^2 &= -6\left(\frac{1}{2}, \frac{1}{2}, \frac{-1}{2}, \frac{-1}{2}, 0, 0, 0, 0\right) + 2\left(0, 0, 0, 0, \frac{1}{2}, \frac{1}{2}, \frac{-1}{2}, \frac{-1}{2}\right) \\ &= (-3, -3, 3, 3, 1, 1, -1, -1). \end{aligned}$$

We found above that $\mathbf{D}^1 = (-1, 1, -1, 1, 1, -1, 0, 0)$. Hence

$$\begin{aligned} \mathbf{f} &= \mathbf{A}^2 + \mathbf{D}^2 + \mathbf{D}^1 \\ &= (8, 8, 8, 8, 6, 6, 6, 6) + (-3, -3, 3, 3, 1, 1, -1, -1) \\ &\qquad\qquad + (-1, 1, -1, 1, 1, -1, 0, 0). \end{aligned}$$

This formula further illustrates the idea of MRA. The full resolution signal $\mathbf{f}$ is produced from a very low resolution, averaged signal $\mathbf{A}^2$ consisting of repetitions of the two averaged values, 8 and 6, to which are added two detail signals. The first addition supplements this averaged signal with enough details to produce the next higher resolution averaged signal $(5, 5, 11, 11, 7, 7, 5, 5)$, and the second addition then supplies enough further details to produce the full resolution signal $\mathbf{f}$.

In general, if the number N of signal values is divisible k times by 2, then a k-level MRA:

$$\mathbf{f} = \mathbf{A}^k + \mathbf{D}^k + \cdots + \mathbf{D}^2 + \mathbf{D}^1$$

can be performed on the signal $\mathbf{f}$. Rather than subjecting the reader to the gory details, we conclude by describing a computer example generated using FAWAV. In Figure 1.3 we show a 10-level Haar MRA of the signal $\mathbf{f}$ shown in Figure 1.1(a). This signal has 2^{10} values so 10 levels of MRA are possible. On the top of Figure 1.3(a), the graph of $\mathbf{A}^{10}$ is shown; it consists of a single value repeated 2^{10} times. This value is the average of

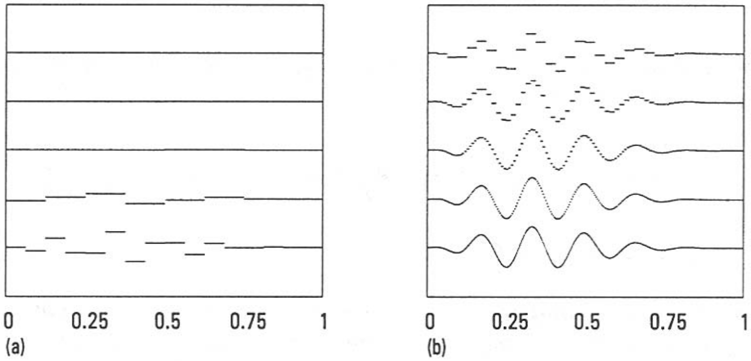

FIGURE 1.3
Haar MRA of the signal in Figure 1.1(a). The graphs are of the ten averaged signals $\mathbf{A}^{10}$ through $\mathbf{A}^1$. Beginning with the signal $\mathbf{A}^{10}$ on the top left down to $\mathbf{A}^6$ on the bottom left, then $\mathbf{A}^5$ on the top right down to $\mathbf{A}^1$ on the bottom right.

all 2^{10} values of the signal $\mathbf{f}$. The graph directly below it is of the signal $\mathbf{A}^9$ which equals $\mathbf{A}^{10}$ plus the details in $\mathbf{D}^{10}$. Each successive averaged signal is shown, from $\mathbf{A}^{10}$ through $\mathbf{A}^1$. By successively adding on details, the full signal in Figure 1.1(a) is systematically constructed in all its complexity.

1.5 Compression of audio signals

In Section 1.2 we saw that the Haar transform can be used to localize the energy of a signal into a shorter subsignal. In this section we show how this redistribution of energy can be used to compress audio signals. By compressing an audio signal we mean converting the signal data into a new format that requires less bits to transmit. When we use the term, *audio signal,* we are speaking somewhat loosely. Many of the signals we have in mind are indeed the result of taking discrete samples of a sound signal—as in the data in a computer audio file, or on a compact disc—but the techniques developed here also apply to digital data transmissions and to other digital signals, such as digitized electrocardiograms or digitized electroencephalograms.

There are two basic categories of compression techniques. The first category is *lossless compression.* Lossless compression methods achieve completely error free decompression of the original signal. Typical lossless methods are Huffman compression, LZW compression, arithmetic compression, or run-length compression. Combinations of these techniques are used in popular lossless compression programs, such as the kind that produce `.zip` files. Unfortunately, the compression ratios that can be obtained with lossless methods are rarely more than 2:1 for audio files consisting of music or speech.

The second category is *lossy compression.* A lossy compression method is one which produces inaccuracies in the decompressed signal. Lossy techniques are used when these inaccuracies are so small as to be imperceptible. The advantage of lossy techniques over lossless ones is that much higher compression ratios can be attained. With wavelet compression methods, which are lossy, if we are willing to accept the slight inaccuracies in the decompressed signal, then we can obtain compression ratios of 10:1, or 20:1, or as high as 50:1 or even 100:1.

In order to illustrate the general principles of wavelet compression of signals, we shall examine, in a somewhat simplified way, how the Haar wavelet transform can be used to compress some test signals. For example, Signal 1 in Figure 1.4(a) can be very effectively compressed using the Haar transform. Although Signal 1 is not a very representative audio signal, it is representative of a portion of a digital data transmission. This signal has 1024 values equally spaced over the time interval $[0, 20)$. Most of these values are constant over long stretches, and that is the principal reason that Signal 1 can be compressed effectively with the Haar transform. Signal 2 in Figure 1.5(a), however, will not compress nearly so well; this signal requires the more sophisticated wavelet transforms described in the next chapter.

The basic steps for wavelet compression are as follows:

Method of Wavelet Transform Compression

Step 1. Perform a wavelet transform of the signal.

Step 2. Set equal to 0 all values of the wavelet transform which are insignificant, i.e., which lie below some *threshold value.*

Step 3. Transmit only the significant, non-zero values of the transform obtained from Step 2. This should be a much smaller data set than the original signal.

Step 4. At the receiving end, perform the inverse wavelet transform of the data transmitted in Step 3, assigning zero values to the insignificant values which were not transmitted. This decompression step produces an approximation of the original signal.

In this chapter we shall illustrate this method using the Haar wavelet transform. This initial discussion will be significantly deepened and generalized in the next chapter when we discuss this method of compression in terms of various Daubechies wavelet transforms.

Let's now examine a Haar wavelet transform compression of Signal 1. We begin with Step 1. Since Signal 1 consists of $1024 = 2^{10}$ values, we

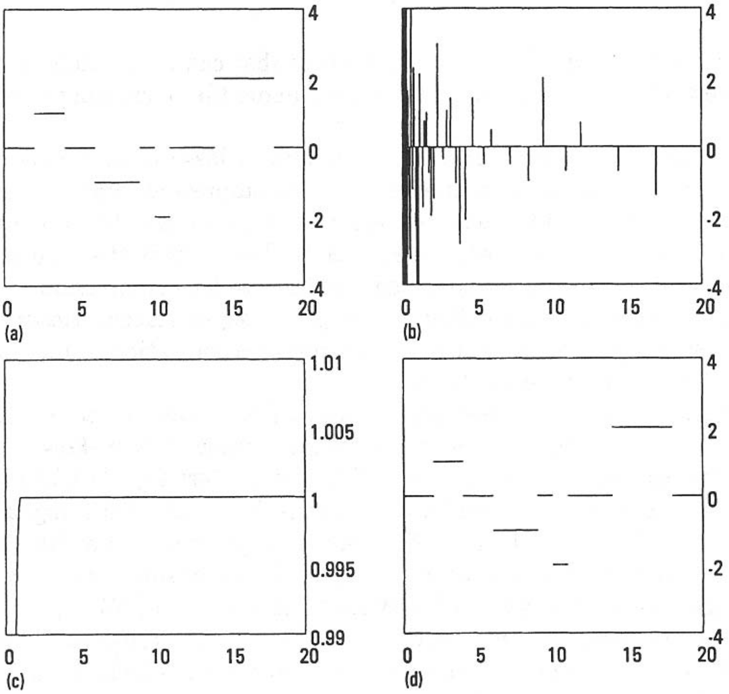

FIGURE 1.4
(a) Signal 1, (b) 10-level Haar transform of Signal 1, (c) energy map of Haar transform, (d) 20:1 compression of Signal 1, 100% of energy.

can perform 10 levels of the Haar transform. This 10-level Haar transform is shown in Figure 1.4(b). Notice how a large portion of the Haar transform's values are 0, or very near 0, in magnitude. This fact provides the fundamental basis for performing an effective compression.

In order to choose a threshold value for Step 2, we proceed as follows. First, we arrange the magnitudes of the values of the Haar transform so that they are in decreasing order:

$$L_1 \geq L_2 \geq L_3 \geq \ldots \geq L_N$$

where L_1 is the largest absolute value of the Haar transform, L_2 is the next largest, etc. (In the event of a tie, we just leave those magnitudes in their original order.) We then compute the cumulative energy profile of this new signal:

$$\left(\frac{L_1^2}{\mathcal{E}_{\mathbf{f}}}, \frac{L_1^2 + L_2^2}{\mathcal{E}_{\mathbf{f}}}, \frac{L_1^2 + L_2^2 + L_3^2}{\mathcal{E}_{\mathbf{f}}}, \ldots, 1 \right).$$

For Signal 1, we show a graph of this energy profile—which we refer to as the *energy map* of the Haar transform—in Figure 1.4(c). Notice that the energy map very quickly reaches its maximum value of 1. In fact, using FAWAV we find that

$$\frac{L_1^2 + L_2^2 + \ldots + L_{51}^2}{\mathcal{E}_{\mathbf{f}}} = .999996.$$

Consequently, if we choose a threshold T that is less than $L_{51} = .3536$, then the values of the transform that survive this threshold will account for essentially 100% of the energy of Signal 1.

We now turn to Step 3. In order to perform Step 3—transmitting only the significant transform values—an additional amount of information must be sent which indicates the positions of these significant transform values in the thresholded transform. This information is called the *significance map.* The values of this significance map are either 1 or 0: a value of 1 if the corresponding transform value survived the thresholding, a value of 0 if it did not. The significance map is therefore a string of N bits, where N is the length of the signal. For the case of Signal 1, with a threshold of .35, there are only 51 non-zero bits in the significance map out of a total of 1024 bits. Therefore, since most of this significance map consists of long stretches of zeros, it can be very effectively compressed using one of the lossless compression algorithms mentioned above. This compressed string of bits is then transmitted along with the non-zero values of the thresholded transform.

Finally, we arrive at Step 4. At the receiving end, the significance map is used to insert zeros in their proper locations in between the non-zero values in the thresholded transform, and then an inverse transform is computed to produce an approximation of the signal. For Signal 1 we show the approximation that results from using a threshold of .35 in Figure 1.4(d). This approximation used only 51 transform values; so it represents a compression of Signal 1 by a factor of 1024:51, i.e., a compression factor of 20:1. Since the compressed signal contains nearly 100% of the energy of the original signal, it is a very good approximation. In fact, the maximum error over all values is no more than 3.91×10^{-3}.

Life would be simpler if the Haar transform could be used so effectively for all signals. Unfortunately, if we try to use the Haar transform for threshold compression of Signal 2 in Figure 1.5(a), we get poor results. This signal, when played over a computer sound system, produces a sound similar to two low notes played on a clarinet. It has $4096 = 2^{12}$ values; so we can perform 12 levels of the Haar transform. In Figure 1.5(b) we show a plot of the 12-level Haar transform of Signal 2. It is clear from this plot that a large fraction of the Haar transform values have significant magnitude, significant enough that they are visible in the graph. In fact, the energy map for the transform of Signal 2, shown in Figure 1.5(c), exhibits a much slower increase towards 1 in comparison with the energy map for the transform of Signal 1. Therefore, many more transform values are needed

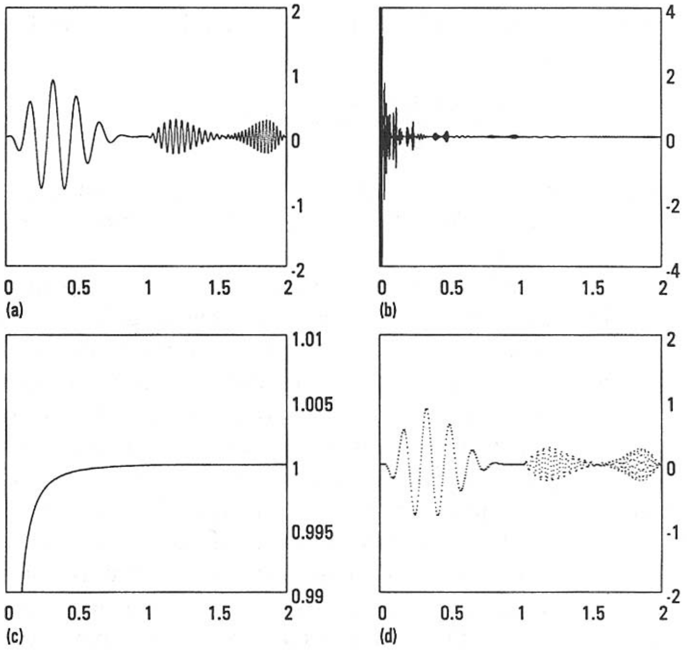

FIGURE 1.5
(a) Signal 2, (b) 12-level Haar transform of Signal 2, (c) energy map of Haar transform, (d) 10:1 compression of Signal 2, 99.6% of energy of Signal 2.

in order to capture a high percentage of the energy of Signal 2. In Figure 1.5(d), we show a 10:1 compression of Signal 2 which captures 99.6% of the energy of Signal 2. Comparing this compression with the original signal we see that it is a fairly poor approximation. Many of the signal values are clumped together in the compressed signal, producing a very ragged or jumpy approximation of the original signal. When this compressed version is played on a computer sound system, it produces a screechy "metallic" version of the two clarinet notes, which is not a very satisfying result. As a rule of thumb, we must capture at least 99.99% of the energy of the signal in order to produce an acceptable approximation, i.e., an approximation that is not perceptually different from the original. Achieving this accurate an approximation for Signal 2 requires at least 1782 transform values. Because Signal 2 itself has 4096 values, this is a compression ratio of only about 2.3:1, which is not very high. We shall see in the next chapter that Signal 2 can be compressed very effectively, but we shall need more high powered wavelet transforms to do it.

A note on quantization

The most serious oversimplification that we made in the discussion above is that we ignored the issue known as *quantization.* The term quantization is used whenever it is necessary to take into account the finite precision of numerical data handled by digital methods. For example, the numerical data used to generate the graphs of Signals 1 and 2 above were IEEE double precision numbers that use 8 bytes = 64 bits for each number. In order to compress this data even further, we can represent the wavelet transform coefficients using less bits. We shall address this issue of quantization in the next chapter when we look again at the problem of compression.

1.6 Removing noise from audio signals

In this section we shall begin our treatment of one of the most important aspects of signal processing, the removal of noise from signals. Our discussion in this section will introduce the fundamental ideas involved in the context of the Haar transform. In the next chapter we shall considerably deepen and generalize these ideas, in the context of the more powerful Daubechies wavelet transforms.

When a signal is received after transmission over some distance, it is frequently contaminated by noise. The term *noise* refers to any undesired change that has altered the values of the original signal. The simplest model for acquisition of noise by a signal is *additive* noise, which has the form

$$(\textit{contaminated signal}) = (\textit{original signal}) + (\textit{noise}). \tag{1.30}$$

We shall represent this equation in a more compact way as

$$\mathbf{f} = \mathbf{s} + \mathbf{n} \tag{1.31}$$

where **f** is the contaminated signal, **s** is the original signal, and **n** is the noise signal.

There are several kinds of noise. A few of the commonly encountered types are the following:

1. *Random noise.* The noise signal is highly oscillatory, its values alternating rapidly between values above and below an average, or mean, value. For simplicity, we shall examine random noise with a mean value of 0.

2. *Pop noise.* This type of noise is heard on old analog recordings obtained from phonograph records. The noise is perceived as randomly

occurring, isolated "pops." As a model for this type of noise we add a few non-zero values to the original signal at isolated locations.

3. *Localized random noise.* Sometimes the noise appears as in type 1, but only over a short segment or segments of the signal. This can occur when there is a short-lived disturbance in the environment during transmission of the signal.

Of course, there can also be noise signals which combine aspects of each of these types. In this section we shall examine only the first type of noise, random noise. The other types will be considered later.

Our approach will be similar to how we treated compression in the last section; we shall examine how noise removal is performed on two test signals using the Haar transform. For the first test signal, the Haar transform is used very effectively for removing the noise. For the second signal, however, the Haar transform performs poorly, and we shall need to use more sophisticated wavelet transforms to remove the noise from this signal. The essential principles, however, underlying these more sophisticated wavelet methods are the same principles we describe here for the Haar transform.

We begin by stating a basic method for removing random noise. Then we examine how this method performs on the two test signals.

Threshold Method of Wavelet Denoising

Suppose that the contaminated signal $\mathbf{f}$ equals the transmitted signal $\mathbf{s}$ plus the noise signal $\mathbf{n}$. Also suppose that the following two conditions hold:

1. The energy of the original signal $\mathbf{s}$ is effectively captured, to a high percentage, by transform values whose magnitudes are all greater than a *threshold* $T_{\mathbf{s}} > 0$.

2. The noise signal's transform values all have magnitudes which lie below a *noise threshold* $T_{\mathbf{n}}$ satisfying $T_{\mathbf{n}} < T_{\mathbf{s}}$.

Then the noise in $\mathbf{f}$ can be removed by thresholding its transform: *All values of its transform whose magnitudes lie below the noise threshold* $T_{\mathbf{n}}$ *are set equal to* 0 *and an inverse transform is performed, providing a good approximation of* $\mathbf{f}$.

Let's see how this method applies to Signal A shown in Figure 1.6(a). This signal was obtained by adding random noise, whose values oscillate between ±0.1 with a mean of zero, to Signal 1 shown in Figure 1.6(a). In this case, Signal 1 is the original signal and Signal A is the contaminated signal. As we saw in the last section, the energy of Signal 1 is captured very effectively by the relatively few transform values whose magnitudes lie above a threshold of .35. So we set $T_{\mathbf{s}}$ equal to .35, and condition 1 in the Denoising Method is satisfied.

Now as for condition 2, look at the 10-level Haar transform of Signal A shown in Figure 1.6(b). Comparing this Haar transform with the Haar transform of Signal 1 in Figure 1.4(b), it is clear that the added noise has contributed a large number of small magnitude values to the transform of Signal A, while the high-energy transform values of Signal 1 are plainly visible (although slightly altered by the addition of noise). Therefore, we can satisfy condition 2 and eliminate the noise if we choose a noise threshold of, say, $T_{\mathbf{n}} = .25$. This is indicated by the two horizontal lines shown in Figure 1.6(b); all transform values lying between ±.25 are set equal to 0, producing the thresholded transform shown in Figure 1.6(c). Comparing Figure 1.6(c) with Figure 1.4(b) we see that the thresholded Haar transform of the contaminated signal is a close match to the Haar transform of the original signal. Consequently, after performing an inverse transform on this thresholded signal, we obtain a denoised signal that is a close match to the original signal. This denoised signal is shown in Figure 1.6(d), and it is clearly a good approximation to Signal 1, especially considering how much noise was originally present in Signal A.

The effectiveness of noise removal can be quantitatively measured in the following way. The *Root Mean Square Error* (RMS Error) of the contaminated signal $\mathbf{f}$ compared with the original signal $\mathbf{s}$ is defined to be

$$\textit{RMS Error} = \sqrt{\frac{(f_1 - s_1)^2 + (f_2 - s_2)^2 + \cdots + (f_N - s_N)^2}{N}}\,. \tag{1.32}$$

Since $\mathbf{f} = \mathbf{s} + \mathbf{n}$, then $\mathbf{n} = \mathbf{f} - \mathbf{s}$. Consequently, the values of $\mathbf{n}$ are formed from the differences of the values of $\mathbf{f}$ and $\mathbf{s}$; so we can rewrite (1.32) as

$$\textit{RMS Error} = \sqrt{\frac{n_1^2 + n_2^2 + \cdots + n_N^2}{N}} = \frac{\sqrt{\mathcal{E}_{\mathbf{n}}}}{\sqrt{N}}\,. \tag{1.33}$$

Equation (1.33) says that the RMS Error equals the square root of the noise energy divided by $\sqrt{N}$, where N is the number of values of the signals. For example, for Signal A the RMS Error between it and Signal 1 is .057. After denoising, the RMS Error between the denoised signal and Signal 1 is .011, which shows that there is a five-fold reduction in the amount of noise. This gives quantitative evidence for the effectiveness of the denoising of Signal A.

Summarizing this example, we can say that the denoising was effective for two reasons: (1) *the transform was able to compress the energy of the original signal into a few high-energy values,* and (2) *the added noise was transformed into low-energy values.* Consequently, the high-energy transform values from the original signal stood out clearly from the low-energy noise transform values which could then be eliminated by thresholding.

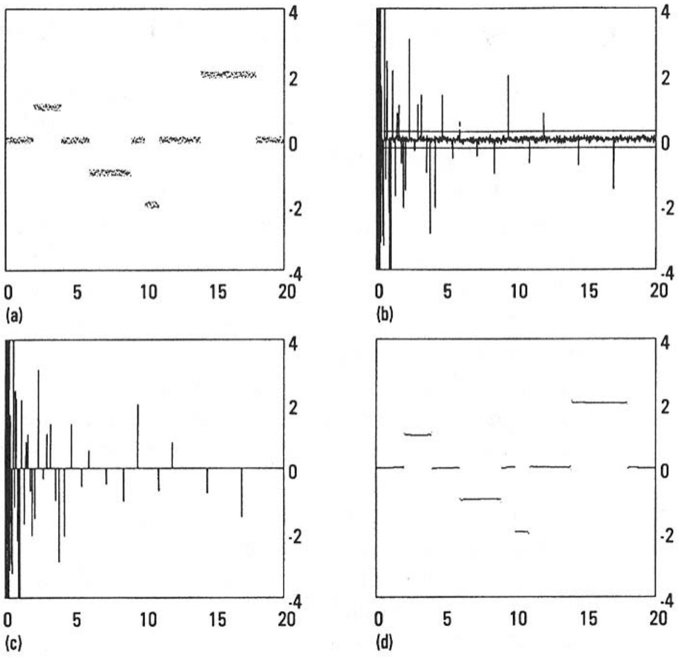

FIGURE 1.6
(a) Signal A, 2^{10} values. (b) 10-level Haar transform of Signal A. The two horizontal lines are at values of $\pm.25$ where $.25$ is a denoising threshold. (c) Thresholded transform. (d) Denoised signal.

Unfortunately, denoising with the Haar transform is not always so effective. Consider, for example, Signal B shown in Figure 1.7(a). This signal consists of Signal 2, shown in Figure 1.5(a), with random noise added. We view Signal 2 as the original signal and Signal B as the contaminated signal. As with the first case considered above, the random noise has values that oscillate between ±0.1 with a mean of zero. In this case, however, we saw in the last section that it takes a relatively large number of transform values to capture the energy in Signal 2. Most of these transform values are of low energy, and it takes many of them to produce a good approximation of Signal 2. When the random noise is added to Signal 2, then the Haar transform, just like in the previous case, produces many small transform values which lie below a noise threshold. This is illustrated in Figure 1.7(b) where we show the 12-level Haar transform of Signal B. As can be seen by comparing Figure 1.7(b) with Figure 1.5(b), the small transform values that come from the noise *obscure most of the small magnitude values that result from the original signal.* Consequently, when a thresholding is done to remove the noise, as indicated by the horizontal lines in Figure 1.7(b), *this removes*

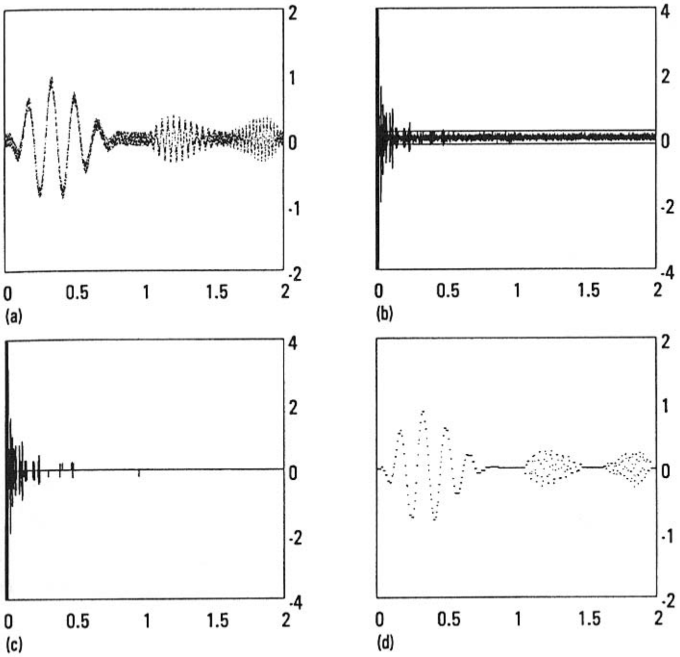

FIGURE 1.7
(a) Signal B, 2^{12} values. (b) 12-level Haar transform of Signal B. The two horizontal lines are at values of $\pm .2$ where $.2$ is the denoising threshold. (c) Thresholded transform. (d) Denoised signal.

many of the transform values of the original signal which are needed for an accurate approximation. This can be verified by comparing the thresholded signal shown in Figure 1.7(c) with the original signal's transform in Figure 1.5(b). In Figure 1.7(d) we show the denoised signal obtained by inverse transforming the thresholded signal. This denoised signal is clearly an unsatisfactory approximation of the original signal. By computing RMS Errors, we can quantify this judgment. The RMS Error between Signal B and Signal 2 is .057, while the RMS Error between the denoised signal and Signal 2 is .035. This shows that the error after denoising is almost two-thirds as great as the original error.

Summarizing this second test case, we can say that the denoising was not effective because *the transform could not compress the energy of the original signal into a few high-energy values lying above the noise threshold.* We shall see in the next chapter that more sophisticated wavelet transforms can achieve the desired compression and will perform nearly as well at denoising Signal B as the Haar transform did for Signal A.

We have tried to emphasize the close connection between the degree of effectiveness of the threshold denoising method and the degree of effectiveness of the wavelet transform compression method. In the next chapter we shall describe how, with more powerful wavelet transforms, a very robust and nearly optimal method of noise removal can be realized.

1.7 Notes and references

More material on the Haar transform and its applications can be found in [RAO]. Besides the lossy compression method described in this chapter, the Haar transform has also played a role in lossless image compression; see [RAJ] or [HER].

For those readers interested in the history of wavelet analysis, a good place to start is the article by Burke [BUR], which has been expanded into a book [HUB]. There is also some history, of a more sophisticated mathematical nature, in the books by Meyer [ME2] and Daubechies [DAU].