Digital Signal Processing: A Computer Science Perspective
Jonathan Y. Stein

Print ISBN 0-471-29546-9 Online ISBN 0-471-20059-X

4

The Frequency Domain

The concept of frequency is clearest for simple sinusoids, but we saw in the previous chapter that it can be useful for nonsinusoidal periodic signals as well. The Fourier series is a useful tool for description of arbitrary periodic signals, describing them in terms of a spectrum of sinusoids, the frequencies of which are multiples of a basic frequency.

It is not immediately obvious that the concepts of spectrum and frequency can be generalized to *nonperiodic* signals. After all, *frequency* is only meaningful if something is periodic! Surprisingly, the concept of spectrum turns out to be quite robust; for nonperiodic signals we simply need a continuum of frequencies rather than harmonically related ones. Thus analog signals can be viewed either as continuous functions of time or as continuous functions of frequency. This leads to a pleasingly symmetric view, whereby the signal can be described in the *time domain* or the *frequency domain.*

The mathematical tool for transforming an analog signal from its time domain representation to the frequency domain, or vice versa, is called the Fourier transform (FT). The name hints at the fact that it is closely related to the Fourier series that we have already discussed. For digital signals we have close relatives, namely the discrete Fourier transform (DFT) and the z transform (zT). In this chapter we introduce all of these, review their properties, and compute them for a few example signals. We also introduce a non-Fourier concept of frequency, the *instantaneous frequency.* The FS, FT, DFT, zT, and instantaneous frequency, each in its own domain of applicability, is in some sense the *proper* definition of frequency.

4.1 From Fourier Series to Fourier Transform

In the previous chapter we learned that the set of harmonically related sinusoids or complex exponentials form a basis for the vector space of periodic signals. We now wish to extend this result to the vector space of all analog signals. The expansion in this basis is the *Fourier transform.*

Looking back at the steps in proving the existence of the Fourier series we see that the periodicity of the signals was not really crucial; in fact the whole periodicity constraint was quite a nuisance! The SUIs form a basis for *all* signals, whether periodic or not. It was only when we introduced the HRSs, sums of which are necessarily periodic, that we had to restrict ourselves to representing periodic signals. It would seem that had we allowed arbitrary frequency sinusoids we would have been able to represent any signal, and indeed this is the case. In fact it would have been just as easy for us to have directly derived the Fourier transform without the annoyance of the Fourier series; however this would have involved a grave break with mathematical tradition that mandates deriving the Fourier transform from the Fourier series.

The basic idea behind this latter derivation is inherent in the FS derived in Section 3.8. There we saw how increasing the period of the signal to be analyzed required decreasing the fundamental frequency of the HRSs. It is a general result that the longer the time duration that we must accurately reproduce, the more frequency resolution is required to do so. Now let us imagine the period going to infinity, so that the signal effectively is no longer periodic. If you find this infinity troublesome just imagine a period longer than the time during which you are willing to wait for the signal to repeat. The required frequency resolution will then become infinitesimal, and at every step of the way the corresponding HRSs form a basis for the signals with this large period. In the limit of aperiodic signals and continuous spectrum we discover that the set of *all* sinusoids forms a basis for the entire vector space of signals. Of course, for our basis signals we can choose to use sinusoids in quadrature $\sin(\omega t)$ and $\cos(\omega t)$, sinusoids with arbitrary phases $\sin(\omega t + \varphi)$, or complex exponentials $e^{i\omega t}$ with both positive and negative frequencies.

We have neglected an essential technical detail—as long as the fundamental frequency is small, but still finite, there are a denumerably infinite number of basis signals, and so the dimension of the space is $\aleph_0$ and expansions of arbitrary signals are infinite sums. Once the spectrum becomes continuous, there are a nondenumerable infinity of basis functions, and we must replace the infinite sums with integrals. The set of 'coefficients' S_k becomes a single continuous function of frequency $S(\omega)$.

The result is an expression for a signal as an integral over all time of a function of frequency times a complex exponential.

$$S(\omega) = \int_{t=-\infty}^{\infty} s(t)\, e^{-i\omega t}\, dt \tag{4.1}$$

This function of frequency is called the Fourier transform (FT). As we shall show, you may think of it as the *spectrum* of a nonperiodic signal. The extension of Fourier's theorem now states that every (not necessarily periodic) function (that obeys certain conditions) can be written as the integral over complex exponentials. The conditions for the convergence of the Fourier transform are almost the same as Dirichlet's conditions for the Fourier series; just remember to increase the region of integration to all times and insist on at most a finite number of extrema and discontinuities in any *finite amount of time.*

Paradoxically, while in normal speech *to transform* usually means to change the form of a quantity without changing its meaning, in mathematics *a transform* is a changing of meaning that does not alter the form. The Fourier transform changes the meaning from time to frequency domain, but the form remains a continuous function. The Fourier series is not a transform since it changes a continuous function into an infinite-dimensional vector of coefficients. We will see later that the discrete Fourier transform translates infinite-dimensional vectors into infinite-dimensional vectors. Specifically, *integral transforms*, like the FT, are representations of continuous functions as

$$F(\omega) = \int f(t)\, K(t, w)\, dt$$

where K is called the *kernel* of the transform.

When dealing with transforms we often use *operator notation*, i.e., we write $S(\omega) = \mathrm{FT}\left(s(t)\right)$ and

$$S(\omega) = \mathrm{FT}\left(s(t)\right) = \int_{t=-\infty}^{\infty} s(t)\, e^{-\mathrm{i}\omega t}\, dt \tag{4.2}$$

and think of FT as an operator that transforms the time domain representation of a signal into the frequency domain representation.

As was the case for periodic signals, the spectrum contains all possible information about the signal, and therefore the signal can be reconstructed from the spectrum alone. Consequently, we can define the inverse Fourier transform (iFT), $s(t) = \mathrm{FT}^{-1}\left(S(\omega)\right)$ where FT^{-1} is the *inverse operator.*

$$s(t) = \mathrm{FT}^{-1}\left(S(\omega)\right) = \frac{1}{2\pi} \int_{\omega=-\infty}^{\infty} S(\omega)\, e^{\mathrm{i}\omega t}\, d\omega \tag{4.3}$$

The form of the iFT is almost identical to that of the transform itself, but it integrates out the frequency variable leaving the time variable. The only differences are the normalization constant (more about that shortly) and the sign of the exponent.

The inverse operator obeys $\mathrm{FT}^{-1}\mathrm{FT} = 1$ where 1 is the identity operator that leaves every signal completely unchanged

$$s(t) = \mathrm{FT}^{-1}\left(S(\omega)\right) = \mathrm{FT}^{-1}\,\mathrm{FT}\,\left(s(t)\right) \tag{4.4}$$

an identity sometimes called the Fourier Integral Theorem. The two representations related by the FT and FT^{-1} operators are called a *Fourier transform pair*. They are both functions of a single continuous variable and contain exactly the same information about the signal, but in different forms. The function $s(t)$ is the *time domain* representation of the signal, while $S(\omega)$ is its *frequency domain* representation.

Let's prove equation (4.4).

$$\begin{aligned}
\mathrm{FT}^{-1}\,\mathrm{FT}\,\left(s(t)\right) &= \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty} S(\omega)\,e^{\mathrm{i}\omega t}\,d\omega \\
&= \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty}\left(\int_{t'=-\infty}^{\infty} s(t')e^{-\mathrm{i}\omega t'}\,dt'\right)e^{\mathrm{i}\omega t}\,d\omega \\
&= \int_{t'=-\infty}^{\infty} s(t')\frac{1}{2\pi}\int_{\omega=-\infty}^{\infty} e^{-\mathrm{i}\omega t'}e^{\mathrm{i}\omega t}\,d\omega\,dt' \\
&= \int_{t'=-\infty}^{\infty} s(t')\delta(t-t')\,dt' = s(t)
\end{aligned}$$

We now see why the exponents have different signs—it's required to get the needed delta function. Incidentally, we see that instead of the normalization constant $\frac{1}{2\pi}$ in the iFT we could have used any constants in both FT and iFT whose product is $\frac{1}{2\pi}$. For instance, physicists usually define a more symmetric pair

$$\begin{aligned}
S(\omega) = \mathrm{FT}\left(s(t)\right) &= \frac{1}{\sqrt{2\pi}}\int_{t=-\infty}^{\infty} s(t)\,e^{-\mathrm{i}\omega t}\,dt \\
s(t) = \mathrm{FT}^{-1}\left(S(\omega)\right) &= \frac{1}{\sqrt{2\pi}}\int_{\omega=-\infty}^{\infty} S(\omega)\,e^{\mathrm{i}\omega t}\,d\omega
\end{aligned} \tag{4.5}$$

but any other combination could be used as well. The DSP convention of putting the constant only in the definition of the inverse transform becomes more symmetric when using the frequency f in Hz (cycles per second) rather than the angular frequency ω in radians per second.

$$\begin{aligned}
S(f) &= \int_{t=-\infty}^{\infty} s(t)e^{-2\pi\mathrm{i}ft}dt \\
s(t) &= \int_{f=-\infty}^{\infty} S(f)e^{2\pi\mathrm{i}ft}df
\end{aligned} \tag{4.6}$$

We have shown that the FT indeed delivers a function of frequency that contains all the information in the signal itself. What we haven't shown is its relationship to the concept of *frequency spectrum* as we understand it. The true spectrum should be prominent at frequencies that are provably significant components of the signal, and should be zero at frequencies not corresponding to any physical aspect of the signal. We *could* show this directly, starting with a single sinusoid such as $s(t) = A\cos(\omega' t)$ and showing that it is

$$\begin{aligned} S(\omega) = \mathrm{FT}\left(A\cos(\omega' t)\right) &= \int_{t=-\infty}^{\infty} A\cos(\omega' t)e^{-\mathrm{i}\omega t}dt \\ &= \int_{t=-\infty}^{\infty} \frac{A}{2}\left(e^{\mathrm{i}\omega' t} + e^{-\mathrm{i}\omega' t}\right)e^{-\mathrm{i}\omega t}dt \\ &= \frac{A}{2}\left(\int_{t=-\infty}^{\infty} e^{\mathrm{i}\omega' t}e^{-\mathrm{i}\omega t}dt + \int_{t=-\infty}^{\infty} e^{-\mathrm{i}\omega' t}e^{-\mathrm{i}\omega t}dt\right) \\ &= \frac{A}{2}\left(\int_{t=-\infty}^{\infty} e^{-\mathrm{i}(\omega-\omega')t}dt + \int_{t=-\infty}^{\infty} e^{-\mathrm{i}(\omega+\omega')t}dt\right) \\ &= 2\pi\frac{A}{2}\left(\delta(\omega-\omega') + \delta(\omega+\omega')\right) \end{aligned}$$

and accordingly has only components at $\pm\omega'$ as expected. Then we would have to invoke the linearity of the FT and claim that for all combinations of sinusoids

$$\sum_{k=0}^{K} A_k \cos(\omega_k t)$$

the FT has discrete lines of precisely the expected relative weights. Next we would have to consider the continuous spectra of nonperiodic signals and show that the FT captures the meaning we anticipate. Finally, we would need to show that the FT is zero for unwanted frequencies. This could conceivably involve forcibly *notching out* frequencies from an arbitrary signal, and observing the FT at these frequencies to be zero.

This prescription is perhaps overly ambitious for us at this point, and in any case there is a shrewd way out. All we really need do is to show that the FT is the proper generalization of the FS for possibly nonperiodic signals. This will ensure that all well-known properties of FS spectra will survive in the FT, and all new properties of the FT will be taken to be the definition of what the true spectrum should be.

We start from slightly modified versions of equations (3.26) and (3.27) for the FS of a periodic signal $s(t)$ with period T

$$\begin{aligned} S_k &= \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} s(t)e^{-\mathrm{i}\frac{2\pi k}{T}t} \\ s(t) &= \sum_{k=-\infty}^{\infty} S_k e^{\mathrm{i}\frac{2\pi k}{T}t} \end{aligned}$$

and define $\omega \equiv \frac{2\pi k}{T}$. We can now think of the FS as S_ω instead of S_k; of course the indices are no longer integers, but there still are a denumerable number of them. They are uniformly spaced with $\Delta\omega = \frac{2\pi}{T}$ between them, and they still run from minus infinity to plus infinity.

$$\begin{aligned} S_\omega &= \frac{\Delta\omega}{2\pi}\int_{-\frac{T}{2}}^{\frac{T}{2}} s(t)e^{-\mathrm{i}\omega t}dt \\ s(t) &= \sum_{\omega=-\infty}^{\infty} S_\omega e^{\mathrm{i}\omega t} \end{aligned}$$

We next envision increasing the period T without limit $T \to \infty$. As we have already discussed, the frequency spacing Δ will become smaller and smaller $\Delta \to 0$, until the sequence $\{S_\omega\}_{\omega=-\infty}^{\infty}$ becomes a continuous function $S(\omega)$. Unfortunately, this definition of $S(\omega)$ is unsatisfactory. Looking back at the equation for S_ω we see that it is proportional to $\Delta\omega$. Assuming the integral approaches a finite value, S_ω will vanish as $\Delta\omega \to 0$. However, the ratio $\frac{S_\omega}{\Delta\omega}$ will remain finite in this limit, and has the pleasing interpretation of being the density of Fourier components per unit frequency.

We therefore propose defining $S(\omega) \equiv \frac{S_\omega}{\Delta\omega}$, in terms of which

$$\begin{aligned} S(\omega) &= \frac{1}{2\pi}\int_{-\frac{T}{2}}^{\frac{T}{2}} s(t)e^{-\mathrm{i}\omega t}dt \\ s(t) &= \sum_{\omega=-\infty}^{\infty} S(\omega)e^{\mathrm{i}\omega t}\Delta\omega \end{aligned}$$

In the limit $T \to \infty$ and $\Delta\omega \to 0$ several things happen. The integral over t now runs from $-\infty$ to $+\infty$. The finite difference $\Delta\omega$ becomes the infinitesimal $d\omega$. The sum over the discrete ω index in the formula for $s(t)$ will of course become an integral over the continuous ω variable. Substitution of these brings us to

$$\begin{aligned} S(\omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} s(t)e^{-\mathrm{i}\omega t}dt \\ s(t) &= \int_{-\infty}^{\infty} S(\omega)e^{\mathrm{i}\omega t}d\omega \end{aligned}$$

which is the FT in an unusual but legitimate normalization scheme. Of course had we defined $S(\omega) \equiv 2\pi \frac{S_\omega}{\Delta\omega}$ we would have obtained exactly (4.2) and (4.3), and $S(\omega) \equiv \sqrt{2\pi} \frac{S_\omega}{\Delta\omega}$ would have produced the physicist's (4.5).

In Section 3.4 we interpreted the FS as the expansion of a periodic signal in the basis of sines and cosines. We have just derived the FT by a limiting process starting from the FS, so it is not surprising that we can interpret the FT as the expansion a nonperiodic signal in a basis. Due to the nondenumerably infinite amount of information in a general nonperiodic signal, it is not surprising that we need a nondenumerable number of basis functions, and that the sum in the expansion becomes an integral.

Reiterating what we have accomplished, we have shown that the FT as we have defined it is the natural generalization to nonperiodic signals of Fourier's expansion of periodic signals into sinusoids. The function $S(\omega)$ has a meaningful interpretation as the Fourier spectral density, so that $S(\omega)d\omega$ is the proper extension of the FS component. The FT is therefore seen to truly be the best definition of spectrum (so far).

EXERCISES

4.1.1 Prove the opposite direction of (4.4), namely

$$S(\omega) = \mathrm{FT}\,\mathrm{FT}^{-1}\left(S(\omega)\right)$$

4.1.2 Find the FT of $A\sin(\omega' t)$. How is it different from that of $A\cos(\omega' T)$?

4.1.3 Find the FT of the rectangular wave of Section 3.8. How does it relate to the FS found there? Find the FT of a single rectangle. How does it relate to that of the first part?

4.1.4 Write a routine that computes the value of the FT of a real signal $s(t)$ at frequency $f = \frac{\omega}{2\pi}$. The signal is nonzero only between times $t = 0$ and $t = T$, and is assumed to be reasonably well behaved. You should use numerical Riemann integration with the time resolution Δt variable.

4.1.5 Generate a signal composed of a constant plus a small number of unrelated sinusoids. Using the routines developed in the previous exercise, plot the real and imaginary parts of its FT for a frequency band containing all frequencies of interest. Vary the time resolution. How is the accuracy affected? Vary the frequency resolution. Are the frequencies of the sinusoids exact or is there some width to the lines? Is this width influenced by the time resolution? How much time is needed to compute the entire FT (as a function of time and frequency resolution)?

4.2 Fourier Transform Examples

The time has come to tackle a few examples of FT calculation. Although it *is* instructive to go through the mechanics of integration a few times, that is not our only motivation. We have selected examples that will be truly useful in our later studies.

The simplest signal to try is a constant signal $s(t) = 1$, and for this signal we almost know the answer as well! There can only be a DC (zero frequency) component, but how much DC is there? The integral in (4.2) is

$$S(\omega) = \int_{t=-\infty}^{\infty} e^{-\mathrm{i}\omega t} dt = \int_{t=-\infty}^{\infty} (\cos \omega t - \mathrm{i} \sin \omega t)\, dt \tag{4.7}$$

(we simply replaced $s(t)$ by 1). Now we are stuck, since the required definite integrals don't appear in any table of integrals. We can't do the indefinite integral and substitute the values at the endpoints, since $\sin(\pm\infty)$ and $\cos(\pm\infty)$ don't approach a constant value; and don't confuse this integral with equation (3.21) for $m = 1$, since the integral is over the entire t axis. Whenever we're stuck like this, it is best to think about what the integral means. When $\omega = 0$ we are trying to integrate unity over the entire t axis, which obviously diverges. For all other ω we are integrating sinusoids over all time. Over full periods sinusoids are positive just as much as they are negative, and assuming infinity can be considered to be a whole number of periods, the integral should be zero. We have thus deduced a delta function to within a constant $S(\omega) = \gamma\delta(\omega)$. To find γ we need to integrate over ω. We know from (4.4) that

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) dw = \mathrm{FT}^{-1}\Big(S(\omega)\Big) = s(t) = 1$$

from which we conclude that $\gamma = 2\pi$.

Let's try the other way around. What is the transform of an analog impulse $s(t) = \delta(t)$? Well it's just

$$\mathrm{FT}\Big(s(t)\Big) = \int_{t=-\infty}^{\infty} \delta(t) e^{-\mathrm{i}\omega t} dt = e^0 = 1$$

using property (A.69) of the delta function. So it works the other way as well—the transform of a delta is a constant. With only minimal additional effort we can find the transform of an impulse at any nonzero time τ. In this case we pick out the exponential at some other time

$$\mathrm{FT}\Big(\delta(t-\tau)\Big) = \int_{t=-\infty}^{\infty} \delta(t-\tau) e^{-\mathrm{i}\omega t} dt = e^{-\mathrm{i}\omega\tau} \tag{4.8}$$

which is a complex exponential in the frequency domain. The interpretation of this sinusoid is slightly different from the usual. Remember that here τ is a constant that we are given and ω is the variable. The sinusoidal behavior is as a function of frequency, and the higher τ is, the more compressed the oscillation becomes. So τ plays the role of frequency here, which is not surprising due to the dual nature of time and frequency.

Conversely, a non-DC complex exponential $s(t) = e^{\mathrm{i}\Omega t}$ has the transform

$$\mathrm{FT}\Big(s(t)\Big) = \int_{t=-\infty}^{\infty} e^{\mathrm{i}\Omega t} e^{-\mathrm{i}\omega t} dt = \int_{t=-\infty}^{\infty} e^{\mathrm{i}(\Omega-\omega)t} dt = 2\pi\delta(\omega - \Omega) \qquad (4.9)$$

(we could interchange the omegas since the delta function is symmetric). Thus the complex exponential corresponds to a single frequency line, as expected.

What about a real sinusoid $\sin(\Omega t)$ or $\cos(\Omega t)$? Using the linearity of the FT and the expressions (A.8) we can immediately conclude that sine and cosine consist of two delta functions in the frequency domain. One delta is at $+\Omega$ and the other at $-\Omega$.

$$\begin{aligned} \mathrm{FT}\Big(\sin(\omega t)\Big) &= \frac{\pi}{\mathrm{i}}\Big(\delta(\omega - \Omega) - \delta(\omega + \Omega)\Big) \\ \mathrm{FT}\Big(\cos(\omega t)\Big) &= \pi\Big(\delta(\omega - \Omega) + \delta(\omega + \Omega)\Big) \end{aligned} \qquad (4.10)$$

The absolute value of the spectrum is symmetric, as it must be for real functions, but sine and cosine differ in the relative phase of the deltas.

The FT decaying exponential can also be useful to know. It is simply

$$\mathrm{FT}\Big(e^{-\lambda t}u(t)\Big) = \frac{1}{\lambda + \mathrm{i}\omega} \qquad (4.11)$$

and actually the same transform holds for complex λ, as long as the real part of λ is positive.

Up to now we have treated rather smooth signals and impossibly singular ones (the delta). We will also need to investigate archetypical jump discontinuities, the sgn and step functions. Since sgn is odd, $\mathrm{sgn}(-t) = -\mathrm{sgn}(t)$, we can immediately deduce that the zero frequency component of sgn's FT must be zero. The zero frequency component of $u(t)$ is obviously infinite and so we know that $u(\omega)$ must have a $k\delta(\omega)$ component. The value of k can be determined from the fact that $u(-t) + u(t) = 1$ and from linearity $\mathrm{FT}(u(-t)) + \mathrm{FT}(u(t)) = \mathrm{FT}(1) = 2\pi\delta(\omega)$; so the DC component is simply $\pi\delta(\omega)$.

Trying to find the nonzero frequency components of either sgn or $u(t)$ we stumble upon one of those impossible integrals, like (4.7). For large ω it

should go to zero since the integral over an even number of cycles of sinusoids is zero; but for smaller ω there is the issue of the end effects. We will be able to prove later that the spectrum decays as $\frac{1}{w}$, i.e., every time we double the frequency the amplitude drops to half its previous value. When displaying the spectrum on a logarithmic scale this translates to a linear drop of 6 dB per octave. Since any signal with a single discontinuity can be considered to be continuous signal plus a step or sgn, all signals with step discontinuities have this 6 dB per octave drop in their spectra.

EXERCISES

4.2.1 Calculate the FT of a complex exponential from those of sin and cos using linearity and equation (A.8).

4.2.2 What is the difference between the FT of sin and cos? Explain the effect of A and φ on the FT of $A\sin(\omega t + \varphi)$.

4.2.3 Find the FT of the single rectangle (equation (3.29)).

4.2.4 Show that $\sum_{n=-\infty}^{\infty} e^{-i\omega nT} = 0$ when ω is not a multiple of $\frac{2\pi}{T}$.

4.2.5 Formally prove that the FT of the impulse train $s(t) = \sum \delta(t - kT)$ is an impulse train in the frequency domain by finding its Fourier *series* and relating the transform to the series.

4.2.6 Our proof of the universality of the Fourier series in Section 3.4 rested on the expansion of shifted delta functions in the basic period in terms of harmonically related sinusoids. Show how this can be simplified using our results for impulse trains.

4.2.7 Prove that the following are FT pairs:

$u(t)$	$\pi\delta(\omega) + \frac{1}{i\omega}$
$e^{-\lambda t}u(t)$	$\frac{1}{\lambda+i\omega}$
$te^{-\lambda t}u(t)$	$\frac{1}{(\lambda+i\omega)^2}$
$\alpha^{\lvert n\rvert}$	$\frac{1-\alpha^2}{1-2\alpha\cos(\omega)+\alpha^2}$
$\lvert t\rvert$	$-\frac{2}{\omega^2}$
$e^{-a\lvert t\rvert}$	$\frac{2a}{a^2+\omega^2}$

4.3 FT Properties

As we saw in the examples, the Fourier transform of a signal may look like just about anything. It is customary to differentiate between continuous and discrete line spectra. When the FT is a continuous smooth function of frequency a nondenumerable number of frequency components are required to reproduce the signal. A FT composed of some number of sharp discrete lines results from a signal that is the sum of that number of sinusoids. In general, spectra may have both continuous and discrete parts. In fact all signals encountered in practice are noisy and so cannot be precisely periodic, and hence some continuous spectrum contribution is always present.

The question of the 'mathematical existence' of the FT is an important one for mathematicians, but one we will not cover extensively. The Dirichlet conditions for the FT require that the integral over all time of the absolute value of the signal be finite, as well as there being only a finite number of extrema and discontinuities in any finite interval. The FT obviously does not exist in the technical sense for periodic signals such as sinusoids, but by allowing delta functions we bypass this problem.

Although we will not dwell on existence, there are many other characteristics of the FT that we will need. Many times we can find the FT of signals without actually integrating, by exploiting known transforms and some of the following characteristics. These characteristics are often closely related to characteristics of the FS, and so we need not derive them in detail.

First, it is important to restate the Fourier Integral Theorem that the inverse FT given by equation (4.3) is indeed the inverse operation.

$$\mathrm{FT}^{-1}\,\mathrm{FT}\,x = x \qquad\qquad \mathrm{FT}\,\mathrm{FT}^{-1}\,X = X \tag{4.12}$$

Next, the FT is linear, i.e.,

$$\begin{aligned} \mathrm{FT}\Big(x(t) + y(t)\Big) &= X(\omega) + Y(\omega) \\ \mathrm{FT}\Big(as(t)\Big) &= aS(\omega) \end{aligned} \tag{4.13}$$

a property already used in our derivation of the FT of real sinusoids.

Speaking of real signals, it is easy to see that the FT of a real signal is Hermitian even,

$$S(-\omega) = S^*(\omega)$$

meaning that $\Re S(\omega)$ is even, $\Im S(\omega)$ is odd, $|S(\omega)|$ is even, and $\angle S(\omega)$ is odd. Conversely the FT of an even signal ($s(-t) = s(t)$) is real, and that of an odd signal is pure imaginary.

There are two properties that deal with changing the clock, namely the time shifting property

$$\text{FT}\Big(s(t-\tau)\Big) = e^{-\mathrm{i}\omega\tau}S(\omega) \tag{4.14}$$

and the time scaling property.

$$\text{FT}\Big(s(ct)\Big) = \frac{1}{|c|}S(\frac{\omega}{c}) \tag{4.15}$$

Conversely, there is a property that deals with shifting the frequency axis

$$\text{FT}\Big(s(t)e^{\mathrm{i}\Omega t}\Big) = S(\omega-\Omega) \tag{4.16}$$

an operation we usually call *mixing*.

What happens when you differentiate $s(t) = e^{\mathrm{i}\omega t}$? You get $\mathrm{i}\omega s(t)$. Similarly, integrating it you get $\frac{1}{\mathrm{i}\omega}s(t)$. It follows that differentiating or integrating an arbitrary signal affects the FT in a simple way.

$$\text{FT}\left(\frac{ds(t)}{dt}\right) = \mathrm{i}\omega S(\omega) \qquad \Big| \qquad \text{FT}\left(\int_{\tau=-\infty}^{t} s(\tau)d\tau\right) = \frac{1}{\mathrm{i}\omega}S(\omega) \tag{4.17}$$

These are surprising results; we think of differentiation and integration as purely time domain operations, but they are even simpler in the frequency domain! We will see in Section 7.3 that the DSP approach to differentiation and integration in the time domain involves first designing a filter in the frequency domain. Note also that differentiation emphasizes high frequencies, while integration emphasizes lows. This is because derivatives involve subtracting nearby values, while integrals are basically averaging operators.

Linearity told us how to find the spectrum when adding signals; what happens when we multiply them? Since we have never tried this before we will have to actually do the integral.

$$\begin{aligned}
\int_{-\infty}^{\infty} x(t)y(t)e^{-\mathrm{i}\omega t}\,dt &= \int_{t=-\infty}^{\infty}\left(\frac{1}{2\pi}\int_{\Omega=-\infty}^{\infty} X(\Omega)e^{\mathrm{i}\Omega t}\,d\Omega\right)y(t)e^{-\mathrm{i}\omega t}\,dt \\
&= \frac{1}{2\pi}\int_{\Omega=-\infty}^{\infty} X(\Omega)\int_{t=-\infty}^{\infty} y(t)e^{-\mathrm{i}(\omega-\Omega)t}\,dt\,d\Omega \\
&= \frac{1}{2\pi}\int_{\Omega=-\infty}^{\infty} X(\Omega)Y(\omega-\Omega)\,d\Omega
\end{aligned}$$

What we did was simply to replace $x(t)$ by its iFT, change the order of integration, and recognize the FT of y. So we have found the following:

$$\text{FT}\Big(x(t)y(t)\Big) = \frac{1}{2\pi}\int_{\Omega=-\infty}^{\infty} X(\Omega)Y(\omega-\Omega)\,d\Omega \equiv X * Y \tag{4.18}$$

Now that we have the answer, what does it mean? The FT of the product of two signals in the time domain is the integral of a strange-looking product in the frequency domain. We hide this strangeness by using the symbol $*$, implying a product of some sort. It's a truly unusual product since the integration variable in Y runs in the opposite direction to that of the X variable. If that is not bad enough, repeating the above computation for iFT of a product in the frequency domain, we find

$$\mathrm{FT}^{-1}\left(X(\omega)Y(\omega)\right) = \int_{T=-\infty}^{\infty} x(T)y(t-T)\,dT \equiv x * y \tag{4.19}$$

where the integration variable in y runs backward in time! We are not yet ready to digest this strange expression that goes under the even stranger name of *convolution*, but it will turn out to be of the utmost importance later on.

A particular case of equation (4.18) is the DC ($\omega = 0$) term

$$\int_{-\infty}^{\infty} x(t)y(t)\;dt = \frac{1}{2\pi}\int_{\Omega=-\infty}^{\infty} X(\Omega)Y(-\Omega)\;d\Omega$$

and by taking $x(t) = s(t)$, $y(t) = s^*(t)$ and changing the name of the integration variable, we get Parseval's relation for the FT.

$$\int_{-\infty}^{\infty} |s(t)|^2\,dt = \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty} |S(\omega)|^2\,d\omega = \int_{f=-\infty}^{\infty} |S(f)|^2\,df \tag{4.20}$$

Parseval's relation tells us that the signal's energy is the same whether we look at it in the time domain or the frequency domain. This is an important physical consistency check.

To demonstrate the usefulness of some of these properties, we will now use the integration rule to derive a result regarding signals with discontinuous derivatives. We know that the FT of the impulse is constant, and that its integral is the unit step $u(t)$. Thus we would expect from (4.17) for the FT of the step to be simply $\frac{1}{\mathrm{i}\omega}$, which is *not* what we previously found! The reason is that (4.17) breaks down at $\omega = 0$, and so we always have to allow for the possible inclusion of a delta function. Integrating once more we get $f(t) = t\,u(t)$, which is continuous but has a discontinuous first derivative. The integration rule tells us that the FT of this f is $-\omega^{-2}$ (except at $\omega = 0$). Integrating yet another time gives us a signal with continuous first derivative but discontinuous second derivative and $\mathrm{i}\omega^{-3}$ behavior. Continuing this way we see that if all derivatives up to order k are continuous but the $(k+1)^{\text{th}}$ is not, then the (nonzero frequency) transform is proportional

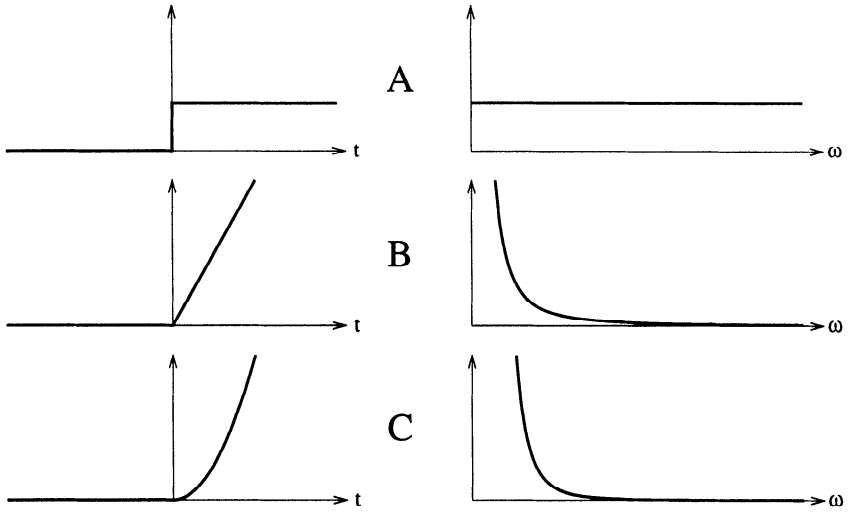

Figure 4.1: The effect of derivative discontinuity on the FT. In (A) the signal itself (the zeroth derivative) is discontinuous and the spectrum is constant. In (B) the first derivative is discontinuous and the spectrum decays as ω^{-2}. In (C) the second derivative jumps and the spectrum decays as ω^{-4}.

to ω^{-k}, and the power spectrum is inversely proportional to ω^{2k}. In other words a discontinuous first derivative contributes a term which decays 6 dB per octave; a second derivative 12 dB per octave, etc. These results, depicted in Figure 4.1, will be useful in Section 13.4.

EXERCISES

4.3.1 Explain why $\int_{-\infty}^{\infty} e^{i\omega t} dt = 2\pi\delta(\omega)$ using a graphical argument.

4.3.2 Show that time reversal causes frequency reversal FT $\big(s(-t)\big) = S(-\omega)$.

4.3.3 Show how differentiation and integration of the spectrum are reflected back to the time domain.

4.3.4 The derivative of $\cos(\omega t)$ is $-\omega \sin(\omega t)$. State this fact from the frequency domain point of view.

4.3.5 Show that we can interchange X and Y in the convolution integral.

4.3.6 Redraw the right-hand side of Figure 4.1 using dB. How does the slope relate to the order of the discontinuity?

4.3.7 Generalize the relationship between spectral slope and discontinuity order to signals with arbitrary size discontinuities not necessarily at the origin. What if there are many discontinuities?

4.4 The Uncertainty Theorem

Another signal with discontinuities is the rectangular window

$$s(t) = \begin{cases} 1 & |t| \leq T \\ 0 & \text{else} \end{cases} \tag{4.21}$$

which is like a single cycle of the rectangular wave. The term 'window' is meant to evoke the picture of the opening a window for a short time. Its FT

$$\begin{aligned} \mathrm{FT}\left(s(t)\right) &= \int_{-T}^{T} e^{i\omega t}\, dt \\ &= \frac{e^{+\mathrm{i}\omega T} - e^{-\mathrm{i}\omega T}}{i\omega} \\ &= 2\frac{\sin(\omega T)}{\omega} = 2T\,\mathrm{sinc}(\omega T) \end{aligned}$$

turns out to be a sinc. Now the interesting thing about this sinc is that its bandwidth is inversely proportional to T, as can be seen in Figure 4.2.

The wider the signal is in the time domain, the narrower it is in frequency, and vice versa. In fact if we define the bandwidth to be precisely between the first zeros of the sinc, $\Delta\omega = \frac{2\pi}{T}$, and relate this to the time duration $\Delta t = 2T$, we find that the *uncertainty product*

$$\Delta\omega\,\Delta t = 4\pi$$

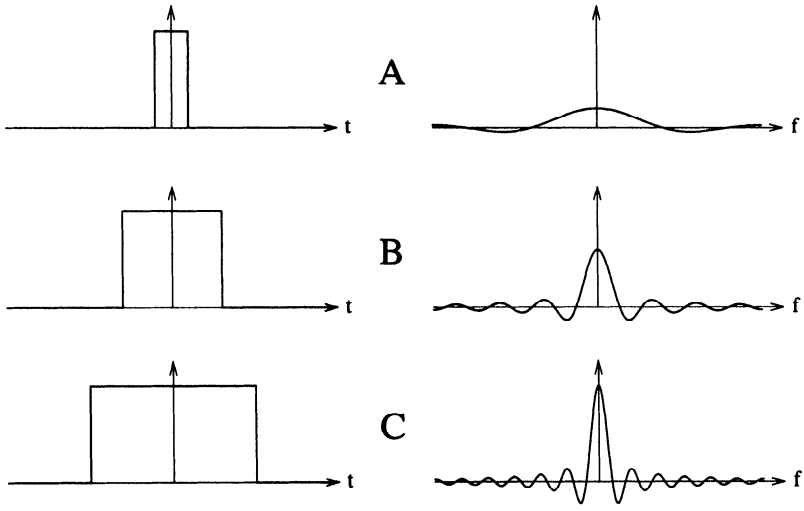

Figure 4.2: Rectangular windows of various widths with their Fourier transforms. Note that the signal energy is not normalized.

although different definitions of bandwidth would change the precise value on the right-hand side.

This is a special case of a more general rule relating time durations to bandwidth. A single sinusoid is defined for all time and has a completely precise line as its spectrum. Signals of finite duration cannot have discrete line spectra since in order build the signal where it is nonzero but cancel it out at $t = \pm\infty$ we need to sum many nearby frequencies. The shorter the time duration the more frequencies we need and so the wider the bandwidth.

It is useful to think of this in a slightly different way. Only if we can observe a sinusoid for an infinite amount of time can we precisely determine its frequency. If we are allowed to see it for a limited time duration we can only determine the frequency to within some tolerance; for all sinusoids with similar frequencies look about the same over this limited time. The less time we are allowed to view the sinusoid, the greater our uncertainty regarding its true frequency. You can convince yourself of this fact by carefully studying Figure 4.3.

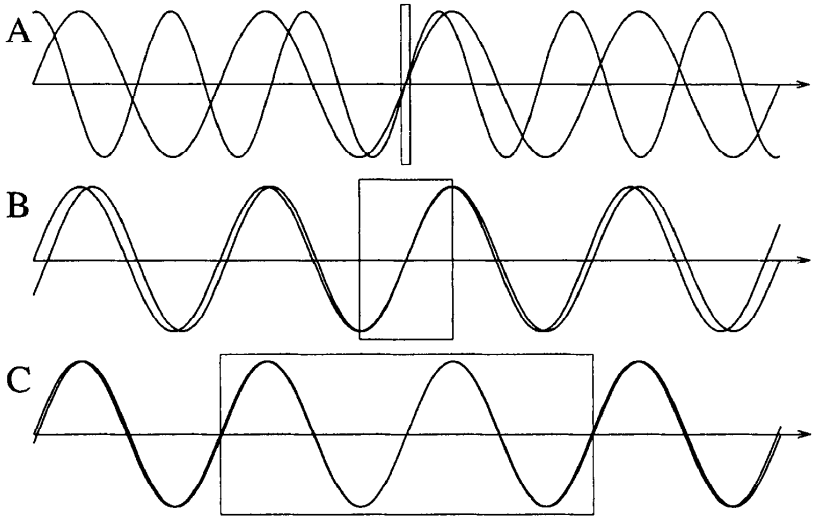

Figure 4.3: The effect of observation window duration on frequency uncertainty. In (A) we only observe the sinusoid for an extremely short time, and hence we can not accurately gauge its frequency. In (B) we observe about half a cycle and can now estimate the frequency, but with relatively large uncertainty. In (C) two full cycles are observed and consequently the uncertainty is much reduced.

As our next example, consider the Gaussian

$$s(t) = Ae^{-\beta t^2} \tag{4.22}$$

whose Fourier transform is

$$S(\omega) = \int_{-\infty}^{\infty} Ae^{-\beta t^2} e^{-\mathrm{i}\omega t} dt = \int_{-\infty}^{\infty} Ae^{-\beta t^2 - \mathrm{i}\omega t} dt \tag{4.23}$$

which doesn't look hopeful. The mathematical trick to use here is 'completing the square'. The exponent is $-(\beta t^2 + i\omega t)$. We can add and subtract $\frac{\omega^2}{4\beta}$ so that

$$S(\omega) = \int_{-\infty}^{\infty} Ae^{-(\sqrt{\beta}t + \frac{\mathrm{i}\omega}{2\sqrt{\beta}})^2} e^{-\frac{\mathrm{i}\omega}{4\beta}} dt = Ae^{-\frac{\mathrm{i}\omega}{4\beta}} \int_{-\infty}^{\infty} e^{-(\sqrt{\beta}t + \frac{\mathrm{i}\omega}{2\sqrt{\beta}})^2} dt \tag{4.24}$$

and a change of variable $u = \sqrt{\beta}t + \frac{\mathrm{i}\omega}{2\sqrt{\beta}}$ gives

$$S(\omega) = Ae^{-\frac{\mathrm{i}\omega}{4\beta}} \int_{-\infty}^{\infty} e^{-u^2} \frac{du}{\sqrt{\beta}} = A\sqrt{\frac{\pi}{\beta}} e^{-\frac{\omega^2}{4\beta}} \tag{4.25}$$

so the FT of a Gaussian is another Gaussian.

Now let's look at the uncertainty product for this case. The best way of defining Δt here is as the variance of the square of the signal. Why the square? Well, if the signal took on negative values it would be more obvious, but even for the Gaussian the energy is the integral of the square of the signal; the 'center of gravity' is the expected value of the integral of t times the signal squared, etc. Comparing the square of the signal $A^2 e^{-2\beta t^2}$ with equation (A.19) we see that the standard deviation in the time domain is $\Delta t = \frac{1}{2\sqrt{\beta}}$, while the same considerations for equation (4.25) lead us to realize that $\Delta\omega = \sqrt{\beta}$. The uncertainty product follows.

$$\Delta t \, \Delta\omega = \tfrac{1}{2}$$

Now it turns out that no signal has a smaller uncertainty product than this. This theorem is called the *uncertainty theorem*, and it is of importance both in DSP and in quantum physics (where it was first enunciated by Heisenberg). Quantum physics teaches us that the momentum of a particle is the Fourier transform of its position, and hence the uncertainty theorem limits how accurately one can simultaneously measure its position and velocity. Energy and time are similarly related and hence extremely accurate energy measurements necessarily take a long time.

The Uncertainty Theorem
Given any signal $s(t)$ with energy

$$E = \int_{-\infty}^{\infty} s^2(t)dt$$

time center-of-gravity

$$\langle t \rangle \equiv \frac{\int_{-\infty}^{\infty} ts^2(t)dt}{E}$$

squared time uncertainty

$$(\Delta t)^2 \equiv \frac{\int_{-\infty}^{\infty} (t - \langle t \rangle)^2 s^2(t)dt}{E}$$

frequency center-of-gravity

$$\langle \omega \rangle \equiv \frac{\int_{-\infty}^{\infty} \omega S^2(\omega)d\omega}{E}$$

and squared frequency uncertainty

$$(\Delta \omega)^2 \equiv \frac{\int_{-\infty}^{\infty} (\omega - \langle \omega \rangle)^2 S^2(\omega)d\omega}{E}$$

then the uncertainty product

$$\Delta t \, \Delta \omega \geq \tfrac{1}{2}$$

is always greater than one half. ∎

Although this theorem tells us that mathematics places fundamental limitations on how accurately we are allowed to measure things, there is nothing particularly mystifying about it. It simply says that the longer you are allowed to observe a signal the better you can estimate its frequencies.

Next let's consider the train of Dirac delta functions

$$s(t) = \sum_{n=-\infty}^{\infty} \delta(t - n\tau) \tag{4.26}$$

depicted in Figure 4.4. This signal is truly fundamental to all of DSP, since it is the link between analog signals and their digital representations. We can think of sampling as multiplication of the analog signal by just such a train of impulses,

$$S(\omega) = \int_{t=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(t - nT)e^{-i\omega t}dt$$

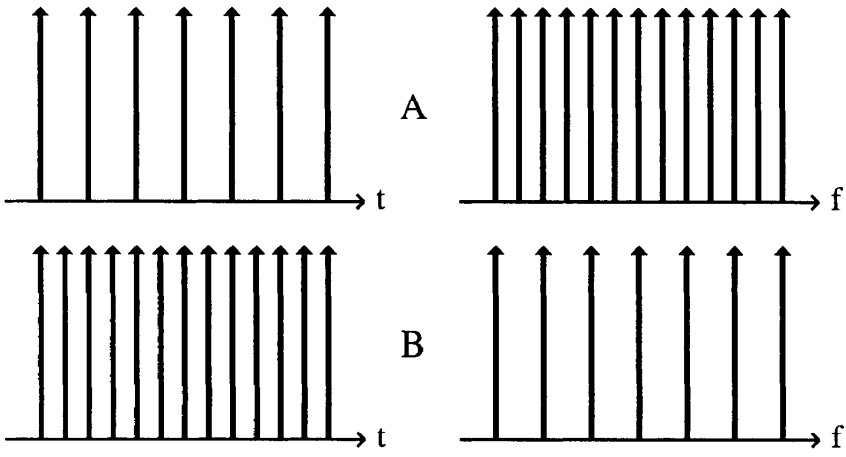

Figure 4.4: Trains of Dirac delta functions in time and frequency domains. Note that the spacing in the time domain is the inverse of that in the frequency domain.

Interchanging the order of summation and integration (we ask permission of the more mathematically sophisticated reader before doing this), we find a sum over the FT of equation (4.8) with $\tau = nT$

$$S(\omega) = \sum_{n=-\infty}^{\infty} \int_{t=-\infty}^{\infty} \delta(t - nT) e^{-i\omega t} dt = \sum_{n=-\infty}^{\infty} e^{-i\omega nT}$$

and once again we are stuck. Looking carefully at the sum we become convinced that for most ω the infinite sum should contain just as many negative contributions as positive ones. These then cancel out leaving zero. At $\omega = 0$, however, we have an infinite sum of ones, which is infinite. Does this mean that the FT of a train of deltas is a single Dirac delta? No, because the same thing happens for all ω of the form $\frac{2\pi}{T}$ as well! So similarly to the Gaussian, a train of impulses has an FT of the same form as itself, a train of impulses in the frequency domain; and when the deltas are close together in the time domain, they are far apart in the frequency domain, and vice versa. The product of the spacings obeys

$$\Delta t \, \Delta\omega \; = \; 2\pi$$

once again a kind of uncertainty relation.

EXERCISES

4.4.1 Prove the *Schwartz inequality* for signals.

$$\left(\int_{-\infty}^{\infty} x^2(t)dt\right)\left(\int_{-\infty}^{\infty} y^2(t)dt\right) \geq \left|\int_{-\infty}^{\infty} x(t)y(t)dt\right|^2$$

4.4.2 Using Parseval's relation and the FT of a derivative prove the following relation involving the uncertainties and the energy E.

$$(\Delta t\,\Delta\omega)^2 = \frac{\int(t - \langle t\rangle)^2 s^2(t)dt \int \left(\frac{ds}{dt}\right)^2 dt}{E^4}$$

4.4.3 Using the Schwartz inequality, the above relation, and integration by parts, prove the uncertainty theorem.

4.5 Power Spectrum

The energy E of a signal $s(t)$ is defined as the integral over all times of the squared values in the time domain. Due to this additive form, we can interpret the integral over some interval of time as the signal's energy during that time. Making the interval smaller and smaller we obtain the power $E(t)$; the signal's energy during a time interval of infinitesimal duration dt centered on time t is $E(t)dt$ where $E(t) = |s(t)|^2$. You can think of the power as the *energy time density*, using the term 'density' as explained at the end of Appendix A.9.

Integrating the power over any finite time interval brings us back to the signal's energy during that time; integrating over all time retrieves the total energy.

$$E = \int_{-\infty}^{\infty} E(t)\,dt = \int_{-\infty}^{\infty} |s(t)|^2\,dt$$

From Parseval's relation we know that the energy is also computable as the integral of squared values in the frequency domain (except possibly for a normalization factor depending on the FT definition chosen). Hence repeating the above arguments we can define the *energy spectral density* $E(f) = |S(f)|^2$, that specifies how the signal's energy is distributed over frequency. The meaning of $E(f)$ is similar to that of the power; the energy contained in the signal components in an interval of bandwidth df centered on frequency f is $E(f)\,df$.

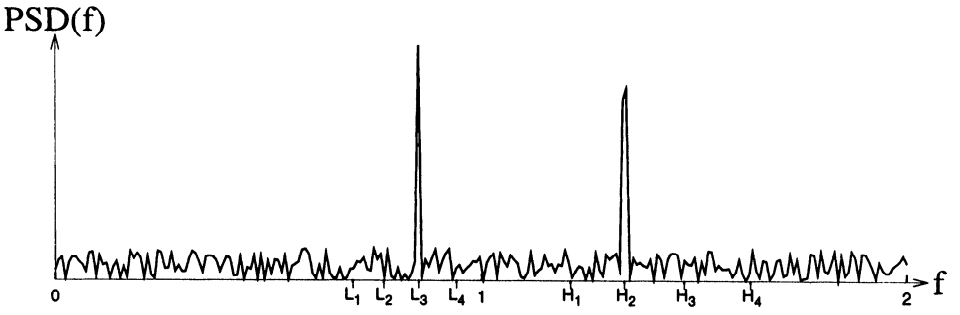

Figure 4.5: Power spectral density for the DTMF digit '8'. The horizontal axis is the frequency in KHz and the vertical axis is a linear measure of the energy density. The eight possible frequencies are marked for convenience.

In the next section we will see that many signals have spectral distributions that vary as time progresses. For such signals we wish to know how much energy is in the frequency range around f at times around t. Since the energy density in the time domain is the power, the desired quantity is called the **P**ower **S**pectral **D**ensity (PSD). PSDs that change in time are so common that we almost always use the term *power spectrum* instead of energy spectrum.

Writing the full FT as a magnitude times an angle $S(f) = A(f)e^{i\phi(f)}$, we see that the PSD contains only the magnitude information, all the angle information having been discarded. At this stage of our studies it may not yet be entirely clear why we need the full frequency domain representation, but it is easy to grasp why we would want to know how a signal's energy is divided among the component frequencies. For example, push-button dialing of a phone uses DTMF signals where two tones are transmitted at a time (see Figure 4.5). The lower tone of the two is selected from four candidate frequencies L_1, L_2, L_3, L_4, and the higher is one of H_1, H_2, H_3, H_4. In order to know that an eight was pressed we need only ascertain that there is energy in the vicinities of L_3 and H_2. The phases are completely irrelevant.

As a more complex application, consider a phone line on which several signals coexist. In order for these signals not to interfere with each other they are restricted by 'masks', i.e., specifications of the maximal amount of power they may contain at any given frequency. The masks in Figure 4.6 are specified in dBm/Hz, where dBm is the power in dB relative to a 1 milliwatt signal (see equation (A.16)). The horizontal scale has also been drawn logarithmically in order to accommodate the large range of frequencies from 100 Hz to over 10 MHz. Although the higher frequency signals seem to be

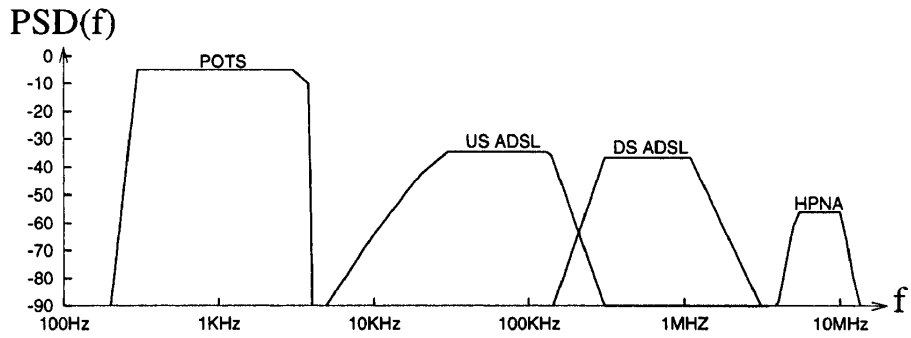

Figure 4.6: PSD masks for several signals on a phone line. The horizontal axis is the frequency in KHz on a logarithmic scale and the vertical axis is the maximum allowed PSD in dBm per Hz. The leftmost signal is the POTS (Plain Old Telephone System) mask, including voice and voicegrade modems. The middle mask is for ADSL, with the lower portion for the 512 Kb/s upstream signal and the upper for the 6 Mb/s downstream signal. At the far right is the mask for the 1 Mb/s Home Phone Network signal.

lower in power, this is only an illusion; the PSD is lower but the bandwidths are much greater.

The mask containing the lowest frequencies is for regular telephone conversations, affectionately called **P**lain **O**ld **T**elephone **S**ervice (POTS). This mask, extending from 200 Hz to about 3.8 KHz, holds for voice signals, signals from fax machines, and voicegrade modems up to 33.6 Kb/s.

The need for high-speed digital communications has led to innovative uses of standard phone lines. The **A**symmetric **D**igital **S**ubscriber **L**ine (ADSL) modem is one such invention. It can deliver a high-speed downstream (from the service provider to the customer) connection of up to 8 Mb/s, and a medium-speed upstream (from the customer to the provider) connection of 640 Kb/s. ADSL was designed in order not to interfere with the POTS signal, so that the standard use of the telephone could continued unaffected. By placing the ADSL signal at higher frequencies, and restricting the amount of power emitted at POTS frequencies, interference is avoided. This restriction may be verified using the power spectrum; the signal phases are irrelevant.

In the same way, after the definition of ADSL the need arose for networking computers and peripherals inside a residence. Of course this can be done by running cables for this purpose, but this may be avoided by using the internal phone wiring but requiring the new 'home phone network' signal to lie strictly above the POTS *and* ADSL signals.

We see that based on the power spectrum alone we may deduce whether signals may coexist without mutual interference. The principle behind this

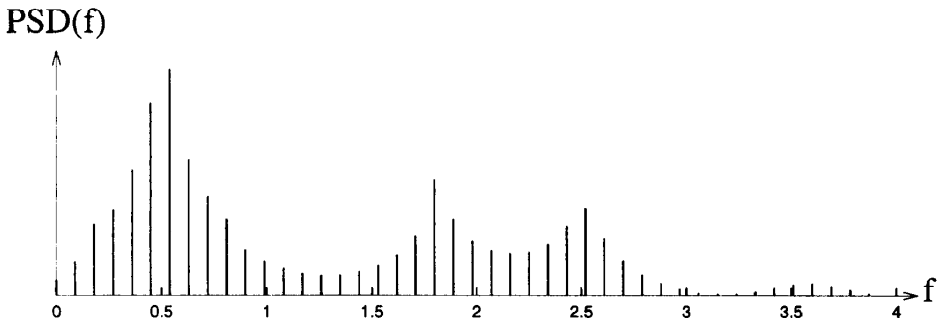

Figure 4.7: Power spectral density for speech, more specifically the sound **eh** pronounced by a male speaker. The horizontal axis is the frequency in KHz and the vertical axis is the energy density in dBm per Hz. The spectrum is obviously made up of discrete lines, and we note that three main resonances at 500, 1820, and 2510 Hz and a weak fourth at a higher frequency.

is that if the frequencies do not overlap the signals may be separated by appropriate filters. Isolation in the frequency domain is a sufficient (but not a necessary) condition for signals to be separable.

A third example is given by the speech signal. Most of the information in speech is encoded in the PSD; in fact our hearing system is almost insensitive to phase, although we use the phase difference between our ears to ascertain direction. In Figure 4.7 we see the spectrum of a (rather drawn out) **eh** sound. The vertical axis is drawn logarithmically, since our hearing system responds approximately logarithmically (see Section 11.2). We can't help noticing three phenomena. First, the spectrum is composed entirely of discrete lines the spacing between which changes with pitch. Second, there is more energy at low frequencies than at high ones; in fact when we average speech over a long time we discover a drop of between 6 and 12 dB per octave. Finally, there seem to be four maxima (called *formants*), three overlapping and one much smaller one at high frequency; for different sounds we find that these formants change in size and location. With appropriate training one can 'read' what is being said by tracking the formants.

In Section 9.3 we will learn that the PSD at a given frequency is itself the FT of a function called the autocorrelation.

$$|S(\omega)|^2 = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} s(t)s(t-\tau)\,dt \right] e^{-i\omega\tau}\,d\tau \tag{4.27}$$

The autocorrelation is a generalization of the idea of squaring the signal, and hence this relation tells us that the squaring operation can be performed either before or after the Fourier integral.

EXERCISES

4.5.1 Write a program that finds the PSD by numerical integration (equation (4.1)) and squaring. Use this program to find the PSD of a rectangular window (equation (4.21)) for several different widths. Repeat the exercise for a sinc-shaped pulse for several different pulse widths.

4.5.2 Build 1024 samples of sine waves of 1, 2, 3, 4, 5, 6, 7, and 8 KHz sampled at 8 KHz. Observe the sines in the time domain; can you see the aliasing for $f > 4$ KHz? Extract the PSD (if you didn't write your own program in the first exercise many programs are readily available for this purpose). Can you read off the frequency? What do you see now for $f > 4$ KHz?

4.5.3 Build 1024 sample points of sine waves with frequencies 1.1, 2.2, and 3.3 KHz sampled at 8 KHz. What happened to the spectral line? Try multiplying the signal by a triangular window function that linearly increases from zero at $n = 0$ to one at the center of the interval, and then linearly decreases back to zero).

4.5.4 In exercise 2.6.4 we introduced the V.34 probe signal. Extract its power spectrum. Can you read off the component frequencies? What do you think the probe signal is for?

4.5.5 Find the PSD of the sum of two sinusoids separated by 500 Hz (use 2 KHz $\pm$ 500 Hz) sampled at 8 KHz. Can you distinguish the two peaks? Now reduce the separation to 200 Hz. When do the two peaks merge? Does the triangular window function help?

4.5.6 In the text it was stated that isolation in the frequency domain is a *sufficient* but not a *necessary* condition for signals to be separable. Explain how can signals can be separated when their PSDs overlap.

4.6 Short Time Fourier Transform (STFT)

The Fourier transform is a potent mathematical tool, but not directly relevant for practical analog signal processing, because the integration must be performed from the beginning of time to well after the observer ceases caring about the answer. This certainly seems to limit the number of FTs you will calculate in your lifetime. Of course, one *can* compute the FT for finite time signals, since they were strictly zero yesterday and will be strictly zero tomorrow, and so you only have to observe them today. But that is only the case for signals that are *strictly* zero when you aren't observing them—small isn't good enough when we are integrating over an infinite amount of time!

In Section 4.2 we found the FT for various infinite time signals. Could we have approximated these mathematical results by numerically integrating over a finite amount of time? Other than the restrictions placed by the uncertainty theorem it would seem that this is possible. One needn't observe a simple sinusoid for years and years to be able to guess its spectrum. Of course the longer we observe it the narrower the line becomes, but we will probably catch on after a while. The problem is that we can't be completely sure that the signal doesn't radically change the moment after we give up observing it. Hence we can only give our opinion about what the signal's FT looked like over the time we observed it. Unfortunately, the FT isn't defined that way, so we have to define a new entity—the **S**hort **T**ime **F**ourier **T**ransform (STFT).

Consider the signal

$$s_1(t) = \begin{cases} \sin(2\pi f_1 t) & t < 0 \\ \sin(2\pi f_2 t) & t \geq 0 \end{cases}$$

which is a pure sine of frequency f_1 from the beginning of time until at time $t = 0$ when, for whatever reason, its frequency abruptly changes to f_2. What is the FT of this signal?

As we have seen, the FT is basically a tool for describing a signal simultaneously at all times. Each frequency component is the sum total of all contributions to this frequency from time $t = -\infty$ to $t = +\infty$. Consequently we expect the power spectrum calculated from the FT to have two equal components, one corresponding to f_1 and the other to f_2.

Now consider the signal

$$s_2(t) = \begin{cases} \sin(2\pi f_2 t) & t < 0 \\ \sin(2\pi f_1 t) & t \geq 0 \end{cases}$$

It is clear that the power spectrum will continue to be composed of two equal components as before since time reversal does not change the frequency composition. Assume now that f_1 and f_2 correspond to a whole number of cycles per second. Then the signal $s_3(t)$

$$s_3(t) = \begin{cases} \sin(2\pi f_1 t) & \lfloor t \rfloor \text{ even} \\ \sin(2\pi f_2 t) & \lfloor t \rfloor \text{ odd} \end{cases}$$

which consists of interleaved intervals of $\sin(2\pi f_1 t)$ and $\sin(2\pi f_2 t)$, must also have the same power spectrum!

The STFT enables us to differentiate between these intuitively different signals, by allowing different spectral compositions at different times. The

FT basically considers all signals to be unvarying, never changing in spectrum, while the STFT is an adaptation of the mathematical idea of the FT to the realities of the real world, where nothing stays unchanged for very long.

The STFT, or more accurately the short time PSD, goes under several different aliases in different fields. A 'musical score' is basically a STFT with a horizontal time axis, a vertical frequency axis and a special notation for durations. The STFT has long been a popular tool in speech analysis and processing, where it goes under the name of *sonogram*. The sonogram is conventionally depicted with a vertical frequency axis, with DC at the bottom, and a horizontal time axis, with time advancing from left to right. Each separate STFT is depicted by a single vertical line, traditionally drawn in a gray-scale. If there is no component at a given frequency at the time being analyzed the appropriate point is left white, while darker shades of gray represent higher energy levels. With the advent of DSP and computer graphics, analog sonographs with their rolls of paper have been replaced with scrolling graphics screens. The modern versions often use color rather than gray-scale, and allow interactive measurement as well.

Figure 4.8 is a sonogram of the author saying 'digital signal processing', with the sounds being uttered registered underneath. With some training one can learn to 'read' sonograms, and forensic scientists use the same sonograms for speaker identification. In the figure the basic frequency (pitch) of about

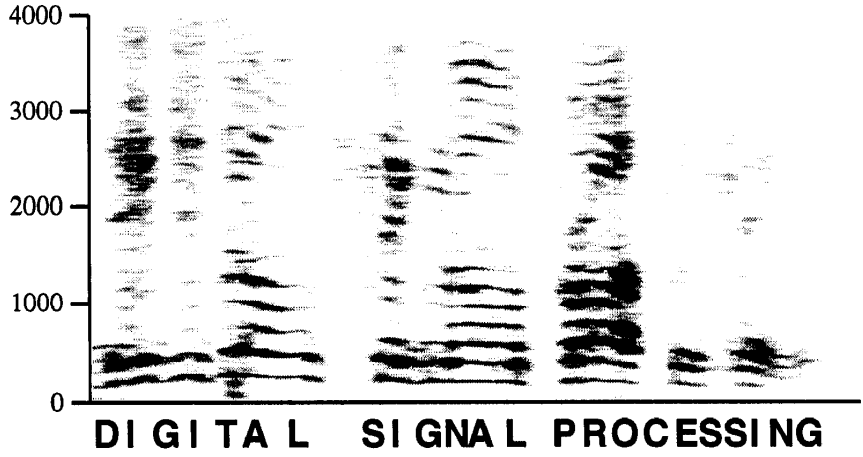

Figure 4.8: Sonogram of author saying 'digital signal processing'. The vertical axis is the frequency from 0 to 4000 Hz, while the horizontal axis is time (approximately 2 seconds). The sounds being uttered at each time are indicated by the writing below.

200 Hz is clearly visible at the bottom, and the difference between vowels and consonants is readily identifiable. You can probably also discern which syllables are accented, and may be able to see similarities among the various **i** sounds. The two **s** sounds in the last word seem to be invisible; this is due to their indeed having low energy, and most of that energy being spread out and at high frequencies, above the bandwidth displayed here. The **ing** is also very weak, due to being unaccented.

Rotating the sonogram by 90° we obtain the *falling raster spectrogram* popular in radar signal processing. Here the horizontal axis represents frequencies in the region of interest, time advances from top to bottom, and gray-scale intensity once again represents the square amplitude of the STFT component. Once the desired range of frequencies is selected, falling raster spectral displays provide intuitive real-time pictures; the display scrolling upwards as text does on a computer terminal.

The transition from FT to STFT requires forcing arbitrary signals to become finite time signals. To accomplish this we multiply the signal by a *window* function, that is, a function $w(t)$ that is strictly zero outside the time of interest. The window function itself should not introduce any artifacts to the spectrum of this product, and will be discussed in more detail in Section 13.4. For now you can think of the simplest window, the rectangular window of equation (4.21). Also commonly used are window functions that rise smoothly and continuously from zero to unity and then symmetrically drop back down to zero.

Of course, the uncertainty theorem puts a fundamental limitation on the precision of the STFT. The longer the time during which we observe a signal, the more precise will be our frequency distribution predictions; but the longer the window duration the more we blur the frequency changes that may be taking place in the signal. The uncertainty inequality does not allow us to simultaneously measure to arbitrary accuracy both the spectral composition and the times at which this composition changes.

The sonogram and similar graphic displays are tools to view the signal simultaneously in the time and frequency domains, yet they do not treat time and frequency on equal footing. What we may really want is to find a function $f(t,\omega)$ such that $f(t,\omega)\,dt\,d\omega$ is the energy in the 'time-frequency cell'. This brings us to define joint time-frequency distributions.

These are derived by considering time and frequency to be two characteristics of signals, just as height and weight are two characteristics of humans. In the latter case we can define a joint probability density $p(h,w)$ such that $p(h,w)\,dh\,dw$ is the percentage of people with both height between h and $h+dh$ *and* weight between w and $w+dw$ (see Appendix A.13). For such

joint probability distributions we require the so-called 'marginals',

$$p(h) = \int p(h,w)\,dw \qquad\qquad p(w) = \int p(h,w)\,dh$$

where the integrations are over the entire range of possible heights and weights, $p(h)dh$ is the percentage of people with height between h and $h+dh$ regardless of weight, and $p(w)dw$ is the percentage of people with weight between w and $w+dw$ regardless of height.

Similarly, a joint time-frequency distribution is a function of both time and frequency $p(t,\omega)$. We require that the following marginals hold

$$s(t) = \int_{-\infty}^{\infty} p(t,\omega)\,d\omega \qquad\qquad S(\omega) = \int_{-\infty}^{\infty} p(t,\omega)\,dt$$

and the integration over both time and frequency must give the total energy, which we normalize to $E = 1$. We may then expect $p(t,\omega)\,dt\,d\omega$ to represent the amount of energy the signal has in the range between ω and $\omega + d\omega$ during the times between t and $t + dt$.

Gabor was the first to express the STFT as a time-frequency distribution

$$p(t,\omega) = \frac{1}{\sqrt{2\pi}} \left| \int_{-\infty}^{\infty} s(\tau) w(\tau - t) e^{-\mathrm{i}\omega\tau} d\tau \right|^2$$

but he suggested using Gaussian-shaped windows, rather than rectangular ones, since Gaussians have the minimal uncertainty product. Perhaps even simpler than the short-time PSD is the double-square distribution

$$p(t,\omega) = |s(t)|^2\ |S(\omega)|^2$$

while more complex is the Wigner-Ville distribution.

$$p(t,\omega) = \frac{1}{2\pi} \int s^*\left(t - \frac{\tau}{2}\right) e^{-\mathrm{i}\omega\tau}\, s\left(t + \frac{\tau}{2}\right)\, d\tau$$

The double square requires computing $|S(\omega)|^2$ by the FT's integral over all time, and then simply multiplies this by the signal in the time domain. It is obviously zero for times or frequencies for which the signal is zero, but doesn't attempt any more refined time-frequency localization. The Wigner-Ville formula looks similar to equation (4.27) for finding the power spectrum via the autocorrelation, and is only one of an entire family of such *bilinear* distributions.

In addition to these, many other distributions have been proposed; indeed Cohen introduced a general family from which an infinite number of different time-frequency distributions can be derived,

$$p(t,\omega) = \frac{1}{4\pi^2} \int\int\int e^{-i\theta t - i\tau\omega + i\theta u} s^*(u - \frac{\tau}{2}) \varphi(\theta,\tau) s(u + \frac{\tau}{2})\, du\, d\tau\, d\theta$$

but none are perfect. Although they all satisfy the marginals, unexpected behaviors turn up. For example, when two frequencies exist simultaneously, some distributions display a third in between. When one frequency component ceases and another commences a short time later, some distributions exhibit nonzero components in the gap. These strange phenomena derive from the bilinear nature of the Cohen distributions. Even more bizarre is the fact that while the short-time PSD and the double-square are always positive, most of the others can take on nonintuitive negative values.

EXERCISES

4.6.1 There is another case for which we can compute the FT after only a finite observation time, namely when someone guarantees the signal to be periodic. Do we need the STFT for periodic signals?

4.6.2 In the text, examples were presented of signals with identical power spectra. Doesn't this contradict the very nature of a *transform* as a reversible transformation to another domain? Resolve this paradox by demonstrating explicitly the difference between the three cases.

4.6.3 Compute the FT by numerical integration and plot the empirical PSD of a sinusoid of time duration T. How does the line width change with T?

4.6.4 A FSK signal at any given time is either one of two sinusoids, one of frequency ω_1, and the other of frequency ω_2. Generate a FSK signal that alternates between ω_1 and ω_2 every T seconds, but whose phase is continuous. Using a sampling frequency of 8000 Hz, frequencies 1000 and 2000 Hz, and an alternation rate of 100 per second, numerically compute the power spectrum for various window durations. You may overlap the windows if you so desire. Plot the result as a falling raster spectrogram. What do you get when a transition occurs inside a window? Does the overall picture match what you expect? Can you accurately measure both the frequencies and the times that the frequency changed?

4.6.5 Repeat the previous exercise with the double-square distribution.

4.6.6 Show that the uncertainty theorem does not put any restrictions on joint time-frequency distributions, by proving that any distribution that satisfies the marginals satisfies the uncertainty theorem.

4.7 The Discrete Fourier Transform (DFT)

We have often discussed the fact that signals are functions of time that have pertinent frequency domain interpretation. The importance of being able to transform between time and frequency domains is accordingly evident. For analog signals we have seen that the vehicle for performing the transformation is the Fourier transform (FT), while in DSP it is the **D**iscrete **F**ourier **T**ransform (DFT).

The DFT can be derived from the FT

$$S(\omega) = \int_{-\infty}^{\infty} s(t)\, e^{-\mathrm{i}\omega t}\, dt$$

by discretization of the time variable. To accomplish this we must first determine the entire interval of time $[t_a \ldots t_z]$ wherein $s(t)$ is significantly different from zero. We will call the duration of this interval $T \equiv t_z - t_a$. If this time interval is very large, or even the entire t axis, then we can partition it up in some manner, and calculate the FT separately for each part. Next divide the interval into N equal-sized bins by choosing N equally spaced times $\{t_n\}_{n=0}^{N-1}$ in the following fashion $t_n = t_a + n\Delta t$ where $\Delta t \equiv \frac{T}{N}$. (Note that $t_0 = t_a$ but $t_{N-1} = t_z - \Delta t$; however, $t_{N-1} \approx t_z$ when $N \gg 1$ or equivalently $\Delta t \ll T$.) If we allow negative n, we can always take $t_a = 0$ without limiting generality. In this case we have $t_n = n\Delta t$. For sampled signals we recognize Δt as the basic sample interval (the inverse of the sampling frequency) $t_s = \frac{1}{f_s}$.

Now we also want to discretize the frequency variable. In a similar way we will define $\omega_k = k\Delta\omega$ with $\Delta\omega \equiv \frac{\Omega}{N}$. It is obvious that short time intervals correspond to high frequencies, and vice versa. Hence, if we choose to use a small Δt we will need a high upper frequency limit Ω. The exact correspondence is given by

$$N\Delta\omega = \Omega = \frac{2\pi}{\Delta t} \qquad or \qquad \Delta\omega\Delta t = \frac{2\pi}{N} \tag{4.28}$$

where we recognize an uncertainty product.

We can now evaluate the FT integral (4.2) as a Riemann sum, substituting t_n and ω_k for the time and frequency variables,

$$S(\omega) = \int_{-\infty}^{\infty} s(t)\, e^{-\mathrm{i}\omega t}\, dt \longrightarrow S_k = \sum_{n=0}^{N-1} s_n e^{-\mathrm{i}(k\Delta\omega)(n\Delta t)}$$

which upon substitution gives

$$S_k = \sum_{n=0}^{N-1} s_n e^{-\mathrm{i}\frac{2\pi nk}{N}} \tag{4.29}$$

which is the DFT. The power spectrum for the digital case is $|S_k|^2$ and each k represents the energy that the signal has in the corresponding 'frequency bin'.

For a given N, it is useful and customary to define the N^{th} root of unity W_N. This is a number, in general complex, that yields unity when raised to the N^{th} power. For example, one square root of unity is -1 since $(-1)^2 = 1$; but $1^2 = 1$ so 1 is a square root of itself as well. Also i is a fourth root of unity since $\mathrm{i}^2 = (-10)^2 = 1$, but so are $-\mathrm{i}$, -1, and 1. There is a unique *best* choice for W_N, namely the trigonometric constant

$$W_N \equiv e^{-\mathrm{i}\frac{2\pi}{N}} = \cos\left(\frac{2\pi}{N}\right) - \mathrm{i}\sin\left(\frac{2\pi}{N}\right) \tag{4.30}$$

which for $N = 2$ is as follows.

$$W_2 = e^{-\mathrm{i}\frac{\pi}{N}} = -1 \tag{4.31}$$

This is the best choice since its powers W_N^k for $k = 0 \ldots N-1$ embrace all the N roots. Thinking of the complex numbers as points in the plane, W_N is clearly on the unit circle (since its absolute value is one) and its phase angle is $\frac{1}{N}$ of the way around the circle. Each successive power moves a further $\frac{1}{N}$ around the circle until for $N = 1$ we return to $W_N^0 = 1$. This is illustrated in Figure 4.9 for $N = 8$.

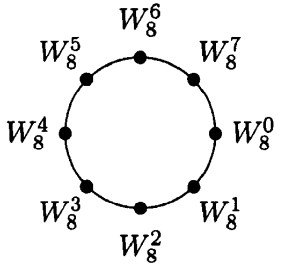

Figure 4.9: The N complex roots of unity displayed graphically. (Here $N = 8$.)

In terms of W_N the DFT can be expressed

$$S_k = \sum_{n=0}^{N-1} s_n W_N^{nk} \tag{4.32}$$

(just note that $(W_N)^{nk} = (e^{-\mathrm{i}\frac{2\pi}{N}})^{nk} = e^{-\mathrm{i}\frac{2\pi nk}{N}}$). The powers $(W_N)^{nk}$ are also on the unit circle, but at integer multiples of the basic angle. Consequently the set of all the powers of W_N divides the unit circle into N equal pieces.

It is illuminating to investigate the simplest DFT, the two-point transform. Substituting $N = 2$ into equation (4.31) we readily find

$$\begin{aligned} S_0 &= \sum_{n=0}^{1} s_n W_2^0 = s_0 + s_1 \\ S_1 &= \sum_{n=0}^{1} s_n W_2^n = s_0 - s_1 \end{aligned} \tag{4.33}$$

which has a simple interpretation. The zeroth (DC) coefficient is simply the sum (i.e., twice the average of s_0 and s_1). The other (high-frequency) coefficient is the difference (the derivative).

How do we return to the time domain given the discrete frequency components S_k?

$$s_n = \frac{1}{N} \sum_{k=0}^{N-1} S_k W_N^{-nk} \tag{4.34}$$

This is easy to show by direct substitution of (4.32).

Equations (4.32) and (4.34) are the main results of this section. We see that the s_n and the S_k can be calculated one from the other, and so contain precisely the same information. They form what is known as the discrete Fourier transform pair. With the equations we have derived one can go back and forth between the time and frequency domains, with absolutely no loss of information.

The DFT as we have derived it looks only at s_n over a finite interval of time. What happens if we take the DFT S_k and try to find s_n for times not in the interval from 0 to $N - 1$? The DC term is obviously the same outside the interval as inside, while all the others are periodic in N. Hence the DFT predicts $s_{N+n} = s_n$, not $s_{N+n} = 0$ as we perhaps expected! There is no way of getting around this paradox; as discussed in Section 2.8 the very act of sampling an analog signal to convert it into a digital one forces the spectrum to become periodic (aliased).

The only way to handle a nonperiodic infinite duration digital signal is to let the DFT's duration N increase without limit. Since the Nyquist frequency range is divided into N intervals by the DFT, the frequency resolution increases until the frequency bins become infinitesimal in size. At this point we have a denumerably infinite number of time samples but a continuous frequency variable $S(\omega)$ (defined only over the Nyquist interval). There is no consensus in the literature as to the name of this Fourier transform. We will sometimes call it the **Long Time DFT** (LTDFT) but only when we absolutely need to differentiate between it and the usual DFT.

The (short time) DFT takes in a finite number of digital values and returns a finite number of digital values. We thus have a true transform designed for digital computation. However, this transform is still a mathematical concept, not a practical tool. In Chapter 14 we will see that the DFT is eminently practical due to the existence of an efficient algorithm for its computation.

EXERCISES

4.7.1 Derive the LTDFT directly from the FT.

4.7.2 Express W_N^{-k} and $W_N^{(N-1)-k}$ in terms of W_N^k. Express $W_N^k + W_N^{-k}$ and $W_N^k - W_N^{-k}$ in terms of sine and cosine. How much is $W_N^{(n+m)k}$? Derive the trigonometric sum formulas (A.23) using these relations.

4.7.3 What is the graphical interpretation of raising a complex number to a positive integer power? What is special about numbers on the unit circle? Give a graphical interpretation of the fact that all powers of W_N are N roots of unity. Write a program that draws the unit circle and all the W_N^k. Connect consecutive powers of each root with straight lines. Describe the pictures you obtain for odd and even N.

4.7.4 What are the equations for 4-point DFT, and what is their interpretation?

4.7.5 Write a straightforward routine for the computation of the DFT, and find the digital estimate of the PSD of various sinusoids. Under what conditions is the estimate good?

4.8 DFT Properties

Some of the DFT's properties parallel those of the FT for continuous signals discussed in Section 4.3, but some are specific to signals with discrete time index. For most of the properties we will assume that the frequency index is discrete as well, but the obvious extensions to the LTDFT will hold.

First, let's review properties that we have already mentioned. We clearly need for the inverse operation defined in equation 4.34 to be a true inverse operation, (i.e., we need a sort of 'Fourier sum theorem').

$$\mathrm{DFT}^{-1}\,\mathrm{DFT}\; s = s \qquad \mathrm{DFT}\,\mathrm{DFT}^{-1}\, S = S \tag{4.35}$$

It is equally important for the DFT to be linear.

$$\begin{aligned} \mathrm{DFT}(x_n + y_n) &= X_k + Y_k \\ \mathrm{DFT}(as_n) &= aS_k \end{aligned} \tag{4.36}$$

Also important, but not corresponding to any characteristic of the FT, are the facts that the DFT and its inverse are *periodic* with period N. For example, when given a signal $s_0, s_1, \ldots s_{N-1}$ we usually compute the DFT for the N frequencies centered around DC. If we want the DFT at some frequency outside this range, then we exploit periodicity.

$$S_{k+mN} = S_k \qquad \text{for all integer } m \tag{4.37}$$

This leads us to our first implementational issue; how should we put the DFT values into a vector? Let's assume that our signal has $N = 8$ samples, the most commonly used indexation being 0 to $N-1$ (i.e., $s_0, s_1, \ldots s_7$). Since there are only 8 data points we can get no more than 8 independent frequency components, about half of which are negative frequency components.

$$S_{-4}, S_{-3}, S_{-2}, S_{-1}, S_0, S_1, S_2, S_3$$

Why is there an extra negative frequency component? Consider the signals

$$e^{\mathrm{i}2\pi fn} = \cos(2\pi fn) + \mathrm{i}\sin(2\pi fn) \qquad \text{where } f = \frac{k}{N}$$

for integer k, which are precisely the signals with only one nonzero DFT component. For all integer k in the range $1 \le k \le \frac{N}{2}$ the signal with frequency $f = +\frac{k}{N}$ and the corresponding signal with negative frequency $f = -\frac{k}{N}$ are different. The real part of the complex exponential is a cosine and so is unchanged by sign reversal, but the imaginary term is a sine and so changes sign. Hence the two signals with the same $|f|$ are complex conjugates. When $k = -\frac{N}{2}$ the frequency is $f = -\frac{1}{2}$ and the imaginary part is identically zero. Since this signal is real, the corresponding $f = +\frac{1}{2}$ signal is indistinguishable. Were we (despite the redundancy) to include both $f = \pm\frac{1}{2}$ signals in a 'basis', the corresponding expansion coefficients of an arbitrary signal would be identical; exactly that which is needed for periodicity to hold.

$$\ldots S_{-4}, S_{-3}, S_{-2}, S_{-1}, S_0, S_1, S_2, S_3, S_{-4}, S_{-3}, S_{-2}, S_{-1}, S_0, S_1, S_2, S_3, \ldots$$

In fact, any N consecutive Fourier coefficients contain all the information necessary to reconstruct the signal, and the usual convention is for DFT routines to return them in the order

$$S_0,\ S_1,\ S_2,\ S_3,\ S_4 = S_{-4},\ S_5 = S_{-3},\ S_6 = S_{-2},\ S_7 = S_{-1}$$

obtained by swapping the first half $(S_{-4}, S_{-3}, S_{-2}, S_{-1})$ with the second (S_0, S_1, S_2, S_3).

Let's observe a digital signal s_n from time $n = 0$ until time $n = N - 1$ and convert it to the frequency domain S_k. Now using the iDFT we can compute the signal in the time domain for all times n, and as we saw in the previous section the resulting s_n will be periodic. No finite observation duration can completely capture the behavior of nonperiodic signals, and assuming periodicity is as good a guess as any. It is convenient to visualize digital signals as circular buffers, with the periodicity automatically imposed by the buffer mechanics.

Now for some new properties. The DFT of a real signal is Hermitian even,

$$S_{-k} = S_k^* \qquad \text{for real } s_n \tag{4.38}$$

and that of an imaginary signal is Hermitian odd. Evenness (or oddness) for finite duration discrete time signals or spectra is to be interpreted according to the indexation scheme of the previous paragraph. For example, the spectrum $S_0, S_1, S_2, S_3, S_{-4}, S_{-3}, S_{-2}, S_{-1} =$

$$7,\ -1+(\sqrt{2}+1)\mathrm{i},\ -1+\mathrm{i},\ -1+\frac{\sqrt{2}-1}{4}\mathrm{i},\ -1,\ -1-\frac{\sqrt{2}-1}{4}\mathrm{i},\ -1-\mathrm{i},\ -1-(\sqrt{2}+1)\mathrm{i}$$

is Hermitian even and hence corresponds to a real signal. This property allows us to save computation time by allowing us to compute only half of the spectrum when the input signal is real.

Conversely, real spectra come from Hermitian even signals $(s_{-n} = s_n^*)$ and pure imaginary spectra from Hermitian odd signals. For example, the DFT of the signal $s_0, s_1, s_2, s_3, s_4, s_5, s_6, s_7 =$

$$7,\ -1-(\sqrt{2}+1)\mathrm{i},\ -1-\mathrm{i},\ -1-2(\sqrt{2}-1)\mathrm{i},\ -1,\ -1+2(\sqrt{2}-1)\mathrm{i},\ -1+\mathrm{i},\ -1+(\sqrt{2}+1)\mathrm{i}$$

will be real.

The properties that deal with transforming the discrete time and frequency axes are the time shifting property

$$\mathrm{DFT}\, s_{n-m} = e^{-\mathrm{i}mk} S_k \tag{4.39}$$

the time reversal property

$$\mathrm{DFT}\, s_{-n} = S_{-k} \tag{4.40}$$

and the frequency shifting (mixing) property.

$$\mathrm{DFT}\left(s_n e^{\mathrm{i}n\kappa}\right) = S_{k-\kappa} \tag{4.41}$$

Of course, for finite-duration DFTs, time shifts can move us to times where we haven't observed the signal, and frequency shifts to frequencies where we haven't computed the DFT. When this happens simply use the periodicity properties. When we use the word 'shift' for digital signals we always mean 'circular shift' (i.e., shift in a circular buffer).

Parseval's relation for the DFT is easy to guess

$$N \sum_{n=0}^{N-1} |s_n|^2 = \sum_{k=0}^{N-1} |S_k|^2 \tag{4.42}$$

and for infinite duration signals the sum on the left is over a denumerably infinite number of terms and the right-hand side becomes an integral.

$$\sum_{n=0}^{\infty} |s_n|^2 = \int_{-\infty}^{\infty} |S_k|^2 \, dk \tag{4.43}$$

The simplest application of Parseval's relation for the DFT involves a signal of length two. The DFT is

$$S_0 = s_0 + s_1 \qquad S_1 = s_0 - s_1$$

and it is easy to see that Parseval's relation holds.

$$S_0^2 + S_1^2 = (s_0 + s_1)^2 + (s_0 - s_1)^2 = 2(s_0^2 + s_1^2)$$

Products of discrete signals or spectra correspond to convolution *sums* rather than convolution integrals.

$$\text{LTDFT}\left(x_n y_n\right) = \sum_{\kappa=-\infty}^{\infty} X_k Y_{k-\kappa} \equiv X * Y \tag{4.44}$$

$$\text{LTDFT}^{-1}\left(X(\omega)Y(\omega)\right) = \sum_{m=-\infty}^{\infty} x_n y_{n-m} \equiv x * y \tag{4.45}$$

When the signals are of finite time duration the periodicity forces us to define a new kind of convolution sum, known as circular (or cyclic) convolution.

$$\text{DFT}\left(x_n y_n\right) = \frac{1}{N} \sum_{\kappa=0}^{N-1} X_k Y_{(k-\kappa) \bmod N} = X \circledast Y \tag{4.46}$$

$$\text{DFT}^{-1}\left(X_k Y_k\right) = \sum_{m=0}^{N-1} x_n y_{(n-m) \bmod N} = x \circledast y \tag{4.47}$$

where the indices $k-\kappa$ and $n-m$ wrap around according to the periodicity. In other words, while the linear (noncircular) convolution of x_0, x_1, x_2, x_3 with y_0, y_1, y_2, y_3 gives

$$\begin{aligned} x * y \;=\; & x_0y_0, \\ & x_0y_1 + x_1y_0, \\ & x_0y_2 + x_1y_1 + x_2y_0, \\ & x_0y_3 + x_1y_2 + x_2y_1 + x_3y_0, \\ & x_1y_3 + x_2y_2 + x_3y_1, \\ & x_2y_3 + x_3y_2 \\ & x_3y_3 \end{aligned}$$

the circular convolution gives the following periodic signal.

$$\begin{aligned} x \circledast y \;=\; & \ldots \\ & x_0y_0 + x_1y_3 + x_2y_2 + x_3y_1, \\ & x_0y_1 + x_1y_0 + x_2y_3 + x_3y_2, \\ & x_0y_2 + x_1y_1 + x_2y_0 + x_3y_3, \\ & x_0y_3 + x_1y_2 + x_2y_1 + x_3y_0, \\ & x_0y_0 + x_1y_3 + x_2y_2 + x_3y_1, \\ & x_0y_1 + x_1y_0 + x_2y_3 + x_3y_2, \\ & x_0y_2 + x_1y_1 + x_2y_0 + x_3y_3, \\ & \ldots \end{aligned}$$

We will return to the circular convolution in Section 15.2.

To demonstrate the use of some of the properties of the FT and DFT we will now prove the sampling theorem. Sampling can be considered to be implemented by multiplying the bandlimited analog signal $s(t)$ by a train of impulses spaced t_s apart. This multiplication in the time domain is equivalent to a convolution in the frequency domain, and since the FT if an impulse train in time is an impulse train in frequency, the convolution leads to a periodic FT. Stated in another way, the multiplication is a sampled signal s_n, and thus we should talk in terms of the DFT, which is periodic. We know that the impulse train in the frequency domain has repetition frequency $f_s = \frac{1}{t_s}$, and so the convolution forces the frequency domain representation to be periodic with this period. The situation is clarified in Figure 4.10 for the case of bandwidth less than half f_s. If the analog signal $s(t)$ has bandwidth wider than $\frac{1}{2}f_s$ the spectra will overlap, resulting in an irreversible loss of information.

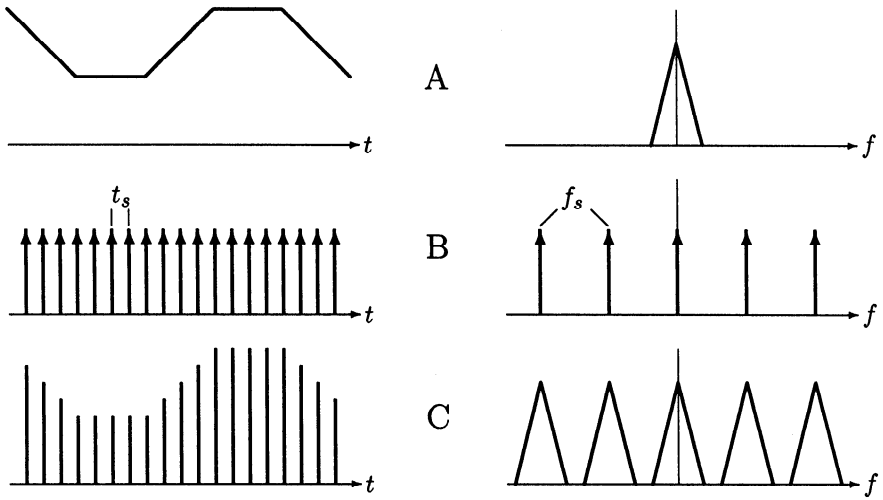

Figure 4.10: The sampling theorem. On the left we see the signals of interest in the time domain, and on the right in the frequency domain. The graphs in (A) depict the original analog signal, those in (B) the sampling impulses, and in (C) the sampled signal.

EXERCISES

4.8.1 Prove all of the DFT's properties stated above.

4.8.2 DFT routines usually return the same number of outputs as inputs, but sometimes we need higher frequency resolution. Assuming that we only have access to N samples, how can we generate $2N$ DFT components? Conversely, assume we have N DFT components and require $2N$ signal values. How can we retrieve them? These tricks seem to create new information that didn't previously exist. How can this be?

4.8.3 Prove that an even time signal has an even DFT, and an odd time signal has an odd DFT. What can you say about real even signals?

4.8.4 Explain why we didn't give the counterparts of several of the properties discussed for the FT (e.g., time scaling and differentiation).

4.8.5 Why does the circular convolution depend on N? (Some people even use the notation $x \circledN y$ to emphasize this fact.)

4.8.6 In Section 2.8 we mentioned the band-pass sampling theorem that holds for a signal with components from frequency $f_0 > 0$ to $f_z > f_0$. Using a figure similar to Figure 4.10 find the precise minimal sampling rate.

4.8.7 What can be said about the FT of a signal that is zero outside the time interval $-T < t < +T$? (Hint: This is the converse of the sampling theorem.)

4.9 Further Insights into the DFT

In this section we wish to gain further insight into the algebraic and computational structure of the DFT. This insight will come from two new ways of understanding the DFT; the first as the product of the W matrix with the signal, and the second as a polynomial in W.

The DFT is a linear transformation of a finite length vector of length N to a finite length vector of the same length. Basic linear algebra tells us that all linear transformations can be represented as matrices. This representation is also quite evident from equation (4.32)! Rather than discussing a function that transforms N signal values s_0 through s_{N-1} into frequency bins S_0 through S_{N-1}, we can talk about the product of an N by N matrix $\underline{\underline{W}}$ with an N-vector $(s_0, \ldots s_{N-1})$ yielding an N-vector $(S_0, \ldots S_{N-1})$.

$$\underline{S} = \underline{\underline{W}}\,\underline{s} \tag{4.48}$$

For example, the simple two-point DFT of equation (4.33) can be written more compactly as

$$\begin{pmatrix} S_0 \\ S_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} s_0 \\ s_1 \end{pmatrix}$$

as can be easily seen. More generally, the W_N matrix is

$$\underline{\underline{W}} = \begin{pmatrix} W_N^0 & W_N^0 & W_N^0 & \cdots & W_N^0 \\ W_N^0 & W_N^1 & W_N^2 & \cdots & W_N^{N-1} \\ W_N^0 & W_N^2 & W_N^4 & \cdots & W_N^{2(N-1)} \\ W_N^0 & W_N^3 & W_N^6 & \cdots & W_N^{3(N-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ W_N^0 & W_N^{N-1} & W_N^{2(N-1)} & \cdots & W_N^{(N-1)(N-1)} \end{pmatrix} \tag{4.49}$$

$$= \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & W_N & W_N^2 & \cdots & W_N^{N-1} \\ 1 & W_N^2 & W_N^4 & \cdots & W_N^{2(N-1)} \\ 1 & W_N^3 & W_N^6 & \cdots & W_N^{3(N-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & W_N^{N-1} & W_N^{2(N-1)} & \cdots & W_N^{(N-1)(N-1)} \end{pmatrix}$$

and since W_N is the N^{th} root of unity, the exponents can be reduced modulo N. Thus

$$\underline{\underline{W_2}} = \begin{pmatrix} W_2^0 & W_2^0 \\ W_2^0 & W_2^1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{4.50}$$

$$\underline{\underline{W_4}} = \begin{pmatrix} W_4^0 & W_4^0 & W_4^0 & W_4^0 \\ W_4^0 & W_4^1 & W_4^2 & W_4^3 \\ W_4^0 & W_4^2 & W_4^4 & W_4^6 \\ W_4^0 & W_4^3 & W_4^6 & W_4^9 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & W_4 & W_4^2 & W_4^3 \\ 1 & W_4^2 & W_4^0 & W_4^2 \\ 1 & W_4^3 & W_4^2 & W_4^1 \end{pmatrix} \quad (4.51)$$

$$= \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \end{pmatrix}$$

and

$$\underline{\underline{W_8}} = \begin{pmatrix} W_8^0 & W_8^0 & W_8^0 & W_8^0 & W_8^0 & W_8^0 & W_8^0 & W_8^0 \\ W_8^0 & W_8^1 & W_8^2 & W_8^3 & W_8^4 & W_8^5 & W_8^6 & W_8^7 \\ W_8^0 & W_8^2 & W_8^4 & W_8^6 & W_8^0 & W_8^2 & W_8^4 & W_8^6 \\ W_8^0 & W_8^3 & W_8^6 & W_8^1 & W_8^4 & W_8^7 & W_8^2 & W_8^5 \\ W_8^0 & W_8^4 & W_8^0 & W_8^4 & W_8^0 & W_8^4 & W_8^0 & W_8^4 \\ W_8^0 & W_8^5 & W_8^2 & W_8^7 & W_8^4 & W_8^1 & W_8^6 & W_8^3 \\ W_8^0 & W_8^6 & W_8^4 & W_8^2 & W_8^0 & W_8^6 & W_8^4 & W_8^2 \\ W_8^0 & W_8^7 & W_8^6 & W_8^5 & W_8^4 & W_8^3 & W_8^2 & W_8^1 \end{pmatrix} \quad (4.52)$$

which can be made explicit using $W_8 = e^{-\mathrm{i}\frac{\pi}{4}} = \frac{\sqrt{2}}{2}(1 - \mathrm{i})$.

The W matrix is symmetric, as is obvious from the above examples, but there are further, less obvious, symmetries as well. For instance, any two rows of the matrix are orthogonal, and the squared length (sum of squares of the elements) of any row is precisely N. Furthermore, there are relations between the elements of W_N and those of W_M when M divides N. It is these relations that make the FFT possible, as will be explained in Section 14.5.

The matrix representation gives us a simple interpretation for the inverse DFT as well. The IDFT's matrix must be the inverse of the DFT's matrix

$$\underline{s} = \underline{\underline{W}}^{-1} \underline{S} \quad (4.53)$$

and

$$\underline{\underline{W}}^{-1} = \frac{1}{N} \underline{\underline{W}}^* \quad (4.54)$$

where the Hermitian conjugate of the W_N matrix has elements

$$(W^*)_N^{nk} = e^{+\mathrm{i}\frac{2\pi nk}{N}} = W_N^{-nk}$$

as can easily be shown.

There is yet another way of writing the basic formula for the DFT (4.32) that provides us with additional insight. For given N and k let us drop the

indices and write $W \equiv W_N^k$. Then the DFT takes the form of a polynomial in W with coefficients s_n

$$S_k = \sum_{n=0}^{N-1} s_n W^n \tag{4.55}$$

which is a viewpoint that is useful for two reasons. First, the connection with polynomials will allow use of efficient algorithms for computation of polynomials to be used here as well. The FFT, although first introduced in signal processing, can be considered to be an algorithm for efficient multiplication of polynomials. Also, use of Horner's rule leads to an efficient recursive computation for the DFT known as Goertzel's algorithm. Second, a more modern approach considers the DFT as the polynomial approximation to the *real* spectrum. When the real spectrum has sharp peaks such a polynomial approximation may not be sufficient and rational function approximation can be more effective.

EXERCISES

4.9.1 Write explicitly the matrices for DFT of sizes 3, 5, 6, 7, and 8.

4.9.2 Invert the DFT matrices for sizes 2, 3, and 4. Can you write the iDFT matrix in terms of the DFT matrix?

4.9.3 Prove that any two rows of the DFT matrix are orthogonal and that the squared length of any row is N. Show that $\frac{1}{\sqrt{N}}\underline{\underline{W_N}}$ is a unitary matrix.

4.10 The z Transform

So far this chapter has dealt exclusively with variations on a theme by Fourier. We extended the FS for periodic analog signals to the FT of arbitrary analog signals, adapted it to the DFT of arbitrary digital signals, and modified it to the STFT of changing signals. In all the acronyms the ubiquitous **F** for Fourier appeared; and for good reason. The concept of spectrum *a la Fourier* is rooted in the basic physics of all signals. From colors of light through the pitch of voices and modes of mechanical vibration to frequencies of radio stations, Fourier's concept of frequency spectrum is so patently useful that it is hard to imagine using anything else.

In the special world of DSP there *is*, however, an alternative. This alternative is entirely meaningless in the analog world, in some ways less meaningful than the Fourier spectrum even in the digital world, and on occasion seems to be a mere artificial, purely mathematical device. It *does* sometimes enhance our understanding of signals, often greatly simplifies calculations, and always includes Fourier's spectrum as a special case.

This alternative is called the *z transform*, which we shall denote zT. This nomenclature is admittedly bizarre since the use of the letter z is completely arbitrary (there was no section in the previous chapter named 'Z Discovers Spectrum'), and it is not really a transform at all. Recall that the FS, which maps periodic analog signals to discrete spectra, is not called a transform. The FT, which maps analog signals to continuous spectra, and the DFT, which makes digital signals into discrete spectra, are. The zT takes an arbitrary digital signal and returns a continuous function. This change of form from sequence to function should disqualify it from being called a *transform*, but for some reason doesn't. Even more curious is the fact that outside the DSP sphere of influence the term 'z transform' is entirely unknown; but a closely related entity is universally called the *generating function*.

As we have done in the past, we shall abide by DSP tradition. After all, every field has its own terminology that has developed side by side with its advances and applications, even if these terms seem ridiculous to outsiders. Computer hardware engineers use *flip-flops* without falling. Programmers use *operating systems* without upsetting surgeons. Mathematicians use *irrational* numbers and *nonanalytic* functions, and no one expects either to act illogically. High-energy physicists hypothesize subatomic particles called *quarks* that have *strangeness*, *flavor*, and even *charm*. When lawyers *garnish* they leave people without appetite, while according to their definitions the victim of *battery* can be left quite powerless. So saying *DC* when there is no electric current, *spectral* when we are not scared, and *z transform* pales in comparison with the accepted terminologies of other fields!

The basic idea behind the classic generating function is easy to explain; it is a trick to turn an infinite sequence into a function. Classic mathematics simply knows a lot more about functions than it does about infinite sequences. Sometimes sequences can be bounded from above or below and in this way proven to converge or not. A few sequences even have known limits. However, so much more can be accomplished when we know how to change arbitrary sequences into functions; specifically, recursions involving sequence elements become algebraic equations when using generating functions.

Given a sequence $s_0, s_1, s_2, \ldots$, its generating function is defined to be

$$s(x) \equiv \sum_{n=0}^{\infty} s_n x^n \tag{4.56}$$

basically an infinite polynomial in x. The variable x itself is entirely artificial, being introduced solely for the purpose of giving the generating function a domain. It is easily seen that the correspondence between a sequence and its generating function is one-to-one; different sequences correspond to different generating functions, and different generating functions generate different sequences. In a way, generating sequences are the opposite of Taylor expansions. A Taylor expansion takes a function $s(x)$ and creates a sequence of coefficients s_n of exactly the form of equation (4.56), while the generating function does just the opposite. The Taylor coefficients give us intuition as to the behavior of the function, while the generating function gives us insight as to the behavior of the sequence.

We can demonstrate the strength of the generating function technique with a simple example, that of the Fibonacci sequence f_n. This famous sequence, invented by Leonardo of Pisa (nicknamed Fibonacci) in 1202, models the number of female rabbits in successive years. We assume that each mature female rabbit produces a female offspring each year and that no rabbit ever dies. We start with a single female rabbit ($f_0 = 1$); there is still only that rabbit after one year ($f_1 = 1$), since it takes a year for the rabbit to reach maturity. In the second year a new baby rabbit is born ($f_2 = 2$), and another in the third ($f_3 = 3$). Thereafter in each year we have the number of rabbits alive in the previous year *plus* those born to rabbits who were alive two years ago. We can deduce the recursive definition

$$f_0 = 1 \qquad f_1 = 1 \qquad f_n = f_{n-1} + f_{n-2} \qquad \text{for } n \geq 2 \tag{4.57}$$

that produces the values $1, 1, 2, 3, 5, 8, 13, 21, \ldots$. However, were we to need f_{137} we would have no recourse other than to recurse 137 times. Is there an explicit (nonrecursive) formula for f_n? At this point we don't see any way to find one, but this is where the generating function can help. Generating functions convert complex recursions into simple algebraic equations that can often be solved.

The generating function for the Fibonacci sequence is

$$f(x) = \sum_{n=0}^{\infty} f_n x^n = 1 + x + 2x^2 + 3x^3 + 5x^4 + 8x^5 + \ldots$$

and this is what we wish to evaluate. To proceed, take the recursion that defines the Fibonacci sequence, multiply both sides by x^n and sum from $n = 2$ to infinity.

$$\begin{aligned}
\sum_{n=2}^{\infty} f_n x^n &= \sum_{n=2}^{\infty} f_{n-1} x^n + \sum_{n=2}^{\infty} f_{n-2} x^n \\
&= x \sum_{n=2}^{\infty} f_{n-1} x^{n-1} + x^2 \sum_{n=2}^{\infty} f_{n-2} x^{n-2} \\
&= x \sum_{n=1}^{\infty} f_n x^n + x^2 \sum_{n=0}^{\infty} f_n x^n \\
f(x) - f_0 x^0 - f_1 x^1 &= x\left(f(x) - f_0 x^0\right) + x^2 f(x) \\
f(x) - 1 - x &= f(x)x - x + f(x)x^2
\end{aligned}$$

Solving the algebraic equation we easily find an explicit expression for the generating function

$$f(x) = \frac{1}{1 - x - x^2}$$

which is plotted in Figure 4.11.

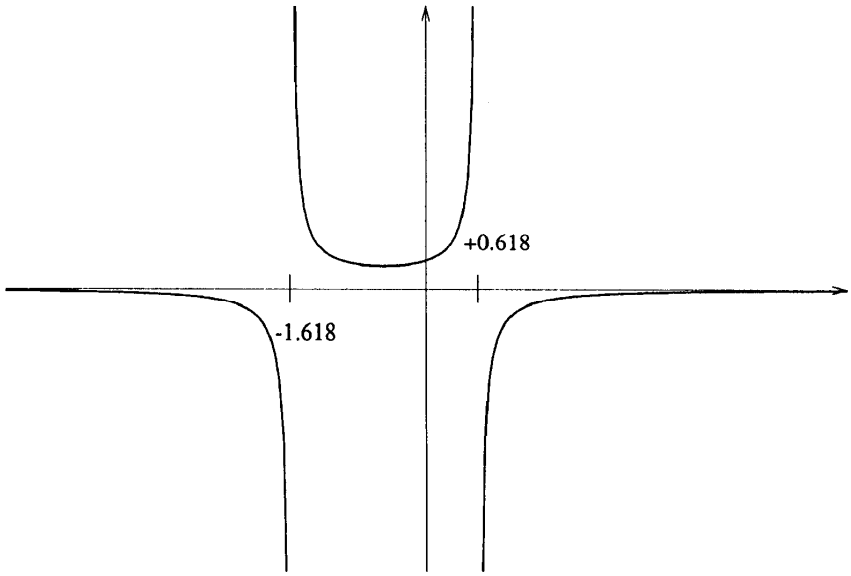

Figure 4.11: The generating function for the Fibonacci sequence. Note the divergences at $-\gamma \approx -1.618$ and $-\gamma' \approx 0.618$.

The zeros of the quadratic in the denominator are easily found to be $-\gamma$ and $-\gamma'$ where $\gamma \equiv \frac{1+\sqrt{5}}{2} = \cos^{-1}(\frac{\pi}{5})$ is the famous 'golden ratio' and $\gamma' = \equiv \frac{1-\sqrt{5}}{2} = -\frac{1}{\gamma}$.

We can now return to our original problem. In order to find an explicit formula for the n^{th} Fibonacci element, we need only to rewrite the generating function as an infinite polynomial and pick out the coefficients. To do this we use a 'partial fraction expansion'

$$f(x) = \frac{1}{(x+\gamma)(x+\gamma')} = \frac{1}{a-b}\left(\frac{a}{1-ax} - \frac{b}{1-bx}\right)$$

where $a + b = -ab = 1$. Utilizing the formula for the sum of a geometric progression $\frac{1}{1-ax} = \sum_{n=0}^{\infty}(ax)^n$ and comparing term by term, we find

$$f_n = \frac{1}{\sqrt{5}}\left(\gamma^{n+1} - (\gamma')^{n+1}\right) \tag{4.58}$$

the desired explicit formula for the n^{th} Fibonacci element.

Most people when seeing this formula for the first time are amazed that this combination of irrational numbers yields an integer at all. When that impression wears off, a feeling of being tricked sets in. The two irrational numbers in the numerator contain exactly a factor of $\sqrt{5}$, which is exactly what is being eliminated by the denominator; but if it is all a trick why can't a formula without a $\sqrt{5}$ be devised? So we are now surprised by our prior lack of surprise! Equation (4.58) is *so* astounding that you are strongly encouraged to run to a computer and try it out. Please remember to round the result to the nearest integer in order to compensate for finite precision calculations.

Now that we have become convinced of the great utility of generating functions, we will slightly adapt them for use in DSP. The z-transform is conventionally defined as

$$S(z) = \mathrm{zT}(s_n) = \sum_{n=-\infty}^{\infty} s_n z^{-n} \tag{4.59}$$

and you surely discern two modifications but there is also a third. First, we needed to make the sum run from minus infinity rather than from zero; second, the DSP convention is to use z^{-1} rather than x; and third, we will allow z to be a complex variable rather than merely a real one. The second change is not really significant because of the first; using z instead of z^{-1} is equivalent to interchanging s_n with s_{-n}. The really consequential

change is that of using a complex variable. Unlike the generating function we saw above, $S(z)$ is defined over the complex plane, called the z-plane. Sinusoids correspond to z on the unit circle, decaying exponentials to z inside the unit circle, growing exponentials to z outside the unit circle. The definition of z in the complex plane makes available even more powerful analytic techniques. The study of functions of complex variables is one of the most highly developed disciplines that mathematics has to offer, and DSP harnesses its strength via the z transform.

Any complex variable z can be written in polar form

$$z = re^{i\omega}$$

where r is the magnitude, and ω the angle. In particular, if z is on the unit circle $r = 1$, and $z = e^{i\omega}$. If we evaluate the zT on the unit circle in the z-plane, considering it to be a function of angle, we find

$$s(\omega) = S(z)\Big|_{z=e^{i\omega}} = \sum_{n=-\infty}^{\infty} s_n z^{-n} = \sum_{n=-\infty}^{\infty} s_n e^{-i\omega n} \tag{4.60}$$

which is precisely the DFT. The zT reduces to the DFT if evaluated on the unit circle.

For other nonunity magnitudes we can always write $r = e^{\lambda}$ so that $z = e^{\lambda + i\omega}$ and

$$S(z) = \sum_{n=-\infty}^{\infty} s_n z^{-n} = \sum_{n=-\infty}^{\infty} s_n e^{-(\lambda+i\omega)n} \tag{4.61}$$

which is a digital version of the **L**aplace **T**ransform (LT). The Laplace transform, which will not be discussed in detail here, expands functions in terms of exponentially increasing or damped sinusoids, of the type described in equation (2.11). Its expression is

$$f(s) = \int_{-\infty}^{\infty} f(t)e^{-st}dt \tag{4.62}$$

where s is understood to be complex (defining the s-plane). Sinusoids correspond to purely imaginary s, decaying exponentials to positive real s, growing exponentials to negative real s. The LT generalizes the FT, since the FT is simply the LT along the imaginary s axis. This is analogous to the zT generalizing the DFT, where the DFT is the zT on the unit circle. Although a large class of analog signals can be expanded using the FT, the LT may be more convenient, especially for signals that actually increase or decay

with time. This is analogous to the DFT being a sufficient representation for most digital signals but the zT often being more useful.

We have been ignoring a question that always must be raised for infinite series. Does expression (4.59) for the zT *converge*? When there are only a finite number of terms in a series there is no problem with performing the summation, but with an infinite number of terms the terms must decay fast enough with n for the sum not to explode. For complex numbers with large magnitudes the terms will get larger and larger with n, and the whole sum becomes meaningless.

By now you may have become so accustomed to infinities that you may not realize the severity of this problem. The problem with divergent infinite series is that the very idea of adding terms may be called into question. We can see that unconvergent sums can be meaningless by studying the following enigma that purports to prove that $\infty = -1$! Define

$$S = 1 + 2 + 4 + 8 + \ldots$$

so that S is obviously infinite. By pulling out a factor of 2 we get

$$S = 1 + 2(1 + 2 + 4 + 8 + \ldots)$$

and we see that the expression in the parentheses is exactly S. This implies that $S = 1 + 2S$, which can be solved to give $S = -1$. The problem here is that the infinite sum in the parentheses is meaningless, and in particular one cannot rely on normal arithmetical laws (such as $2(a + b) = 2a + 2b$) to be meaningful for it. It's not just that I is infinite; I is truly meaningless and by various regroupings, factorings, and the like, it can seem to be equal to anything you want.

The only truly well-defined infinite series are those that are *absolutely convergent*. The series

$$S = \sum_{n=0}^{\infty} a_n$$

is absolutely convergent when

$$A = \sum_{n=0}^{\infty} |a_n|$$

converges to a finite value. If a series S seems to converge to a finite value but A does not, then by rearranging, regrouping, and the like you can make S equal to just about anything.

Since the zT terms are $a_n = s_n z^n$, our first guess might be that $|z|$ must be very small for the sum to converge absolutely. Note, however, that the sum in the zT is from negative infinity to positive infinity; for absolute convergence we require

$$A = \sum_{n=-\infty}^{\infty} |s_n||z|^n = \sum_{n=-\infty}^{-1} |s_n||z|^n + \sum_{n=0}^{\infty} |s_n||z|^n = \sum_{n=1}^{\infty} |s_{-n}||\zeta|^n + \sum_{n=0}^{\infty} |s_n||z|^n$$

where we defined $\zeta \equiv z^{-1}$. If $|z|$ is small then $|\zeta|$ is large, and consequently small values of $|z|$ can be equally dangerous. In general, the **R**egion **O**f **C**onvergence (ROC) of the z transform will be a ring in the z-plane with the origin at its center (see Figure 4.12). This ring may have $r = 0$ as its lower radius (and so be disk-shaped), or have $r = \infty$ as its upper limit, or even be the entire z-plane. When the signal decays to zero for both $n \to -\infty$ and $n \to \infty$ the ring will include the unit circle.

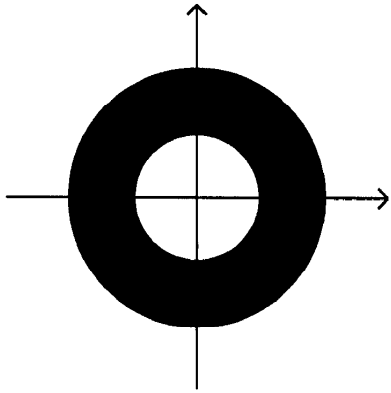

Figure 4.12: In general, the region of convergence (ROC) of the z transform is a ring in the z-plane with the origin at its center.

The z-plane where the zT lives, with its ROCs, poles, and zeros, is a more complex environment than the frequency axis of the FT. We will learn a lot more about it in the coming chapters.

EXERCISES

4.10.1 The zT is an expansion in basis functions $z^n = re^{i\omega n}$. Show that this basis is orthogonal.

4.10.2 Derive the generating function for a Fibonacci sequence with initial conditions $f_0 = 1$, $f_1 = 2$. What is the explicit formula for f_n?

4.10.3 The integer recursions for the two families of sequences g^+ and g^-

$$g^{\pm}_{n+1} = \begin{cases} 3g^{\pm}_n \pm 1 & g^{\pm}_n \text{ odd} \\ \frac{1}{2}g^{\pm}_n & g^{\pm}_n \text{ even} \end{cases}$$

may eventually lead to $g = 1$, or may oscillate wildly. For example, for the g^- case, $g_0 = 5$ leads to a cycle $5, 14, 7, 20, 10, 5$; no cycle has ever been found for the g^+ case (the Collatz problem). Compute numerically the generating functions $g^{\pm}(x)$ for $0 \leq x < 1$ and starting values $g_0 = 2 \ldots 10$. Can you tell which initial values cycle from the generating function?

4.10.4 Consider the infinite series $S = 1 - 1 + 1 - 1 + \ldots$. Writing this $S = (1 - 1) + (1 - 1) + \ldots = 0 + 0 + \ldots$ it would seem to converge to zero. Regroup to make S equal something other than zero. Is S absolutely convergent?

4.10.5 Show that if the zT of a signal is a rational function of z then the locations of poles and zeros completely specifies that signal to within a gain.

4.10.6 Show that the LT of $s(t)$ is the FT of $s(t)e^{-\lambda t}$.

4.10.7 Find the Laplace transforms of the unit impulse and unit step.

4.10.8 Derive the zT from the LT similarly to our derivation of DFT from FT.

4.10.9 According to the *ratio test* an infinite sum $\sum_{n=0}^{\infty} a_n$ converges absolutely if the ratio $|\frac{a_{n+1}}{a_n}|$ converges to a value less than unity. How does this relate to the zT?

4.11 More on the z Transform

Once again the time has come to roll up our sleeves and calculate a few examples. The first signal to try is the unit impulse $s_n = \delta_{n,0}$, for which

$$S(z) = \mathrm{zT}(\delta_{n,0}) = \sum_{n=-\infty}^{\infty} s_n z^{-n} = 1 \cdot z^0 = 1$$

which is analogous to the FT result. The series converges for all z in the z-plane. Were the impulse to appear at time $m \neq 0$, it is easy to see that we would get $S(z) = z^{-m}$, which has a zero at the origin for negative times and a pole there for positive ones. The ROC is the entire plane for $m \leq 0$, and the entire plane except the origin for $m > 0$.

What is the zT of $s_n = \alpha^n u_n$? This signal increases exponentially with time for $\alpha > 1$, decreases exponentially for $0 < \alpha < 1$, and does the same but with oscillating sign for $\alpha < 0$.

$$S(z) = \mathrm{zT}(\alpha^n u_n) = \sum_{n=-\infty}^{\infty} \alpha^n \, u_n \, z^{-n} = \sum_{n=0}^{\infty} (\alpha z^{-1})^n$$

Using the equation (A.47) for the sum of an infinite geometric series, we find

$$S(z) = \frac{1}{1 - \alpha z^{-1}} = \frac{z}{z - \alpha} \tag{4.63}$$

which has a pole at $z = \alpha$. The ROC is thus $|z| > |\alpha|$, the exterior of disk of radius α. When does the FT exist? As a general rule, poles in the z-plane outside the unit circle indicate explosive signal growth. If $|\alpha| < 1$ the ROC includes the unit circle, and so the FT converges. For the special case of the unit step $s_n = u_n$, we have $\alpha = 1$, so the zT is $\frac{z}{z-1}$ with ROC $|z| > 1$; the FT does not exist.

We can shift the signal step to occur at time m here as well. In this case

$$S(z) = \sum_{n=-\infty}^{\infty} \alpha^{n-m} u_{n-m} z^{-n} = \sum_{n=m}^{\infty} \alpha^{n-m} z^{-m} z^{-(n-m)}$$

which after a change in variable from n to $n - m$ gives

$$S(z) = z^{-m} \sum_{n=0}^{\infty} \alpha^n \, z^{-n} = z^{-m} \frac{1}{1 - \alpha z^{-1}} = \frac{z^{1-m}}{z - \alpha}$$

with poles at $z = \alpha$ and $z = 0$, and ROC unchanged.

What about $s_n = \alpha^{-n} u_n$? This is a trick question! This is the same as before if we write $s_n = (\frac{1}{\alpha})^n u_n$ so $S(z) = \frac{1}{1-\alpha^{-1}z}$ with ROC $|z| > |\alpha^{-1}|$. Since the sum we performed is true in general, the α used above can be anything, even imaginary or complex. Hence we know, for instance, that the zT of $e^{i\omega n}$ is $\frac{1}{1-e^{i\omega}z^{-1}}$ with ROC $|z| > |e^{i\omega}| = 1$.

We can perform a calculation similar to the above for $s_n = \alpha^n u_{-n}$.

$$\begin{aligned} S(z) = \; & \textstyle\sum_{n=-\infty}^{\infty} \alpha^n u_{-n} z^{-n} \; = \sum_{n=-\infty}^{0} (\alpha z^{-1})^n \\ = \; & \textstyle\sum_{n=0}^{\infty} (\alpha^{-1} z)^n \; = \frac{1}{1 - \alpha^{-1} z} \end{aligned}$$

The ROC is now $|z| < |\alpha|$, the interior of the disk. Shifting the ending time to $n = m$ we get

$$\begin{aligned} S(z) = \sum_{n=-\infty}^{\infty} \alpha^{n-m} u_{-(n-m)} z^{-n} &= \sum_{n=-\infty}^{m} \alpha^{n-m} z^{-m} z^{-(n-m)} \\ = \sum_{n=0}^{\infty} (\alpha^{-1} z)^n &= z^m \frac{1}{1-\alpha^{-1} z} \end{aligned}$$

with an extra pole if $m < 0$. It will often be more useful to know the zT of $s_n = \alpha^n u_{-n-1}$. This will allow covering the entire range of n with no overlap. It is convenient to remember that the zT of $s_n = -\alpha^n u_{-n-1}$ is exactly that of $s_n = \alpha^n u_n$ but with ROC $|z| < |\alpha|$. The desired transform is obtained by noting that multiplication of s_n by anything, including -1, simply causes the zT to be multiplied by this same amount.

Rather than calculating more special cases directly, let's look at some of the z transform's properties. As usual the most critical is *linearity*, i.e., the zT of $ax_n + by_n$ is $ax(z) + by(z)$. The ROC will always be at least the intersection of the ROCs of the terms taken separately. This result allows us to calculate more transforms, most importantly that of $\cos(\omega n)$. We know that $\cos(\omega n) = \frac{1}{2}(e^{i\omega n} + e^{-i\omega n})$, so the desired result is obtained by exploiting linearity.

$$S(z) = \tfrac{1}{2}\left(\frac{1}{1-e^{i\omega} z^{-1}} + \frac{1}{1-e^{-i\omega} z^{-1}}\right) = \frac{1-\cos(\omega) z^{-1}}{1-2\cos(\omega) z^{-1} + z^{-2}}$$

The next most important property of the zT is the effect of a time shift. For the FS and FT, shifting on the time axis led to phase shifts, here there is something new to be learned. In the cases we saw above, the effect of shifting the time by m digital units was to multiply the zT by z^{-m}. In particular the entire effect of delaying the digital signal by *one* digital unit of time was to multiply the zT by a factor of z^{-1}. This is a general result, as can be easily derived.

$$\begin{aligned} \mathrm{zT}(x_{n-1}) = \sum_{n=-\infty}^{\infty} x_{n-1} z^{-n} &= \sum_{n=-\infty}^{\infty} x_n z^{-(n+1)} \\ = \sum_{n=-\infty}^{\infty} x_n z^{-1} z^{-n} &= z^{-1} \sum_{n=-\infty}^{\infty} x_n z^{-n} = z^{-1}\,\mathrm{zT}(x_n) \end{aligned}$$

Accordingly the factor of z^{-1} can be thought of as a unit delay *operator*, as indeed we defined it back in equation (2.21). The origin of the symbol that was arbitrary then is now understood; delaying the signal by one digital unit

of time can be accomplished by multiplying it by z^{-1} in the z domain. This interpretation is the basis for much of the use of the zT in DSP.

For example, consider a radioactive material with half-life τ years. At the beginning of an experiment $n = 0$ we have 1 unit of mass $m = 1$ of this material; after one half-life $n = 1$ the mass has dropped to $m = \frac{1}{2}$ units, $\frac{1}{2}$ having been lost. At digital time $n = 2$ its mass has further dropped to $m = \frac{1}{4}$ after losing a further $\frac{1}{4}$, etc. After an infinite wait

$$\tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \tfrac{1}{16} + \ldots = 1$$

all of the material has been lost (actually converted into another material). The mass left as a function of time measured in half-lives is

$$m_n = \tfrac{1}{2}^n$$

an exponentially decreasing signal. Now a scientist measures the amount of mass at some unknown time n and wishes to predict (or is it postdict?) what the mass was one half-life back in time. All that need be done is to double the amount of mass measured, which is to use the operator z^{-1} with z being identified as $\frac{1}{2}$. This example might seem a bit contrived, but we shall see later that many systems when left alone tend to decrease exponentially in just this manner.

What about time reversal? For the FT this caused negation of the frequency; here it is straightforward to show that the zT of s_{-n} has its z variable inverted, $\mathrm{zT}(s_{-n}) = S(z^{-1})$. If the original signal had a ROC $R_l < |z| < R_h$, then the time-reversed signal will have a ROC of $R_l^{-1} > |z| > R_h^{-1}$. The meaning of this result is not difficult to comprehend; the inversion of $z = re^{\mathrm{i}\omega}$ both negates the ω and inverts r. Thus decaying exponentials are converted to exploding ones and vice versa.

You must be wondering why we haven't yet mentioned the inverse zT (izT). The reason is that it is somewhat more mathematically challenging than the other inverse operations we have seen so far. Remember that the zT's range is a ring in the complex z-plane, not just a one-dimensional line. To regain s_n from $S(z)$ we must perform a *contour integral*

$$s_n = \frac{1}{2\pi i} \oint S(z) z^{n-1} dz \tag{4.64}$$

over any closed counterclockwise contour within the ROC. This type of integral is often calculated using the residue theorem, but we will not need to use this complex mechanism in this book.

Many more special zTs and properties can be derived but this is enough for now. We will return to the zT when we study signal processing systems. Systems are often defined by complex recursions, and the zT will enable us to convert these into simple algebraic equations.

EXERCISES

4.11.1 Write a graphical program that allows one to designate a point in the z-plane and then draws the corresponding signal.

4.11.2 Plot the z transform of $\delta_{n,m}$ for various m.

4.11.3 Prove the linearity of the zT.

4.11.4 Express $\mathrm{zT}(\alpha^n x_n)$ in terms of $x(z) = \mathrm{zT}(x_n)$.

4.11.5 What is the z transform of the following digital signals? What is the ROC?

1. $\delta_{n,2}$
2. u_{n+2}
3. $a^n u(n)$
4. $a^n u(-n-1)$
5. $\frac{1}{2}^n u_n + \frac{3}{2}^n u_{-n}$

4.11.6 What digital signals have the following z transforms?

1. z^{-2}
2. z^{+2}
3. $\frac{1}{1-2z^{-1}}$ ROC $|z| > |2|$

4.11.7 Prove the following properties of the zT:

1. linearity
2. time shift $\mathrm{zT}s_{n-k} = z^{-k}S(z)$
3. time reversal $\mathrm{zT}s_{-n} = S(\frac{1}{z})$
4. conjugation $\mathrm{zT}s_n^* = S^*(z^*)$
5. rescaling $\mathrm{zT}(\alpha^n s_n) = S(\frac{z}{\alpha})$
6. z differentiation $\mathrm{zT}(ns_n) = -z\frac{d}{dz}S(z)$

4.12 The Other Meaning of Frequency

We have discussed two quite different representations of functions, the Taylor expansion and the Fourier (or z) transform. There is a third, perhaps less widely known representation that we shall often require in our signal

processing work. Like the Fourier transform, this representation is based on frequency, but it uses a fundamentally different way of thinking about the concept of frequency. The two usages coincide for simple sinusoids with a single constant frequency, but differ for more complex signals.

Let us recall the examples with which we introduced the STFT in Section 4.6. There we presented a pure sinusoid of frequency f_1, which abruptly changed frequency at $t = 0$ to become a pure sine of frequency f_2. Intuition tells us that we should have been able to recover an *instantaneous frequency*, defined at every point in time, that would take the value f_1 for negative times, and f_2 for positive times. It was only with difficulty that we managed to convince you that the Fourier transform cannot supply such a frequency value, and that the uncertainty theorem leads us to deny the existence of the very concept of instantaneous frequency. Now we are going to produce just such a concept.

The basic idea is to express the signal in the following way:

$$s(t) = A(t)\cos\Big(\Phi(t)\Big) \tag{4.65}$$

for some $A(t)$ and $\Phi(t)$. This is related to what is known as the *analytic representation* of a signal, but we will call it simply the *instantaneous representation*. The function $A(t)$ is known as the *instantaneous amplitude* of the signal, and the $\Phi(t)$ is the *instantaneous angle*. Often we separate the angle into a linear part and the deviation from linearity

$$s(t) = A(t)\cos\Big(\omega t + \phi(t)\Big) \tag{4.66}$$

where the frequency ω is called the *carrier frequency*, and the residual $\phi(t)$ the *instantaneous phase*.

The *instantaneous frequency* is the derivative of the instantaneous angle

$$2\pi f(t) = \frac{d\Phi(t)}{dt} = \omega + \frac{d\phi(t)}{dt} \tag{4.67}$$

which for a pure sinusoid is exactly the frequency. *This* frequency, unlike the frequencies in the spectrum, is a single function of time, in other words, a signal. This suggests a new world view regarding frequency; rather than understanding signals in a time interval as being made up of many frequencies, we claim that signals are fundamentally sinusoids with well-defined instantaneous amplitude and frequency. One would expect the distribution of different frequencies in the spectrum to be obtained by integration over the time interval of the instantaneous frequency. This is sometimes the case.

Consider, for example, a signal that consists of a sinusoid of frequency f_1 for one second, and then a sinusoid of nearby frequency f_2 for the next second. The instantaneous frequency will be f_1 and then jump to f_2; while the spectrum, calculated over two seconds, will contain two spectral lines at f_1 and f_2. Similarly a sinusoid of slowly increasing instantaneous frequency will have a spectrum that is flat between the initial and final frequencies.

This new definition of frequency seems quite useful for signals that we usually consider to have a single frequency at a time; however, the instantaneous representation of equation (4.65) turns out to very general. A constant DC signal can be represented (using $\omega = 0$), but it is easy to see that a constant plus a sinusoid can't. It turns out (as usual, we will not dwell upon the mathematical details) that all DC-less signals can be represented. This leads to an apparent conflict with the Fourier picture. Consider a signal composed of the *sum* of the two sinusoids with close frequencies f_1 and f_2; what does the instantaneous representation do, jump back and forth between them? No, this is exactly a *beat* signal (discussed in exercise 2.3.3) with instantaneous frequency a constant $\frac{1}{2}(f_1 + f_2)$, and sinusoidally varying amplitude is with frequency $\frac{1}{2}|f_1 - f_2|$. Such a signal is depicted in Figure 4.13. The main frequency that we see in this figure (or hear when listening to such a combined tone) is the instantaneous frequency, and after that the effect of $A(t)$, *not* the Fourier components.

We will see in Chapter 18 that the instantaneous representation is particularly useful for the description of communications signals, where it is the basis of *modulation*. Communications signals commonly carry informa-

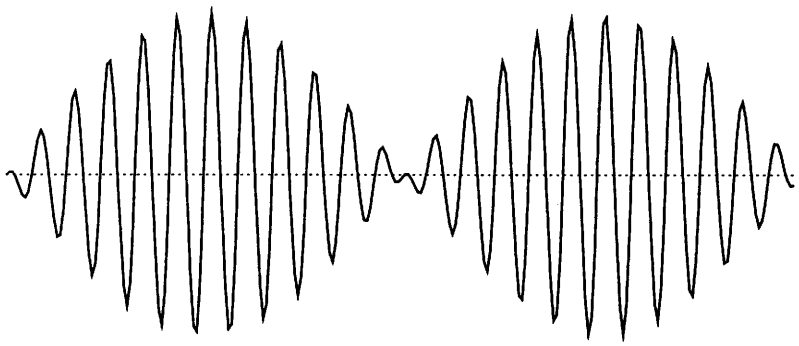

Figure 4.13: The beat signal depicted here is the sum of two sinusoids of relatively close frequencies. The frequencies we see (and hear) are the average and half-difference frequencies, *not* the Fourier components.

tion by varying (modulating) the instantaneous amplitude, phase, and/or frequency of a sinusoidal 'carrier'. The carrier frequency is the frequency one 'tunes in' with the receiver frequency adjustment, while the terms *AM* (**A**mplitude **M**odulation) and *FM* (**F**requency **M**odulation) are familiar to all radio listeners.

Let us assume for the moment that the instantaneous representation exists; that is, for any *reasonable* signal $s(t)$ without a DC component, we assume that one can find carrier frequency, amplitude, and phase signals, such that equation (4.65) holds. The question that remains is *how* to find them. The answering of this question is made possible through the use of a mathematical operator known as the Hilbert transform.

The Hilbert transform of a real signal $x(t)$ is a real signal $y(t) = \mathcal{H}x(t)$ obtained by shifting the phases of all the frequency components in the spectrum of $x(t)$ by 90°. Let's understand why such an operator is so remarkable. Assume $x(t)$ to be a simple sinusoid.

$$x(t) = A\cos(\omega t)$$

Obtaining the 90° shifted version

$$y(t) = \mathcal{H}x(t) = A\cos\left(\omega t - \frac{\pi}{2}\right) = A\sin(\omega t)$$

is actually a simple matter, once one notices that

$$y(t) = A\cos\left(\omega\left(t - \frac{\pi}{2\omega}\right)\right) = x\left(t - \frac{\pi}{2\omega}\right)$$

which corresponds to a time delay. So to perform the Hilbert transform of a pure sine one must merely delay the signal for a time corresponding to one quarter of a period. For digital sinusoids of period L samples, we need to use the operator $z^{-\frac{L}{4}}$, which can be implemented using a FIFO of length $L/4$.

However, this delaying tactic will not work for a signal made up of more that one frequency component, e.g., when

$$x(t) = A_1\cos(\omega_1 t) + A_2\cos(\omega_2 t)$$

we have

$$y(t) = \mathcal{H}x(t) = A_1\sin(\omega t) + A_2\sin(\omega t)$$

which does *not* equal $x(t-\tau)$ for any time delay τ.

Hence the Hilbert transform, which shifts all frequency components by a quarter period, independent of frequency, is a nontrivial operator. One way of implementing it is by performing a Fourier transform of the signal, individually shifting all the phases, and then performing an inverse Fourier transform. We will see an alternative implementation (as a *filter*) in Section 7.3.

Now let us return to the instantaneous representation

$$x(t) = A(t)\cos\Big(\omega t + \phi(t)\Big) \tag{4.68}$$

of a signal, which we now call $x(t)$. Since the Hilbert transform instantaneously shifts all $A\cos(\omega t)$ to $A\sin(\omega t)$, we can explicitly express $y(t)$.

$$y(t) = \mathcal{H}\,x(t) \;=\; A(t)\sin\Big(\omega t + \phi(t)\Big) \tag{4.69}$$

We can now find the instantaneous amplitude by using

$$A(t) = \sqrt{x^2(t) + y^2(t)} \tag{4.70}$$

the instantaneous phase via the (four-quadrant) arctangent

$$\phi(t) = \tan^{-1}\frac{y(t)}{x(t)} - \omega t \tag{4.71}$$

and the instantaneous frequency by differentiating the latter.

$$\omega(t) = \frac{d\phi(t)}{dt} \tag{4.72}$$

The recovery of amplitude, phase, or frequency components from the original signal is called *demodulation* in communications signal processing.

We have discovered a method of constructing the instantaneous representation of any signal $x(t)$. This method can be carried out in practice for digital signals, assuming that we have a numeric method for calculating the Hilbert transform of an arbitrary signal. The instantaneous frequency similarly requires a numeric method for differentiating an arbitrary signal. Like the Hilbert transform we will see later that differentiation can be implemented as a filter. This type of application of numerical algorithms is what DSP is all about.

EXERCISES

4.12.1 We applied the Hilbert transform to $x(t) = \cos(\omega t + \phi(t))$ and claimed that one obtains $y(t) = \sin(\omega t + \phi(t))$. Using trigonometric identities prove that this is true for a signal with two frequency components.

4.12.2 Even a slowly varying phase may exceed 2π or drop below zero causing nonphysical singularities in its derivative. What should be done to phases derived from equation (4.71) in such a case?

4.12.3 What is the connection between the instantaneous frequency and the spectrum of the signal? Compare the short time power spectrum calculated over a time interval to the histogram of the instantaneous frequency taken over this interval.

4.12.4 Show that given a signal $s(t)$ and any amplitude signal $A(t)$ an appropriate phase $\Phi(t)$ can be found so that equation (4.65) holds. Similarly, show that given any phase an amplitude signal may be found. The amplitude and phase are not unique; the $x(t)$ and $y(t)$ that are related by the Hilbert Transform are the *canonical* (simplest) representation.

4.12.5 Find an explicit direct formula for the instantaneous frequency as a function of $x(t)$ and $y(t)$. What are the advantages and disadvantages of these two methods of finding the instantaneous frequency?

4.12.6 We can rewrite the analytic form of equation (4.68) in quadrature form.

$$x(t) = a(t)\cos(\omega t) + b(t)\sin(\omega t)$$

What is the connection between $a(t)$, $b(t)$ and $A(t)$, $\phi(t)$? We can also write it in sideband form.

$$x(t) = \big(u(t) + l(t)\big)\cos(\omega t) + \big(u(t) - l(t)\big)\sin(\omega t)$$

What are the relationships now?

Bibliographical Notes

The DFT and zT are covered well in many introductory texts, e.g., [187, 252, 167], while the Hilbert transform and analytic representation are confined to the more advanced ones [186, 200]. An early book devoted entirely to the zT is [125], while tables were published even earlier [97].

The uncertainty theorem was introduced in quantum mechanics by Heisenberg [99]. Another physicist, Wigner [282], derived the first example of what we would call a time-frequency distribution in 1932, but this mathematical achievement had to be translated into signal processing terms. The article by Leon Cohen [39] is the best introduction.