Fundamentals of the Discrete Fourier Transform

Mark H. Richardson
Hewlett Packard Corporation
Santa Clara, California

The Fourier transform is a mathematical procedure that was discovered by a French mathematician named Jean-Baptiste-Joseph Fourier in the early 1800's. It has been used very successfully through the years to solve many types of engineering, physics, and mathematics problems. The Fourier transform is defined for continuous (or analog) functions, and is usually applied in situations where the functions are assumed to be continuous. More recently however, it has been implemented in digital form in various types of analyzers. These analyzers compute digital (or sampled) forms of power spectrums, frequency response functions, and other types of frequency domain functions from measured (sampled) time domain signals.

The implementation of the Discrete Fourier Transform, or DFT, became practical in 1965 when Cooley and Tukey described an algorithm for computing the DFT very efficiently. Their algorithm (and others like it) has become known as the Fast Fourier Transform (FFT).

Using the FFT algorithm, present day mini-computer based analyzers can compute a DFT in milliseconds where it used to take hours using standard computational procedures.

Direct computation of the DFT on an N-point complex valued function requires N^2 operations; where an operation is defined as one multiplication plus an addition. The Cooley-Tukey algorithm takes approximately $N \log_2 N$ operations; where N is a power of 2. Table 1 indicates how much longer it takes to compute a DFT by direct computation compared to the Cooley-Tukey algorithm, for typical data record sizes.

N	$N^2/N \log_2 N$
256	32
512	57
1024	102
2048	186
4096	341
8192	630

Table 1-Direct vs. FFT computation of DFT.

Many other methods for efficiently computing the DFT have since been discovered. However, all methods which require on the order of N log N operations have become known as FFT's.

The properties of the Fourier transform, and its cousin the Laplace transform are quite extensively documented, and their use as mathematical tools is taught in most undergraduate engineering curriculums today. However the use of the DFT, and the problems encountered with its application to measured time domain signals are not generally understood.

In this section all the fundamental concepts associated with the use of the DFT are presented.

We begin by examining the Fourier transform and some of its properties, and then show how a fundamental concept called "windowing" can be applied to the Fourier transform to derive the DFT and all of its properties. Using the convolution property, or as we will call it here, the windowing rule of Fourier transforms, we will define the concepts of sampling, aliasing, leakage and the wrap-around error. These are all important concepts which must be understood in order to avoid significant errors in the application of the DFT to measured data.

The Fourier Transform - The forward Fourier transform is defined as the integral

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \qquad (1)$$

The inverse Fourier transform is defined as the integral

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi ft} df \qquad (2)$$

X(f) is the (complex) Fourier transform of x(t), where f and t are real variables. We will assume that t is the time variable (in seconds) and f is the frequency variable (in Hertz), although this transform can be used in many other applications where these variables have different meanings. Normally x(t) is a real valued function of time but this restriction is not at all necessary. X(f) represents its corresponding frequency domain function.

The two functions x(t) and X(f) are known as a Fourier transform pair. There is a unique Fourier transform X(f) corresponding to each function x(t). Thus knowing X(f) is equivalent to knowing x(t) and visa-versa. X(f) and x(t) are really two different representations of the same phenome-

non. If the phenomenon is known in terms of x(t), then equation (1) shows how X(f) is represented in terms of x(t). Likewise if X(f) is known, equation (2) shows how x(t) is represented in terms of X(f).

Table 2 lists some commonly used Fourier transform pairs.

Time Domain Function	Frequency Domain Function
Auto Correlation	Auto Power Spectrum
Cross Correlation	Cross Power Spectrum
Impulse Response	Frequency Response

Table 2 - Fourier transform pairs.

Correlation Functions and Spectral Products - A *power* or *energy density spectrum* is defined as follows:

$$G_{xx}(f) = X(f)X^*(f) \qquad (3)$$

$G_{xx}(f)$ is a spectrum obtained by multiplying X(f) by its own conjugate $X^*(f)$. $G_{xx}(f)$ is real and positive at all frequencies. The inverse Fourier transform of $G_{xx}(f)$ is called the *autocorrelation function* of x(t) and is therefore written as the integral

$$R_{xx}(t) = \int_{-\infty}^{\infty} G_{xx}(f)e^{j2\pi ft}df \qquad (4)$$

This autocorrelation function is usually real valued, but not necessarily positive for all values.

The *cross power spectrum* is defined as

$$G_{yx}(f) = Y(f)X^*(f) \qquad (5)$$

where both Y(f) and X(f) are Fourier transforms obtained from the functions y(t) and x(t). The inverse transform is called the *cross-correlation function*, and can be written as the integral

$$R_{yx}(t) = \int_{-\infty}^{\infty} G_{yx}(f)e^{j2\pi ft}df \qquad (6)$$

This quantity is usually real valued.

The *frequency response function* (also called the transfer function) is computed as the ratio of the cross power spectrum over the auto power spectrum, i.e.

$$H(f) = G_{yx}(f) \Big/ G_{xx}(f) \qquad (7)$$

and its inverse Fourier transform is *the impulse response function*.

Another frequency domain function often used in conjunction with the frequency response function is the *coherence function*. It is computed as the ratio of the magnitude squared of the cross power spectrum divided by the product of the input and output auto power spectrums, i.e.

$$\gamma^2(f) = \frac{|G_{yx}(f)|^2}{[G_{xx}(f)G_{yy}(f)]} \qquad (8)$$

The coherence function is real valued having values between zero and one.

Figure 1 shows one of the transform pairs, the impulse response and frequency response function, plotted along the time and frequency axes respectively.

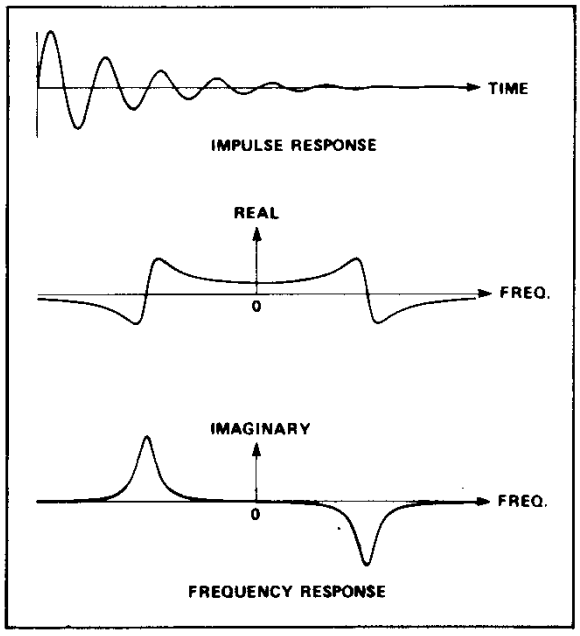

Figure 1—Plot of Fourier transform pair.

Notice that the frequency function X(f) exhibits some symmetry about the origin (f = 0) of the frequency axis. That is the real part *Re* [X(f)] satisfies the equation

Re [X(f)] = *Re* [X(-f)]

and the imaginary part *Im* [X(f)]satisfies

Im [X(f)] = - *Im* [X(-f)]

Re [X(f)] is called an even function and *Im* [X(f)] an odd function.

Another way of saying this is that the function value for negative frequencies is the complex conjugate of the function value for the corresponding positive frequency, i.e.

$X(f) = X^*(-f)$

This property is called *Hermitian symmetry*. It is the result of the following general rule.

Symmetry Rule: The Fourier transform of a real valued function is Hermitian symmetric about the origin in the other (transform) domain.

Because of the Hermitian symmetry, normally only the values of the frequency function X(f) for non-negative frequencies are displayed on the CRT of a Fourier analyzer. It is important to keep in mind however that the complete function X(f) also includes the conjugate values for negative frequencies.

Windowing - In practice we usually measure some signal x(t) which corresponds to the phenomenon we wish to analyze, e.g. motion, pressure, temperature, etc. This analog signal x(t) is commonly measured as an electrical voltage, and it could be processed using analog techniques to perform the indicated multiplication and integration of equation (1) to obtain the Fourier transform of the signal. However this is impractical and is not done in commercially available instrumentation today.

Moreover, a more serious drawback to obtaining a Fourier transform is that we can't measure the signal x(t) over the infinite interval (-∞, ∞), but only over some finite interval (t_1,t_2) as shown in Figure 2. Hence we never measure the entire signal x(t) but only a "windowed" version of it, $\bar{x}(t)$.

The windowed

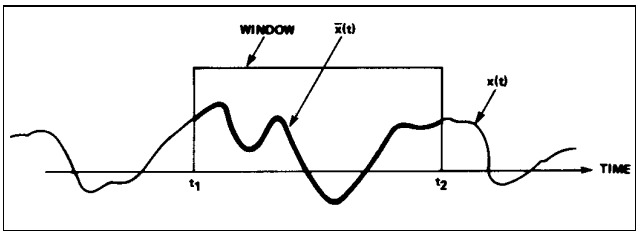

Figure 2 - Truncating the time signal.

signal x(t) can be thought of as the entire signal x(t) multiplied by the windowing function w(t), where w(t) is equal to 1 in the interval (t_1,t_2) and zero elsewhere.

Secondly, the DFT works with digital or sampled data. This sampling process can also be thought of as a multiplication of the continuous signal by a sequence of unit amplitude impulses as shown in Figure 3. Hence the sampled data is really the continuous data multiplied by a "sampling" window or function.

This windowing of the signal brings into effect a fundamental rule of the Fourier transform.

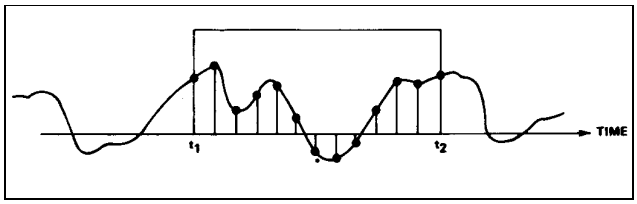

Figure 3 - Sampling the time signal.

Windowing Rule: If two functions are multiplied together in one domain their Fourier transforms are convolved together in the other domain.

Convolution is a simple mathematical procedure but is somewhat difficult to grasp conceptually without working out an example. Convolution is defined by the integral

$$c(t) = \int_{-\infty}^{\infty} x(\tau)y(t-\tau)d\tau$$

Figure 4 depicts the process for two rectangular functions x(t) and y(t). The function c(t) is computed by sliding the function y(-t) in the time direction and summing up (integrating) the products of the two functions where they intersect. After y(-t) has completely passed by x(t) the function c(t) is complete.

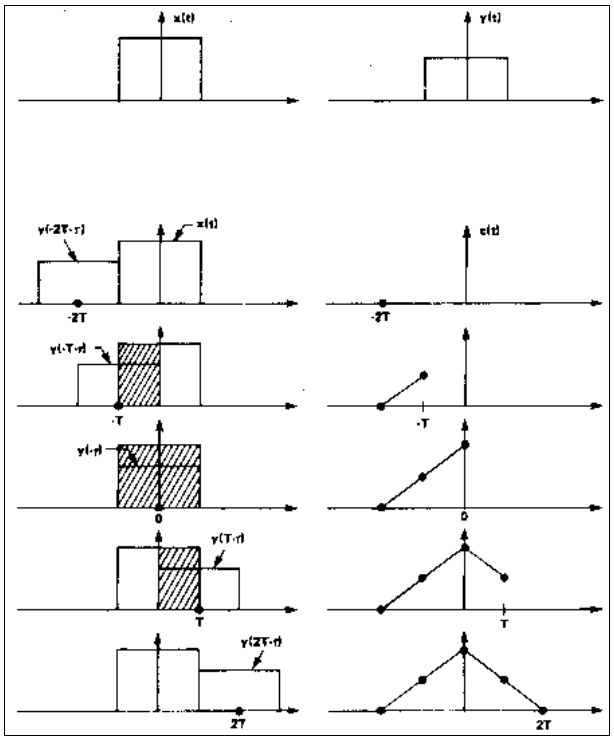

Figure 4 - Convolution of two functions.

The Discrete Fourier Transform (DFT) - The need for calculation of the Fourier transform has become increasingly important over the years, partly because the cost of computers has been steadily declining, and partly because

the difficulty and sophistication of our measurements has been steadily increasing. The continuous transform theory is very useful for theoretical work, but is not suited for calculation with instrumentation techniques. To compute the DFT we must work with sampled versions of our functions in both the time and the frequency domains, and our functions are of limited duration in both domains.

It is possible to either develop the DFT from basic axioms, or to derive it from the continuous Fourier transform. We will take the latter approach, because one or our main goals will be to relate the results of our discrete measurements and calculations to the results that we expect from application of the Fourier transform to continuous signals.

Sampled versus Continuous Data

There are three modifications that must be made to the time function x(t) and its transform X(f), in order to represent these functions in digital instrumentation.

It is therefore convenient to describe this process of converting continuous data to discrete data as three distinct steps or operations. A thorough understanding of these processing steps, and the order in which they occur, should eliminate most of the confusion that might arise concerning the interpretation of various DFT results.

1. x(t) must be multiplied by a time window w(t) of duration T to obtain a time record of finite extent. This results in the convolution of the spectrum X(f) with the transform of w(t). This transform is called the line shape of w(t) and denoted by L(f).

2. $x(t) = x(t)\, w(t)$ must be sampled N times at Δt intervals in anticipation of storage in a digital memory. Thus, $T = N\Delta t$. This is done by multiplying the windowed time function by a second "sampling function" known as the SHAH function. This causes replication of the frequency function at intervals $1/\Delta t$ along the frequency axis.

3. Finally, it is necessary to restrict the resulting frequency function to a finite number of samples in order to store the result in a digital memory. The frequency function is sampled N times at Δf intervals. Thus $2F_{max} = N\Delta f$ if we sample the function over an interval $(-F_{max}, F_{max})$. This sampling is done by multiplying the frequency function by another SHAH function which causes replication of the time function at intervals $1/\Delta f$ along the time axis.

These three essential steps are illustrated in both domains by the example in Figure 5. Here, the original time function is a cosine function of frequency (f_o), and its true frequency spectrum is a pair of delta functions located at frequencies $\pm f_o$. We have applied a rectangular window w(t) in the time domain, corresponding to the frequency

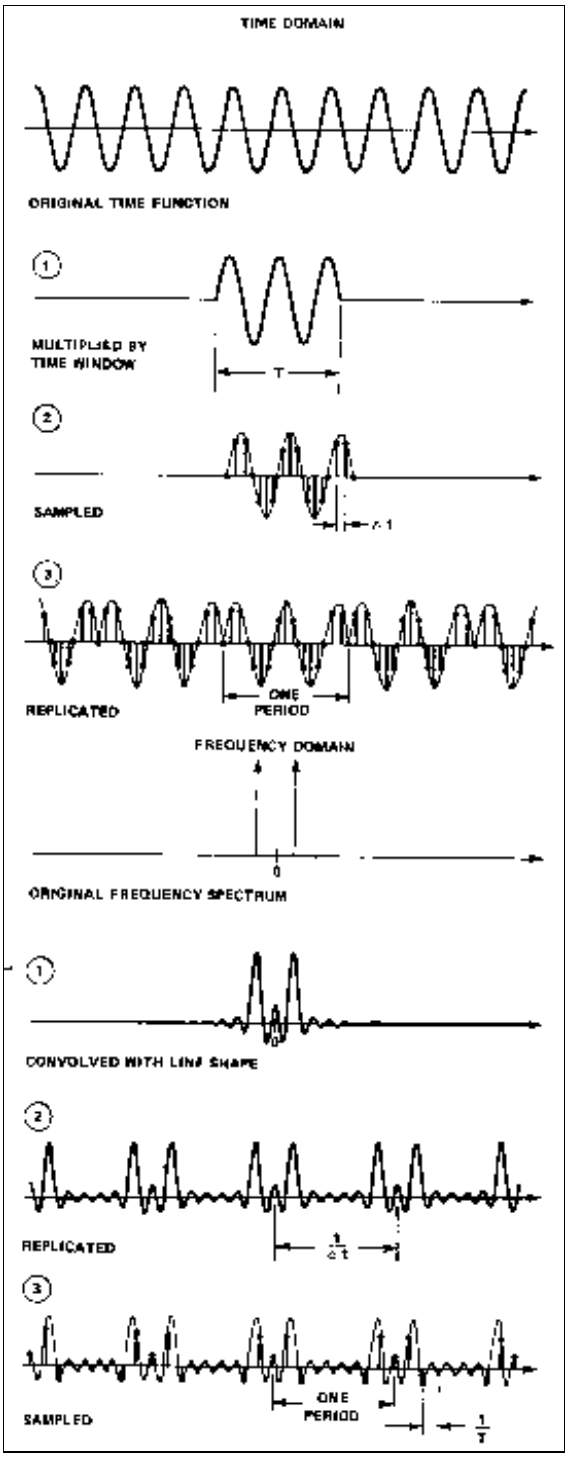

Figure 5 - The three steps to convert continuous data to discrete data.

domain convolution with the line shape $L(f) = (\sin \pi f)/ \pi f$. Notice that each delta function is "smeared" by L(f) and causes intermixing of various frequencies. This phenomenon is called *leakage*, and will be discussed in more detail later.

The time function is sampled at intervals of Δt, and the smeared frequency spectrum is reproduced at Intervals of $1/\Delta t$ along the frequency axis. Notice that there is some over lap of these replicated spectra. This overlap effect is called *aliasing*, and will also be discussed more fully in a later section.

Finally, the replicated spectrum is sampled at Δf intervals, causing the truncated time function to be replicated at $1/\Delta f$ intervals, and becoming periodic with a $1/\Delta f$ period. The net result is a periodic sequence of N samples per period in each domain. Since each function can be described entirely if one period is known, it is only necessary to store the N samples of one period.

It should be apparent that the final discrete frequency spectrum appears to be substantially different from the true spectrum. However in the following sections where each of these steps is examined in more detail, techniques for minimizing these errors will be discussed.

Returning to the definition of the Fourier transform eq. (1), if we now let $f = m\Delta f$, $t = n\Delta t$ and if we assume that x(t) has non zero values only at the times $t = n\Delta t$ for N values of n and that X(f) has non-zero values only at the frequencies $f = m\Delta f$ for N values of m then the Discrete Fourier Transform becomes

$$X(m\Delta f) = \Delta t \sum_{n=0}^{N-1} x(n\Delta t) e^{-j2\pi mn/N}$$

$$m = 0\ldots, N-1$$

Likewise the Inverse Discrete Fourier transform is written

$$x(n\Delta t) = \Delta f \sum_{m=0}^{N-1} X(m\Delta f) e^{j2\pi mn/N}$$

$$n = 0\ldots, N-1$$

where $T = N\Delta t = 1/\Delta f$. The independent variables are now the pair of indices m and n. $X(m\Delta f)$ has discrete values at intervals of $\Delta f = 1/T$, and is periodic with period $1/\Delta t$, while $x(n\Delta t)$ has discrete values at intervals of Δt, and is periodic with period T. Each sequence comprises N numbers, which may be real or complex. If $x(n\Delta t)$ is real, then $X(m\Delta f)$ will be Hermitian. Thus, only the positive half-period of $X(m\Delta f)$ need be calculated. There are exactly N independent numbers that describe the transform function in each domain. In the time domain, there are N real samples of the time function, and in the frequency domain there are (N-1)/2 complex numbers for the range $1<m<(N/2)-1$, and 2 real numbers for m = 0, N/2. If $x(n\ \Delta t)$ is complex, then there are 2N real numbers in each domain, and a full period of $X(m\Delta f)$ is needed.

It should be apparent that the DFT is ideally suited for handling digital data. The DFT is, strictly speaking, a relation between sequences of sample coefficients, whereas the actual time and frequency functions are comprised of sequences of delta functions. As previously discussed, the functions can be obtained by scaling the impulses of a SHAH function with these coefficients over the appropriate interval in either domain. Hence, the sequence of sample coefficients can be thought of as a sampled function, with the implication that the SHAH function may be introduced when necessary.

It should also be emphasized that the DFT gives the *correct* Fourier transform for sampled periodic data. If the original time function x(t) is indeed periodic, such that the time window width T is an integer number of periods, and if the frequency spectrum is band-limited so as not to exceed $\pm 1/(2\Delta t)$, then the DFT will produce the correct spectrum without any leakage or aliasing. Sometimes these restrictions are met, but generally there is both leakage and aliasing to contend with. These anomalies are caused by the sampling process however, and not by the DFT.

Time Windows - The first step in converting a long continuous time function into a finite sequence of data samples is truncation of the time signal. We define a time window w(t), such that w(t) = 0 outside some time interval of duration T. This interval may be anywhere along the time axis, but we generally either begin the interval at t = 0, or center the interval about the origin. Regardless of where we choose the interval, only the phase of the frequency function is altered by time displacement, and the magnitude is unchanged. For most of the following discussion on windows, we will assume an interval centered at t = 0.

Applying the windowing rule, we know that multiplication in the time domain corresponds to convolution in the frequency domain. Thus, the frequency function after windowing is the original function convolved with the line shape of the window L(f).

As an example of a common window line shape pair, consider the rectangular window w(t) and its corresponding line shape L(f) = (sin Tf) / (f) = T sinc(Tf). These functions are illustrated in Figure 6.

If the true spectrum X(f) is convolved with L(f), the result will be the sum of a sequence of X(f) shapes, each slightly displaced in frequency, and weighted by the value of L(f) at that displacement. Suppose, for example that x(t) = 1 before any windowing is attempted. Then $X(f) = \delta(f)$ is a delta function at the frequency origin. The introduction of a rectangular window produces X(f) = T sinc(Tf), instead of the delta function. Thus, the convolution with a line shape "smears" the true frequency spectrum over a significant range of adjacent frequencies. This general behavior is typi-

cal of all line shapes that result from finite time windows. The line shape comprises a

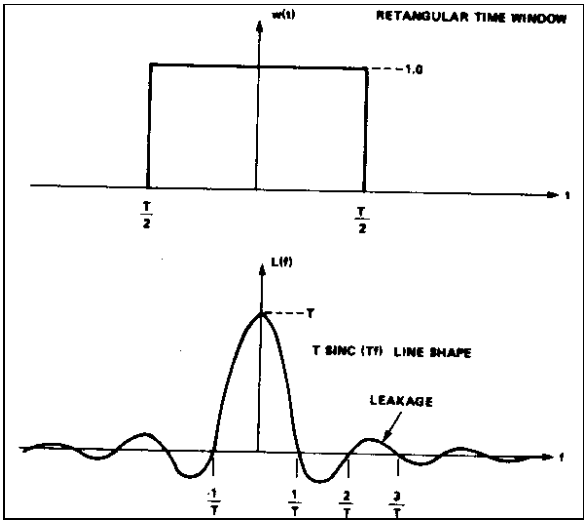

Figure 6 - A rectangular time window and its corresponding frequency line shape.

main lobe (analogous to antenna or optical theory) centered about the frequency origin, and a number of smaller side lobes that persist over the remaining frequency range. When we convolve a spectrum with a function of this sort, we find that the energy or power that is intended to be associated with a particular frequency, is in reality spread over a wide range of adjacent frequencies. We call this phenomenon leakage. It should be rather obvious that leakage can be a serious source of error in our measurements, unless we take steps to constrain this parasitic spread of energy.

Leakage, and What To Do About It

Leakage is always undesirable! Fortunately, it can sometimes be completely eliminated, and can always be reduced to an arbitrarily small amount with a suitable window.

Leakage is completely absent when a rectangular window is applied to data that is exactly periodic in the window interval T, and the resulting spectrum is sampled at intervals of $\Delta f = 1/T$. The reason becomes apparent when we observe that frequency domain sampling causes time domain replication, and that the replication of one cycle of a periodic function simply reproduces the original function. In the frequency domain, the line shape sinc(ft) has nulls at all multiples of 1/T along the frequency axis (except at f = 0), and the spectrum of the original periodic time signal has values only at frequencies which are multiples of 1/T. Thus, the convolution with sinc(fT) does not produce any interaction between these frequencies at multiples of 1/T.

Obviously, if existing time signals are periodic, it is possible to eliminate leakage if the window width can be adjusted properly to encompass an integer number of data periods. In some cases, measurements can be made on a physical system by applying a controlled source of excitation to the system, in which the excitation is made suitably periodic in the window interval. It is necessary that the period be held as close to T as possible throughout the measurement interval, or else the spectral samples will occur at points where the sinc(fT) line shape is not precisely zero.

Thus, for signals that are periodic in the window interval T. or for transients (whether repetitive or not) that completely decay within the window interval, the rectangular window is optimum. We assume that non-periodic noise is negligible in these cases. Other windows may be used, but the result will be a considerable and unnecessary loss in frequency resolution.

When x(t) contains a significant amount of random noise, such noise is not periodic in the window interval T. Thus, the measured noise spectrum will have considerable leakage. If the shape of the noise spectrum is of particular interest, it is necessary to choose a window with less leakage than the rectangular shape.

When x(t) is not periodic in the window interval T. we must use a window whose line shape is a compromise between small side lobes and a narrow main lobe. The side lobes can always be made arbitrarily small, at the expense of main lobe broadening. Thus, it is always necessary to balance the desired frequency resolution against the deleterious effects of leakage. The optimum choice depends on the application. One window that is commonly used with non-periodic random data is the Hanning window, shown in Figure 7.

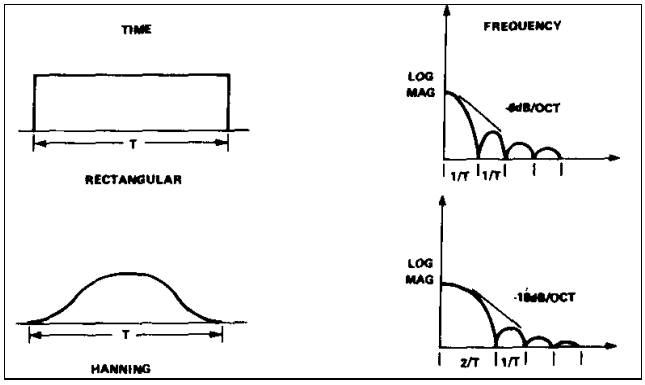

Figure 7 - Windows and their line shapes.

Sampling In the Time Domain - Since the DFT operates on digital or sampled data, the measured time domain signal must be sampled, and the resulting numbers stored in memory.

Ideal sampling is the process of observing a continuous signal only at discrete instants of time. The FFT algorithm assumes that the signal is sampled uniformly, i.e. the time period between successive samples remains constant.

The sampling process may be viewed as the multiplication of a continuous signal by a sequence of unit amplitude delta functions. This sequence of delta functions, called the SHAH function, is shown together with its Fourier transform in Figure 8a. Note that if we sample the signal with a period t between samples, the Fourier transform of the SHAH function is another SHAH function with impulses separated by a frequency equal to the sampling frequency ($fs = 1/\Delta t$).

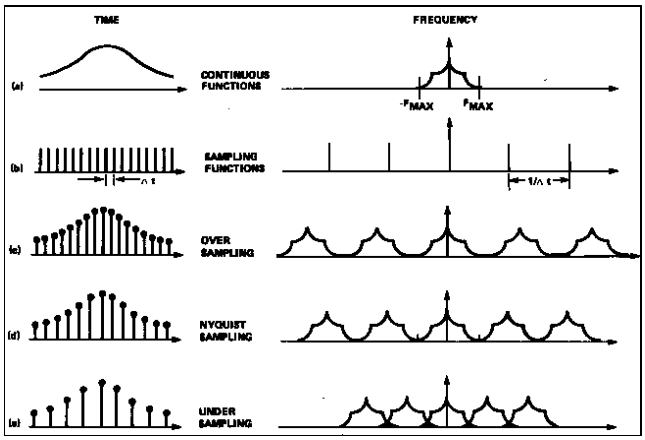

Figure 8 - The effect of sampling.

Applying the windowing rule, we see that multiplication of the continuous signal by a SHAH function in the time domain results in a replication of the Fourier transform of the signal in the frequency domain, due to the convolution of its transform with another SHAH function.

Figure 8a shows the spectrum of a "band limited" signal, i.e. its spectrum (Fourier transform multiplied by its conjugate) contains non-zero values only in the interval ($-F_{max}$, F_{max})

Figures 8c, 8d and 8e show the effects of sampling the time domain signal at slower and slower rates. It is clear that as t becomes larger, 1/t becomes smaller and the replicated Fourier transform of the signal begins to overlap upon itself, as shown in Figure 8e. This overlapping of the frequency domain function upon itself is called *aliasing*.

In practice two different methods can be used to prevent aliasing

1. Sample a known band limited signal at a high enough rate. If a signal is known to be band limited in the interval ($-F_{max}$, F_{max}), then sampling it at a rate $fs = 2F_{max}$ is sufficient to prevent aliasing. F_{max} is called the Nyquist rate.

2. If the bandwidth of the signal is unknown, band-limit it by passing it through an analog low pass (antialiasing) filter before sampling it. Then follow step 1 above.

Most antialiasing filters can be characterized by a line shape in the frequency domain as shown in Figure 9. Their rolloff rate varies with design but most commercially available filters will adequately bandlimit a signal if their cutoff frequency (f_c) equals $F_{max}/2$. Hence a safe rule of thumb for preventing aliasing in a signal of unknown bandwidth is to analog filter it using a cutoff frequency of $F_{max}/2$ and sample it using a sampling frequency of 2 F_{max}. Useable data is then in the range ($-F_{max}/2$, $F_{max}/2$).

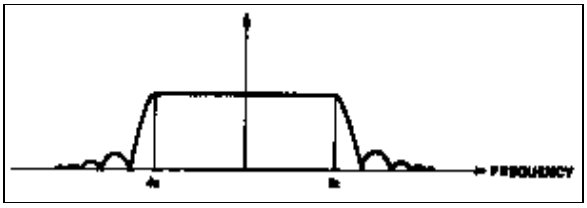

Figure 9 - Antialiasing filter line shape.

Frequency Domain Sampling - The last of the three processing steps is to multiply the frequency function by some suitable sampling function in the frequency domain. This step is necessary so that we can evaluate the Fourier transform with a finite number of calculations, and so that we can store the results in a finite digital memory. We know that this technique of frequency sampling is permissible, provided we sample at uniform intervals no greater than $\Delta f = 1/T$, because this implies that the corresponding time function will be replaced with a period no smaller than T. Since our time window width is T, this sampling interval will insure that no time overlap occurs during this replication step.

Quadratic Frequency and Correlation Domains - We are often interested in measurements involving the product of two frequency functions, either in the form of an auto-power spectrum, or a cross-spectrum between two related quantities. As discussed earlier, the inverse Fourier transforms of these frequency quantities are called correlation functions.

Applying the windowing rule once again, multiplication of two functions together in the frequency domain causes convolution of the inverse transforms in the time domain.

The convolution process for the discrete transform differs from the continuous transform in that it is *circular* in nature. That is, we are always convolving periodic functions, so the convolution with one period of a function overlaps that of adjacent periods. It might be helpful to visualize one period of each function wrapped around a cylinder of the appropriate radius. All displacements associated with the convolution operation can be pictured as incremental rotations of one waveform with respect to the other around the circumference of this cylinder. Thus, as a portion of a function moves past the t = T boundary of the time window, it immediately reappears at t = 0.

In Figure 10, we see that a rectangle of width T. convolved with itself, produces a triangle of width 2T. However, when we sample the resulting spectrum at intervals of 1/T, we replicate this composite time function at intervals of T. and the sum of these multiple images is a constant. This is obviously quite different from the true triangular shape. The "tails" of the expected triangle are "wrapped-around" the cylinder, and fill in the remaining portion of the triangle, producing a rectangle.

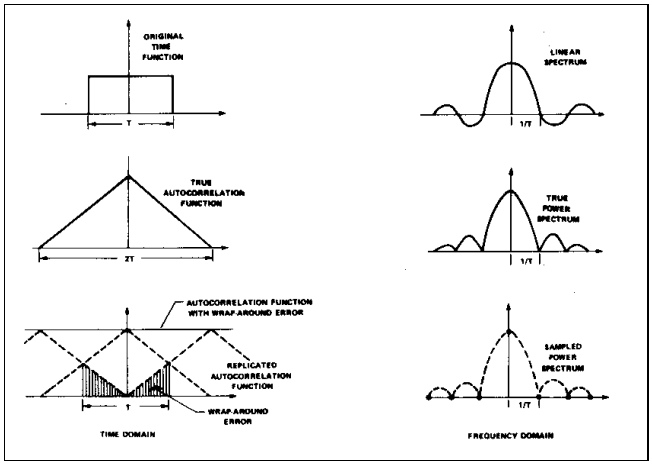

Figure 10 - The wrap-around effect caused by sampling a quadratic frequency spectrum at 1/T intervals instead of 1/2T intervals

When overlap of periodic time functions occurs this is called *wrap-around* error. This is completely analogous to aliasing in the frequency domain and is sometimes called time domain aliasing. As with aliasing, this phenomenon is caused by undersampling the product of two functions which have been multiplied together in the frequency domain. If some wrap-around does occur, it means that the multiplication of two frequency domain functions has produced a frequency spectrum that cannot be adequately represented by samples spaced 1/T apart. Generally speaking, the product of k functions can introduce additional fine detail into the result which requires finer sampling by a factor k to completely represent the product.

In many cases, the sum of the widths of the individual time functions is effectively less than T, so when these functions are convolved together, the composite width is also effectively less than T, and the wrap-around effect is absent, or at least negligible. Otherwise, this phenomenon causes significant errors in the convolution or correlation process, particularly near the ends of the time interval.