Behavior Research Methods & Instrumentation
1983, Vol. 15(2), 228-241

SESSION VII
AN INTRODUCTION TO THE THEORY AND APPLICATIONS OF FAST FOURIER TRANSFORMS

Correlations, contrasts, and components: Fourier analysis in a more familiar terminology

HOWARD L. KAPLAN
Addiction Research Foundation, Toronto, Ontario M5S 2S1, Canada

The Fourier transform partitions the energy in a waveform into the sum of the energies of simpler components. This process is the same as the partitioning of variance into linear contrasts and is a way of measuring the correlation between the waveform and each member of a family of prototype model waveforms. Such a partitioning will often, but not always, result in a meaningful decomposition of the original waveform.

The mathematics of the fast Fourier transform (FFT) are often considered arcane or mysterious. While the formulas and computer programs used to implement FFTs are often complex, taking obscure shortcuts in order to reduce computation time, the fundamental principles are already familiar from elementary statistics, but they are sometimes not recognized in a different context: They are the same principles used in calculating correlation coefficients, best-fitting curves, and analysis-of-variance contrasts.

CORRELATIONS AND CONTRASTS

To begin, we will need a few definitions. A vector is an ordered list of numbers. For example, in an experiment with 12 treatment conditions, the treatment means vector is simply the ordered list of treatment means. We will be concerned with three types of vector. Data vectors consist of raw or averaged data, such as the treatment means vector. Prototype vectors consist of arrangements of numbers to which we wish to compare data vectors. A prototype vector represents a hypothesis or model of the shape of a data vector. For example, we might want to compare our treatment means vector to a prototype consisting of 12 increasing, equally spaced numbers, in order to evaluate the model that our data consist of a steady increase or decrease. That is the same as comparing our data to a graph that is a straight line. Alternatively, we might wish to compare our data to a hypothetical model that says that the first six elements are all equal to some constant and the last six are equal to a different constant. Because these two constants must be equally far from the grand mean in different directions, this hypothesis says that the data vector has the same shape as a vector consisting of six −1s followed by six +1s.

Component vectors are multiples of prototype vectors, and they represent that multiple of a prototype that provides the best fit to the data. To the extent that the hypothesis is true, we can add some multiple of that prototype vector to a constant and get 12 numbers that are close to the 12 terms of our actual data vector. When we compute the slope and intercept of the best-fitting line from a correlation and apply that slope and intercept to each value of our predictor variable, the resulting set of numbers is a component vector. Figures 1-4 show, in graphic form, a data vector, the prototype vectors corresponding to the two hypotheses above, and the fit of the second hypothesis to the data, resulting in a component vector.

Let us look at an example of a data vector with 12 elements:

10 13 12 14 16 19 25 24 27 32 29 35

and compare it to this typical straight-line increase:

1 2 3 4 5 6 7 8 9 10 11 12

A straightforward, naive approach would be to ask how

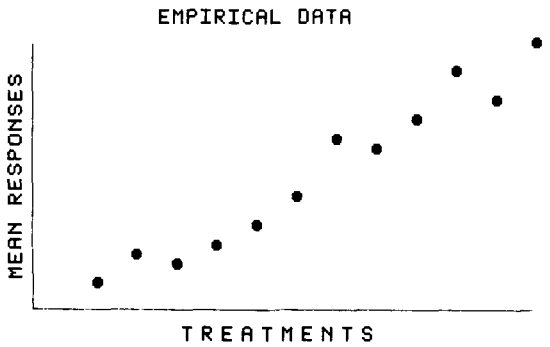

Figure 1. A data vector is an ordered set of responses or response means. One purpose of analysis of variance is to represent such a vector as a sum of multiples of prototype vectors.

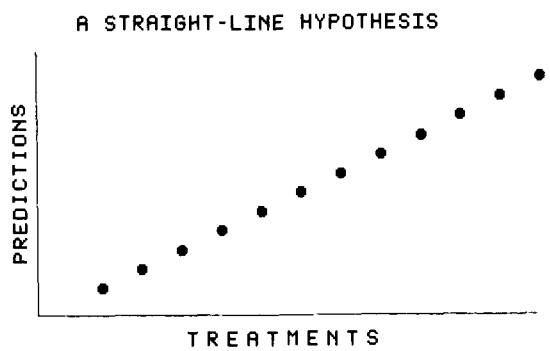

Figure 2. A common model used in representing data vectors is that of a linear increase or decrease, as shown here.

well the data correlate with the prototype vector. As part of the calculation, we would build up a sum of cross-products:

$$(1*10) + (2*13) + (3*12) + \ldots + (12*35).$$

After adjusting for the fact that neither vector has a mean of 0 or a standard deviation of 1, the size of this sum of cross-products would become the correlation coefficient. A large coefficient means a strong resemblance between the data and a straight-line increase. Beyond simply calculating the correlation, we can actually produce the equation for the best-fitting prediction line. The 12 dependent values predicted by that equation then form a component vector, the best-fitting estimate of the part of the variance (or information) in the data that can be represented as a straight-line increase. If the correlation is high, then that line provides a good approximation to the actual data graph.

Someone more comfortable with analysis of variance might go to a table of linear contrasts and find these coefficients:

$$-11 \quad -9 \quad -7 \quad -5 \quad -3 \quad -1 \quad 1 \quad 3 \quad 5 \quad 7 \quad 9 \quad 11$$

As before, each data point is to be multiplied by the corresponding coefficient, the result is summed, adjustments are made for the standard deviation of the prototype vector and the cell sizes, and the result is squared and used as a test statistic. Again, we could derive that multiple of the contrast prototype vector that, added to the grand mean, gives the best approximation to the data vector. A contrast, in other words, is a prototype vector, or a hypothesis about the shape of the data.

There are important differences between these two approaches, primarily concerning the interpretation of the error term. However, there are also fundamental similarities in the two approaches. If we take the second prototype vector (−11 to +11), add 13 to each term, and multiply the result by .5, we get the first prototype vector (1 to 12). However, the addition and multiplication by constants (which are nothing more than a change of measuring scale) are operations that do not change the correlation coefficient, and thus the two approaches are computing the same information. The primary difference is that the linear contrast approach also uses some additional information, the within-group variance, in assessing the statistical significance of that extracted information. If we forget about the hypothesis-testing part of analysis of variance and concentrate only on its role in describing a data vector in terms of prototypes

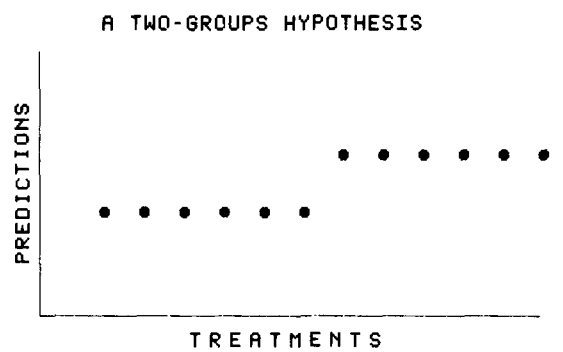

Figure 3. Another common model used in representing data vectors is that of a constant response within each of several treatment groups. In the model, the magnitude of the group difference is arbitrary, as the model is multiplied by the best possible constant when fitting it to the data.

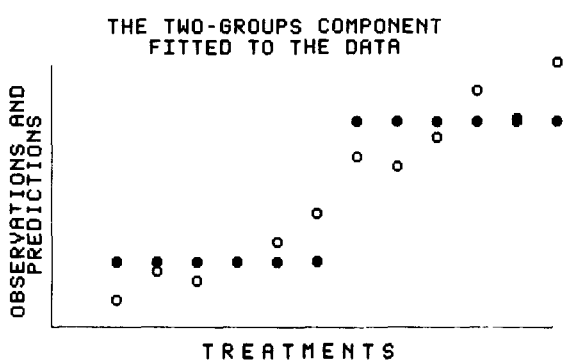

Figure 4. The two-groups hypothesis, shown as dark circles, provides a reasonable representation of the data, shown as open circles. The average discrepancy, or residual, between the data and the model is considerably smaller than the average discrepancy between the data and the grand mean.

or components, then we have two ways of performing the same step in such an analysis. Emerson (1983) provides further information about formal tests of significance of these hypotheses. At this stage, we are more concerned about measuring the degree of resemblance between the data and various hypotheses than about conducting formal significance tests.

Let us suppose that we want to consider the hypothesis that the data have a parabolic shape, a symmetric bowl with a minimum at the center of the vector (Figure 5). Once again, we could perform a naive correlation with any parabolic curve or we could look up the quadratic contrast coefficients for a 12-observation series. The latter coefficients are simply a typical parabola chosen to make certain computational adjustments convenient, but again, the contrast is really only asking how well the data correlate with a parabolic prototype. All polynomial contrasts are the same, nothing more than correlations with curves generated by polynomial terms.

Any one contrast provides a measure of the resemblance between the data vector and one prototype and yields a component vector with the best least squares fit to the actual data vector. It is quite rare when a contrast can exactly fit the data vector. There are almost invariably differences between the component, or model, and the data. If the original data vector is replaced with the differences, or residuals, between itself and the prediction, we have a new data vector consisting of information that could not be represented by the terms of the model, or prototype vectors, introduced so far. In many cases, it is necessary to account for those differences with more components, which represent resemblances between the data and additional prototypes or models. In order to avoid continually readjusting for the effects of the grand mean, contrasts are generally rescaled to have a mean of 0 and are applied to the data vector after it has had its own grand mean subtracted from each element, resulting in a new grand mean of 0. In other words, we generally begin by fitting the data to a vector of all +1s, a model that says that the data all equal the grand mean. Once this is done, only a multiplicative adjustment needs to be made to rescale each new contrast as it is fit to the residuals.

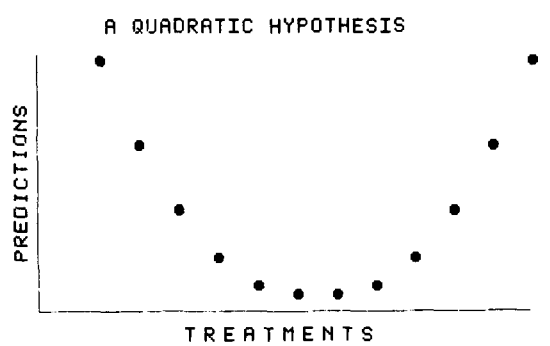

Figure 5. A third model commonly used in representing data vectors is a quadratic trend, or a part of a parabola. Higher order polynomial trends are also often used.

Whenever two or more contrasts are applied to the same data, traditional analysis of variance requires them to be orthogonal. That is, if any of the contrasts is applied to another contrast's coefficients (instead of to the data vector), the resulting sum of cross-products (or correlation) must be 0. This is the definition of orthogonality.

What are the implications of orthogonality when analyzing a data vector? In informal terms, it ensures that any two contrasts are extracting uncorrelated, or nonoverlapping, information from the original data. That is, it guarantees that the order in which the contrasts are applied is immaterial: Each contrast will yield the same result, the same component vector, whether applied to the original raw data or to the residuals left over after applying any or all of the other orthogonal contrasts in the set being used. More formally, orthogonality guarantees the additivity of sums of squares. If we have two orthogonal contrasts, then we can add their associated components together, term by term, to get a better prediction of the data than either component alone could give. Furthermore, the sum of squares associated with that combined component is equal to the sum of the two separate sums of squares associated with the original contrasts or components. The converse is also true: The sums of squares for two prototypes will add to the sum of squares for the sum of the prototypes only if the two are orthogonal. If a complete set of $n - 1$ orthogonal contrasts is applied to a data vector, then the sums of squares from the various contrasts will add to the sum of squares for differences among treatment means, that is, be a complete partition of the information about such differences. (For those familiar with the terminology, any $n - 1$ orthogonal contrasts form an orthogonal basis for the $n - 1$-dimensional space of n-element vectors all of whose grand means are 0.)

Given any vector of more than two data points, there are literally an infinite number of families of contrasts that can be used to extract information about differences among the elements. Although each such family contains all of the information, not all extract it with equal usefulness for a particular experimental design. In a one-way treatment group design, the usual contrasts emphasize a hierarchy of differences among groups: placebo vs. drug, low dose vs. high dose, and so on. Even when the groups differ on a qualitative dimension, such as drug administration route (oral vs. injected vs. inhaled), it is still possible to label the groups, say, 1 to 4, and extract linear, quadratic, and cubic trends. The information so obtained is generally meaningless to us, but none has been lost, and the components could be used to reconstruct the four group means. Use of polynomial contrasts would more typically be useful

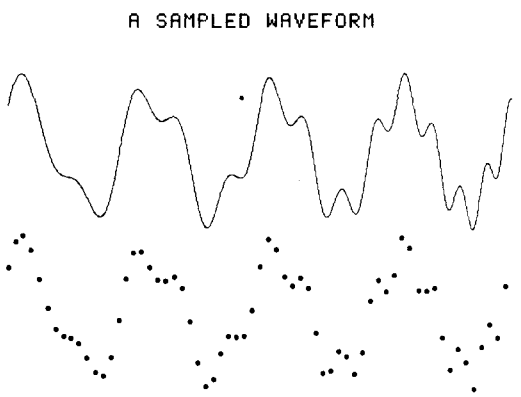

Figure 6. A continuous process is often sampled at discrete points in time, generally equally spaced. A spatial dimension may also be the abscissa. Typical ordinates for physical waveforms are voltage, pressure, and displacement, although any variable that can be measured on an interval scale may be used as the ordinate.

in psychophysical work, in which the response is often a polynomial function of a physically measurable stimulus property. Even when there is no theoretical reason for a response curve to be a polynomial function of the stimulus, the concepts of linear and quadratic trend are often concise descriptions of the shape of the response curve and can be usefully employed. Although an analysis of variance may reveal that deviations from linearity are significant, a linear approximation to the data often provides a good rule of thumb for predicting the response.

While experimental treatments in an analysis-of-variance framework are generally discrete values of some independent variables, we must often deal with a continuous independent variable, such as time. In such a case, the dependent variable cannot be measured for all values of the independent variable, but instead we must sample the response at specific points in time. Figure 6 shows a continuous process being sampled in such a way. Although there are techniques for dealing with irregular sampling intervals, in this paper we will be concerned only with sampling at fixed intervals, such as 256 times/sec or 1 time/year. It is data like these, a continuous process sampled at regular intervals, that are often explained best by a family of contrasts other than the familiar polynomial contrasts.

THE FAST FOURIER TRANSFORM AS A FAMILY OF CONTRASTS

Some sets of observations are believed to be samples of fundamentally periodic processes. For example, if the behavior of a marine animal is sampled every 30 min, cycles with a period of 24 h may occur, indicating diurnal variation, or cycles with a period of 12 h 25.5 min may occur, suggesting variation with the tides. Over a short time, less than one cycle, a linear or quadratic contrast may account for most of the variance in the data, but over many cycles it is increasingly difficult to provide a meaningful account of the data with polynomial contrasts. Instead, what is needed is a comparison of the data vector with a fundamentally periodic prototype vector, such as a sequence of points sampled from a sinusoid at points spaced equally in time. Figure 7 shows a linear, a cubic, and a sinusoidal prototype for 36-point vectors. Note particularly that while the cubic prototype provides a reasonable fit to a response that waxes and wanes near the center of the range of times sampled, it becomes unbounded rather than cyclic beyond the range shown. The sinusoid, on the other hand, continues to copy its central shape indefinitely in both directions along the abscissa.

What, exactly, is a sinusoid? It is a continuous function, such as the ones shown in Figure 8 and sampled in Figure 7. Formally, there are numerous possible defini-

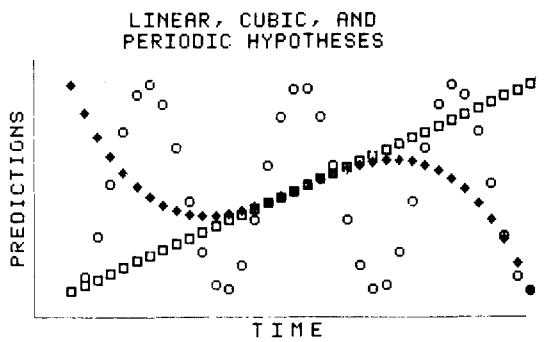

Figure 7. In a local region, polynomial hypotheses such as linear (open squares), quadratic (not shown), and cubic (filled diamonds) may provide good approximations to a fundamentally periodic process. Over several cycles, however, only periodic prototypes, such as sinusoids (open circles), can provide an adequate account of the data.

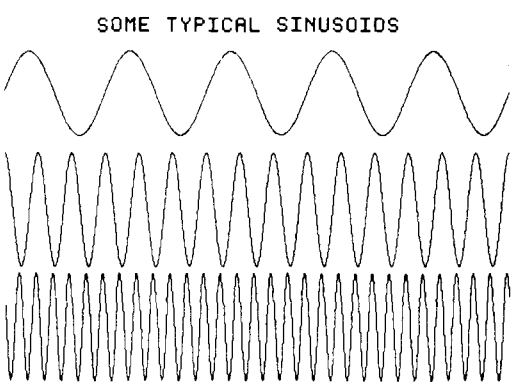

Figure 8. Sinusoids are a familiar shape in science and engineering. In the lower two waveforms in this figure, and in some of the smaller waveforms in subsequent figures, the grain with which the plotting hardware is able to represent sinusoids makes them appear to be composed of many straight-line segments, rather than appearing as smooth curves. This is an example of the kind of distortion that can be introduced by sampling a continuous process. The process by which we interpret these figures as sinusoids is analogous to filtering out the higher frequency, distortion components.

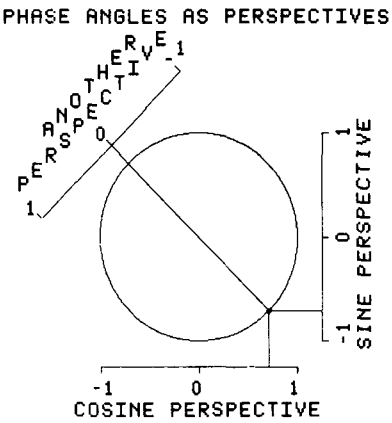

Figure 9. As the point proceeds counterclockwise around the circle at a fixed rate in time, its projection onto any axis is a sinusoid as a function of time. The amplitude is equal to the radius of the circle, and the frequency is equal to the number of revolutions the point makes per unit time. The inverse of the frequency can also be used to characterize the waveform. If the abscissa is time, then the inverse is the period; if the abscissa is a spatial dimension, then the inverse is the wavelength.

tions, some algebraic and some geometric. The best one for our purposes is that a sinusoid is a perspective view, or a projection, of a point moving around a circle at a fixed rate (Figure 9). It is conventional to let the motion be counterclockwise, and to let the point be at the extreme right (3 o'clock) at Time 0, the starting time. If that is done, the point's projection onto the ordinate (the sine perspective) is going through 0 toward +1 at Time 0 and is moving upward at its maximum velocity. Such an event is called a positive-going zero crossing. At the same time, the point has just reached its rightward peak position on the other (cosine) perspective shown here and is about to begin moving downward from its maximum. Since a circle projects the same image in all directions of the plane, both perspectives will see the exact same cyclic process as the point revolves. However, the timing of the processes will differ for the two views. Everything seen by the cosine perspective will be seen by the sine perspective .25 cycles later.

A number of interrelated terms can be used to measure sinusoids. The diameter of the generating circle becomes the peak-to-peak amplitude, or difference between the largest and smallest values attained by the function. The amplitude (without the "peak-to-peak" prefix) is equivalent to the radius of the generating circle and is the most commonly used measure of the magnitude of a sinusoid. When a standard sinusoid is needed, for example, as a periodic prototype, an amplitude of 1 is generally chosen. It is also reasonable to speak of sinusoids whose amplitudes are negative; these are simply sinusoids whose ordinates are below 0 whenever the standard sinusoid is above 0. The frequency of any sinusoid is the number of revolutions of the generating point about its circle per unit time, or number of cycles of the sinusoid per unit time, and the time required to complete one cycle is its period.

In addition to the two standard sine and cosine perspectives on the point, we can also consider other ones, such as the one that sees the point moving down from −.3 at Time 0. Because a circle is a two-dimensional figure, any other such perspective can be represented as a linear combination, or weighted sum, of two fundamentally different perspectives, such as the sine and cosine perspectives shown. Because the sine and cosine perspectives are geometrically orthogonal, we can compute the weights for such a sum by drawing right triangles (Figure 10). The triangle side lengths and the weighting coefficients are both proportional to the amplitudes of the waveforms. The sine and cosine perspectives are the two orthogonal legs; any other perspective is the hypotenuse, and the angle of the axis onto which the rotating point must be projected is the angle of the hypotenuse. This process converts rectangular coordinates, sine and cosine projections or components, into polar coordinates, radius and angle. If we know the amplitudes of sine and cosine components, then the Pythagorean theorem allows us to compute the amplitude of the resultant sum.

The difference in time between various perspectives is measured in terms of fractions of a cycle of the underlying revolution and called "phase." We can speak of the cosine perspective as being 90 deg, or .25 cycle, ahead of the sine perspective, as each ordinate point on the sine occurs that much earlier on the abscissa of the cosine. We can also describe the cosine as being 270 deg behind the sine in phase, as each sine ordinate point will occur .75 cycles later for the cosine. Any two sinusoids 180 deg out of phase are negative multiples of

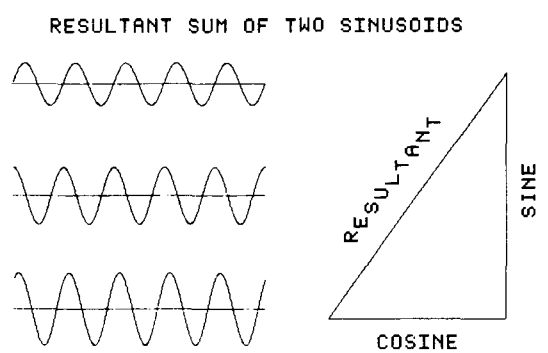

Figure 10. The sum of two sinusoids at the same frequency is a resultant sinusoid at the same frequency, with amplitude and phase determined by the length and angle of the third side of a triangle representing the vector sum of the component sinusoids. When the two are 90 deg apart, as in sine and cosine waveforms, the relevant triangle is a right triangle and the resultant is represented by the hypotenuse. In all cases, the triangle can be rotated so that the angles of the three sides correspond to the axes onto which a revolving point is projected. In other words, there is a very natural mapping between the family of sinusoids at a given frequency and a two-dimensional real vector space.

each other along the ordinate, and any two sinusoids 360 deg (or any multiple of 360) out of phase are positive multiples of each other. We can speak of a sinusoid as being in sine phase, being in cosine phase, being 180 deg out of phase with sine, being 10 deg ahead of cosine, and so on. If no reference phase is stated, then a waveform's phase angle is with respect to sine phase. In summary, we can specify a sinusoid of any frequency by either of two pairs of other statistics: by the amplitude of the unique sine and cosine components whose sum is the waveform, or by its amplitude and its phase with respect to sine.

Given a periodic process that is not a simple sinusoid, we might wish to decompose the process into the sum of sinusoidal components (Figure 11). This, for example, is what a listener does when he hears the simultaneous playing of several tuning forks not only as a chord but also as separate notes. Why do we use sinusoids as our basis for describing complex periodic waves, rather than decomposing them into sums of square waves, triangular waves, or other apparently simple periodic processes? There are several answers to this. One answer is that many physical systems are approximately linear, meaning that the response to the sum of two inputs is the sum of the responses. It can be shown that whenever a sine wave is used as input to a linear system, the output is a sine wave of the same frequency. The same is not necessarily true if the input waveform has any other shape. Therefore, the response of such systems can be characterized more simply by detailing their response to sine-wave inputs than to any other family of inputs, and that is a major reason why the sine wave is so fundamental to the analysis of periodic processes. Another reason for using sinusoids is that they make it easy to decompose a waveform, regardless of the phases of its components. In the representation of a sinusoid as a projection of a revolving point, all views are the same except for phase. If we represent any other periodic waveform as projections of a point moving along the boundary of any other shape, we can no longer make that claim. For example, if a point moves along the perimeter of a square, then there are four views directly at one side, four views directly at one diagonal, and so on. The shape of the waveform, not only its starting point, depends on the viewing direction, and thus we cannot describe an arbitrary linear combination of two waveforms at the same shape and frequency as being the same waveform, except for its starting phase: It is a different waveform. That is another reason why sinusoids are typical components for decomposition of a periodic process.

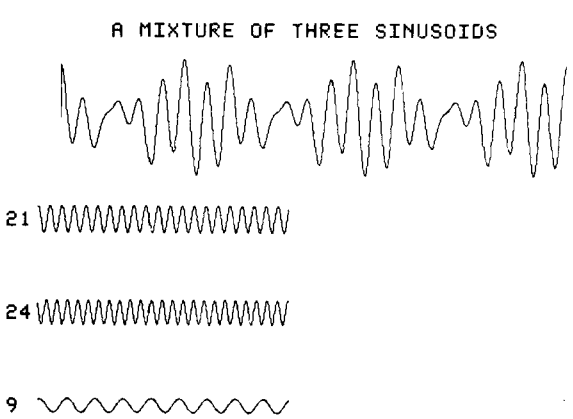

Figure 11. The three sinusoids shown here in smaller scale, at frequencies of 21, 24, and 9 cycles, sum to form the sinusoid shown in larger scale. As the greatest common divisor of the three frequencies is 3, the resultant sum has a fundamental frequency of 3; that is, it repeats three times for each 21 repetitions of the first component, and so on.

The most important reason for using sinusoids as prototypes in analyzing periodic events is that sets of orthogonal sinusoids are very straightforward to generate. Although periodic events theoretically continue indefinitely, they are generally measured over a finite sampling duration, such as 1 sec, 6 min, or 200 years. Some periods exactly divide such a sampling duration, and others do not. For example, if the sampling duration is 1 day, then events with periods of 12 h, 6 h, and 30 min all occur an integral number of times during the day, whereas events whose periods are tidal (12 h 25.5 min), 5 h, or 17 min do not occur an integral number of times each day. Over any sampling duration, two sinusoids with different frequencies, both of which exactly divide the sampling duration, are always orthogonal. For example, over 1 day, sinusoids with periods of 3 and 4 h are orthogonal. This orthogonality holds for any regular sampling frequency that also divides the day. For instance, if the two sinusoids are sampled once per hour, a 24-point vector is obtained for each; and if they are sampled every 12 min, a 120-point vector is obtained for each. In both cases, the vectors are orthogonal. As some sophistication in mathematics is required for a formal proof of the assertion that two sinusoids whose different periods both exactly divide the sampling duration are always orthogonal, no such proof will be provided here.

Given a sinusoid at any frequency, there is essentially only one other sinusoid at that same frequency that is orthogonal to it. These two sinusoids have the relationship of sine and cosine or are .25 cycle out of phase with each other. This fact makes it very easy to construct a complete set of orthogonal, periodic contrasts for use in decomposing any set of points sampled at equal intervals in time: Use the sine and cosine components corresponding to periods that divide the sampling frame. The lowest frequency is equal to the total sampling duration, and the highest frequency has a period equal to two sampling intervals. The total number of such components so generated is the number of degrees of freedom among the points of the vector, or one less than its length. (Sine and cosine collapse into a single component for the highest frequency, whose period is two sampling intervals long.) For 20 points sampled over 1 sec, we can measure all of the periodic components up to 10 Hz (cycles per second) using

the 19 prototypes of sine and cosine components at 1, 2, . . . , 9 Hz, and the one component at 10 Hz, all orthogonal to each other. (For those who understand such terminology, it can be stated that the complete set of sinusoids so generated forms an orthonormal basis for the space of data vectors' residuals around their grand means.)

In dealing with a set of points that represent a waveform sampled at equally spaced points in time, what statisticians call variance, or sum of squares, is what engineers call energy. Except for some scale factors that depend on the physical system being considered, the energy in a waveform is measured by the sum of squares of its deviations from its mean value. If we divide a waveform into components, we may be able to understand its structure better. If we divide it into orthogonal components, then the energies of the components sum to form the energy of the original waveform. This makes the fraction of a wave's energy at a given frequency analogous to the proportion of variance accounted for by a pair of contrasts. One frequency is analogous to a pair of contrasts, instead of just one, because the energy at any frequency may be associated with two components, in sine and cosine phase. In turn, the energy of those two components is additive because they are orthogonal. The same additivity rules hold for power (energy per unit time), another common term from engineering.

Thinking either in terms of energy or in terms of sum of squares, we can partition the information about variance in any sampled waveform by using orthogonal sine and cosine components. Just as any two orthogonal contrasts exhaust the information in a three-group analysis of variance, so do the two orthogonal components, sine and cosine, exhaust the information at one frequency. To extract the information at other frequencies, we can compare the data to sine and cosine prototype vectors at those other frequencies. If there are 2n data points, then a complete set of orthogonal contrasts consists of sine and cosine waves with periods of 2n, 2n/2, 2n/3, . . . , 2 data points.

Components whose frequencies are integral multiples of some base frequency (or whose periods are integral divisors of some base period) are called harmonics, and the base frequency is called the fundamental. Therefore, the components of a complete partitioning of energy are the fundamental frequency (one period for the entire data vector) and all of its harmonics up to the harmonic for which each cycle is only two data points long. This way of partitioning the information in a set of data points, analyzing it by contrasts that are harmonics of a fundamental sinusoid, is called a discrete Fourier transform. That is, the Fourier transform of a set of data points is nothing more than the analysis of those data by a set of contrasts, or correlations, with sinusoidal prototypes, and their subsequent representation as a sum of components that are proportional to those prototypes.

A convenient and standard way of graphing such an analysis of energy is in the form of a spectrum. A spectrum is a graph whose abscissa is frequency and whose ordinate is energy at that frequency. Newton's prism is just an optical device for forcing the component energies in light to bend by different amounts, therefore spreading themselves along a physical abscissa. The spectrum of a pure sine wave consists of a graph that is zero everywhere except at the frequency of the wave, just as the spectrum of light produced by a laser consists of a single thin bar. In many applications, the phase information in a spectrum is uninteresting, and only the energies of the various components are shown. In this paper, the vertical bars that represent the energy at various frequencies will be shown divided by a horizontal line along the abscissa. The total height of the bar is the total energy at the frequency, and the portions above and below the horizontal line represent the proportions of that energy in the sine and cosine phases. Figure 12 shows a waveform and its spectrum, and Figure 11 shows the three component frequencies (in a smaller scale), along with their sum, which exactly reproduces the waveform. Clearly, if a waveform must be decomposed into a large number of components, the spectrum is a more efficient form of display.

It is easier to believe that some waveforms are sums of sinusoids than to believe that of other waveforms. Figure 11, for instance, shows a waveform that clearly looks like the sum of sinusoids, and its spectrum shows that it is the sum of three sinusoids. One period of a sawtooth waveform, however, does not look much like a sum of sinusoids, which are smooth, whereas the sawtooth has angles. Figure 13 shows such a waveform and its spectrum, which actually contains an infinite number of components. As we include more and more of the components from this infinite family, the result bears an increasingly greater resemblance to the original waveform. Figure 14 shows the sum of the six largest components of the waveform. The approximation is still visibly different from the original, but it is also

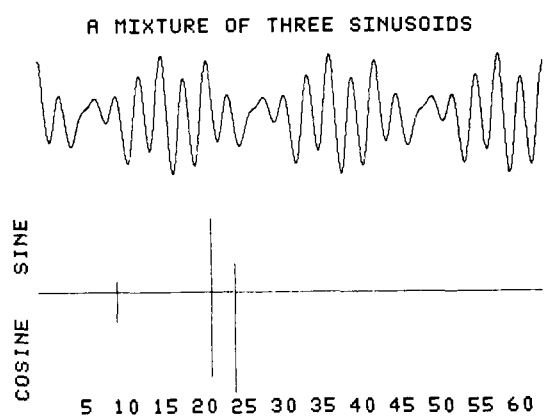

Figure 12. The sum of three sinusoids has a spectrum consisting of energy only at the frequencies of its three components.

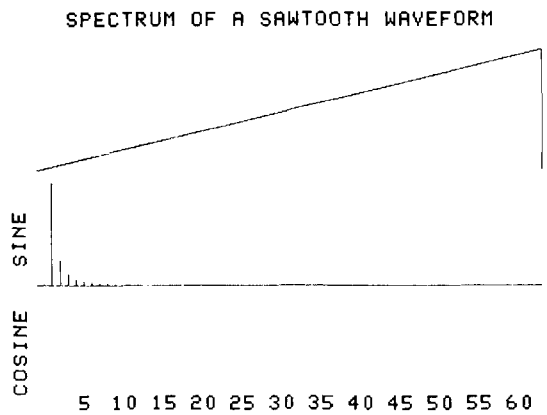

Figure 13. A sawtooth waveform that completes one cycle in the space shown here has components of decreasing amplitude at the successively higher harmonics of the fundamental frequency.

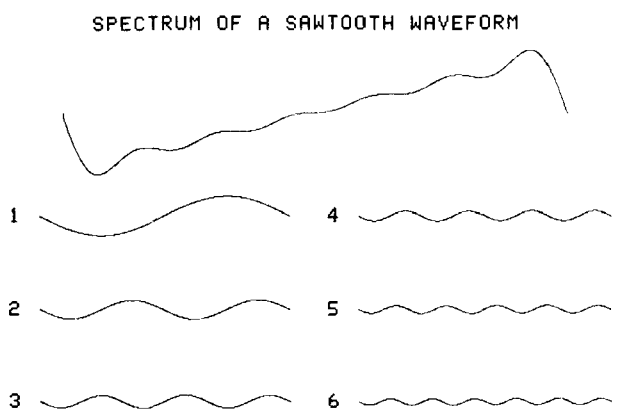

Figure 14. The sum of the first six components of the sawtooth wave shown in Figure 13 provides a reasonable approximation to the shape of the waveform. The corners are noticeably rounder in the sum than in the original; this is because corners are basically areas where the effects of higher frequency components are most evident.

clearly getting quite close. If we are dealing with a sampled waveform, having only 2n points in the sample vector, then we can always sum the grand mean and 2n – 1 periodic components, as above, to reproduce the original waveform at all of the 2n points that were sampled. That is, the restriction of the sampling to 2n points makes it impossible to detect any components whose period is shorter than two sampling intervals. This relationship between the sampling interval and the maximum detectable component frequency (or minimum component period) is known as Nyquist's theorem.

The discrete Fourier transform of 2n data points can, in theory, be calculated by performing 2n – 1 correlations with prototype vectors. In practice, very few Fourier transforms are so calculated. Prior to the computer age, Fourier transforms were rarely calculated on actual data; instead, the results were derived analytically for certain waveforms of interest in engineering. For example, a square wave can be shown to consist of only odd harmonics of its fundamental frequency. Just as most calculus problems are solved by using analytic transforms rather than by actually dividing up a graph into many narrow rectangles, so Fourier transforms were, for many years, performed by similar analytic tools that bypassed the need for actual computation of sums of cross-products. However, with the advent of fast computers for both collecting data vectors and analyzing them, it has become both practical and important to devise efficient ways of performing the calculations on actual data, rather than just on theoretical waveforms.

A fast Fourier transform (FFT) implements the calculation of a discrete Fourier transform by using some of the redundancy inherent in the calculation of the sine and cosine prototype vectors. For example, in calculating the vectors for frequency k, all of the numbers sine(kt) and cosine(kt) will be calculated for t = 0, 1, . . . and will be multiplied by their corresponding data points. Later on in the calculations, for twice that frequency, we will need sine(2kt) and cosine(2kt) for the same sampled data points. However, if we use trigonometric identities, we can see that sine(2kt) = 2*sine(kt)*cosine(kt) and that cosine(2kt) = cosine(kt)* cosine(kt) – sine(kt)*sine(kt). In other words, most of the work of calculating the prototype for the higher frequency, and of multiplying it by the data vector, has already been done at the lower frequency. By using trigonometric identities such as these, much of the work of calculating the cross-product sums with the higher frequencies can reuse calculations from lower frequencies. There are many FFT algorithms, but all are variations on this theme of reusing the calculations for some frequencies in order to save work at other frequencies. It happens to be the case that many of the algorithms work most efficiently when the number of data points is a power of 2, and that is why one often reads about 4,096-point transforms rather than 4,000-point transforms, and so on. As an example of the speed at which modern computers can calculate an FFT, the 5-year-old laboratory minicomputer that I generally use can calculate a 1,024-point transform in 600 msec. A modern number-crunching computer, such as those in most university computer centers, would be orders of magnitude faster. It is interesting also to note that the speed of integrated circuit multipliers is increasing so rapidly that multiplications no longer monopolize the time of a frequency analysis, and some recent hardware is saving time and money on deciding which multiplications to do next and simply doing the redundant, complete set of correlations with raw sinusoids.

WHAT CAN WE LEARN FROM A FAST FOURIER TRANSFORM?

The FFT characterizes any periodic waveform as a sum of sinusoidal components. Of what use is such an analysis? The primary use of such an analysis is in characterizing the properties that perfect communications channels must have in order to transmit certain

classes of signals, and in characterizing the kinds of information that imperfect channels remove from their signals or add to their signals. The human auditory system, for example, is a communications channel transmitting information about sound waveforms. Let us imagine that we want to answer the apparently simple question, "Can a person hear everything that is happening onstage at a symphony concert?" One interpretation of this question is, "Is the auditory system sensitive to all of the frequencies generated by the instruments of the standard orchestra?" One part of answering this question is a determination of what frequency components are contained in those instruments, and a second part is a determination of whether the observer is sensitive to all of those frequencies. Some tool for measuring frequency content is certainly part of the orchestra measurement, and many contemporary electronic spectrum analyzers do in fact use FFT circuitry to report what frequencies are present in their input. The FFT may or may not be useful in characterizing the observer, depending on what kind of response is being measured. If the response is not an attempt to match the periodic sound input with a similar periodic output, then the FFT is of no use. For instance, if an observer is asked to press a button when he hears a tone in his earphones, then only a frequency synthesizer is needed to measure the observer, not a frequency analyzer. However, if the response being measured is a neural frequency coding, in which signals pass through the nervous system with a frequency matching that of sound input to the ears, then a frequency analysis may be useful in detecting whether the input frequency is present in more than usual proportion among the signals passing through the nerves.

Many electronic communications systems remove some of the frequencies present in their input when producing their output. For example, the telephone does not transmit all of the frequencies present in the voice, but it restricts the ones transmitted in order to fit more voices into the capacity of the communications networks. If we use a signal including components from the upper part of our hearing range as input to a telephone and perform an FFT on the signal actually transmitted, we find that the higher frequencies are not present in the output. The voice quality perceived by the listener is somewhat reduced, indicating that the listener is sensitive to frequencies that the telephone system rejects. The process of removing some frequencies present in the input signal, or of reducing their amplitudes, is known as filtering. Figure 14, which shows the sum of the first six components of the sawtooth waveform, can also be seen as showing what happens to the sawtooth if it is passed through a filter that removes all frequencies higher than that of the sixth component.

Consider a visual pursuit experiment in which an observer simply tracks a point moving back and forth horizontally across the screen of an oscilloscope, and we measure eye position. The point moves back and forth sinusoidally, at a rate of 1 cycle/sec (1 Hz), and subtends a visual angle of 10 deg. We can measure both the input and the output, the stimulus and the response, on a common scale: degrees from center. If the response is an exact copy of the stimulus, then there is no need to measure the ways in which they differ. Whenever tracking is not perfect, we can ask how one might measure the difference between the two waveforms. One method is to obtain a single global measure of the amount of discrepancy. At every sampling instant, the difference in degrees between the target and the direction of sight can be measured, and the mean discrepancy can then be calculated. Because this mean may tend toward 0 when positive and negative differences are averaged, it is more usual to square the differences before adding them and to take the square root of the average. This gives a statistic called the root mean square difference between the input and the output. (This is, in other words, the variance in the output not accounted for by a nonrescaled copy of the input as a component.) One of the major limitations of this statistic is that it does not distinguish among various causes of such a difference. For example, both an attenuation of the output, in which eye position is a perfect sinusoid whose width is not the full 10 deg expected, and a phase shift of the output, in which the eye is always looking where the stimulus was 70 msec earlier, can produce the same root mean square tracking error. One of the advantages of a frequency analysis approach is that it can distinguish among various sources of tracking error.

Figure 15 shows a sample sinusoidal stimulus in which the waveform has sine phase; that is, at the point at which we begin our measurements, the input is at its positive-going zero crossing, or the point at which it is moving rightward at its maximum velocity. Time is the abscissa, and the ordinate represents visual angle left or

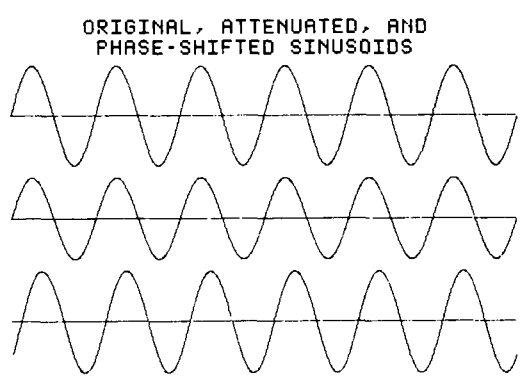

Figure 15. In an experiment in which a subject is expected to copy a stimulus sinusoid with a motor response (such as visual or manual tracking), it is possible to produce a perfect sinusoid that is still an imperfect copy of the original. An attenuated copy (shown in the center) has less amplitude, and a phase-shifted copy reproduces the original with a consistent lead or lag in time.

right. If tracking is perfect, then the response waveform will match the stimulus one, and each will have an FFT consisting of a single component at the stimulus frequency in sine phase. In addition, the mean of the response waveform will be 0 deg, that is, properly centered. One possible way for the subject to fail his task is to produce a perfect sinusoid matching in time, but with less amplitude than in the original. If the stimulus travels 10 deg to either side of center, the subject may look only 8 deg to either side of center. This effect is called attenuation; it is the opposite of amplification. Such attenuation is shown in the center graph of Figure 15. Another possibility is that the waveform has the proper amplitude but is centered at some point other than 0 deg. This can be called bias, but another name from engineering is that the response has a dc component. The term "dc" is an electrical concept, direct current, as opposed to sinusoidal deviations from that center, or alternating current. Both apparent attenuation and bias can be the result of improper calibration of the response-measuring apparatus. Even when changes in the response voltage that indicates eye position are proportional to changes in visual angle, meaning that the system has no nonlinear distortions, failure to properly estimate the slope and intercept of the voltage-position relationship may lead to the appearance of bias or attenuation.

Even in an improperly calibrated but linear system, we can detect the other three components of failure to track properly: phase shift, distortion, and noise. Phase shift implies that the eye is sweeping back and forth at the appropriate stimulus frequency, but not at the proper time. For instance, it may always be looking where the point was located 20 msec earlier. At 1 Hz, 20 msec is .02 of a complete cycle of 360 deg, or a 7.2-deg phase lag. For a faster stimulus, the same number of milliseconds of phase lag would represent a greater proportion of a cycle, and hence a larger angle of lag, such as 15 deg. The bottom graph of Figure 15 shows a phase lag of less than 90 deg. Phase lag is indicated in the FFT by the presence of a nonzero cosine component at the stimulus frequency; the sum of the sine and cosine components is a sinusoid whose phase is different from both.

The subject's response may be a waveform that is not a pure sinusoid of any phase or amplitude but includes additional components. For example, if the subject always stopped moving his eye when it was 7 deg from center and waited for the target to return to the 7-deg range before following it again (Figure 16), the response graph would look like a flattened, or clipped, sine wave. Such clipping would show up in the FFT as frequency components that are multiples of the fundamental. The added components are not at the fundamental, because if they were, the result would be another sinusoid. If the added components were not multiples of the fundamental, they would not have an identical effect on every cycle, as they would get out of step with the fundamental. For example, any component with a period 1.5 times that of the fundamental might reach its own peak at the peak of one cycle of fundamental and a trough at the peak of the subsequent cycle of fundamental. By a parallel argument, any added components at multiples of the fundamental will add the exact same distortion to every cycle. Thus, any consistent distortion (such as clipping) must arise from the presence of multiples, or harmonics, of the fundamental. By the other side of the argument above, any deviation from a pure wave that is inconsistent from cycle to cycle must be the result of frequency components other than multiples of the fundamental. We can use the term "distortion" to describe systematic deviations from the target, or energy at harmonics of the fundamental, and we can use the term "noise" to describe energy at any other components. Figures 17 and 18 show an example of noise, a component not a multiple of the fundamental, added to a sinusoid.

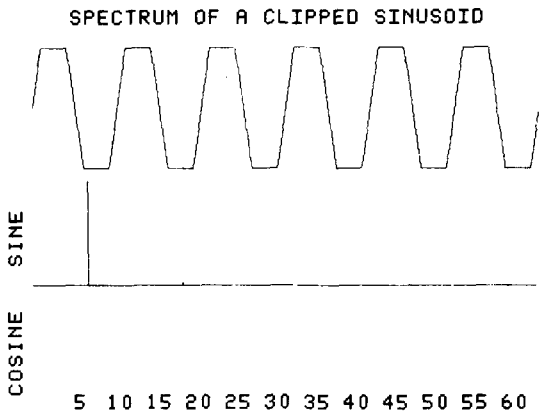

Figure 16. A clipped sinusoid is one in which all points beyond a constant limit are replaced with that constant. Such a sinusoid can be produced by reproducing the sinusoid through an electronic or biological amplifier that does not have sufficient dynamic range. The spectrum of a clipped sinusoid shows small amounts of energy at the odd harmonics of the fundamental; in this figure, only the energy at 18 cycles can be seen as significant, although other components also exist at smaller amplitudes. When it is important to display the relatively small energies of some components in a spectrum, a logarithmic ordinate can be used.

The sine and cosine components at the target frequency can be rescaled as energy and phase of a single sinusoid. The energy of the harmonics of the fundamental gives another measure, and the remaining energy gives a third energy measure. These four measures are all orthogonal, and therefore, the energy at the target frequency, the noise, and the distortion sum to the total energy of the response. We thus have a mathematical method of partitioning the response energy into three independent components, plus an additional statistic representing phase shift. The Human Responses Laboratory at the Addiction Research Foundation once planned such a visual pursuit experiment, intending to use these measures to characterize the tracking response and to see whether any or all changed under ethanol.

Since our basic sampling interval was 10 msec, we planned to use a 1,280-msec, or 128-point, period for our sinusoidal stimulus. Unfortunately, in pilot studies, some subjects appeared to track that frequency perfectly even when under the influence of ethanol, and others could not track it smoothly when entirely sober. That suggested that fixed-frequency tracking should be abandoned altogether and that a stimulus should be used whose frequency changed from .33 Hz to 1 Hz or higher over the course of about 20 sec. Doing this left us with a driving stimulus that was locally almost a pure sinusoid, but which globally covered a wide range of frequencies. Figure 19 shows a graph of such a stimulus and its spectrum. The FFT of such a stimulus is a mess,

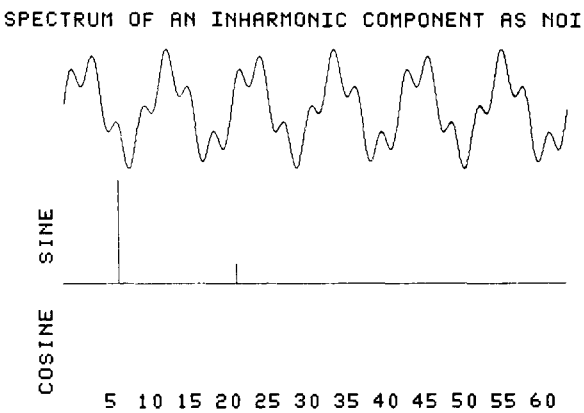

Figure 17. We can consider this waveform to be three exact repetitions of its first third or to be six inexact repetitions of its first sixth. The latter would be the more reasonable representation if we knew that this waveform represented output from a six-cycle sinusoidal input. In that case, we would call the representation "noisy" rather than "distorted." When a copy of a sinusoid differs from the original, any difference that is consistent from cycle to cycle (as in Figure 16) can be called "distortion," whereas any difference that changes from cycle to cycle can be called "noise." The former can be represented as the addition of harmonics to the fundamental; the latter can be represented as the addition of nonharmonic components. As 21 cycles is not a harmonic of 6, this figure shows noise.

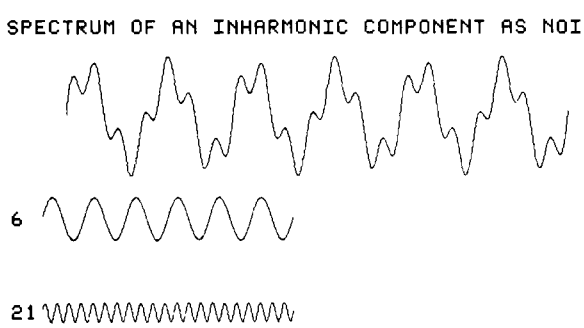

Figure 18. The two components shown here in a smaller scale will sum to form the waveform shown at the top of the figure. Because 21 is a half-integral multiple of 6, the same points at alternate cycles of the 6 are added to peaks and troughs of the 21. This alternation can be seen at the top of the figure, where three peaks have accentuated points and the other three have small dips in them.

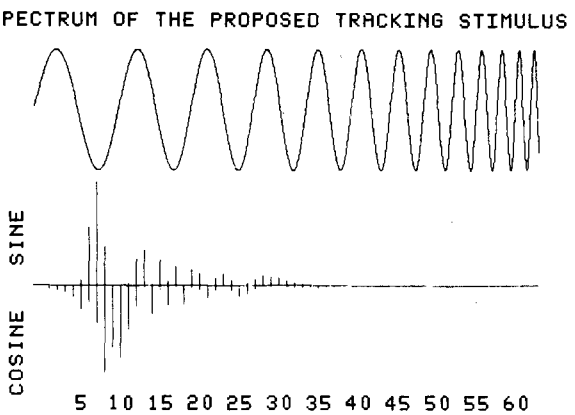

Figure 19. This is a representation of a stimulus once used in an experiment, a sinusoid whose frequency accelerates in time. In any local segment, the waveform is approximately sinusoidal, but considered as a whole, its spectrum shows a large number of significant components. The phases of the components are quite critical; it is possible, for example, to make the waveform run backward in time by reversing the phases (negating the amplitudes) of all of the sine components and leaving the cosine components unchanged. Replacing the phase angles of all components with random angles would result in a sum that does not even resemble a single sinusoid changing frequency, but rather a burst of noise. Clearly, the spectral representation of this waveform contradicts its apparently simple structure. This happens because a single spectrum tries to represent each component of the wave as persisting unchanged infinitely into the past and the future, while to our minds the simplest explanation is that we have a single component whose frequency does change.

suggesting that the transform of the response would be at best such a mess and at worse an absolutely uninterpretable mess. The problem is that it becomes impossible to determine whether any particular component of the response is an accurate copy of one part of the stimulus or an inaccurate copy of another part. It would not be at all easy to perform the analysis of attenuation, phase shift, and distortion at individual frequencies, as described above. To see what happened to our plans to use the FFT, it is necessary to consider some of its more subtle limitations when applied to certain waveforms.

SOME SUBTLE LIMITATIONS OF THE FAST FOURIER TRANSFORM

Over a finite sampling time, any set of data points can be considered as at least one complete repetition of a periodic process. Even if there is no apparent periodicity or sinusoidal component in the data, the FFT can produce a set of sinusoids whose sum is equal to the given waveform. If the sinusoids are extended beyond the measurement duration, they will repeat copies of the arbitrary waveform into the infinite past and future. Sometimes it is sensible to consider such infinite repetition. For example, consider a sawtooth wave consisting of a straight-line increase from −1 to +1 over the course of 10 msec (Figure 13). If an FFT is performed

on this nonperiodic sample of data, it reports the existence of frequency components at 100 Hz, 200 Hz, 300 Hz. and so on. In fact, if we convert the data to a voltage and repeatedly play the result through earphones, an observer will hear a sawtooth wave with a fundamental frequency of 100 Hz and overtones at 200 Hz, 300 Hz, and so on. In other words, any set of data points can be repeated many times in order to create a periodic process, and the FFT then tells us the frequency spectrum of that process.

An FFT will treat any data vector as if it were one or more exact repetitions of a periodic process and will represent that vector as the sum of underlying, sinusoidal periodic processes. Whenever the data vector does not fit those assumptions, the resulting set of components has only limited utility in describing the data. In particular, when the data vector represents a locally near-periodic process that is changing some important characteristic (such as frequency or amplitude) during the sampling time, such a change is represented as the interference among periodic processes at other frequencies, and the phase relationships among the interfering components are very critical in reconstructing a sum with appropriate characteristics.

Figure 20 shows a classic example of change in a periodic process being represented as the interference of two other periodic processes. The bottom waveform is a sinusoid whose amplitude decreases and increases, and the top two waveforms are its only components. This effect is known as beats in sound, as moire patterns in silk, or as the vernier scale on a machinist's caliper. Let us consider two sine waves being played simultaneously through earphones, with equal amplitude, at frequencies of 100 and 101 Hz. (For convenience, Figure 20 displays frequencies of 10 and 11 cycles.) At the beginning of the playing, the two waves are both in phase and are not very different from a 100.5-Hz sinusoid. The 100-Hz wave is like a 100.5-Hz wave but is slowly getting farther behind, and the 101-Hz wave is slowly getting farther ahead. After .5 sec, the 100-Hz wave is .25 cycle behind where 100.5 Hz would be, and the 100-Hz wave is .25 cycle ahead of where 100.5-Hz would be. The result is that the two sinusoids are .5 cycle out of phase with each other and, therefore, cancel each other: One is at its peak while the other is at its trough. As the second progresses, the 100-Hz wave eventually gets .5 cycle behind the 100.5-Hz wave, the other gets .5 cycle ahead, and the two waves are a full cycle out of step and, therefore, reinforce each other. What the ear hears is not the two separate waves, but rather, a single 100.5-Hz wave changing in amplitude. If we change the phase of either component by .5 cycle, we get the opposite amplitude envelope: diminished at the ends and increased in the center.

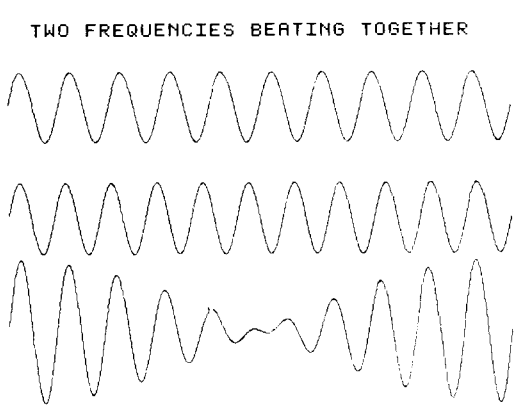

Figure 20. A classic example of the difference between global and local representations of waveforms occurs in the phenomenon of beats. When two sinusoids of almost identical frequency are added, some local portions reinforce and other cancel, resulting in a waveform that appears to have intermediate frequency and a variable amplitude: The envelope, or local amplitude, varies in time.

When two frequencies beat together, we do not resolve the frequency difference between the components, but we interpret the input as a single frequency changing in amplitude. This is typical of what happens when non-steady-state processes are subjected to an FFT: The FFT synthesizes the appearance and disappearance of waves by the reinforcement and cancellation of sinusoids of similar frequency. Each of the components has a constant amplitude over time, but our interpretation of their sum may be as a different wave with a changing amplitude. The notion of a component that changes its amplitude during the sampling frame is not part of the logic of the FFT.

This phenomenon can be considered a conflict of time scales. The FFT always looks at the correlations between the data vector and sinusoids that persist for the entire sampling duration. Even though it may sound as if a sinusoid of a given frequency is occurring only during part of the duration, the FFT will represent it as occurring during the whole duration, along with other sinusoids. The other sinusoids may cancel what we hear during part of the duration and may reinforce it during other parts of the duration. Our local time scale may perceive sinusoids that start, change amplitude, and stop, but the FFT can interpret data only as mixtures of continuous sinusoidal components. Its view is global, whereas our view, or the view of some system we are studying, may be much more local. In any small fraction of the second in which we consider two beating sine waves, the resultant has an almost constant amplitude, but globally, its change in amplitude cannot be represented as any one component but must be synthesized by the interference among components.

Waveforms that suddenly change their composition as a sum of sinusoids are common in our lives. Many of the waveforms that are important in our lives are not continuous, like the tides, but are discontinuous, such as speech. Over the course of a fraction of a second, much of speech (the vowels) is essentially a periodic

pressure wave, well described by a frequency analysis into its components. However, over the course of seconds, the important events in speech are discontinuities, such as consonants, and order information, such as which word precedes which other word. While an FFT does, in theory, capture such discontinuous and order information, it is not a straightforward matter to interpret an FFT in such a way as to recover that information. If our interest is simply in knowing the predominant frequencies, for purposes such as knowing what kind of frequency range we must perceive in order to understand speech, then the FFT provides a useful view of the information. However, while a computer could reconstruct the speech from its FFT, the computational effort in doing so is considerable. A human being cannot look at the spectrum of a sentence and state what was spoken: For a complex waveform such as a sentence, the FFT simply does not reduce the data to an easily understood form. There are steady-state sounds, such as the sound produced by a woodwind instrument after the first few milliseconds of onset transition. For such a waveform, the FFT produces an analysis that is close to our experience: a fundamental with overtones. It is simply not meaningful, although it may be formally correct, to think of an entire sentence as being the continuing presence of a very large family of overtones, which cancel each other out most of the time so that only a few frequencies are heard at any instant. Clark, Dooling, and Bunnell (1983) describe one technique for using a family of overlapping FFTs to represent such non-steady-state waveforms.

Not all of the limitations of the FFT are related to obvious discontinuities in the waveform. Problems can also occur when the process being measured has a fundamental period that is not an exact fraction of the sampling duration. For example, let us return to the marine animal whose behavior depends on the 12-h 25.5-min tidal period. If we sample the behavior for 7 days, we wind up with about 14.5 cycles of activity. Figure 21 shows sinusoidal behavior over such a time period. If we started sampling at a positive peak, then we end sampling at a negative peak. Because the FFT considers its input as one repetition of an infinite process into the past and the future, it treats the data as if the last half-cycle were followed immediately by the start of a full cycle. In other words, sampling the data for anything other than an exact number of periods introduces a discontinuity into the periodic process, and that discontinuity results in an analysis that does not fully reveal the underlying periodicities. The FFT of 14.5 cycles of sine wave contains a mixture of frequency components, clustered around the true frequency but with some degree of spreading that depends on how many repetitions of the cycle occurred before the discontinuity. If our only interest is in the approximate frequencies contained in the spectrum and if we do not need to know whether energy is exactly a harmonic of some other frequency, then such approximations and spreading are quite tolerable. There is also a technique known as windowing that can reduce the maximum amount of frequency spreading induced by a bad sampling duration, but at the cost of creating some spread even around components that are exact divisors of the duration.

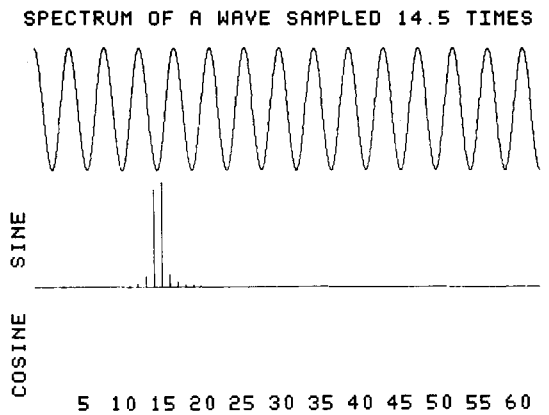

Figure 21. When a sinusoid is sampled a nonintegral number of times, its spectrum consists of a number of components clustered around the true, nonintegral frequency. As the number of repetitions of the waveform increases, the energy of these additional components approaches 0 as a fraction of the energy representing the fundamental. For this reason, spectral measurements of an unknown frequency should include a large number of repetitions of the fundamental.

Signals that are only approximately periodic are usually useful candidates for FFT analysis. For example, the alpha-rhythm component of the spontaneous EEG is an irregular electrical signal with a frequency near 10 Hz. If the overall dominant, or average, frequency shifts up or down, that may be of experimental or clinical significance, and so a reasonably accurate estimate of that frequency is important. With the use of a window and a long sampling frame, uncertainty about that frequency can be reduced to a small fraction of 1 Hz, which is rather smaller than any clinically significant difference in the dominant frequency. Also, the FFT is a good tool for estimating the total power in the alpha rhythm, that energy around 10 Hz, even though it may not be able to give an exact measure of the dominant frequency.

In summary, the FFT is most useful in these situations: (1) When the event being measured has an inherently variable spectral composition, such as EEG activity or hand tremor—The underlying waveform is a mixture of closely related frequencies, and if the FFT returns a result describing such a mixture, that is appropriate. (2) When only the frequency range of the waveforms is needed, in order to design a system capable of coping with that frequency—The frequencies and order of sounds produced by an orchestra are very exact, but only the range of frequencies and intensities may be necessary information in choosing an appropriate

microphone. (3) When a large number of repetitions of an exact, periodic waveform is included in the sampling duration–If one samples several hundred repetitions of some wave, the frequency spread around the fundamental, as a fraction of the fundamental frequency, is small compared to the components representing the fundamental frequency. (4) When there is an external driving force, such as the tides or a pendular visual stimulus, whose frequency is known exactly, so that the FFT's assumption of a fragment of an infinite periodic process of exactly known period is not unreasonable–This is especially useful when the sampling period consists of an exact number of repetitions of the underlying, driving waveform, so that no problems of frequency spread occur. (5) When the waveform to be analyzed can be decomposed into a number of shorter segments, each one of which satisfies the same one of the assumptions above.

REFERENCES

CLARK, C., DOOLING, R. J., & BUNNELL, T. Analysis and synthesis of bird vocalizations: An FFT-based software system. *Behavior Research Methods & Instrumentation,* 1983, **15**, 251-253.

EMERSON, P. L. Analysis of variance with Fourier analysis of coherent data. *Behavior Research Methods & Instrumentation,* 1983, **15**, 242-250.