How to Understand Fourier Theory

first principles

Let's say you start with one little sinusoidal waveform. It's a sine wave or it's a cosine wave. It doesn't *do* very much. All sines and cosines look the same except some are taller or shorted, and some are wider or narrower. A sinusoid is the shape that results from the tip of an arrow turning in a circle. A sinusoid, in fourier theory, has no amplitude of its own. A sinusoid always has a maximum amplitude of 1 or -1. A sinusoid, in fourier theory, is infinitely long--it has no beginning or end.
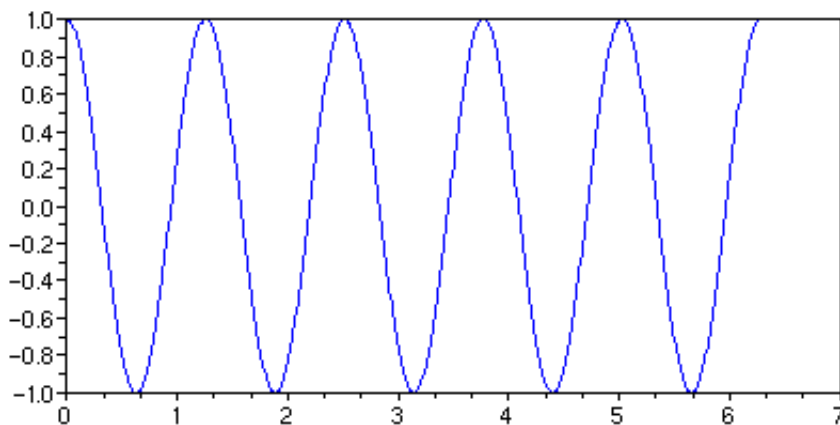
All this means that there is very little information contained in the fourier definition of a sine or cosine wave. The only information it offers is the relative distance between its peaks. When we say that the sinusoid is infinitely long, we simply mean that there is no information available about its endpoints. There is no literal reality to infinity--it is a fiction of the mathematician, a symbol, a placeholder that represents an absence of information.

Now let's say we have a second little sinusoid. This other sinusoid has a slightly different period between its peaks--a slightly different frequency. What happens when we add them together?
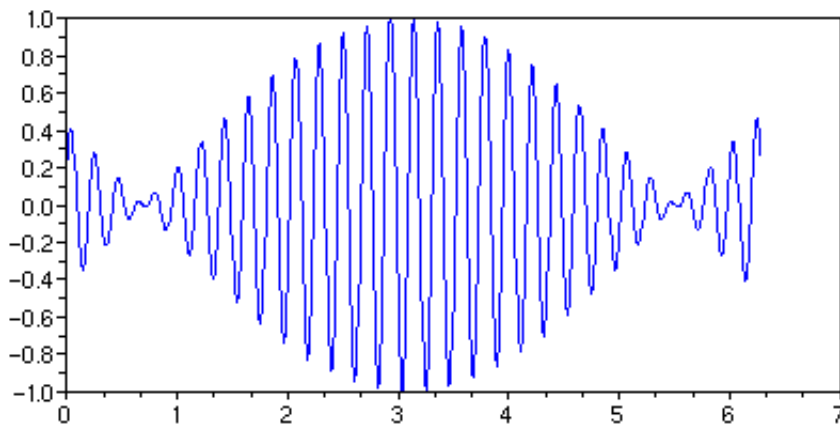
They combine to form a *beat frequency*, just like when you tune your guitar strings. This is the most important concept in fourier theory. Adding two sinusoids had the same effect as one sinusoid being modulated by another sinusoid. This is a special case of what is called *the modulation theorem*. There is a simple trig identity that will help you to remember how this works:

$$\cos(u) + \cos(v) = 2 \cos((u+v)/2) \cos((u-v)/2)$$

And the reverse is true: modulating one sinusoid by another sinusoid has the same effect as adding two sinusoids, which may be seen by the inverse trig identity of the above:

The two sinusoids above were deliberately chosen to have frequencies nearly equal to show the modulation effect. It is interesting the repeat this experiment, but gradually increase the difference between between the frequencies.
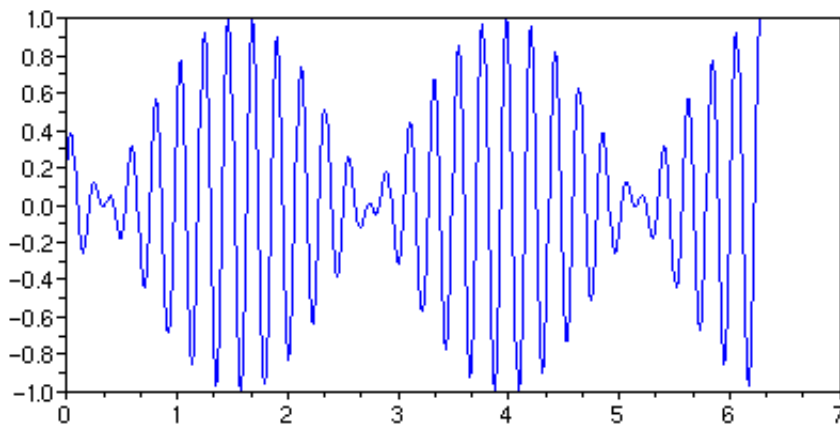

The result is that the modulating "envelope" begins to increase in frequency.

As we increase the frequency difference, the resulting wave shape becomes more convoluted, and it becomes increasingly difficult to distinguish the modulator from the base frequency.
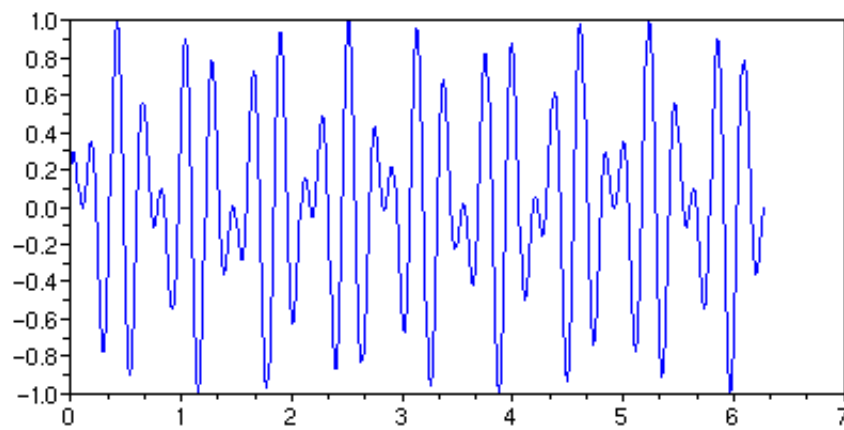
If we continue to increase the frequency difference between these two frequencies until one of them has a frequency twice as large as the other, then an interesting thing happens.
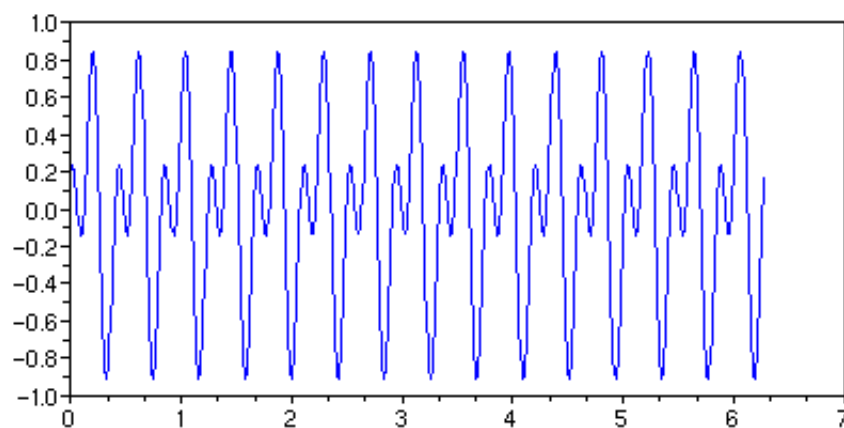
This has the same effect as if the period of the modulator matches the period of the base waveform. In other words, the harmonically varying waveform is equivalent to a simple sinusoid *with an amplitude envelope whose periods begin and end in phase with the base waveform*. This must have been why somebody called this the modulation theorem, because of its overreaching implications. It explains fourier theory. It says that frequency components are due to a modulation, or that a modulation can be resolved into frequency components.

repeating waveforms

Sinusoidal waveforms repeat: their fundamental shape is repeated indefinitly. The combination of *any* two sinusoids produces a waveform that repeats.

In the example shewn above, we saw that the resulting waveforms had one repeating shape superimposed an another repeating shape. But it was not obvious that the resulting waveform had a pattern that *exactly repeats* itself. When the two frequency components were chosen to be integer multiples of each other we saw clearly that the pattern exactly repeated itself. But what about the cases in which the difference in frequency was chosen arbitrarily? The answer is that they, yes even they, have a pattern that exactly repeats itself.

To see this, imagine two nested circles. Imagine that the smaller circles turns inside the larger circle. Each circle has a benchmark at the bottom so that we can see its orientation as the

small circle turns inside the larger circle.
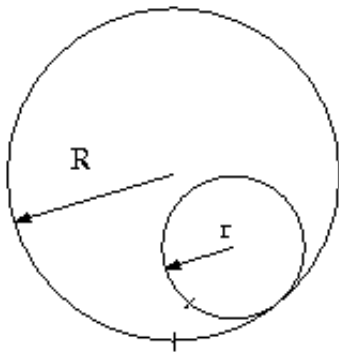
The larger circle is like the slower frequency, and the smaller circle is like the higher frequency. Where does the combination of two sinusoids repeat itself? The question is analogous to the problem of how many times the inner circle turns within the larger circle before it arrives at its original orientation. The answer to the geometrical problem depends on the ratio of the radii of the circles. In the fourier problem this corresponds to the ratio of the two frequencies. There will be some ratio of whole numbers which will give the repetition condition for these two waveforms. In the case of figure 2 above which has the slow beat frequency, the ratio was 22:23. The composite waveform repeats itself exactly after 22 cycles of the slower frequency (and 23 of the higher). This is not immediately evident from the plot because the smaller oscillations have a frequency that is the average of the two component frequencies.

The addition of any two sinusoids results in a waveform that repeats itself. It follows that the superposition of *any* repeating waveform results in a waveform that also repeats itself. This is the key to understanding fourier theory. This interval between which the composite waveform repeats itself exactly is its *interval of orthogonality*. This interval is where you apply your FFT calculation to minimize the "smeared out" frequencies (the modulation artifacts). This interval is where you perform your fourier series calculation to obtain a spectral component. This interval is where you perform your fourier integral (although you're not given much choice for the latter!).

The repetition feature is important in many ways for understanding the fourier calculations. Firstly, it is assumed that the component waveforms all have a constant amplitude. Unlike the Laplace Transform, whose kernel includes a decaying exponential function, there is no concept of amplitude variation in fourier theory. Any variation in the resulting waveform, for example, the "beats" in figure 2 above, is due to the additive interference between the components. Figure 2 shows an amplitude variation, but figure 5 does not. Figure 5 resulted from components whose frequency ratios divide cleanly into integer factors, and this is the typical application for the *fourier series* formulas. But figure 2 resulted from components whose frequency ratios result in rational numbers (floating point for programmers). Figure 2 illustrates how an amplitude envelope on a sound is resolved in fourier theory: it is presumed that the end-to-end sound envelope *is just one interval that repeats itself*. The envelope is like a more complicated version of the "beat" amplitude. And since we know how to get a beat amplitude--by mixing two components with nearly the same frequency--we know what to look for in the fourier spectrum of the sound. We look for frequency components that are very close together with respect to frequency.

the basic idea of fourier theory

Figure 2 and 5 above show us the "basic idea of fourier theory", which is described as follows:

> Small variations in a signal are due to widely-spaced frequency components, and slow variations in a signal are due to frequency components that are close to each other.

For musicians studying sound waves this rule translates into the following rule:

> The harmonics of a sound are due to the frequency components that are integer multiples of the fundamental, and the amplitude envelope is due to frequency components that are grouped near the harmonics.

We have seen that the combination of *any* set of frequency components exactly repeats itself at some interval that depends on the ratios of their frequencies. This interval at which the resulting waveform repeats itself has a special significance. In this interval, each of the component waveforms contain an integer number of complete repetitions--a different number for each, but complete repetitions. This interval is the *interval of orthogonality* for its fourier transform calculation. This is the interval over which the inner product of one of the components with the resultant waveform causes all the other components to vanish. For those of you who are new to "inner product," see the previous tutorial, <u>what is fourier theory</u>.

This interval of orthogonality shows up in the pages of fourier calculations in various places. These are the endpoints on the integral formulas for the fourier series formulas. These endpoints appear variously in the literature as

```
-pi  to  pi

-L   to  L

-T   to  T
```

and so on. The fourier series formulas were devised with the idea of integral multiple frequency components. The interval of orthogonality can be easily seen in that case, because the fundamental frequency component is always where the resultant cycle repeats itself.

the flavors of fourier theory

The fourier calculations appear with different symbols and different contexts.

- Fourier Series
- Fourier Transform
- Discrete Fourier Transform

How does the FFT relate to fourier theory, and whay are there different flavors of the fourier transformation as shewn above?

The FFT is what you use for practical work, and the fourier transformations above that contain calculus symbols in their formulas have almost no use outside the classroom.

The fourier series formulas, as noted in the previous paragraph, were designed specifically with the idea of component waveforms whose frequencies are integer multiples of the fundamental. But it's greatest drawback is that it is defined with calculus symbols: its members belong to a make-belive world, that exists only in the mind of the mathematician, called continuous-valued functions. The universe is discrete--everything else is analog.

The fourier transform formulas are a special case of the fourier series formulas in which the interval of orthogonality, that is, the interval at which the waveform exactly repeats itself, has been magnified to extend from minus infinity to plus infinity. As I have stated over and over, the combination of any sinusoids produces a waveform that repeats itself. In this case, the waveform repeats itself infinitely far away. The price that is paid, mathematically speaking, for this wizardry is that now the frequency spacing between the component frequencies is arbitrarily small. These ideas are mathematical gaming--they have no literal reality. You can study the derivation of the fourier integral for yourself in the textbooks. Try Arken, "Mathematical Methods for Physicists." One interesting way to look at the fourier transform

integral is from the study of Moebius Transforms. Imagine that the interval from -pi to +pi is the circumference of a circle. Draw a line next to the circle. The circle can be "mapped" to the line, that is, for every point on the circle a corresponding point is found on the line. Zero on the circle maps to zero on the line; pi on the circle maps to infinity on the line. It's a pictorial representation of the relationship between an arc and its tangent function.

The Discrete Fourier Transform (DFT) is the basis for the FFT calculation. Why should working with discrete (digital) values require inventing a whole new branch of fourier theory? After all, the world is discrete--the calculus is continuous. But such is the mind set of the people teaching this subject. The only thing you need to worry about is not sampling higher than the Nyquist frequency. Never mind all that z-transform filler.

There is a considerable amount of space in the textbooks dedicated to a "kaiser window". The subject of windowing a data set is considered to be an advanced topic in digital signal processing. The point is this: waveforms are considered to be comprised of fourier components that are infinitely long. But no set of sampled data can be infinitely long. So the truncation of the infinitely ling wave form results in the spectrum being *convoluted* with a sinc function (more or less). Therefore, a truncated waveform has its frequency information "smeared out". The application of a window has the purpose of minimizing this frequency smearing. It should be obvious now that you have read this tutorial what is wrong with that picture. Fourier components are not infinitely long, they just lack information about their end points. Components can only be retreived with calculations that are performed between data points where the waveform repeats itself exactly. If you perform the calculation at arbitrary data points, then the calculation does not reproduce the fourier components that were combined to make the waveform. And if the waveform data has no regions that repeat (exactly)? Ach so, then the fourier transform is a misapplication of mathematics to the problem, isn't it? Almost nobody in the engineering profession understands this subject: data windows are applied arbitrarily as if by recipe. Whatever you do, don't *ever* zero-pad the data.

summary

We have seen in this tutorial how the combination of constant-amplitude sinusoids produce waveforms that repeat exactly. The segment of waveform between which the waveform repeats exactly is its *interval of orthogonality* with respect to fourier calculations. The time variation of a given waveform, that is, its amplitude envelope, is the result of frequency components close to each other with respect to frequency (a truncation is a kind of envelope). The fast variation of the waveform, that is, what is typically thought to be "frequencies" in the waveform, is the result of components that are far apart with respect to frequency. *The Modulation Theorem* provides the surpassing concept of how a fourier spectrum is to be understood.

Music Synthesis and Physics Home