Chapter 1: Learning to use the SR770

This section teaches you the first-time use of a powerful Fourier-analyzer tool, the SR770. In this section, we'll specialize to the study of steady (cw, or continuous-wave) signals, and the emphasis will be on how to configure the instrument. Many of the capabilities you learn here will display their payoff in later sections. Later sections will also instruct you in the 'how' and the 'why' of what goes on inside the 770; for now, the goal is mostly to get it to do things for you.

We suggest a learning environment in which you produce a signal with a benchtop signal generator, and you send its output to both an oscilloscope and to the SR770. The generator need only to cover the range 0-100 kHz, and it ought to be capable of generating sine, square, and triangular waveforms. It is not important whether it is ancient or modern, analog or digital, in its construction.

You should set that generator to produce sine waves with frequency near 50 kHz, and amplitude near 1 Volt (that is, with excursions of 2 Volts, peak-to-peak), and you should use your 'scope to confirm that you are getting such a waveform. Questions: Do you know how to set a triggering mode for your 'scope, and can you get a stable display? Why is 500 mV/div a good choice for vertical-scale setting? Why is 10 μs/div a good choice for horizontal-scale setting?

When you have this signal appearing on the 'scope, and you have also connected it to the input BNC terminal marked Signal In, A at the lower right of the 770, you can turn on the 770 (the power switch is at lower left of the display screen). Watch the 770 go through its self-test exercise, ending with an Offset Calibration. When the 770 shows a spectral display, it's ready for you to set it up in a 'standard configuration'. To do so, notice at right the MENU buttons or 'hardkeys'; there are also 'soft menu' buttons or 'softkeys' just right of the display screen, whose function changes according to the labels shown on the display screen at its right side.

Here's a suggested list of standard settings, many of which will already be in place from a previous user's configuration of the instrument. The bullet points below mark the 'hard MENU' functions you can select in turn.

- MEAS , (for measure), brings up a soft menu;
  Press the Measure Menu softkey, and within that menu, choose Spectrum and then press Return;
  Press Display Menu, choose Log Mag(nitude) and Return;
  Press Units Menu, choose Volts P(ea)k; and (lower down), choose Volts rather than EU, and then press Return;
  Press Window Menu, choose BMH and press Return;
  Skip the Calculator menu.
- INPUT , for configuring the input stage, brings up a soft menu:
  Input Source: choose A, rather than A-B;
  Grounding: choose Ground;
  Coupling: choose DC;

Input Range: press this, and see a number in ‘dBV’ light up. Confirm that you can change this number with the (sole) rotating knob on the 770, and use that knob to set it to 10 dBV (see Appendix A3 for the meaning of the dBV scale). Or, try using the ENTRY keypad to do so;
Trigger: press for a menu, set to Cont(inuous) among the five choices, then Return;
Auto Offset: choose Off, to suppress periodic dc-offset calibrations of the input circuitry that would otherwise interrupt your work

- ANALYZE , for choice of analysis options, brings up a soft menu, from which you can select None.
- SOURCE , for controlling the 770’s internal signal generator, brings up a soft menu, which you can set to Off.
- STORE/RECALL allows you to save settings to a file, which you might want to do. But you will not yet *need* to do so, since the 770 will come back on in the future with the settings at which you have left it.

There are further hard-menu buttons to set:

- FREQ , for setting the frequency range of coverage, brings up a soft menu; for now, just find and press the Full Span softkey, and note the display’s horizontal axis now displays limits of 0 and 100 kHz on its scale.
- DISPLAY sets the display-screen parameters, chosen by soft-menu options:

  for Format, choose Single;
  for Marker, choose On;
  for Marker Width, choose Spot;
  for Marker Seeks, choose Max(imum);
  for Grid Dv/Screen, chose 8;
  for Graph Style, choose Line;
  and for any of these, you can check what happens if you make alternative choices.

Somewhere amid this procedure, you’ll have noticed the Fourier spectrum of your input signal is being displayed, in real time. You should be seeing a narrow tall peak centered at 50 kHz, in the middle of the horizontal scale. Question: How does this peak respond, if you change the generator’s frequency setting? or if you change the generator’s amplitude setting? Ensure that you understand how and why generator-setting changes will show up (though in different ways!) on both your ‘scope and on the 770.

Back to the last hard-menu buttons:

- SCALE brings up soft-menu choices for the scales of the display. An easy alternative to all these options is to push the AUTOSCALE button among the ENTRY buttons at the center of the 770.
- AVERAGE brings up soft-menu choices; for now, use the uppermost softkey to set averaging to Off.
- SYSTEM brings up soft-menu options akin to system-preference choices for a computer, none of which you need to select just yet.

Now turn your attention away from hard-menu buttons to the CONTROL buttons of the 770. First press the one labeled MARKER. This couples the rotating knob to the vertical dashed marker-line visible on the display. Turn the knob and watch the marker move, and move it until it is centered on the peak you see in the spectral display. Now note that you can read off, at the top of the display, the frequency to which you've set the marker, and the amplitude of the Fourier component that you have thereby selected. This is how the 770 can serve as a 'Fourier voltmeter', giving you the magnitude (but, so far, not the phase) of any spectral component you select by pointing to it with the marker. Confirm that changes of your generator settings create changes in your 770 display, *and* at the marker's readout, of the sort you would expect.

That completes the long list of set-up choices; most of your future encounters with the 770 will require much less attention to these options. You will find that for any new measurement, the AutoRange and AutoScale buttons are very good at adapting the 770's input-stage's sensitivity and its display scale (respectively) to the new signal you might be measuring. Now turn your attention from 770 set-up options to some actual physics that you are newly able to see in the frequency domain.

For another route to familiarization with the SR770 used in its most basic mode, refer to the SRS Operating Manual's section on 'Getting Started', and follow its step-by-step instructions on pp. 1-1 through 1-5.

There may be times when you get stuck with the SR770, and can't make sense of what it is doing. In such cases, refer to the SRS Operation Manual in the section 'Getting Started', at p. 1-47, for some suggestions less drastic than a full power-up and re-boot.

Leaving the 770 as you have configured it, set your generator to produce a frequency near 10 kHz. That will move the spectral peak you've been seeing to the 10-kHz location on the horizontal scale, but it is also likely to make visible some *other* peaks. Use the marker to identify these new frequency components, and see how many of them correspond to harmonics of the fundamental, ie. how they occur at integer multiples of the fundamental's frequency. You have just discovered the level of harmonic distortion produced by your signal generator; that is to say, you are now sensitive to the extent to which the generator's output waveform *departs* from a pure sinusoid. If the generator's output is periodic in time, all the extra spectral content is guaranteed (by Fourier's series expansion) to lie at harmonic frequencies; but any departure from a pure sinusoid will show up as *non-zero amplitudes* for these harmonics.

Signal generators differ widely in the purity of their sine-wave output, so if your fundamental shows up with Volt-level amplitude, you might find the harmonics show up with amplitudes below a milliVolt (for a modern low-distortion digital-synthesis generator), or at the 10 to 100-mV level (for some analog generators). Note that your 770's vertical scale is logarithmic, so that relatively tiny amplitudes (for example, 1 mV compared to 1 Volt) show up with apparent prominence, pitilessly exposing the harmonic distortion (if any) of your generator. Also note that there are reasons that even-numbered harmonics might be relatively small, but odd-numbered harmonics might be relatively larger, for some types of generator circuitry, so you might see an odd-even staggering of peak heights for these harmonics.

To see a case in which harmonic content is deliberate, rather than a small departure from the ideal, change your signal generator to produce *square* waves (but still of frequency f near 10 kHz). Now you should see prominent *odd* harmonics (at frequencies $3f$, $5f$, $7f$, $9f$, etc.), which are required parts of a square-wave waveform; you might also see some non-zero strength of *even* harmonics (which would be absent from an ideal square wave). Now you can do some quantitative measurements, since a square wave attaining voltage levels of ±1 Volt has a calculable Fourier spectrum. The predicted amplitudes of the Fourier components are, for n = {1, 2, 3, 4, 5 etc.}, the numbers { $4/\pi$, 0, $(1/3)(4/\pi)$, 0, $(1/5)(4/\pi)$, etc.}. Apart from an overall scale factor set by the strength of the square wave, note the predicted proportions {1, 0, 1/3, 0, 1/5, etc.} for the amplitudes of the harmonics. Read off and tabulate the amplitudes you observe, and compare them to this prediction.

Change your generator again to a *triangle* wave. Relative to the amplitude of the fundamental, how *should* the amplitudes scale, according to Fourier-series theory? How *do* they scale in fact, according to your observations?

There are good reasons for the amplitudes of Fourier components of a waveform to drop as n^{-1}, or n^{-2}, or some other power law. To see such power laws more visually, try arranging for a log-log display. Your vertical scale is already logarithmic, because you've selected Log Mag(nitude) in the Measure menu. But in the Scale menu, the bottom softkey brings up the option of a logarithmic, rather than a linear, horizontal scale. For best results, lower your generator frequency to 2 kHz, so there is room for lots of harmonics to fit into your 0-100 kHz scale. Now alternately select square vs. triangular wave input waveforms, and notice the many spectral peaks you can observe. The tops of the peaks fall along a straight line, but a line whose *slope* changes when you change the waveform. That straight line (on a log-log display) is a tip-off of a power-law dependence for harmonic content. For a square wave, the (odd) Fourier coefficients drop off as n^{-1}, while for a triangular wave, the (odd) Fourier coefficients drop off as n^{-2}, which accounts for both for the linear trends, and the differing slopes, of the patterns you see. (See Appendix A4 for more insight into the reason for the difference in power-law exponent.)

In summary, you have learned to use the SR770 for some initial measurements. The sine-wave measurement may look mundane to you, since for a simple sinusoid, you could *already* have read its amplitude and its frequency from a 'scope's display of the time-domain signal. But you should now be aware that the frequency-domain view you can get on the 770 will glaringly reveal harmonic distortion in a wave which looks sinusoidal on a 'scope. Similarly, the Fourier analyzer will test quantitatively the harmonic content of any periodic waveform, and you have tested this capability against expectations with square and triangular waveforms.

Chapter 2: Learning to use the SR770's internal waveform source

This section teaches you some more capabilities of the SR770. In this section, we'll address particularly the internal waveform generator that's built into the 770, and learn how to use it. For the first time, you'll also be using the Electronic Modules provided in the TeachSpin instrument case. There are lessons of great generality to be learned here, including one version of the frequency-time 'uncertainty principle'. In particular, you'll see that there are limits to the extent to which Fourier analysis can resolve (that is, can see as separate) two closely-spaced frequency components.

The 770's internal waveform Source

First, the internal waveform generator. There is an actual analog waveform *output* on the 770, marked Source Out, among the four BNC connectors on the lower front panel. To configure the source, press the Source button among the hard-menu choices at the right side of the panel. This will bring up a soft menu, with choices among the sort of waveform to be generated within the 770 and exported via this BNC output. Change from the previous setting of None to Sine, and then use the bottom softkey to Configure the sine-wave source by another soft menu. In that sub-menu, you can push a softkey to highlight either frequency or amplitude; and then you can use either the rotary dial to increment the present value up or down, or the Entry keypad to enter any desired value. Note that the frequency can be set anywhere in the 0-100 kHz range, and that the amplitude anywhere in the 1-1000 mV range.

To confirm that this internal Source is really working, use a BNC cable to convey its output to an oscilloscope. When you've confirmed that there is a signal emerging, use a BNC splitter (a 'T' or 'F' connector) to convey the same Source signal also to the Input A of the 770. You should see that the 770's version of a sine wave source is very nearly free of harmonic distortion. But you might speculate that this is not a fair test, that the 770 'knows what it should depict' and is getting the right result via internal software, rather than external hard-wired, connection.

To dispel that fear, and to learn the lessons of this section, you now want to use your former signal generator, and the new Source inside the 770, to create an arbitrary superposition of *two* sinusoids, independent as to frequency and amplitude. To do this, we provide the Summer module among the Electronic Modules part of the TeachSpin equipment. This module will create, out of two inputs, the (negative of the) *analog sum* of the two waveforms. This only requires that the frequency content of both inputs lie below about a MHz, and the voltages involved lie in the ±12-V range (but there is no requirement that either input be sinusoidal, as in the present application). Since this is an active circuit, the Electronics Module will need to be getting power. Provide this via a power-line connection of the external power supply of the Electronic Modules, connecting that power supply's output to the rear-panel connector on the Electronic Modules box. The power supply's green light, and the front-panel red Power indicator, should both light up.

Now test the Summer unit via a ‘scope test. With the 770 serving as a source of a sine wave of 50-kHz frequency and 1-V amplitude, send its output via a BNC cable to one of the Summer’s inputs. Now configure your benchtop signal generator to be a sinewave source of frequency 40 kHz, and amplitude about 2 Volts, and send its output to the other input of the Summer by a second cable. Use a third cable to convey the Summer’s output to a ‘scope.

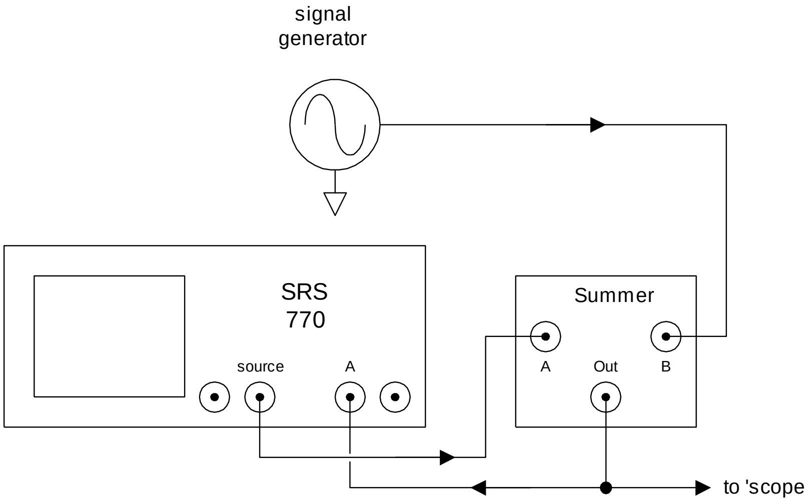


Fig. 2.1: A block diagram of one use of the Summer module

You should see on the ‘scope the superposition signal, which is quite literally the instant-by-instant sum of the two waveforms you are generating. In the time domain, that sum looks quite complicated, and it can display some quite strange-looking effects if the two sources’ amplitudes are picked to be similar, and their frequencies are picked to be very close together.

But when you have confirmed (perhaps by temporary disconnection of one, then the other, of the two input cables) that this superposition is really present at the summer’s output, use a BNC splitter at the Summer to convey this superposition signal also to the A-Input of the 770. As in the previous section, you are now able to see simultaneous time-domain and frequency-domain views of one and the same signal, on the ‘scope and 770 respectively.

What’s the point? The same signal which looks messy on the ’scope has a *very simple* appearance in the spectrum, as it should. That’s because the Fourier transform is a linear mathematical operation: the Fourier transform of a sum is the sum of the Fourier transforms. So two sources, each of which has a single-line spectrum, combine to a sum which is complicated in the time domain, but whose two-line spectrum is just the sum of the two formerly separate line spectra.

But there is much more than mathematics here! Confirm that you can separately modify the *frequency* of each source, with the expected results in the spectrum. Confirm also that you can separately modify the *amplitude* of each source, with the expected results in the spectrum. Next, imagine that in each of the two cables conveying sine-waves to the Summer, you had an on-off switch which acted like a telegraph key, turning that signal from full-on to full-off at will. Each ‘telegrapher’ would be able, independently of the other, to control the height of his or her line in the two-line spectrum. But that means that *two separate telegraph messages could simultaneously be sent over one cable,* the cable from the Summer to the 770.

This is called ‘frequency multiplexing’, and it applies far beyond telegraphy. You can see that there is ‘space’ in the frequency spectrum, and things can be made to happen side-by-side and independently in frequency space, even though they are physically overlapping in one and the same cable.

Once you’ve seen this can be done with two signals, there is nothing in principle from it working with many more signals. In fact, you might wonder if in the 0-100 kHz coverage of the 770, there might be room for (say) 100,000 separate telegraph channels to exist in parallel. The answer is yes in principle, but no in practice; and ‘no’ for two distinct reasons.

The first of these is that the SR770 presents the frequency content of whatever signal it is displaying as the contents of 400 adjacent channels. In your present 0-100 kHz ‘full span’ mode of using the 770, what looks like a continuous line spectrum is in fact a line connection of 400 dots on the screen, each one depicting the (log of the magnitude of) the spectral content in a ‘channel’ that’s 250 Hz wide. You can see that channel-width by pressing the FREQuency hard-menu button, bringing up the soft menu which displays an entry called Linewidth. That’s where you’ll see the 250-Hz number (for the settings you’re now using); it’s really the inter-channel spacing of the frequency sorting that the Fourier transform is doing.

The second problem with multiplexing 10^5 signals down one single wire has to do with modulation, which you’ll study in Section 3. But briefly put, you’ll find that a 40-kHz sine wave, once it’s modulated by turning it on and off with a telegraph-key switch, will no longer have a spectrum which is a single line. The spectrum of the modulated 40-kHz signal will take up more *room* in frequency space, and you’ll soon learn how much more.

For more exercises in using the Source function within the SR770, refer to the SRS Operating Manual, in its section ‘Getting Started’, at p. 1-31 in general, and pp. 1-33 through 1-36 in particular on the Sine source.

Resolving sinusoids of similar amplitude

Now let’s look into the question of how close together two frequencies can be, and still be resolved as two spectral peaks in a frequency-domain view. In the language of spectroscopy, we’re talking about the ‘resolving power’ of a Fourier spectroscope. You are all set up to send to the 770 a superposition of two sinusoids, of independently adjustable frequencies and amplitudes. For these first tests, set the amplitudes to be near-equal. Now have a look at what the spectrum looks like when you hold one frequency fixed, and vary the other so that the two frequencies involved are close together in frequency space. For the settings you have used thus far (‘full span’ coverage of 0-100 kHz, and the choice of the BMH window), you will find that two frequencies cannot be resolved if their frequency difference is under about 500 Hz, and that they are barely resolved (into two separate peaks) if their frequency difference is about 1 kHz. Question: How can you test the claim that it’s the frequency *difference*, and not the frequency *ratio*, that matters? Consider two frequencies of 1 and 2 kHz (differing by 100%) and another two frequencies of 94 and 95 kHz (differing by 1%) – are these pairs equally resolvable?

Now why is such a large frequency difference required for resolution, and can we do better? The answer is: Easily, *but it takes longer*. If you use the FREQuency hard-menu button, you will see that you have configured the 770 to a 100-kHz ‘span’ (ie. width of frequency coverage), and that the acquisition time is listed as 4 ms, 0.004 s. This means that everything you see on the screen is computed from the acquisition of data in blocks of duration 4 ms. (The actual acquisition is ‘voltage sampling’, at a fixed sampling rate of 256 kSa/s, thus 256 samples per millisecond, and 1024 samples during the 4-ms acquisition period.) Now consider two frequencies of 50 kHz and 51 kHz; these numbers give a count of whole cycles per second, so there are 50,000 vs. 51,000 cycles per millisecond, or 200 vs. 204 cycles during the 4-ms acquisition window. A Fourier transform is certainly able to resolve the distinction between these two sinusoids, precisely because **they differ by more than one whole cycle occurring during the acquisition time**. But if the frequency difference is smaller, you’ll see not two (incipiently) resolved peaks, but rather one, broadened, spectral peak, unresolved into its separate frequency constituents.

But this identification of the cause also shows the cure. To achieve greater frequency resolution requires capturing more cycles of the sinusoids involved, and that requires spending *more acquisition time*. In the 770, this is achieved by using, in the FREQuency hard menu, the Span entry of the soft menu that comes up. If you press the Span softkey, it’ll highlight, and then you can control it using the rotary knob. If you dial that knob downwards, you’ll find you can reduce the span from 100, to 50, or 25, or 12.5, etc. kHz; that is to say, you can create successive *halvings* of the frequency span of coverage. But as you do so, you’ll see the acquisition time is simultaneously *doubling*: a 50-kHz span is achieved by using an acquisition interval of duration 8 ms, and so on.

Once you have called for (say) a 50-kHz span, it is still your choice where to put that 50-kHz of coverage in frequency space. To do this, use either the Start Frequency (or Center Frequency) softkeys, and the Entry keypad, to set a start (or center) frequency for your spectral coverage. For example, if you are studying a superposition of 50 kHz and 51 kHz, you might set the 770 to use a center frequency of 50.5 kHz. Now have a look at he real-time Fourier spectra of your superposition, as you successively reduce the span by factors of two. You will see the two peaks, formerly unresolved, get as resolved as you wish; but you will also see that updates to the display come less frequently. By the time you get down to a span of 1.56 kHz (ie. 100 kHz/2^6), you will have your two spectral peaks near opposite edges of your display, *very* well resolved – *but* you’ll be needing a quarter-second acquisition time to do so.

This tradeoff between frequency resolving power and data-acquisition time is generic to all of physics; you could call it the frequency-duration 'uncertainty principle', and write it as

(frequency resolution achievable) · (acquisition time required) ≥ a number ,

where the exact value of the pure number on the right-hand side depends on your definition of the frequency-resolution condition (and also, as you'll see, the 'windowing' choice you make in the 770). If you multiply this equation through by Planck's constant, you get a relation sometimes called the energy-time uncertainty principle; but you've now seen it's a fact about classical waveforms in time.

Practice setting two new frequencies elsewhere in the spectrum, and then resolving them, using the full-span and then 'frequency-zooming' technique you've now learned. Find out what is the minimum value to which you can set the 770's span, and what acquisition time that would entail.

Resolving sinusoids of quite different amplitude

Thus far you've used a superposition of two sinusoids of similar amplitude, but evidence for *that* sort of superposition is easy to obtain using a mere oscilloscope. The true power of Fourier methods shows up when sinusoids of markedly different amplitude are involved. So pick two frequencies about a kHz apart (say 50.1 and 51.1 kHz), and form their superposition; leave your signal generator's amplitude at about a Volt, but this time, set your 770's internal Source to give an amplitude of a milliVolt, 1 mV. Your two components of the superposition now differ by 10^3 in amplitude, or by 10^6 in actual power, and a 'scope view of the superposition will reveal absolutely *nothing* of the existence of the weaker component. What does your 770 show?

The answer depends on a choice of configuration of your 770 called 'windowing'. Pick a frequency span that shows your two peaks resolved, and then use the MEASure menu button, and the Window softkey, to see the four choices provided by the 770: Uniform, Flattop, Hanning, and BMH. Press these in turn, and you'll see quite distinct views of the spectrum. Each of these 'window' choices (and there are many more that can be used in Fourier signal processing) makes a different trade-off between resolving power (at the signal peak), computed lineshape (in the 'skirts' of a spectral line), and amplitude accuracy. (See Appendix A5 for more about windowing.)

Here's some guidance of what to look for in your spectral views:

- Uniform window: peaks optimally sharp at their tops, but the stronger peak has 'skirts' extending far and wide to both sides, nearly obscuring the presence of the weaker peak
- Flattop window: peaks with an optimally 'flat top', but also showing steep sides.
- Hanning window: peaks with rather sharp tops, and steep sides, with some flaring skirts near the bottoms of the peaks

- BMH window: peak tops not as sharp as the Hanning window, but very steep-sided peaks with no skirts.

The exact meaning of the 'windowing' operation is left to Appendix A5, and it's a great example of what really goes on in a Fourier transform applied to a set of data captured over a finite duration of time. But the implications of what you've already seen, and more things that you will see, can be summarized as a list of recommendations:

- Uniform window:
  *the required choice for transient (as opposed to steady) waveforms; also
  *optimal for resolving two very-closely-spaced peaks of near-equal amplitude
- Flattop window:
  *the optimal choice for quantitative accuracy of measurement of peak heights, ie. amplitudes of spectral lines
- Hanning window:
  *good choice for spectral resolution;
  *not quite so good for accuracy of amplitude measurement, or seeing very weak peaks near strong ones
- BMH window:
  *good compromise choice for steady waveform; not the highest resolution at peaks' tops, nor perfect amplitude accuracy, but very good for seeing weak peaks near a strong one.

Spectral Leakage, and the Windows cure

You've now seen the effect, if not the theory, of windowing in computing Fourier transforms, in the context of trying to resolve two closely-spaced frequency components, whether of near-equal or very unequal amplitude. This (optional) section shows you another motivation for the use of windowing.

For this demonstration, you need only a single-frequency source, but ideally it should be a digital-synthesis signal generator, capable of 1-Hz frequency settability say around 50 kHz. Any amplitude, say 1 Volt, of output will suffice. Now send that signal to your 770, and set the 770 for its full-span mode, covering 0-100 kHz. You should see a prominent spectral peak at mid-scale. Now set the 770, under the MEASure menu, and the Windows soft menu, to the Uniform-window mode.

The test consists in looking at the displayed version of the Fourier spectrum as you vary the frequency from 50.000 kHz, through 50.125 kHz, to 50.250 kHz. What you will see is a spectral peak undergoing not very visible changes, but that peak rising up out of a background which does undergo prominent changes. Notice how narrow and isolated the peak is when 50.000 or 50.250 kHz is used; notice how much spectral energy gets splattered into the wings of this spectral line when you use frequencies intermediate between these two.

This effect is called ‘spectral leakage’, and it is one good reason for avoiding the routine use of the Uniform window for cw (steady) signals. The reason from spectral leakage is tightly connected to the actual process of voltage sampling, for a finitely-long acquisition time, that is characteristic of digital frequency analysis.

Recall that (as you are using it) the 770 takes 256 voltage samples per millisecond for 4 ms, and then computes a Fourier transform of that data array. Internally the 770 generates values for the spectral energy at each of 512 frequencies, starting at dc, and spaced uniformly from 0 to half the sampling frequency (128 kHz in this case). Then it displays values for 400 components, at frequency locations of the form (integer)·250 Hz, which are the frequencies 0.25, 0.50, 0.75, 1.00, . . . 99.50, 99.75, 100.00 kHz. You can verify this claim by using the marker to point to various points in the spectral display, and seeing that the frequencies you can point to, and the spectral components you can read out, at the top of this display, fall onto this frequency list.

The reason for that list? The 770 takes data only for a 4-ms time duration, and then computes the Fourier series corresponding to that set of data. In the process, it implicitly assumes there is a Fourier series matching each of the sampled data values, which is, in turn, to assume that the actual input is periodic in time, with period exactly 4 ms. That is to say, the 770 takes data for one 4-ms-wide acquisition interval, and then ‘pastes copies’ of this data into adjacent 4-ms-long blocks of time, creating the (fictitious) ‘periodic extension’ of the data it has captured. That fictitious and infinite data set is periodic by construction, even if what is actually occurring physically is confined to a one-time pulse. So the fictitiously-extended data set *does* have an exact Fourier-series representation, and that’s what you see displayed on the 770.

That works perfectly for a 50.000-kHz input signal, because it has a 20.000-μs period, and exactly 200 cycles of that waveform fit into the 4-ms = 4000-μs time window. So when copies of this 4-ms of data are ‘pasted onto’ either side of the actual data, the resultant mathematical function is *continuous*, in value and slope, at every joining of the 4-ms-wide frames of data.

The same is true to an input frequency of 50.250 kHz, since that fits exactly 201 cycles into the acquisition interval. In this and the previous case, the fictitious infinite data set is just what you’d get from an eternally existing single-frequency sinusoid; and sure enough, there should be only one non-zero Fourier coefficient in its expansion. That’s why you see a beautifully narrow spectral peak, standing on a very low background, in these two cases. But other frequencies, including 50.125 or 50.100 kHz, do <u>not</u> fit an integer number of cycles into the acquisition interval. The particular case of 50.1 kHz its exactly 200.4 cycles into the 4-ms interval. The periodic extension of that 4-ms set of data looks like the figure below, and shows discontinuities in value at the ‘pasting points’ of the extended data set.

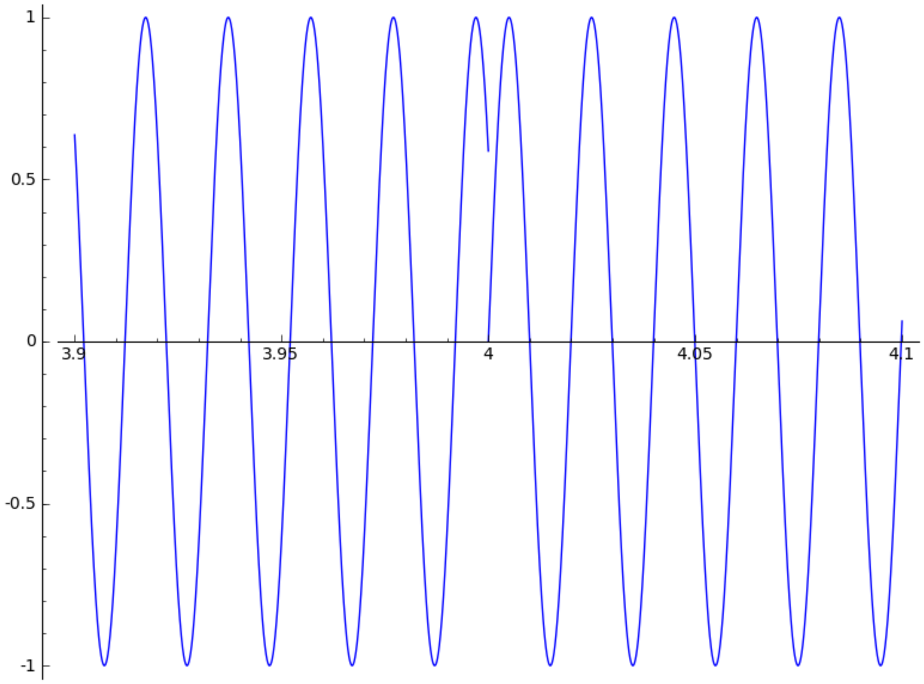


Fig. 2.2: The behavior of a 50.1-kHz sine wave, acquired in a 4.00 ms window and periodically extended, viewed near t = 4.00 ms, showing the *dis*continuous function to which the Fourier series converges

So the Fourier transform ought to give, and does give, the Fourier composition of an extended signal which is periodic, but periodic with period 4.00 ms. The Fourier coefficients exist at each integer multiple of 0.25 kHz, from 0.25 to 100.00 kHz. But there is no guarantee that all the coefficients vanish except for a single non-zero one (as in the previous cases of coefficient #200, or #201). Instead, it takes *lots* of non-zero Fourier coefficients to synthesize the (discontinuous) fictitious function, extended in time, on which the calculation is implicitly being done.

The cure is the use of a *non*-uniform window. To the 4-ms-long data set, *before* the process of Fourier transformation, there is applied a weighting function, which assigns high weights to data values near the middle of the time window, and smaller weights to values near the start and end of the window. *Then* the Fourier transform is applied to this (weighted) data, and the window function can be chosen to assign minimal weight to those values near the boundary times – those very places where discontinuities in the periodic extension cause the spectral leakage. The figure below shows a stylized set of data for sampling, and what it looks like after a certain sort of window is applied.

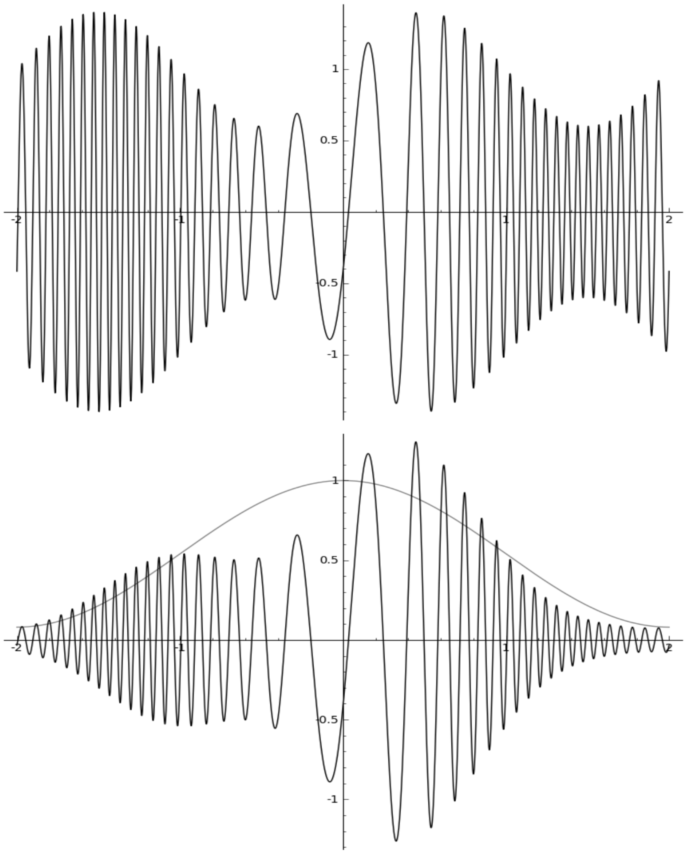


Fig. 2.3: An underlying data set (above), and what it looks like after a Hamming window is applied (below)

Amplitude measurement: another motivation for windowing

One of the many uses of a Fourier analyzer is to serve as a ‘multi-frequency ac voltmeter’, since such a device can measure the magnitudes of *each* of the spectral peaks you find at a number of frequencies of interest. This exercise will show you that if accuracy of amplitude measurement is your goal, you should use the ‘Flattop window’ capability of the 770.

To illustrate this, you need only have a waveform generator, whose frequency you can vary by small increments, while keeping its output amplitude constant. For the example below, you could arrange for the frequency to vary from 500 to 750 Hz, or from 50,000 to 50,250 Hz, or some other range of width 250 Hz. You can of course use an oscilloscope to monitor the sine wave-wave output as you vary the frequency, to confirm that the amplitude, as produced, is staying stable. You might pick an amplitude of about 1 Volt.

Now view your waveform with the 770 as well as the ‘scope, using the ‘full span’ mode with frequency coverage from 0-100 kHz. You should now be familiar with how to see the spectral peak that should result; recall the usefulness of the AutoRange and AutoScale functions built into your device. Now position the marker to find the peak you see on the display, and you'll learn some things:

- The peak location is 'quantized' in units of 250 Hz (so if the actual frequency is 600 Hz, the frequency bin showing the biggest Fourier amplitude will be 500 Hz instead).
- The peak will show an amplitude which is *not* invariant as you very the signal frequency through a 250-Hz range.

The first of these features could of course be cured by 'zooming in' in frequency span, until you had a minimal frequency spacing (called 'linewidth' in the soft menu found under the FREQuency menu) which was much smaller – then you would have higher resolution of frequency readout.

But the second of these features is a real issue, related to spectral leakage. In general, an input signal which does not exactly match the frequency of one of the 'Fourier bins' implicit in you choice of frequency span, will show up with some of its spectral energy spread over a *number* of frequency channels. For an extreme example, if the frequency span is 100 kHz, so the bin width is 250 Hz, and if the input signal is of frequency 50,125 Hz, then equal amplitudes will be indicated at the two nearest bins at 50,000 Hz and 50,250 Hz – but neither of these amplitudes will be the true amplitude of the input waveform! The indicated amplitude in either bin will be well *under* the true amplitude.

Confirm that this effect exists, moving your input frequency to several values spread out over a 250-Hz range, and then repeat the test for all four Window choice the 770 offers you. You should find this leakage effect is worst in the Uniform window, and a smaller problem in the Hanning and BMH windows, but it is almost perfectly cured in the Flattop window – which is, by design, optimized for this sort of measurement. See if you can find any limit on the inaccuracy caused by this sort of bin-mismatching when using the Flattop window.

Clearly the implication is that you should use the Flattop window in any experiment in which amplitude accuracy of a generic single-frequency signal (as opposed to dynamic range, or frequency resolution) is your chief goal.

Chapter 3: Modulated Waveforms – Amplitude Modulation

This section will show you how to create an 'amplitude-modulated' wave, and how such a wave looks in the time domain and in the frequency domain. This method of modulation has given its name to 'AM radio', and it is one of the simplest ways to let a high-frequency (eg. 540 – 1600 kHz) 'carrier' wave be the bearer of low-frequency (eg. audio, 0.3 – 6 kHz) program content. The biggest lesson to be learned is that modulating a carrier wave, even one of fixed single frequency, leaves that wave with a Fourier spectrum which is no longer of zero width. You will see that a modulated carrier now has *sidebands*.

[To show you the relevance of this claim, here's a question for you. Consider a monochromatic laser of optical frequency f_0 , shining its beam through a 'chopper', which alternately opens its window, and then closes it, on a 50:50 basis, at a rate of 10^6 open/shut cycles per second. Now think of the light beam emerging from the chopper, which clearly has only half the average power as before. The question is – what is the optical *spectrum* of that emerging beam? The wrong answer seems logical: Only those photons which get through the chopper can contribute to that spectrum, and each of those photons came out of a laser which produces (only) photons of frequency f_0. So the beam is at half power, but still monochromatic. Plausible, but *false*. The right answer is something you'll learn in this chapter: the light beam emerging from the chopper will have a spectrum which includes frequency f_0, but *also* frequencies such as $f_0 \pm 1$ MHz.]

To work on these exercises electronically, you'll need two signal generators, and we suggest, for the 'carrier frequency', that you use the 770's internal generator, set to give sine waves, of amplitude 1 V, and of frequency near 50 kHz. For the lower frequency, which provides the 'program content', we suggest that you use some function or signal generator, capable of various waveforms, with amplitude of up to 5 V, and frequencies of a few kHz.

Here's what you want to do. Instead of a 'pure carrier' waveform of fixed amplitude A, and fixed carrier frequency f_c, you want to create a waveform given by

$$V(t) = [A\,(1+\alpha\cos(2\pi f_p\, t))\,]\cdot\cos(2\pi f_c\, t) \quad .$$

The final factor describes the carrier wave, described by a cosine oscillating at frequency f_c. But the initial, bracketed, factor shows that the formerly-fixed amplitude A is being replaced by a (slowly) time-varying amplitude $A(1 + \alpha \cos (2\pi f_p\, t))$. So the 'instantaneous amplitude' is now varying slowly between $A(1-\alpha)$ and $A(1+\alpha)$. Here α is a pure number, called the 'modulation index', and often kept to be <1. Furthermore, the time-scale of this variation is what conveys the information content in the slower, 'program-content' waveform. For simplicity, we're choosing, for 'program content', a simple sinusoid of frequency f_p .

Here's the theory of why this AM waveform ends up showing sidebands. Simple trigonometric identities allow the $V(t)$ above to be written as

$$V(t) = A\cos(2\pi f_c t) + (A\alpha/2)\cos(2\pi(f_c - f_p)t) + (A\alpha/2)\cos(2\pi(f_c + f_p)t) \quad ,$$

which now is seen to be, no longer the product of two, but rather the sum of *three*, sinusoids. Of these three terms, the first is just unmodified carrier, and the other two describe sidebands – that is, terms which have two new frequencies (f_c - f_p) and (f_c + f_p), and which therefore lie at two new and distinct locations in frequency space. In fact, they are 'satellites' of the carrier peak, and lie equidistant from it, forming the 'lower sideband' and 'upper sideband' respectively. Notice that the amplitude of the sidebands depends on the overall scaling factor A, but also on the modulation index α, and hence the amplitude of the program content.

Before going on to interpret this theory, it's time to turn it into practice. Note that the operation of *multiplication* is essential to create the AM waveform above. So there's a hardware, analog, Multiplier module among the electronic modules in your apparatus. For any two inputs $V_A(t)$ and $V_B(t)$, this produces a real-time output which is the (scaled) product of the two inputs, with scaling factor

$$V_{out}(t) = [V_A(t) \cdot V_B(t)]/(10\,V) \quad .$$

This model applies, provided that both inputs are kept within the voltage range ±10 Volts, and that their frequency content is limited to about 1 MHz or lower.

But that's not quite enough to give the AM we want. Suppose we have as program-content a waveform coming from an external generator, $P\cos(2\pi f_p t)$, where P is perhaps up to 5 Volts, and f_p is a few kHz. We first need to sum this waveform with a constant dc voltage, perhaps also of size +5 Volts, to form the sum ($5\text{ V} + P\cos(2\pi f_p t)$), and then send that sum into one input of the Multiplier. So find the adjustable DC Voltage supply among your electronic modules, and adjust its level to +5 Volts, and send that to one input of the Summer, and your 'program content' from a signal generator to the other. Build this combination of modules, and practice drawing a useful block diagram of what you have built.

Now take the Summer's output to one of the Multiplier's inputs, and take your carrier wave (from the 770's Source output) to the other input of the Multiplier, and the output of the Multiplier should now be giving you an amplitude-modulated waveform, namely

$$\begin{aligned} V_{out}(t) &= (5\,V + P\cos(2\pi\ f_p\, t)) \cdot (1V\cos(2\pi\ f_c\, t))/(10\,V) \\ &= (\frac{5\cdot 1}{10}V)(1 + \frac{P}{5V}\cos(2\pi\ f_p\, t)) \cdot \cos(2\pi\ f_c\, t) \quad . \end{aligned}$$

This is an AM waveform with a modulation index of $\alpha = P/(5\text{ V})$, with a carrier frequency and a program-content frequency that you can control.

To see that this is really happening, look first with a two-channel ‘scope in the time domain. Let your external-generator output, the low-frequency sine wave, also run ch. 1 of the ‘scope, and have it trigger the ‘scope with this signal. Set the time base for about 0.2 ms/div, so that a few cycle of your program-content waveform fill the scale, and put this display on the upper half of your screen. Now use ch. 2 of the ‘scope to display the Multiplier-module’s output on the lower half of your screen. What do you see? What do you expect to see? How does the display change if you change the amplitude of the modulating waveform? How do you see the cycles of the carrier waveform? Can your eye supply, in the ch. 2 waveform, the ‘envelope’ of the rapidly-varying individual cycles of the carrier wave? You might try out acquisition modes of your ‘scope called single-sweep (to give a one-shot view), or peak (to show not $V(t)$ values, but the $V(t)$ extrema).

When you’ve used this method to check that the Summer, and Multiplier, modules are doing their jobs, take the Multiplier output and send it also to the 770’s signal input. Now you should see the frequency-domain view of what you’d previously seen only in the time domain on your ‘scope. You should see the carrier peak, and the two sidebands. In this picture, it is much easier to see what happens when

- you vary the amplitude of the carrier (by adjusting the 770’s source-out)
- you vary the frequency of the carrier (similarly, by adjusting the 770)
- you vary the amplitude of the program content (by adjusting the output amplitude of your external generator)
- you adjust the frequency of the program content (similarly, by adjusting the frequency of your external generator)
- you change the waveform of the program content (say, to triangle or square wave – and can you now simulate the ‘chopper’ function, of alternately having the carrier present and then absent?)

Can you understand what happens for each change, and why it happens?

Now that you have a picture of sidebands, you can understand why a modulated carrier starts to ‘take up room’ in frequency space. A pure carrier might be monochromatic, infinitely narrow in frequency space; but a modulated carrier takes up more space. If program content is just a sinusoid, of frequency f_p, still the modulated carrier has a frequency content of three peaks, of full width $2f_p$ in frequency space. If the carrier is to be modulated by variable program content, of frequencies up to 6 kHz, then a block of frequency space at least 12 kHz wide has to be allocated to this AM channel. In broadcasting practice, a bit more room is allowed, and as a result, the frequency width of the AM band of (1600 – 540) kHz = 1060 kHz can support at most 1060/20 = 53 separate AM channels. This ‘using up’ of frequency space is one reason that room in frequency space has to be allocated, rationed, or even *bought and sold* in the marketplace (look up ‘spectrum auction’ to see the financial scale). It is also characteristic of any method of modulation – the transmission of information at any given rate, by any given modulation scheme, will take up a bandwidth proportional to the rate of transmission of information. This marks the place where economics – the allocation of an exhaustible resource – comes into communication theory.

Finally, a few words about modulation index. For the numbers listed above (including a dc offset of +5 V into the Summer, and an amplitude P of program-content sinusoid), you have seen that the modulation index comes out to α = P/(5 V).

Let's see first what happens when α is confined to <1; the limiting case of P = 5 V gives α = 1, a 'fully-modulated' AM wave. Under these conditions, the modulated waveform can be written as

$$V(t) = A[\cos(2\pi f_c t) + (1/2)\cos(2\pi(f_c - f_p)t) + (1/2)\cos(2\pi(f_c + f_p)t)] \quad ,$$

which shows that the three-peak frequency spectrum has sidebands of amplitude *half* that of the carrier. This says that the sidebands each have *one-quarter* the power of the carrier, so that of a total power budget of 6 units, the power allocation is: 1 unit to the lower sideband, 4 units to the carrier, and 1 unit to the upper sideband. That is to say, fully 2/3 of the total power lies in the carrier, which transmits no program content at all. This represents a relatively wasteful use of power. And that is a best-case scenario, since *real* program content is more complicated than a simple sinusoid, and the modulation index for 'ordinary content' has to be <<1 if the modulation index is to remain <1 for the strongest program content.

What's so sacred about keeping α < 1? There is no problem in principle, or in Fourier space, in choosing (for example) α = 2; the trigonometry which turns a product-of-two-terms into a sum-of-three-terms is just as valid for any value of α . At the value of α = 2, you can show that both sidebands reach the same amplitude as the carrier, so now 2/3 of the power lies in the sidebands, and only 1/3 lies wasted in the carrier.

The problem with this sort of over-modulation comes in the *time*-domain view of the output waveform, and the use of the envelope of the modulated carrier as a surrogate for the program content. Have a look on your 'scope's ch. 2 to see what happens when you transgress from the α <1 to the α >1 domain. You'll see that for α <1, the envelope of ch. 2 is a faithful copy of the actual program content displayed on ch. 1; but that for α >1, this is no longer true. In the simplest scheme for de-modulation of AM, over-modulation causes a *distortion* of the program content which is eventually to be extracted back out of the received waveform.

Chapter 4: Modulated Waveforms – Heterodyning and Mixing

This Chapter assumes you've worked through Ch. 3 on amplitude modulation, and now introduces a subtly different modulation scheme. The motivation for the difference depends on the context: in communications, it is a more efficient use of power; in signal processing in general, is to make clear the difference between adding two signals (superposition) and multiplying two signals (mixing). In this Chapter you'll learn the very general technique of 'frequency heterodyning' which is used in a host of applications.

Let's start with the language of a higher-frequency 'carrier wave' and a lower-frequency signal, or 'program content'. Let's use the same resources as in Ch. 3: for the carrier, use the 770's internal source, configured to a sine wave of amplitude 1000 mV and frequency near 50 kHz; and for the signal, use an external function generator, for starters, set to give a sine wave of amplitude 5 V and frequency a few kHz.

Now let's *skip* the step of adding a dc offset to the low-frequency signal, and instead, run both the program signal, and the carrier, directly into the two inputs of the Multiplier module. What should emerge? The expected output has the form

$$V_{out}(t) = [(5V)\cos(2\pi\ f_p\ t)]\cdot[(1V)\cos(2\pi\ f_c\ t)]/(10V) \quad ,$$

according to the model previously introduced for the analog multiplier, and using f_{p} and f_{c} for the frequencies of program content and carrier respectively.

Now this waveform is *not* a sum, or superposition; it's a *product*, and a standard trigonometric identity (can you find it? can you *derive* it?) allows it to be rewritten as

$$V_{out}(t) = \frac{5V\cdot 1V}{10V}\cdot\frac{1}{2}\{\cos[2\pi\,(f_c + f_p)\,t] + \cos[2\pi\,(f_c - f_p)\,t]\} \quad ,$$

which is now seen to be the sum of two terms, of new and different frequencies. We can call them the 'sum frequency' and the 'difference frequency'. They appear at exactly the locations you'd expect for the two sidebands occurring in amplitude modulation, but now there is *no* unmodified carrier term midway between them. This motivates the term 'suppressed carrier', and cures the energy inefficiency mentioned in Ch. 3.

To see this in action, set up an arrangement (as in Ch. 3) that will let you see this output, both in the time- and frequency domains. In the time-domain view on a 'scope, ensure as before that you have a ch. 1 display of your 'program content', also triggering the 'scope for a stable display, and devote ch. 2 to displaying the modulated carrier. This time you will see that the envelope of the modulated carrier does *not* have the same shape as the program content. The implication is that re-creation of the program content from the modulated carrier would require something more complicated than the 'envelope detection' which suffices for the $\alpha < 1$ AM method. There are various clever ways to 'de-modulate' your ch. 2 signal, which would be called a 'double-sideband, suppressed-carrier' or DSB-SC signal. Can you invent one of them?

Now step back a bit from communications methods, and consider what the multiplier has done. From two signals, of frequencies f_p and f_c, you have produced an output which contains (ideally) no Fourier components at either f_p or at f_c, but consists of only two components at the novel frequencies, the sum and difference frequencies, $f_c + f_p$ and $| f_c - f_p |$. In fact, it is a searching test of the conformance of the analog hardware multiplier to the pure-product model to look for traces of 'punch-through' of the two frequencies f_p and f_c in the $V_{out}(t)$ signal. You can predict, from the theory above, the expected amplitude of your sum- and difference-frequency signals, and you can check that these 'intermodulation products' appear with the expected amplitude. Relative to these two desired outputs, by how many dB are the two punch-through signals suppressed? You will probably see non-zero amplitudes of these two signals, but they should be much smaller than the two expected signals. Notice that even if the undesired signals *are* present, they are present at frequency locations *distinct* from the desired sum and difference frequencies: so from a frequency-domain point of view, they are easily separated from the desired signals.

Out of two original frequencies, you have created two new frequencies, the sum and difference frequencies. In the trade, this is called 'heterodyning' (heterodyne a word coined to express 'different frequency'). The technique is *very* widely used, most typically to 'down-convert' some high-frequency signal into a frequency range in which electronic processing is easier. Suppose, for example, that you had an incoming signal at frequency 12.345 678 GHz, in the microwave region of the spectrum. It is not so easy to modify, filter, amplify, and process such signals (for one thing, you'd need X-band waveguides merely to conduct them from place to place), and you certainly could not do so with operational amplifiers on a laboratory protoboard! But *if* you had a fixed-frequency microwave source (called a 'local oscillator' in the trade), and *if* could set it to produce a frequency of 12.345 600 GHz, and *if* you had a two-input, one-output device which acted like an analog multiplier for these high-frequency signals, *then* you'd expect to appear at its output a signal composed of only two frequencies: the sum frequency near 24.7 GHz, and the difference frequency, arranged in this case to lie at 0.000 078 GHz = 78 kHz. The first thing to notice is that it would be trivially easy to filter away the 24.7-GHz signal, and leave only the low-frequency 78-kHz signal. The next thing to notice is that it's *really easy* to process a 78-kHz signal – not only can you handle it with ordinary BNC cables and op-amp circuitry, you could put it directly into your 770 and quantify it in detail.

Note too that if the incoming signal were to change by 1% in amplitude, so would your down-converted signal at 78 kHz; furthermore, if the incoming signal were to change instead by rising just 1 kHz in frequency (which is a change of less than 1 part in 10^7 of its frequency!), the down-converted signal would *also* rise in frequency by 1 kHz – but this is now more than a 1% change in the difference frequency, and trivially easy to see. Perhaps you can see why down-conversion by heterodyne methods is a technique so widely used in radio, microwave, and optical communication, and in so many other forms of physics and astronomy instrumentation as well.

You might also check Appendix A6 to learn the difference between the 'difference frequency' produced by mixing, and the 'beat frequency' detectable in a mere superposition.

Mixer technology

Of course, heterodyning depends thus far on a product-operation conducted in an analog multiplier. Your Multiplier module does a very accurate job of realizing the product operation, but only for input signals of frequencies under a few MHz. One of the attractions of the heterodyne technique is that (provided perfect multiplication is not required), there are other ways to form sum and difference frequencies, which can be conducted with simpler circuits than actual multipliers. In particular, the desired mathematical operation can be conducted with 'diode mixers', and these in turn can be made to operate at truly high frequencies, up to dozens of GHz. This immensely enlarges the applicability of heterodyne methods to other regions of the spectrum.

[In fact, *any* electrical device which contains some non-linearity will serve, to some degree, as a frequency mixer. Suppose you had a one-input, one-output device which mapped an input voltage V_{in} to an output voltage V_{out}, according to some functional relationship $V_{\text{out}}(V_{\text{in}})$, and suppose that $V_{\text{out}}(V_{\text{in}}=0) = 0$. Then a Taylor expansion of the functional relationship would give a series

$$V_{out}(V_{in}) = a\,V_{in} + b\,V_{in}^{\;2} + c\,V_{in}^{\;3} + \dots \quad ,$$

where non-linearity consists in non-zero sizes of the *b* or *c* coefficients shown. Now if such a device has even a quadratic non-linearity, and if it is driven with a superposition of two terms, so $V_{\text{in}} = V_1 + V_2$, then the output includes not only terms in V_1, V_2, and V_1^2 and V_2^2, but also (from the quadratic term) the product $V_1\,V_2$. And if V_1 and V_2 have two distinct frequencies, then this product term will contain the desired sum and difference frequencies.]

In many regions of the spectrum, non-linearity is all that's available, or needed – non-linear optics uses laser beams in bulk materials to produce *optical* sum or difference frequencies. But in the realm of electronics, there are diode-based mixers which not only produce the sum- and difference-frequencies, but also do so with adequate efficiency and reasonable suppression of the two input frequencies. They are called 'double-balanced mixers' (balanced, that is, to suppress the punchthrough of both the input frequencies), and there are two of them among your electronic Modules.

The two mixers are adapted to different regions of the frequency spectrum. The one called Audio Mixer works best for frequencies applied to inputs (labeled A and B) in the range 20 Hz – 50 kHz. The other mixer is intended for input frequencies in the 0.025 – 200 *MHz* range; that's 25 kHz to 200,000 kHz. [For historical reasons, the two inputs of this mixer are labeled RF (for 'radio frequency') and LO (for 'local oscillator'). The same tradition calls the output of this mixer the IF (for 'intermediate frequency', referring to its typical use in producing the difference frequency.]

A circuit diagram for a typical diode mixer is more than a little opaque, but it does reveal several features:

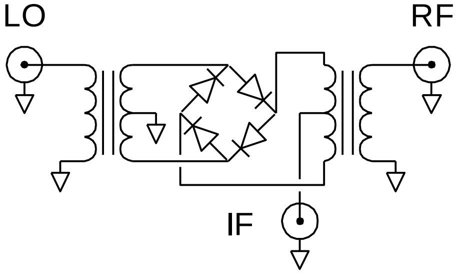


Fig. 4.1: The internal circuit diagram of a generic diode mixer.

- The two inputs are both transformer-coupled to the device's innards, which is why there's a lower frequency limit to the mixer's operation.
- The single output is dc-coupled, so operation down to a difference frequency of *zero* is feasible.
- The internal works of the mixer is not a 'bridge rectifier' of four diodes, but has the topology of a 'ring modulator' – can you see the difference?
- There is no connection to any power supply(!); instead, operation depends on the non-linearity of the diodes as 'passive devices', and the output signal's energy must come from the input signals.
- Operation will require that at least one of the two inputs provide enough voltage to the diodes to get them to conduct.
- There needs to be a return path for currents from the output port; for the Audio Mixer, there's an internal 1-kΩ resistor in parallel with the output for this purpose, but for the High-Frequency Mixer, you need to **add a load**, typically a 50-Ω termination at the far end of the output-signal cable.

Rather than agonize over the behavior of the internal circuitry, we suggest instead the observation of a mixer's behavior from an <u>external</u> point of view. In keeping with the theme of this entire Manual, we'll suggest that you observe a mixer's operation in both the time, and the frequency, domains.

Here's an interesting view of a mixer's behavior in the time domain. Start by characterizing your Audio Mixer, since that permits you to use conveniently low frequencies. Use, as a 'fast' waveform, a 25-kHz sine wave from an external signal generator, and choose an amplitude of about 700 mV (which is enough to get those diodes to conduct). Send that to the *A*-input of the mixer. Now pick, as a 'slow' waveform, a sine wave of frequency around 1 kHz and amplitude of 1 V (or lower), and send that to both the *B*-input of the mixer, and to ch. 1 of a dual-trace oscilloscope. Arrange the 'scope to trigger on this slow waveform, and choose a sweep rate so that one or two full cycles of the sine wave will display. Now send the output of the mixer, marked $A \otimes B$, to ch. 2 of your 'scope.

What do you look for? In your ch. 2 trace, look for a 'chopped' version of the ch. 1, *B*-input signal. You can see that the *A*-input, which goes through full cycles every 40 μs, alternately 'switches' the Output from being a positive-but-reduced copy, to a negative-and-reduced copy, of the signal at the *B*-input. (Change the slow waveform *B*-input temporarily from sine, to triangle, wave, to see this 'copying' operation at work.) You might think of the mixer as a switchable *de*amplifier, alternating between gains near +0.4 and -0.4, in response to the signal at the *A*-input.

The rather sudden chopping you see is the result of the mixer's diodes being put alternately into their conducting and non-conducting regimes. If you vary the amplitude of the *A*-input, you'll find that going up to about 1.2 Volt makes little difference, but going much below 0.7 V will give domains in which the diodes fail to switch properly (particularly for low-amplitude waveforms at the *B*-input). Keep this in mind when you use a diode mixer – the manufacturer will have designed it for a relatively definite level of input signal at one of the inputs (the one that would be called the 'local oscillator' in radio-frequency practice). As another way to see this 'chopping' action, you could try changing the *A*-input drive from sine wave to square wave.

While a diode mixer requires that one input be of sufficient amplitude to turn diodes on, the other input can cope with signals of much lower amplitude. To show this, change your *B*-input signal from Volt-level to 100-mV, then 10-mV, levels of signal. Change your 'scope gains to persuade yourself this 'chopped copying' operation also works for weak signals at the *B*-input. In practice, input signals much smaller even than this can be used.

While these time-domain views show you that the diodes are executing the switching behavior that is intended, they do not teach you about the frequency-domain characterization of the mixer. The 'chopping' action of the diode mixer ought to hint to you that the mixer's output will contain a wealth of high-frequency content. To investigate the frequency content of your mixer output, we suggest that you choose (from your 770 and your external generator) two sine waves, chosen so that their two frequencies, <u>and</u> their sum, <u>and</u> their difference, will all lie in the 2-20 kHz range, and all at distinct and identifiable places. For example, if you choose 1 and 20 kHz for input frequencies, you expect sum and difference frequencies of 21 and 19 kHz respectively, and 'punchthrough' of the two input frequencies at 20 and 1 kHz (ie. all identifiable).

Continue using your Audio Mixer, and remember that at least one of your two inputs needs an amplitude of order 0.7 to 1.0 Volts to get the mixer to operate properly. (The same applies to a radio-frequency mixer, which might recommend that its input labeled with designation LO = local oscillator get an input signal of +7 dBm; that stands for a power level of 5 mW, say into a 50-Ω impedance, and represents a signal of amplitude 0.7 V.) Now send the mixer output not just to a 'scope, but also to the 770. Configure that for either a 0-100 kHz, or later a 0-25 kHz, frequency span.

You will see that your mixer output's frequency-domain content displays a *host* of spectral lines, each at a particular location in frequency space. You'll see just how many such lines occur if you use a Log scale for the vertical-axis display. When you drive a diode mixer with inputs of frequencies f_1 and f_2 ,

- 'punchthrough' of the inputs will show up at f_1 and at f_2 ;
- harmonic distortion, either in the sources or in the mixer, will give terms at $2f_1$, $3f_1$, . . . and at $2f_2$, $3f_2$, . . . (and odd harmonics may predominate);
- the desired outputs, ideally the strongest peaks in the spectrum, will be at $f_1 + f_2$ and at $| f_1 - f_2|$;
- higher-order modulation products can appear too, due to higher orders of nonlinearity of the mixer – look for $2f_1 – f_2$, or $3f_2 + f_1$, or other small-integer combinations of f_1 and f_2 ;
- the generic peak will appear at frequency $m \cdot f_1 + n \cdot f_2$, where m and n are (positive or negative) integers; generally, the smaller the integers, the stronger the peak (and odd integers are favored, too).

If you want to track down the *m*,*n*-assignment of any particular peak, you can 'zoom in' on it in frequency space, and then see how its frequency *changes* when you create a small change Δf_1 or Δf_2 in either of the input frequencies. You expect the peak to move by amount $m \cdot \Delta f_1$ or $n \cdot \Delta f_2$, and so the frequency shift can tell you the values of m and of n.

You can do a rather complete catalog of the peaks you see, and assign to each its integers m and n, and you could try to understand the intensity of each of the many output frequencies in terms of these integers. In practice, all the peaks other than the main sum- and difference-frequency peaks might well be nuisances to you. If you want some justification for neglecting them, change your display from Log to Linear Magnitude, and use the AutoScale button, and you'll get another view of the prominence of the desired peaks.

Now that you know more about what to look for, try the High-Frequency Mixer among your Electronic modules, using two frequencies in the range 30-50 kHz. Here the terminology is different, as the inputs are called RF for radio-frequency and LO for local-oscillator, and the output is called IF for intermediate-frequency. Recall you'll need some load, or current-return path, attached to the output of this mixer.

The most typical application of a mixer is to aid in the detection of a weak incoming signal, using an adequately strong and fixed-amplitude local-oscillator signal. Here the typical technological application is to isolate the difference frequency output, called 'intermediate frequency' (*not* because it's intermediate between f_1 and f_2, but because it is intermediate between these two high frequencies and dc or zero frequency). For example, an FM table radio accepts input signals in the range 88-108 MHz, and mixes them with a local-oscillator frequency which can be tuned through the range 77-97 MHz. The difference frequencies fall in the range under 31 MHz, but only those stations which give rise to difference frequencies near 11 MHz will create signals which will be amplified by the 'intermediate frequency amplifier' which follows the mixer in the signal-detection chain.

And that brings up another illustration of the value of heterodyne detection. Suppose you had a *weak* rf (radio frequency) signal of small amplitude A_{rf}, and frequency f_{rf}, and a local oscillator of amplitude A_{lo} and frequency f_{lo}. If you were to try to detect the weak rf signal *directly*, you'd have a signal of very small power $\propto A_{\mathrm{rf}}^{\ 2}$. If, on the other hand, you put the two signals into a multiplier or mixer, you can get out the difference frequency $|\, f_{\mathrm{rf}} - f_{\mathrm{lo}}\, |$, which has not only been arranged to have a more convenient location in frequency space, but is *also* going to appear with a not-so-small amplitude $\propto A_{\mathrm{rf}}\, A_{\mathrm{lo}}$. This is now *linear*, not quadratic, in the small quantity A_{rf}. This is yet another advantage of the heterodyne technique, in making possible the detection of weak signals.

Mixers as phase detectors

One of the technical specifications of any mixer is the frequency coverage intended for its two inputs (traditionally labeled RF and LO), and for its output (traditionally labeled IF). Because of the use of transformer coupling to the internal diode circuitry, frequency coverage will be limited for any given mixer.

One of the reasons you might care about the frequency coverage of the IF output is the clever use of a mixer to 'mix' two signals of the same frequency. This means the ordinary sum- and difference-frequencies that are expected to emerge at the output will be (respectively) double the input frequency, and the dc or zero frequency. Your Multiplier, and your two Mixer units, all *do* have dc-coupled capability at their outputs. So for example when the Multiplier's *A*- and *B*-inputs are driven by $V_A \cos(2\pi f_A t)$ and $V_B \cos(2\pi f_B t - \phi)$, the output is expected to be

$$\begin{aligned} V_{out}(t) &= [V_A \cos(2\pi f_A t)]\cdot[V_B \cos(2\pi f_B t - \phi)]/(10\,V) \\ &= \frac{V_A V_B}{10\,V}\cdot\frac{1}{2}\{\cos[2\pi(f_A - f_B)t + \phi] + \cos[2\pi(f_A + f_B)t - \phi]\} \quad , \end{aligned}$$

so that if the two input frequencies are equal ($f_A = f_B$), the result becomes

$$V_{out}(t) = \frac{V_A V_B}{20\,V}\{\cos[\phi] + \cos[2\pi(2 f_A)t - \phi]\} \quad .$$

Apart from the term of high frequency $2f_A$ (which is easily filtered away), the result is a term proportional to the (cosine of the) phase difference ϕ of the two inputs. The (filtered) Multiplier output, in other words, displays a constant value, which is positive and maximal for two signals in phase, negative and minimal for two signals 180° out of phase, and zero for the intermediate case of two inputs 90° out of phase. This detection of the relative phase of two signals is often used in 'phase-locked loops', in which the feedback which forces the phase-detector output to stay at value zero has the effect of locking the two wavetrains to possess the same frequency (and fixed to have 90° relative phase).

If you have two digital signal generators, each of which can be trusted to produce frequencies with part-per-million precision and stability, you can set both to 5- or 10-Volt amplitudes, and both to the same (nominal) frequency, perhaps 12,345. Hz. If you send the two generator outputs to channels 1 and 2 of a ‘scope, and you trigger the ‘scope on ch. 1, you’ll see a stable ch. 1 display. On ch. 2, however, you may see a slow drift of the ch. 2 waveform, precisely because it comes from a generator running at its *own* idea of 12,345. Hz; this may be different (on the order of a few parts-per-million, here of order 10^{-2} Hz) compared to the other generator. If you also run these two outputs to your Multiplier, and attach a digital multimeter to monitor the Multiplier’s output, you will see the slow variation in output which arises from the slowly varying phase difference (which is also visible in real time by comparing your ‘scope’s chs. 1 and 2 displays).

Try the same experiment with your Audio Mixer, except this time, set the two input frequencies deliberately to be about 1 Hz apart (and change the signal amplitudes to about 1 V). Send the mixer’s output to a ‘scope, and you’ll see some complicated behavior. Part of that behavior will be due to the presence of the *sum* frequency, here about 25 kHz. So to clean that up, send the mixer output through the Filter module, set to a frequency of 1 kHz, a Q of 0.7, and using the Low-Pass output. Now you’ll have filtered away the sum frequency, and left behind the difference frequency, which should display a sine wave of about 1 Hz frequency. The period of this waveform is a *very* sensitive indicator of the exact frequency difference between the two generators. Now set the two generators back to the *same* nominal frequency, and use the ‘scope with a slow 10 s/div sweep speed, to watch the output reveal the *very* small frequency difference the two generators might actually exhibit.

Mixers as frequency doublers

What else happens if you put the same frequency into both inputs of a mixer? The previous section focused on the resulting ‘dc’ difference frequency, but now think instead about the *sum* frequency, which will be at double the frequency you put into the two outputs. The ‘multiplier model’ of a mixer shows you that

$$\cos(2\pi\ f\ t)\cdot\cos(2\pi\ f\ t)=\cos^2(2\pi\ f\ t)=\frac{1}{2}+\frac{1}{2}\cos(2\pi\cdot 2f\cdot t) \quad ,$$

so if you filter out the dc component, what’s left is a signal of frequency 2*f*. Or, if you can put a 90° phase shift into the path of one of the two inputs, and you’ll get an application of

$$\cos(2\pi\ f\ t)\cdot\sin(2\pi\ f\ t)=\frac{1}{2}\sin(2\pi\cdot 2f\cdot t) \quad ,$$

which gives ‘pure doubling’, and an output lacking any dc offset. Either way, you can get a waveform with a frequency in a domain which might have been *in*accessible to you without the use of the mixer. There are regions of the electromagnetic spectrum for which this technique provides the easiest way to reach a given target frequency.

Chapter 5: Modulated Waveforms – Frequency Modulation

You are culturally aware that information is broadcast via electromagnetic waves by both 'AM' and 'FM' methods. You've seen the Fourier view of AM spectra in Section 3, and now you'll see how FM, frequency modulation, looks in the frequency domain. The relevance of this treatment extends beyond broadcasting technology to a host of physical systems which can be understood by a frequency-modulation model.

[But first an exercise to get you to think about FM. You know about the Doppler shift: how a source of electromagnetic radiation of frequency f_0 (when at rest), but moving at velocity v, will deliver, to a stationary receiver, radiation which is Doppler-shifted to a new frequency $f = f_0\ (1 \pm v/c)$, 'red shifted' or 'blue shifted' according to a receding or approaching source. What if the source is alternately doing both, by oscillating in position, alternately approaching and receding? This really happens, for example when a ^{57}Fe excited nucleus emits a γ-ray photon, of energy 14.4 keV, from inside an atom which is vibrating because it's in a crystal lattice at room temperature. Now compute: that γ-ray energy gives the photons a frequency of $f_0 = 3.48 \times 10^{18}$ Hz. It's easy to show (from the equipartition theorem) that the lattice vibrations entail oscillating motions, at speeds v of about ±200 m/s, at lattice vibrational frequencies of order 10^{12} cycles per second. Thus you see that the ratio v/c takes on a continuous *spread* of values, from -0.7×10^{-6} to $+0.7 \times 10^{-6}$. So due to the Doppler effect, you might expect to receive, from such a source, a *spread* of frequencies f, ranging up and down from f_0 by an amount of about $(0.7 \times 10^{-6})(3.5 \times 10^{18}$ Hz) or about $\pm 2.5 \times 10^{12}$ Hz. Right? No, quite wrong – the γ-rays you receive will include a component whose spectrum is *un*spread, of spectral width less than 10^7 Hz, centered right at f_0. This is better than *10^5* times more monochromatic than you expected – but why does this occur?]

Thus far you've seen an unmodulated 'carrier' waveform written as

$$V(t) = A\cos(2\pi f_c\, t - \phi) \quad ,$$

and you might suppose that the technique of FM is to write

$$V_{FM}(t)\ ?=?\ A\cos[2\pi f(t)\, t - \phi] \quad ,$$

where $f(t)$ is a time-varying frequency. But a more fundamental view is to think of the entire argument of the cosine function as the 'phase function' of the oscillation, φ(t), and to write an unmodulated carrier as

$$V(t) = A\cos[\varphi(t)] \quad , \quad \text{with} \quad \varphi(t) = 2\pi\ f_c\, t - \phi \quad .$$

Note that φ(t) grows *linearly* with time for an unmodulated carrier; perhaps you can see why the $-\phi$ part of the phase function is called the 'phase constant'. (See Appendix A2 for details.) But you'll note that for this model,

$$\frac{1}{2\pi}\frac{d\varphi(t)}{dt} \equiv f_c \quad ,$$

and we hereafter use this to *define* what we mean by the 'instantaneous frequency' of a carrier wave. So if we want frequency modulation, we need not a constant frequency f_c, but instead a varying (instantaneous) frequency, such as

$$\frac{1}{2\pi}\frac{d\varphi(t)}{dt} = f_{inst}(t) = f_c + \delta f \cos(2\pi \; f_m \; t) \quad .$$

This expression is the simplest example of FM, in which the instantaneous frequency varies sinusoidally, centered about a central value f_c, with a 'peak frequency deviation' δf, covering the frequency range f_c - δf to f_c + δf. The other new variable is f_m, the 'modulating frequency'; this is the number which would take on the value 440 Hz for the transmission whose program content is an orchestra tuning up. (Notice that the choice of δf is a separate matter of technique, and will be limited by the amount of 'frequency space' the FM technology is allocated.)

Now we can integrate the phase equation with respect to time to get

$$\frac{1}{2\pi}\varphi(t) = f_c \; t + \delta f \cdot \frac{1}{2\pi \; f_m}\sin(2\pi \; f_m \; t) + const \quad ,$$

so with a choice of the constant (equivalent to a choice of time-origin), we get

$$\varphi(t) = 2\pi f_c \; t + \beta \sin(2\pi \; f_m \; t) \quad .$$

Here the dimensionless number β is called the 'modulation index', and is given by

$$\beta \equiv \text{(peak frequency deviation) / (modulating frequency)} = \delta f / f_m \; .$$

Then the waveform, still assuming a constant (and thus *un*modulated) amplitude A, can be written as

$$V_{FM}(t) = A\cos[2\pi f_c \; t + \beta \sin(2\pi f_m \; t)] \quad .$$

This is a complicated waveform (a sine function inside a cosine function's argument!) for generic β-values; but for small β, we can work out its Taylor expansion:

$$\begin{aligned} V_{FM}(t,\beta) &= V_{FM}(t,0) + \frac{1}{1!}\beta^1 \frac{\partial V_{FM}}{\partial \beta} + \frac{1}{2!}\beta^2 \frac{\partial^2 V_{FM}}{\partial \beta^2} + ... \\ &= A\cos[2\pi \; f_c \; t + 0] + \beta A \cdot -\sin[2\pi \; f_c \; t + 0][0 + 1\cdot \sin(2\pi \; f_m \; t)] + ... \\ &= A\cos(2\pi \; f_c \; t) - \beta A \sin(2\pi \; f_c \; t)\sin(2\pi \; f_m \; t) + O(\beta^2) \quad . \end{aligned}$$

The terms through order β^1 can be written, via a trigonometric identity, as

$$V_{FM}(t) \cong A\cos(2\pi \; f_c \; t) - \frac{A\beta}{2}[\cos(2\pi(f_c - f_m)t) - \cos(2\pi(f_c + f_m)t)] \quad .$$

This form shows us the frequency content:

the carrier, at frequency f_c, with amplitude *A*;

and two sidebands, at frequencies $f_c \pm f_m$, with amplitudes $\pm A\cdot\beta/2$.

[The <u>sign</u> of the amplitude coefficients is significant – see Appendix A7 for an interpretation of the *phases* of the sidebands relative to the carrier.]

This expansion can be carried to higher orders in β by Taylor-expansion methods, but the results are much more compactly written using a wonderful Bessel-function expansion discussed in Appendix A8. The result is

$$V_{FM}(t) = A\cos[2\pi\, f_c\, t + \beta\sin(2\pi f_m\, t)] = \sum_{n=-\infty}^{\infty} J_n(\beta)\cos[2\pi(f_c + n\, f_m)\, t] \quad .$$

Now we see, for general β, an infinite series of sidebands, where the $n = 0$ term gives the carrier at frequency f_c, the $n = \pm1$ terms the sidebands at $f_c \pm f_m$, the $n = \pm2$ terms giving sidebands at $f_c \pm 2f_m$, and so on. In principle the sidebands extend indefinitely far from the carrier, but in practice, the amplitudes are very small for $|n| > \beta$. The results $J_0(0) = 1$, $J_1(\beta) \approx \beta/2 + \ldots.$, and $J_{-n}(\beta) = (-1)^n J_n(\beta)$ give back the results derived above by Taylor expansion. But the Bessel-function values are calculable for *any* values of β and *n*, so **the spectrum of a frequency-modulated waveform can be predicted in detail.** The prediction is a central carrier, and (fanning out around it) a symmetrical family of sidebands, carrying an increasing amount of spectral power as the modulation index β grows.

<u>Generating an FM waveform</u>

We have made provision, among the Electronic Modules, for the systematic investigation of frequency modulation. The 'Voltage Controlled Oscillator' is a module capable of delivering sinusoidal output waveforms, of knob-variable amplitude *A*, with frequency in the 0-100 kHz range. The instantaneous frequency is controlled by a range switch, by the frequency fine-adjustment knob, *and* by a control voltage you can inject. It is easy to test the operation of the VCO: turn its range switch to the 20-100 (kHz) position, set its fine-adjustment dial to mid-range, and look, using a 'scope and the 770, at the Output. You should see a sinusoidal output waveform, with frequency about 60 kHz. You may turn the fine-adjustment dial until you get a spot-on 50-kHz output.

There still remains the Modulation Input to use as an independent variable; now you can try driving this input using the DC Voltage module to get a steady voltage lying in the ±5-V range. You should verify that for every setting of the control voltage, you get a sinusoidal output with the same knob-chosen *amplitude*, and furthermore, that this control voltage steers the *frequency* of this sinusoid within the 20-100 kHz range. For control voltages not too far from zero, the mapping you'll find is linear, and so you can model the VCO as producing an output frequency

$$f_{out} = 50 \text{ kHz} + s \cdot V_{in} \quad ,$$

where s is the sensitivity of the oscillator's frequency to control-input voltages. You might find $s \approx 3$ kHz / Volt, so that each Volt of control input moves the output frequency by about 3 kHz.

Now the response of the VCO to changes in V_{in} is nearly instantaneous – on the range you're using, f_{out} adapts to changes in V_{in} within a few μs. (Can you find a way to test this claim? Think about square-wave input to V_{in}, and a single-shot 'scope view of the VCO's output around the time that this square wave changes sign.) So your choice of V_{in} is *not* limited to dc values, but could include time-dependent voltages, of which the simplest form is

$$V_{in}(t) = 0 + a\cos(2\pi \ f_m \ t) \quad .$$

Here a is an amplitude (up to a few Volts) and f_{m} is a modulating frequency (perhaps 440 Hz or some audio frequency, but well below the carrier frequency of 50 kHz), so now you expect the VCO to deliver an instantaneous output frequency of

$$f_{inst}(t) = 50 \ kHz + s \cdot a \cos(2\pi \ f_m \ t) \quad .$$

This ought to give the VCO output the character of pure FM, producing a waveform of a fixed amplitude A you have already found, with a carrier frequency of $f_{\text{c}} = 50$ kHz, but with a whole family of sidebands located in frequency space at $f_{\text{c}} \pm n\, f_{\text{m}}$. This waveform has peak frequency deviation $\delta f = s \cdot a$, so its modulation index is $\beta = \delta f / f_{\text{m}} = s \cdot a / f_{\text{m}}$, and thus the magnitudes of the Fourier components in the spectrum of the VCO output are predicted to be,

at frequency $f_{\text{c}} + n{\cdot}f_{\text{m}}$, a component of strength $A\, J_{\text{n}}(\, s \cdot a / f_{\text{m}})$.

In this expression, the VCO's output amplitude A and central frequency f_{c} are set by knob choices, and its sensitivity to voltage-control s is fixed, but the amplitude and frequency, a and f_{m}, of the *modulating* waveform are free for you to choose.

We suggest that you start with f_{m} fixed, perhaps to 2 kHz, using an external signal generator as a source. Now you can view that 2-kHz generator output on a 'scope, and vary its amplitude a at will. We suggest you start with a small (<< 1 Volt) amplitude.

Now look as the VCO's output in the frequency domain, using the 770 as a 'Fourier spectrometer'. A span of 50 or 25 kHz, centered around 50 kHz, is appropriate. Since you'll be trying to measure the magnitudes of Fourier components, the 'Flattop window' is appropriate – use the FREQuency menu button, and Window softkey, to select this.

Get a look at the spectrum on your 770, and watch it change in real time as you increase the amplitude a. You should see a growing family of sidebands, with a recognizable progression of spectral power out of the carrier at f_{c}, into the sidebands at $f_{\text{c}} \pm n\, f_{\text{m}}$. (Are the sidebands *in fact* spaced this way, as the mathematical model claims?) In fact, you can raise the modulating amplitude a far enough to *extinguish* the carrier: since $J_0(x)$ has its first zero at $x = 2.4048\ldots$, you can predict

$s \cdot a / f_m \approx 2.40$ for ‘carrier suppression’

and with s = 3 kHz/Volt, and f_m = 2 kHz, an amplitude of $a \approx 1.6$ Volt should bring you to this condition. (Appendix A8 describes the physics of why this carrier-suppression occurs.)

In fact you can map, systematically as a function of *a*, the strength of each sideband, and compare to the Bessel-function model above. You’ll find you can go beyond ‘carrier suppression’; you can bring the innermost sidebands from near-zero, to a maximum, to *their* first suppression condition. You’ll find you can create *many* sidebands, spreading power over a wide spectral range.

Thus far in raising the modulating amplitude *a* up from zero, you’ve started with FM of low modulation index, and then watched as this β-parameter grew upward from zero. Now here’s a way to see the opposite limiting case, reducing β from infinity, and also to make some connection with intuition.

For this case, fix your modulating amplitude *a* to some value (such as 1 Volt), and imagine starting with some *very* low modulating frequency (such as f_m = 0.1 Hz). (What value does this predict for the modulation index β?) Now, over a 10-second period, the modulating voltage varies cyclically between -1 and +1 Volt, and the VCO’s output frequency ought to vary between (say) 47 and 53 kHz. (Why?) In fact, it ought to vary slowly enough that you can see the VCO’s output spectrum as a single peak, but a peak whose location wobbles back and forth in the 47-53 kHz range, requiring 10 seconds for a full cycle of variation. To see this, set the 770 to a frequency span of 12.5 kHz or more (so that the 770 gets a fresh acquisition of data every 32 ms or less, giving at least 30 fresh frames per second to give you a movie-like view of ‘live action’).

Once you’ve seen this migrating spectral peak, start raising the modulating frequency f_m from 0.1, to 0.2, 0.5, 1., 2., 5., 10, 20, 50, 100 Hz and more. What do you see? At first, a peak with a wobbling center location; later, the range $f_c - \delta f$ to $f_c + \delta f$ getting ‘filled in’ with spectral power. But eventually, your spectral resolution lets you see that the power actually lies (only!) in the family of sidebands predicted above. For any acquisition or averaging time longer than $1/f_m$, the VCO’s output waveform goes through its whole cycle of frequency variation, and its Fourier spectrum is restricted to a family of sidebands – resolvable, or not, depending on how you have set your 770. To confirm this, go back to a low modulating frequency, such as 20 Hz, where you *seemed* to get a *continuum* of spectral power in the range 47 to 53 kHz. Now zoom in, in frequency span, until you can see, ie. resolve, this spectrum, to test the claim that FM, even of high modulation index, gives *not* the smear of frequencies you might have expected, but a family of resolvable sidebands around the carrier.

[And going back to ^{57}Fe γ-rays: notice that so long as the modulation waveform is periodic, the frequency-modulated spectrum, viewed at sufficient spectral resolution, always includes a spectral peak at the unshifted *and unbroadened* carrier frequency. The success of the famous Mössbauer experiment done with these γ-rays can now be attributed to there being due to a substantial fraction of the spectral power left as-is in the unmodified carrier, instead of being splattered into vibrational sidebands by the Doppler effect attributable to thermal vibrations.]

Now stepping back from the details of Bessel functions, here's another problem for you to consider. Suppose your co-worker had used such a VCO as yours to generate a frequency-modulated carrier, centered around 50 kHz, conveying some program content unknown to you. If all you could receive was this frequency-modulated carrier,

- what would it look like on a 'scope?
- what would it look like on your 770?
- how could you tell if it was a modulated (as opposed to *un*modulated) carrier?
- how would you *de*-modulate it, that is, recover the encoded program content?

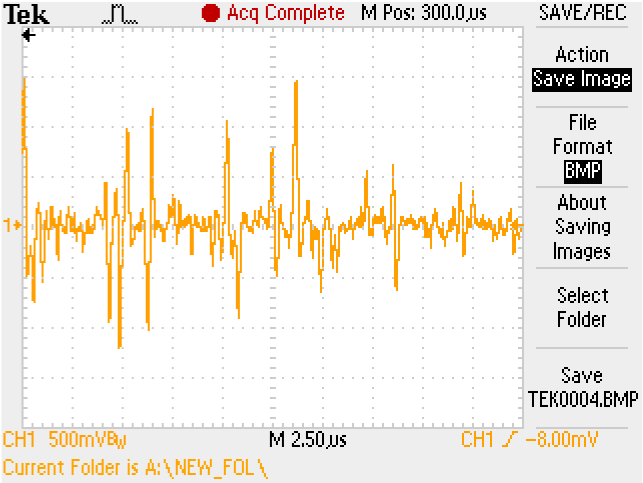
One method for demodulating an FM carrier is illustrated in Chapter 18 as an optional project.

Chapter 6: Noise Waveforms

Thus far you've used an oscilloscope for time-domain views, and the SR770 for frequency-domain views, of sinusoidal or periodic signals. The Fourier spectra of such signals are line spectra: there are narrow 'spectral lines' at the fundamental frequency f_1 [f_1 = 1/(period T)] , and also at its integer multiples, the harmonics of the fundamental. But there is nothing in *between* these isolated narrow peaks – in optical terms, these are bright lines sitting against a dark background. Now you are about to see that non-periodic waveforms will yield Fourier spectra which are continuous, not discrete.

The ability to display and quantify waveforms with continuous spectra is especially important for the *noise waveforms* that are so common in physics, engineering, and communication. Whether noise is of interest in its own right, or whether it is just a problem in competition with a desired signal, it still needs to be detected and measured. In this section, you'll see how that's done, and the new units in which it's done.

But first you can have a look at an actual noise waveform. Among your Electronic Modules is one we've called Buried Treasure, with a single BNC output. (The 'treasure' you'll discover is a sinusoidal signal, and the 'burial' is under stronger electronic noise. See Ch. 15 for a fuller description.) Connect that to a 'scope and to the 770, and set the module's rotary switch to the Noise position, and now look at what your 'scope shows. You should see a fuzzy waveform resembling Fig. 6.1 below; note that this is a view using 500 mV/div vertically, and 2.5 μs/div horizontally. (The waveform's appearance will *change* if you try much faster or much slower sweep rates – try it out.) If you arrange to trigger your 'scope on occasionally-occurring large positive excursions of this signal, and use a fast enough sweep, you'll see the waveform is made of a irregular series of voltage spikes, of both polarities, occurring at apparently random times. [These spikes are due, in the case of this module, to breakdown in reverse-biased Zener diodes, so the unpredictability is quantum-mechanical.] That is enough to make the whole waveform unpredictable, and yet of stable statistical properties, as is characteristic of 'stationary noise'. In this case, the noise is also 'spectrally white', at least out to frequencies of about 2 MHz.



M Pos: 300.0μs

Action Save Image

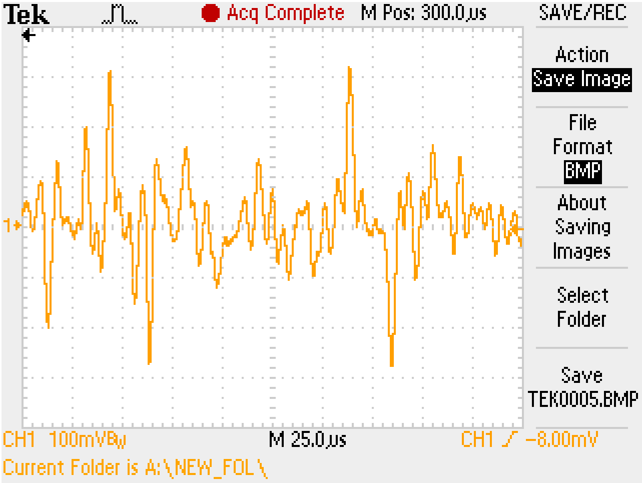
File Format BMP

CH1 500mV Bw M 2.50μs CH1 ƒ −8.00mV

Current Folder is A:\NEW_FOL\

Fig. 6.1: A view of the Noise output, at 500 mV/div and 2.5 μs/div.

Those sharp spikes in the time domain correspond to high-frequency content in the frequency domain. But noise waveforms, like any other waveforms, can be filtered, and in this case you'll want to change the rotary switch to the Filtered Noise position. This will filter out the highest-frequency components of the noise, but leave unchanged all the components under 100 kHz to which your 770 is sensitive. In the time domain, the 'scope view of this signal display less spiky behavior, and also a smaller characteristic size – note the new units of vertical (and horizontal) scale in this view:



CH1 100mV B_W M 25.0µs CH1 ⟋ −8.00mV
Current Folder is A:\NEW_FOL\

Fig. 6.2: A view of the Filtered Noise output, at 100 mV/div and 25 µs/div.

To get the 770's idea of the Fourier spectrum of this waveform, make the following selections: use the MEASure button, and under the Measure softkey menu, choose PSD or power spectral density; under the Units menu, chose Volts rms; under the Window menu, chose BMH. Also use the FREQuency button, and the softkey menu choose Full Span (0-100 kHz coverage). Finally, use the AutoRange and AutoScale buttons to get a good display of the spectrum. The 770 will digitize and Fourier-transform the incoming data stream using the same hardware and software as formerly – it can't know you're calling this a noise waveform. But it will present the results with the newly appropriate vertical-scale units, labelled Vrms/√Hz. You'll soon learn what this strange unit means. The display will also show the *randomness* of the noise via the fluctuations you see, with values on the plot varying considerably both from frequency to frequency, and from acquisition to acquisition. This is characteristic of noise – even if its long-term statistical properties are entirely stable, you will still get fluctuations from one particular statistical sample to the next.

In such circumstances it is very convenient to assume that long-term averages do exist, and to find them by using the AVERAGE button of the 770. Choose among the softkeys to put averaging ON, and to use 64 averages, and to use the rms averaging type, and the Exponential averaging mode. You'll see that it takes some time (of order 64 times the 4-ms acquisition time you're using), but that you'll get a running-average of the spectrum with fluctuations that are now 'averaged down' (by a factor of √64 =8, as it happens).

In the averaged view of the power spectrum, look for the emergence of a 'flat spectrum', denoting equal spectral power in each of many frequency bins of equal width. This is the meaning of 'white noise'. (The phrase is meant to be parallel to 'white light', ie. light having an (equal) combination of all the spectral colors.)

For more exercises in measuring 'power spectral density' with the SR770, refer to the SRS Operating Manual, in its section 'Getting Started', at pp. 1-7 through 1-9 for an exercise in measuring the spectrum of noise.

[If you'd like to practice on another noise source, use the Source Out function of the 770, and configure it to produce, not sinusoids as formerly, but noise. You can specify 'white' or 'pink' noise, and can specify the amplitude (actually, the rms measure) of the noise in the 1-1000 mV range. Convey the source-output to the usual A-channel input, and look at the spectrum with the tools you've now learned to use. To see the difference between 'white' and 'pink' noise, it might be best to use the option of the logarithmic scale for the horizontal axis. Just as 'pink light' is like white light, but with extra red, or low-frequency, content, so 'pink noise' has a spectral density which is not flat-in-*f* (in the 0-100 kHz range) but having an excess at low frequencies. See if you can establish a power-law dependence in the voltage spectral density you are seeing, and find what power-law exponent was chosen in this realization of pink noise.]

For more exercises in using the Noise configuration of the SR770's internal Source, refer to the SRS Operating Manual, in its section 'Getting Started', at pp. 1-39 through 1-41, for instruction in using its internal noise source.

<u>The units of power spectral density and voltage spectral density</u>

What of those obscure units, Vrms/√Hz ? Let's think about a monochromatic sinusoid, perhaps of 1-Volt amplitude. You've used vertical-scale units such that you can read this amplitude directly. You could of course change the vertical-scale units from Volts to Volts(rms), and for this sinusoid, you'd have gotten 0.707 V for the 'rms measure' of the sinusoid. But amplitude is not a useful measure for noise, whose waveform might have a probability distribution, but certainly lacks any useful amplitude or maximal positive excursion. But it still has an rms measure, and here's why:

What a noise waveform certainly has is an instantaneous voltage function $V(t)$, whose mean value is zero, but which would nevertheless convey instantaneous power $[V(t)]^2/R$ if sent into a resistance R. And the time-averaged power, written as $< [V(t)]^2 / R >$, is definite and non-zero: the mean of the voltage-squared is *not* zero (even if the square of the voltage-mean *is* zero). In fact <u>mean-square voltage</u> is a measure of a signal which is applicable to any waveform, periodic or not, predictable or noisy.

For periodic signals, it has become conventional to compute this mean-square value, and then to take the square root of that, to produce the *root*-mean-square or rms measure of a signal. For noise signals, it's more suitable to stick with the mean-square measure of a signal, even if it has the units of Volts-squared or V^2, since that number is a direct surrogate for power.

Mean-square voltage also an additive measure for noise. That is to say, if you were to sum two independent noise signals, to form the superposition

$$V(t) = V_1(t) + V_2(t) \quad ,$$

then the mean-square measure of the result is given by

$$\left\langle V^2(t) \right\rangle = \left\langle [V_1(t) + V_2(t)]^2 \right\rangle = \left\langle V_1^2(t) \right\rangle + 2\left\langle V_1(t)\, V_2(t) \right\rangle + \left\langle V_2{}^2(t) \right\rangle \quad .$$

In this last expression, you see a cross-term < $V_1(t)$ $V_2(t)$ >, and this time average will be *zero* for uncorrelated signals. In fact it's used as a *definition* of the absence of correlation, since to get < $V_1(t)$ $V_2(t)$ > greater than 0, we'd have to require that $V_1(t)$ and $V_2(t)$ were more often of the same sign (giving a positive product) than of opposite signs (giving a negative product) – and that sort of cooperative-sign conspiracy is not possible for statistically-independent signals. So for independent noise signals, noise power *in the mean-square sense* is additive.

Furthermore, this mean-square measure is divisible, in that the power actually present in the noise signal can always be dis-aggregated or 'binned' on the basis of its frequency content. You can imagine a signal $V(t)$ being sent into a whole parallel bank of filters, in which

Filter #1 passes all, but only, the 0-250 Hz components of $V(t)$;
Filter #2 passes all, but only, the 250-500 Hz components of $V(t)$;
etc.
and Filter #400 passes all, but only, the 99,750-100,000 Hz components of $V(t)$;
and so on.

Each of 400 output signals $V_j(t)$ would have its own mean-square measure < $[V_j(t)]^2$ >, and the sum of those mean-square measures would give the total power, present in the 0-100 kHz band, in the original signal. This sort of fanciful exercise in filter hardware is a very close approximation to what your 770 is actually doing, but via FFT software.

For a filter whose 'passband' is of width Δf, we'd expect that the value of < $[V_j(t)]^2$ > emerging from it would itself be proportional to Δf. (By power additivity, you'd expect that doubling the width of a frequency bin to which noise is applied would indeed double the mean-square measure of the signal emerging.) This motivates forming the quotient,

$$< [V_j(t)]^2 > / \Delta f \, ,$$

which provides a useful measure of 'power spectral density', with units of Volts-squared per Hertz, or V^2/Hz, usually denoted S for spectral density. This local density of spectral power might be a constant, as in white noise, or it might depend on frequency, as in $S(f)$, but the point of S's definition is that

$$\int_0^\infty S(f)\, df = \left\langle V^2(t) \right\rangle \quad .$$

This says that the total power in the signal $V(t)$, given via its mean-square measure, can be dis-aggregated into its frequency distribution, and that integrating over all frequencies gives back the total. That is to say, *S*(*f*) *df* gives the part of the mean-square measure of *V*(*t*) which is attributable to frequency components in the range (*f*, *f* + *df*).

Finally, taking the square root of the numerical measure of *S* gives a result, sometimes called 'voltage spectral density', with units of $(V^2/Hz)^{1/2}$ = Volts per square-root-of-Hertz, or V/√Hz. The 770 labels these units Vrms/√Hz, the idea being that squaring the number (undoing the 'root' part) gives back a result which is a mean-square voltage per Hertz of bandwidth.

To see this work on the Filtered-Noise waveform you've been watching, use the Average function, and watch your Fourier spectrum settle to give a voltage spectral density of about 220 μVrms/√Hz, approximately uniform over the 0-100 kHz frequency range. (Results from *your* Zener sourrce may differ somewhat from this nominal value.) The square of this number is $(2.2 \text{ x } 10^{-4} \text{ V}/\sqrt{\text{Hz}})^2 = 4.8 \text{ x } 10^{-8} \text{ V}^2/\text{Hz}$, which is a power spectral density, valid over the 0-100 kHz range (and an unknown frequency distance higher in frequency space). So the integral

$$\int_0^{100\,kHz} S(f)\,df \approx (4.8x10^{-8}\,V^2/Hz)\cdot(10^5\;Hz) = 4.8x10^{-3}\,V^2 \quad [=(0.07\,V)^2]$$

would give the mean-square measure of the signal which is the 0-100 kHz filtered version of *V*(*t*).

This complicated procedure is needed precisely because a noise waveform lacks an amplitude, and also fails to contain any power 'right at' any particular frequency. Rather, **the power in a noise signal is real, but distributed continuously across the spectrum, and it has to be handled by a *density***, not a local value.

Exercise in spectra, and spectral density

In this exercise, we combine the Filtered Noise source you've been using with Amplification and further Filtering, and then we sum it with an externally-generated sinusoid, to produce a waveform of the sort often seen in the lab: a signal polluted by, or even buried under, broadband noise. And then we see how to quantify that signal, and quantify the noise, and how to use the 770 to see that signal amid the noise, in a picture vastly clearer than anything visible in the time domain.

Start with your Buried-Treasure module's Filtered-Noise output, and amplify it by x10 and x2.5 using the Wide-Band Amplifier module. Now further filter that output, by sending it to the input of your Filter module. Set that filter to frequency 20 kHz and Q = 0.71, and look at the Low-Pass output of the filter module. The original noise spectrum has been initially filtered free of very-high-frequency components, then amplified, and now further filtered, basically to have

> very nearly the amplified content, for frequencies 0 < *f* < 20 kHz,
> but rapidly dropping content for frequencies *f* above 20 kHz.

In fact, the behavior of this filter for white noise is akin to that of a sharp-edged or 'brick-wall' filter, which would multiply the frequency content of the input

by ·1 for 0 < *f* < 22.2 kHz,
and by ·0 for *f* > 22.2 kHz.

That is to say, the 'equivalent noise bandwidth' of this particular filter is 22.2 kHz.

Now send the filter's output to the Summer unit; to its other input, send a sinusoid from an external generator. Set that generator to a frequency near 10 kHz, and an amplitude of 0.3 Volts. The output of the summer will resemble this 'scope view:

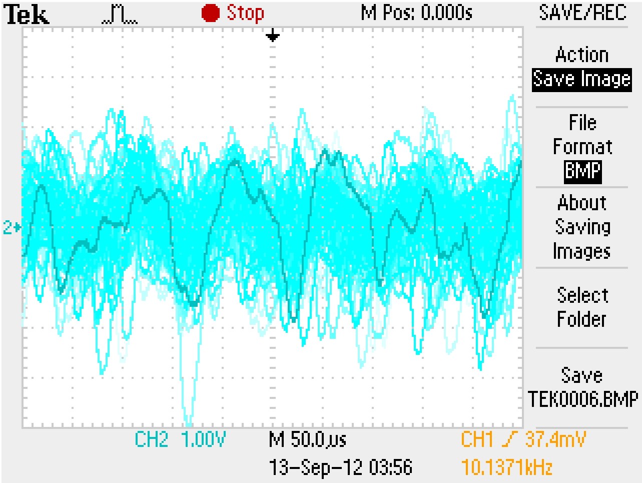


Fig. 6.3: A sinusoidal signal, of period 100 µs and amplitude 0.3 V, well-buried under noise.

This noisy waveform hides a sinusoid of amplitude 0.3 V behind noise that splatters all over the ±2-V range, and beyond. Apply that signal-plus-noise waveform to your 770, and look as follows:

1: In full frequency Span, notice the 'noise floor' across the display; because of the filters' actions, it is spectrally flat from 0 to 20 kHz, and then drops monotonically in the 20-100 kHz range. (Try temporarily using a logarithmic scale for the frequency axis to help see this better.) Confirm, by temporarily lowering the Amplifier gain 10-fold, that your noise module really is the source of this noise background. (The noise floor should drop by 20 dB, ie. 100-fold in power, due to 10-fold smaller voltage gain at the amplifier.)

2: Be sure you've set the Measure softkey to PSD, and units to Volts rms, and use Averaging to estimate the voltage spectral density of the nose floor. You might find a value near 4 mVrms/√Hz.

3: Now look at the spectrum in the vicinity of 10 kHz and spot the 'spectral line' due to your signal generator, lying above the 'floor' of continuous spectrum from the noise source. Temporarily change the frequency, or kill the amplitude, of your generator, to confirm the origin of this feature. Note how prominent the spectral line is, in this frequency-domain view, compared to the signal's total *submergence* under noise in the time-domain view of Fig. 6.3.

4: Switch to the appropriate function to quantify this 'signal peak': use the Measure softkey to change from PSD to Spectrum, and (since you're trying to measure the height of a peak) use the Window softkey to change to a Flattop window. To see that you really have a stable rms measure for this signal, now zoom in on the frequency axis, reducing the Span in steps of 2, eventually using a span of 390 Hz or less, centering your signal peak in the display. The rms measure of this signal will scarcely change in the process – particularly as the frequency bins get narrow, and your peak height is minimally contaminated by noise. Even the power spectral density of the noise floor won't change in its Vrms/√Hz units as you do this zooming process – check out that claim also. But Vrms and Vrms/√Hz are literally incommensurate units, and the *prominence* of a signal peak against a noise floor will rise, dramatically, when you zoom in by this method.

[Why? Because all the spectral power in the signal ends up in a single spectral bin, whether that bin is 250 Hz or <1 Hz wide. But the amount of spectral power of the noise part of the waveform which ends up in any particular bin drops *linearly* with the frequency-width of the bin, and you've lowered that bin-width by 256-fold during this exercise.]

[This lunch is not free! Notice that he narrower you make your frequency span, and the more the signal peak stands up above the noise floor, the *longer* the acquisition time which is required.]

5: Now learn to predict the rms measure of the signal-plus-noise waveform, using

$$\left\langle\, [V_{sig}(t) + V_{noise}(t)]^2 \,\right\rangle = \left\langle V_{sig}{}^2(t) \right\rangle + \left\langle V_{noise}{}^2(t) \right\rangle \quad .$$

(and why is there no cross term in this formula?)

- here $<V_{\text{sig}}^2(t)>$ is the mean-square measure of the signal, given by the square of its rms measure. (Squaring undoes the 'rooting', leaving just the mean square.)
- here $<V_{\text{noise}}^2(t)>$ is given instead by an integral over a density, via

$$\int_0^\infty S(f)\, df \approx (S_{noise}) \cdot \Delta f = (S_{noise}) \cdot (22.2\, kHz)$$

where S_{noise} is the approximately-constant power spectral density you've measured on the flat part of the noise floor, and Δf is the equivalent noise bandwidth (22.2 kHz, for this particular filter) over which that S_{noise} value is applicable.

When you've predicted $<V^2(t)>$ by combining these results from the 770, ie. when you've done an estimate wholly from observations conducted in the *frequency* domain, find a way to measure the mean-square value of $V(t)$ in the *time* domain. You might find a 'true rms' voltmeter which does the rms measurement correctly, even for signals with frequencies up to >20 kHz; or you might use the 'measure' capability of a 'scope. A voltmeter will effectively average over its update time of half-a-second or so; a 'scope will average over its full-screen acquisition time – you might use 50 μs/div or more on the sweep speed, to give >500 μs of voltage-logging time.

Once either of these tools has given you an rms value, square that number to get a mean-square value $<V^2(t)>$ in Volts-squared, derived from time-domain measurements. Now for two questions:

a) Does you time-domain measure *agree* with your frequency-domain estimate?
b) In the sum

$$\left\langle V_{sig}{}^2(t)\right\rangle + \left\langle V_{noise}{}^2(t)\right\rangle \quad ,$$

what fraction of the sum is due to the first, the signal, term?

For the data of Fig. 6.3, that fraction is under 11%, so the waveform is *more than 89% noise* on a power basis. Yet a frequency-domain view of that signal-under-noise, viewed at sufficient spectral resolution, can show a signal peak standing >30 dB, ie. *a thousand-fold in power*, above the local noise floor.

This exercise should give you a vivid picture of why frequency-domain methods are used very broadly for 'signal recovery' in the presence of noise. For more fun with such exercises, see the projects in Ch. 15 of this manual.

6: Now you can vary the sizes of signal and noise independently – vary the signal strength with your generator's amplitude adjustment, and the noise level by changing the gain in the broadband-amplifier module. You can independently remove either contribution from the signal-plus-noise superposition by removing one or the other input cable to the summer unit. You can use a true-rms digital voltmeter at the summer's output to quantify the output in rms measure, and you can square that number to give a mean-square value.

For fun, prepare a signal-plus-noise combination with equal (rms) contributions of signal and noise. Perhaps you can adjust each to give 0.7 V (rms). What happens when both waveforms are going into the summer? A look with a 'scope will show you a waveform with 'signal-to-noise' ratio of 1:1, in which you can see the signal by eyeball, despite the noise. But what will the true-rms meter read?

Answer: *not* 0.7 + 0.7 = 1.4 Volts (rms). Instead, do the computation this way: if the square root of the mean square, ie. the rms value, is 0.7 Volts, then upon squaring, you get the plain mean-square, with value $(0.7\text{ V})^2 = 0.49\text{ V}^2$. So you have two waveforms, each of which has been set to give a mean-square of 0.49 V^2. And since mean-square values are additive (for these uncorrelated waveforms), you expect $0.49 + 0.49 = 0.98\text{ V}^2$ as the mean-square measure of the output. The square root of this mean-square is a root-mean-square, rms, value, of 0.99 V. And *that* is what your true-rms meter should read.

Final question: if you've prepared a noise signal with dc-to-20 kHz coverage, and if its rms value is 0.7 V, what will its power spectral density be, upon measurement with the 770? Expected answer: the spectrum is flat, ie. 'white', from dc upwards, and starts dropping at a 'corner' at 20 kHz, decreasing rapidly (though not abruptly) to zero at higher frequencies. We claim that the filtered spectrum acts as if it were flat to 22.2 kHz, and zero thereafter. Now we have a measure of 0.7 V (rms), so squaring, we conclude the waveform has a mean-square value of 0.49 V^2. So its spectral density is

$$S = (0.49\ \mathrm{V}^2) / (22.2 \times 10^3\ \mathrm{Hz}) = 22.1 \times 10^{-6}\ \mathrm{V}^2/\mathrm{Hz} \quad .$$

The square root of this is the voltage spectral density, 4.7 x 10^{-3} V/√Hz or 4.7 mV/√Hz. The 770 should show this result when set to units of Volt-rms, indicating 4.7 mVrms/√Hz. Try it out.

Chapter 7: The LCR system, introducing transfer functions

You are about to see time-domain and frequency-domain views of the *transfer function*, which is an attribute of a host of physical systems. Suppose you have a physical system, a 'black box', which you can stimulate with a drive, or excitation signal, and from which you can extract an output, or response, signal. Now if the physical system possesses the properties of linearity and time-invariance, then it can be described by a transfer function. It doesn't matter if the system is electronic, mechanical, acoustical, optical, or gravitational in character.

We've included an example, a very important example, of such a linear and time-invariant system among the Electronic Modules; it's called the 'LCR Circuit', and it's just a series combination of an inductor (L = 68 mH), a resistor (R, variable through 0 - 1 kΩ), and a capacitor (C = 10 nF = 0.010 μF). The artwork on the panel of the Electronic Modules shows just how the components are hooked up to each other and to external connectors. There's provision to drive it electrically at an input BNC connector marked Drive, and also to see the potential difference across the capacitor at an output connector marked Response.

To see that this LCR circuit works, set up an external signal generator to give ±10-Volt square waves of frequency 40 Hz or less, and send them to ch.1 of a 'scope, and also to the input of your LCR module. (You'll notice that the amplitude will drop to about ±0.2 V when driving the module, because of the Module's 1-Ω input resistor which is now loading down your generator's 50-Ω output impedance.) Now view the output of the circuit using ch. 2 of your 'scope, and arrange for an over-and-under dual-trace view of stimulus (above) and response (below). Pick a sweep time of order 5 ms/div for your 'scope display. You should see that the sharp edges of the square wave provoke the 'ringing' behavior of the LCR circuit, and the 'ring-down time' depends on your choice of *R*-value in the circuit.

Now set *R* to its minimum value [which is 49 Ω, from the dc resistance of the inductor], and change from square-wave to triangle-wave excitation. For 40-Hz triangle waves, the output resembles the input, and similarly for 400-Hz waves; but not so for 4-kHz triangle waves. If you scan upward from 4 kHz, you'll see some new features: there's a 'resonance' near 6 kHz, and in the vicinity of the resonance, the response to a triangle-wave drive is a sinusoidal-looking response! So the mathematical property of linearity is **not** a requirement of simple proportionality between input and output, as you might first suppose.

So if the mathematical property of linearity is not the same as mere proportionality, what *is* it? Given drive or stimulus $S(t)$ and its response $R = R(t) = R[S(t)]$, linearity does *not* require that the response be a multiple of the stimulus:

not necessary that $R\,[\,S(t)\,] \propto S(t)$.

Instead, linearity requires two things: first, that the response to a sum be the sum of the responses,

$$R\ [\ S_1(t) + S_2(t)\ ] = R\ [\ S_1(t)\ ] + R\ [\ S_2(t)\ ] \quad ;$$

and second, that response to a scaled stimulus be the scaling of the response,

$$R\ [\ \alpha\ S(t)\ ] = \alpha\ R\ [\ S(t)\ ] \quad \text{for any constant } \alpha \quad .$$

And what is time-invariance? Merely the requirement that the time-of-day doesn't matter, or that the properties of the system don't change with choice of origin of time.

No physical system is exactly linear, and most can be 'over-driven' into a range of non-linearity. But many systems have a range of linearity, and the techniques of the transfer function are *indispensable* for describing the system within that range.

Now subject your system to a sinusoidal excitation, again starting at a low frequency like 40 Hz. Again, raise the frequency, and use your 'scope over-and-under view to see what happens – change the sweep rate of the 'scope to keep a few cycles of the drive in your view. You will see that resonant behavior again, near 6 kHz. But along the way, note something special: *un*like the case of excitation by square or triangle wave, **the response to a sinusoidal drive is always a sinusoidal response, and one which has the same frequency as the drive.**

[You've been changing the frequency of the drive waveform; but why aren't you separately trying changes in the amplitude of the drive? Answer: if your system is linear, then doubling or halving the drive just doubles or halves the response. That remains true even when the drive and response have totally different shapes – confirm that claim for excitation by slow square waves.]

Sinusoidal excitation therefore has a *highly privileged* status for linear time-invariant systems, since a drive or stimulus

$$S(t) = A \cos\,(2\pi\, f\, t)$$

has to provoke a response which is *also* sinusoidal, and also at frequency *f*. All that can change between the input and the output is the amplitude, and the phase, of the sinusoid. Since the output amplitude has to scale (by linearity) with the input amplitude, we can write the response as

$$R(t) = A\, M(f) \cos\,(2\pi\, f\, t - \varphi(f)\,) \quad ,$$

where M is the magnitude of the transfer function, and φ is the phase shift of the transfer function. Both of these are functions of frequency, but neither can be a function of time, nor of the amplitude of excitation used.

You will learn how to measure this magnitude and this phase shift, for your LCR system, first in the time domain, and then (with amazing efficiency) in the frequency domain. Before you measure them, have a look at them on your ‘scope. To make your system not so spectacularly resonant, set the *R*-adjustment to the ‘9 o'clock’ position on its dial. Arrange for the ‘scope to trigger on the zero-crossings of the stimulus (= drive = input) trace, and set your sweep rate to 50 μs/div. Now arrange to vary the frequency of sinusoidal excitation from 3 to 10 kHz, while watching what happens:

- On the upper trace, the drive, you’ll see constant amplitude, but growing frequency.
- On the lower trace, watch for the increase in amplitude, by a factor of about eight, as you pass through resonance.
- Also on the lower trace (and requiring you to *ignore* the amplitude-resonance you’ve just seen), look instead at the times-of-zero-crossings of the response to see evidence of its phase shift. Well below resonance, you’ll see the response is in phase with the stimulus; well above resonance, you’ll see the response is (smaller and) inverted, ie. 180°-phase-shifted, relative to the stimulus. If you scan upward in frequency through resonance while watching this, you’ll see that the phase shift grows monotonically (though not uniformly) from 0° degrees, through 90° right at resonance, to 180°at high frequencies.

Now that you know what to look for, you can take data systematically. You might stick with your present choice of *R*-value, and take data at a variety of frequencies *f*. You do not need to make your frequency settings equally-spaced on the *f*-axis; instead, you might want to take points more closely-spaced near resonance. Remember that at any frequency, the magnitude of the transfer function is just a ratio of amplitudes,

$$M(f) = (\text{amplitude of response}) / (\text{amplitude of drive}) \quad , \quad \text{both at frequency } f \quad .$$

The phase shift is a bit trickier to measure, but you can measure it via the time delay Δt between (the upward-going zero-crossing of the drive) and (the upward-going zero-crossings of the response), and using the defining proportionality

$$\text{phase shift } \varphi(f) / (2\pi) \equiv (\text{measured } \Delta t) / (\text{waveform's period } T) \quad .$$

Theory of the transfer function for the LCR system

The data you’ve acquired, point by point in frequency, is worth comparing with theory. What follows is an abbreviated derivation of the theory, which (under other guises) is common to the mechanical spring-mass (or any other single-mode) resonant system.

If you draw a vertically-oriented schematic diagram for your LCR circuit, you’ll see that the drive $V_{\text{in}}(t)$ is applied to the ’top end’ of the series connection of inductor, resistor, and capacitor, and that the ‘bottom end’ of this series combination is at ground potential. If you let $q(t)$ represent the charge on the upper plate of the capacitor, then the current $i(t)$ passing through the inductor and resistor has to be given by $i(t) = dq/dt$.

Now the drive voltage $V_{\rm in}(t)$ at the 'top of the string' obeys the equation

$$V_{in}(t) - L\frac{di}{dt} - i(t)R = V_{out}(t) = \frac{1}{C}q(t) \quad ,$$

and this can be re-arranged to give the differential equation for the charge $q(t)$:

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{1}{C}q(t) = V_{in}(t) \quad ,$$

Now the charge $q(t)$ is indirectly visible, in that the definition of capacitance gives

$$C = \frac{q(t)}{V_{out}(t)} \quad , \quad \text{so} \quad V_{out}(t) = \frac{1}{C}q(t) \quad .$$

Using this relationship allows us to write a (linear, second-order, constant-coefficient, inhomogeneous) differential equation for $V_{\rm out}(t)$ in terms of the drive, $V_{\rm in}(t)$:

$$L\cdot C\,\ddot{V}_{out} + R\cdot C\dot{V}_{out} + V_{out}(t) = V_{in}(t) \quad ,$$

To turn this into a standard form, it is conventional to define a 'natural (angular) frequency' ω_0 , and a (dimensionless) damping parameter γ according to

$$\omega_0 = \frac{1}{\sqrt{LC}} \quad \text{and} \quad 2\gamma\,\omega_0 = \frac{R}{L}$$

It turns out that the textbook case of 'critical damping' is marked by $\gamma = 1$. Now the differential equation is fully described by one natural-frequency value and one damping parameter, and would have exactly the same character if it had arisen in a mechanical or optical context:

$$\ddot{V}_{out} + 2\gamma\,\omega_0\,\dot{V}_{out} + \omega_0^{\;2}\,V_{out}(t) = \omega_0^{\;2}\,V_{in}(t) \quad .$$

Exercise: for the nominal values of L and C, find the value, and the units, of ω_0, and find the ordinary frequency (in Hertz) f_0 to which this corresponds. (Does this match the location at which you found the resonance?)

Exercise: show that the damping parameter γ can also be written as

$$\gamma = \frac{R}{2\,\omega_0\,L} = \frac{R}{2\sqrt{L/C}} \quad ,$$

and estimate the range over which you can vary it – work out $2\sqrt{(L/C)}$ (which must have dimensions of resistance), and recall that the effective R value is the 0-1 kΩ setting of the variable resistor, plus the inductor's dc resistance of about 49 Ω. You'll see that $1/(2\gamma)$ also gives the quality factor or 'Q' of the resonance – what range of Q-values can you cover?

The differential equation we've derived is very well known, and there are many ways to solve it. But it's an even simpler task to solve this equation for the only case we really need – that of sinusoidal drive. It's easiest of all to write the actual drive A cos (ω t) in a complex version, A exp (-i ω t) , and to find the solution A [M(ω) e $^{i\varphi(\omega)}$] exp (-i ω t), which gives the transfer function [in brackets] in its complex form. But by one means or another, the result for the magnitude response can be derived:

$$M(\omega) = \frac{\omega_0{}^2}{\sqrt{(\omega_0{}^2 - \omega^2)^2 + (2\gamma\,\omega\,\omega_0)^2}} \quad \text{or} \quad M(f) = \frac{1}{\sqrt{[1-(f\,/\,f_0)]^2 + [2\gamma\; f\,/\,f_0]^2}} \quad .$$

You should confirm that this predicts that

- $M(f << f_0) \approx 1$, so low-frequency components pass through unaffected;
- $M(f \approx f_0) = 1/(2\gamma) = Q$, so frequencies around resonance are enhanced Q-fold; and
- $M(f >> f_0) \approx (f_0/f)^2$, so that high frequency components are suppressed and drop off as f^{-2}.

The phase response is less intuitive, but it can be expressed in an inverse-cosine form which correctly describes its smooth and continuous variation between 0 and π radians.

$$\varphi(\omega) = \cos^{-1}\frac{\omega_0{}^2 - \omega^2}{\sqrt{[\omega_0{}^2 - \omega^2]^2 + [2\gamma\,\omega\,\omega_0]^2}} \quad \text{or} \quad \varphi(f) = \cos^{-1}\frac{1 - f^2\,/\,f_0^2}{\sqrt{[1-(f\,/\,f_0)^2]^2 + [2\gamma\; f\,/\,f_0]^2}} \quad .$$

Note that theorists' derivations are given in terms of a fixed circuit parameter ω_0 and a variable drive (angular) frequency ω , but that experimenters will want to express the circuit parameter as an (ordinary) frequency $f_0 = \omega_0/(2\pi)$, and their chosen drive frequency as f (both of these in Hertz). You should plot the predicted magnitude and phase responses as functions of frequency, over a relevant range, for various values of the damping parameter γ. You will gain more intuition if you see these graphs plotted on both linear and logarithmic frequency scales.

Now you should compare your observations for magnitude and phase shift with these predictions. You could try a least-squares fit to your data, using f_0 and γ as fitting parameters, or you could try to measure L, C, and R-values and overlay your data with a zero-free-parameters *prediction*. Once you've done this, you'll understand where you could have taken more (or fewer) data points. Once you've modeled your data, you can vary the choice of R-value in your model, to understand how the data would have changed if you had taken it with a different choice of R.

Transfer functions by Fourier methods

But the central point of this chapter is to give you a way to get this sort of data *all at once* by Fourier methods. The idea is to drive the LCR circuit *not* with one sinusoid at a time, but with *all sinusoids at once* – and one name for such a superposition is 'white noise'.

The method is to use white-noise excitation and to look at the emerging voltage and Fourier-analyze it, into all the sinusoids of which it's composed. This is amazingly easy to do, and it recreates for electrical circuits a technique which was first applied to infrared spectroscopy – giving the Fellgett advantage of looking at all the frequencies in parallel.

The reason it works is of course linearity. The response to a sum is the sum of the responses; here the sum is white noise considered as a sum of sinusoids. For white noise, each term in the sum has equal strength in the drive. The response of the physical system to each term is a sinusoid of unchanged frequency, but modified in magnitude and shifted in phase. So now the sum of the responses is no longer white noise, but a noise waveform with *un*equal coefficients for each sinusoidal term. In fact, the coefficient for each term is just $M(f)$, the transfer function's magnitude at that frequency. So the Fourier spectrum of the output waveform is a direct picture of the entire plot of $M(f)$, for a whole range of frequencies all at once.

To use this method requires a source of white noise that can drive your LCR circuit. One such source is the Buried-Treasure module, which you can switch to the Filtered-Noise selection to get noise which is spectrally flat from dc to over 100 kHz. To give that signal more strength, you can use the Power Audio Amplifier module, set to gain about 6, and connect its output to the LCR circuit's input Drive connection. Put a BNC splitter at this input and bring out a cable to the 770's input, so you can check the spectrum of the input noise that is actually driving the circuit. Use the MEASure button to bring up the Measure menu, and choose PSD = power spectral density, as is appropriate to the continuous spectrum you're viewing. You can choose a frequency span covering 0-25 kHz, and AutoRange and AutoScale as usual, to get a useful display. Now use a suitable amount of averaging, and use the SCALE button to set the vertical scale to 1 dB/div, to confirm that you have a 'flat spectrum' going into your circuit. Now you might want to change back to 10 dB/div, and you might even adjust the Top-Reference softkey so that your flat spectrum lies exactly halfway up your scale.

Then you're just one step away from a revelation. Instead of monitoring the input waveform to your LCR circuit, connect the 770 to the output waveform of your circuit. You need no changes in scales, levels or settings, but need merely to observe the modified noise spectrum. It's been modified by gain for all the frequencies for which $M(f) > 1$, and by loss where $M(f) < 1$, and it in fact gives you a real-time graph of $M(f)$. To see the near-real-time response, change the R-value to vary the circuit's damping, and watch how fast are the changes in the spectrum you're seeing. (The more averaging you're doing, the slower will be the response, but the smaller will be the statistical scatter.)

You might want to try both linear and log scales for the frequency axis while you're trying these changes. Notice that (contrary to many persons' intuition) the change in damping affects the $M(f)$ function *scarcely at all* in either the $f << f_0$ or in the $f >> f_0$ regions – only the immediate neighborhood of the resonance gets affected.

If you like this technique, you can get its results out of the 770 by learning how to save-to-USB the graph you're seeing. You'll get out 400 'spot values' of Fourier spectrum, and if you do the same for the input spectrum, and do a division for each frequency of

(output Fourier magnitude) / (input Fourier magnitude) ,

then you'll have 400 spot-values of *M*(*f*), all obtained at the same time through the magic of the superposition principle. You can compare these values to theory, just as you did for the former values you tediously obtained one-by-one.

If you'd like to try this technique on other sorts of circuits, replace the LCR circuit with the Filter module as your next 'device under test'. For starters, drive the filter with white noise, and try out the band-pass output of a filter set to frequency 10 kHz. Vary the filter's setting of Q-value and see what happens. The point is that though the filter is an electronic circuit which might have nothing to do with resonance in an LCR circuit, it is built to be a linear time-invariant system, and you are measuring its transfer function. In fact, you're getting a view of the magnitude *M*(*f*) of its transfer function, in real time. So you can now *rapidly* try the effect of distinct frequency-settings on the filter, and you can rapidly see how (for fixed frequency and Q-setting) the low-pass, band-pass, and high-pass outputs look in frequency space.

This method tells you, all at once, about the magnitude factor in the transfer function; but you'll notice that it tells you nothing about the *phase* response. That's because you're using the 770 to measure only the magnitude of the Fourier spectrum emerging from your noise-excited system-under-test. At a deeper level, that's because you don't have a way to measure *simultaneously* the time record of the white-noise input and the modified noise output. If you were able to do this, and if you then Fourier-transformed the two simultaneously-obtained time records, the difference of

(output-term phase) - (input-term phase)

for the coefficients of each term in the Fourier transform would give you a graph of the phase-shift function φ(*f*), just as a ratio of magnitudes of these coefficients would give you *M*(*f*). [For a second method for getting the phase information of the transfer function, see the use of transient waveforms in Chapter 9; and for a third method, which actually carries out a version of the subtraction above, see Appendix A16.]

Chapter 8: Transfer functions of acoustic systems

This section introduces you to some apparatus ‘outside the box’ of Fourier Methods, to show you the power of Fourier thinking in understanding a physical system. In this case, the system-under-test is an acoustic resonator of cylindrical shape, with an actuator (a small speaker) at the center of the bottom end, and a sensor (a tiny microphone) at the center of the top end. This is a model system, in that it displays a series of ‘normal modes’, at which it will respond resonantly, ie. much more strongly than at general frequencies. You’ll learn how to find those resonances, how to find them *efficiently,* and how to understand them. Along the way, you’ll see again how white noise can be your friend, rather than your enemy, in some kinds of investigations.

Find and open your cylindrical resonator (using a straight-up lift to do so), and locate the speaker and the microphone. Notice that the resonator’s inner diameter of 2.810" = 71.37 (± 0.03) mm is fixed, while its internal length can be adjusted upward from a minimum value of 0.787" = 19.99 (± 0.03) mm by using either or both of the two extra rings provided, and also by using the nylon thumbscrews which allow the top cover to be ‘backed out’ by up to 10 mm more. The resonances you will excite correspond to motions in the air trapped inside the resonator (*not* due to resonances in its metal walls – and how would you show that?) Since the speaker excites, and the microphone detects, only on the cylinder’s axis of symmetry, the modes you will detect are expected to be those of cylindrical symmetry.

Here’s the old-fashioned way of finding a resonance in your structure: You drive the speaker with a fixed-amplitude, variable-frequency sine-wave generator, and you also send that generator signal to ch. 1 of a dual-trace oscilloscope. Then you use ch. 2 of that ‘scope to monitor the signal picked up by the microphone. Finally, you scan over a range of generator frequencies, and try to find one (or more) frequencies at which there is a locally much greater response from the microphone.

To set up to do this, it’s handy to have an external function generator. Pick a sinusoidal waveform with an amplitude of about 1 V, and drive the speaker via the connections and settings shown in Fig. 8.1 below. Choose a generator frequency of about 1 kHz, and temporarily open up your resonator to confirm that you can hear the speaker ‘singing’. Now connect the microphone by the arrangement also shown in the figure – this connection provides +5-V power to the pre-amplifier that’s integrated into the microphone, and also sends back (via the same wire) the ac signal corresponding to what the microphone picks up. For a reality check, whistle into your microphone, and confirm you can get signals of amplitude of order 10 mV to display on your ‘scope’s ch. 2.

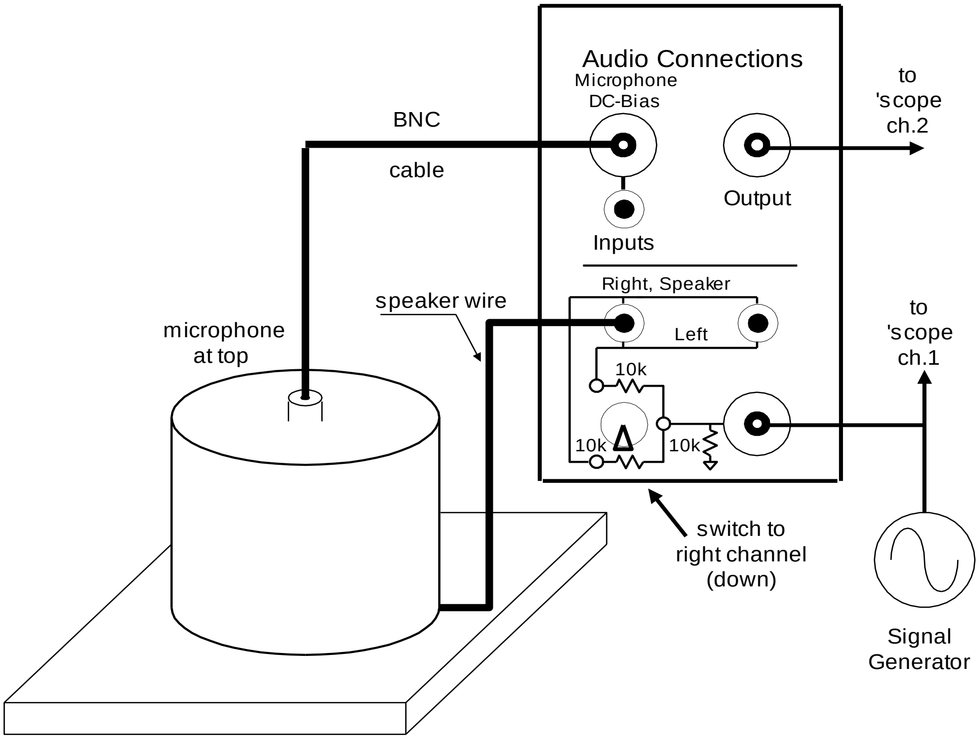


Fig. 8.1: Using the Electronic Module allowing easy connections to the resonator's speaker and its piezoelectric microphone.

Now reassemble your resonator, set to its minimum length, and start with your generator at 1 kHz, and slowly dial up its frequency, while monitoring the ch. 2 microphone signal with your 'scope. Choose a time base such that a few full cycles of the drive appear on ch. 1, and use the highest sensitivity on ch. 2 to look for pick-up by your microphone. What you hope to see is a signal which has the same frequency as the drive (but perhaps with a shifted phase), and whose amplitude is resonantly enhanced right in the neighborhood of a particular (acoustic normal-mode) frequency.

If you manage to find a resonance in the range 3-6 kHz, scan through it with sufficient resolution to see if you can find the center of the resonance. In the process, you will appreciate how narrow in frequency the resonance is, and perhaps also how easy it would be to *miss* the resonance in a scan search like this (for example, if you were to scan too quickly through the resonance). Experience this by searching for the next resonance up in frequency, and see what it takes to be sure you haven't missed any intermediate resonances.

Of course, if you are interested in detecting a whole *family* of resonances, and eventually locating each of them in frequency and characterizing each as a mode, *and* if you'd like to do that for a variety of length-of-resonator choices, you will soon tire of this hand-searching procedure, or have doubts about its completeness.

So now it's time to see the whole family of resonances *all at once*, by using white noise as a drive, and Fourier methods for detection. So substitute, for the previous sine-wave source, the 770's white-noise source – see Ch. 6 for how to set this up. Your 'scope's ch. 1 will show the garbage-looking waveform that is now going into the speaker. If you open the resonator, you may *hear* this white-noise waveform as a faint hiss. Now conduct the microphone output signal not only to the 'scope's ch. 2 as before, but also to the signal input of the 770. What you will be dong is exploiting the linear behavior of the acoustic train of speaker → resonator → microphone, in that you will be exciting the speaker (and resonator) simultaneously with all the frequencies present in the white-noise source. The resonator will respond selectively, conveying to the microphone with extra strength all (but only) those frequencies which match acoustic resonances of the device. Then the microphone will be generating a signal which is also a superposition of many frequencies (one for each resonance excited), but the 770 will tease those apart, ie. Fourier-analyze them, showing each one as a separate peak in frequency space, with a height indicating the strength of the resonance.

When your 770's display shows any peaks at all, you might start to notice the spectrum show fluctuations. These are due to the randomness that's present in the noise you're using for excitation. You can minimize the effects of these fluctuations by using the AVERAGE menu button and the softkeys that it brings up. At the cost of adding to your waiting time, this will smooth out the statistical fluctuations.

Once you have an averaged view you like, you might want to zoom in on the location of a detected, or suspected, resonance peak. Use the FREQ menu, and the Span and Center choices, to have a close look at the lowest-frequency peak you have detected. Continue zooming in until you can see that the peak exhibits a non-zero width. At this point, you will be able to use the Marker knob to read off the frequency location of this peak to a few-Hz precision. As you zoom in like this, you will also see that the Fourier spectrum you are viewing gets 'noisier', ie. subject to greater fluctuations, and for a good reason too. Your white-noise source is generating a fixed amount of power, and is spreading that power uniformly over all of the 0 - 100 kHz frequency space. When your 770 as frequency analyzer is zooming in, it is becoming sensitive to an ever-smaller fraction of that total power. If you are using a span of 780 Hz, you get about 2-Hz spectral resolution in your display, but you are also being sensitive to only 780/100,000 of the total white-noise power you're sending into your speaker.

There is a cure for this problem, too: that is to change from a white-Noise source to a 'Chirp' noise source in your 770's internal source. Now the 770 will generate a waveform which consists of *only* those frequency components which fall within its span of analysis (so that if you have chosen a Start frequency of 10,100 Hz, and a Span of 780 Hz, the chirp waveform consists of a summation of 400 equal-amplitude sinusoids, spread out uniformly as a 'frequency comb' in the range 10,100 to 10,880 Hz). That is to say, all of the 770's source power is now devoted to generating sinusoids lying in the very range which you are analyzing. [See Appendix A9 for more details on how it's done, and why it's called 'chirp'.] *But* for Fourier-analysis using the Chirp source, you will need to change the Window choice to **Uniform**.

See if this process will enable you to 'zoom in' more successfully on suspected resonant peaks, and start to tabulate the frequencies at which peaks occur. You are doing 'acoustic spectroscopy', finding the spectrum of spectral response of your resonator just as an atomic-optical scientist would find the spectral fingerprint of an atom or molecule. Record the location of at least a dozen resonant frequencies that occur in the 2 - 20 kHz region of the spectrum. (You cannot expect that the speaker and microphone to give good response above the 20-kHz maximum frequency for which they were designed.)

Now repeat that 'fingerprinting' operation for another choice of resonator length you can achieve – recall that you can 'stack' the extension rings which lengthen the resonator. Once you have two or more such fingerprints, we suggest you make a 'stick diagram' of resonances as a function of frequency, and make a new diagram for each new length you use. Now 'stack' the stick diagrams so that frequency runs across, and multiple diagrams are lined up in a stack vertically. You might be able to see two kinds of resonances:

- those whose location stays fixed in frequency, ie. those independent of length *L*;
- those which shift, systematically, with changes in *L*.

For any particular mode, you can zoom in on its resonant peak, and then see if changing *L* (by using the thumbscrews) will 'move' that peak.

In fact acoustical theory (see below) predicts that the mode-frequency vs. length relationship ought to be more transparent if your stick diagram is arranged on a frequency-*squared* axis, and if your 'stacking' goes as inverse-length-squared. That is to say, the prediction (from solving a partial differential equation) of the mode frequency *f* for a given mode as a function of *L* is that $f^2 = a^2 + b^2/L^2$ for constants *a* and *b*.

This means the 'sticks' which appear in your diagrams will start to take on identities, such that you could trace them in your mind as having locations varying *continuously* with *L*. In fact, using the thumbscrew adjustments, you *can* vary *L* continuously – once the screws engage, they will give 1/32" = 0.7938 mm of extra length per full turn. You could become sufficiently assured of the individual modes' identities to assign them labels. Notice that the raw data by itself does *not* do this labelling! The further analysis of these modes is of course best conducted with some theoretical guidance, which is provided below.

But the important feature of this lesson is called the 'Fellgett advantage': Fourier methods allow you to search a whole region of frequency space *simultaneously and in parallel*, rather than sequentially and one-frequency-at-a-time, with tremendous advantages in time spent, and completeness achieved. The whole resonance spectrum, or vibrational spectrum, or NMR spectrum, or infra-red spectrum . . . of a system can thus be obtained all-at-once, compared to the scanning method that you tried first.

For more exercises in using the Chirp configuration of the SR770's internal Source, refer to the SRS Operating Manual, in its section 'Getting Started', at pp. 1-42 through 1-46, for instruction in using its internal chirp source.

Theory for acoustic resonances in a cylindrical resonator

A suitable treatment of acoustic resonances within a cylindrical volume of air is found in M. J. Moloney, *Plastic CD containers as cylindrical acoustic resonators,* Am. J. Phys.

77, 882 (2009). There (or elsewhere) you'll find that the behavior of a sound field in air can be described via a single scalar field, labeled *p* for pressure, giving the pressure deviation from a constant ambient pressure. Subject to some reasonable approximations, you'll find that the pressure field for a normal mode, ie. a solution of one definite frequency for the equations of motion for sound, has to obey the Helmholtz equation,

$$(\nabla^2 + \frac{\omega^2}{c^2})\, p(\vec{r}) = 0 \quad .$$

Here ω is the angular frequency of the oscillation in the normal mode, and *c* is the speed of sound in air. To solve this equation in a cylindrical geometry, subject to the appropriate boundary conditions, requires that

$$p(\vec{r}) = p(r,\theta,z,t) \propto J_m(k_r\, r)\cos(m\theta)\cos(\frac{\pi l z}{L})\cos(\omega t) \quad ,$$

where *z* is height measured along the axis of the cylinder, *r* is the radial coordinate perpendicular to that axis, θ is the azimuthal angular coordinate around the *z*-axis, and *t* is the time. The boundary conditions assumed are those of zero air-velocity components perpendicular to the end-walls or side-walls of the resonator, which in turn requires that the pressure field *p* have no pressure gradient in a direction perpendicular to a bounding surface. The azimuthal symmetry of excitation and detection in the TeachSpin resonator means that only modes having *m*=0 will be well-detected, so the sound pressure field becomes

$$p(r,\, no\ \theta,\, z,\, t) \propto J_0(k_r\, r)\cos(\frac{\pi l z}{L})\cos(\omega t) \quad ,$$

Here $J_0(k_r\, r)$ is the ordinary Bessel function of zeroth order (whose appearance in solving a second-order differential equation in cylindrical coordinates is not unexpected). Now to satisfy the boundary conditions at both the *z*=0 and the *z*=*L* endplates of the cavity, the 'quantum number' *l* above needs to be an integer, *l*=0, 1, 2, . . . ; and to satisfy the boundary condition at the sidewalls *r*=*R* of the cavity requires that the first derivative of the J_0-function vanish at argument $k_r\, R$. If we let x_{0n}' be the *n*th zero of the J_0-function's derivative (where $x_{00}' = 0.0$, $x_{01}' = 3.8317$, $x_{02}' = 7.0156$, $x_{03}' = 10.1735$, etc.), then we can write this condition as

$$k_r\, R = x_{0n}' \quad .$$

Finally, for the Helmholtz equation to be solved, the mode frequency ω has to be related to other constants by

$$\frac{\omega_{0nl}{}^2}{c^2} = k_r{}^2 + (\frac{\pi\, l}{L})^2 \quad .$$

Here the ω-value is triply subscripted, with the 0-subscript denoting modes of *m*=0, ie. lacking azimuthal variation, the *n*-subscript of *n* = 0, 1, 2, 3, . . . denoting the character of radial variation, and the *l*-subscript of *l* = 0, 1, 2, . . . denoting the character of variation along the central axis. In fact, the subset of modes detectable in your apparatus has ‘quantum numbers’ m=0, n, and l. This prediction for angular frequencies can of course be turned into a prediction for ordinary frequencies, of the form

$$f_{0nl}{}^2 = c^2\,[\,(\frac{x_{0n}{}'}{\pi}\frac{1}{2R})^2 + (l\,\frac{1}{2L})^2\,] \quad .$$

This is the prediction which generated the earlier suggestion of laying out $f^{\,2}$-values along an axis, and stacking multiple such plots, according to $1/L^2$-values, along a perpendicular direction.

Given these general results, you might now lay out a theoretical ‘stick diagram’ for modes, first for l=0, and recognize that these are given by

$$f_{0n0} = c\,(\frac{x_{0n}{}'}{\pi}\frac{1}{2R}) \quad .$$

Since that Bessel-function-derivative’s root has location $x_{0n}' \approx (n + ¼)\pi$ for large n, these frequencies become, for large n, approximately equally spaced along the frequency axis.

Now add to this l=0 diagram the prediction for l=1 modes, and you’ll see each l=0 mode gains a ‘satellite’, on the higher-frequency side. Further add in the l=2 predictions, and you’ll see another satellite of higher frequency still. Soon you’ll be able to understand the predicted spectrum of modes, for your fixed radius R and your chosen length L. This should allow you to assign, to each mode you’ve observed experimentally, a unique set of ‘quantum numbers’. Ideally, you should have no missing modes, and have no leftover unassigned modes, when you’re done with this assignment procedure. You’ll also have introduced a good deal of theoretical understanding to your otherwise-unlabelled experimental data.

You might also see some modes with a very curious lineshape, not symmetrical around a peak frequency, but skewed. These ‘Fano profiles’ show up when a weak resonance lies in the ‘wings’ of a strong one, in which case the speaker excites both modes, and the microphone picks up the superposition of both modes – a superposition which can give destructive, as well as constructive, interference, depending on *phases* of waveforms. (You can find a pair of modes for which the thumbscrews can be used to ‘tune’ one mode so its frequency moves right through that of another mode, as you vary L using the thumbscrews; this will let you see the changing shape of a Fano profile.)

Finally, you can see if one single value of the speed of sound c will allow your theoretical mode-frequency pattern to match your experimental data for mode frequencies. In the process, you might have to investigate how the textbook speed-of-sound value is predicted to vary with the air pressure and temperature. You’ll also find that your experimental resolution is quite high – pick a mode which is relatively strong, and relatively isolated, and see if you can reliably create, and detect, variations in its resonant frequency by (say) 2 Hz out of about 10-20 kHz. At this level of measurement precision, you will have to pay attention to the measurement and control of independent variables.

Chapter 9: Fourier transforms of transient waveforms

You have used the SR770 to digitize and Fourier-transform both periodic and non-periodic waveforms, but in both cases, these have been steadily continuing waveforms. Not all waveforms have this character – and now you'll see how *transient* waveforms can be captured and Fourier-transformed. This capability is crucial in many forms of 'Fourier-transform spectroscopy', and it brings up a few new issues. These include the issues of triggering, ie. setting a particular time as $t = 0$; but that setting in turn makes it feasible to get *phase information* from a Fourier transform.

As usual, we'll start with a concrete phenomenon occurring within your Electronic Modules, and you'll see it first in the time domain. As an analog to the sort of transient waveform typical of pulsed nuclear-magnetic-resonance (NMR) signals, we'll look at the ring-down transient that can be excited and observed in the LCR module of the electronics package. Signals such as these have the character of being provoked by an initiating event under the experimenter's control, and then dying away so that, within a finite time, the entire signal can be captured.

To run the LCR module, set up an external function generator to produce square waves at ±10-V levels, and connect its output to the input of the module (and also to ch. 1 of an oscilloscope). The 1-Ω resistor connected at the input will give you (for a generator of 50-Ω internal impedance) a loaded-down square wave at ±0.2-V levels, still easily sufficient to excite the inductor-capacitor system. Take as your output the voltage across the capacitor in the system, and convey that by cable to ch. 2 of your 'scope. Now make sure you know how to set the 'scope to trigger on the rising edge of the ch.1 signal (by setting source, slope, and level among the trigger settings of your 'scope). This marks ch.1 as the cause, and the rising edge of the exciting square wave as the time-origin, of the LCR physics that is now visible on ch. 2.

Now find the *R*-adjustment of the LCR circuit, and set it to its minimum value (so the transient decays as slowly as possible, ie. lasts as long as possible). In principle the exponentially-decaying waveform is never really over, but you can find a time at which it's finished for all practical purposes. Of course, you'll have to set your square-wave generator's frequency *low* enough, to make its half-period *long* enough, so that the LCR system can *finish* its transient response before being re-energized by another square-wave transition.

Now on your 'scope you are getting a time-domain view of the waveform you want to capture, and the 'time window' in which you want to capture it. To see the frequency-domain view, you need (for the first time) to bring *two* cables over to the 770. Let a copy of your driving square-wave waveform be one of these, and bring it to the Trigger-In BNC input of the 770. The output waveform of the LCR circuit, with its transient ring-down, is the signal you want to send to the usual A-input of the 770.

Before you do any of the ordinary set-up, you need to learn how to get the 770 to trigger properly. Using the INPUT menu button, you can use the Trigger softkey to bring up a menu with a new set of softkeys. Use the topmost one of those to make the choice Ext for triggering by an external waveform, namely the waveform you've hooked up to the Trigger-In BNC connector. Since you have a waveform which is making the transition from -0.2 Volts to +0.2 Volts, it is sensible to use the 0.0-V level, and a positive slope, as the trigger criterion, so set those too. You can set the Trigger Delay to zero, and ignore the Arming menu, and then you can push Return to exit the Trigger menu.

Now with these triggering settings, you have chosen the upward transition of the square wave driving the LCR circuit as marking the $t = 0$ point for both the 'scope and the 770. Now you can set up for Fourier-transforming in the usual way. *But* there is one *un*usual feature, particular to transient waveforms, which you need to select: under the MEASure menu, find the Window softkey, and set it to **Uniform**. This is essential for transient waveforms, so as to weight equally all the data points acquired within the acquisition window – the other Window choices will wipe out nearly all the weight of the brief-duration transients you want to capture.

Otherwise, the set-up is familiar: use the AutoRange and AutoScale buttons to configure the 770's front end and its display, and start by choosing the full-span, 0-100 kHz range of the FREQuency scale. From the MEASure menu, use the Measure softkey to select Spectrum, and the Display softkey to select the Log Magnitude mode of display. You'll note that the choice of full span enforces an acquisition time window of 4 ms duration, and comparison to what you see on your 'scope will reveal that you are not capturing the full transient (in this mode). That means the spectrum you'll see does not have quite as good spectral resolution as you'd expect. Already you should see a single peak, of non-zero width, in your display – that's the frequency-domain view of your transient ring-down signal.

Now cut the frequency span in half a few times, so as to 'zoom in' on this peak. Each time you halve the span, the acquisition time the 770 requires will *double*. Each transient gets captured in an acquisition window which starts at your $t = 0$ triggering event, and continues for the duration noted. Cross-reference to your 'scope to ensure that what the 770 is capturing is what you want it to get. By the time you choose a span from 0 to 12.5 kHz, the acquisition window has reached 32 ms duration. To see how this matters, temporarily increase the frequency of your square-wave drive until a second square-wave transition falls within 32 ms of your trigger event. On the 'scope, nothing strange happens when you do this; but on the 770, the Fourier spectrum exhibits curious features, arising from the interference between the 'old transient' and the 'fresh transient' that are now both present in the time window. Remember this as a general lesson – when working with transients, the combination of time-domain view (what are you capturing?) and frequency-domain view (what spectrum does it have?) is crucial for valid data-taking.

Now the frequency location of the peak of your Fourier (magnitude) spectrum is easy to read using the Marker function, and you should make a check against the near-periodicity your 'scope can reveal. If the peak occurs at 6 kHz, you'd expect the cycle of the ring-down waveform ought to occur with 1/6-ms, ie. ≈170 μs periodicity. Do yourself the favor of making a reality check on this claim, using the 'scope!

For more exercises in using the triggering capability of the SR770, refer to the SRS Operating Manual, in its section 'Getting Started', at pp. 1-11 through 1-13, for instruction in using its Trigger input and triggering menus.

The complex character of the Fourier transform

Only when you have a definite location for $t = 0$ do you have the chance to talk about the *phases* in a Fourier spectrum. In previous chapters you have used a Continuous mode of triggering, in which the 770 does fresh data acquisitions 'all the time'. That is to say, it starts fresh acquisition windows at times which are *random* relative to the physics of the waveform. But now you have a definite $t = 0$ instant enforced by triggering, occurring at a time which is *stable* relative to the transient signal (right at the start of the ringdown, in this case.)

Here's why that matters – we can use the language of Fourier integrals to define a transform, according to

$$\tilde{V}(f) = \int_{-\infty}^{\infty} V(t)\,\exp(2\pi i\; f\; t)\; dt \quad ,$$

which together with

$$V(t) = \int_{-\infty}^{\infty} \tilde{V}(f)\,\exp(-2\pi i\; f\; t)\; df$$

define a valid Fourier-transform pair. [Leading factors of 1/√(2π) are *not* missing from these equations – rather, they are unnecessary when ordinary frequencies *f* are used instead of theorists' angular frequencies ω.] Now suppose that we shift the origin of time by amount t_0, by putting the time-shifted $V(t - t_0)$, instead of $V(t)$, into the recipe which produces the transform $V^{\sim}(f)$. You can show that the new result is related to the old result by

$$\tilde{V}_{new}(f) = \tilde{V}_{old}(f) \cdot \exp(i \cdot 2\pi\; f\; t_0) \quad ,$$

which shows a phase shift by angle $2\pi f t_0$ (in radian measure). The first thing that this demonstrates is that the old, and the new, transforms have identical magnitudes – which is why triggering, ie. the choice of $t = 0$, did not matter when the magnitude of the Fourier spectrum was all you were computing. But when you're trying to get the actual complex value (ie. including the magnitude *and the phase*) of the Fourier transform, timing does matter. Not only does a time shift by t_0 change the phases, it changes the phases *differently* for each separate frequency *f*.

But you're in an environment where you do have a physically-significant and electronically-enforced choice of $t = 0$, so your Fourier transform does contain useful and valid phase information. You can arrange for the 770 to display phases directly, in degrees of angle, but it is often better to get two separate views of the real and imaginary parts of the Fourier coefficients. (Naturally, the real and imaginary parts of any complex number fix the magnitude and the phase of that number – but they have the advantage of not being subject to 'wrap-around' by 360°, and not having artificial boundaries at (say) -180° and +180°.)

To get the 770 to give these displays, first use the DISPLAY button's top softkey to select the Up/Dn (rather than Single) display. Now you'll get *two* graph panels on your display, which share the same frequency scale, but can otherwise be configured independently. Use the ACTIVE TRACE control button to toggle between these two windows. Having selected one window, you can use the MEASure button and the Display softkey to set the upper window to Real part; similarly you can set the lower window to display Imaginary part. Now you have a real-time view, as a function of frequency, of both the real and imaginary parts of the Fourier transform. For one choice of the damping, the data will look like the two traces in the plot below:

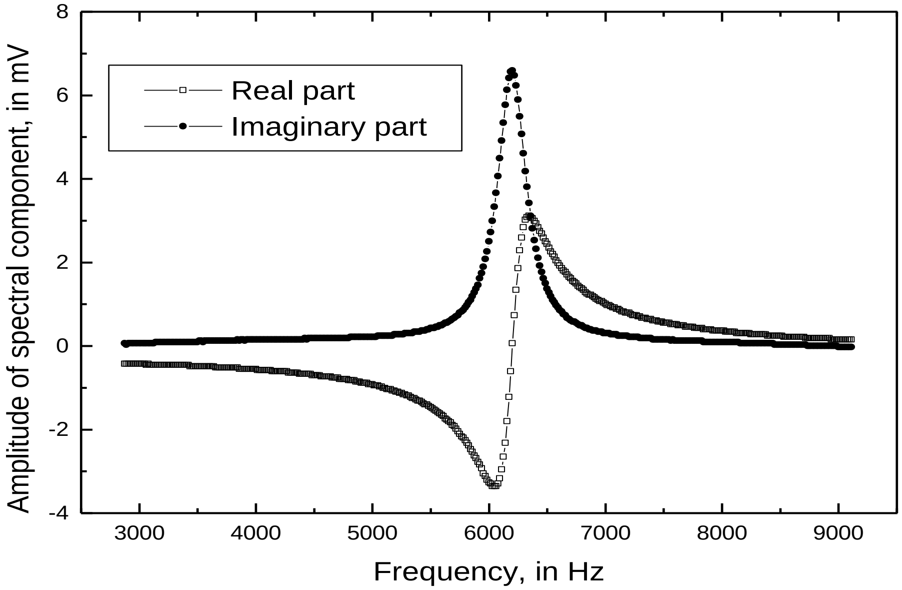
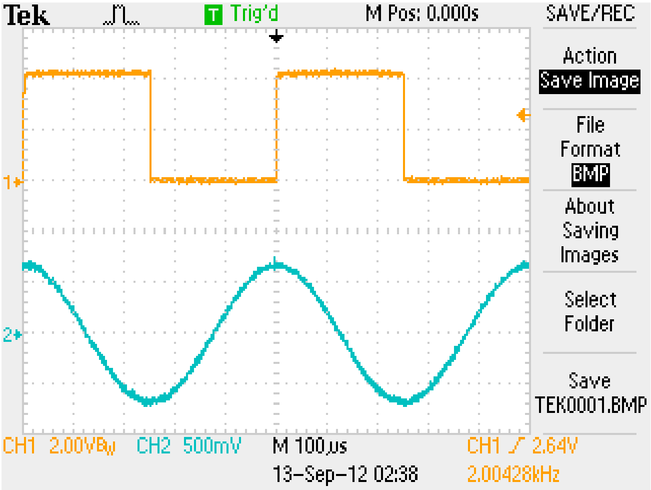


Fig. 9.1: The real and imaginary parts of the Fourier transform of the 'ring-down' signal from an LCR circuit, obtained from a single ring-down in an acquisition of duration 64 ms.

Before we go on to discuss the theory of these curves, it is worth finding out *empirically* just what 'real' and 'imaginary' mean in this context. To do so, all that's needed is a signal generator with a 'sync' or other reference output, as well as a sinusoidal output.

The goal is to create, relative to a definite trigger criterion, first a 'cosine' type of wave, and then a 'sine' type of wave – and in each case, to see what the 770 makes of them.

Below is an example of a 'scope view of a 'TTL output' (above) and a sinusoid (below) as emerging from a particular signal generator. (As it happens, the sinusoid's frequency is near 2 kHz.) Clearly, relative to the rising-edge of the upper waveform, the lower waveform is a cosine type of wave.
So with the upper trace being used to trigger the 770, the lower trace gets digitized, Fourier-transformed, and displayed – and it turns up as a signal chiefly in the Imaginary display. (A slight adjustment, by a very small fraction of the period of the wave, of the Trigger Delay setting will put the signal *entirely* into the Imaginary channel – further illustrating that the Fourier phases are sensitive to the setting of time origin.)



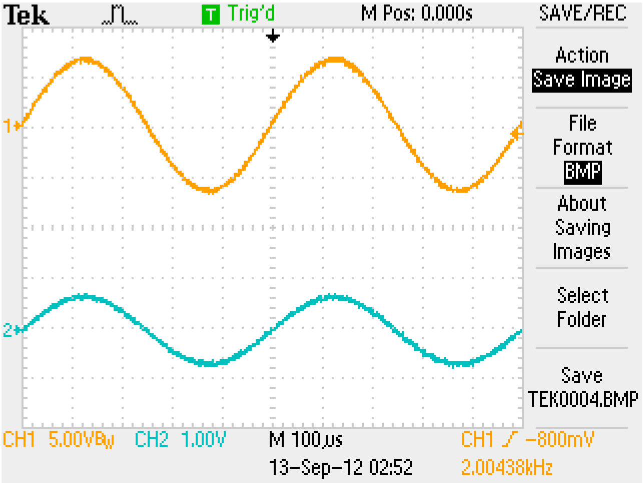
M Pos: 0.000s

CH1 2.00V BW CH2 500mV M 100µs CH1 ⟋ 2.64V

13–Sep–12 02:38 2.00428kHz

Fig. 9.2: Oscilloscope views of trigger (above) and signal (below) waveforms, used to see how the 770 treats a 'cosine' type of wave

Here's a second example of a high-level output of a signal generator (above) and a low-level output of the same generator (below), where the zero-crossings of the upper trace can serve as a trigger for the 770. The lower trace shows the unambiguously sine-type sinusoid that the 770 Fourier-analyzes.


Tek Trig'd M Pos: 0.000s

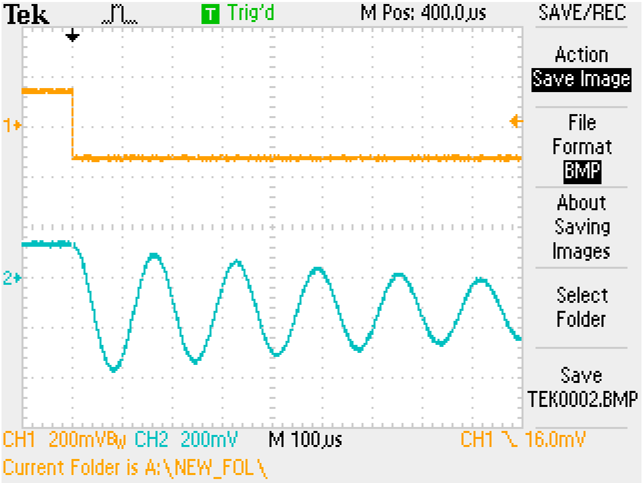
CH1 5.00V BW CH2 1.00V M 100µs CH1 ⟋ −800mV

13−Sep−12 02:52 2.00438kHz

Fig. 9.3: Oscilloscope views of trigger (above) and signal (below) waveforms, used to see how the 770 treats a ‘sine’ type of wave

Trying this out, the Fourier peak turns up chiefly in the Real display. Again, a small adjustment of the Trigger Delay will put the signal entirely in the Real channel.

Now back to Fig. 9.1, in which it’s clear that relative to the triggering used there, the resonant-peak signal turns up in the Imaginary channel, which means that from the test of Fig. 9.2, it’s a cosine-type of wave. And checking in the time domain reveals that this is true – Fig. 9.4 shows the LCR’s input square-wave excitation, whose falling edge triggers the ‘scope and 770, and it also shows the start of the ring-down transient, which clearly has ‘cosine character’ relative to the chosen origin of time.


Tek Trig'd M Pos: 400.0µs

CH1 200mV BW CH2 200mV M 100µs CH1 ⟍ 16.0mV

Current Folder is A:\NEW_FOL\

Fig. 9.4: Early stage of the falling edge of the square-wave excitation (above) and the LCR circuit’s response (below), which establish the ‘cosine’ character of the response.

This empirically-demonstrated behavior of the Real and Imaginary response of the 770 is consistent with its internal use of a definition of the discrete Fourier transform whose analog in Fourier integrals is

$$\tilde{V}(f) = \int_{-\infty}^{\infty} [i\,V(t)]\exp(-2\pi i\, f\, t)\,dt \quad ,$$

and the inverse transformation is then required to be

$$V(t) = \int_{-\infty}^{\infty} [-i\,\tilde{V}(f)] \exp(+2\pi i\, f\, t)\, df \quad .$$

These definitions give the same results for Fourier-transform *magnitudes* as the ones more familiar to physicists, and they also can be used to show that (in the case of phase-sensitive triggered operation) a sine-type input will show up in the Real spectrum, and a cosine-type wave will show up in the Imaginary spectrum.

This illustrates another general point – it's easy for an instrument to get the magnitude of a Fourier transform correct (although it can be a toss-up whether the magnitude, or the magnitude-squared, gets plotted), but it takes some care with signs and factors-of-*i* to get the *phases* correct. A wise experimenter will therefore check out an instrument's phase assignments by first doing a 'control experiment' on a simple device-under-test, and only then the experiment on the system of actual interest, before being confident about phase assignments.

The physics of the complex Fourier transform of the LCR transient

Everything above used the LCR transient as a convenient example of a system giving a transient, but it was focused on the data-acquisition and Fourier-transformation technique being used. Now we temporarily focus on the LCR physical system itself, which is simple, and important, enough, to work out in detail.

The differential equation describing the LCR system in use was derived in Chapter 7, where it was solved for the case of sinusoidal drive. But the equation can also be solved for the conceptually-important cases of drive by a 'unit step' or a 'unit impulse' excitation. You might think that the unit-step response is even more specialized than response to sinusoidal excitation, but it's not actually true -- you'll learn that the unit step response of a linear system *fully characterizes* it. That is to say, **if you know the unit-step response, you can predict the response of the system to any stimulus whatsoever**(!). Put another way: if two (linear and time-invariant) systems have the same unit-step response, then they will have the same response to *any other* stimulus waveform as well. (Details on these connections are found in Appendix A10.) [Conversely, if two (linear and time-invariant) systems have the same transfer function, then they have the same unit-step response, and indeed the same response to any stimulus. (An illustration of the use of the transfer function is found in Appendix A15.)]

By one means or another, the unit-step response of the LCR system's differential equation can be found. For an initially quiescent system, subjected at time $t = 0$ to a one-time fall of V_{in} from level +1/2 to level -1/2, the output voltage for $t > 0$ obeys

$$V_{out}(t) = -\frac{1}{2} + 1 \cdot e^{-\gamma\omega_0 t}\,[\cos(\omega_0\, t\sqrt{\ }) + \frac{\gamma}{\sqrt{\ }}\sin(\omega_0\, t\sqrt{\ })], \text{ where } \sqrt{\ } \equiv \sqrt{1-\gamma^2} \quad ,$$

where as usual $\omega_0 \equiv 2\pi f_0$ defines the natural (angular) frequency of the system, and γ gives its (dimensionless) damping parameter. This form clearly shows the boundary condition of $V_{\text{out}}(t{=}0) = +1/2$, and the approach of $V_{\text{out}}(t)$ to a limiting value of -1/2 as t gets large. It also shows that the frequency of the damped oscillations is not exactly ω_0 , but instead the slightly smaller (damped) frequency $\omega_0\sqrt{}(1 - \gamma^2)$. Because of triggering, what the 770 will see is <u>none</u> of the $t < 0$ behavior of V_{out}; and furthermore, the 770 will not be sensitive to the constant, ie. the dc, term present in this mathematical solution. Then with this analytic result, it's feasible to compute the Fourier transform of $V_{\text{out}}(t)$ according to the (770's choice of) Fourier-integral representation given above. The result is

$$\tilde{V}_{out}(f) = \int_{t=0}^{\infty} i\cdot\{e^{-\gamma\omega_0 t}[\cos(\omega_0\, t\sqrt{\ }\,) + \frac{\gamma}{\sqrt{\ }}\sin(\omega_0\, t\sqrt{\ }\,)]\}\exp(-2\pi i\, f\, t)\, dt \quad .$$

The lower limit of the integral has been changed from $t = -\infty$ to $t = 0$, since (due to triggering) we are *in*sensitive to $t < 0$ behavior. This integral is still a bit messy, but can be performed by replacing the cosine and sine by imaginary exponentials. Doing so reveals some terms that are resonant when $2\pi f + \omega_0 \sqrt{} \approx 0$, ie. at 'negative frequencies', and other terms which are resonant when $2\pi f - \omega_0 \sqrt{} \approx 0$, ie. at the positive frequencies we care about. Ignoring the non-resonant terms, we get

$$\tilde{V}_{out}(f) = \frac{i}{2}(1-\frac{i\gamma}{\sqrt{\ }})\frac{i}{i\gamma\,\omega_0 - (2\pi f - \omega_0\sqrt{\ }\,)} \quad .$$

For relatively small damping, we can nearly ignore ($-i\,\gamma/\sqrt{}$) relative to 1 in the first parenthesis, and then we extract

$$\text{Re}\,\tilde{V}_{out}(f) = \frac{1}{4\pi}\frac{f - f_0\sqrt{\ }}{(f - f_0\sqrt{\ }\,)^2 + (\gamma\, f_0)^2} \quad \text{and} \quad \text{Im}\,\tilde{V}_{out}(f) = \frac{1}{4\pi}\frac{\gamma\, f_0}{(f - f_0\sqrt{\ }\,)^2 + (\gamma\, f_0)^2} \quad .$$

This is a justifiably famous, and very applicable, pair of equations! You'll notice:

- the two expressions have the same denominator, which goes to a minimum value, of $(\gamma f_0)^2$, at the location in frequency space where f matches the damped frequency $f_0 \sqrt{}(1 - \gamma^2) = f_0 \sqrt{}$.
- the leading coefficients of the Real and Imaginary terms are identical, because they are connected by Kramers-Kronig relationships.
- the Imaginary term has a constant numerator, and it is therefore a function symmetrical about the center frequency, and it has a characteristic 'Lorentzian' shape.
- the denominator goes to double its minimum value when $f - f_0 \sqrt{} = \pm \gamma f_0$, or when $f = f_0 \sqrt{} \pm \gamma f_0$. This gives the FWHM, or full-width at half-maximum, of the Lorentzian curve, as FWHM = $2\gamma f_0 = f_0/Q$, another famous and very general result.
- the Real term is by contrast *anti*-symmetrical about the same center frequency, and has the shape called 'dispersive'. It has the value zero at resonance; and it reaches a minimum below, and a maximum above, the resonance; these are respectively located at the frequencies $f_0 \sqrt{} \pm \gamma f_0$.

- the location of the zero-crossing of this dispersion curve has *first*-order sensitivity to the location f_0 of the resonance; by contrast, the value of the Lorentzian near its peak has only second-order sensitivity to shifts in f_0.

The result is that either the half-maximum points of the Lorentzian, or the minimum and maximum points of the dispersion curve, enable you to *read off* the FWHM of the resonance from the Fourier spectrum.

There is another and very general connection between time- and frequency-domains lying behind the width of this Fourier transform. We've seen the imaginary part of the Fourier transform of the decaying waveform has a full-width at half-maximum of

$$(2\,\gamma)\,f_0\ \ .$$

In the time domain, the amplitude of the oscillation decays according to envelope function exp $(-\,\gamma\,\omega_0\,t)$, so oscillations decay to half their original amplitude in a time

$$(\ln 2)\,/\,(\gamma\,\omega_0)\ \ .$$

The product of these two numbers is

$$\mathbf{(FWHM)\cdot(half\text{-}life) = (2\,\gamma)\,\boldsymbol{f_0}\cdot(\,\ln 2\,)\,/\,(\gamma\,\omega_0) = (\ln 2)\,/\,\pi \approx 0.2206}\ \ .$$

Note that this result is *in*dependent of either the damping, or the resonant frequency, involved (so it applies to *any* resonance of this character, ie. having exponential decay in the time domain, which corresponds to a Lorentzian lineshape in the frequency domain). Thus it expresses a general trade-off between duration of a transient on the one hand, and the width of its Fourier transform on the other. This is a purely classical feature, although multiplying through by Planck's constant gives a relation which could be written as

$$(h\,\delta f)\cdot(t_{1/2}) = (\ln 2\,/\,\pi)\,h = (2\ln 2)\,\hbar\ \ ,$$

which is sometimes interpreted as

$$(\text{energy 'uncertainty'})\cdot(\text{half-life})\ \approx \hbar\ \ .$$

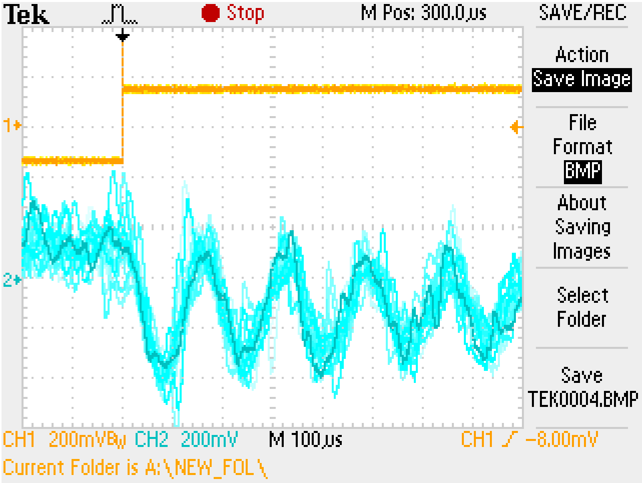
Properly speaking, the FWHM of a Fourier spectral peak is *not* an uncertainty, it's a width. Notice that the experimental uncertainty to be attached to the location of the center of a distribution can be a very small fraction of its FWHM – just *how* small a fraction would depend on statistical (signal-to-noise) and also systematic effects. But independent of any 'uncertainty principle' interpretation, and independent of any quantum effects, the trade-off still stands between a transient's duration and its Fourier transform's width.

Transient waveforms and noise

One of the new capabilities you've learned in this Chapter is triggered acquisition of transient waveforms. You've seen that a trigger event for the start of data acquisition serves to define a $t = 0$ instant, and this makes possible the acquisition of *phase* as well as magnitude information in the Fourier transform. Now in this (optional) section, we take up the issue of noise-polluted transient waveforms, and the possibility of ‘vector averaging’ in the frequency domain.

So in this section we revert to regarding the LCR system as just another physical system which can be excited by a unit-step voltage. So you can go back to the use of a low-frequency square wave as a way both to excite the LCR system, and to provide the necessary $t = 0$ trigger information to both a ‘scope and the 770. The novelty here will be to see how *noise* in the transient signal can be dealt with. Though in the research environment a transient signal might be ‘born noisy’, here in this example we will do what we’d never do in the lab – we’ll deliberately add some noise to the transient, to see how noise-reduction techniques will work.

So use your ‘Buried Treasure’ module, set to Filtered Noise, as a source of a white-noise waveform, and bring its output to the Wide Band Amplifier module for amplification. Use x1 rather than x10 settings on the switches, and set the Variable gain knob to about 3 on its scale. Now bring that amplified noise to one input of the Summer module, and bring the LCR transient signal to the other input of the Summer. (You have the option of filtering the noise on its way to the Summer, using for example a 20-kHz low-pass filter; this will *not* change the noise spectral density in the neighborhood of the 6-kHz resonance, but it will reduce the rms measure of the noise waveform.)
Now bring the summed output to the ‘scope, and get a time-domain view of the now-noisy ringdown transient signal. Your ‘scope is still triggered by the same $t = 0$ criterion, but each new transient will get a fresh overlay of noise. Your ‘scope can display the results, and you can adjust the gain in your ‘noise channel’ to go all the way from (full signal and no noise), through (full signal and a bit of noise), to (full signal and its burial under noise). Fig. 9.5 below displays a transient signal of initial value 200 mV, immersed under noise which has an rms value of about 200 mV as well.



M Pos: 300.0µs

Action Save Image

File Format BMP

CH1 200mV B_W CH2 200mV M 100µs CH1 ⨍ −8.00mV

Current Folder is A:\NEW_FOL\

Fig. 9.5: The LCR-circuit’s ringdown transient, now subjected to the addition of white noise

Of course the deliberate addition of noise to a signal is *not* an end in itself, but it’s done here just to provide a model for downstream signal processing which is also applicable to transient signals for which the noise is intrinsic, or built-in. What noise reduction techniques are available in these contexts?

In the time domain, there is a capability, on most digital ‘scopes, for the averaging of multiple acquisitions. This depends on the ‘scope being triggered, so that each successive acquisition has the *same* time history of transient signal, but a fresh and random overlay of noise. Under such circumstances, acquisition and averaging of N noisy traces yields a result with the full signal, but the noise reduced by order $N^{-1/2}$. For $N = 256$, this gives a 16-fold improvement in the signal-to-noise ratio of the result. (Of course it also entails a 256-fold slowdown of the response of your display to any actual physical changes in the system.) Try this out, and also try changing the R-setting of your LCR module to produce a change in the ringdown response, and watch the tradeoff between noise reduction and promptness of response.There are parallel methods of noise reduction which can be conducted in the frequency domain. Given that your 770 is getting a $t = 0$ trigger for each acquisition, here too you can acquire, digitize, Fourier-transform, and display the results. Here the results will be spectra, but you have a choice of what to display, and should try both methods a) and b) below.

a) If you choose to display the Magnitude (or Log Magnitude) of the spectrum of the LCR transient, you expect to see the resonant peak corresponding to the frequency content of the ringdown. If white noise is also present, you expect the display to reflect that presence. What will be the difference? Since the noise is white, it has a Fourier spectrum which is flat-in-frequency, but non-zero. So the Magnitude plot will display the LCR circuit’s resonant peak, but now lying atop a flat background of non-zero size. Now if you use the Average function, you’ll see that *fluctuations* in the flat background (and the peak) will diminish – but you’ll see the background settle toward a *non*-zero constant, which is the average noise power per spectral bin.

b) If instead you choose to display the Real and Imaginary parts of the spectrum of the LCR transient, you will see a *different* result upon averaging. Now you are sensitive to the *phase*, and not just the magnitude, of the noise. And since (relative to the $t = 0$ trigger) the phase of the noise is random, each new noisy acquisition is as likely to give a negative contribution to that average as it is to give a positive one. So now you’ll see the effect of averaging is similar (in that fluctuations will be reduced), but also different (in that averaging reduces the effect of noise *toward zero*, and not toward the non-zero average noise power).

This technique is sometimes called ‘vector averaging’, and it’s another motivation for using the phase sensitivity that the triggered acquisition of transients can bring you. Using either technique, fluctuations are reduced by a factor of $N^{1/2}$ for N acquisitions; but it can be very useful to average toward a zero noise contribution, instead of toward the non-zero average noise power.

Chapter 10: Modulated waveforms – modulation by a pulse

You'll recall that Chapters 3, 4, and 5 of this Manual addressed three types of modulated waveforms, and now we take up a fourth type. At one level, we have in mind here the simplest type of modulation, as we just 'turn a carrier-wave on and off' after the fashion of Morse code. At another level, we are describing what happens in the digital modulation of a light-beam in fiber optics, turning on and off a laser beam. At a third level, we are going to get to conclusions of immensely broad applicability, including to fields as different as optical diffraction, radio-frequency interferometry, and atomic spectroscopy.

Again we will adopt the language of radio transmission, and think about a carrier wave (here, a sine wave, of perhaps 50 kHz frequency, ie. of 20 μs period) and a modulating wave (here, an on-vs.-off sort of wave, which has 'on' and 'off' intervals that might be milliseconds, or seconds, long). And we can be very concrete about these things, since we can get the carrier wave using the Source Out function of the 770, selected for Sine waveshape, 50-kHz frequency, and (maximal) 1000-mV amplitude. The modulating waveform we want is square-wave in character, except we want a wave whose two levels are 0 and height H, not the usual $-A$ and $+A$. There are several ways you might get such a wave:

- you can use your external generator in its square-wave mode, but then add a zero-offset, so that by adding a dc offset of $+B$, you change a square wave alternating between $-B$ and $+B$ into one alternating between 0 and $+2B$;
- or, if you have a generator lacking a dc-offset control, you can bring its output to the Summer module of your electronics, and then use the -5-V to +5-V DC Voltage knob at another of your Modules to provide the other input to the Summer. The right adjustment of that dc level will then give the desired waveform.

But both these options leave you with a '50% duty cycle', in which the modulating waveform spends as much time high as it does low. You'd like to be able to vary those proportions, so you could

- look for a duty-cycle adjustment on your generator,
- or you might be able to find a <u>pulse</u> generator, preferably one with a quiescent level of 0 V, relative to which you can get steep-edged pulses of 5 or 10-V amplitude, occurring at a rate, and with a width, which are separately knob-adjustable,
- or you might have a protoboard, or a computer-based system, in which you can control a *digital* output line to vary from logic-low to logic-high. If it's a TTL-compatible output, then it will deliver <0.4 V when low, and >4.0 V when high. You can put that signal into the Summer as well, if you want to bring the shifted low-level to zero Volts.

Any one of these methods will provide a modulating signal sufficient for most of the exercises below.

Now you're going to use your analog Multiplier module as a 'gate', by sending your sine-wave carrier into one input, and your modulating signal into the other. If you have a sine wave of form $A \cos(2\pi f_c t - \phi)$, and a modulation signal that alternates between levels 0 and M, then the multiplier output will be

$$V_{out}(t) = [A\cos(2\pi\ f_c\, t - \phi)]\cdot\{0 \text{ or } M\} / (10\,V) = \{\frac{A\cdot M}{10\,V} \text{ or } 0\}\cos(2\pi\ f_c\, t - \phi) \quad .$$

So the output is either a constant times the carrier, or it is zero. The question is – what does this modulation do to the *spectrum* of $V_{out}(t)$? The wrong answer says that the output is either zero (giving no spectrum) or the constant carrier (giving a delta-function spectrum), and hence that the spectrum remains a delta-function at frequency f_c. The right answer shows that you expect a modified, in fact a *broadened*, spectrum, even if the modulating waveform is just a one-time or single pulse.

[There is another way to produce a finite-duration burst of carrier, quite distinct from using a pulse to get a multiplier to 'open a gate'. That alternative is to use a digital-synthesis signal generator to produce a 'tone burst'. Typically you set such a generator to produce sine waves of the frequency and amplitude you wish, and then select the 'burst mode', and specify an integer number of cycles of sinusoid to produce. If you have such a generator, check to see if you can get a one-time burst, or a series of bursts separated adequately in time. Check if you can get a trigger-out pulse when the burst occurs, or if you have to provide a trigger-in pulse to initiate the burst. Finally, check to see if you can specify the initial phase of the waveform – to select a sine-type start, a cosine-type start, or some other starting phase to the tone burst.]

Now once you have set up one of these methods to produce a burst of carrier wave, you want to see in the time domain what's going on. Your modulating waveform (which produces the burst) is on for perhaps 1-3 ms, and then off for a long time, such as >150 ms. That waveform goes to the Multiplier, but also to the trigger inputs of both a 'scope and the 770. You want both of these configured to trigger on the rising edge of your modulating waveform. You definitely want to use the MEASure button, and the Windows softkey, of the 770, to select the **Uniform** window, so that you will be uniformly sensitive to input voltages during the whole of its acquisition time.

Now think, and look with a 'scope, in the time domain first. At the rising edge of your modulating signal, you have enforced the $t = 0$ point for both 'scope and 770. Then for 1-3 ms thereafter, the modulating signal is high, and the multiplier output will show the sine-wave oscillations of the carrier coming through. Meanwhile the 770 will be acquiring data during the whole of its acquisition window, which is 4 ms long (for a choice of frequency span of 100 kHz) or longer still (for a narrower frequency span). You do want to be sure that the modulating waveform does not go to the high level again anywhere during this acquisition window. You should get facile with your 'scope, and look on both the 1-ms and the 100-ms time scales to be sure you understand when everything is happening. You might also make a hand-drawn 'timing diagram' to assure yourself that (for each trigger event) the 770 is seeing exactly one brief 'pulse' of the presence of the carrier. Remember also the much *shorter* time scale, of about 20 μs, which is the period of one cycle of the 50-kHz carrier.

Now the 770 should show you the Fourier spectrum of that pulse-modulated carrier, and the spectrum will *not* be a delta-function at f_c. Instead, it will display a curious pattern, centered on f_c, and with a fixed shape spread symmetrically around f_c, but with a width controlled by the duration of your modulating pulse. For a modulating pulse of on-duration 1.0 ms, your spectrum will display a peak at 50 kHz, but deep dips at 49 and 51, 48 and 52, . . . kHz. Have a look at that spectrum, using spans as wide as 100 kHz and as narrow as 3.125 kHz. [Be aware that if you choose *too* narrow a span, you will have chosen too long an acquisition time, and you will not be detecting a single pulse of carrier-wave – your ‘scope will show you why.]

The lesson is that a steady carrier has a delta-function spectrum, but a carrier that is present only for a one-time burst of duration τ has a *broadened* spectrum. We’ll now compute why that happens, assuming the pulse lasts from time 0 to τ, that it doesn’t recur during the acquisition time in question, and also that the carrier wave has a $t = 0$ phase that’s some particular value ϕ. We’ll do that by computing the Fourier transform according to

$$\tilde{V}(f) = \int_{-\infty}^{\infty} V(t) \exp(+2\pi i\, f\, t)\, dt \quad .$$

To stipulate that a pulse occurs only once, we can change the limits of integration from the mathematical (-∞, ∞) to the physical (0, τ), since that brief interval is the only time the carrier is non-zero. The result is

$$\begin{aligned}\tilde{V}(f) &= \int_0^{\tau} A\cos(2\pi\, f\, t - \phi)\exp(+2\pi i\, f\, t)\, dt \\ &= \tau \frac{A}{2} e^{i\phi}\, e^{\pi i (f - f_c)\tau}\, \mathrm{sinc}(\pi(f - f_c)\tau) + \mathrm{extra} \quad .\end{aligned}$$

Here sinc(x) ≡ (sin x) / x is a function repeatedly arising in Fourier methods, and the term marked ‘extra’ is a similar one, but with argument $\pi(f + f_c)\tau$ instead, and which is (for that reason alone) often negligible. If we care only about the magnitude of the Fourier transform, then we get the prediction

$$\left|\tilde{V}(f)\right| = \tau \frac{A}{2} \left|\mathrm{sinc}(\pi(f - f_c)\tau)\right| \quad .$$

It is no surprise that this grows with the amplitude A, or the duration τ, of the pulse-of-carrier-wave that is being sent in. It is also no surprise that the result is symmetrical in frequency around the carrier frequency f_c. The novelty is that this predicts for the spectrum not a delta-function, but a distribution of non-zero width around f_c. In fact, the first zeroes on either side of f_c come when the argument of the sinc, and hence the sine, function reaches ±π. That comes at frequency locations $f = f_c \pm 1/\tau$. Test these predictions by separately varying f_c and τ; now you understand why the use of τ = 1 ms gave deep minima at ±1, ±2, . . . kHz from the carrier.

Notice that the modulated carrier is *no longer monochromatic*. Its center frequency matches that of the un-modulated carrier, but it now is distributed over a range of frequency space. The shorter the pulse duration, the greater is this spread. This effect is very real, and ultimately limits (for any kind of carrier) how fast information can be transmitted by modulating that carrier. It certainly should tell you that two carriers, assigned to two distinct locations in frequency space, cannot co-exist if either is modulated too fast, because of the resulting overlap of the spectra of the now-modulated carriers.

A further prediction of the theory is that the magnitude of the Fourier spectrum does *not* depend on the phase of the carrier wave. It is very easy to test his prediction, since you are using the 770 in a triggered mode, and are therefore in a position to be sensitive to the phase, and not just the magnitude, of the Fourier transform.. You might try a two-panel display, showing on one panel the (linear) magnitude spectrum, and on the other panel the Real part of the Fourier transform. You should see a stable |sinc|-function for the linear-magnitude plot, but for the real-part plot, a curious pattern which fluctuates, and looks different on every successive trigger. The explanation lies in the effectively-random phase ϕ, which (after all) captures the state-of-phase of the carrier at the instant the modulating waveform turns on. As long as the carrier wave and the modulating wave come from independent generators, it is more than likely that this phase will be effectively random.

Now that you've worked on the experiment, and the theory, for this single-pulse-of-carrier case, it would be well to understand these results from a deeper point of view. What you are doing is literally multiplying a carrier wave $A \cos (2\pi f_c t - \phi)$ and some modulation signal $M(t)$, and producing an analog output voltage $V_{out}(t)$. But there is an elegant prediction for the Fourier transform of such a product, namely that it is a <u>convolution</u> of the two transforms separately:

$$F[\, M(t) \cdot A \cos (2\pi f_c t - \phi)] = F[\, M(t)\,] * F[\, A \cos (2\pi f_c t - \phi)\,] \quad .$$

Here we're using $F[]$ to mean 'take the Fourier transform of', and * to represent the convolution operation. Of the terms on the right-hand side, it is easy to understand that the Fourier transform of the carrier wave is a delta-function at $f = f_c$. (Actually, there is another delta-function at $f = - f_c$, but you're not seeing its effects.) Then the convolution of that delta-function just gives back the same shape as $F[M(t)]$, but moves its center to the frequency location $f = f_c$.

So now we know what to compute for <u>more general</u> modulating-waveform functions $M(t)$ – we can have them start at $t = 0$, or be centered at $t = 0$, and it won't matter for the magnitude of the Fourier transform. For any $M(t)$-choice, we need only compute its Fourier transform (without any reference to the carrier) to know what lineshape it will create for the Fourier transform of the modulated carrier.

For example, if we revert to a one-time pulse that is τ wide in time, and M high in voltage, we ought to reproduce the previous calculation by computing

$$F[M(t)] = \int_{-\infty}^{\infty} M(t) \exp(+2\pi i\ f\ t)\ dt = \int_{-\tau/2}^{\tau/2} M \exp(+2\pi i\ f\ t)\ dt$$
$$= \tau\ M\ \mathrm{sinc}(\pi f \tau) \quad .$$

Sure enough, the sinc-function emerges; this result, convolved with the carrier's delta-function at $f = f_c$, will give back the lineshape previously predicted.

This sort of calculation could easily be done for a one-time pulse of any other shape as well. You could imagine a modulating waveform which is a one-time *Gaussian* pulse centered at $t = 0$; the prediction would be a Fourier spectrum centered at f_c, but broadened to another Gaussian shape in frequency space. It is somewhat harder to check this experimentally, as you'd need to produce a Gaussian pulse in time, with about a millisecond characteristic duration. Various 'arbitrary waveform generators' or software-controlled computer systems can provide such a pulse. The technical advantage of such a modulating pulse is that the lineshape it produces in the carrier's spectrum will no longer display so much 'frequency splatter' outwards from its center at f_c. For signals which have to exist within a restricted frequency bandwidth, this is important.

There is another modulating waveform which might be easier for you to generate, and that is to make $M(t)$ a <u>double</u> pulse. You want $M(t)$ to be 'on' for a duration τ, and then off for a time T, and then to return with another 'on' pulse of duration τ again. The result of using this pulse-pair to modulate a carrier is clear in the time domain: what arrives is a burst of carrier, followed by nothing, and then another burst taken from the <u>same</u> carrier. What is not obvious is what the *spectrum* of such a waveform will be – those two 'pulses' of carrier are non-overlapping in time, but they are both pieces of the same carrier, and the cycles arriving in one pulse are thus phase-related to, and phase-coherent with, the cycles arriving in the later pulse. It turns out that the spectrum of this double-pulse waveform, while still centered at $f = f_c$ in frequency space, displays some features which depend on τ, and other features which depend on T. Here's the easiest computation, performed by the method introduced above. We need

$$F[M(t)] = \int_{-\infty}^{\infty} M(t) \exp(+2\pi i\ f\ t)\ dt$$
$$= \int_{T/2-\tau}^{-T/2} M \exp(+2\pi i\ f\ t)\ dt + \int_{T/2}^{T/2+\tau} M \exp(+2\pi i\ f\ t)\ dt$$
$$= \exp(2\pi i\ f\ (-\frac{T}{2} - \tau))\ \tau\ M\ \mathrm{sinc}(\pi f \tau) + \exp(2\pi i\ f\ (\frac{T}{2}))\ \tau\ M\ \mathrm{sinc}(\pi f \tau) \quad .$$

This is a complicated expression, but you can see that there are two parts of the integration region which give non-zero contributions to the result. You can see that sinc-type factors emerge from both sub-integrals, each of duration τ. And you can see that the final Fourier transform will involve a sum of two complex exponentials, which will add constructively or destructively, depending on the values of f and T:

$$F[M(t)] = \tau\ M\ \mathrm{sinc}(\pi f \tau) \{ \exp(2\pi i\ f\ (-\frac{T}{2} - \tau)) + \exp(2\pi i\ f\ (\frac{T}{2})) \}$$
$$= \tau\ M\ \mathrm{sinc}(\pi f \tau)\ e^{-\pi i f \tau} \cdot 2 \cos(\pi\ f\ (T + \tau)) \quad .$$

Now the magnitude of the Fourier transform is easy to see, and it gives

$$\left| F[M(t)] \right| = \tau\, M \left| \mathrm{sinc}(\pi f \tau) \cdot 2\cos(\pi\, f\,(T+\tau)) \right| \quad .$$

This is not yet the Fourier transform of the modulated carrier which you'll see on the 770, but it is very simply related to what you will see. That's because the convolution of the result above with the delta-function at f_c just shifts this whole function, until it's centered on f_c instead of around $f = 0$ as above.

The exciting part is that it so clearly shows the presence of two factors. One is the same sinc-factor you've seen before, in the single-pulse-of-carrier case. That factor depends only on τ, the width of the single pulse(s). The factor that's *new* to the two-pulse experiment is the cosine term, and this factor depends on T+τ, the center-to-center separation in time of the two pulses of carrier. If you choose T to be rather larger than τ, then this cosine factor gives closely-spaced 'wiggles' which then get 'enveloped' by a sinc-type function of less rapid variation.

An example of the result is plotted below:

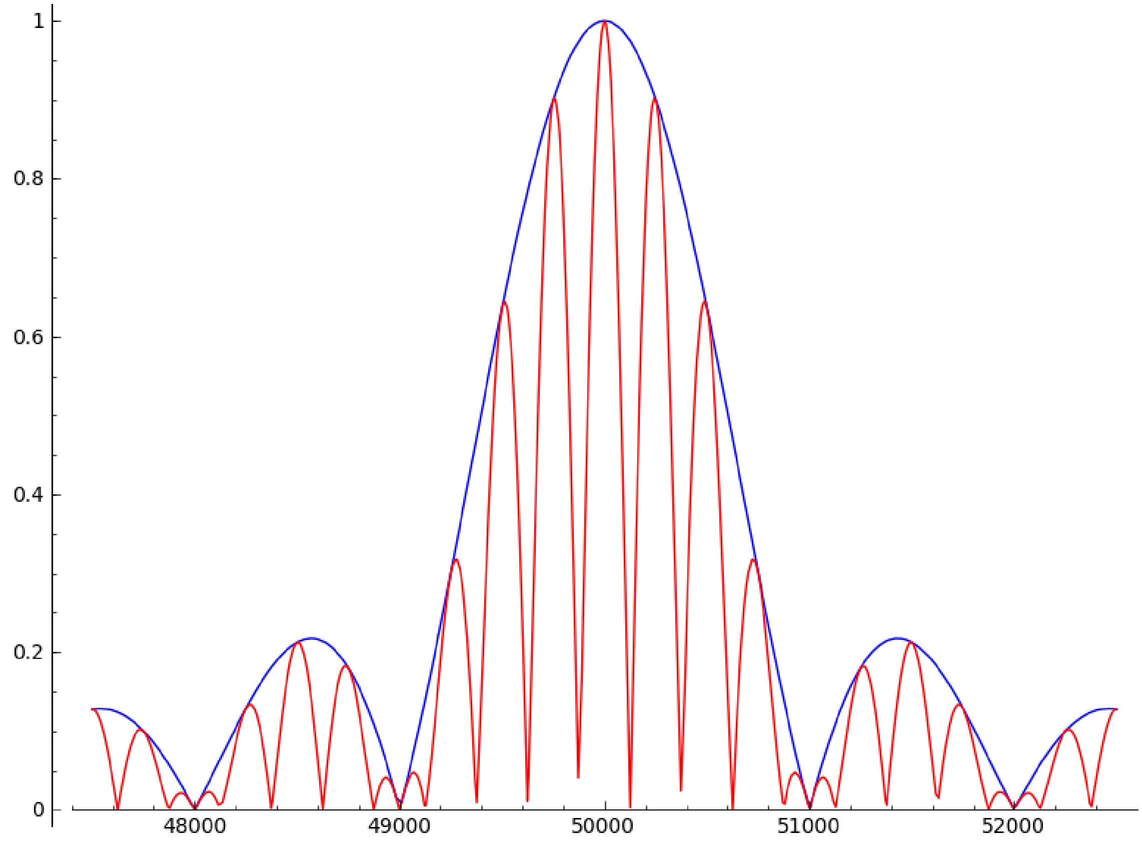


Fig. 10.1: The (linear) magnitude of the Fourier spectrum, predicted for carrier frequency 50 kHz, and assuming for modulation two 1-ms pulses, with a 3-ms separation between them. The 'envelope' function also shown is the predicted shape for a single-pulse modulation.

There is some remarkable physics going on in this graph, since there now appears some 'fine structure' in the frequency spectrum. If you needed to locate the center frequency of the spectrum above to high precision, it is clear that you could do better using the closely-spaced 'fringes' of the two-pulse pattern than using the broad envelope which would result from the one-pulse pattern. The reason, of course, is that a one-pulse pattern corresponds to being sensitive to the carrier only for duration τ (here, 1 ms). But the two-pulse pattern represents a measurement in which carrier waves are arriving for (parts of) a full duration of $T+2\tau$ (here, 5 ms). In fact, you should observe, or compute, the spectra you'd get from two alternative methods:

- one pulse of carrier, but now for a total duration of 5 ms; and
- two pulses of carrier, using two 1-ms pulses with 3-ms 'off time' between them.

These two methods do *not* give the same spectrum, and the spectral resolution you could achieve is nearly two-fold *better* using the two-pulse sequence – even though the measurement time is limited to a total of 5 ms though in both cases.

In atomic physics, this technique is called 'Ramsey's method of separated oscillatory fields'. It involves exposing an atomic system to interrogating radiation to induce a quantum transition, and it achieves a high spectral resolution by making that exposure come in two episodes of duration τ, separated in time by interval T.

But the same physics is present in many other contexts. If you were to plot the magnitude-squared instead of the magnitude (as above), you'd see a clone of a two-slit interference pattern(!) from optics, in which the width chosen for each slit corresponded to τ, and the center-to-center separation of the slits corresponded to $T+\tau$.

Another illustration of the same physics comes from synthetic-aperture radio telescopes. If you were not content with the angular resolution of a radio telescope of 10-m diameter, you could try to enlarge that aperture to 50 m. Or, you could build a 'synthetic aperture' having a width that large, by using two 10-m telescopes, laid out with a 30-m gap between them. The latter choice is not only vastly cheaper than the single huge telescope, it also gives better angular resolution – and the separation or 'baseline' can be varied, and extended from 50 meters to kilometers, or even intercontinental dimensions.

This will teach you that Fourier methods go beyond connecting the time- and frequency-domains. In physics there are many other pairs of conjugate variables which are connected by a Fourier-transform pair. In quantum mechanics, you may have seen position x and wavenumber $k = p/\hbar$ form such a pair. In optics, the pair is transverse position, and transverse wavenumber (related to angle-off-axis of propagation). The same pair of variables shows up in radio astronomy, which is optics under another guise. One of the values of 'Fourier thinking' is that it lets you form powerful analogies connecting quite distinct branches of physics, and therefore permits you to transport your intuition from familiar, into novel, contexts.

Chapter 11: Down-conversion and demodulation of AM radio

In this Chapter we apply many of the concepts gathered from previous sections, with the goal of turning your Electronic Modules into a working AM-band receiver. You should have done the experiments of Chapter 3, so as to understand the concepts of amplitude modulation, and sidebands. You will need an external signal generator capable of producing sine waves of 1-Volt amplitude in the 0-2 MHz band, but otherwise you have all the parts needed. The goal is not just to duplicate what any cheap table radio can do, but for you to understand *how it's done*.

Reception

AM radio in North America is transmitted via electromagnetic waves of frequencies 540 - 1600 kHz. This entails wavelengths of 500 down to 200 meters, much longer than any plausible antenna structure. So the electric and magnetic fields which arrive are very nearly uniform-in-space over the dimensions of your antenna. The easiest antenna to use is a length of wire, in which electrons will experience forces due to the E-field of the radio wave. These will create radio-frequency (rf) currents within the wire, flowing between the antenna's 'open end' and its grounded end.

Because the antenna couples to the field largely in a capacitive manner, the rf currents will be enhanced if the antenna is 'loaded', with some series inductance added to the capacitance-to-E. Find the Fourier Methods antenna structure included among your equipment, and spot the long antenna wire, and the series inductors that can be included in the rf-current path. There's a jumper-wire which will allow you to choose how much inductance to put in series; that choice will shift the (rather broad) resonance of this LC-resonant system, adapting it to the frequency you wish to receive. For starters, choose to put 2 or 3 inductors in series with your antenna.

Because the LC-resonance is broad, the antenna responds rather *un*selectively, so the antenna currents will be a superposition of signals at a host of frequencies. The amplitudes are under the 1-mV level, so you'll first use the Wide-Band Amplifier module to amplify them, at least to the point where you can see them on an oscilloscope. But you'll see *nothing* like a textbook picture of an AM waveform, since you are still receiving a whole band-full of AM stations.

Down-conversion

You could try to isolate a single station's broadcast by building a frequency-selective, ie. narrow-band, amplifier, for example, one with high gain in the 800-820 kHz region (and low gain elsewhere) for picking up a station assigned to frequency 810 kHz. But it's easier to impose this frequency selectivity *later* in the process, and to apply a downconversion process <u>first</u>. (Look up the 'superheterodyne' concept.) So you want to send the superposition of AM signals to one input, and the single-frequency sine wave from a local oscillator to the other input, of the High-Frequency Mixer module you have.

[See Chapter 4 on diode mixers to learn how this mixer works. Your signal generator serves as the LO or local-oscillator source, and for it to operate the mixer, you need it to deliver a sinusoid of amplitude 0.7 to 1.1 Volts at the mixer.]

What frequency do you choose for the LO? That depends on, or rather that *controls*, what station you're going to select. If your target station is assigned a frequency of 810 kHz, then an LO set either to 730 kHz or to 890 kHz will yield, at the mixer's output, a 'difference frequency' of 80 kHz. Of course the mixer's output will also contain sum-frequency terms, and sum- and difference-frequencies due to many other radio stations.

But only two frequencies on the 'AM spectrum' can be the source of what emerges from the next stage – the IF or intermediate-frequency amplifier, an amplifier of adequate frequency selectivity and fixed pass-band frequency. For an IF amplifier, use your Filter module, set to 80-kHz frequency, and Q set to 8, and use its *band*-pass output. The filter serves as an amplifier of gain 8 (near 80 kHz), but gain <<8 elsewhere. [The Filter's input impedance of about 10 kΩ also serves as the necessary output load for the High Frequency mixer.] The full-width at half-maximum of the Filter's response is 80 kHz/8 = 10 kHz, so its passband is the 75-85 kHz range. So only two AM-band channels could contribute to the emerging signals – one station located 80 kHz above your LO frequency, and the other located 80 kHz below. (If you're in a region where *both* these channels are assigned to actual stations, you have a problem – selectivity of the LC-tuned antenna, or of an rf-amplifier stage upstream of the mixer, would be the cure.) You'll 'tune your radio' by changing the LO frequency.

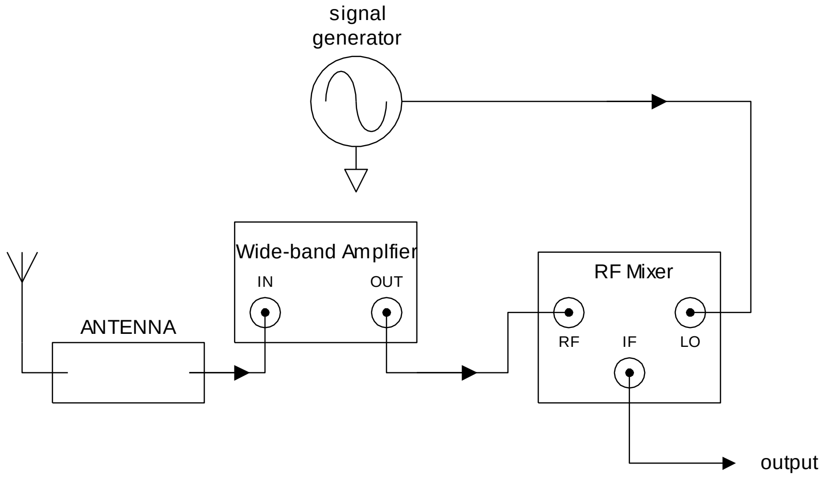


Fig. 11.1: A block diagram of some of the modules needed for AM reception

Look with a 'scope at the output of your Filter section, and look for evidence of signals with frequency about 80 kHz (ie. period about 12 μs). If you see anything at all, you'll get a vastly more informative view if you send the Filter output to the SR770, set to its full-span (0-100 kHz) coverage, and set for Log Magnitude display. As usual, you can use the AutoRange and AutoScale buttons to adapt the 770 to your signal strength.

It is not automatically true that you'll see anything significant, because you might have 'tuned your radio' to an empty place on the dial. So do the calculation needed such that your favorite AM station, or the locally-strongest AM station, will be properly tuned in –

there will be *two* LO settings which would work, and you can choose either. Now look on the 770, in the 70-90 kHz region, for evidence that you're seeing the AM signal from (a single) station. You should see the 'down-converted carrier' as a spectral line at 80 kHz, and you should see the 'down-converted sidebands' symmetrically disposed to either side of it, spreading out in continua about 5 kHz either side of the carrier. You should see that the carrier is steady, but that the sidebands fluctuate – varying with the program content being broadcast by the station, of course.

Before you go on, try increasing your LO frequency by 5 kHz. Does the downconverted spectrum move in frequency space? How far? And in which direction? And why? You can also use this occasion to 'tune your antenna', ie. maximize your signals by varying the inductance in your antenna system. You might also change the gain in your high-gain amplifier section; you'd like to amplified signal from the antenna to reach the 1-Volt level when it reaches the mixer, but not be any greater than this – else you'll get clipping in the mixer.

Tuning

Now that you know what to look for, first find back the station you're targeting with the use of the *other* choice of LO frequency that will also give a 80-kHz difference frequency. Now try a systematic variation of the LO frequency from (540 - 80) kHz to (1600 + 80) kHz. Changing the LO frequency by 10 kHz at a time is sufficient (because 'AM channels' are assigned with 10-kHz quantization). During the course of your search, any station you can detect should appear *twice*. (Why?) Make a note the LO frequencies that 'work', infer the AM-station carrier frequencies that are available for reception, and categorize the stations into stronger vs. weaker. For some of the stronger stations, try antenna-tuning for optimal signal strength – you should expect higher-frequency signals to require smaller series inductance to bring them to LC-resonance.

De-modulation

After the downconversion process, you have imposed selectivity on the AM spectrum by using the bandpass filter, but you have *not* created an audio signal. What's needed is de-modulation, the recovery of the modulating signal that went into the original production of the broadcast AM waveform. There are various methods for doing this.

a) envelope detection
If you have a waveform of 80-kHz carrier plus sidebands showing up on your 770, and if it has >10-mV amplitude, you can see it directly on your 'scope. Choose, not a sweep speed that would show individual cycles (of about 12 μs period) of the 80-kHz carrier, but instead a slower sweep speed, perhaps 1 ms/div. Now find out how to get your 'scope to give you a Single Sweep (rather than continuous triggering), and then manually acquire and examine a series of these single-sweep acquisitions. Each one should show a 'smear' of cycles of downconverted carrier, with 80 full cycles of oscillation fitting into each millisecond of time. What you're looking for is time variation, on the millisecond timescale, of the *amplitude* of these oscillations.

(The ‘peak detection’ mode of your ‘scope will be helpful here.) For example, the presence of 1-kHz program content in the AM station’s broadcast would be the reason for a 1-ms period of such fluctuations. You should see the (rare) quiet times of AM broadcast giving a steady *non*-zero amplitude of 80-kHz oscillations; and the non-quiet program content will raise, and lower, this amplitude on a ms-timescale.

If you can see these oscillations with their fluctuating envelope, you could execute the textbook ‘envelope detection’ method of extracting the program content. It’s only easy if you can get Volt-level amplitudes of the 80-kHz oscillations, because then a trivial one-diode half-wave rectifier, followed by a bit of capacitance, will give you a dc level which basically follows the successive peaks of each (say, positive) excursion of the 80-kHz signal. Apart from having a dc offset, that envelope gives back the audio waveform which you want to send to the Power Amplifier and Speaker modules.

If the 80-kHz signal has an amplitude <0.1 Volt, you’d need something more clever than just a diode. You might care to build, on a protoboard, an op-amp circuit called a ‘precision rectifier’. It could be a half-wave, or a full-wave rectifier (also known as an absolute-value circuit). Either method will give you the ‘envelope detection’ mentioned in any textbook.

b) heterodyne to ‘zero beat’

The problem with the signal you have is that all the audio content of the broadcast is frequency-shifted up to the 80-kHz region. But you could change that, by tuning your LO frequency to *match* the station’s carrier frequency. Now the carrier and LO get mixed to a zero-frequency signal, and each sideband (say, 440 Hz to either side of the carrier) get mixed to produce a 440-Hz signal. Of course after this sort of mixing, you want a filter section which passes audio frequencies, from zero upwards. So change the Filter section to a 3-kHz frequency, a Q of 0.71, and use the *low*-pass output; this will send onward all, but only, frequencies in the 0-3 kHz range.

In principle, you can match your LO frequency exactly to the carrier frequency you’re receiving, at which point the difference (or ‘beat’) frequency gets right down to zero. To be concrete: if an AM station is assigned carrier frequency 810 kHz, and transmits as program content a 440-Hz concert-A tone, then its spectrum includes a carrier at 810.00 kHz, and two sidebands at 810.00 ± 0.44 kHz. The previous use of an LO frequency of 730 kHz downconverted all three frequency components (simultaneously, by linearity) to the frequencies 79.56, 80.00, and 80.44 kHz. But now if you tune the LO to 810 kHz, the new values of difference frequencies are |-0.44|, 0.00 and 0.44 kHz. In principle, you should get a signal, coming right out of the low-pass filter, which is a dc level plus (two copies of) a 0.44-kHz = 440-Hz tone.

So try this out – pick the strongest station you’ve found, tune up your antenna, tune the LO to give ‘zero beat’, send the filter’s output on a scope and the 770, and send it to the Power Amp and Speaker modules, and see if you can hear any broadcast content. You will find that you hear annoying fluctuations (called ‘motorboating’), even if you tune your LO very carefully (meaning, at the 1-Hz level). Why is this?

The problem is the exact matching required. Suppose the AM station really does broadcast exactly at 810 kHz = 810,000 Hz; then in our example its sidebands are at 809,560 and 810,440 Hz. If your LO is set to be *even 1 Hz off* of ‘zero beat’, to 809,999 Hz, then the downconverted signal includes three components which emerge from the low-pass filter:

the carrier, downconverted to 1 Hz;
the upper sideband, downconverted to 441 Hz
the lower sideband, downconverted to 439 Hz.

You can’t hear the 1-Hz signal, though a (dc-coupled) ‘scope would show it as the dominant signal emerging from the filter. The speaker does produce, and you can hear, and the 439- and 441-Hz signals. But that’s the problem – these two signals will alternately re-inforce and cancel each other, giving the annoying ‘beats’ or fluctuations in intensity that you hear. The problem gets only worse if the mis-tuning of the LO reaches a few Hz, or a few dozen Hz. You’d prefer a demodulation scheme which does *not* require part-per-million matching of LO and station frequency.

c) product detection

We suggest you go back to method a), reverting to the use of an LO frequency 80 kHz away from the station you want to receive, and reverting to the use of the 80-kHz, Q=8, bandpass filter. In our concert-A example, the filter’s output now consists of frequencies 79.56, 80.00, and 80.44 kHz.

Now if you had a signal which was *phase-locked* to that 80.00 kHz, you could use a mixer, or your Multiplier module, to mix

(79.56 & 80.00 & 80.44) kHz with that (80.00) kHz ,

and you’d get the |-0.44|, 0, and +0.44 kHz frequency components you want. And the phase-locking would ensure that the ‘0’ really was zero, and that the |-0.44| matched the 0.44 kHz, to abolish the motorboating.

While a phase-locked loop can be built, there is a simpler way. Instead, you can use your Multiplier module to mix

(79.56 & 80.00 & 80.44) kHz with *itself* .

Not only does this provide the output you want, it is also immune to 1-Hz or even 10- or 100-Hz error in the LO tuning!

[Example: if the LO were to be set not to the desired 730.00 kHz but 10 Hz = 0.01 kHz higher, the downconverted spectrum would include components at (79.55 & 79.99 & 80.43) kHz. Now you’d be mixing that spectrum with itself, ie. with (79.55 & 79.99 & 80.43) kHz. The difference frequencies which emerge include (three) exactly-zero frequencies, and (four) exactly-0.44-kHz or 440-Hz frequencies, with*out* any 10-Hz errors. (There would also be some 880-Hz content, which would (for sizeable modulation index) constitute 2nd-harmonic distortion of the output signal.)]

Because of the use of a multiplier function, this is called ‘product detection’, and it is robust again moderate detunings of the LO [as method b) was *not*].

Chapter 12: Deterministic chaos, in time- and frequency-domains

You are perhaps culturally aware of 'the butterfly effect', or the technical term 'chaos', and in this Chapter you'll get to study one example of a chaotic system by the powerful combination of time- and frequency-domain methods. We've including among your Electronic Modules one called Chaos, which is a working electronic realization of the so-called 'Lorenz attractor'.

What is the 'Lorenz attractor'?

The Lorenz attractor is the name given to the solutions of a set of differential equations. It played an important role in the discovery of *chaos*, and the realization that a system evolving entirely deterministically could nevertheless show a degree of unpredictability which would previously have been attributed to randomness.

Stripped of its context, the Lorenz system is an initial-value problem, a set of first-order differential equations for three real-valued functions conventionally called $x(t)$, $y(t)$, and $z(t)$, evolving from initial conditions $x(0)$, $y(0)$, and $z(0)$. In their usual dimensionless form, the system of equations is

$$\frac{dx}{dt} = s\,[- x(t) + y(t)] \quad ,$$

$$\frac{dy}{dt} = r\,x(t) - y(t) - x(t) \times z(t) \quad ,$$

$$\frac{dz}{dt} = x(t) \times y(t) - b\,z(t) \quad ,$$

where the parameters s and b have canonical values of 10 and 8/3 respectively, and where r is a 'control parameter' which may take on values ranging from <1 to >100.

Given any t=0 set of initial conditions, a system like this evolves *deterministically*. That is to say: in the entire mathematical universe, there is exactly one triple of functions $\{x(t), y(t), z(t)\}$ which evolves from the specified initial conditions, and which solves the Lorenz equations. The entire future evolution of the x, y, and z is therefore fully fixed and determined by the $t = 0$ initial conditions, with no influence of randomness or indefiniteness.

What came as a surprise to physicists whose intuition grew up within Newtonian mechanics was that, despite the determinism of the equations, the Lorenz solutions could display boundedness (staying contained within finite limits) yet an entire *absence* of periodicity, and the solutions could also exhibit a rapid loss of effective predictability. The features of the solutions include 'sensitive dependence on initial conditions' and a 'strange attractor' in *xyz*-space, and these have become signatures of deterministic chaos. To produce such behavior requires a system of at least three first-order differential equations, and the presence of at least one non-linearity in the differential equations. Many other systems exhibiting chaos are now known.

Why is it important?

The Lorenz system is important for several reasons. One is that it arose in (an approximation to) a mathematical model of an actual, and interesting, physical system. Another is that there is nothing mathematically suspect, pathological, or discontinuous about its specification. A third reason is that it has been investigated at sufficient depth, and in the right contexts, to enable general lessons to be drawn from it. In addition, it was historically important in the emergence of chaos as a paradigm for understanding time evolution of systems. As a result, the Lorenz system helped to make it clear to physicists that deterministic time evolution after the fashion of Newtonian mechanics could nevertheless lead to solutions which exhibited the kind of unpredictability which would previously have been expected only from quantum mechanics or thermodynamic randomness. To see unpredictability-in-practice emerge from a *classical* system of only *three* degrees of freedom constituted a genuine surprise, and an enduring lesson, to physicists.

It is worth describing the physical system which gave rise to the Lorenz model, in part to understand why it might be expected to show certain kinds of behavior. It came from an idealized description of a problem in fluid mechanics, in which a horizontal layer of fluid is in contact with a hotter surface below, and a cooler surface above. For small enough differences between these two boundary temperatures, such a fluid can stay at rest, transporting heat from bottom to top purely by conduction. But beyond a threshold value of the temperature difference, motion can arise in the fluid, due to thermal expansion and buoyancy. In a two-dimensional idealization of the fluid layer, the first mode of convection predicted to break out is a series of rolling convective cells, characterized by 'unit cells', in each of which there is an ascending column of heated fluid that cools at the top surface and descends in falling columns of fluid. An approximate treatment of this problem uses a dimensionless $x(t)$ function as a measure of the rate of fluid rotation in each such cell, uses $y(t)$ as a measure of the temperature difference between the rising and falling columns of fluid, and uses $z(t)$ as a measure of the deviation, from the linear profile characteristic of pure conduction, of the vertical temperature profile at the cell's center.

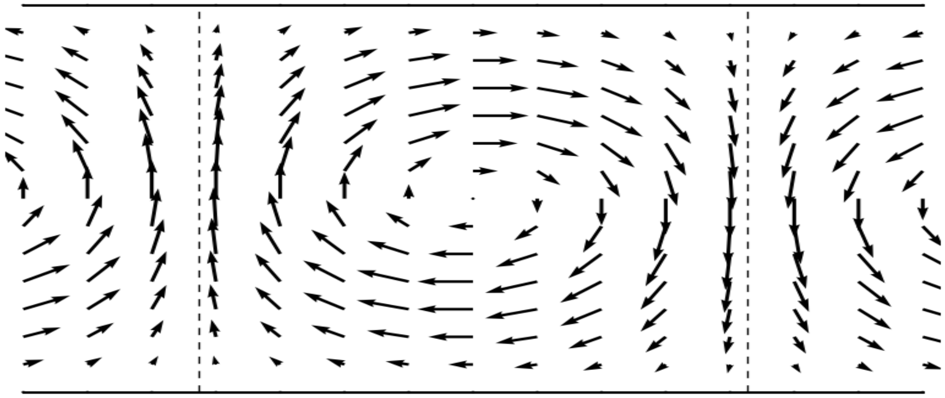


Fig. 12.1: A sketch of velocity flow vectors in some 'convecting cells' of a fluid contained between two planes, and heated from below

For the scaling which Lorenz adopted, the parameter choice s = 10 and b = 8/3 emerged as suitable for the description of *air* as a fluid. The application Lorenz had in mind was the earth's atmosphere, heated from below, by contact with sun-lit land, and becoming convectively unstable when the degree of heating-from-below (quantified by the r-parameter) became sufficiently large.

It is feasible to establish some results for solutions to the Lorenz system. From a physical point of view, we'd call the choice $0 < r < 1$ the 'conductive regime', because in this range, any initial conditions evolve toward a solution (or 'attractor') $x = y = z = 0$. So in the steady state, the fluid is not rotating, shows no temperature difference across a cell, and shows a purely linear temperature profile in the vertical direction.

Similarly, any choice $1< r < 24.73$ leads to a 'steady convection' regime. In this domain, it is possible to show that any initial condition evolves toward the steady-state values

$$x = \pm\sqrt{b(r-1)} \quad , \quad y = \pm\sqrt{b(r-1)} \quad , \quad z = r - 1 \quad .$$

So after initial transients die away, the system settles toward a state of either clockwise or counter-clockwise rotation in each cell, with a steady angular velocity and a steady (though non-linear) vertical temperature profile. In these solutions, x and y have the same sign, which corresponds to ordinary convection, with the warm fluid rising.

[The problem of small oscillations about these steady-state solutions can also be worked out analytically, and this leads to the prediction of a transition point at $r = 470/19 \approx 24.7368$. Short of this r-value, such oscillations are exponentially damped; but at this r-value, the decay rate passes through zero and changes *sign*. (The frequency of these small oscillations about the steady-state solution is also computable, and predicts a period-in-t of 0.6528 units of dimensionless time, in the vicinity of this transition point.)]

Thus for a choice of $r > 24.74$, the system does *not* settle into one of the two stable states corresponding to rolling convection, but instead the x, y and z-functions all undergo oscillations of *growing*, not decaying, amplitude. The result is large fluctuations in the three functions, with x and y both alternating erratically *in sign* (corresponding to reversal of the sense of rotation in the convective rolls).

Above this threshold in r, for values such as $r = 28$, the novel feature of *chaos* emerges. It is still true that the system of equations, together with the initial conditions, fully determines the future of the system. But two successive 'runs' of the system, starting from initial conditions of arbitrarily small separation, will evolve in divergent fashion, with differences (in at least one dimension) growing exponentially in time. The result is that predictability degrades exponentially in time, such that any degree of imprecision, however small, in the specification of the initial conditions will lead, within a finite time, to the inability to predict (for example) even the *sign* of the x-function at some given future time. In other words, **the deterministic system of differential equations soon becomes of *no more use than a coin toss* for predicting** whether the fluid will be convecting in one direction as opposed to the other.

The nature of the solution can be understood as the 'motion' of a point in an *xyz*-phase-space. As *x*, *y*, and *z* evolve in time, this point traces out a continuous curve in space, a trajectory in three dimensions. The trajectory connects deterministically to its initial condition; so if the trajectory were ever to cross itself, then the system would be guaranteed to be periodic in time. But amazingly, within the chaotic regime, the system does *not* exhibit any periodicity. So while evolving within a bounded regime, the phase-space point somehow traces out a continuous curve which fills a sub-region (called the strange attractor) without ever passing twice through any point in phase space. Remarkably, motion from *any* initial condition will tend toward this strange attractor.

More surprising still is that within the chaotic regime, there are 'windows' in the choice of *r*, regions (of non-zero width in *r*) in which the strange attractor collapses to an ordinary closed curve in *xyz*-space. For such *r*-values, any choice of initial conditions (other than $x = y = z = 0$) evolves toward a curve which is then traced out periodically. Thus after an initial transient dies away, the phase-space point repeatedly traces out a closed curve or 'orbit'.

How to build an electronic realization of the Lorenz attractor

It was historically important that the Lorenz system, initially studied via approximate numerical solutions of its differential equations, was soon also solved by forming physical solutions on 'analog computers'. That is to say, a set of three variables proportional to *x*, *y*, and *z* was encoded as the behavior of three voltage functions in a real-time analog-electronics circuit, itself connected so as to be described by the same set of three differential equations. The physical behavior of these voltages was found to exhibit the chaotic phenomena first found via numerical digital computation.

This observation prevented physicists from blaming either round-off errors in finite-precision digital representations, or the approximations implicit in algorithms used in the numerical solution of differential equations, for the emergence of chaos. Instead, it forced them to see that actual physical systems – such as a set of three voltage functions evolving continuously and deterministically in real physical time – could also display all the attributes of chaos. And just as digital solutions of differential equations have issues of round-off or precision issues, so too analog representations of the Lorenz system have inevitable issues such as component tolerances. The fact that chaos with parallel properties emerged from both digital and analog realizations of the Lorenz system pointed to chaos as a *robust* property of the system, not one requiring Platonic perfection for its appearance.

The TeachSpin realization of the Lorenz system executes the 'analog' method of solving the differential equations. In this example of what is a more general methodology, the dimensionless *xyz*-functions of the canonical system of differential equations are scaled to three voltage functions V_x, V_y, V_z, according to scaling choices

$$\frac{x(t)}{401} = \frac{V_x}{10\,V} \quad , \quad -\frac{y(t)}{401} = \frac{V_y}{10\,V} \quad , \quad \frac{z(t)}{401} = \frac{V_z}{10\,V} \quad ,$$

which are suggested by technical convenience. Similarly the dimensionless time-variable *t* of the canonical system of equations is translated to physical time *T* according to

$$t = \frac{T}{t} \quad ,$$

where τ is a time-scale with a value fixed by choice of an RC time-constant, switch-selected to be τ = 0.47 ms (fast), τ = 10.5 ms (medium), or τ = 331 ms (slow). The further parameter choice of *r* is set by a single adjustment knob, the TeachSpin design making accessible the range $0 < r < 364$. With these choices of scaling, the Lorenz system (with usual coefficients $s = 10$ and $b = 8/3$) requires the voltages to evolve in physical time *T* according to

$$\tau \frac{dV_x}{dT} = -10\,(V_x + V_y) \quad ,$$

$$\tau \frac{dV_y}{dT} = -\,r\,V_x - V_y + 401\frac{V_x \cdot V_z}{10\,V} \quad ,$$

$$\tau \frac{dV_z}{dT} = -\,401\frac{V_x \cdot V_y}{10\,V} - \frac{8}{3}V_z \quad .$$

Behavior of three voltages according to these equations can be enforced on the functions by building those voltages into circuits which (via feedback) cause them to satisfy these differential equations. This depends on the ability to execute ‘operations’ with voltages, a task for which ‘operational amplifiers’ are ideally suited. In particular, such circuits can create fixed numerical multiples or submultiples of voltages, and (through the use of analog-multiplier devices) can also form voltages which are the real-time product of two other input voltages. With the devices actually used in the circuitry, we form the products

$$V_{xy}(t) \circ V_x(t) \times V_y(t)/(10\,V) \quad , \quad V_{xz}(t) \circ V_x(t) \times V_z(t)/(10\,V) \quad .$$

Finally, we formally integrate the differential equations above with respect to time *T*, and get the system

$$V_x = -\frac{1}{\tau}\int dT\,(10V_x + 10V_y) \quad ,$$

$$V_y = -\frac{1}{\tau}\int dT\,(r\,V_x + V_y + 401V_{xz}) \quad ,$$

$$V_z = -\frac{1}{\tau}\int dT\,(401V_{xy} + \frac{8}{3}V_z) \quad .$$

These equations can be enforced electronically because the time-integrations needed on the right-hand sides can *also* be conducted with operational amplifiers. In fact, a good way to view this set of equations is to imagine a process of ‘boot-strapping’. If we imagine that we had three wires carrying voltages $V_x(t)$, $V_y(t)$, and $V_z(t)$, we could form from these three signals all the multiples, products, sums, and integrals indicated on the right-hand-sides of the equations above. In particular, three op-amps would be producing at their output terminals the three right-hand-sides above. Then we create the desired feedback by connecting those three outputs so as to *drive* the very three wires previously merely labelled as the V_x, V_y, and V_z wires. Thus the physical evolution of the voltages is forced by the laws of electronics to obey a set of differential equations which are a (scaled) representation of the Lorenz system.

Once the feedback is in operation, the electronic system is autonomous, free of any external drive. It therefore runs on its own, evolving from initial conditions deterministically. The only intervention necessary or possible is the adjustment of one knob, a 10-turn dial whose setting controls to the choice of r-parameter, with the scaling

$$\frac{r}{364} = \frac{R}{10} \quad ,$$

where R is the turns-count set on the 10-turn dial (so that r increases from 0, and by 36.4 units per turn of the dial). Meanwhile, the voltages V_x, V_y, and V_z are available for external study or display in real time. The only other adjustment is the choice of the time scaling τ ; this is expected to change only the rate in physical time at which the voltage-analog of the mathematical solution evolves.

We have also included a toggle switch, flipping which to Reset will send, and hold, the system to a point very near (V_x, V_y, V_z) = (0, 0, 0) Volts. When the switch is toggled back to Run, the time evolution according to the differential equations commences, with a characteristic rise of the voltage $V_z(t)$ from near-zero values providing an indication of the start of the time evolution.

Now any analog realization of the Lorenz system is subject to a mismatch with the mathematical equations. The biggest deviations of the analog electronic system from the ideal include

a) component tolerances, particularly in the values of capacitances used in the three integrators. These are of 5% tolerance, so the three integrators used might have time constants mismatched to this degree;

b) zero offsets, particularly in the outputs of the multiplier circuits (even when their inputs are zero). These are visible as non-zero values taken on by V_x, V_y, V_z-values, even in the Reset condition.

The results of such imperfections do *not* prevent the generation of chaos, but they do mean that numerical values you observe might differ slightly from the values quoted below.

How to use the Chaos module

If you are using the Lorenz-attractor module for the first time, we suggest you set the time-scale switch to *medium*, and the 10-turn r-dial to the bottom (r = 0) end of its range. You should also connect a dual-trace oscilloscope to view the $V_x(t)$ output on the 'scope's ch. 1, and the $V_z(t)$ output on ch. 2. Arrange for the scope to run in its automatic-triggering mode, at a rate of 10 ms/division. Set the toggle switch to Reset, and you should see $V_x(t)$ and $V_z(t)$ both take on, and maintain, outputs near 0 Volts.

Now flip the switch to Run, and try adjusting the 10-turn dial by 0.25 turns (ie. raise the r-value form near 0 to near 9.1). You should see V_x and V_z both change, and settle to new, non-zero, but steady values. [In fact at the new setting of $r = 9.1$ we can predict $x \rightarrow \sqrt{}b(r\text{-}1) = \sqrt{}(8/3 \cdot 8.1) = 4.65$, and $z\rightarrow r - 1 = 8.1$, which under the scaling previously adopted ought to give $V_x\rightarrow 0.116$ V, $V_z\rightarrow 0.202$ V. In practice, you'll see the effects of zero-offsets in these voltages.] To see that the analog-electronic system has some dynamics to it, toggle the reset switch briefly to Reset, then back to Run, and observe the V_x- and V_z-values take some *time* to reach their steady values.

To see the oscillations about steady state, raise the r-value to near 12 (0.33 turns on the dial), and notice the new V_z -value to which the system settles. Set the 'scope's trigger mode to normal, and set it to trigger on a ch. 2 signal (ie. on V_z) which is rising through a level of about 0.20 V. Now when you flip briefly to Reset and then back to Run, the 'scope should trigger on, and capture, the time history of $V_z(t)$ rising from 0, and experiencing decaying oscillations, while settling to its steady-state value.

When you have seen this, and have picked a proper time scale for the 'scope to display this, raise the r-value to 18 or 24 (0.49 or 0.66 turns). Now by this same Reset-Run method you should see oscillations (of a somewhat different period) which have a *longer* decay time. When you have optimized your view of these longer-lasting transients, continue to raise the r-value just a bit more. Ideally, near 0.68 turns on the dial, you ought to see the oscillations cease to decay, ie. start to *grow* instead. This growth will rapidly lead to *non*-small oscillations, and the system will break into (chaotic) oscillations. The R- and hence r-value you find for threshold might differ from these ideal numbers, but you will find a threshold for chaotic behavior.

For oscillations not far above this threshold, you can check that $V_x(t)$ falls into the range $-0.6 \text{ V} < V_x(t) < +0.6 \text{ V}$, $V_y(t)$ falls into the range $-0.7 \text{ V} < V_y(t) < +0.7 \text{ V}$, and $V_z(t)$ falls into the range $+0.4 \text{ V} < V_z(t) < +1.8 \text{ V}$. To get a view of the oscillations, continue to trigger on V_z-values, but set the trigger level to near the *top* end of the range which $V_z(t)$ covers. You've arranged to view the behavior of the system which follows the attainment of relatively rare large values of $V_z(t)$. With this sort of triggering on $V_z(t)$, use the 'scope's second trace to give a simultaneous display with $V_x(t)$ [or alternatively $V_y(t)$], to see that all three outputs display chaotic oscillations. Arrange to display many cycles of individual oscillation, so that you can see $V_x(t)$ and $V_y(t)$ are both functions which alternate, quite irregularly, in sign. (By contrast, you'll see that $V_z(t)$ takes on only positive values.)

To see that the chaos in this system is 'self-starting', set the $V_z(t)$ trigger level to small positive values, so you can see what follows a switch out of the Reset mode. If you have a digital 'scope with an Accumulate or Persistence mode, you can acquire and compare several traces, all evolving from nominally identical initial conditions. See if you can gain evidence for the 'sensitive dependence on initial conditions' which cause such overlaid trajectories to diverge from each other, and estimate the time-scale over which this happens. Notice that each fresh start of the system evolves to a qualitatively similar behavior, despite the inevitable differences in detail.

Now switch your ‘scope display to its XY-mode, and return to automatic triggering. You should get a 2-d display with $V_x(t)$ across, and $V_z(t)$ upwards. Set scales and zero-offsets so as to center the locus of (V_x, V_z) pairs on the screen. Now use the persistence mode of the ‘scope so as to view some time history of the curve which the instantaneous point $(V_x(t), V_z(t))$ is tracing out. You should see the famous owl-face or ‘butterfly’ diagram, a view of the Lorenz system’s ‘strange attractor’.

You’re seeing a 2-d projection of a curve which is a trajectory in a 3-d voltage space. You can of course view (V_y, V_z) or (V_x, V_y) pairs by the same technique, to see two other 2-d projections of the same 3-d trajectory. It is not so easy to see the true three-dimensionality of the trajectory in real time (unless you can find a ‘scope which offers a back-panel ‘z-axis modulation’ as well as XY-display capability.) But it is true that, despite appearances in the 2-d projections, the trajectory never crosses itself in 3-d phase space.

In this mode, you are making ‘parametric plots’ such as of $(V_x(t), V_z(t))$ pairs, and the only effect of time-scale setting is the rate at which the curves are being traced out on the screen. In this mode, you can try the three positions of the speed switch, to see if (on the slow setting) you can follow by eye the time evolution of the flying spot, or if (on the fast setting) you get an apparently steady curve. The claim is that the shape of the attractor on your screen should not change, but only the rate at which the curve is being traced out in physical time. The ‘fast’ setting is of course ideal for showing, in short order, what happens when you change the r-parameter’s value.

When you have your favorite projection of the Lorenz attractor in a live view, repeat the Reset toggling to confirm that each fresh evolution (from *very* subtly distinct initial conditions) nevertheless leads to an attractor of the same appearance. Then with the attractor running in real time, slowly raise the r-adjustment to see how the attractor changes. You should see the attractor expand to fill a larger range of V_x and V_y, and get even more positive in V_z. Adjust scales and offsets as needed to keep it in view.

To first appearance, the attractor only changes in detail, not in character, as a function of the r-setting. But if you are careful, you should spot, near $r \approx 100$ (≈ 2.8 turns on the 10-turn dial) a region in which the attractor *collapses* to a closed curve. You should confirm that this ‘window of periodicity’ extends over a small region of r-adjustment. To see that this curve is an attractor (now a *non*-strange attractor), toggle to and from Reset a few times, and see if (after a transient) the system settles to this same periodic orbit after each Reset.

There are other windows of periodicity you can find for other values of the r-parameter. To find them, continue raising the setting on the 10-turn r-adjuster. Some of these ‘windows’ are wider than others, and therefore give more robust periodic behavior than in the first window you might have found.

Time- and frequency-domain views of the operation of the Lorenz attractor

You have now had a 'tourist's view' of some of the features of the Lorenz attractor, either in the time domain (as in $V_x(t)$ and $V_z(t)$ plots) or in phase-space projection (as in the locus of (V_x, V_z) values). Other things can be learned from these two views, but even more can be learned in the frequency domain, by taking the Fourier transform of (any one of) the outputs of the Lorenz-attractor module.

To practice on this, we suggest that you set the 10-turn dial to just *below* the onset of chaos. You'd like to see $V_z(t)$ rise, from its near-zero value at a Reset, to larger positive values, and then settle toward a constant by decaying oscillations. When you practice this with a time-domain view on a 'scope, measure the period of the oscillations as they approach steady-state, and infer from this a frequency of oscillations. (The frequency will depend on your setting of the speed switch, of course.)

Now send the $V_z(t)$ signal to the SR770 spectrum analyzer, and arrange to trigger the 770 as you have triggered the 'scope. The settings you'll want are those suited to observing a transient which fits in the 0-to-10-Volt range, and has frequency content centered at the frequency you computed. Now you should be able to trigger on, and capture, the transient which follows a Reset-and-Run combination. Since this is a decaying (and not a continuous) oscillation, you should see a spectral peak on non-zero width, centered on a frequency matching what you've computed.

Once you've configured the 770 to succeed in this (harder) task, adjust the *r*-setting to take you into the regime of chaos. Now the time development of $V_z(t)$ is ongoing, so you can change the 770 to a continuous-triggering mode. Now the oscillations are not even close to sinusoidal, so there will be frequency content to higher frequency than you've seen so far. So extend the span of frequency coverage to show this. Now the oscillations are *not* periodic, so you expect to see not a 'line spectrum' of a fundamental plus its harmonics, but instead a 'continuous spectrum'. The spectrum you see will drop off rapidly with frequency. So the spectrum will *not* be white noise, and is in fact expected to drop off faster than f^{-n} for any power of *n*. To see this, arrange for a logarithmic display of the frequency scale on your 770 (the vertical scale is already logarithmic in spectral power density, if it's displaying in dB units). A power-law spectrum $S(f) \propto f^{-n}$ would give a down-sloping straight line in the view you're now seeing – your continuous spectrum should drop faster than that. In fact, it drops exponentially with frequency. (See Appendix A4 for discussion of the reasons for these differences.)

You are seeing the spectrum of a time-domain waveform so close to random that the power spectrum will show the fluctuations to be expected in any random-noise spectrum. You can arrange to use the Average mode (over 16 or 64 or more acquisitions) to see how these fluctuations average away to give you a stable view of the underlying power spectrum of the waveform you're analyzing.

The power spectrum changes its character, in several ways, as you adjust the *r*-parameter of the system in real-time operation.

1. As you increase r in the $r > 25$ range, you recall that $V_z(t)$ oscillates erratically but with larger amplitude. So you can expect more spectral power as you raise r: this will lift the spectral-density curve *up*ward on your display.

2. As you increase r, you are 'driving' the system harder, and thus causing things to happen faster in time. So it is not surprising that spectral content should extend to higher frequency: that spreads the spectral-density curve *right*wards in your plots.

3. As you increase r (carefully), you can fall into windows of periodicity which are dramatically different in their time-domain behavior. Even more dramatic are the consequences in the frequency domain. (For quicker response time to changes in r-settings, you might *dis*able the averaging mode for this investigation.) Whenever the strange attractor collapses to a periodic orbit, the continuous spectrum has to collapse to a line spectrum. You will see spectral power at the 'fundamental frequency', whose reciprocal is the period of the system in time, and you will see harmonics of this fundamental. Are they located at integer multiples of the fundamental, as they should be? How narrow, in frequency, are the peaks? (Use the BMH setting of the Window menu for an optimal view, and reduce the frequency span and narrow in on a peak to investigate.) How narrow in frequency *should* these peaks be? How high (in dB) does the fundamental peak stand above the 'noise floor' of your measurement system? How much better does this figure of merit become when you use spectral averaging?

4. There is another phenomenon best studied in the frequency domain, even if it can also be seen in the time domain. This is called 'period doubling', and it is a phenomenon which has been observed in the approach to chaos of *many* dynamical systems. In the Lorenz system, it can be observed in the periodic windows mentioned above. In your Lorenz attractor, it is best to start with the periodic behavior found in the window near $r \approx 100$. The question is, what happens for r-values within the periodic window (ie. not in the chaotic regime), but not at the center of the window? The phenomenon is best seen in frequency space, since periodicity with period T causes the continuous spectrum to collapse to a line spectrum, with peaks at frequency $f_1 = 1/T$ and its harmonics $2f_1$, $3f_1$, etc. When you have found an r-value yielding such a spectrum, try some very small changes in the r-value, and look for change in the power spectrum in the vicinity of the frequency $(1/2)f_1$. The right setting for r will give a spectrum with a (weak) spectral peak at $(1/2)f_1$, a (still-strong) peak at f_1, another (weak) peak at $(3/2)f_1$, and so on. Since this set of peaks is a line spectrum consisting of the harmonics of a <u>new</u> fundamental, of frequency $(1/2)f_1$, this corresponds to a time-domain waveform which must have a period of $2T$, not T as before. Hence the term 'period doubling'.

If you can find a condition in which the $(1/2)f_1$ content is as big as possible relative to the spectral line at f_1, it is worth looking for this period-doubling in the <u>time domain</u>. In a $V_z(t)$ graph, for example, get a 'scope display of two full cycles, with three successive peaks in $V_z(t)$, and see if you can establish that the 1st and 3rd cycles are clones, but the intermediate 2nd cycle is a bit different from both of them. So instead of successive identical cycles of length T of the poetic form AAAA . . . , you should look for the pattern ABAB . . . , so that the actual periodicity is formed by repetitions of the (AB) motif, for an period of $2T$.

An alternative way to see period doubling is in the phase plane, which will show a single closed orbit for the period-*T* signal. For operation under period-doubled conditions, the orbit ought to split into two nearly, but not quite, overlapping traces for part of its closed curve.

You will find that the frequency domain is the better place to spot period-doubled behavior, since the presence of even a *weak* spectral line at frequency (1/2)f_1 stands out so well from the noise floor of the system. A peak that is even 60 dB down (10^{-6} in power, 10^{-3} in Fourier amplitude) of the main peak at frequency f_1 is readily detected, particularly when using the averaging mode and the BMH choice of windowing function. By contrast, a fractional change of order 10^{-3} between alternating cycles in the time domain is *very* difficult to spot.

Now suppose you've found the *r*-value (call it r_1) at which you get period-*T* behavior, and a nearby one (call it r_2) for which period-doubled behavior can be observed. In principle, a further change in *r*, of size even smaller than $|r_1 - r_2|$, will take you to an r_4-value, at which the period will have doubled again, to 4*T*. If you can find such an r_4-value, it'll show up best in the frequency domain, giving a spectral peak at frequency (1/4)f_1. And so on: there is an ever-more-closely-spaced sequence of *r*-values r_1, r_2, r_4, r_8, . . . lying along the 'period-doubling route to chaos'. So closely spaced do these *r*-values become that you may see an apparently sudden change from period-2*T* or -4*T* behavior directly to chaos. But it is important that you should have seen at least one 'generation' of period doubling, because this is a progression on the route from periodicity to chaos characteristic of many dynamical systems.

Chapter 13: Harnessing harmonic distortion -- the fluxgate magnetometer

What is a fluxgate magnetometer?

A 'fluxgate magnetometer' is a device for measuring magnetic field strength, based on the non-linear properties of ferromagnetic materials. It is sensitive to a single component of the vector magnetic field, and can sense the magnitude and sign of that component. Among the range of technologies that could be used to measure magnetic fields, the fluxgate method is distinguished by its high sensitivity, fast response, and low power consumption. Devices based on fluxgate technology are robust and reliable enough to be put onto spacecraft, and sensitive enough to detect the nanoTesla fields found in interplanetary space.

To make clear the niche that fluxgate devices can fill, consider first the range of field strength one might want to measure. High-current electromagnets and solenoids generate fields up to 20 Tesla; the ambient magnetic field near the earth's surface has strength 30 - 60 μT (1 μT = 10^{-6} Tesla); and the most sensitive magnetometers measure down to the pT (= 10^{-12} T) range. In the high-field regime, techniques of nuclear magnetic resonance provide a suitable detection technology, while in the low-field limit, the best techniques depend on superconducting quantum-interference devices (SQUIDs), on or advanced optical-pumping techniques. In the mT-to-T range, Hall-effect or other solid-state devices can be used. But to measure μT fields like the earth's, and with nT (= 10^{-9} T) resolution, millisecond response time, and considerable accuracy, using compact, low-power, and ambient-temperature electronics of modest cost, the fluxgate technique is likely to be the method of choice. And in connection with Fourier Methods, you will also see that the operation of a fluxgate magnetometer is a wonderful illustration of the power of thinking in the frequency domain.

Applications for measuring a 50-μT (= 50,000-nT) field with nT resolution include geomagnetic and geophysical exploration, and other reconnaissance technologies, civilian and military.

How a fluxgate magnetometer works

A fluxgate magnetometer contains, in its sensor proper, a ferromagnetic material which is subject to the external magnetic field B_{ext} (to be measured), and also to an ac magnetic field generated by a sinusoidal current sent into a solenoid wound around the sample. Then the presence of B_{ext} is 'read out' via the emf that is generated (according to Faraday's Law) in a secondary or 'pick-up coil' wound around the sample and the excitation solenoid. Everything else is done with clever electronics, with no moving parts, nor any need for vacuum, cryogenics, etc.

To understand the sensor's operation, it is simplest to think of a long thin rod-like sample, aligned along the direction of the external field $\boldsymbol{B}_{\text{ext}}$. Then the physics of the sensor can be understood in a one-dimensional calculation, in terms of the triad of fields B, H, and M. Here M is the magnetization density of the ferromagnetic material, the magnetic moment per unit volume in the material. Here *B* is the 'magnetic induction' which appears in the Lorentz force law. Finally *H* is the auxiliary 'magnetic field', connected to *B* and *M* (in SI units) by

$$\vec{B}=\mu_0(\vec{H}+\vec{M}) \quad .$$

The advantage of introducing the *H*-field is that in this thin-rod geometry, the value of *H* is fully determined by the current *i*(*t*) in the solenoid, and the number of turns *n* per unit length in the solenoid, according to

$$H(t)=n\,i(t) \quad ,$$

despite the complicating effects of the ferromagnetic response of the material filling the solenoid. In particular, in a ferromagnetic material, *H* is a good choice of *in*dependent variable, since it is under an experimenter’s direct control via the current *i*(*t*).

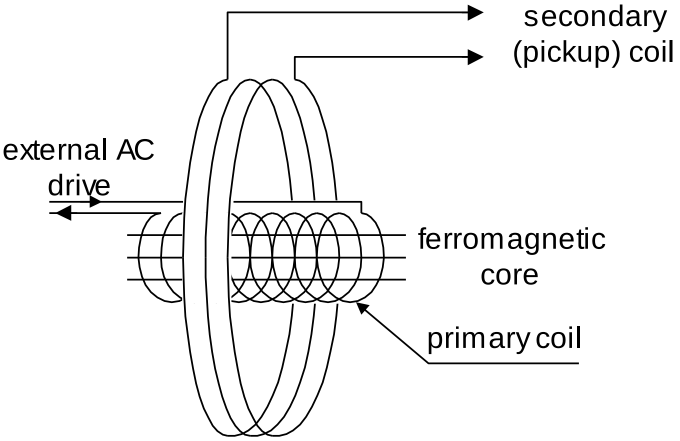


Fig. 13.1: A block diagram of the simplest possible fluxgate magnetometer

The sample's response can be predicted via the *M*-vs.-*H* plot appropriate to the material. In SI units, *M* and *H* have the same dimensionality (both are in Amperes/meter or A/m), and para- and dia-magnetic materials display simple, and linear, response to *H*:

$$M \propto H \quad ,$$

with constants of proportionality which are positive for paramagnets, negative for diamagnets, and of size far below one in either case. But ferromagnetic materials have much more complicated responses, described by *M*(*H*) functions or curves, with physical effects including saturation and hysteresis. Most importantly, a relatively small *H* (from external solenoid currents) can 'leverage' the ferromagnetic material to produce a *much* larger *M*-value, by re-orienting the magnetization of pre-existing ferromagnetic domains in the material.

Now while a real fluxgate sensor might use a ferromagnetic material exhibiting both hysteresis and saturation, all that is *essential* to its operation is some amount of non-linearity in the *M*(*H*) function, as will be illustrated below.

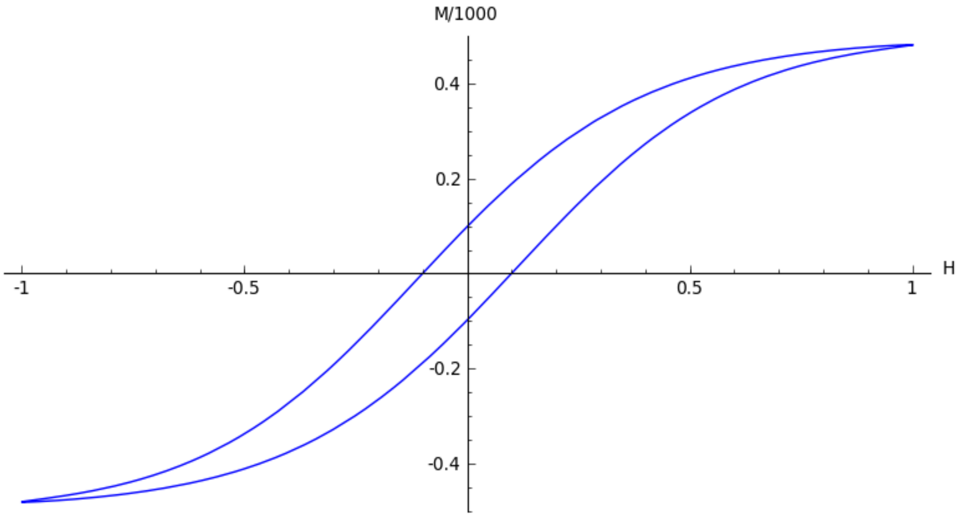


Fig. 13.2: A plot of magnetization *M* as a function of magnetic intensity *H* for a generic ferromagnetic material, taken through a cycle of variation in *H* – axes not to scale.

In the plots above and below we have drawn curves appropriate to an isotropic material, one whose properties are unchanged under an end-for-end reversal. (A sample with permanent magnetization along one direction would *not* be such a material.) For an isotropic material, the response *M* to a reversed excitation -*H* has to be the opposite of the response to excitation +*H*. That is to say, the *M*(*H*) function or curve must obey

$$M(-H) = -M(+H) \quad ,$$

which expresses the 'inversion symmetry' displayed in the diagrams above and below. So if *M*(*H*) is a single-valued function, it needs to be an odd function of *H*. Then the simplest model for an *M*(*H*) function with non-linearity has, for its small-*H* behavior, the expansion

$$M(H) = a_1 H^1 - a_3 H^3 \quad .$$

This cubic expansion is of course to be trusted only in the vicinity of *H* = 0. We could write it in the alternative form

$$M(H) = (\mu - 1)\, H\, [1 - (H / H_n)^2] \quad .$$

In this form, the quantity μ would be called the 'initial permeability', and the interest of ferromagnetic materials lies in the fact that μ might be as large as 10^{+4} - 10^{+6}. In this form, there is a second parameter, a characteristic *H*-strength H_n scaling the onset of non-linearity, and again, the interest in 'soft' ferromagnetic materials lies in the fact that H_n might be of order 10^3 A/m, so that $\mu_0 H_n$ is of order 10^{-3} T. For *H*-values rather smaller than H_n (which is the only region in which the cubic-nonlinearity model above can be trusted), the *M*(*H*) curve has the shape shown below.

Finally, for use in the fluxgate magnetometer with ac excitation, in order to keep eddy-current losses low, the ferromagnetic materials used need to be poor electrical conductors or insulators, so ferrite materials are of interest.

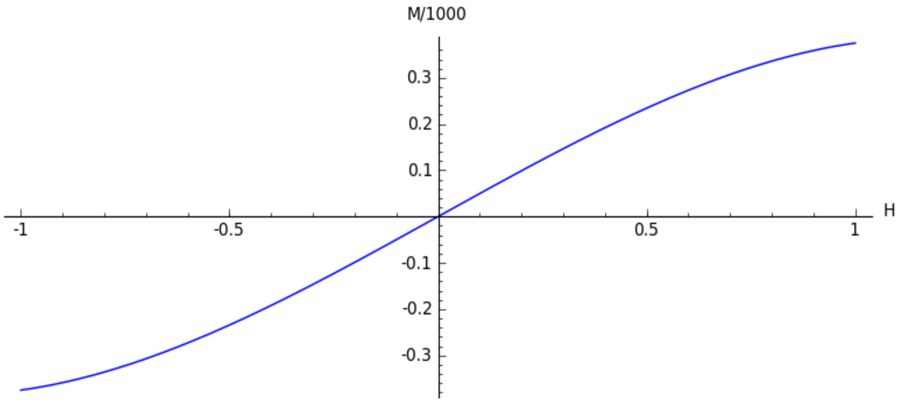


Fig. 13.3: A schematic plot of an *M*(*H*) curve exhibiting cubic non-linearity near the origin.

Now the operation of a fluxgate magnetometer can be understood graphically. We imagine the solenoidal coil of turn-density *n* is driven by a sinusoidal current of amplitude i_0 and (angular) frequency ω, so that the *H*-field in the sample is given by

$$H(t) = n\,i(t) = n\,i_0 \sin \omega t \quad .$$

This is the time history of the independent variable, and the *M*(*H*) curve we have adopted will *map* it to the time history of a dependent variable *M*(*t*). For *H*(*t*) of a size which drives the sample into a regime where non-linearity starts to have an effect, this entails that *M*(*t*) will *not* be a pure sinusoid, but will suffer some distortion. For an *H*(*t*) drive which is symmetrical about zero, this distortion has the form shown below:

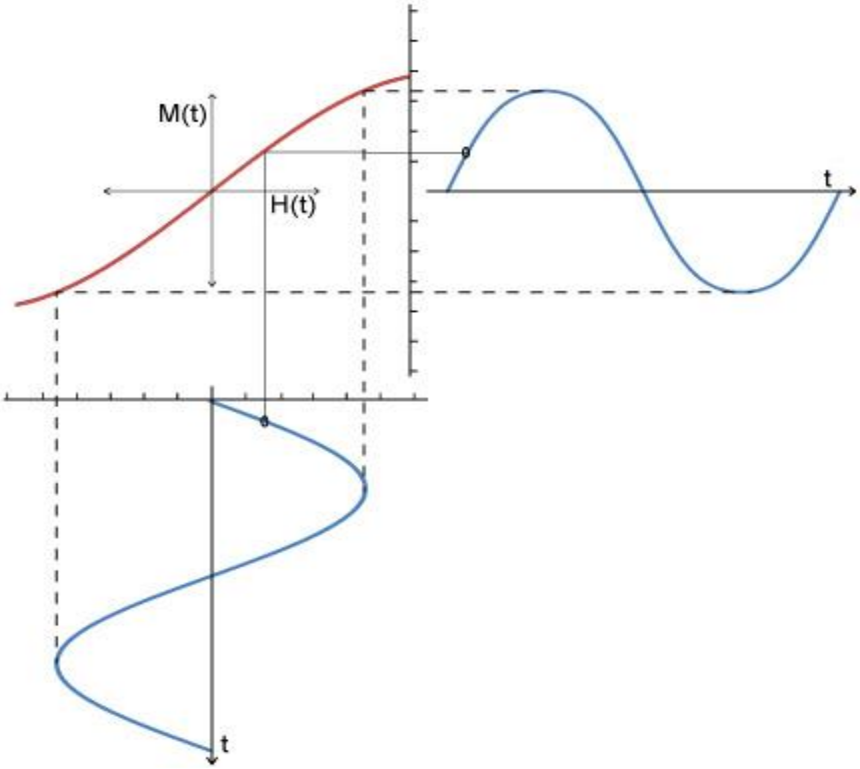


Fig. 13.4: A schematic diagram of how a non-linear *M*(*H*) relationship (upper left) maps a sinusoidal excitation *H*(*t*) (below) into a *non*-sinusoidal response *M*(*t*) (at right) .

We will see below that such an $M(t)$ function contains 3rd-harmonic distortion, in addition to the main or fundamental term of frequency ω. But the interest of this device as a sensor comes when $H(t)$ contains in addition a static, or offset, term

$$H(t) = H_e + n\, i(t) = H_e + n\, i_0 \sin \omega t \quad ,$$

corresponding to the steady effect of the external field, with $H_e = B_{ext}/\mu_0$. Now the excursions of the independent variable $H(t)$ are *not* centered on the inflection point of the $M(H)$ function, so the mapping from H to M produces an $M(t)$ function with a *new* kind of distortion, shown in the plot below:

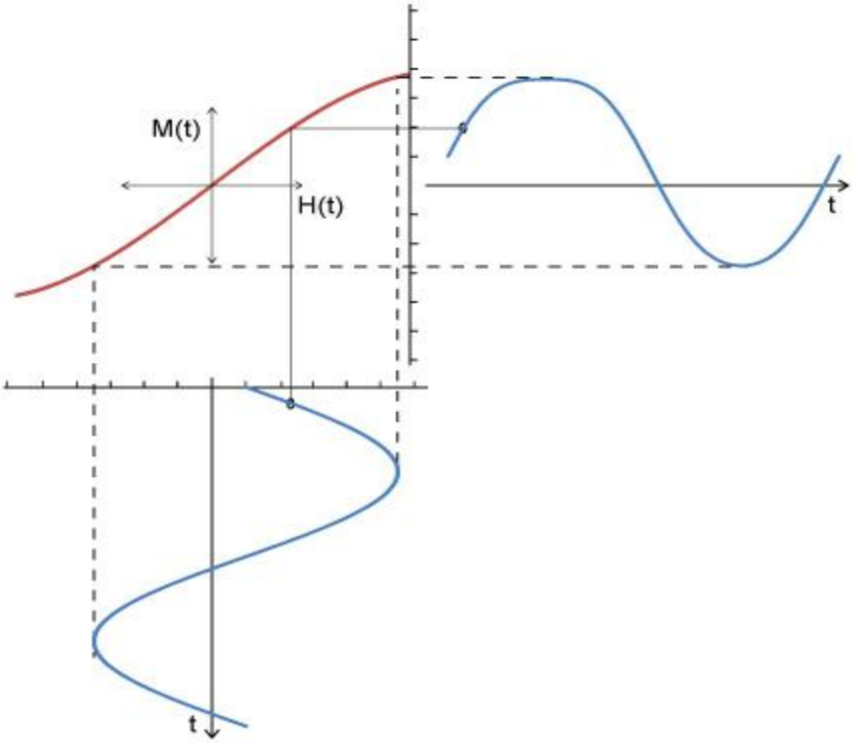


Fig. 13.5: A schematic diagram of the mapping of $H(t)$ (below) into $M(t)$ (at right), in the case of an $H(t)$-function which is *not* symmetrical around zero.

It is already clear geometrically above, and will be shown algebraically below, that in addition to the fundamental term, and the previously-discussed 3rd-harmonic distortion, this new sort of $M(t)$ response contains a 2nd-harmonic distortion, a term at a novel location in frequency space. (The visible cue to this 2nd-harmonic term is the distinct curvature in the positive vs. the negative lobe of the $M(t)$ function at right in the figure above.) This Fourier component turns out to have a predictable phase, and an amplitude proportional to the external static field B_{ext} that is to be measured, *so the detection of the 2nd-harmonic term is the basis of the fluxgate's operation as a magnetometer.*

This response in the magnetization of the ferromagnetic material might seem to be very arcane, but it is directly detectable via pick-up in the secondary coil. By Faraday's Law, the emf in a secondary coil of N turns will be

$$\varepsilon(t) = -N \frac{d\Phi_B}{dt} = -N\, A \frac{d}{dt} B = -\mu_0\, N\, A \frac{d}{dt}[H(t) + M(t)] \quad .$$

Here *A* is the cross-sectional area of the ferromagnetic sample where all the magnetic flux is concentrated. The term in *dH*/*dt* is identical to the response that would be generated in the absence of a sample. But in the presence of a ferromagnetic sample, the term in *dM*/*dt* is much larger, and furthermore, it contains the effects of the harmonic distortion created by the response of the sample. So the abstract and solid-state quantity *M* can be 'read out' via the much more concrete and electronically-accessible emf detected in the pick-up coil.

We now work out the functional form of that emf, under the assumption of the offset *H*(*t*) function, and the non-linear *M*(*H*) function, that we have introduced above. The non-linearity we've assumed in *M*(*t*) generates terms in $\sin^2 \omega t$ and $\sin^3 \omega t$, which can be transformed into Fourier components using the identities

$$\sin^2 \omega t = \frac{1}{2} - \frac{1}{2}\cos 2\omega t \quad , \quad \sin^3 \omega t = \frac{3}{4}\sin \omega t - \frac{1}{4}\sin 3\omega t \quad .$$

Then the panoply of terms arising can be gathered according to frequency, giving

$$\begin{aligned} M(t) &= [a_1 H_e - a_3 H_e^3 - \frac{3}{2} H_e (n\,i_0)^2] + [a_1 n\,i_0 - 3a_3 H_e^2 n\,i_0 - \frac{3}{4} a_3 (n\,i_0)^3] \sin wt \\ &\quad + [\frac{3}{2} a_3 H_e (n\,i_0)^2] \cos 2wt + [\frac{1}{4} a_3 (n\,i_0)^3] \sin 3wt \quad . \end{aligned}$$

Of these terms, the constant terms will give no contribution to the Faraday's-Law emf. The particular case of $H_{\text{ext}} = 0$ gives

$$M(t) = [a_1 n\,i_0 - \frac{3}{4} a_3 (n\,i_0)^3] \sin \omega t + [\frac{1}{4} a_3 (n\,i_0)^3] \sin 3\omega t \quad ,$$

displaying the terms of frequency ω and 3ω mentioned previously. But there is also, in general, a single term of frequency 2ω, and it gives

$$\begin{aligned} B_{2w}(t) &= m_0 \, [H_{2w}(t) + M_{2w}(t)] \\ &= m_0 \frac{3}{2} a_3 H_e (n\,i_0)^2 \cos 2wt \quad , \end{aligned}$$

and therefore gives an emf

$$\begin{aligned} e_{2w}(t) &= - \, N A \frac{dB_{2w}}{dt} = - \, N A \frac{3}{2} m_0 \, a_3 \, H_e \, (n\,i_0)^2 (- \, 2w) \sin 2wt \\ &= 3 N A \, a_3 \, B_{ext} (n\,i_0)^2 \; w \sin 2wt \quad . \end{aligned}$$

In terms of the material parameters μ_{m} and H_{n} previously defined, this can be written as

$$\varepsilon_{2\omega}(t) = 3 N A (\mu - 1) B_{ext} (\frac{n\,i_0}{H_n})^2 \; \omega \sin 2\omega t \quad ,$$

which makes it clear that a material with large initial permeability μ, and a small value for the onset on non-linearity H_{n}, is a desirable choice for a fluxgate magnetometer.

Notice that the amplitude of this Fourier component of the emf is directly proportional to B_{ext}, so that if the rest of the constants are known (or are found by calibration, under a known field), this emf's amplitude becomes a surrogate for B_{ext}, and that allows the external field to be measured.

How it's built

Given a model for the non-linear behavior of a ferromagnetic sample, we now have a prediction for the amplitude of the frequency-2ω Fourier component of the emf induced in the pick-up coil. For a small-scale realization of such a fluxgate sensor, suppose we have a 'soft' ferrite material with initial permeability $\mu = 10^{+3}$, and with $H_n = 10^{+3}$ A/m modeling its non-linearity. [Materials with even more favorable properties are known.] Suppose we have for a sample a rod of diameter 6 mm, so cross-sectional area $A = 28$ mm^2, and we wind it with a solenoidal coil of $n = 5$ turns/mm, ie. $n = 5 \text{ x } 10^3$ /m. If we excite that coil with a sinusoidal current of amplitude $i_0 = 10$ mA = 0.01 A, at a frequency of 1 kHz (so $\omega = 6.3 \text{ x } 10^3$ /s), we get a prediction for the amplitude of the frequency-2ω term expected in a pick-up coil of $N = 80$ turns:

$$\varepsilon_{2\omega}(t) = 3\cdot 80\cdot 28x10^{-6}m^2\cdot 10^3\cdot B_{ext}\,(\frac{50\,A/m}{10^3A/m})^2(6x10^3/s)\sin 2\omega t \quad .$$

Thus in a field of $B_{\text{ext}} = 50$ μT, we expect the 2ω-term in the emf to have an amplitude of 5000 μV or 5. mV. This is readily detectable, the more so since it occurs at an isolated location in frequency space, and since we know the frequency and the phase with which it is predicted to occur.

The are several ways to make it even more easily detectable. We note first that the sinusoidal current with which we drive the primary solenoid needs to be a pure sinusoid, or at least needs to have very low levels of 2nd-harmonic distortion. Otherwise there'd be some $\varepsilon_{2\omega}$ signal already present in the pick-up coil even in the absence of any effect of B_{ext} computed above.

Next we note that our $\varepsilon_{2\omega}$ signal needs to be detected in the presence of a much larger signal at the fundamental frequency ω. In principle this is not a difficulty (the two frequencies are easily resolved), but in practice it puts large demands on the dynamic range of the circuitry processing the emf from the pick-up coil. A much more clever idea is to double the $\varepsilon_{2\omega}$ signal, while nearly cancelling the signal at the fundamental frequency, by using two side-by-side samples lying inside one single pick-up coil.

We wind identical primary coils around the two samples, and arrange for equal, but opposite, currents $i(t)$ to pass through the coils, simply by putting the two primaries in 'reverse series' as shown below. Now the pick-up coil is sensitive to the changes in *two* bundles of magnetic flux, equal and opposite if the samples are matched. In fact, any

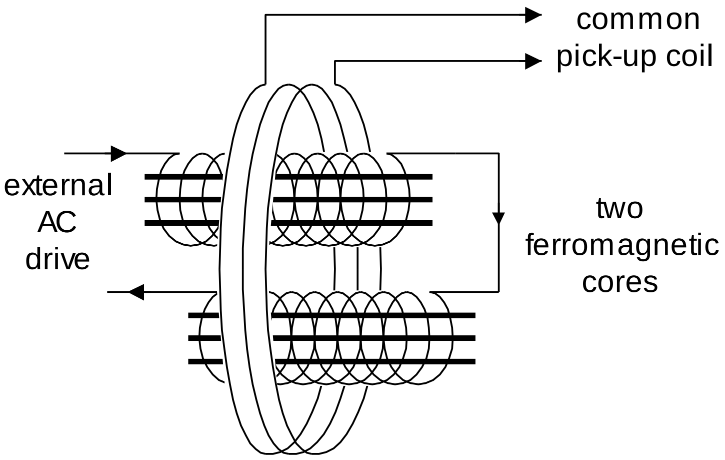


Fig. 13.6: The use of two (longitudinally-staggered) ferromagnetic elements, each wound with a solenoid as primary coil, and with one common secondary coil surrounding both

terms in the emf $\mathcal{E}(t)$ which are linear (or cubic) in the amplitude i_0 are predicted to cancel out in this new arrangement, whereas terms in i_0^2 are doubled in the total emf. On paper, all the terms at both frequencies ω and 3ω are of the form expected to cancel, while the 2ω-signal used in fluxgate magnetometry is expected to double. Even if the cancellation of the fundamental signal that is actually achieved is only at the 1% level, this still markedly reduces the frequency-ω signal that is picked up, and makes the desired signal at 2ω relatively much more prominent to the downstream electronics. In practice, there is a deliberate longitudinal offset in the positions of the two primaries, so a longitudinal sliding of the secondary coil can be used to find that location at which the net frequency-ω signal does vanish.

After all this discussion of an idealized model of a ferromagnetic material, it is worth pointing out that materials likely to be used in commercial fluxgate sensors are not well described by the nonlinear but single-valued $M(H)$ function used in the development above. Real materials certainly display hysteresis, and may also undergo saturation, so their $M(H)$ relation has the character of a hysteresis loop. But it is still the case that there is a signal with a Fourier component at frequency 2ω, induced in a pick-up coil wound around the sample and excited at frequency ω, and it is still the case that the sign and magnitude of the signal at frequency 2ω is diagnostic of the external magnetic field B_{ext}. What is not readily calculable in such a case is the coefficient which relates this measureable amplitude to the value of B_{ext}. This illustrates that fluxgate sensors are of their nature in need of calibration; but this is relatively easy to accomplish.

Using the TeachSpin fluxgate transducer

From a purely empirical point of view, the fluxgate sensor is a sort of transformer. In its TeachSpin realization, one BNC cable marked Primary will accept an ac current drive, and the other BNC cable marked Secondary will produce an ac emf. It is the second-harmonic component of this emf which is expected to be sensitive to the value of the magnetic field in which the sensor is immersed.

We suggest exciting the primary at a frequency *f* of about 1 kHz, and we suggest getting this frequency from the Source-Out capability of the SRS770 (because that can produce a sinusoid with very little second-harmonic content of its own). Because the 770's Source will produce at most a 1-V amplitude, we suggest that you use the Power Audio Amplifier module, with this sinusoid as input, to drive the primary of the fluxgate, and that you raise the gain until you get about a 6-V amplitude while driving the primary.

Now you can connect the Secondary winding directly to the 770, and look at the spectrum of the emf, using the 0-3.125 kHz Span. You expect to see a 'noise floor'; you might also see 60- or 120-Hz interference, and many harmonics thereof; but the spectrum should be dominated by spectral peaks at *f*, 2*f*, and 3*f* .

> [If the peak at *f* is larger than 10 mV, it might be worth improving the cancellation of this peak. That is achieved by finding the secondary coil – visible near the tip of the fluxgate sensor, and wound on a white plastic form – and sliding it longitudinally along the sensor. It is crucial to do this with minimal rotation, lest you tear the fine-wire connections to the secondary. Very small, sub-mm adjustments, are all that will be required.]

Next use the AutoRange control of the 770 so that it uses as high an input sensitivity as the signal at *f* permits; you're now optimized to look at the small signal at 2*f*, which is where the magnetic-field sensitivity is predicted to occur. For a first reality check, try placing a small permanent magnet at a place where it creates a magnetic field along the axis of the fluxgate, and then vary its distance from the fluxgate – you hope to see a variation in the strength of this 2*f* signal.

This is also the place to profit from a 'Fourier view' of the landscape, the noise floor, against which the desired signal at frequency 2*f* lies. The point of this exercise is to choose the value of the excitation frequency *f* so that the desired signal at 2*f* lies at a place where the noise spectrum is at a low, or floor, level. If the noise floor is 'white', ie. independent of frequency, then one choice of *f* is as good as any other. But very commonly there are peaks standing above the noise floor, and you can now steer your 2*f* signal *away* from those regions by changes in your choice of *f*. In particular, you can look for, and avoid, spectral lines of interference, commonly at harmonics of the local ac-line frequency. [The same considerations apply whenever you use lock-in detection.]

You will now be measuring the magnitude of this 2*f*-harmonic signal, so for best accuracy you should use the Flattop window choice. You should also be aware that the size of this signal is predicted to depend on (the square of) the current in the primary, so you need to control this independent variable to a known and stable value. Once you have a signal that's under control, and sensitive to the magnetic field, you are ready to calibrate your sensor's sensitivity to magnetic field.

The double solenoid

The theoretical development above shows that a detectable signal (the amplitude of the frequency-2ω term in the output emf) is predicted to be a linear function of the external magnetic field B_{ext}. But the constant of proportionality depends on the values of some parameters, including the permeability and nonlinearity constants of a magnetic material, which are not easy to measure. As a result, a fluxgate magnetometer, once built and working, needs to be calibrated. The TeachSpin fluxgate sensor is therefore built to fit into a special solenoid, to make this calibration easy.

The solenoid is mounted in a frame, so that by rotation of the wooden base on a tabletop, and adjustment of the clamping screws, the axis of the solenoid can be arranged to lie parallel to the local magnetic field. Of course you'll need a compass and dip needle to determine that direction in 3-d space. (At an indoor location inside a steel-framed building, the magnetic field might not point to 'magnetic north', and that direction in turn may differ from geographical north. And the field vector $\boldsymbol{B}_{\text{ext}}$ will certainly not lie in a horizontal plane.)

The solenoid is wound on a tube, into either end of which the fluxgate sensor will fit snugly – there's an O-ring in place for a friction fit. When the sensor is fully installed into the coil form, the active elements near the sensor's tip will be located at the center of the solenoid.

The solenoid is wound with *two* electrically-separate windings (using two-color ribbon-connector wire) which share the same volume, and turn density. Each solenoid is wound in two layers, and in each layer the 'pitch' of the helical windings in nominally 0.100" = 2.54 mm. (You can measure the actual pitch of the outer layer by counting the visible turns of one color against a ruler.) The reciprocal of that pitch gives the turn density n, the number of turns (in one layer, of one of the solenoids) per unit length; and in the limit of a long solenoid, the expected field inside the solenoid is

$$B_{\text{ext}} = 2 \cdot \mu_0\, n\, i \quad ,$$

where i gives the current in the wire, and the factor of 2 accounts for the two layers. The nominal turn density is n = 1 (turn) / (0.00254 m) = 394 m^{-1}, and this predicts

$$B_{\text{ext}} / i = 2 \cdot (4\pi \text{ x } 10^{-7} \text{ T m/A})\ \ (394 \text{ /m}) = 990\ \mu\text{T/A} \quad .$$

This computed value is subject to 'end corrections' in the actual solenoid, because it is not infinitely long. The windings of each two-layer solenoid fit between an inner diameter of 1.00" = 25.4 mm and an outer diameter of 1.14" = 29.0 mm, and they fit into a total length of 5.74" = 146 mm, and these dimensions will permit the calculation of end corrections (of about -2%).

The solenoid may be used in the (-2,+2)-A range, or briefly in the (-3, +3)-A range. The motivation for the double windings on the solenoid is to allow the separate but simultaneous production of dc and ac fields. The wires of the two electrically-separate solenoids are brought out to binding-post connections at the base of the solenoids' frame.

Modeling the output

Now if you use either one of the double-solenoid's windings, you have a way to create, along the sensitive axis of the fluxgate, a field which is the sum of the local earth's-field , plus the solenoid's field. As a function of this independent variable, you can record the dependent variable, which is the magnitude of the $2f$-component of the emf produced by the secondary. Because you are measuring the magnitude only, your results will all be positive. How should they depend on the field?

In a simple model, the $2f$-component would have an amplitude A which is a linear function of the field, with

$$A = S\,B_{\text{ext}} \quad .$$

Here S is the sensitivity of the device, with units of say μV/μT. Of course what you're measuring is the magnitude M of this spectral component, so your model would give

$$M = S\,|\,B_{\text{ext}}\,| \quad .$$

In practice, you'll see a result different from this, and the reason is that (even in the absence of any external magnetic field) the sensor is likely to be producing some signal at frequency $2f$ (due to harmonic distortion in either the source of the excitation, or the amplifier, or the magnetic properties of the sensor itself). So now there is a slightly complicated problem of 'phasor addition', since what you detect is (the magnitude of) the sum of this background $2f$ signal and the one produced due to B_{ext}. What you want is the magnitude of the sum of

$$A_{bgr}\cos(2\omega t + \phi_{bgr}) + S\;B_{ext}\cos(2\omega t + \phi_B) \quad ;$$

note that both terms are of frequency $2f$, but they occur with possibly different phases.

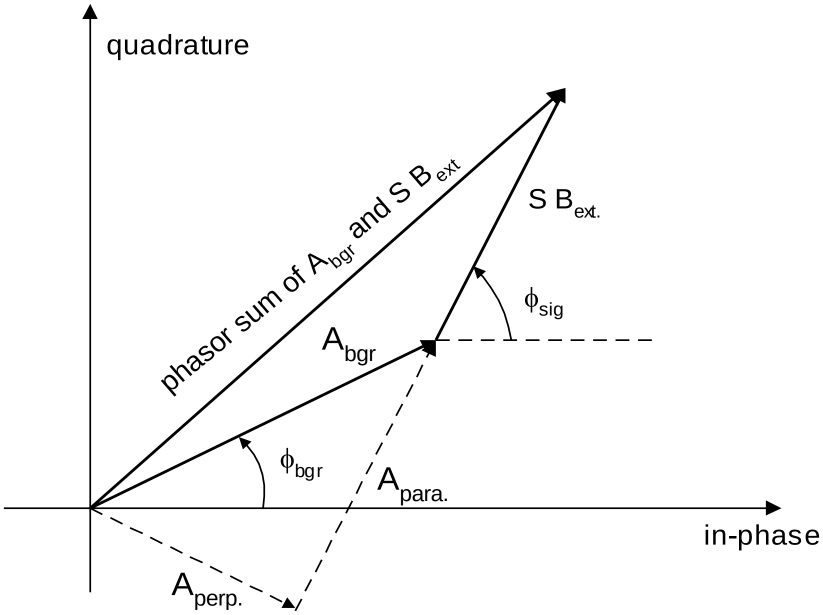


Fig. 13.7: A phasor model for the resultant of two components, each of frequency $2f$, but of independent amplitudes (A_{bgr} and $S{\cdot}B_{\text{ext}}$) and phase (ϕ_{bgr} and ϕ_{sig}).

The phasor diagram above shows how to understand the magnitude of their sum, which can be written as

$$M = \sqrt{A_{perp}{}^2 + (A_{para} + S\ B_{ext})^2}$$

where A_{perp} and A_{para} are the components of the phasor A_{bgr} perpendicular to, and parallel to, the phasor of the signal due to B_{ext}. This predicts a plot of M as a function of B_{ext} which is a hyperbola, having minimum value of A_{perp} , and two asymptotes with slopes $\pm S$.

In practice, this suggests that you make a quadratic fit to the *square* of the measured signal magnitude, since this can be written as

$$M^2 = S^2\ [(B_{ext} + \frac{A_{para}}{S})^2 + (\frac{A_{perp}}{S})^2] = S^2\ [(B_{ext} + a)^2 + b^2] \quad .$$

Of the three parameters returned by the fit, the sensitivity S is what you really care about, while a and b model the background emf that you wish were absent.

Because part of the background emf, namely a, masquerades as if it were genuine magnetic field B_{ext}, you need some way to account for this ‘zero offset’. Perhaps the best way is to install the fluxgate sensor into the opposite end of the solenoid, and repeat your calibration with no other changes. Relative to the sensor, you expect to have the same values of non-idealities a and b, but effectively B_{ext} (including both the earth’s field, and the solenoid field) will have been turned around. Or, if you plot your new data on the same axis for current, or for solenoid field, the values of the non-idealities a and b will have been reversed. That will offset two calibration curves by a horizontal difference of $2a$, and that offset will tell you how big a is. The two curves should have equal, non-zero, minima, and those minima will tell you b’s value.

Once you have a model like this, you have a calibrated fluxgate magnetometer. Now you can bring a permanent magnet into its vicinity (it’s easiest to place the magnet on-axis with the fluxgate, and with its magnetic moment also parallel to this axis), and see the effect that it has. You can flip that magnet end-for-end to reverse its contribution to the fluxgate signal. You can explore the validity of the expected $1/r^3$-law for the file due to a small dipole source. You can explore the limits of linearity of response of the fluxgate to field. Perhaps best of all, you can reflect on how you *understand* that all of this operation depends on the deliberate use of second-harmonic distortion.

Detecting *ac* magnetic fields with your fluxgate

You’ve seen that a fluxgate, excited by a frequency f, produces an output signal at frequency $2f$, with amplitude related to the (steady) magnetic field B_{ext}. In fact (apart from two ‘zero offsets’) the output signal can be written as

$$V_{out}(t) = S\ B_{ext} \cos(2\pi \cdot 2 f\ t) \quad .$$

Now what if the field B_{ext} is not steady, but is itself varying according to

$$B_{ext} = B_{ext}(t) = B_{dc} + B_{ac}\cos(2\pi \cdot F\,t) \quad ?$$

If you trust the model above, at least for ac-field frequency F not too large, you're led to expect

$$V_{out}(t) = S\,[B_{dc} + B_{ac}\cos(2\pi \cdot F\,t)]\cos(2\pi \cdot 2f\;t) \quad ,$$

and you should recognize this as an example of amplitude modulation (just as in Ch. 3). That, in turn, should lead you to think of the *spectrum* of $V_{\text{out}}(t)$, which ought to consist of a 'carrier' at frequency 2*f*, with two 'sidebands' at 2*f* ± *F*. In fact, given the model above,

$$V_{out}(t) = S\;B_{dc}\cos(2\pi \cdot 2f\;t) + S\;B_{ac}\,\frac{1}{2}\{\cos[2\pi(2f+F)t] + \cos[2\pi(2f+F)t]\} \quad ,$$

so the prediction for the spectrum is:

- the central peak's height depends only on B_{dc} (not on B_{ac} or *F*);
- the sidebands' *locations* depend only on (*f* and) *F*, the ac-field's frequency (but not on B_{dc} or B_{ac});
- the sidebands' *magnitudes* depend only on B_{ac}, the ac field's amplitude (and not on B_{dc} or *F*), exhibiting one-half the sensitivity you've measured for dc fields.

These are exciting (and very little-known) results!

You can check them immediately, if you've set up your fluxgate in its calibration mode. You've been looking at the magnitude of the spectral peak at 2*f*, as it responds to dc current in the solenoid. But now you can add *ac* current to the other winding of the (double) solenoid, and in response to it, you should immediately see new spectral sidebands at 2*f* ± *F*. Best of all, you'll see these peaks competing with only a noise floor, without any pollution by the A_{para} and A_{perp} non-idealities mentioned above. [A sinusoidal ac field is mathematically simplest, but you could think about a square-wave ac field as providing another kind of 'control experiment' – you can look for the fluxgate's response alternately with, and without, the extra current in the second solenoid. The value of Fourier analysis is that this response shows up at a *new location in frequency space*, in particular at a location where it is not competing with the non-idealities modeled in the zero offsets.]

You can trust the model above for the sensitivity of your system to ac fields, or you can calibrate it by an *a priori* method. If doing so, you'll need to measure the ac current in the (second) solenoid. If you do this with a multimeter, note that ac ammeters have a restricted range of frequency coverage, and note that by convention they tell you the rms measure of the current – the amplitude is of current waveform is √2-fold bigger.

Once you have your device calibrated for ac fields, it is fun to see how small an ac field you can detect. In pursuing this 'hero experiment', you want to optimize your 770 for sensitivity. So you might pick a frequency *F* which is of order 10 Hz, so sidebands do not suffer from the 'tails' of the carrier peak. You might try the various Window choices of the 770. You can check if you have the Input sensitivity set to as sensitive a scale as will not overload the 770's input circuitry. You can also try the Averaging mode of the 770. You should be easily able to detect ac fields of amplitude 1 μT – what timescale does it take to do so reliably? If you have more averaging time to spend, how much better can you do? If you change to the PSD = power spectral density choice of the MEASure menu, and you change your vertical-scale units to Vrms/√Hz, and use your conversion factor between μV of response and μT of field, you can convert the height of the noise floor into units of μT/√Hz or nT/√Hz. Now you're in position to read the research literature about high-sensitivity magnetometers, and understand the vocabulary of the state-of-the-art in weak-field detection.

You can also vary the frequency F of your ac field. At low values ($F \approx 1$ Hz), you'll need to narrow the 770's frequency span in order to resolve the sidebands from the carrier. At higher values ($F \approx 100$ Hz), it is not certain that the amplitude-modulation model above is applicable – but you're in position to find out if it is. If you prove that your device has sensitivity at $F = 51$ or 59 Hz, then you can look for sidebands, at frequencies $2f \pm 50$ Hz, or at $2f \pm 60$ Hz, which are present even when you put *no* ac current into your solenoid. What do you suppose might be causing them?

Finally, to appreciate yet another advantage of Fourier methods, try operating with the sensor in the solenoid, but subject only to the ambient and steady earth's field. Now you have *both* solenoid windings at your disposal, and you can arrange to excite them both, with two *separate* ac generators at two distinct frequencies F_1 and F_2. That should give you a carrier and *two* sets of sidebands on the 770's display. This tells you that your fluxgate is a real-time ac-field spectrometer – one which is simultaneously sensitive to the magnitudes of *all the frequencies* that might be present in the local ac magnetic field, with a uniform-in-frequency sensitivity that you have established. This combination thus represents the use of Fourier methods to give a novel capability in ac-field diagnostics.

Using a lock-in amplifier with your fluxgate

The fluxgate magnetometer encodes its indication of an external (steady, or dc) magnetic field via the magnitude (and phase) of a signal at frequency $2f$, where f is the frequency at which it's excited. That means its output could be 'decoded' by the use of a lock-in amplifier. This *optional* section describes how that might be done.

A lock-in amp requires a 'reference input', and depending on the lock-in you might have on hand, there are at least two ways to use it with a fluxgate:

a) you can send to the lock-in's reference input the frequency-f signal you're using to drive the fluxgate, and then set the lock-in to operate in its '$2f$ mode'; or

b) if your lock-in lacks a $2f$-mode, you can send the ≈1-kHz, ≈6-V signal you're using to drive the primary coil of the fluxgate to both inputs of the Multiplier module as well. The Multiplier will then act as a squarer, and its output will consist of a dc offset plus a frequency-$2f$ signal. If your lock-in's reference input is ac-coupled, it'll be getting just the $2f$-signal it needs.

Once you've 'locked' the lock-in to this reference signal, you can attach the fluxgate's secondary-coil output directly to the lock-in's signal input. You can choose suitable filtering at the input (to reject dc and low-frequency signals), and you can choose a full-scale sensitivity of order 1 mV. You can set the 'time constant' or averaging time of the lock-in to 0.1 s (or more). The least intuitive choice comes in setting the *phase* of the lock-in (which is, after all, a phase-sensitive amplifier). Here are at least two options:

a) If you have a 2-channel lock-in, then you can see, simultaneously displayed on two outputs, the in-phase (I) and quadrature-phase (Q) signals at frequency 2*f*. That is, relative to the phase supplied and defined by your reference input, you can see indications of A_I and A_Q in a measurement of the input signal,

$$V_{in}(t) = (\sqrt{2}\ A_I)\cos(2\pi \cdot 2f\ t) + (\sqrt{2}\ A_Q)\sin(2\pi \cdot 2f\ t) \quad .$$

The values of A_I and A_Q will have signs as well as magnitudes (conventionally expressed in rms measure), and together they locate the tip of the phasor of Fig. 13.7. If you obtain from the lock-in the two dc output voltages proportional to the I and Q outputs of the lock-in, and send them to the X- and Y-inputs of a 'scope (set for XY-display), then you'll have a 'vector voltmeter' which displays that phasor-tip in real time. It is now *very* instructive to see this XY-spot move as you vary the B_{ext} field (say, by changing the current in the calibration solenoid). It will trace out a linear path in I-Q space on the XY-display. And, if you change the phase setting on the lock-in, you will *rotate* this linear path in XY-space. Choose the phase setting at which this linear spot-motion is all in one channel of your two-channel lock-in. (The other channel's output will *not* be zero, but will have been made to be a constant.) Now one channel will be registering a signal *in*dependent of B_{ext} akin to A_{perp}, while the other channel's output will be a measure of $A_{para} + S\ B_{ext}$, ie. the desired output directly related to the magnitude *and sign* of B_{ext}. (So finally you'll be able to distinguish positive and negative field values.)

b) If you have a single-channel lock-in, here's a procedure to follow. You'll want the fluxgate mounted in its calibration solenoid, and you'll want to be able to vary the solenoid current smoothly in the 0-to-1-A range. Now conduct that current variation, up and down, while watching the lock-in output.

For some choices of phase, the lock-in output will *rise* with current; for other settings, it will *fall* with current. There will be two phase settings (in the 0-to-360° range) at which the lock-in output will be (non-zero but) *independent of* the coil current. Find (either one) of these. Now, finally, fix the phase setting to be ±90° away from that value. At this phase setting, the lock-in output will again be a measure of $A_{para} + S\ B_{ext}$.

Now the use of the solenoid and its known coil-constant will allow you to calibrate the sensitivity *S* of the fluxgate. To find the offset A_{para}, mount the fluxgate sensor in the other end of the solenoid and repeat the calibration; you should get the same A_{para} value, but the effect of B_{ext} (both earth's and solenoid's contributions) will have been reversed.

Of course you can now test the system for linearity of response to field B_{ext} (does that extend to the ±2 mT or ±3 mT range you can reach with the solenoid?) You can also test the system for sensitivity to field (does a ±1-μT change in field show up? against what noise level? with what choice of time constant?)

Chapter 14: Frequency-domain views of audio waveforms

This Chapter requires only Chapters 1 and 2 as preparation, and is very open-ended in character. It first addresses the practical questions of getting audio waveforms out of the consumer-appliance world and into the laboratory-measurement world. Then it takes up some open-ended ideas for investigations which you can perform, once you have the capability of seeing the Fourier spectra of audio waveforms.

Techniques for interconnecting the audio and laboratory worlds

Among the Electronic Modules, you'll see at upper left a module called Audio Connections, whose purpose is to bring audio signals from various sources into the 'BNC world' where you can process them with other modules, and with your SR770. It is divided into two sections, which are separately covered below

a) Microphone signals

Your auxiliary equipment includes two microphones, one built into the Acoustic Resonator, and the other (with a lapel pin) free to use in the open air. These are alike in their construction, but they differ in their cable connections. Neither can be used 'by itself', but either can be used via the upper-half section of the Audio Connections module.

Here's a diagram showing how to get audio signals converted by microphone into BNC-accessible electronic signals:

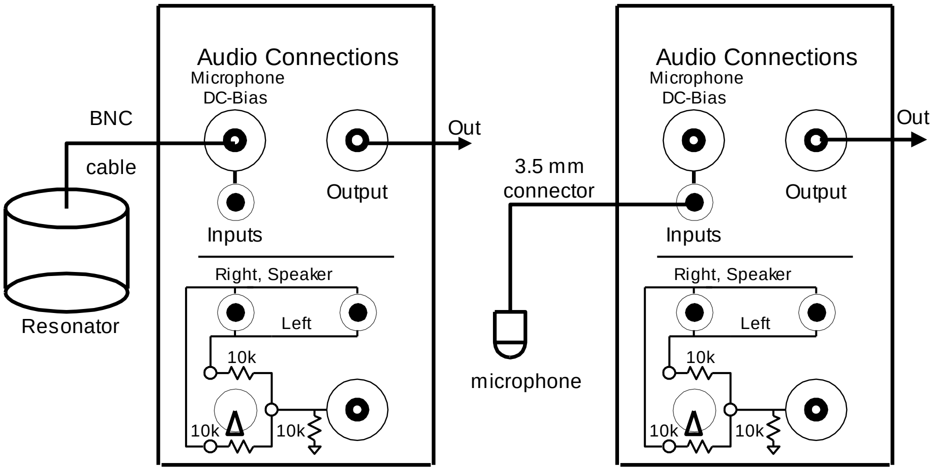


Figure 14.1: (left) Using the microphone in the Acoustic Resonator; (right) using the free-standing microphone

You will appreciate that the Audio-Resonator microphone is attached via a BNC cable, while the free-standing microphone uses the consumer-market audio standard of a (mono) 3.5-mm audio jack. The microphone connections marked in your Modules accommodate either of these.

This Module also provides the electrical ‘bias’ which both microphones require. Unlike some textbook microphones, which are simple sources of induced current or emf (in response to air-pressure fluctuations), both of your microphones are *active* devices which require power to be supplied to them. The conversion of acoustic signals to electronic ones proceeds in two stages. First, the air-pressure fluctuations create a waveform in electric charge via a piezoelectric transducer. But rather than send those (very small) charge signals into the large (and unpredictable) input capacitance of a cable plus electronics at its far end, each microphone contains, within its small volume, an active electronic circuit acting as a charge-to-voltage converter. This device, like an op-amp, requires dc electrical power for its operation. Ingeniously, that power is supplied *to* the microphone by the same wire which also conducts the audio electrical signal *back* from the microphone. So the cable to either microphone contains a common, or ground wire, and a single active wire. The Module provides a dc ‘bias’ or positive potential of 5 Volts (to power the q-to-V converter), and dc-couples that to the active wire. But there is also an ac-coupled signal pick-off from the same wire, which makes available to you, at the BNC connector labeled Output, the voltage waveform which represents the audio pressure-fluctuation waveform.

Remember, **neither microphone will work at all** if you connect it directly to an ordinary passive input, such as that of a ‘scope or the 770. Your monitoring device would be ‘looking at’ the output of the <u>un</u>powered amplifier built into the microphone, and that output will be inactive in the absence of that +5-Volt bias.

[Microphone inputs in many kinds of consumer electronics, including suitably configured sound-card inputs in computers, may provide the needed +5-V bias for microphones such as these. If you’re not sure whether this bias is present, you can monitor a microphone *input* with just a voltmeter, to see if its center connection is maintained at a +5-V potential relative to its outer shell.]

b) Speaker signals

The lower section of the Audio Connections module has two 3.5-mm plugs and a single BNC connector, and it can be used in two ways: either to send waveforms from the BNC world to speakers with 3.5-mm jacks, or to get (mono or stereo) signals born in the 3.5-mm world into (mono) signals in the BNC world. That is to say, this part of the Module can ‘work either way’.

Here’s a first example: if you have a signal source (such as an external generator, or the Source-Out function of the 770) with a BNC output, and want to drive a speaker (such as the speaker in the Acoustic Resonator) which is equipped with a 3.5-mm plug. In this case, connect the cable from the generator to the BNC connector, connect the speaker’s 3.5-mm plug into the jack marked Right/Speaker, and set the toggle switch to the <u>down</u> position.

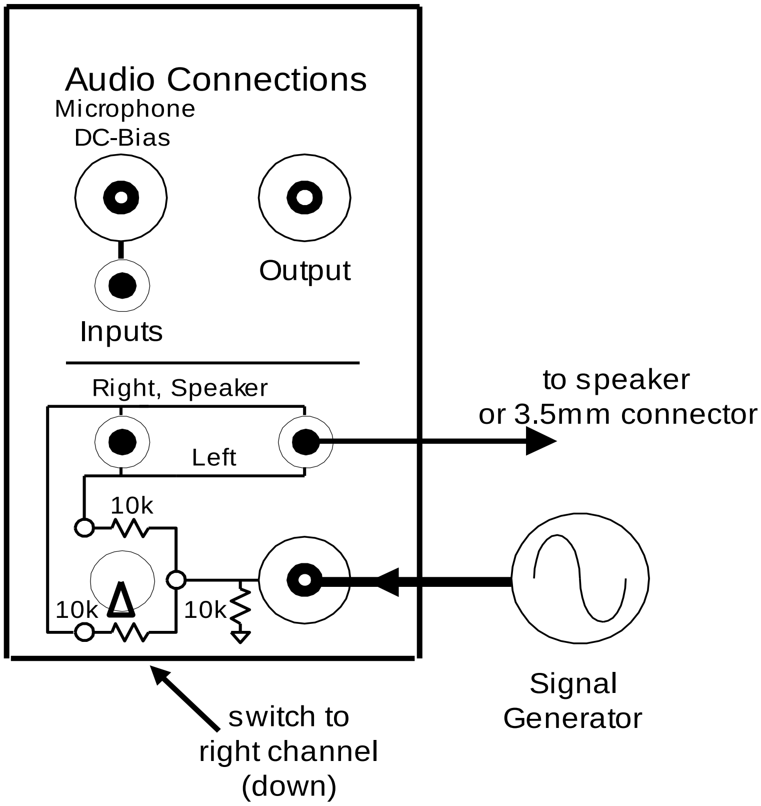


Fig. 14.2: Using an external generator to drive a 3.5-mm-equipped speaker; note that the switch is set to short out a series 10-kΩ resistor.

Below is a second example: if you have a signal source (such as a personal music player) which sends its output via a cable whose far end is a 3.5-mm plug, you can get that signal to emerge from the BNC connector on the panel. In fact, if that 3.5-mm plug conveys a stereo signal, you can get either its right-channel, or left-channel, or sum-signal information to appear at the BNC connector, just by setting the toggle switch to one of its three positions. Note also that the signal being monitored in the BNC world can simultaneously be sent along to another device in the 3.5-mm world.

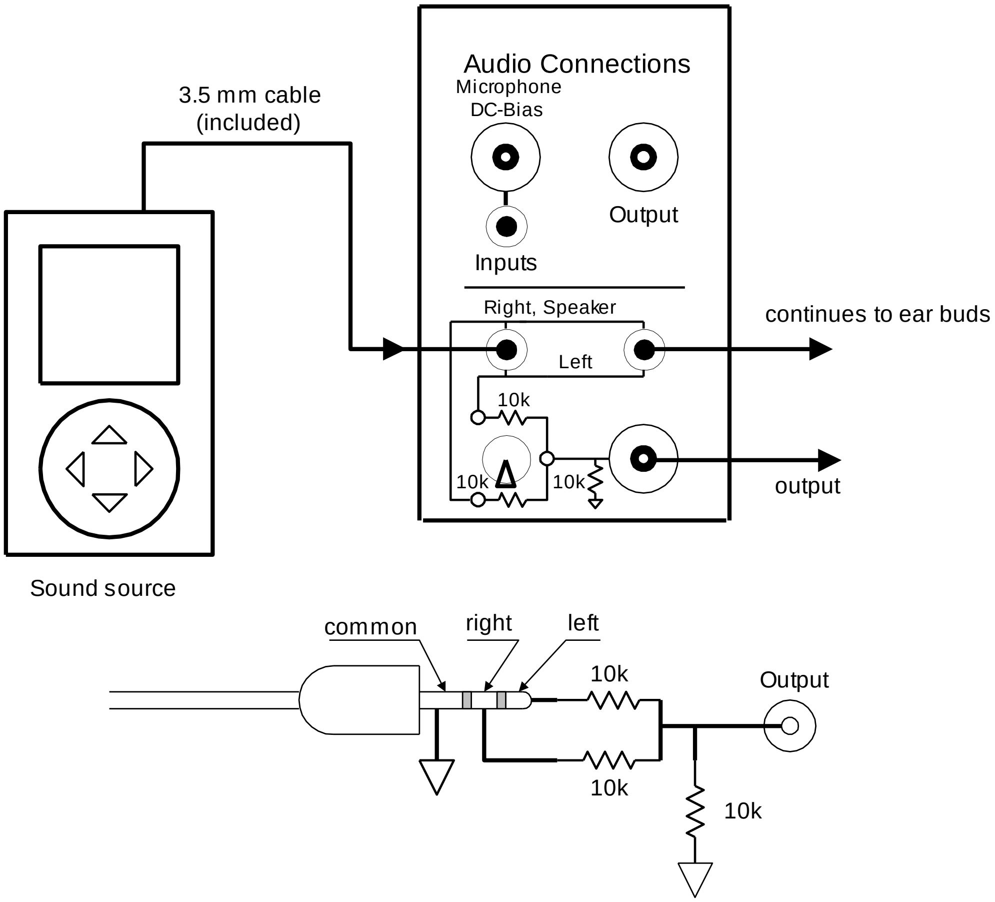


Figure 14.3: Using a signal source with 3.5-mm output to drive a BNC connector; note that in the central position, the switch uses a series resistor of 10-kΩ in each channel, and allows a summed-current to flow in the 10-kΩ resistor to ground.

The goal of these ‘practical means’ is to afford you convenient ways to interconnect signals between the 3.5-mm and the BNC worlds, without having to resort to alligator clips or other kludges.

Signals of audio origin, and their Fourier analysis

Supposing you have surmounted these wiring-technique obstacles, you now have multiple sources of audio electronic signals, and can bring them by BNC cabling to any other part among your Modules, or direct to the 770 for Fourier analysis. The question now is what to do with such signals.

If the signals correspond to audible sounds, then by convention they lie in the 20 Hz - 20 kHz band. The microphones supplied are claimed to work, uniformly to ±3 dB, across the full audio spectrum. The lower limit of that spectrum enables you to use ac-coupling at the input of the 770, while the upper limit means you could use the 770 on its 0-25 kHz span without any loss. If you use the Continuous mode of triggering the 770, you will get a freshly updated spectrum every 16 ms, so you’ll have a real-time display of the waveform’s spectral content.

The first audio signals you might want to study ought to be steady signals, for convenience. You can get such signals from singing, from whistling, or from playing certain musical instruments. Once you have a steady signal, you should see a steady display of its frequency-domain content on the 770, and a BNC splitter will allow you to see the time-domain view simultaneously on an oscilloscope.

If your signal is not only steady, but also periodic, then it will have a 'line spectrum' of a fundamental and its harmonics. If you use the appropriate choice of Frequency Span to get a good view of the fundamental, then you can use the Marker to give a numerical value for f_1, the fundamental frequency. From that fundamental frequency value, you can correctly compute the period of the signal, from $T = 1 / f_1$. That, in turn, will enable you to set the 'scope display so as to display one or two full cycles of the waveform.

Once you know the fundamental frequency, you can identify what harmonics are present, and with what magnitudes relative to the fundamental. Sounds differ a great deal in their composition, and this Fourier view is the key to understanding the musical concept of timbre. In particular, this view can answer a question which might have troubled you from your first introduction to sound: how is it that two waveforms of the same frequency, and the same power, can nevertheless be so easily distinguished?

[For example, an oboe will sound a 440-Hz note to tune up an orchestra, and a flute can and will be tuned to produce the same frequency and the same volume – but no-one could possibly fail to distinguish their sounds. For another example, a singer can project and maintain a vowel sound indefinitely (unlike consonantal sounds, which are characterized by their transients), and a singer might sing the vowel sound 'aaah' at frequency 880 Hz, and then, without change in frequency or volume, change to the vowel sound 'eeee'. But every human listener will notice the change, even if a secondary-school model for sound fails to account for it.]

The answer lies not in the power in the waveform nor frequency of the fundamental, but in the pattern of harmonic content, of the sounds involved. Depending on your interests in audio sounds, here are some projects you could undertake.

a) Musical timbre

If you have access to musical instruments, you can practice playing them at a given pitch, and then practice acquiring views of their waveforms, in the time and frequency domains, using a 'scope and the 770. Note that a 'scope may have a run/stop button to push (when you have reached the steady state), and the 770 has the same 'freeze' function available with its Pause button. You will find musical instruments whose output is concentrated in the fundamental, and others which are dominated by certain harmonics. The distribution of power, ie. how much is in what harmonic, may also change as the instrument is used to generate different musical notes of the scale.

If you have a measurement of the distribution, among harmonics, of the power in a sound, then you could try to synthesize two musical sounds, of the same frequency and total power, but which are nevertheless distinguishable in timbre from each other. Such syntheses are nowadays rather easy to execute in software, and might be ‘played’ directly via the sound-card output of a computer. One simplification in the modeling is that (to a good approximation) your ear is insensitive to the phases assigned to each Fourier component, so that the Fourier magnitude spectrum that’s easy to measure on the 770 gives all the information needed for this sort of synthesis.

If you try this sort of project, it is worth configuring your sound synthesizer so it can be switched at will between an ‘A source’ and a ‘B source’, perhaps from oboe-like to flute-like sound. A simple synthesis model will deceive no musician, since it will wholly lack all the other cues and content in real musical sounds. You model will lack the initial transient (the ‘attack’) and the modulation (‘vibrato’) present in actual musical sounds. But at least an A/B comparison will enable you to understand what makes two distinct syntheses distinguishable by ear.

b) Vocal formants

In this exercise, you need access that most ‘human’ of all instruments, a trained singer’s voice. Ideally, you’d like a singer who can maintain a long singing of one musical sound, at rather steady level of volume, while gliding in pitch over a wide range (such as an octave). And then you’d like that exercise repeated for several distinct choices of pure vowel sounds (that is, neither diphthongs nor consonants), such as those represented phonetically in English by ä (as in f**a**ther), ē (as is b**ea**t or **ea**sy), ü (as in b**oo**t), e (as in s**e**t or r**e**d), and ö (as in b**o**ne). (Non-singers will need some practice to produce these vowel sounds at indefinite length, and also at variable pitch.)

Now each such vowel sound (and there are others, too – think of the o-sound in s**aw**), sung at a given pitch or choice of fundamental frequency, will show up in a Fourier spectrum as a series of peaks. It is the *envelope function* in which this series of peaks lie which characterizes human vowels, and to see that envelope, you want to watch a live Fourier spectrum while a singer is changing the pitch being sung. Clearly, during the course of a 10% or 50% reduction of the pitch that’s sung, every peak will move downward in frequency. What you want to look for is the fixed-in-frequency envelope in which the peak-tops move. The peaks in this envelope function are called formants, and the location-in-frequency of the formants is what allows human listeners to distinguish distinct vowel sounds.

For example, the ü (as in boot) sound is characterized by having its main formants at 320 and 800 Hz. Those location are fixed, ie. absolute in frequency, and are thus not going to move, even as the singer changes his fundamental frequency (say from 200 Hz to 100 Hz). [‘His’ is the right choice of word here – a male singer is likely to do better at producing these low frequencies.] Similarly, the ä (as in father) sound is characterized by having its lowest-frequency formants at 700 and 1150 Hz.

It follows that a ü-sound sung at 150 Hz will have extra strength in its 2nd and 5th harmonics, while the ä-sound sung at the *same frequency* will have extra strength in its 5th and 8th harmonics.

It also follows that a note sung at a high frequency may *fail* to have spectral peaks anywhere in the vicinity of particular formant frequencies, and this will render certain high-pitched vowel sounds difficult to distinguish. [Ask a trained soprano singer about this problem.]

Finally, you might wonder how it is that humans can produce a fundamental frequency, and a set of formant frequencies, which can be independently controlled. The simplest explanation is that fundamental frequency is controlled by tension in the vocal cords, whereas harmonic content is controlled via resonances in the vocal tract, eg. the throat and mouth. It will be a measure of your understanding of Fourier methods if you can say why the periodic release of air between stretched vocal cords leads to sound with harmonic content in the first place. But given such as-produced harmonic content, you can imagine modeling the rest of the vocal tract by a transfer function, whose maxima-in-frequency can be varied, and which define the formant desired.

Chapter 15: Signal Recovery for signals-under-noise

From a frequency-domain point of view, you now know that a sinusoid is the simplest possible signal, in the sense that it delivers all its energy at one point in frequency space. By contrast, noise has its energy spread out over a whole range of frequencies, a continuum in frequency space. This Chapter addresses the question of detecting a sinusoidal signal when it is affected by, immersed in, or even buried under broad-band noise. This problem comes up frequently in research, and in the real world – consider trying to detect optical pulsations in the light received from a neutron star, or needing to detect the radio signal-beacon broadcast by a distant spacecraft.

It's important at this stage to distinguish between two real, but quite different, experimental cases. There are cases in which the experimenter has exact knowledge of the frequency, and even the phase, at which to expect the sinusoid. For example, the detected *radio* emission from a neutron star can give a real-time and essentially exact value of the frequency at which any pulsations in visible light ought to be present. For a more common example, the signal emerging from a table-top experiment will have an exactly known frequency, *if* that experiment is itself being driven, stimulated, or excited by an external source of known frequency. When you have access to a 'reference signal' whose frequency is an exact predictor of the frequency of the weak, noise-ridden signal you're trying to detect, then you can apply signal-recovery methods equivalent to lock-in detection.

The other and opposite case is 'blind detection', in which you lack access to any 'reference signal', and have only *one* wire or experimental channel coming from the experiment, which conveys to you the noise-infested signal. In such cases, not only the amplitude of the hypothetical signal, but also its frequency, are unknown numbers. In such cases, lock-in methods are not applicable, and Fourier methods can be used instead.

As a concrete introduction to this sort of search and discovery process, we have included among your Electronic Modules one called 'Buried Treasure', which delivers broad-band noise, and can also have signals immersed under it, or added to it. This module has only one output, and one selector switch, with the following functions:

- setting 'Noise' delivers broadband noise, from dc to about 2 MHz;
- setting 'Filtered Noise' delivers that noise, but filtered to the dc-to-0.2 MHz band;
- settings A, B, C, D each delivers that filtered noise, but in each case, a sinusoid of a particular frequency and amplitude has been added to it.

[That frequency *differs* among the four settings; the amplitude differs too, and it drops, with the signal strongest in A, weakest in D, position**.** The four frequency values you'll get are repeatable (and the A-choice does put the frequency into the audible range), but the four frequency values will change if you (or your instructor) change a piggy-back programming board on the printed-circuit board behind the front panel of the instrument.]

Now how does one look for a monochromatic signal among broadband noise? It's a 'needle in a haystack' problem, and success depends on the needle being sharp, while the haystack is broad. Or less metaphorically, the frequency spectrum of the noise is broad, even flat for white noise, while the spectrum of the signal is sharp, in principle a delta-function in frequency. So you'd like a view of the spectrum, and you're looking for a sharp signal peak to be seen standing up above a flat and uniform noise floor.

It would be a good introduction for you to 'spike' the experiment with a signal under your control. Do this by conveying your Filtered-Noise output to the Summer module, and there add to it a sinusoidal signal, perhaps of 20-mV amplitude, and with a known frequency in the 2-4 kHz range, derived from an external signal generator. If you look at the Summer's output on a 'scope, you'll see a mass of positive- and negative-going pulses of noise, with lots of occurrences of values more than 200 mV away from zero. So the signal, of amplitude only 20 mV, is well-buried in under the noise.

The first hint that such a signal is 'recoverable' can (in this example) come by *ear*. Send the Summer's output to the Power Audio Amplifier module, and set its gain knob to about 2 on the 0-11 scale, and connect its output to the Speaker connection. Now dial up the amplifier gain until you hear a hiss at the speaker (that's the sound of white noise), and listen for a high-pitched *tone* among the white noise. Your hearing is *amazingly adept* at picking out such a tone from noise, and the more so if you change the frequency of the tone, say among the values 2, 3, and 4 kHz – in this demonstration, the frequency is after all under your direct and immediate control. (You could now temporarily try position-A for the switch, instead of the external generator, to listen for a signal whose frequency you *don't* know or control.)

To *see* the signal (instead of hearing it), send the Summer's output to the 770 instead, and configure that for Full Span (0-100 kHz) coverage. Use AutoRange to adapt its input sensitivity to the noise level that's arriving (it'll select a sensitivity such that even rather infrequent noise peaks stay on-scale it its internal electronics), and use the MEASure button, and the Measure softkey, to select PSD or power spectral density. Use Return to get back to the menu, and use softkeys to select Log Magnitude under Display, and Volts RMS under Units. Press the AVERAGE button, but for now, turn averaging Off. You should get a display showing the 0-100 kHz spectrum of the input signal, and you will likely not see any spectral peak leap out at you. Try changing your generator's frequency setting from 2 or 4 to about 40 kHz, and the spectrum will still show a host of fluctuations and no obvious signal peak. Now finally use the Average button to select averaging On, and set 16 averages, and use RMS-type averaging, and Exponential rather than Linear mode of averaging. Suddenly the signal peak will be *easy* to see, even if you change the signal's frequency back to 3 kHz. What sets the borderline between invisible and easy-to-see?

There are several issues involved in the blind detection of a weak signal in the presence of noise. There is first of all the amplitude *A* of the sinusoidal signal in question – you'd like *A* to be big. There is next the power spectral density *S* of the noise that's competing with the signal, and you'd like *S* to be small. But you also have, as an independent variable, the number *n* of spectra you take and average together; the larger *n* is, the better for detection (but larger *n* entails a longer waiting time). Finally, there is another variable: δf, the width-in-frequency of each spectral bin of the Fourier spectrum, which for the 770 is always 1/400 of the frequency span you have chosen. If you pick a smaller span, then δf gets smaller too, and you'll see that this helps in blind detection.

For example, suppose that by averaging you can easily see a signal at (say) 43 kHz. Now if you turn the averaging off, the signal will get buried among the fluctuations in the noise floor. But try picking a span of 6.25 kHz (1/16th of what you've been using), taking care to center this new span around 43 kHz. With the averaging still off, you can easily see the signal again. There's no magic here – you might notice that for a Span of 100 kHz, the 770 needs an acquisition time of 4 ms, so averaging 16 such spectra gives you access to 64 ms of data. By contrast, when you lower the span to 6.25 kHz, the acquisition time becomes 64 ms, so one *un*-averaged acquisition at this span setting gives you as long an averaging time as 16 acquisitions at the original full span. So a narrow span is good for detectability – but if the signal's location in frequency space is only known to be somewhere in the 0-100 kHz range, the use of a span of 6.25 kHz will cover only 1/16th of that range, and so 15 clones of such a search will be needed to assure full coverage of the spectrum.

This teaches you another lesson, just as applicable as in the needle-in-haystack problem: if you know about where to look, the search time needed will drop, and in direct proportion to *how much* you know about where to look.

Once you've learned to re-discover, by Fourier methods, a signal that you know that you've put into your system, it's time to think about finding an 'unknown' signal. But rather than guesswork searching, it's possible to work out quantitatively just how hard a problem this can be. Everything depends on computing the mean-square voltage that's sorted into each frequency bin of width δf by the process of Fourier analysis. In any bin that gets noise-only, the expected value is

$$<V_n^2(t)> = \int S(f)\,\mathrm{d}f \rightarrow S\,\delta f \ ,$$

whereas that single lucky bin into which the signal (as well as some noise) falls will give the expected result

$$< [V_n(t) + A\cos(\omega\, t - \phi)\,]^2 > = S\,\delta f + A^2/2 \ .$$

So you can imagine 399 bins each showing result $S\,\delta f$ for mean-square voltage, and one bin, as an outlier, showing instead the result $S\,\delta f + A^2/2$. But in the presence of random noise, even the signal-free noise-only bins will show statistical fluctuations in their contents, from bin to adjacent bin, and from acquisition to subsequent acquisition. So what you really want is the excess power in a bin (due to the signal) to be large compared to the *standard deviation* of the values in the 'noise floor'. For noise, that standard deviation (for $n = 1$, one acquisition) turns out to be about equal in size to the mean, and for n acquisitions, it drops as $n^{-1/2}$. So what we need, for an 'N standard deviation' detection, is

$$A^2/2 \ > N\,(\,S\,\delta f / \sqrt{n}\,) \ .$$

Now the 770 as configured above is ready to read *S*, though indirectly. Whether or not you are seeing a signal peak in the spectrum, you are seeing a noise floor. If you raise the number of averages *n* adequately, the noise floor's value, at the marker's location, will settle to a number which you can read at the top of the display. Using the Buried Treasure noise source, you might see a result for 'voltage noise density' coming out as about 200 μVrms/$\sqrt{}$Hz, or 2×10^{-4} V/$\sqrt{}$Hz. The <u>square</u> of that number, which is 4×10^{-8} V^2/Hz, gives the noise power density *S*, as mean-square voltage per unit frequency interval So now it's possible to see what is the predicted threshold for a $N = 5$-sigma detection: we need

$$A^2/2 \ > 5\ (4 \times 10^{-8}\ V^2/Hz \cdot 250\ Hz\ /\ \sqrt{1}\ )\ \text{ or } A > 10 \times 10^{-3}\ V = 10\ mV.$$

This computation assumes full frequency span (which is what fixes δf to 250 Hz, and no use of averaging (which is what sets *n* to 1).

In practice this calculation is inexact, in part because the statistics of bin-contents are not Gaussian, and in part because you're looking for a direct eyeball spotting of the 'lucky bin', rather than finding it after-the-fact as an outlier in a histogram of bin-content values. But this does show why it takes *time* to be sure of the detection of a weak signal. If *A* drops in size, then to maintain this inequality, you need either to reduce the bin-width δf (by lowering the frequency span, which raises each acquisition time) or raise the number of averages *n* (which raises the time needed to complete the averaging).

With this guidance, and with past success in detecting signal A buried under noise, try in succession to find the signals B, C, and D. By design, the tasks will grow harder along this list. A mark of success is to be sure you have detected a spectral-line signal, and to have established its frequency and its amplitude.

In real life, you finally become convinced of a positive detection of a signal by means of *reproducibility*. That is to say,

- you look for a spectral peak to stay steadily above the fluctuations in the noise floor;
- you look for it to come and go, as you switch access to it on and off (using the A/B/C/D switch to go onto, and off of, the signal);
- you look for the peak to reappear in a different bin (though at the same location in frequency) when you make a small change in the Start Frequency of the frequency span.

The difficult decisions come when the amplitude *A* is really weak, so that after you've expended all the observation time you've been allotted, you are *still* not sure if you have a genuine outlier, or just a fluctuation. You'd have to do a careful and detailed statistical argument to decide what is the level of possible amplitude *A* that you have definitively excluded from being present in the frequency range you've explored.

Supposing the contrary, that you have become convinced that you do have a genuine detection of a signal. Here's a way to use the 770 to measure its amplitude. You'd like to use the MEASure button to pick, for Window, the Flattop choice, so that a measured amplitude will not be affected by whether the signal falls at the center, or near the edges, of a bin. You'd like to use either a narrow enough frequency span, or a large enough number of averages, so that the spectral peak (the 'needle') stands up by at least 20 dB above the noise floor (the 'haystack'). Then the bin with the signal will have contents, in the mean-square sense, which is dominated (>99%) by signal power, rather than noise power. Now if you go back to measuring Spectrum (rather than PSD), and choose units of Volts Pk (rather than Volts RMS), you can put the marker at the spectral peak, and read off directly the amplitude you're after. [If you can't get the peak to stand up this far, you can compute mean-square power per bin, and subtract the noise-only value from the bin-contents with signal-plus-noise value, giving a corrected value for signal-alone.]

Notice that you need to set the 770 to two distinct choices of MEASure menu to measure the noise, as opposed to measuring the amplitude of the signal. The signal is a sinusoid, and it has an amplitude (in Volts), and you want to measure the Fourier spectrum to find it. But the noise is a continuum in frequency space, and it does not *have* an amplitude; instead, it has a voltage spectral density (in V/√Hz) or a power spectral density (in V^2/Hz). That is to say, even though you can see the signal peak and the noise floor simultaneously on one and the same display, the two quantities are incommensurate in their units, and so it requires two distinct kinds of measurements to find the values of these two quantities individually. That's the reason that 'zooming-in in frequency' by reducing the frequency span can give you a signal peak which stands up, more and more, above the noise floor – you are *not* just magnifying the horizontal scale of an underlying single graph.

In the process of zeroing-in, on the frequency scale, to the spectral peak of whose reality you have become convinced, you can also get a good estimate for the frequency of the signal, and for the uncertainty of that estimate. You might change the Window choice to Uniform for optimal frequency resolution, and you will still need to know the bin-width δf to estimate your uncertainty. All you can know from the display you are getting is that the signal falls (somewhere) within the bin of width δf where you've see it.

In principle, there is no limit (except your patience) to the weakness of a signal you can recover by this method. That's because the search time needed to find and confirm the presence of a signal of amplitude A varies approximately as $1/A^2$, so searches for two-fold weaker signals will take about four times as long. But another problem in the world of actual signals comes with the assumption that the signal is truly monochromatic. If you could tolerate an acquisition time of 1000 s, then each frequency bin of your spectrum would be 1 mHz (milliHertz) wide, and your method of detection assumes that the signal stays in *one bin*, of 1-mHz width, during the whole of your acquisition time. That is to say, the signal would need to stay stable in frequency to 1 mHz or better. If the signal frequency is of order 50 kHz, this would require frequency stability of one part in fifty million(!). Some sources of signal *are* that stable in frequency, but some are not. If the signal frequency varies *unpredictably*, then there is a rather sharp lower limit to how weak the signal can be, and still be detected by this method.

Chapter 16: Coupled Oscillators

Previous chapters have had you learn Fourier methods, but now you'll apply that learning, using Fourier methods to study a system of enormous generality: coupled oscillators. As it happens, the Coupled-Oscillator system included in Fourier Methods is mechanical, and the oscillators are coupled to each other, and the outside world, via magnetism. But the general concept of coupled oscillators applies to *any* system which has interacting 'normal modes', whether it be mechanical, electronic, electromagnetic, acoustical, or even quantum-mechanical in character. You'll use Fourier methods to learn concepts and terminology which are *portable* across all these domains.

First, what is a 'normal mode'? For a very simple mechanical oscillator, like a spring-and-mass system, there's only one coordinate, and its natural motion is sinusoidal, and at a particular frequency. As soon as a system has more than one moving part, it needs multiple coordinates to describe it, and these coordinates can undergo motions much more complicated than a single sinusoid. But there are special kinds of system-wide motion in which each coordinate *does* oscillate as a pure sinusoid, and for which one single frequency prevails across the whole system. Such a pattern in space is called a normal mode of the system.

Clearly a strung and tuned violin is a multi-particle system, and very complicated motions of its parts can be imagined. But if one string is carefully excited in its fundamental mode, then each part of that string, and of the bridge, and also of the body of the violin will oscillate sinusoidally, all at one common frequency. And there are certainly other normal modes of the violin system; three of them, with three new frequencies, have spatial patterns of oscillation whose motions are concentrated in the other three tuned strings.

For this Coupled Oscillator module of Fourier Methods, we've provided a simpler mechanical system than a violin, chiefly characterized by only two coordinates, and therefore possessing two normal modes of interest. But as is the case in general, you'll learn that a given normal mode involves motions of all the parts of a system.

Torsional-Reed Oscillators

The mechanical system which is the basic oscillator in this device is torsional in character, and is based on the elastic twisting of a thin ribbon, or reed, made of phosphor-bronze. One end of the reed is clamped, and the other end is free, but has attached to it a mass providing rotational inertia. As isometric view of such a system in Fig. 16.1 illustrates the coordinate system used in this Chapter. The reed is clamped back at $x = -L$, but near $x = 0$ the reed, and the mass it bears, is free to move. There's a 'wiggling mode', in which the end of the reed moves sideways, with the mass moving mostly in the $\pm y$-direction. But while that motion is easily visible, it occurs at a rather low natural frequency (of order 6 Hz), and it is not the mode of interest to your studies.

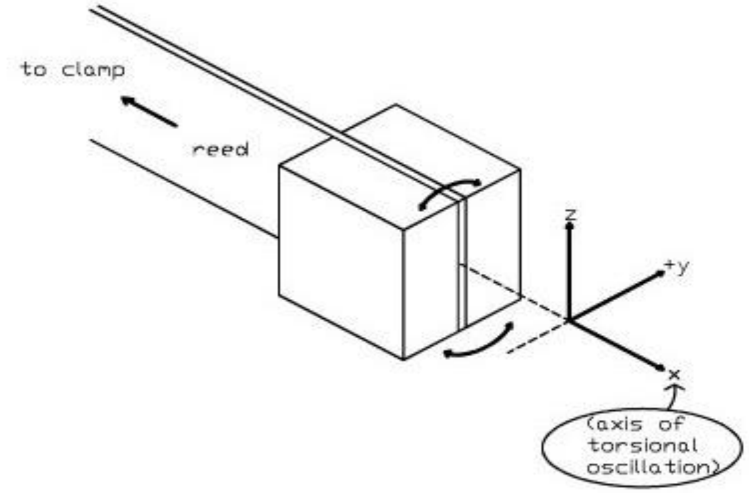


Fig. 16.1: An isometric view of (one) torsional-reed oscillator, and an xyz-coordinate system to describe it.

Instead, you'll study a different mode, in which the center-of-the-mass stays fixed in space, and the free end of the reed *twists* instead of wiggling. In such a motion, the 'central fiber' of the reed doesn't move at all, and each part of the mass moves with x = constant, but with y and z changing due to rotation about the x-axis. This mode is described by an angle of rotation (of the mass) about the x-axis, which we will call θ. This mode has a natural frequency near 150 Hz, and at the low amplitudes you'll be using, its motion is nearly invisible to the eye.

[Since the mass undergoes rotation about the x-axis in this mode, it certainly has a rotational inertia. If we model the mass as a cube of side s (and s = 1/4 inch = 6.35 mm for your mass), and of density ρ (and ρ = 7.51 g/cm^3 for your material), then it has mass $m = \rho\, s^3 = 1.92$ g in your oscillator. The rotational inertia of this object, for rotation about its center of mass and around the x-axis, is given by $I = (1/6)\, \rho\, s^5 = (1/6)\, m\, s^2$, with value estimated at $I \approx 0.129$ g cm^2 = 1.29 x 10^{-8} kg m^2.]

[That gives the 'inertia term' of this oscillator; the 'spring-constant' part of it comes from the torsional elasticity of the phosphor-bronze strip. We describe this by a torsional constant κ, which gives the torque arising per unit (radian-measure) rotation of the reed's free end. The torsional reed has width, or height-in-z, of w = 1/4 inch = 6.35 mm, and thickness-in-y of t = 0.010 inch = 0.254 mm. Its length-in-x is $L \approx 4$ inches or 100 mm. Elasticity theory then predicts a torsion constant of approximately

$$\kappa = G\frac{wt^3}{3L} \quad ,$$

where G is the shear modulus of the material, about 41 GPa = 41 x 10^9 N/m^2 for phosphor-bronze. This predicts κ ≈ 0.014 N·m, ie. the reed will act back with a torque of 0.014 N m per radian of twist applied to its end.]

The I- and κ-values computed above are only estimates based on material properties, and neither is easy to measure individually. But torsional oscillations provide a good check on a combination, since the (angular) frequency of torsional oscillations is predicted to be

$$\omega = \sqrt{\kappa / I} \quad .$$

The estimates above give the prediction $\omega = (1.4 \text{ x } 10^{-2} \text{ N m}/1.3 \text{ x } 10^{-8} \text{ kg m}^2)^{1/2} \approx 1040/\text{s}$, which predicts an (ordinary) frequency of oscillation of $f = \omega/(2\pi) \approx 160$ Hz, close to frequencies that you'll observe.

How the reed's torsional motion is excited and detected

The torsional mode near 150 Hz would be hard to observe or detect, except that in your oscillator the mass is actually a permanent magnet. Its magnetic moment is perpendicular to the square faces of the two slabs attached to the reed, and the magnetic-moment vector **μ** points along the y-axis when the reed is untwisted. [The size of the magnetic moment is $\mu = M\ V$, where $M \approx 1.0 \text{ x } 10^6$ A/m is the magnetization of the NdFeB material used, and $V = s^3$ is the volume of the magnet. This gives $\mu \approx 0.26 \text{ A m}^2$.]

Now a magnetic moment **μ** interacts with a magnetic field ***B*** to give a torque

$$\vec{\tau} = \vec{\mu} \times \vec{B} \quad ,$$

and an interaction energy $U_{\text{mag}} = -\ \boldsymbol{\mu} \cdot \boldsymbol{B}$. Now look at one of your (two) torsional-reed oscillators, to see that the oscillator's reed is holding the magnet-block inside a coil wound around a white cylindrical form having an axis pointing along z. When a current i runs through this coil, it generates a field in the z-direction at the magnet's location, of magnitude $k_z\ i$ (where $k_z \approx 6.6$ mT/A can be predicted from the coil geometry and number of turns). Now with **μ** in the y-direction, and ***B*** in the z-direction, the cross-product gives a torque in the x-direction as desired: that's just the direction of torque needed to twist the reed. It has the predicted size $|\boldsymbol{\tau}| = \mu\ B \sin 90°$, and for a current of $i = 0.1$ A, this gives B = 0.66 mT and a torque of $172 \text{ x } 10^{-6}$ N m. Applied to a reed of torsion-constant κ, we expect (from a dc current and a static B) a static angular deflection

$$\Delta\theta = \tau / \kappa \approx (172 \text{ x } 10^{-6} \text{ N m}) / (1.4 \text{ x } 10^{-2} \text{ N m}) \approx 12 \text{ x } 10^{-3} \text{ rad} \quad ,$$

which is only about two-thirds of a degree. Though that static deflection would be small, a torque of this magnitude generated by an *alternating* current $i(t)$, and one oscillating at the reed-mass system's resonant frequency, could certainly 'pump up' the oscillator to a larger amplitude of oscillation.

So this coil system, driven by a current, allows the oscillation to be excited; but in reverse, this magnet-in-coil combination also allows the oscillation to be *detected*, electronically. That's due to Faraday's Law: a magnet with **μ** quiescently along the y-axis, but now in oscillatory motion in angle $\theta(t)$, can create a changing magnetic flux through any turn of the coil. That flux is zero quiescently, but varies around zero as the rotation varies around $\theta = 0$. The emf generated in the coil is given by the time rate of change of the flux, and (by a reciprocity theorem) it turns out the emf per unit angular velocity of rotation is equal to the previously computed torque per unit current, $\tau / i = \mu\, k_z \approx 1.7 \text{ x } 10^{-3}$ N m/A = $1.7 \text{ x } 10^{-3}$ V per rad/s.

Now if the reed develops a sinusodal oscillation in θ given by

$$\theta(t) = A\cos(\omega t) \quad ,$$

where $\omega \approx 1040$ /s gives the angular frequency of oscillation at resonance, and *A* is the amplitude of the angular motion, then the angular velocity is

$$\frac{d\theta}{dt} = -\omega A \sin(\omega t) \quad ,$$

and this has a peak value of ωA. Even for an amplitude of oscillation of only 1° = 0.017 rad, this has a magnitude of 18 rad/s. Then the emf in the coil is an ac voltage, at the resonant frequency, with a voltage amplitude

$$\mathcal{E} = (\mu\, k_z)\frac{d\theta}{dt} = (1.7\, x\, 10^{-3}\, \frac{\text{V}}{\text{rad/s}})(18\,\text{rad/s}) = 31\,\text{mV} \quad ,$$

which is easily detectable (especially using the SRS770 and Fourier methods!).

<u>Tuning the oscillator</u>

Thus far the natural frequency of the oscillator is set by the rotational inertia of the magnet's mass, and the torsional elasticity of the reed, but it is not otherwise adjustable. To give the system another independent variable, there is a provision to change the effective torsion constant. That's provided by an external and static magnetic field, which immerses the whole oscillator in a field $\boldsymbol{B}_y$ in the *y*-direction.

The effect of that field is to supplement the system's elastic potential energy $\kappa\, \theta^2/2$ with another term, $U_{mag} = -\, \boldsymbol{\mu} \cdot \boldsymbol{B}_y = -\, \mu\, B_y \cos\theta \approx -\, \mu\, B_y\, (1 - \theta^2/2) = \text{const} + \mu\, B_y\, \theta^2/2$. So the total potential energy becomes $(\kappa + \mu\, B_y)\, \theta^2/2$, and the effective torsion constant becomes $\kappa + \mu\, B_y$. So now the predicted torsional-oscillation frequency becomes

$$\omega = \sqrt{\frac{\kappa + \mu B_y}{I}} \quad .$$

and this predicts that $\omega^2 = (2\pi\, f)^2 = (\kappa + \mu\, B_y) / I$.

This B_y is generated by another set of coils, called the Tuning Coils, wound on two rectangular white plastic frames, and B_y will be proportional to the current i_T in these coils: $B_y = k_y\, i_T$. This new coil constant can be easily computed in the limiting case that the tuning coils are long in the *x*-direction, in which case their wires form a '4-wire field'. Along the center of such a structure (where, in our case, the magnet-blocks lie), the field is easily calculated, and the geometry of the structure and the number of turns give $k_y \approx$ 3.4 mT/A. So now we have the prediction that the experimentally-accessible quantity ω^2 ought to be a linear function of the coil current i_T, with intercept κ / I and slope $\mu\, k_y / I$. If a dc current of $i_T = \pm 2$ A is sent through these coils, we get $(\kappa + \mu\, B_y) \approx (0.014 \pm 0.0018)$ N m, showing that the effective torsion constant can be changed by order ±13%, and the frequency of the resonance changed by order ±7%. That represents the ability to 'fine-tune' the frequency of torsional oscillation, using a non-contact and real-time external parameter to do so.

Coupling two oscillators

If you look at your Coupled-Oscillator unit, you'll see *two* of these reed-based torsional oscillators in place, with their magnet-masses in each other's proximity. If you check using a little compass, you'll find that the two oscillators have their magnetic moments pointing in *opposite* directions (one along the +*y*, the other along the -*y*, direction). But the oscillators are otherwise nominally identical. To a good approximation, each one is excited by, and detected by, its own vertical-axis coil; but the two oscillators live in a single common B_y field generated by the Tuning Coil. Now if we name these oscillators #1 and #2, and if we assume they have the same rotational inertia *I*, but perhaps slightly different torsion constants κ_1 and κ_2, we can imagine that the two oscillators have two separate oscillation frequencies, given by

$$\omega_1^{\,2} = \frac{\kappa_1 + \mu\, B_y}{I} \quad \mathit{and} \quad \omega_2^{\,2} = \frac{\kappa_2 - \mu\, B_y}{I} \quad .$$

The signs differ precisely because one magnet has its $\boldsymbol{\mu}$ lying along $\boldsymbol{B}_y$, while the other has its $\boldsymbol{\mu}$ lying *opposed* to $\boldsymbol{B}_y$. But B_y is a field magnitude common to both oscillators, and it is controlled by a single external current i_T in the tuning coil. So a plot of ω_1^2 and of ω_2^2 , both as functions of i_T, ought to display two straight lines, of similar *y*-intercept, but of *opposite* slopes. In particular, there is a value of i_T at which these two lines ought to *cross* – that is the say, there's a current i_T which produces a field B_y which 'tunes' the two oscillators to have a single, common, oscillation frequency (despite, for example, trifling differences in their construction).

That would be called a 'mode crossing' because of the crossing of two normal-mode frequency lines in a plot, but the crossing does *not*, in fact, occur. Instead, you'll see an '**avoided crossing**', which is the single most general feature that can be found in coupled-oscillator systems across the breadth of physics. If your two oscillators each interacted with the field B_y, but not with each other, you'd see a crossing – but they are deliberately situated, with their magnets separated by distance *r* along the *x*-direction, so that they *do* interact. The interaction is the direct dipole-dipole coupling of two magnets, which gives anther interaction energy U_{int} which we'll now compute.

Let **μ**$_1$ be a point magnetic dipole regarded as the source of a magnetic field. Then located at vector displacement ***r*** from this source, a second magnetic dipole **μ**$_2$ finds itself immersed in a field

$$\vec{B}_1(\vec{r}) = \frac{\mu_0}{4\pi}\,[3\frac{\vec{\mu}_1\cdot\vec{r}}{r^5}\vec{r} - \frac{\vec{\mu}_1}{r^3}] \quad ,$$

and therefore experiences a magnetic interaction energy

$$U_{\text{int}} = -\vec{\mu}_2\cdot\vec{B}_1(\vec{r}) = \frac{\mu_0}{4\pi}\,[-3\frac{(\vec{\mu}_1\cdot\vec{r})(\vec{\mu}_2\cdot\vec{r})}{r^5} + \frac{\vec{\mu}_1\cdot\vec{\mu}_2}{r^3}] \quad .$$

For the case at hand, even during oscillations of each dipole, the magnetic moments **μ** lie in the *y*-*z* plane, while the inter-dipole displacement ***r*** is along the *x*-direction, so two of the dot-products vanish, and this interaction energy reduces to

$$U_{\text{int}} = \frac{\mu_0}{4\pi}\,[\frac{\vec{\mu}_1\cdot\vec{\mu}_2}{r^3}] \quad .$$

Now if we adopt the angular coordinates shown in Fig. 16.2 below, the dot product gives

$$U_{\text{int}} = -\frac{\mu_0}{4\pi}\,\frac{\mu_1\,\mu_2}{r^3}\cos(\theta_1-\theta_2) \approx -\frac{\mu_0}{4\pi}\,\frac{\mu_1\,\mu_2}{r^3}[1-\frac{(\theta_1-\theta_2)^2}{2}] \quad .$$

where we've used the small-argument form of the cosine function to simplify the expression.

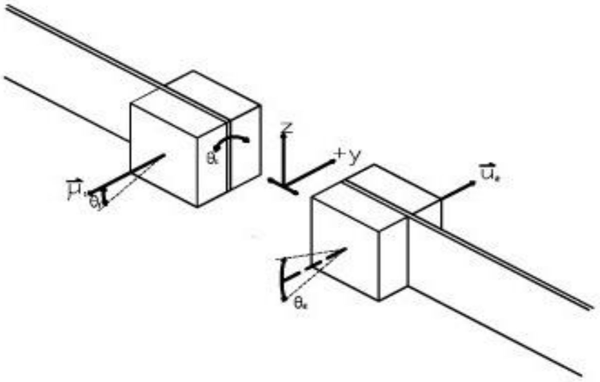


Fig. 16.2: Two magnetic moments, each on its own reed oscillator, each with its own angular coordinate for torsional oscillations.

This expression includes a leading term which depends on *r*, but not on the angles (it gets more negative as *r* is decreased, corresponding to the attractive forces that these oppositely-directed moments exert on each other). We care here about the fixed-*r* but angle-dependent term, which is lowest when the angles are equal, corresponding to the magnetic torques that these moments exert on each other, which tend to co-align the angles θ_1 and θ_2 we've introduced. Most importantly, expanding this interaction term, we'll find a term in the *product* $\theta_1\,\theta_2$, which will couple these two oscillators together and thereby totally change their behavior.

Now finally we can write the total angle-dependent potential energy of this system, including the elastic terms for each reed, each magnet's interaction with the field B_y, and the two magnets' interaction with each other. That gives

$$U_{net} = \frac{1}{2}\kappa_1\,\theta_1^2 + \frac{1}{2}\kappa_2\,\theta_2^{\,2} + \frac{1}{2}(\mu B_y)\theta_1^2 + \frac{1}{2}(-\mu B_y)\theta_2^{\,2} + \frac{1}{2}[\frac{\mu_0}{4\pi}\,\frac{\mu^2}{r^3}](\theta_1 - \theta_2)^2 \quad .$$

Here we have made the small-angle approximation, and we have ignored any numerical difference between μ_1 and μ_2. Recall the numerical magnitudes: the torsion constants κ are about 0.0140 N m, the coupling to the B_y-field gives μ B_y of about 0.0018 N m at a tuning-coil current of $I_T = 2$ A, and now the coupling constant

$$\lambda \equiv [\frac{\mu_0}{4\pi}\,\frac{\mu^2}{r^3}]$$

turns out to have magnitude about 0.0008 N m when the center-to-center separation of the two dipoles is taken to be $r = 20$ mm. So the coupling of the two oscillators to each other seems to be really weak, with $\lambda/\kappa \approx 0.06$, but nevertheless it turns out to have crucial consequences. The reason is that we can 'tune to the crossing', ie. put the system into a condition in which the two oscillators would otherwise have matching frequency. In such a case, motion of one oscillator will drive the other oscillator *resonantly*, so even this weak coupling can have large consequences.

Using the coupled-oscillator system: sinusoidal drive

It's time to set theory temporarily aside, and actually excite this coupled-oscillator system. For now, we'll ignore the Tuning Coil, and we'll start with ordinary electronic excitation and detection methods. You'll want an external sine-wave generator which you can hand-tune over the range 100-200 Hz, and you'll want a dual-trace oscilloscope for detection.

Before you do any electrical measurements on your Coupled Oscillator, you want to ensure it's properly adjusted. Use a fingertip or thin plastic probe to deflect one of the reeds sideway (ie. in the *y*-direction) and then 'let it go', which will excite the 'wobble motion' of that reed. The attractive force of interaction between the two magnets ensures that both reeds will participate in this wobble. What you want is for both reeds to be free to move; in particular, check that their top and bottom edges are not rubbing against the vertical-cylinder coil forms which enclose both magnets. If the wobble motion does not go on for many seconds, there is some friction involved. To eliminate it, find the screws which clamp the far ends of a reed, and loosen them just enough that the reed can be moved in its clamp. You can tilt the reed (ie. allow the magnet move in the *z*-direction) to free it from scraping the coil form.

After you have both reeds properly clamped, you can check that the clamps themselves are properly oriented relative to the frame of the apparatus. A screw on the top of each clamp allows the clamping block to be rotated about the *z*-axis direction. You may want to loosen both clamping blocks, and check their rotation about the *z*-axis so as to ensure that the mechanical equilibrium position of both magnets correctly centers them in their cylindrical coil forms – this adjusts their equilibrium position in the *y*-coordinate. Again, the magnets are sufficiently coupled that each reed will affect the other reed's zero-position. (If you want to *de*-couple the reeds, temporarily loosen one reed's end clamp, and pull that reed along its length to move its magnet farther away from the other magnet.)

Once you have confirmed that the 'wobble motions' are properly centered and mechanically free, you can hand-damp out any remaining wobble motion. But now you can be quite sure that the torsional motion you'll be investigating is also free and very nearly undamped.

Below is a circuit which seems to excite only one oscillator. You might set a sine-wave generator to about 1-Volt amplitude, and apply it to one drive coil's BNC connector so that current flows through the built-in 1-kΩ resistor as shown. This turns the generator into a 'current source', of about 1-mA amplitude. The connection shown to the 'scope's ch. 1 is then a voltage surrogate for the drive current being applied to the coil. Meanwhile ch. 2, connected via a BNC adapter to the banana plugs wired directly to this coil, will show the sum of two voltages: one is the $i(t)\cdot R$ resistive drop across the drive coil, and the other is the Faraday's-Law emf generated in the coil by the torsional motion of the oscillator. At resonance, this 'back emf' will *dominate* over the resistive part of the signal.

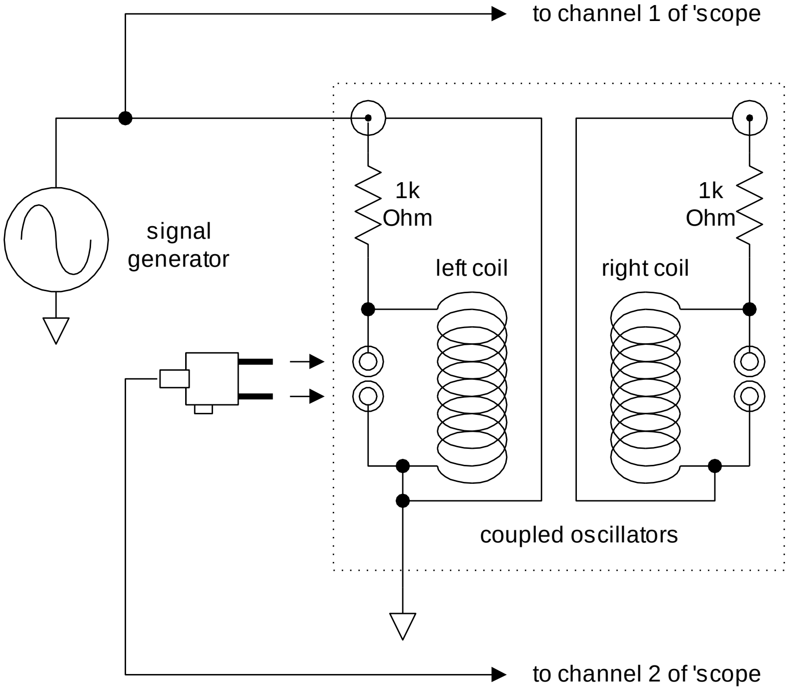


Fig. 16.3: A first way to use a sine-wave generator to excite, and an oscilloscope to monitor, the state of one oscillator.

Now set your ‘scope so the ch. 1 signal triggers the ‘scope, and displays a few cycles across the screen. Set ch. 2 to display as well, and set it to a sensitivity of about 10 mV/div. For a generic value of the drive frequency, you expect on ch. 2 to see the $i{\cdot}R$ drop across the coil, which might have an amplitude of about 10 mV, and will certainly be in phase with the drive being displayed on ch. 1. What you want to do is to vary the drive frequency, quite slowly, over the 100-200 Hz range, and look for a *departure* from this mere $i{\cdot}R$-drop behavior. That departure will show up with several characteristics:

- it will be of larger amplitude, perhaps up to 40 mV;
- it will show a departure in *phase* compared to the drive waveform; and
- it will take some *time* to develop, since you’re driving a high-Q resonant system.

See if you can close in on the resonant frequency, at which the ch. 2 signal should be maximal in amplitude, and back in phase with the drive. You will have to tune to a *fraction of 1 Hz* to locate the center of the resonance.

But after you find ‘the resonance’, go looking for another – it might be about 10 Hz away in frequency. It will have similar characteristics as the one you found first, though it might differ somewhat in size.

Even though you think you’re exciting, and detecting, only one of your two oscillators, they form a coupled system, and the system has two normal modes. *Either* mode can be excited by resonant drive at any point in the system, as in this case where you first excite one, then the other, normal mode, using a drive coil that is nominally coupled only to *one* magnet-on-reed oscillator.

There is a useful display that you are now in a position to use – configure your ‘scope to the XY-mode, in which the ch. 1 signal drives the display spot left-right, and the ch. 2 signal drives it up-down, on the display. Away from resonance, you’ll see a nearly ‘flat line’. Near, but not at, resonance, you’ll see an open curve, a titled ellipse, which indicates phase shifts between drive current and emf response. At resonance, the ellipse will collapse back to a line, but a tilted line

This titled-line plot is also very useful for checking to see if you are over-driving the system. If you’re on resonance, and reduce the generator’s amplitude by a factor of two, the whole plot you’re seeing should shrink by a factor of 2. If instead it causes the line to open up to an ellipse, this says the resonance frequency for larger drive *differs* from the resonance frequency for smaller drive. That sort of anharmonicity is a sign of overdriving the system. You might use the new and smaller amplitude to repeat the search, and find more precise values for the two resonant frequencies of the two normal modes of the system.

Here is an alternative, and even more informative, way to interrogate the system: leave the generator and ch. 1 connections just as they are, but now use the BNC-to-banana adapter to connect *ch. 2* of the ‘scope to the drive/pick-up coil of the *other* oscillator. Now there will be no $i{\cdot}R$-drop at all, and the ‘scope will display, on ch. 2, purely an emf signal, directly related to the angular velocity of the reed/magnet you’re <u>not</u> exciting directly. Now only the coupling of the two oscillators provides a pathway for the ch. 1

drive to communicate to the ch. 2 response. But you should see very similar indications of two resonances, with the advantage that there is no $i{\cdot}R$-background under your ch. 2 signal. There will also be, in this mode, this distinction between the two resonances: at the center of the second resonance, the phase of the response will be 180 degrees different, ie. flipped in sign, compared to the case of the first resonance you saw. If you use the XY-display method on the 'scope, then the ellipse that you see in general will collapse to a line at the center of each resonance, but that tilted line will have an up-slope for one resonance, and a down-slope for the other.

You will soon tire of the fine-tuning required on your signal generator, so instead of tediously moving the frequency by 1 Hz or even 0.1 Hz at a time, over a range of frequencies 10 or 100 Hz wide, it's time to *try all the frequencies at once*. That is to say, it's time to use Fourier methods.

Exciting the system by noise, and by chirp, waveforms

The method you're about to apply to the Coupled-Oscillator system is the same as you might have used in Ch. 8 on an acoustic system. The idea is to send white noise – a superposition of sinusoids at *all* frequencies – into the system, and then to Fourier-analyze the results, to see which frequencies are preferentially transmitted by the system.

In the present case, you can use the Source Out of the 770, configured with the SOURCE button to be white noise of 1000 mV rms measure, as your source of noise. You can send that, through the 1-kΩ resistor, to one of the drive coils of your Coupled Oscillator. Then the emf generated in the *other* vertical-axis coil, serving as pick-up coil for the other oscillator, is the signal you'll want to Fourier-analyze. Convey it to the A-input of the 770, and set up as usual to measure a spectrum. Remember to use the AutoRange function to optimize the 770's input for the (rather low) level of signal you'll be picking up.

You know that the resonances you care about lie in the 100-200 Hz range, so you might pick a frequency span of 390 Hz (and a start frequency of 0 Hz). This will entail an acquisition time of 1.024 s. You will see a Coupled-Oscillator's full spectrum in a second, and upon repeated acquisitions you'll see the statistical fluctuations to be expected from noise excitation. (Use the Continuous mode of Triggering, under the INPUT button, to get ongoing updates of the spectrum.) You can use the Average function to reduce these fluctuations, though at a cost in update time.

The MEASure button will let you use the Measure softkey to choose to measure spectrum, or power spectral density. The Display softkey will let you pick Log Magnitude, which will give you a fine view of the large dynamic range you can cover. The Scale button will let you pick Top Reference and dB/div settings for your vertical scale. Recall that 80 dB of vertical scale stands for a voltage range of 10^4 which you can cover. With so large a range of sensitivity, you're sure to see other modes than the two you previously discovered by ‘scope and scanning. Here are some clues about ‘other peaks’:

- Look for signals at 60, 120, 180 Hz (or integer multiples of your local line frequency) -- these represent interference (for example, direct inductive or capacitive coupling to your pick-up coil) and might persist even if the noise drive were to be removed.
- Look for signals about 6 (or 12) Hz, or for sidebands 6 (or 12) Hz away from the main peaks – these may be due to the modulation of the torsional modes by the ‘wobble’ motion of the reeds. If you damp away that wobble using temporary by-hand intervention, you might see these diminish.
- Look for the non-zero *width*-in-frequency of the main modes -- this is due either to finite acquisition time, or to the finite lifetime of the torsional modes due to their damping.
- Look for unexplained or unassigned modes – this is where Fourier methods allow unexpected (or undesired) *discoveries*.

For the best measurement of the frequencies of the main torsional modes, you’ll want to decrease the 770’s frequency span even more, perhaps by three more factors of 2. You can also change the Start Frequency of the span, so as to center its coverage where you want it. You’ll see two separate kinds of disadvantages come with this zoomed-in view:

- The acquisition time goes up by that many factors of 2; for a 49-Hz span, it will have risen to 8.192 s. Averaging will further slow the response of the system to any changes.
- The spectra you see will look (and be) ‘noisier’, ie. subject to larger fluctuations. That’s because you are spreading a fixed amount of noise power across more bins in frequency, so there’s less noise power (and thus greater fluctuation) per bin.

You can cure this latter problem, and very dramatically, by changing the Source configuration from Noise to Chirp. In the Noise mode, an rms output of 1000 mV has to contain frequency components over the full 0-100 kHz range, no matter what span you actually choose to analyze. But in the Chirp mode, the Source synthesizes a noise-like equal superposition of *only* those frequencies you are configured to analyze. Since it can now (for example) devote 1000 mV of output range to a span of 49 Hz rather than 100 kHz, the spectral power per bin analyzed can go up by 2^{11} = 2048 (!). You may have to use the AutoRange button to deal with the much larger signals that result. *But* when using the Chirp waveform, you will also want to change the Window mode to **Uniform**.

You can also deal with part of former problem. Chirp source or not, a span of 49 Hz still requires an acquisition time of >8 seconds. But it does *not* require Averaging to reduce the fluctuations, since the Chirp waveform is synthetic and deterministic, it is in fact *periodic* with the period given by the acquisition time. So what you get on one acquisition is what you’ll get on any other.

Clearly with your fine views of spectral peaks, you’re now in position to use the Marker function to estimate the resonant frequencies to about 0.1-Hz precision and accuracy.

Tuning the normal modes

There is no special interest in the exact numbers you get for the resonant frequencies you've measured, except if they can be understood and changed systematically. Clearly the use of different length, thickness, or width of the torsional reeds would have put the resonances at different locations in frequency.

But the fact that there are two main resonances is no accident; it's due to the fact that there are two oscillators, and that they're coupled together to give two normal modes. Your next goal is to 'tune' the individual oscillators' frequencies, to see the systematic changes this creates in the two normal-mode frequencies.

The Tuning Coil previously mentioned is your way to achieve this. You can use up to ±2 Amp currents continuously, or ±3 A briefly, of dc current in these coils. There is a self-resetting fuse in series with the coil, which will interrupt the current if it gets too hot. If you find that the current suddenly drops to zero, turn the supply voltage down to zero, wait about a minute, and then watch the current as you dial up the supply voltage again.

The current you send through the Tuning Coils will create a *y*-directed field (of about 3.4 mT/A) in the region of the reed-mounted magnets. But before taking spectroscopic data on their effect, you ought first to check the centering of the tuning coils relative to the oscillators.

> Here's what to do, and what to look for. Have the tuning coils connected to a dc power supply, and be ready to hand-dial that from $i_T = 0$ to 2 Amperes and back. Now look down from above at your two oscillators' reeds (no need to have either drive coil connected at this point), and watch to see if the magnets *move* (in the ±*y*-direction) when you dial up the current.
>
> If you see motion, it's due to the imperfect centering (in the *y*-direction) of the coils relative to the magnets. On the centerline of the coils, the B_y-field is gradient-free, so that field exerts the desired torques, but not any forces, on the magnets. But if the magnets find themselves not on centerline, then gradients exist in B_y, creating forces on the permanent magnets. You might see one magnet pulled in the +*y*-, the other in the -*y*-direction (because the same field gradient is acting on the *oppositely*-directed magnetic moments of the two magnets).
>
> Once you see this effect, you want to eliminate or minimize it. To do so, you want to change the *y*-position of the tuning coil's centerline. You can do so most easily by loosening two brass round-head clamping screws, and then sliding sideways, or tilting the coil forms slightly (shimming underneath one or the other of the plastic forms) to get the centering you want.
>
> When you can see no ±*y*-motion of the magnets upon changing i_T from 0 to 2 Amperes, you are adequately aligned. In particular, you can now be sure that

tuning the coils will not move the magnets and reeds to the point of touching the drive-coil forms (and thereby spoiling their high Q of torsional oscillations).

Given proper alignment of the tuning coils, you can now restore Noise or Chirp drive of one oscillator, and the Fourier-analysis of the pick-up from the other, to do the spectroscopy of your resonant modes. But now you have an independent variable, the tuning-coil current i_T, to affect the two dependent-variable frequencies you're seeing. You should first do an exploration over the full -2-A to +2-A range of i_T, to confirm that the two frequencies do indeed change. You will find a range of current over which the frequencies scarcely change at all; that's the range over which the two frequencies attain a minimum separation. Outside of this range of i_T, expect the two frequencies to move apart from each other. Note well that the two frequencies do *not* 'coalesce' or cross as functions of i_T.

Gather data, for a series of choices of i_T, of the higher (f_+) and lower (f_-) normal-mode frequencies. When you've learned how to do this over the (-2, +2)-A range, get a few points outside this range. Limit the tuning-coil current to the ±3-A range, and spend less than a minute at ±3-A levels before allowing the coils to cool.

Theory of the normal-mode frequencies, and the avoided crossing

The 'Fourier spectroscopy' you've now done has given you a data table – for each of a number of values of tuning current i_T, you have measured f_+ and f_-, the higher- and lower-frequency normal-mode resonant frequencies. What can be done with that data?

Theory (below) suggests forming some new plots. First form the values $\omega_+ = 2\pi f_+$, and $\omega_- = 2\pi f_-$, and then the combinations $(\omega_+^2 + \omega_-^2)$ and $(\omega_+^2 - \omega_-^2)^2$, and plot those as functions of I_T. Theory predicts the emergence of a *constant* for the first, and a parabola for the second. On these new plots, especially with the fits added, any discrepant data points will stand out.

Recall that for *un*-coupled oscillators you expected two separate frequencies ω_1 and ω_2 for oscillators #1 and #2, and you expected the quantities ω_1^2 and ω_2^2 each to display a straight-line, but differently-sloping, variation as a function of I_T. But that would give rise to a crossing of the two straight lines, and as a result, you'd have gotten for $(\omega_+^2 - \omega_-^2)^2$ a parabola, but one having a minimum value of zero. The data for coupled oscillators says otherwise – the quantities ω_+^2 and ω_-^2 reach a minimum, but *non*-zero, separation. Why?

The answer comes from the finding the system's equations of motion, and solving them. Using the complete potential-energy expression for U_{int} above, we can get the response to the torques on the two oscillators via

$$I\,\ddot{\theta}_1 = \Sigma\tau = -\partial U_{net}/\partial\theta_1 + (\mu k_z)\; i_1(t) \quad ;$$

$$I\,\ddot{\theta}_2 = \Sigma\tau = -\partial U_{net}/\partial\theta_2 + (\mu k_z)\; i_2(t) \quad .$$

Here i_1 and i_2 are the currents send into the two vertical-axis drive coils. These become

$$I\,\ddot{\theta}_1 + (\kappa_1 + \mu\,B_y)\,\theta_1 + \lambda\,(\theta_1 - \theta_2) = (\mu\,k_z)\ i_1(t) \quad ;$$
$$I\,\ddot{\theta}_2 + (\kappa_2 - \mu\,B_y)\,\theta_2 + \lambda\,(\theta_2 - \theta_1) = (\mu\,k_z)\ i_2(t) \quad .$$

which are two coupled 2nd-order differential equations.

Now we look for a normal mode of the *un*driven system by setting $i_1 = 0 = i_2$, and asserting the existence of oscillations at a single *common* frequency ω:

$$\theta_1(t) = A_1 \exp(-i\,\omega t) \quad \text{and} \quad \theta_2(t) = A_2 \exp(-i\,\omega t) \quad ;$$

Notice that we claim purely sinusoidal motion of both θ-coordinates, and at one single system-wide frequency.

Under this assumption, the coupled differential equations for $\theta_1(t)$ and $\theta_2(t)$ turn into coupled *linear* equation for amplitudes A_1 and A_2. They can be written in matrix form as

$$\begin{bmatrix} -I\omega^2 + (\kappa_1 + \mu\,B_y + \lambda) & -\lambda \\ -\lambda & -I\omega^2 + (\kappa_2 - \mu\,B_y + \lambda) \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad .$$

This homogeneous set of equations does not give A_1 or A_2, but instead gives (two versions of) the ratio A_1 / A_2. Those two versions are <u>in</u>consistent, unless the determinant of the matrix above vanishes. Writing out that determinant gives a quadratic equation for the unknown ω^2, which will give two solutions for ω^2. To simplify the form of that solution, consider two frequencies the system could have:

let Ω_1 be the frequency at which θ_1 would oscillate, if θ_2 were held fixed at 0;
and let Ω_2 be the frequency at which θ_2 would oscillate, if θ_1 were held fixed at 0.
From the equations of motion we find that these imaginable frequencies are given by

$$I\,\Omega_1{}^2 = \kappa_1 + \mu\,B_y + \lambda \quad \text{and} \quad I\,\Omega_2{}^2 = \kappa_2 - \mu\,B_y + \lambda \quad ,$$

and this gives a simpler form to the vanishing of the determinant, which becomes

$$I\,(\Omega_1{}^2 - \omega^2)\cdot I\,(\Omega_2{}^2 - \omega^2) - \lambda^2 = 0 \quad .$$

The quadratic equation to solve is

$$\omega^4 - (\Omega_1{}^2 + \Omega_2{}^2)\,\omega^2 + \Omega_1{}^2\,\Omega_2{}^2 - \lambda^2 / I^2 = 0 \quad .$$

There are two solutions for ω^2, and we'll call the larger of them $\omega_+{}^2$, and the smaller $\omega_-{}^2$; the two are given by

$$\omega_\pm{}^2 = [\,\Omega_1{}^2 + \Omega_2{}^2 \pm \sqrt{(\Omega_1{}^2 + \Omega_2{}^2)^2 - 4(\Omega_1{}^2\,\Omega_2{}^2 - \lambda^2 / I^2)}\ ]/2$$
$$= [\,\Omega_1{}^2 + \Omega_2{}^2 \pm \sqrt{(\Omega_1{}^2 - \Omega_2{}^2)^2 + (2\lambda / I)^2)}\ ]/2 \qquad .$$

This is the form which motivates plotting the two linear combinations mentioned above, since it predicts

$$\omega_+{}^2 + \omega_-{}^2 = \Omega_1{}^2 + \Omega_2{}^2 = \frac{1}{I}(\kappa_1 + \mu B_y + \lambda + \kappa_2 - \mu B_y + \lambda) = \frac{1}{I}(\kappa_1 + \kappa_2 + 2\lambda) \quad ,$$

and since B_y drops out of this prediction, it predicts a flat-line plot as a function of I_T.

The predictions for $\omega_\pm{}^2$ can also be used to give the *squared* difference of squares,

$$(\omega_+{}^2 - \omega_+{}^2)^2 = (\Omega_1{}^2 - \Omega_2{}^2)^2 + (2\lambda / I)^2 \quad ,$$

and since $\Omega_1{}^2$ and $\Omega_2{}^2$ are both linear in B_y, and hence in I_T, this is predicted to display a *parabolic* dependence on the tuning current. That parabola, in turn, is predicted to reach down to a minimum value of $(2\ \lambda / I)^2$, directly related to the coupling λ between the two oscillators. The *x*-axis location of the parabola's minimum is the condition

$$\Omega_1{}^2 = \Omega_2{}^2 \quad , \quad \text{ie.} \quad (\kappa_1 + \mu B_y + \lambda) = (\kappa_2 - \mu B_y + \lambda) \quad ,$$

or at $2\ \mu\ B_y = \kappa_2 - \kappa_1$; this just gives the value of the B_y -field at which the two oscillators have been 'brought into tune'.

At this minimum-separation point, we have $(\omega_+{}^2 - \omega_-{}^2)\ {}^2 = 0\ {}^2 + (2\ \lambda / I)^2$, so we get $(\omega_+{}^2 - \omega_-{}^2) = (+)\ (2\ \lambda / I)$. Meanwhile, we can put $\kappa_{avg} \equiv (\kappa_1 + \kappa_2) / 2$ and write $\omega_+{}^2 + \omega_-{}^2 = (2\ \kappa_{avg} + 2\ \lambda) / I$, so now it's easy to solve for the values of $\omega_+{}^2$ and $\omega_-{}^2$ at the minimum separation:

$$\omega_+{}^2 = (\kappa_{avg} + 2\lambda) / I \quad \text{and} \quad \omega_-{}^2 = \kappa_{avg} / I \quad \text{(at minimum separation)} \quad .$$

So quite directly from the data, you can read off the values of κ_{avg}/I and λ/I, and using the computed value of rotational inertia I will then give deduced values for κ_{avg} and for λ, to be compared to earlier estimates.

Away from minimum-separation, we can define $B_y{}^{\min} = (\kappa_2 - \kappa_1) / (2\mu)$ as that B_y-value needed to attain the minimum-separation condition, and then it's easy to show that

$$(\omega_+{}^2 - \omega_+{}^2)^2 = (2\mu / I)^2 (B_y - B_y{}^{\min})^2 + (2\lambda / I)^2 \quad ,$$

This form displays the parabola most clearly, and it shows that a best-fit of the data to a parabola (against B_y) will give parameter values $(2\ \mu / I)^2$, $B_y{}^{\min}$, and $(2\ \lambda / I)^2$. Again using the computed value of I, these will give deduced values for μ and for λ. With values for those parameters in hand, the entire theoretical model is specified, and plots of $\omega_+{}^2$ and $\omega_-{}^2$, or of ω_+ and ω_- , or of f_+ and f_- , can be generated and laid atop the data. For a canonical view of the data, plot measured $\omega_+{}^2$ and $\omega_-{}^2$ data-points, overlay the theory's predicted curves for $\omega_+{}^2$ and $\omega_-{}^2$, and add in the plots of the $\Omega_1{}^2$ and $\Omega_2{}^2$ functions, to see the asymptotes toward which the theoretical curves tend. You'll be getting the best view of the system's avoided crossing.

One motivation for making such plots is that it's then easy to model how they would change if the coupling λ were to vary. (Notice that in the hardware, it is feasible – by loosening three screws – to adjust the separation r between the two magnets, and notice also that modest changes in r will make substantial changes in λ, due to the r^{-3} dependence.) You should find that a smaller λ-value entails that the two curves for ω_+^2 and ω_-^2 will approach each other more closely – they 'more narrowly avoid' a crossing. But in principle, for any degree of coupling, however small, the curves do not cross.

Another prediction of the theory can be worked out, and that is for the ratio A_1 / A_2 of amplitudes of the two separate coordinates θ_1 and θ_2. This is algebraically easy only at the minimum separation point $B_y = B_y^{\min}$, and there you will find

in the ω_- mode, $A_1 = +A_2$, so that $\theta_1(t) = +\theta_2(t)$,
but in the ω_+ mode, $A_1 = - A_2$, so that $\theta_1(t) = -\theta_2(t)$.

In fact this makes it clear why (at minimum separation) we found $\omega_-^2 = \kappa_{\text{avg}} / I$, *in*dependent of the coupling λ; since the two magnets oscillate in unison in the ω_- mode, the coupling energy $\lambda\,(\theta_1 - \theta_2)^2 / 2$ is, and remains, zero. The surprise is that the two oscillators can and do oscillate in unison, at equal amplitude, even though the drive is being applied to only one of them, and coupling seems to be 'doing no work'.

Away from the minimum-separation condition, $| A_1 / A_2 | \neq 1$, so each normal mode is characterized by having the motion 'concentrated' in one of the two oscillators. Your spectroscopy has given you numerical values for all the constants you need to make a complete theoretical prediction for how this amplitude-ratio varies with the tuning.

Other investigations of the normal modes

Now that you've seen the glamorous avoided crossing in a coupled-oscillator system, you might operate at the minimum-separation point and consider what *else* you can do with this system.

One sort of fun is to use burst-mode sine-wave excitation (see Ch. 10 for discussion of the spectral properties of such bursts.) You could use the circuit of Fig. 16.3 again, but send in a burst of N cycles of a sinusoid, at a chosen frequency, instead of a continuous signal. Knowing what you know about the frequency content of a burst, you could even arrange your burst to have a spectral peak at the frequency of one normal mode, and a spectral null at the location of the other. That is to say, you could selectively excite one normal mode.

Then a 'scope view of the emf generated after the burst, in either *or both* of the two drive/pick-up coils, could show you the ring-down of that normal mode in the time domain. In this way you could also investigate the phase relationship of the two magnets' torsional oscillations, which is predicted to be 180° different for one normal mode compared to the other.

If you can measure the ring-down time for a normal mode, you can check to see if the damping is the same, or different, for the two normal modes. You could also see if the damping time (measured in the time domain) and the resonance width (measured by Fourier methods) are related as they ought to be.

It is also possible to tailor a burst-mode excitation which sets up a *superposition* of both normal modes. This would require a burst with equal Fourier components at the two normal-mode frequencies, which can be achieved by the right choice of the frequency, and duration, of the burst. You'll find that such a burst needs to be of relatively short duration. But there is a payoff, since the time evolution of what you excite is quite dramatic. Since you are exciting the system through a single drive coil, the combination of normal modes that you're exciting must be one which has a large motion of the magnet that's located in the coil you're driving, and which entails rather little motion of the other magnet. The time evolution of that superposition of normal modes is highly structured in time. Suppose you're at the minimum-separation condition (where the effect is the most dramatic). Recall that at that condition, the ω_- mode has $\theta_1(t) = +\theta_2(t)$, while ω_+ mode has $\theta_1(t) = -\theta_2(t)$. Supposing you're exciting oscillator #1 near time $t = 0$, the superposition you set up must have a time evolution described by

$$\theta_{1.}(t) = A\exp(-i\,\omega_-\,t) + A\exp(-i\,\omega_+\,t) \quad ,$$
$$\theta_2(t) = A\exp(-i\,\omega_-\,t) - A\exp(-i\,\omega_+\,t) \quad .$$

Notice that near $t = 0$, θ_1 is oscillating with an amplitude near $2A$, while θ_2 is scarcely oscillating. But as time evolves, precisely because the frequencies ω_+ and ω_- are different, the phase relationships of the two terms involved changes. In the vicinity of a time t_{swap} defined by

$$\omega_+\, t_{\text{swap}} = \omega_-\, t_{\text{swap}} + \pi \quad ,$$

the larger-frequency term will have gained an entire half-cycle on the smaller one, and the phases will have become such that the two terms in $\theta_1(t)$ are now nearly cancelling, while the two terms in $\theta_2(t)$ are now adding. That is to say, during the time between $t = 0$ and t_{swap} , the energy of the motion has been 'swapped' from being largely in oscillator #1 (where it was injected by the drive), to being largely in oscillator #2 (where it arrived via the coupling between the oscillators). This marks the first half of a repeating cycle of energy exchanges back and forth between the oscillators. But note that from the viewpoint of normal modes, each normal mode is steadily evolving at a fixed amplitude, and this swapping is just a manifestation of the superposition.

You can show that the 'swap time' is inversely proportional to the coupling coefficient λ (which you can vary), and that it is rather smaller than the ring-down time due to damping, so you can hope to see many cycles of this swapping phenomenon. The signature of this behavior is the emf in oscillator #2's pick-up coil, which should start near $t = 0$ with little signal, and then have the energy come (and go) by this transfer phenomenon.

The transfer function for the coupled-oscillator system

Strictly speaking, a normal mode is the pattern in space, of single-frequency sinusoidal oscillation that an excited but otherwise isolated system will display. So a temporary drive, followed by an autonomous ringdown, of a system would be the purest exhibition of normal-mode behavior. What you've been seeing, via the noise- or chirp-driven method of Fourier spectroscopy, is instead something like the steady-state behavior of the system, which is driven with coil #1, and monitored using coil #2. In fact, with the original sine-wave-and-'scope method of excitation, what you were seeing was the *transfer function* of the apparatus, its steady-state response to a single sinusoid. Now it's time to extract predictions about that transfer function, by solving the equations of motion for a drive of coil #1, and making them predict the emf produced in coil #2. You should not be surprised to learn that this transfer function has two tall peaks, located at the two normal-mode frequencies. But away from those peaks, this calculation ought to predict the full panorama, the details of the line-shape, of what you have detected in your Fourier spectroscopy.

Here's the method: into the equations of motion, we model the drive conditions by putting $i_1(t) = i_{\mathrm{dr}} \exp(-\,i\,\omega\,t)$ and $i_2(t) = 0$; here ω is now an independent variable, a drive frequency that can be freely chosen. Then we look for steady-state solutions sharing this frequency, of the form

$$\theta_1(t) = A_1 \exp(-\,i\,\omega\,t)\ , \quad \text{and} \quad \theta_2(t) = A_2 \exp(-\,i\,\omega\,t)\ \ .$$

The observable signal is the emf generated in coil #2, which is given by

$$\varepsilon_2 = (\mu\,k_z)\,d\theta_2 / dt = (\mu\,k_z)\cdot -i\,\omega A_2 \exp(-i\,\omega\,t) \quad .$$

Putting the assumed solutions into the original equations of motion, we get

$$\begin{aligned} &I\,\ddot{\theta}_1 + (\kappa_1 + \mu\,B_y)\,\theta_1 + \lambda\,(\theta_1 - \theta_2) = (\mu\,k_z)\ i_{dr} \exp(-i\,\omega t) \quad ; \\ &I\,\ddot{\theta}_2 + (\kappa_2 - \mu\,B_y)\,\theta_2 + \lambda\,(\theta_2 - \theta_1) = (\mu\,k_z)\ 0 \quad , \end{aligned}$$

and these becomes an algebraic set of equations,

$$\begin{aligned} &-I\omega^2\,A_1 + (\kappa_1 + \mu\,B_y)\,A_1 + \lambda\,(A_1 - A_2) = (\mu\,k_z)\ i_{dr} \quad ; \\ &-I\omega^2\,A_2 + (\kappa_2 - \mu\,B_y)\,A_2 + \lambda\,(A_2 - A_1) = 0 \quad . \end{aligned}$$

These can be put into matrix form, now an *in*homogeneous set of linear equations for the unknown amplitudes A_1 and A_2:

$$\begin{bmatrix} -I\omega^2 + I\,\Omega_1^{\,2} & -\lambda \\ -\lambda & -I\omega^2 + I\,\Omega_2^{\,2} \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} \mu\,k_z\ i_{dr} \\ 0 \end{bmatrix} \quad .$$

Now the matrix can be inverted, which gives the solution for A_1 and A_2:

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \frac{1}{\det}\begin{bmatrix} I\,(\Omega_2{}^2-\omega^2) & \lambda \\ \lambda & I\,(\Omega_1{}^2-\omega^2) \end{bmatrix}\cdot\begin{bmatrix} \mu\, k_z\; I_{dr} \\ 0 \end{bmatrix} \quad .$$

Here the denominator is given by the determinant

$$\det = I(\Omega_1{}^2-\omega^2)\cdot I(\Omega_1{}^2-\omega^2)-\lambda^2 \quad ,$$

which can be shown to be equal to

$$\det = I^2(\omega^2-\omega_+{}^2)(\omega^2-\omega_-{}^2) \quad ,$$

in terms of the two normal-mode frequencies previously found.

So the result can finally be written in reasonably compact form,

$$\varepsilon_2(t) = \lambda\,(\frac{\mu\, k_z}{I})^2\;\frac{-i\,\omega}{(\omega^2-\omega_+{}^2)(\omega^2-\omega_-{}^2)}\;i_1(t) \quad .$$

As it's written, this expression *diverges* when the drive frequency matches either of the two normal-mode frequencies. In practice, the relevant term in the denominator goes not to zero, but to a small (and pure-imaginary) value, set by the damping that is sure to be present in the system. That is to say, in the presence of damping, the factor

$$\omega^2-\omega_+{}^2 \rightarrow \omega^2-\omega_+{}^2+2i\,\gamma_+\,\omega\,\omega_+ \quad ,$$

where γ_+ is the (dimensionless, positive-real, and rather small) damping constant of the ω_+ normal mode; and similarly for the other factor. With these changes, the expression above gives the full complex-valued transfer function of the driven system, including phase information. It is easy to see that when the drive frequency matches either of the two normal-mode frequencies, the emf response is *in phase* with the drive, but that this phase shift goes through 90° over the range of drive frequencies corresponding to the full-width at half-maximum of the resonance.

While the complex transfer function computed above thus contains the full magnitude and phase information about the system, comparison with the Fourier spectroscopy conducted above will require the extraction of the *magnitude* of the predicted transfer function. That, in turn, will require estimates of the various parameters in the theory, but the only critical ones are the two normal-mode resonance frequencies ω_+ and ω_-.

That completes the theory for the transfer function of this coupled-oscillator system. For a truly elegant way to acquire experimentally the frequency panoramas of the magnitude *and the phase* of the transfer function, making full use of the 770's chirp capabilities, see Appendix A16.

Chapter 17: Fourier methods for detecting non-linearity

There is a class of physical systems, called 'linear time-invariant' systems, with very special properties. Such systems, when excited by a steady sinusoid, also respond with a pure sinusoid, perhaps of modified amplitude and phase, but of *un*changed frequency. But many physical systems are deliberately or accidentally non-linear, and Fourier methods provide a very sensitive way to detect such non-linearity. In this Chapter, we'll see a very simple electronic system which can exhibit non-linear behavior, and we'll see some of the Fourier methods for detecting that non-linearity.

One of the reasons for this sort of study is that non-linearity can be very desirable in some branches of physics. Both in electronics and optics, a non-linear system, when driven by a sinusoid of frequency *f*, can exhibit an output which contains (perhaps weak) terms at frequencies 2*f* and/or 3*f*. This frequency-doubling or -tripling can make available frequency outputs in an otherwise inaccessible regime. For example, the now-common 'green laser pointer' involves a (diode-laser-pumped) neodymium laser oscillating at wavelength near 1064 nm, whose radiation is frequency-doubled (and thus wavelength-halved) to give its visible green output at 532 nm.

The physical system you'll study is much simpler. It's the Intermodulation Distortion section of your Electronic Modules, and it is just a voltage divider with selectable non-linearity.

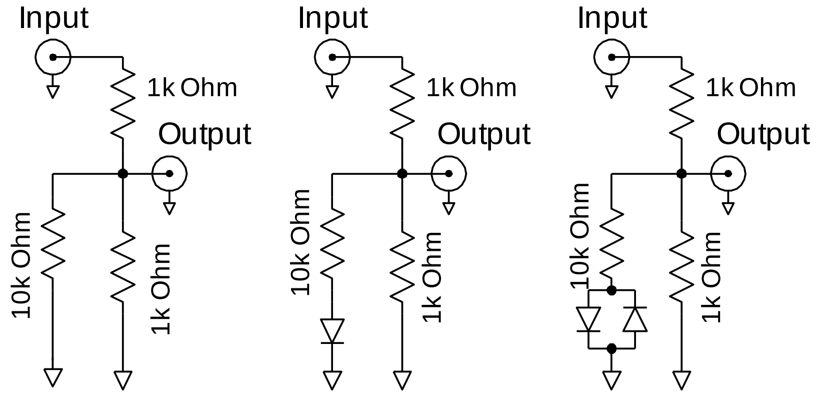


Figure 17.1: Schematic diagrams for the Intermodulation-Distortion module, with the circuits left, center, and right (respectively) giving the results selected by the switch settings 'none', 'asymmetric', and 'symmetric'.

In the 'none' position of the toggle switch, the circuit is a purely-resistive voltage divider. Given a low-impedance voltage source to create V_{in}, and a high-impedance voltmeter to measure V_{out}, you can see that the expected behavior is

$$V_{out} = \frac{R_1}{R_1 + R_2} V_{in} = \frac{0.91k}{0.91k + 1.00k} V_{in} = 0.476\, V_{in} \quad .$$

This proportional behavior does indeed satisfy the requirements (see Ch. 7) for a linear time-invariant system.

But if the switch is in one of its other two positions, then one or two diodes are involved in the voltage divider. For input voltages of order 1 Volt, there will be of order 0.5 V across the diode(s), causing them to conduct to some degree, which modifies the mapping from V_{in} to V_{out}. For any frequencies below (say) 10 MHz, we still expect V_{out} to depend on V_{in}'s present value only, but the functional relationship will no longer be a simple proportion. Instead, there will be a non-linear relationship, and a functional dependence $V_{\text{out}}(V_{\text{in}})$. If we Taylor-expand that functional form for small values of V_{in}, we expect a series of the form

$$V_{out}(V_{in}) = 0.500\,V_{in} + a_2\,V_{in}{}^2 + a_3\,V_{in}{}^3 + \ldots \quad .$$

In fact you can see that in the switch position marked 'symmetric', the presence of back-to-back diodes entails an (anti)symmetry between positive and negative input voltages, and in this position we expect V_{out} to be an odd function of V_{in}. Then the Taylor expansion we expect is

$$V_{out}(V_{in}) = 0.500\,V_{in} + a_3\,V_{in}{}^3 + \ldots \quad .$$

That is to say, we expect the a_2-coefficient to be zero. In the switch position marked 'asymmetric', only one diode is involved, and there's no odd-parity argument forcing a_2 to vanish.

Here's an old-fashioned way to detect non-linearity: you just take data on the function $V_{\text{out}}(V_{\text{in}})$, either by dc or ac methods.

The dc method has you put in a succession of static dc voltage levels, say in the -2 to +2-Volt range, reading these V_{in} values, and the resulting V_{out} values, with dc multimeters. Then you plot the results, and try an appropriate polynomial fit to the data. In the range (-1, +1) Volt, you might expect an adequate fit using

- for the 'none' position, $V_{\text{out}} = 0 + a_1 V_{\text{in}}$ (and with $a_1 \approx 0.48$);
- for the 'symmetric position, $V_{\text{out}} = a_1 V_{\text{in}} + a_3 V_{\text{in}}^3$ (and with $a_1 \approx 0.50$);
- for the 'asymmetric' position, $V_{\text{out}} = a_1 V_{\text{in}} + a_2 V_{\text{in}}^2 + a_3 V_{\text{in}}^3$ (and with $a_1 \approx 0.50$).

The ac method for seeing this behavior is rather qualitative, but more immediate and visual. Here, you use a triangle wave falling in the (-1, +1)-Volt range, and use it to drive your module's V_{in} and also ch. 1 of an oscilloscope. You pick a frequency of (say) 100 Hz or higher, to get a real-time display, with a full new cycle every 10 ms or less. Then you look with the 'scope's ch. 2 at the V_{out} of your module. Finally, you use the 'scope's XY-display mode to see a real-time plot of the V_{in}-across, V_{out}-upwards dependence of the function $V_{\text{out}}(V_{\text{in}})$. Now changing the module's toggle-switch position will tell you that any non-linearity is rather weak, since it takes a careful look at the XY-plot to see if this makes any changes at all. Put another way, the proportional model

$$V_{out}(V_{in}) \approx \frac{1}{2} V_{in}$$

is a pretty good zeroth-order description of the system in all three of its switch positions. But you should be able to look more closely, to see how and where the mapping changes when you do flip the switch.

By contrast to the dc and ac methods just mentioned, the goal of modern Fourier methods is to find a signature of non-linearity which is *not* just a small curvature on an otherwise nearly-linear plot, but a more dramatic manifestation which is direct evidence of non-linearity. The first such signature is frequency doubling or tripling.

Suppose that into a system modeled by the third-order expansion

$$V_{out}(V_{in}) = \frac{1}{2}V_{in} + a_2\, V_{in}{}^2 + a_3\, V_{in}{}^3 + ...$$

we inject a sinusoidal input,

$$V_{in}(t) = A\cos(2\pi\, f\, t) \quad .$$

Then at the output, under this model, we expect

$$V_{out}(t) = \frac{1}{2}A\cos(2\pi\, f\, t) + a_2(A\cos(2\pi\, f\, t))^2 + a_3(A\cos(2\pi\, f\, t))^3 \quad .$$

Trigonometric identities allow this to be written as

$$V_{out}(t) = \frac{A}{2}\cos(2\pi\, f\, t) + a_2\, A^2 \cdot \frac{1}{2}[+\cos(2\pi \cdot 2f \cdot t)] + a_3A^3 \cdot \frac{1}{4}[3\cos(2\pi\, f\, t) + \cos(2\pi \cdot 3f \cdot t)] \quad ,$$

Separating terms by their distinct frequency (because is just how they will be separated by the process of Fourier analysis), we find

$$V_{out}(t) = \frac{a_2\, A^2}{2} + (\frac{A}{2} + \frac{3a_3A^3}{4})\cos(2\pi\, f\, t) + \frac{a_2\, A^2}{2}\cos(2\pi \cdot 2f \cdot t) + \frac{a_3A^3}{4}\cos(2\pi \cdot 3f \cdot t) \quad .$$

The novelty is an output with a frequency-2*f* term (if $a_2 \neq 0$), and a frequency-3*f* term (if $a_3 \neq 0$). Sure enough, non-linearity can lead to frequency doubling or tripling.

[This sort of behavior is very dramatic in optical contexts. If we have a laser beam of say 100 mW = 10^{-1} W of power, at the eye-invisible wavelength of 1064 nm, it's easy to confirm that there is not even a nW = 10^{-9} W of any green-light content in the beam. (Or if there is any such, it's easy to filter it away.) Then if we send that beam into a frequency-doubling crystal of an optically-nonlinear material, the presence in the output beam of even 1 μW = 10^{-6} W of 532-nm green light will be glaringly obvious, and enormous compared to any <<10^{-9} W previous background of green light.]

In the case of electronic systems, it's not always so easy. If we're seeking 2*f*- or 3*f*-content in a $V_{\text{out}}(t)$ signal, we need to worry if that is due to non-linearity in our module-under-test, or if instead it *might already have been present* in the $V_{\text{in}}(t)$ signal. That latter case would be due to harmonic distortion in the source of the original frequency-f waveform, and for many generators, this will already be present at the level of 10^{-3} of the amplitude of the drive waveform.

That problem is what motivates a new and truly sensitive test for non-linearity. Into a system suspected on non-linearity, we inject a ‘two-tone’ signal, of the form

$$V_{in}(t) = A_1 \cos(2\pi\ f_1\, t) + A_2 \cos(2\pi\ f_2\, t) \quad .$$

There might well be harmonic distortion in the source, so that weak terms of frequencies $2f_1$, $3f_1$, . . . and $2f_2$, $3f_2$, . . . might also be entering our device-under-test. More terms of this character will be generated if the device-under-test does exhibit non-linearity. But the novelty, the sensitive test for non-linearity in the device-under-test, will be the *additional* presence of terms at the novel frequencies such as f_1+f_2, $|f_1-f_2|$, $2f_1+f_2$, f_1+2f_2, and so on. Because they involve two (independently-variable) frequencies, these terms are not likely to be present in the output of either generator #1 or #2 involved in the two-tone source, and (by choice of f_1 and f_2) they can be arranged to fall at novel (and empty) locations on the frequency scale. Thus they can be detected, in a background-free way, by Fourier analysis of the $V_{\text{out}}(t)$ signal.

We’ll work out the response of a quadratically non-linear system, specified by the model

$$V_{out}(V_{in}) = \frac{1}{2} V_{in} + a_2\, V_{in}^{\ 2} \quad ,$$

when it is driven by the two-tone signal above. (We leave as an exercise for the reader the response of a *cubically*-nonlinear system to the same excitation.) We find

$$V_{out}(t) = \frac{A_1}{2}\cos(2\pi\ f_1\, t) + \frac{A_2}{2}\cos(2\pi\ f_2\, t) +$$
$$a_2\ \{\frac{A_1^{\ 2}}{2}[1+\cos(2\pi\cdot 2 f_1\cdot t)] + \frac{A_2^{\ 2}}{2}[1+\cos(2\pi\cdot 2 f_2\cdot t)] + 2A_1A_2\cdot\frac{1}{2}[\cos(2\pi|\ f_1 - f_2\,|\,t) - \cos(2\pi(\ f_1 + f_2)t]\} \quad .$$

The function $V_{\text{out}}(t)$ is dominated by the main terms of frequency $f_1$and f_2. It also includes some dc terms, and terms at the doubled frequencies $2f_1$ and $2f_2$. Thus far, there is nothing novel; we are looking at the response to a sum (the two-tone signal), and the terms already mentioned are just the sum of the responses to the two tones individually. But the novelty is the presence of ‘cross terms’, terms which arise *only* because of the joint presence of the two tones. These occur at the otherwise ‘dark frequencies’ $|f_1-f_2|$ and f_1+f_2, the difference and sum frequencies. Each of these should appear with amplitude $a_2 A_1 A_2$, jointly proportional to amplitudes A_1 and A_2, and to the coefficient of quadratic non-linearity a_2.

Testing one module for non-linearity

The SR770 is equipped to make this sort of non-linearity test very easy, because its Source-Out internal generator can be configured to produce a two-tone waveform of exactly this type. The SOURCE hard-menu button gives access to softkeys which will allow the selection of 2-Tone, and its configuration. The frequencies f_1 and f_2 can be independently chosen (in the 0-100 kHz range), and the amplitudes A_1 and A_2 can also be chosen (in the 1-500 mV range). For first tests, we suggest two nearby frequencies values (such as f_1 = 21 kHz, f_2 = 23 kHz), and two equal amplitudes (such as $A_1 = A_2$ = 400 mV).

For more exercises in setting up the Two-Tone mode of the internal Source of the SR770, refer to the SRS Operating Manual, in its section 'Getting Started', at pp. 1-37 through 1-38, for instruction in using its internal source to generate a superposition of two monochromatic sinusoids.

Sending that Source-Out waveform into the input of your module, and then sending the module's output to the 770's usual Input-A, we suggest a first look using full span (0-100 kHz), and the usual AutoRange and AutoScale commands. We also suggest the use of a Log Magnitude display, perhaps with choice of 15 dB/div for the vertical scale. For the eventual detection of weak spectral peaks in the presence of strong ones, we suggest the use of the Hanning or BMH mode of the Windows menu.

Start with your module set to the 'none' position, in which you expect only linear behavior. You should see the output display spectral peaks at frequencies f_1 and f_2 only (and with what amplitudes?). Now you can use the Average function to get the cleanest possible look at the noise floor above which these peaks rise, and you can look at the suspect locations, namely $| f_1 - f_2 |$ and $f_1 + f_2$, to get a measurement (or an upper bound) on any signal you see.

That completes the 'control group', against which you will now compare the 'experimental group'. Switch your module from 'none' to 'asymmetric'. Do any new peaks appear? Indeed they do. Among them are peaks at $2f_1$ and $2f_2$; but what are the most prominent of the new peaks? Are their frequencies as you expect? What are the amplitudes of the main peaks at f_1 and f_2? And of the peaks at $| f_1 - f_2 |$ and $f_1 + f_2$? How do the amplitudes of these sum- and difference-peaks depend on your choice of A_1 and A_2?

And what about all the other peaks you see? Each is expected to be at a location $m \cdot f_1 + n \cdot f_2$, where m and n are integers (positive, negative, or zero). See if you can identify the next-most-prominent peaks you see – one method for doing so is to make a small change Δf_1 or Δf_2 in the source frequencies, and then to see how far your peak-under-investigation moves. In our model, we expect Δf_{out} to show a variation of $m \cdot \Delta f_1$ or $n \cdot \Delta f_2$, as the case may be.

Once you've seen this behavior, switch to the 'symmetric' model of the mapping $V_{out}(V_{in})$. You'll see whole families of outputs diminish (though not quite to zero, since the odd-order (anti)symmetry required for them to vanish entirely cannot be perfect). What are the dominant peaks now? (Still at f_1 and f_2). What are the most important peaks indicative of non-linearity? (If you work out the theory, you'll find, among others, terms at $2f_1 + f_2$ and $f_1 + 2f_2$ – with what coefficients? And along with what other terms?) If you see a peak at $2f_1 + f_2$, you should be able to show, in theory and by experiment, that its amplitude is proportional to $a_3 A_1{}^2 A_2$; can you confirm this? Can you extract an a_3-value from a comparison of theory and experiment?

Finally, the name 'intermodulation distortion'. We've seen non-linearity, even of the simplest kind, can lead to novel frequency combinations. In response to a two-tone input, the new frequencies which emerge may remind you of the action of a multiplier or a mixer (as in Ch. 4). It's the action of one frequency (say f_1), modulating the presence of another frequency (say f_2) in the output, thereby producing novel frequencies, which motivates the name 'intermodulation distortion'. (Equally, the presence of f_2 modulates f_1.) But who would care about this obscure effect? Audiophiles do, and for good reason.

Many sources of musical sound are periodic, so their waveforms consist of fundamental-plus-harmonics. Given the typically strong harmonic content in musically-interesting sounds, some additional harmonic distortion (due, for example, to non-linearity in an amplifier) has a surprisingly inaudible effect in the sound as perceived. But intermodulation distortion is another matter. If we have a musically-tolerable two-tone signal, it will more than likely be harmonious: that is to say, it will have f_1 and f_2 in the ratio of small integers. But intermodulation distortion can lead to the generation of new frequencies, such as the difference frequency, which are *not* in the harmonic series of either f_1 or f_2, and which might be judged to be *not* harmonious with either f_1 or f_2. Your ears, in other words, are much more sensitive to, and intolerant of, intermodulation distortion than they are to harmonic distortion.

Testing other units for non-linearity

Now that you know how to conduct these tests, you can use other modules than the Intermodulation-Distortion one for testing. You'd like to test modules which are 'meant to be linear', and these include the Summer, the Filter section, and both Amplifier sections. In each case, you are performing a 'null test', in which certain spectral peaks are predicted to be *absent* for a module behaving linearly. So the detection of certain spectral peaks is direct evidence of non-linearity, at some level. The choice of modules, and of frequencies and amplitudes for the two tones, is left up to you.

Chapter 18: Demodulation of FM signals

To appreciate this Chapter, you'll need to have worked through the formation and analysis of frequency-modulated signals using Chapter 5. You'll recall that using a voltage-controlled oscillator, you found that an external 'programming' voltage could move the frequency of that oscillator (while keeping its amplitude constant). Now we take up the inverse problem – if you were picking up a constant-amplitude but varying-frequency sine wave, how could you extract the program content which lay behind the original frequency modulation?

This Chapter will show you one example of a method for 'demodulating FM', one which will exploit your understanding in time- and frequency-domain vocabularies, and which can also be carried out using the tools of your Electronic Modules.

The system as a whole will need

- a 'transmitter', a unit which transforms a program-content waveform into a frequency-modulated carrier wave;
- a 'link' which would ordinarily be achieved with a radiated electromagnetic wave, but here will be mere conveyance through a coaxial cable; and
- a 'receiver' which will accept the FM waveform, and re-create from it a near-facsimile of the original program content.

First, the transmitter unit. The heart of generating an FM waveform is again the Voltage-Controlled Oscillator module. For reasons you'll soon see, you'll now use that VCO on its 2-10 kHz range. Monitor the output of your VCO with a 'scope and the 770, and see a staircase-version of a sinusoid emerging from it. Use the frequency-adjust knob of the VCO to get an output frequency of 6 kHz, and use the amplitude-adjust knob to give that sinusoid an amplitude of about 5 V. (Your 'scope shows you the output is not a pure sinusoid; the 770 will show you that beyond the 6-kHz fundamental, the next important Fourier components are the 9th and 11th harmonics – and you'll see why these won't matter at all. Appendix A12 tells you more about the Fourier content of 'staircase functions'.)

Now use the DC Voltage module as a knob-adjustable source of control voltages in the range (-5, +5) Volts. Send that voltage to a multimeter and to the VCO's control input, and use the 770 to monitor the VCO's frequency output, and confirm that you can use the DC Voltage knob to 'steer' the VCO's frequency output. Make some measurements to show that this voltage-to-frequency mapping has a sensitivity of about 0.15 kHz per Volt. So now any control voltage in the range (0 ± 3) Volts will map the VCO's output frequency into the range (6.0 ± 0.5) kHz.

Temporarily set up an external generator to produce triangle waves of (say) 220-Hz frequency and 3-V amplitude. If you send that waveform (*instead* of the knob-adjustable DC Supply) into the control input of the VCO, then its output should be frequency modulated, with

- center, or 'carrier', frequency f_c = 6.0 kHz,
- peak frequency deviation δf = 0.5 kHz,
- modulation frequency f_m = 0.22 kHz

(so the 'modulation index' $\beta = \delta f / f_m \approx 2.3$).

As a result, the VCO output should show up on the 770's spectral display with a central carrier peak at 6.0 kHz, and a family of multiple sidebands, spaced by 0.22 kHz, to either side of the carrier.

That completes the transmitter unit. The 'link' will be just a BNC cable carrying this varying-frequency, fixed-amplitude sine-like wave to a 'distant' receiver. In practice, you'll build your receiver unit from more pieces of your Electronic Modules, so there's no distance to speak of – but use a long, or a distinctive, cable for this link, just to make the point that your 'link' cable will serve as the *sole* connection between 'transmitter' and 'receiver' parts of your system.

Now for the 'receiver' or demodulator. All that it gets, via the link, is a signal of fixed amplitude *A* and (relatively slowly) varying frequency f_{var}, of sinusoidal form

$$V_{rec}(t) = A_{fixed}\cos(2\pi\ f_{var}\ t) \quad .$$

As we've set it up, the amplitude is fixed near 5 V, and the frequency varies in the range (6.0 ± 0.5) kHz. How can we electronically deduce the changes in frequency? We'll use the LCR module to build a 'frequency discriminator'.

How an LCR circuit can be a frequency discriminator

The LCR circuit can achieve our goal because its magnitude *and phase* responses both vary, as functions of frequency, in the vicinity of its resonant frequency f_0. (Chapter 7 deduces these responses as parts of the transfer function of the LCR system.) In this use of the LCR system, we'll have the transmitter's carrier frequency f_c lying centered on the LCR's system's resonant frequency, and that brings up a problem:

The magnitude response $M(f)$ of a resonant circuit is peaked at $f = f_0$, and as a result $M(f)$ has only *second*-order variation when f departs from f_0. Not only does this make the change in response small, it also ensures that changes in f with $f < f_0$, and with $f > f_0$ both give decreases in the magnitude of the response. But we want changes that are *linear*-in-$(f - f_0)$, not quadratic-in-$(f - f_0)$, and we'll get them from the *phase* response $\varphi(f)$ of the LCR circuit. Recall that the phase shift for an LCR circuit has the behavior

$\varphi(f) \approx 0$	for $f \ll f_0$;
$\varphi(f) = \pi/4$	for $f = f_0 - f_0/(2Q)$;
$\varphi(f) = \pi/2$	for $f = f_0$;
$\varphi(f) = 3\pi/4$	for $f = f_0 + f_0/(2Q)$;
$\varphi(f) \approx \pi$	for $f \gg f_0$.

In particular, recall that the phase shift φ(*f*) is a *linear* function of frequency in the vicinity of $f = f_0$. We can write

$$\varphi(f) \approx \frac{\pi}{2} + k\,(f - f_0)$$

to express the linear variation of phase shift in the vicinity of f_0. You could work out the value of the constant *k*, but clearly it's of order

$$k = \frac{d\varphi}{df} \approx \frac{\Delta\varphi}{\Delta f} = \frac{3\pi/4 - \pi/4}{f_0(1+1/2Q) - f_0(1-1/2Q)} = Q\frac{\pi/2}{f_0} \quad .$$

Now how can we use this linear-in-frequency variation? Suppose that for our LCR module we have an input waveform

$$V_{in}(t) = A\cos(2\pi\ f\ t) \quad .$$

and it produces a phase-shifted output waveform with modified magnitude,

$$V_{out}(t) = A \cdot M(f)\cos(2\pi\ f\ t - \varphi(f)) \quad .$$

If we send these two waveforms to our Multiplier module, we'll get an output of

$$V_{mult}(t) = \frac{V_{in}(t)\cdot V_{out}(t)}{10\,V} = \frac{A\cdot AM(f)}{10\,V}\cos(2\pi\ f\ t)\cos(2\pi\ f\ t - \varphi(f))$$

$$= \frac{A^2}{10\,V}M(f)\cdot\frac{1}{2}[\,\cos(\varphi(f)) - \cos(2\pi\cdot 2f\cdot t - \varphi(f))\,] \quad .$$

For our application, the term at frequency 2*f* lies at about 12 kHz, so it's easily filtered away (using, say, the 1-kHz low-pass Filter module). What's left is

$$\langle V_{mult}(t)\rangle = \frac{A^2}{20\,V}M(f)\cos(\frac{\pi}{2} + k(f - f_0)) = \frac{A^2}{20\,V}M(f)\sin(k(f - f_0)) \quad ,$$

where here we've used the <…> notation to denote the low-pass-filtered version. And now if the frequency deviation $f - f_0$ is not too large, we can treat *M*(*f*) as a constant, and also use the small-argument approximation in the sine function, and get

$$\langle V_{mult}(t)\rangle \approx \frac{A^2}{20\,V}M(f_0)\cdot k(f - f_0) \quad .$$

That is to say, we have an electronic combination which accepts a variable-frequency, fixed-amplitude sinusoidal input, and produces a low-frequency output which is approximately a linear function of the frequency deviation $f - f_0$.

Building and testing an FM demodulator

Now for the details of building such a demodulator, using Modules available to you. Let us suggest that you conduct this exercise using an increment-and-test approach, in which you add one module at a time to an emerging system, and that you test each new module for functionality as you add it.

So start with a working 'transmitter', controlled by the DC Supply's knob changing the frequency of your VCO. The test is that your 'link' cable should show, on the 770, a spectral peak which you can move about, in the 5.5-to-6.5 kHz range, using your knob.

That link delivers a quasi-sinusoid of fixed amplitude (about 5 V) to the receiver you are about to build. You want that received signal to drive the multiplier, and *also* the LCR module. But to deal with the 1-Ω input impedance of the LCR module, we suggest you use the Power Audio Amplifier module, set to gain about 1, to supply the current-hungry input of the LCR circuit.

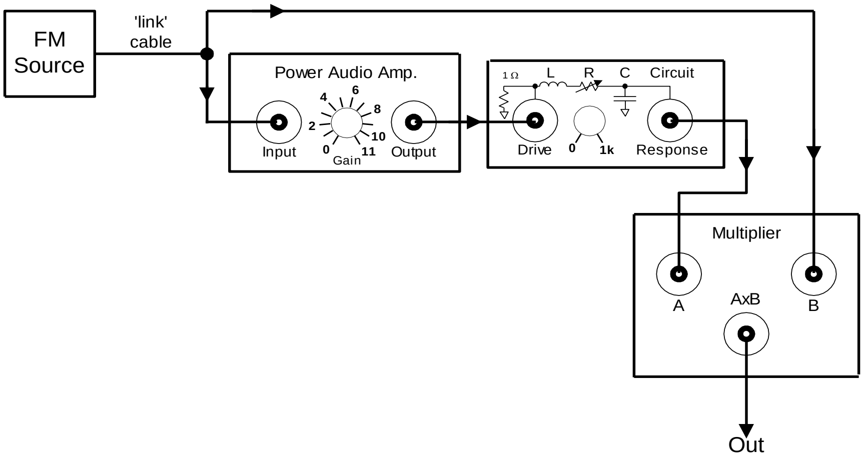


Fig. 18.1: The 'link' cable arriving at the 'receiver', and there driving both the Multiplier and (in a power-boosted fashion) the LCR module.

The LCR circuit has a fixed resonant frequency near 6 kHz (which is what has motivated this value as the choice of the carrier frequency for your transmitter). You can adjust its damping resistor *R*, and hence its Q; we suggest that you start with this resistor set to the middle of its range. (That gives the LCR circuit a Q-value of about 5, so its FWHM is about 6 kHz/5 = 1.2 kHz, so its half-power points are near 6.0 ± 0.6 kHz, about the range of frequency excursions you'll be using.)

An appropriate test for the receiver modules thus far is to use the knob-adjust of the DC Supply as a way to change the transmitted frequency, and to watch with a 'scope the output of the LCR module as you do so. You should see a (cleaned-up) sinusoid, whose frequency will vary, and whose magnitude will peak, as you 'dial through resonance'. Adjust the gain of the Power Amp so that, at resonance, you get an amplitude of about 6 Volts here.

Next, send this LCR output to the second (the other) input of the Multiplier. Now use your ‘scope to watch the Multiplier’s output, which should display a waveform of curious shape. There is lots of 12-kHz content here to distract you, but look instead at the average value of this waveform – it should vary as you dial the DC Supply knob. (Why?)

If that Multiplier output is present, convey it to the Filter module’s input. We suggest you start with the filter set to a 1-kHz corner frequency, and a Q-value of 1. Now monitor the Low-Pass output of the Filter with a ‘scope; it should be nearly free of 12-kHz content. But the near-dc level displayed by the ‘scope should vary, in a nearly linear fashion, with the DC Supply knob which is controlling the transmitter frequency. (Here’s a place where you *must* have the ‘scope’s input set to dc-coupling!)

You now have a system which can be drawn schematically as

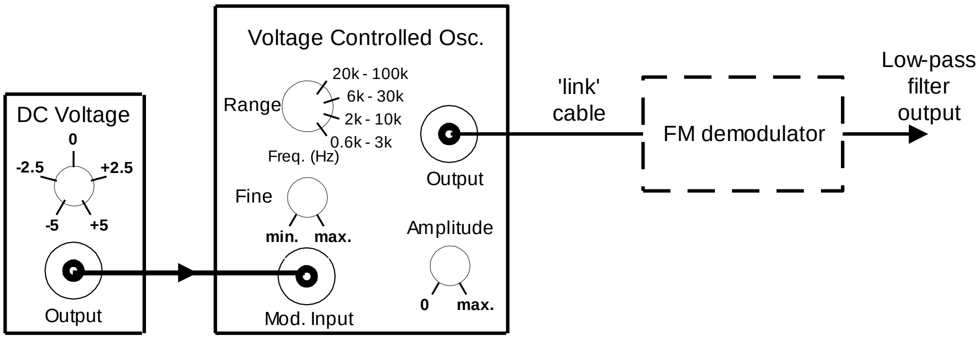


Fig. 18.2: A block diagram of a ‘dc test’ of the transmitter, link, and receiver sub-sections

The end-to-end sensitivity of this system is not very interesting; you might find that a ΔV(at input) of 1 Volt causes a Δf (in the link) of 0.15 kHz, and that in turn creates a ΔV(at output) of order 0.1 V. The important thing is that a change-at-left gives a proportional change-at-right, via the intermediary of the link. Note too that the input at left, and output at right, are both dc values, but the link is carrying a pure-ac waveform of fixed amplitude – so how *is* the knob information being encoded and sent?

If this dc-test works, it’s time to replace the DC Supply knob with a more interesting drive source. We suggest a triangle wave from an external waveform generator, set to be symmetrical around zero, of amplitude 2 or 3 Volts, and frequency of 100 or 200 Hz. Send that to your ‘scope’s ch.1 and also to the VCO control input. Now send your ‘receiver’ output (ie. the Low-Pass output of the Filter) to the ‘scope’s ch. 2.

You should be seeing what you’re sending on ch. 1, and what you’ve received and demodulated on ch. 2. There are many tests you should try to persuade yourself this system is working:

- Disconnect the ‘link’ cable. (Result: a flat-line dead response on ch. 2)
- Change the frequency of the external generator. (Result: parallel change in the frequency displayed on ch. 2)
- Change the amplitude of the generator. (Result: parallel change in the amplitude displayed on ch. 2 – but with failing proportionality beyond about 3-V amplitude)

- Change the dc offset of the generator. (Result: you should see both ch.1 and ch.2 signals undergo parallel changes on the 'scope)
- Change the waveform of the generator, say from triangle, to sine, to square. (Result: parallel changes in the ch. 2 waveform, but with indications – especially from the square wave – of limited frequency response of the system)
- Change your source from the signal generator to something more interesting, such a speech waveforms recorded onto some audio-player. All you want is a waveform lying in voltage range ±2 Volts, and with frequency content concentrated in the dc-to-1 kHz range. (Result: tolerably understandable transmission of speech signals – so now try some musical content instead)
- Or, as an alternative to an audio waveform, use the Chaos module as a source of waveforms. We suggest the 'medium' speed setting, and the X-channel output, with the 10-turn dial set to 2 or 3 turns, as a source of a random-looking waveform which you can use in a test for faithful transmission by your FM link.
- Or (while monitoring the FM-transmission of ±1-Volt, 100-Hz square waves) change the Q of your LCR system. (Result: higher Q gives higher sensitivity, but creates a smaller range of useable frequency deviation)

Finally, when you're persuaded this link really works as a method for transmission, in encoded form, of your choice of program content, you can do some further projects:

- With a generator giving low-amplitude, 200-Hz sine waves as source, use the 770 to monitor the Fourier spectrum of the signal flowing in the link cable. Using a span of 0 - 12.5 kHz, you should see a characteristic FM spectrum, displaying the carrier and multiple sidebands. Now raise the amplitude of the 200-Hz program content progressively, and you'll see the family of sidebands growing in characteristic ways. Once you've seen and understood this, try raising the *frequency* of your program content, and see how that affects the spectrum.
- With some speech or music waveform providing the program content, try to 'listen to the link'. That is, find a way to get the FM waveform (of about 5-V amplitude, centered at 6 kHz) that is flowing in the link cable, to drive a speaker or headphones. What will you hear? For quiescent moments of program content, a pure 6-kHz sine wave; for non-quiet moments, a very strange audio signal indeed. Once you've heard, in live audio, the screeching chaos of the waveform that is coming along the link cable, you can be the more amazed that the demodulator can extract, out of all that, a facsimile of the original program-content waveform. Better still, you can be gratified that you can understand, and model, what is going on at every step along the way.
- Now that you know how this is done, you could supply your *own* LCR module with a resonant frequency of your choice. If you pick 50 or 80 kHz, you can 'broadcast', ie. send along the link cable, a higher-frequency signal than heretofore. That, in turn, will allow you to modulate at higher frequencies, to use a higher corner frequency for the low-pass filter, and to get higher-fidelity transmission of musical sound. You'll have to decide what *L*, *C*, and *R*-values to use, and you'll be able to avoid the use of that 1-Ω input resistor. As a result, you won't need the Power Audio Amplifier module, and you could devote that module instead to driving the speaker.

- Independent of the previous idea, you might appreciate that everything depends on the frequency (ie. the location of zero-crossings), and not on the amplitude, of the ≈6-kHz signal in the link cable. So, to what degree is this demodulation technique independent of the *waveform* of the linking signal? To find out, you could try a 'digitized' version of the 6-kHz frequency-modulated waveform: you could convert that staircase-sinusoid into a logic signal, which is logic-High when the sinusoid is positive, and logic-Low when the sinusoid in negative. (A comparator chip is the tool you'd need.) Then you could find out if the receiver system you've built will still work, and (in the process) learn why a 'limiter' function is built into commercial FM receivers.

Appendix A1. How a Fourier Transform is defined and computed

This Appendix describes, in the simplest case, exactly what a digital Fourier analyzer such as the SR770 really does with the input voltage, so as to produce the frequency-spectrum display that you've seen. We confine the discussion to the 770's simplest mode, in which the analyzer's 'full span' of 100 kHz is used, and in which the data-acquisition time is 4.00 ms.

The 770 uses a fixed 'sampling rate', and therefore does *not* read the input voltage continuously, but rather at fixed intervals separated in time by $\Delta t \equiv 4$ ms/1024 = 3.90625 μs. At the set of times $\{t_0, t_0 + \Delta t, t_0 + 2\ \Delta t, \ldots t_0 + 1023\ \Delta t\}$, the device forms a 16-bit digital representation of the input voltage $V(t)$, giving a list of 1024 16-bit numbers. Everything that follows depends (only) on these 'voltage samples', and is performed in digital electronics by an internal computer.

If *you* had the list of voltage samples $\{V(t_0), V(t_0 + \Delta t), V(t_0 + 2\ \Delta t), \ldots V(t_0 + 1023\ \Delta t)\}$, you could imagine making a least-squares fit of this data set to a <u>constant</u>. This might be a poor model for the actual data, but the point is that there exists an algorithm which would return one parameter value, fixing it by a formula (namely, the mean value) whose input consists of the list of sampled voltages. Or, you could get a (somewhat better) fit of the same data by a <u>linear</u> model, in this case getting *two* output parameters (intercept and slope), both of them fixed by formulae whose input (again) consists of the list of sampled voltages. Naturally, you could also try fitting the data to a higher-order <u>polynomial</u>, and achieve a still-better fit, at the cost (or, with the return) of more fitting-parameter values. In fact, if you allowed yourself a model with a full 1024 free parameters, you could perhaps fit each and every sampled point <u>exactly</u>. That is to say, your new 'best fit' would be a *perfect* fit. For many reasons, a **Fourier expansion** is much more suited than a high-order polynomial as a fitting function, so in practice, the data is fit to the model

$$V_F(t) = \sum_{n=1}^{512} [C_n \cos(2\pi\ n\ f_1\ t) + S_n \sin(2\pi\ n\ f_1\ t)] \quad ,$$

which has 512 cosine coefficients C_n, and 512 sine coefficients S_n, as its adjustable parameters. The frequency f_1 is *not* adjusted, but is instead fixed at 250 Hz = 1/(4 ms), so that each trigonometric term, of frequency $n \cdot f_1$, goes through exactly n full cycles during the 4-ms data-acquisition window. Ignoring the fine point that this model lacks a dc-average-value term C_0, this model is capable of matching the sampled data *exactly* at each of the time-sampling points.

Now just as for the one-parameter fit to an average value, or the two-parameter fit to a straight-line model, so here too there exists an explicit formula which gives, for each n-value, the C_n- or S_n-value in terms of the list of sampled voltages. Those formulae are called the 'discrete Fourier transform', described in Appendix A11. The formulae involve multiplying the data-list entries by constants, and summing. For each of 1024 coefficients to compute, the sum involves 1024 multiplications, so it would seem that $(1024)^2$ multiplications are needed to perform the whole calculation.

But a very clever algorithm, popularized by Cooley and Tukey, re-arranges that calculation so that in fact only about $1024 \cdot \log_2(1024)$ multiplications are needed. That cuts the number of (computationally expensive) multiplications from $(1024)^2 \approx 10^6$ to only about 10^4, and this 100-fold speed-up is what makes the 'fast Fourier transform' algorithm so justifiably famous.

Once the coefficients are thereby calculated, it's easy to form the combinations, for each *n*, of

$$M_n = (C_n{}^2 + S_n{}^2)^{1/2} \quad \text{and} \quad \varphi_n = \tan^{-1}(S_n / C_n) \quad ,$$

which can be taken as the magnitude M_n and the phase φ_n of the *n*th Fourier term, ie. the term at frequency $n\,f_1$. In the case at hand, those frequency components come at f_1 = 250 Hz, 2 f_1 = 500 Hz, . . . 400 f_1 = 100 kHz, and so on. The 770 displays only the first 400 (magnitude) coefficients; the dc average value is not computed or displayed, and neither is the extra information available between 100 and 128 kHz.

Notice that the model $V_\mathrm{F}(t)$, which fits the sampled data-points perfectly in the 4-ms window of its acquisition, is a periodic-in-time model, so that the model not only fits the data-set perfectly, it would also continue onward and repeat periodically (as the actual voltage $V(t)$ might not) in each subsequent 4-ms window.

Notice also that while the model $V_\mathrm{F}(t)$ matches the actual voltage $V(t)$ at each of the sampling points of ≈3.9-μs separation, there is no guarantee in general that it would *also* match the actual voltage at times intermediate between the sampling points. But this matching can be guaranteed, by Shannon's sampling theorem, *provided* that the frequency content of the actual voltage waveform is confined to lie in the interval (0 to 128 kHz), ie. that it lies between dc and half the sampling frequency. [See Appendix A12 or A13 for two illustrations of why this suffices.] This is actually assured in the 770, because the input voltage applied at the front panel is first filtered, so as to preserve unmodified all its frequency content in the 0-to-100 kHz range, but to filter *out* all its content above 128 kHz, before it is sent to the digitizer. So the voltage actually sampled and digitized *does* meet the conditions of the sampling theorem, and the result is that the frequency spectrum displayed is fully faithful to the frequency content of the voltage in the 0-to-100 kHz region.

That input filter is called an anti-aliasing filter, and it is executed by real-time all-analog electronic hardware. It is the *absence* of such anti-aliasing filters in generic voltage-sampling apparatus, such as oscilloscopes, which gives rise to various pathologies called 'aliasing' in the spectral displays they can be arranged to create.