PHYS 331: Junior Physics Laboratory I

Notes on Noise Reduction

When setting out to make a measurement one often finds that the signal, the quantity we want to see, is masked by noise, which is anything that interferes with seeing the signal. Maximizing the signal and minimizing the effects of noise then become the goals of the experimenter. To reach these goals we must understand the nature of the signal, the possible sources of noise and the possible strategies for extracting the desired quantity from a noisy environment..

Figure 1 shows a very generalized experimental situation. The system could be electrical, mechanical or biological, and includes all important environmental variables such as temperature, pressure or magnetic field. The excitation is what evokes the response we wish to measure. It might be an applied voltage, a light beam or a mechanical vibration, depending on the situation. The system response to the excitation, along with some noise, is converted to a measurable form by a transducer, which may add further noise. The transducer could be something as simple as a mechanical pointer to register deflection, but in modern practice it almost always converts the system response to an electrical signal for recording and analysis.

In some experiments it is essential to recover the full time variation of the response, for example the time-dependent fluorescence due to excitation of a chemical reaction with a short laser pulse. It is then necessary to record the transducer output as a function of time, perhaps repetitively, and process the output to extract the signal of interest while minimizing noise contributions. Signal averaging works well when the process can be repeated, and we will examine the method in some detail.

Alternatively, the signal of interest may be essentially static. Consider, for example, measuring the optical absorption of a beam of light traversing a liquid as a function of temperature or optical frequency. Any time variation is under the experimenter's control and does not convey additional information, so it is only necessary to extract the essentially steady signal from whatever noise is present. For this situation, drifts and instabilities in the transducer are often important, along with noise introduced in the experimental system. The usual strategies to deal with this combination of problems include chopping, filtering and, ultimately, lock-in detection, all of which we will consider.

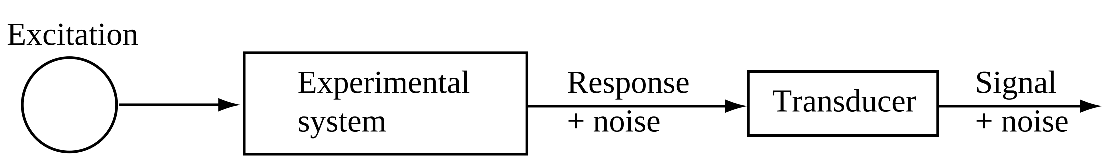

Fig 1. Schematic of a generalized experiment.

A. Sources and characteristics of noise

Noise can be roughly categorized as follows: "Interference" is at a definite frequency. Examples would be 60 Hz fields from power lines, high frequency fields from nearby radio or TV transmitters, and scattered laser light in an optics set up. Broad band, or "white noise" is more or less uniform over a wide range of frequencies. It can arise from multiple sources which blend together to give a uniform appearance, or, ultimately and inevitably, from the thermal vibrations of the apparatus. "Flicker" or "1/f noise" is a broad band noise that increases with decreasing frequency. It is quite commonly observed, but the level depends on the system. For example, all electrical resistors of the same value exhibit the same level of white noise but the additional 1/f contribution depends on the material. The overall situation is summarized in Fig. 2, which shows a typical noise spectrum and some interference sources. (The appendix to these notes includes a brief discussion of power spectra.)

The type of noise afflicting the measurement, as well as the measurement requirements, will suggest an appropriate strategy. First one should attempt to strengthen the signal, if possible. This might involve using a stronger optical source, larger telescope, bigger force or whatever else might be possible and appropriate. After that one can try to reduce external interference by putting the apparatus in a metal box to block high frequency disturbances, by proper grounding to minimize power-line pickup, by cleaning lenses to reduce scattered light, and so on. Once the

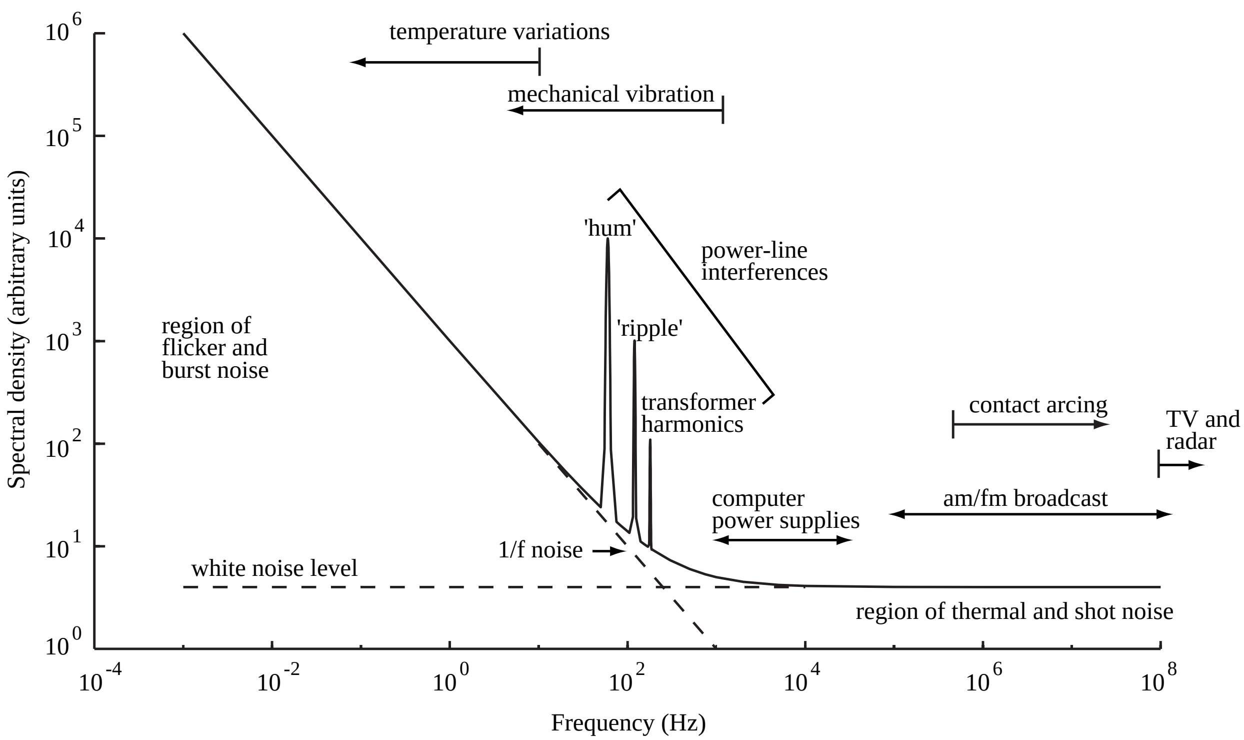

Fig. 2 Schematic spectrum of noise and interference.

signal has been enhanced and shielded as much as possible, it is appropriate to turn to signal recovery procedures.

B. Signal averaging

Signal averaging exploits the fact that if one makes a measurement many times the signal part will tend to accumulate but the noise will be irregular and tend to cancel itself. More formally, the standard deviation of the mean of N measurements is smaller by a factor of $\sqrt{N}$ than the standard deviation of a single measurement. This implies that if we compute the average of many samples of a noisy signal, we will reduce the fluctuations and leave the desired signal visible. There are, of course, a number of complications and limitations in practice.

We model the averaging process by assuming a signal voltage $V_S(t)$ which is contaminated with a noise voltage $V_n(t)$ that is comparable or larger than V_s. The voltage actually measured is then the sum of these two parts

$$V(t) = V_S(t) + V_n(t) \qquad (1)$$

If we record $V(t)$ over some time interval we have a record which contains information about the time variation of V_S. The crucial point is to average many such records of $V(t)$ so that the noise, which is equally likely to be positive or negative, tends to cancel out while the signal builds up. This obviously requires that we start each measurement at the same relative point in the signal, a requirement we will consider when we implement the averaging.

To calculate the actual noise reduction we need to compute the standard deviation of the average of $V(t)$ after N records have been averaged. For convenience, we assume that the noise is characterized by $\overline{V}_n = 0$ and standard deviation σ_n. If the noise mean is not zero, we can simply include $\overline{V}_n$ as part of the signal and subtract it later. The signal is assumed to repeat exactly from trial to trial, so $\sigma_S = 0$. With these assumptions,

$$\overline{V}(t) = V_s(t) \quad \text{and} \quad \sigma_{\bar{v}} = \sigma_n / \sqrt{N} \qquad (2)$$

which is the expected outcome. This result is often phrased in terms of a voltage signal to noise ratio, given by

$$\frac{\overline{V}}{\sigma_{\bar{v}}} = \frac{V_s}{\sigma_n}\sqrt{N} \qquad (3)$$

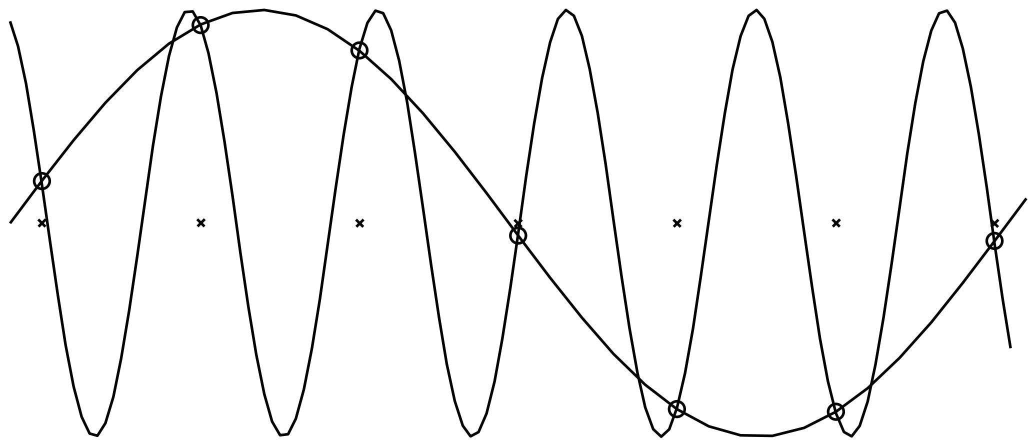

Fig. 3 Sampling of sinusoidal waves. Note that there is no way to distinguish between the two sine waves on the basis of the regularly-spaced samples.

This relation is the basis for the common statement that averaging N records improves the original signal/noise ratio by a factor of $\sqrt{N}$. Note that this improvement is bought at the expense of measurement time. Specifically, the signal to noise ratio improves only as the square root of time spent averaging, a result we will see again.

The averaging process could be done using a continuous record of $V(t)$, but it is more common to sample the input with an analog to digital converter at regular intervals Δt so that the required computations can be done digitally. Unfortunately the sampling process introduces the possibility of aliasing, as illustrated in Fig. 3. Evidently input frequencies from zero to $1/(2\Delta t)$ are all distinguishable, but if higher frequencies are present they will be confused with some lower frequency. The highest distinguishable frequency is called the Nyquist frequency, $1/(2\Delta t)$. Experimentally, it is important to avoid aliasing by low-pass filtering the input before sampling and choosing Δt small enough that the Nyquist frequency is well above the pass limit.

Several factors limit the degree of S/N enhancement attainable by averaging. For example, digital systems measure inputs to a specified accuracy, perhaps 12-14 bits, and do arithmetic to a similar fixed precision. Comparable effects occur in analog averaging schemes. Ultimately this must limit the accuracy of the computed average. There may also be time or voltage-dependent drifts in the instrument, which have the effect of adding spurious variation to the true signal. It is also possible for the signal itself to change with time, perhaps due to internal changes in the system under study or the slow variation of external conditions. Whatever the cause, one always finds that the gain in S/N does not increase indefinitely with averaging time.

C. Filtering and modulation

In the case of a quasi-static signal, where we do not need to measure the actual waveform, it may be possible to use a frequency-selective filter to suppress noise. Consider again the power spectrum of noise shown in Fig. 2. If the desired signal occupies a fairly narrow range of frequencies somewhere in this region we could improve the situation by measuring the voltage only in that range, and throwing out fluctuations at all other frequencies. For example, in a laser scattering experiment we might use an optical filter at the laser frequency to reject room light and other disturbances. Any noise contributions that happen to fall in the same frequency range as the signal will still interfere, but we will have removed a great deal of other noise.

The filters used are commonly either low-pass or band-pass. A low-pass filter rejects all inputs above a certain frequency, and is therefore most useful when the signal is essentially DC. A DC voltmeter is a good example of this approach, in that it integrates the input for some finite time, effectively canceling any input component that changes sign during the integration period. Alternatively, if the signal has a known periodicity it may be possible to design a filter to reject all inputs outside of a small band of frequencies near the signal, as in the laser example above.

If the signal is inherently static it may be advantageous to impose a periodic variation. Referring again to Fig. 2, the spectral density of noise tends to decrease with increasing frequency, so measurements made at higher frequencies are inherently less noisy. The output of DC amplifiers also tends to drift with temperature and other environmental variables, adding further uncertainty to the measurement. These problems can be addressed by interrupting or modulating the excitation so that the signal portion of the transducer output has a definite frequency. In the laser light scattering experiment the modulation might be done by interrupting the laser beam with a toothed wheel. The electrical output of the detector can then be filtered to minimize noise and other disturbances added by the detector or amplifiers.

The improvement in signal to noise ratio due to filtering can be easily quantified for the simple situation shown in Fig. 4. The input noise spectrum is assumed to be flat or "white" from zero frequency up to the bandwidth B_I. The signal *s(t)* is visible as a bump on the total input spectrum, and has a mean-square intensity $\overline{s^2(t)}$. The total mean-square intensity of the noise is just the area under the spectral density curve, or $P_N B_I$, so the mean-square signal to noise ratio at the input is

$$SNR_I = \frac{\overline{s^2(t)}}{P_N B_I} \tag{4}$$

By design, the filter transmits all the signal but attenuates strongly over most of the noise bandwidth. If the effective bandwidth of the filter is B_O the mean-square signal to noise ratio at the output will be

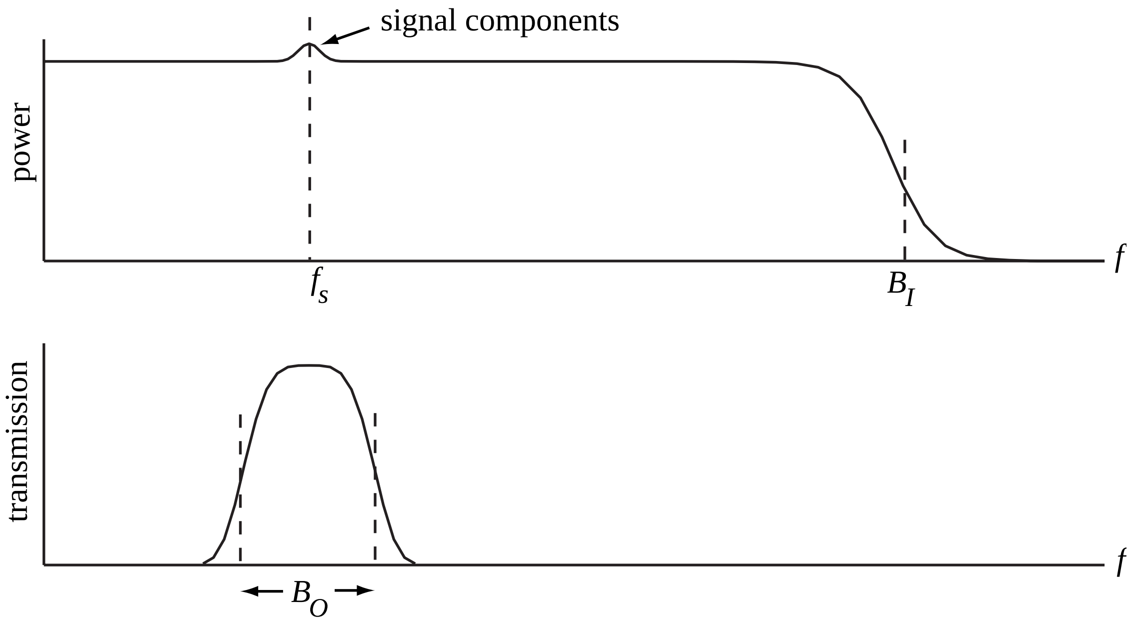

Fig. 4 Power spectrum of signal and noise (above), and the percent-transmission characteristic of a bandpass filter (below).

$$SNR_O = \frac{\overline{s^2(t)}}{P_N B_O} \tag{5}$$

and the improvement in mean-square *SNR* is just the ratio of bandwidths

$$\frac{SNR_O}{SNR_I} = \frac{B_I}{B_O} \tag{6}$$

Instruments are usually designed to display rms voltage, rather than mean-square voltage, so the improvement read on the output meter will be the square-root of this value.

As always, there are some practical limits. For maximum enhancement it is obviously desirable to make the filter bandwidth as small as possible, but it must be left broad enough to pass all signal frequencies. Depending on the implementation, very narrow-band filters may also drift off the desired frequency, with disastrous consequences for the signal. In a favorable situation with broad-band noise, as assumed here, it might be possible to gain a factor of ten or so in rms *SNR* by filtering.

D. Lock-in detection

For bandwidth minimization to be effective we require that the signal occur within a restricted range of frequencies, preferably well above 1/f noise and away from interference

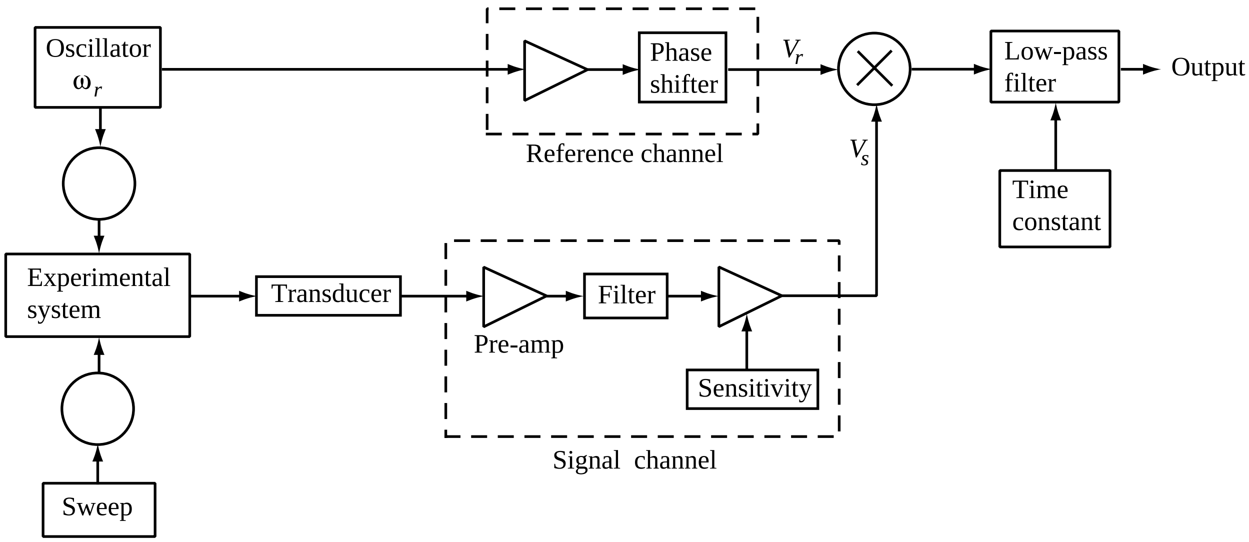

Fig. 5 Block diagram of a phase-sensitive detection system connected to an experiment.

sources. We then insert some sort of filter, most commonly electronic, before our measuring apparatus to pass the signal frequency range and block other frequencies. The lock-in detector is a particularly ingenious implementation of these principles because it allows us to create a very narrow-band filter which will automatically follow small changes in modulating frequency.

The basic principle behind the lock-in detector is shown in Fig. 5. The experimental system is our apparatus and sample, as before. We want to measure some property of this system as a function of an external parameter like temperature or magnetic field which we can slowly 'sweep' over a desired range. An external oscillator is used to modulate a system parameter in such a way that the quantity we wish to measure varies at frequency ω_r. This could be the same parameter that is being varied slowly, such as the wavelength of the incident light in our optics example, or an independent parameter such as the intensity of the light. The object in either case is to impose a variation at ω_r on the desired signal, without modulating the noise that you want to eliminate.

The response of the system is converted to a varying voltage with an appropriate transducer and applied to the phase detector as V_s, which includes noise from the experiment and transducer. The phase detector multiplies this voltage by a reference voltage V_r at frequency ω_r, producing a product voltage V_rV_s. Now comes the crucial point: The desired signal and the reference voltage are at the same frequency and are always in phase, so a Fourier analysis of their product contains a constant term. Other inputs, like noise, will be at different frequencies and phases. Fourier analysis of their products will disclose time-varying components, but almost no DC part. The multiplier voltage is filtered by the low-pass circuit, which passes the DC signal voltage but suppresses the time-varying noise voltages to yield the final output V_o. (For completeness, we note that the slow sweep must be slow enough that all of its Fourier components can get through the low pass filter. Otherwise the desired signal will be partially removed along with the noise.)

There are several advantages to lock-in detection. First, by modulating the right parameter we distinguish the quantity we wish to measure from other fluctuations which are not sensitive to the varying parameter. Second, the desired response is now at a controllable, relatively high frequency. Other disturbances, such as power line interference and mechanical vibrations tend to occur at specific lower frequencies which we can avoid. Finally, the combination of the phase sensitive detector and low-pass filter allow us to work with extremely small bandwidth. Even if we could directly construct a filter with similar bandwidth it would be formidably difficult to keep it centered on the oscillator frequency because of drifts in component values.

To do a quantitative analysis of the multiplier - filter combination we imagine Fourier analyzing V_s and V_r to obtain their frequency components. The reference voltage is a sine wave, so it has only the single component

$$v_r \sin(\omega_r t + \phi_r) \tag{7}$$

while V_s has significant amplitude over a broad range

$$v_s(\omega)\sin(\omega t + \phi_s) \tag{8}$$

The components at and near ω_r represent the desired signal plus any noise that happens to be at the same frequency. Components at other frequencies are due to the noise we want to remove. The result of multiplying one frequency component of V_s by V_r is then

$$v_s(\omega)v_r\left\{\cos\left[(\omega-\omega_r)t + (\phi_i - \phi_r)\right] - \cos\left[(\omega+\omega_r)t + (\phi_i + \phi_r)\right]\right\} \tag{9}$$

The phase detector output is applied to a low pass filter which will reject the second term entirely because it is at the reference frequency or above. The filter will pass the first term, provided that (ω - ω_r) is sufficiently close to zero.

The signal occurs at ω_r, so the output from the phase detector - filter combination due to the desired signal is just a constant amplitude

$$v_i v_r \cos(\phi_i - \phi_r) \tag{10}$$

which we will maximize by adjusting the relative phase of the reference to zero when we set up the instrument. Incidentally, the sensitivity to signal phase is sometimes useful in itself, and is the reason the device is called a phase-sensitive detector.

Other frequency components for which (ω - ω_r) falls within the band pass of the filter will also appear in the final output. The usual low-pass filter is an RC circuit with time constant τ and 3dB bandwidth $1/\tau$, so signals within roughly $\pm 1/\tau$ of ω_r will pass to the output. Effectively, then, the multiplier and low-pass filter act as a bandpass filter of width $2/\tau$ centered at the reference frequency. Note that if the reference frequency shifts somewhat, the effective center frequency shifts as well. It is sometimes said, therefore, that this is a 'lock-in' amplifier or detector.

One final point to consider is the response of the system to a temporal change in signal amplitude. The response time of any linear filter is inversely proportional to the bandwidth. This means that as we narrow the bandwidth the filter output takes longer and longer to respond to changes in the input. An explicit calculation of the response of an RC filter to a step change in input is fairly easy, with the result shown in Fig. 6. Note that the output of even a single RC stage takes about 2τ to come within 10% of the final value. Changes in signal level due to the slow sweep of an external parameter must, therefore, occur over many time constants to avoid significant distortion, and the improvement in *SNR* comes at the expense of measurement time. In fact, since the bandwidth in inversely proportional to τ, Eq. 6 tells us that the improvement in rms *SNR* is proportional to $\sqrt{\tau}$. Since the measurement time increases directly with τ, the *SNR* increases only as the square root of the measurement time, exactly as in signal averaging.

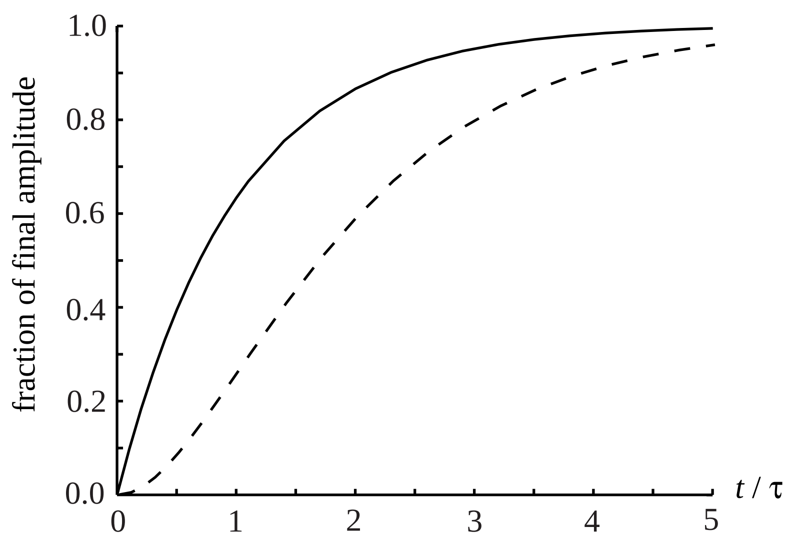

Fig. 6 Response of one- (solid) and two-stage (dashed) RC filters to a sudden change in input.

E. Appendices

1. Mean-square quantities

It is sometimes desirable to have a measure of the intensity of a time-varying voltage. For a sine-wave signal the intensity is conventionally taken as the square of the amplitude, since the power dissipated in a resistor *R* connected to a sine wave voltage of amplitude *A* is proportional to A^2. We can generalize this idea by defining the intensity of an arbitrary time-varying voltage *V(t)* as the time-average of the square of the instantaneous voltage

$$\overline{V^2} = \frac{1}{T}\int_0^T V^2(t)\,dt \tag{11}$$

where the average runs over either one full cycle or over a long time if the voltage is irregular. The quantity $\overline{V^2}$ is called the mean square voltage, with units of volts squared, and is exactly half of A^2 for a sine wave. It is easy to show that the average power dissipated by the voltage *V(t)* in a resistor *R* is just $\overline{V^2}/R$, further reinforcing the idea that this is an intensity.

The square root of $\overline{V^2}$ has units of volts, and is called the root mean square or rms voltage. This quantity is useful because it sometimes behaves like a voltage in circuit calculations but is not limited to sine waves.

2. Power spectra

The rms voltage describes the strength of a signal, but is it sometimes helpful to have information about the time variation as well. The Fourier transform

$$\mathsf{F}(\omega) = \int_{-T/2}^{T/2} V(t)e^{i\omega t}dt \tag{12}$$

is useful for that purpose. To avoid issues of convergence, the integral here runs over some long but finite time. Like the set of Fourier series coefficients, the complex function $\mathsf{F}(\omega)$ gives the amplitudes and phases of the sine waves at frequencies ω which will add up to the original signal. We are usually interested only in intensity information, not phase, so we can employ the magnitude-squared of the transform

$$P(\omega) = \lim_{T\to\infty}\frac{1}{T}\left|\mathsf{F}(\omega)\right|^2 \tag{13}$$

where $P(\omega)$ is called the power spectral density of the signal, with units of volts^2/Hz. For real-valued $V(t)$, the only case of physical interest, $P(\omega)$ is a an even function of ω, so all the physical information is contained in the positive frequency domain. Integrating $P(\omega)$ over all frequencies, we recover the mean-square voltage

$$\overline{V^2} = 2\int_0^\infty P(\omega)\, d\omega \tag{14}$$

further reinforcing the connection to intensity.

More physically, the power part of the name comes again from the proportionality to voltage squared, and it is a spectral density because it is the voltage squared per unit frequency interval. Effectively, $P(\omega)$ tells us how much power our signal has at a specified frequency. This information is particularly useful when considering the action of circuits which selectively attenuate or amplify certain frequency bands.

3. RC filters

One and two-stage RC circuits, as shown in Fig. 7, are frequently used as low-pass filters. Their frequency response can be derived by standard circuit analysis techniques. For a sinusoidal input voltage V_i, the output voltage from one stage is

$$V_o = V_i \frac{Z_C}{Z_C + R} \tag{15}$$

where $Z_C = 1/i\omega C$. In practice we usually need only the amplitude ratio, which is

$$\frac{V_o}{V_i} = \frac{1}{\left[1+(\omega\tau)^2\right]^{1/2}} \tag{16}$$

where $\tau = RC$ is the time constant of the circuit. This ratio is plotted in Fig. 8. The low-pass

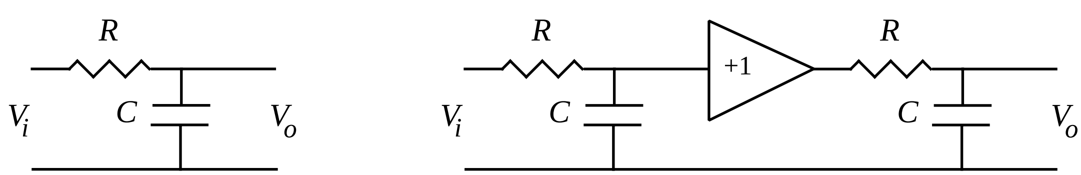

Fig. 7 One and two-stage RC filter circuits. A unity gain amplifier prevents the second stage from loading the first stage.

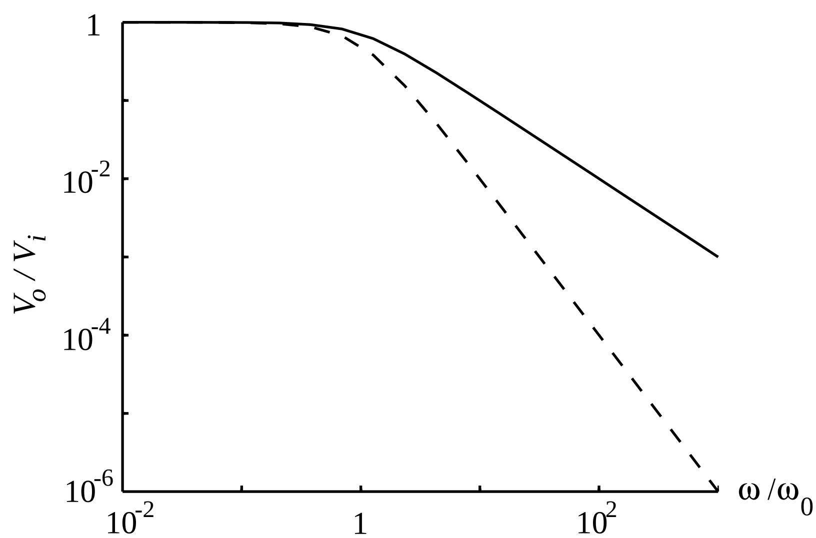

Fig. 8 Frequency response of single (solid line) and two-section (dashed line) RC filters.

filtering action is evident, with the 3 dB point at a frequency of $\omega_0 = 1/\tau$. (The 3 dB point is the frequency at which the power output is reduced by a factor of 2 or, equivalently, the voltage is reduced by a factor of $\sqrt{2}$).

When two RC circuits are connected in series with a buffer amplifier the output voltage of the first becomes the input to the second. The final output is therefore just a product of factors like Eq. 16, or

$$\frac{V_o}{V_i} = \frac{1}{\left[1+(\omega\tau)^2\right]} \tag{17}$$

for two identical sections. Note that the low frequency response is essentially the same as for the single RC circuit, but the attenuation falls off more rapidly with increasing frequency above ω_0.

The time response of the filter can be found by setting up and solving the differential equations for the circuit. For a single-stage RC with an input voltage going from zero to V_i at $t = 0$, the output voltage is

$$V_o = V_i\left(1 - e^{-t/\tau}\right) \tag{18}$$

The corresponding result for the two-stage RC is

$$V_o = V_i\left[1 - \left(\frac{t}{\tau} + 1\right)e^{-t/\tau}\right] \tag{19}$$

Both of these relations have been plotted in Fig. 6. Comparing Figs. 6 and 7, we see that the improvement in high-frequency rejection with the two-stage filter has been obtained at the cost of much longer response time. The application will determine which factor is more important.