Introduction to Signals

Outline

I. Elementary Signal Concepts

- A. Signal Definition and Signal Descriptions
- B. Elementary Signal Characteristics
    1. Signal Support Characteristics
    2. Signal Value Characteristics (Signal Statistics)
    3. Signal Shape Characteristics
- C. Two-Dimensional Signals

II. Elementary Signal Operations

- A. Elementary Operations on One Signal.
- B. Elementary Operations on Two or More Signals

III. Signal Similarity Measures

- A. Mean-squared error
- B. Signal Correlation

IV. Basic Signal Processing Tasks

- A. Signal Recovery/Noise Reduction
- B. Signal Detection/Classification/Recognition
- C. Signal Digitization

I. Elementary Signal Concepts

Reading Assignment: Chapter 1 and these notes. It is recommended that you review these notes every now and then throughout the term. Some of these elementary concepts will only be needed later in the course, and some will only be well understood after you have had more experience with signals.

A. Signal Definition and Signal Descriptions

Definition: A "signal" or "waveform" is a time-varying numerical quantity. More precisely, a signal is a function of time. That is for each value of time t there is number called the[1] *signal value at time* t.

Notation: We typically use lower case letters like x, y, s or subscripted letters like x_1 to represent signals, i.e. functions of time.

Most frequently, we show time t as the argument of such function, as in $x(t)$.

Beware of the Ever-Present Notational Ambiguity: When you see "x(t)" written, sometimes the writer intends you to think of the value of the signal at the specific time t, as in $x(3.1)$, and sometimes $x(t)$ means the whole signal -- that is, the writer intends you to think about the whole signal, i.e. the signal values at all times. When it is essential that reader think about the whole signal, writers will sometimes write x or $\{x(t)\}$ instead of $x(t)$.

Continuous-Time and Discrete-Time Signals: If the time variable ranges over a continuum of values, we say that the signal is *continuous-time*. If the time variable ranges over a discrete set of values we say the signal is *discrete-time.*

More specifically, we assume unless stated otherwise that every continuous-time signal $x(t)$ has time t ranging over all real numbers from $-\infty$ to $+\infty$. In mathematical terms we say that the *domain* of the function x is the interval $(-\infty,\infty)$.

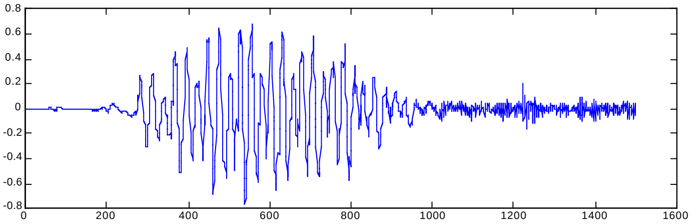
Similarly, unless stated otherwise, every discrete-time signal is assumed to have time t ranging over the set of all integers: $\{..., -2, -1, 0, 1, 2, ...\}$. That is, the domain of the function x is the set of all integers. When dealing with discrete-time signals it is most common to use one of the symbols $i, j, k, l, m,$ or n to denote time rather than t. It is also common to put the time variable within square brackets '[ ]', rather than ordinary parentheses. For instance, the following are examples of the notation used for discrete-time signals $x[n], y[k], z_1[i]$.

Signal Descriptions: Sometimes signals are described with formulas and sometimes they cannot be so described.

Examples of continuous-time signals described with formulas:

$$x(t) = t^2, \qquad y(t) = 3\sin(47t), \qquad z(t) = \begin{cases} 2, & t<0 \\ t^2, & 0\le t\le 1 \\ 3\sin(4t), & t>1 \end{cases}$$

Example of a continuous-time signal that is not describable with a formula:



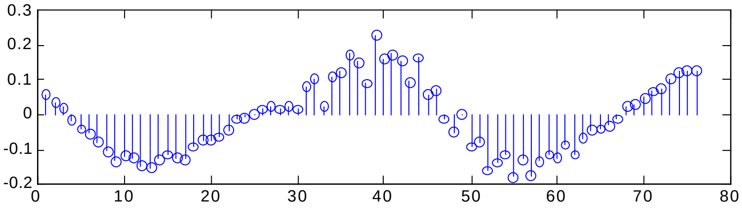
[1] *Italics* is used when a technical term is used or introduced for the first time.

Food for thought: The signal shown above is a recording of someone speaking a couple of words. Everything that one would hear is embodied in the function plotted above.

Examples of discrete-time signals described with formulas:

$$x[n] = n^2, \qquad y[n] = 3\sin(47n), \qquad z[n] = \begin{cases} 2, & n<0 \\ n^2, & 0 \le n \le 10 \\ 3\sin(4n), & n>10 \end{cases}$$

Example of a discrete-time signal that is not describable with a formula:



Are signals described by formulas more "real" or "authentic" than signals that are not so describable? What does it mean to "describe a signal with a formula"? Over the centuries, it has been found useful to give names to certain basic mathematical operations, such as '+', '-', '×' '/', x^2, ln(x), e^x, |x| etc. and certain basic functions, such as sin(x), cos(x), Γ(x), etc. To "describe a signal with a formula" is simply to say that it can be expressed in terms of previously defined operations and formulas. A signal that is not describable by a formula may simply be a function waiting to be blessed with its own name. Or it may be a function that has not previously occurred and may never occur again. Generally, we do not consider signals described by formulas to be any more real or authentic than those that are not so describable.

Note that a formula describing a signal can be quite complex, as in

$$s(t) = \sum_{i=1}^{N} a_i \cos(b_i t + \phi_i)$$

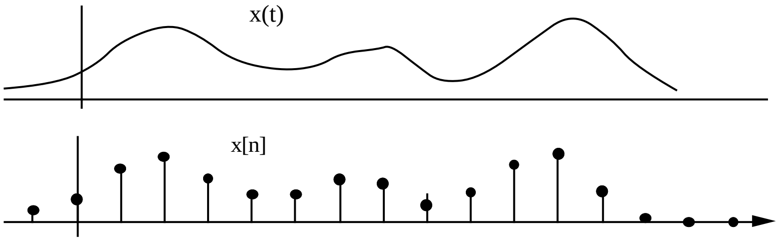
where N, $a_1,\ldots,a_N$, $b_1,\ldots,b_N$, $\phi_1,\ldots,\phi_N$ are "signal parameters", i.e. constants or variables that one needs to know in order to fully determine the signal. It will be important that to develop the skill of being able to work with complex signal formulas. For example, when you see the summation sign Σ, you should recognize that it is just an abbreviation for a sum of N terms. Indeed, to help you to better understand the signal described by a summation, it is often useful to write it in its unabbreviated form, e.g.

$$s(t) = a_1 \cos(b_1 t+\phi_1) + a_2 \cos(b_2 t+\phi_2) + \ldots + a_N \cos(b_N t+\phi_N)$$

Discrete-Time Signals from Continuous-Time Signals via Sampling: Frequently discrete-time signals are produced by *sampling* a continuous-time signal. That is, if x(t) is a continuous-time signal and T_s is a positive number then the discrete-time signal produced by sampling x(t) with *sampling interval* T_s is the signal x[n] defined by

$$x[n] = x(nT_s) .$$

For example, if $T_s = 1.5$, then x[0] = x(0), x[1] = x(1.5), x[2] = x(3), x[3] = x(4.5), etc. The quantity $f_s = 1/T_s$ is called the *sampling frequency* or *sampling rate*, because it represents the frequency or rate (in samples per second) at which samples are taken. For example a continuous-time signal x(t) and a discrete-time signal x[n] produced by sampling x(t) are shown below.



Indeed, as will be discussed a great deal later in the course, we often work with continuous-time signals by working with their samples, i.e. with a discrete-time signal produced by sampling. For example, we often display continuous-time signals by displaying their samples.

On the other hand, there are also discrete-time signals that are not obtained by sampling a continuous-time signal. For example, consider the signal x[n], where x[n] denotes the height of the nth person standing in a certain line.

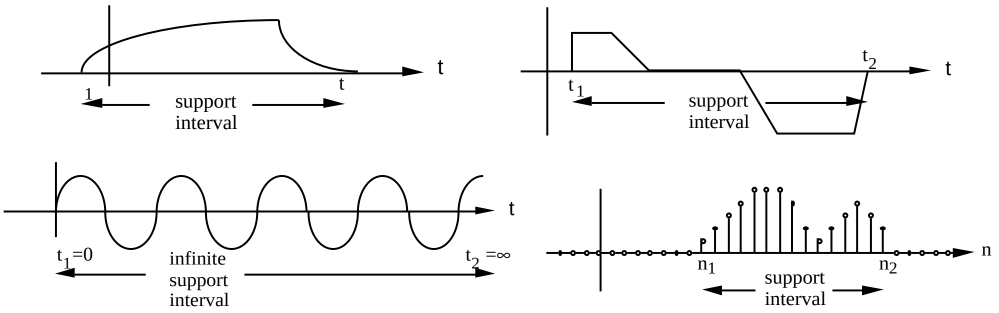
B. Elementary Signal Characteristics[2]

We will primarily present the characteristics of continuous-time signals. There is a discrete-time version of each of these, which will be presented later.

1. Signal Support Characteristics

These are signal characteristics related to the time axis.

Support Interval: Roughly speaking the *support interval* of a signal x(t) is the set of times such that the signal is not zero. More precisely the support interval of a signal x(t) is the smallest interval[3] of times $[t_1,t_2]$ such that the signal is zero outside this interval. We often abbreviate and say simply *support* or *interval* instead of support interval. Several examples are shown below.



Duration: The *duration* or *length* if a signal x(t) is the length of it support interval. Some signals have finite duration and others have infinite duration. For example, the first two signals above have finite duration, and the third signal has infinite duration.

Outside of EECS 206, one will occasionally encounter situations where signals are considered to be *undefined* at times outside their support interval. However, within EECS 206, unless explicitly stated otherwise, we assume the signal value to be 0 outside the support interval. Indeed, we will often define a signal simply by describing its values in some interval, with the presumption that the signal is zero for all times outside this interval. For example, if we introduce a signal as

$$x(t) = t^2, \ \ 1 \le t \le 2 \ ,$$

[2] You do not need to memorize all of these. Rather you need to be aware of the existence of these characteristics, so you can look up and apply the appropriate ones at the appropriate times.

[3] Intervals can be open as in (a,b), closed as in [a,b], or half-open, half-closed as in (a,b] and [a,b]. For continuous-time signals, in almost all cases of practical interest, it is not necessary to distinguish the support interval as being of one type or the other.

then it should be understood that $x(t) = 0$ for $t < 0$ and $t > 2$.

Pulses: Signals with short duration are often called *pulses*. Note that "short" is a subjective or relative designation.

Negative times and time zero: In some of the examples above the signal interval included negative times. What is the significance of negative time? To answer this, one must first answer the question: What is *time zero*? Basically, time zero is just some convenient reference time. Accordingly, a negative time simply represents a time prior to the reference time. For example, a radar system sends a pulse and waits to record the return times of reflections of this pulse from distant objects. It is usually convenient to let "time zero" be the time at which the original pulse was transmitted. Then $t = -1.8$, means 1.8 units of time before the reference time.

2. Signal Value Characteristics, a.k.a. Signal Statistics

We now consider the values a signal $x(t)$ takes.

Maximum and minimum values: If $x(t)$ denotes some generic signal, then it has a *maximum value* x_{max} and a *minimum value* x_{min}. If these are both finite, i.e. $x_{max} < \infty$ and $x_{min} > -\infty$, then the signal is said to be *bounded*.

What do negative vs. positive signal values represent? The answer depends on the application. As an example, when a microphone responds to a sound, there is usually a diaphragm that moves back and forth, tracking the fluctuations in air pressure that constitute the sound. When the diaphragm is pushed one way, the microphone produces a positive voltage; when pulled the other way, it produces a negative voltage.

Average or mean value: A signal also has an average value, also called a mean value. Specifically, the *average* or *mean value* of $x(t)$ over the interval from t_1 to t_2 is

$$M(x) = \frac{1}{t_2-t_1} \int_{t_1}^{t_2} x(t) \, dt \ .$$

Typically a microphone recording has average equal to zero, or very nearly so. In electrical systems, $M(x)$ is often called the *DC value*, where DC stands for *direct current*. If the interval over which the average is sought is infinite, then the average needs to be defined as a limit. For example, the average of the interval $[0,\infty)$ is

$$M(x) = \lim_{T\to\infty} \frac{1}{T} \int_{0}^{T} x(t) \, dt \ ,$$

and the average over the interval $(-\infty,\infty)$ is

$$M(x) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t) \, dt \ .$$

When a signal average is indicated but an interval is not specified, we mean the average over the entire support of the signal.

Absolute value: Quite often, when a signal has values that are both positive and negative, we are interested in a measure of the signal strength apart from its positive or negative sign. With signal strength in mind, one can compute its *magnitude* or *absolute value*, denoted $|x(t)|$.

Squared value, a.k.a. instantaneous power: In most physical situations, the square of $x(t)$, i.e. $x^2(t)$, is a more useful measure of signal strength a time t than magnitude, because it is proportional to the instantaneous power in the signal $x(t)$ at time t, and because power is a quantity of fundamental importance. For such reasons, we often refer to $x^2(t)$ as the *instantaneous power* of $x(t)$ at time t. However, one must remember that the actual power is a constant times this, where the constant depends on the specific physical situation. For example, if $x(t)$ represents the current in amperes flowing at time t through a resistor with resistance R ohms, then the instantaneous power absorbed by the resistor is $R\,x^2(t)$ watts.

Mean-squared value, a.k.a. average power: Whereas $x^2(t)$ is an excellent measure of signal strength at an individual time t, quite frequently we need an aggregate measure of signal strength that applies to the whole signal, or to the signal over some specified time interval. In such cases, we will typically use the *mean-squared value* (MSV). Specifically, the MSV of a signal $x(t)$ over the interval t_1 to t_2 is

$$MS(x) \;=\; \frac{1}{t_2-t_1}\int_{t_1}^{t_2} x^2(t)\;dt\;.$$

This is also called the *average power* in $x(t)$ over the interval t_1 to t_2. As with the definition of average value of x, this definition needs to incorporate a limit when the interval is infinite. And when no interval is specified, the entire support interval is intended.

As an example, mean-squared value is useful when measuring the strength of the signal received by a radar antenna. If it is large in an interval equal to the length of a radar pulse, then we assume that a reflected pulse has been received during this interval, and determine that this pulse is due to an object whose distance is the elapsed time since the original pulse was transmitted times the speed of light. If it is very small, then we can assume that no reflected pulse has been received during this interval, i.e. there is no object at the corresponding distance.

As another example, mean-squared value is used by electric utility companies to determine how much to charge you for the electricity they have supplied. This is because the amount of fuel required by them to supply your electricity is proportional to the mean-squared value of the current supplied to your home.

As a last example, we mention that mean-squared value is often used as a signal quality measure. For example, suppose $x(t)$ is the signal coming from the leftmost of two microphones that are recording an orchestral concert, and suppose $y(t)$ is the signal fed to the left speaker of your stereo receiver after transmission by an FM radio station. Let $e(t) = x(t) - y(t)$ denote the difference between the two signals, which we consider to be an error signal. Then the MSV of $e(t)$ is a good measure of the quality of the system that records and transmits $x(t)$ to you. It is usually called *mean-squared error.*

RMS Value: A closely related quantity is the root mean-squared value (RMSV), which is simply

$$RMS(x) \;=\; \sqrt{MS(x)} \;=\; \sqrt{\frac{1}{t_2-t_1}\int_{t_1}^{t_2} x^2(t)\,dt}\;.$$

On the one hand, RMSV is nicer than MSV in that its value is easier to interpret because it is like a typical signal value, whereas the value of the MSV is harder to interpret because it is like the square of a typical signal value. On the other hand, it is usually easier to work with MSV, because it avoids the square root.

Signal Energy: Another closely related quantity is the *energy* of the signal $x(t)$ in the interval t_1 to t_2, which is

$$E(x) \;=\; \int_{t_1}^{t_2} x^2(t)\;dt\;.$$

By comparing this, with previous definitions, we see that energy is the integral of instantaneous power. It is also the average power multiplied by the length of the interval. Alternatively, average power is energy divided by the length of the interval over which it is computed. A little thought will convince you that it is energy for which an electric utility company actually charges.

Since signal energy and average power (MSV) are related by a constant, the choice of which to focus on is often a matter of taste. If you focus on one, you can easily compute the other.

However, for signals infinite duration often have infinite energy (over their entire support). For such signals, power is usually a more interesting quantity than energy.

Variance and Standard Deviation[4]: The mean-squared value of $x(t)$ minus its average value is called the *variance* of x. The square root of the variance is called the *standard deviation*. That is, the variance[5] of x over the interval t_1 to t_2 is

$$\sigma^2(x) = MS(x-M(x)) = \frac{1}{t_2-t_1}\int_{t_1}^{t_2} (x(t)-M(x))^2 \, dt$$

and the standard deviation is

$$\sigma(x) = RMS(x-M(x)) = \sqrt{\frac{1}{t_2-t_1}\int_{t_1}^{t_2} (x(t)-M(x))^2 \, dt}$$

The variance and standard deviations are useful measures of how "variable" is the signal. A signal with small variance or standard deviation stays close to its average value most of the time, whereas a signal with large variance or standard deviation does not. As with MSV vs. RMSV, standard deviation values are usually easier to interpret because their values are commensurate with signal values. On the other hand, variances are usually easier to compute and work with.

Relationship Between Mean-Squared Value, Variance and Average Value: The following is a useful relationship.

$$MS(x) = \sigma^2(x) + M^2(x) \ (x)$$

Derivation:

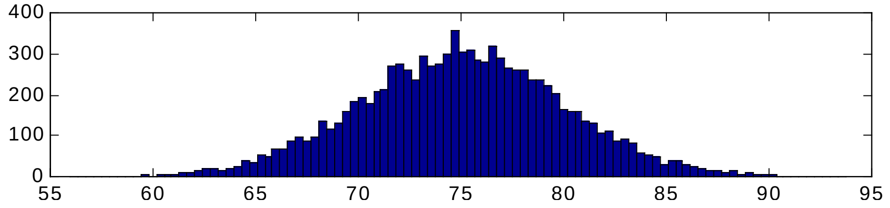
$$\begin{aligned}
\sigma^2(x) &= \frac{1}{t_2-t_1}\int_{t_1}^{t_2} (x(t)-M(x))^2 \, dt \\
&= \frac{1}{t_2-t_1}\int_{t_1}^{t_2} (x^2(t)-2M(x)x(t)+M^2(x)) \, dt \\
&= \frac{1}{t_2-t_1}\int_{t_1}^{t_2} x^2(t) \, dt - \frac{1}{t_2-t_1}\int_{t_1}^{t_2} 2\,M(x)\,x(t)\,dt + \frac{1}{t_2-t_1}\int_{t_1}^{t_2} M^2(x)\,dt \\
&= MS(x) - 2\,M(x)\frac{1}{t_2-t_1}\int_{t_1}^{t_2} x(t)\,dt + M^2(x)\frac{1}{t_2-t_1}\int_{t_1}^{t_2} dt, \\
&\qquad \text{since M(x) is constant we,bring it outside integrals} \\
&= MS(x) - 2\,M(x)\,M(x) + M^2(x) \\
&\qquad \text{by definition of M(x) and by doing the integral} \\
&\qquad \text{on the right-hand side} \\
&= MS(x) - M^2(x)\,, \quad \text{which is the desired relationship.}
\end{aligned}$$

Signal Value Distribution and Histograms: The minimum, maximum, average, and mean-squared value are numbers that each tell us something about the values that

[4] Variance and standard deviation will not be needed early in the course. You can skim them now, and return to them when needed.

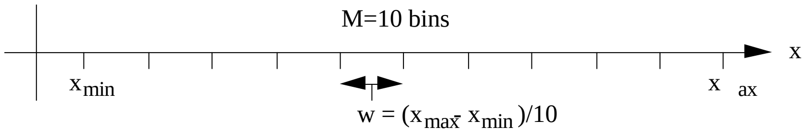
[5] The use of the term σ^2 for variance and σ for standard deviation is traditional.

appear in the signal. The *signal value distribution* gives a more complete picture. Before introducing it, let us review the general meaning of the word *distribution*. As one example, consider the collection of grades resulting from an exam. If we speak of the "distribution of these grades", we mean a plot like that shown below. The horizontal axis shows the possible grades, and the height of the plot above a given grade is proportional to the number of exam papers with that grade. As another example, consider the distribution of incomes of residents of Michigan. Again this is a plot like the one shown below. In this case, the horizontal axis shows the possible incomes, and the height of the plot above a given income is proportional to the number of people with that income.

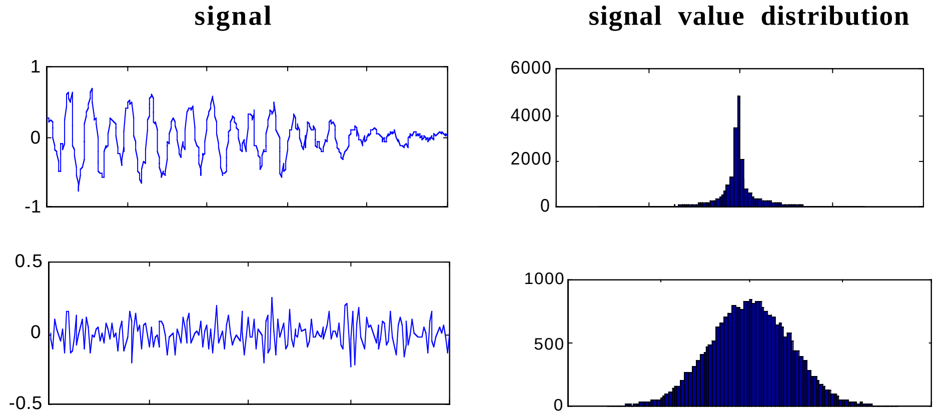


One may similarly consider the distribution of many, many quantities. Not surprisingly, in signals and systems, we are often interested in the distribution of values of a signal x(t), which we call its *signal value distribution*. That is, for a given signal x(t) we want a plot whose horizontal axis shows the signal values and whose height above a given signal value is proportional to the frequency with which that value[6] occurs in the signal.

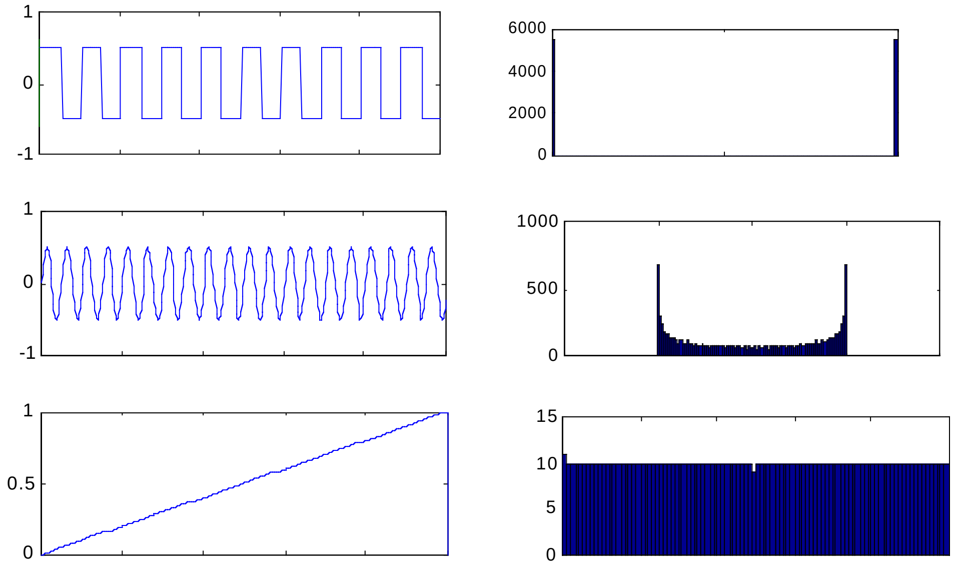
How do we plot the signal value distribution of a signal x(t)? The most common way is make and plot a *histogram*. Specifically, we divide the range of signal values from x_{min} to x_{max} into M equal width *bins*, as illustrated below, where M is some integer, usually in the range 10 to 1000.



If the signal is discrete-time, we count the number of signal values that lies within each bin. We then plot each count above the bin, as illustrated below. If the signal is continuous-time, then we repeat the same procedure on samples of the image. That is, we repeat the procedure on the set of values x(T), x(2T), x(3T), ... where T is the sample spacing. As examples, several signals and their signal value distributions are shown below.



[6] Strictly speaking it is not the frequency of individual values that matter. Rather for any value x, we want the frequency with which signal values lie in a small neighborhood of x, say from x-Δ to x+Δ, where Δ is a small constant.



These histograms were computed with Matlab using the command hist(X,M), where X is a vector containing signal samples, and M is the desired number of bins.

We now justify the statement made earlier that the signal value distribution gives a more complete picture of the signal values than its minimum, maximum, average and mean-squared values. We do this by showing that these latter quantities can be determined, at least approximately, from a histogram. First, the minimum and maximum values will be readily apparent from the histogram. For example, the maximum value is approximately equal to the largest bin center for which the histogram is not zero.

Next, let us show how the average value $M(x)$ can be computed from the histogram. Let $x[1], x[2], \ldots, x[N]$ denote the signal samples. If the histogram has B bins, then the width of each bin will be $W = (x_{max}-x_{min})/B$. The first bin is the interval $(x_{min},x_{min}+W)$, the second bin is the interval $(x_{min}+W,x_{min}+2W)$, and so on. Let C_i denote the center of the ith bin. That is, $C_i = x_{min} + iW - W/2$, for $i =1,\ldots,M$. Let N_i denote the number of signal values that lie in the ith bin. Then the histogram is simply a plot of the points (C_i,N_i), $i = 1,\ldots,B$. The average value of the N signal samples is

$$M(x) = \frac{1}{N} \sum_{n=1}^{N} x[n]$$

Now we observe that we can approximately compute the sum in the above in a different matter. Since there are N_i signal values in the ith bin, we know that there are N_i signal values that approximately equal C_i. The sum of these values is approximately $N_i C_i$. Making this approximation for each of the bins leads to

$$\sum_{n=1}^{N} x[n] \cong N_1C_1 + N_2C_2 + \ldots + N_B C_B .$$

Therefore,

$$M(x) \cong \frac{1}{N} \sum_{i=1}^{B} N_i C_i = \sum_{i=1}^{B} \frac{N_i}{N} C_i$$

That is, the average signal value $M(x)$ is approximately the weighted average of the C_i's (the bin centers), where the weight multiplying C_i is the fraction of samples that lie in the ith bin.

In an entirely similar fashion one may show that

$$MS(x) \cong \sum_{i=1}^{B} \frac{N_i}{N} (C_i)^2 \ .$$

Then from the mean and the mean-squared value, one may directly compute the RMS value, the variance and the standard deviation.

The mean value, mean-squared value, RMS value, variance and standard deviation for a continuous-time signal are each approximately equal to the corresponding quantity for the discrete-time signal produced by sampling the continuous-time signal. Thus, they too may be estimated from a histogram.

In summary, for both discrete-time and continuous-time signals, all of the basic signal value characteristics can be determined, at least approximately, from the signal value distribution.

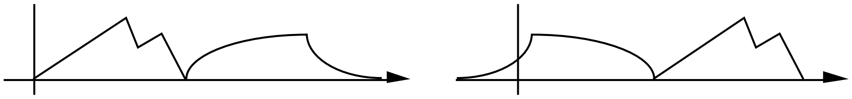
Summary of Signal Value Characteristics

The following table shows the definitions of the signal characteristics mentioned previously, with the exception of signal value distribution, which is not easily summarized in table form. It also lists the analogous characteristics for discrete-time signals.

	Continuous-time signal x(t)	**Discrete-time signal x[n]**
support interval	$[t_1,t_2]$	$\{n_1,n_1+1,...,n_2\}$
duration	t_2-t_1	n_2-n_1+1
maximum value:	$x_{max} = \max_{t} x(t)$	$x_{min} = \max_{n} x[n]$
minimum value:	$x_{min} = \min_{t} x(t)$	$x_{min} = \min_{n} x[n]$
mean value:	$M(x) = \frac{1}{t_2-t_1}\int_{t_1}^{t_2} x(t)\ dt$	$M(x) = \frac{1}{n_2-n_1+1}\sum_{n=n_1}^{n_2} x[n]$
magnitude:	$\lvert x(t)\rvert$	$\lvert x[n]\rvert$
squared value,a.k.a. instantaneous power:	$x^2(t)$	$x^2[n]$
mean-squared value, a.k.a. average power:	$MS(x) = \frac{1}{t_2-t_1}\int_{t_1}^{t_2} x^2(t)\ dt$	$MS(x) = \frac{1}{n_2-n_1+1}\sum_{n=n_1}^{n_2} x^2[n]$
RMS value:	$RMS(x) = \sqrt{MS(x)}$	$RMS(x) = \sqrt{MS(x)}$
energy:	$E(x) = \int_{t_1}^{t_2} x^2(t)\ dt$	$E(x) = \sum_{n=n_1}^{n_2} x^2[n]$
variance:	$\sigma^2(x) = MS(x-M(x))$	$\sigma^2(x) = MS(x-M(x))$
standard deviation:	$\sigma(x) = \sqrt{MS(x-M(x))}$	$\sigma(x) = \sqrt{MS(x-M(x))}$
relationship:	$MS(x) = \sigma^2(x) + M^2(x)$	$MS(x) = \sigma^2(x) + M^2(x)$

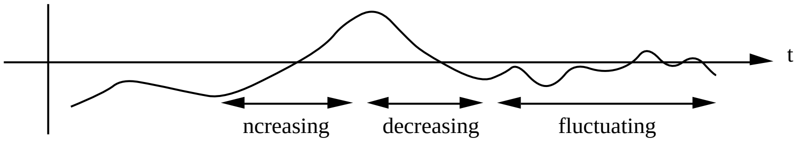
3. Signal Shape Characteristics

In this section we consider signal characteristics related to what we loosely call signal "shape". Note that the signal value characteristics considered previously have nothing to do with signal shape, as one can see by noticing that very different signals can have the same signal value distribution, and consequently, the same min, max, average and mean-squared values. One may also observer that interchanging or time-reversing segments of a signal has no effect on signal value characteristics, but definitely affects signal shape. For example, the following two signals have the same signal value distribution.



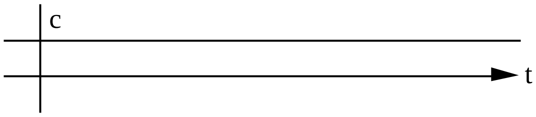
We will first focus on continuous-time signals and later comment briefly on the analogous characteristics for discrete-time signals.

Local shape characteristics: When examining a signal x(t), we often look at segments of it to see if it is *increasing*, *decreasing* or *fluctuating*, as illustrated in the example below.

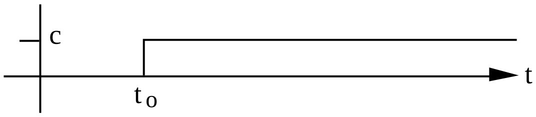


Common signal shapes: The following is a listing of some common signal shapes. These can occur by themselves, or as segments of signals. That is, they may be thought of as local characteristics. The symbols b, c, d, t_o and t_1 represent parameters that need to be specified in order that the signals be completely determined.

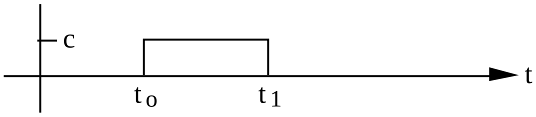
constant: $x(t) = c$



step[7]: $x(t) = \begin{cases} c, & t \geq t_o \\ 0, & t < t_o \end{cases}$



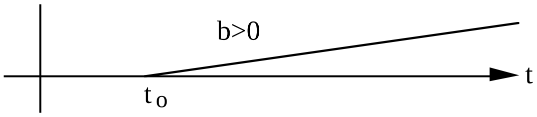
rectangular pulse[8]: $x(t) = \begin{cases} 0, & t < t_o \\ c, & t_o \leq t \leq t_1 \\ 0, & t > t_1 \end{cases}$



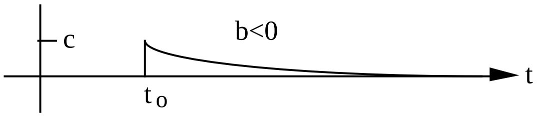
[7] Note that since the value of x at time t_o is c, strictly speaking, we should simply plot the value c at time t_o. Instead, we have drawn a vertical line, from 0 up to c. This emphasizes the change in x as it goes from $x(t) = 0$ for $t<t_o$ to $x(t) = c$ for $t>t_o$. This convention of drawing vertical lines where a function has a step change in value is quite common. We should also note that in the real world, there is no signal cam make a perfect instantaneous step from one value to another, as the formula for the step signal indicates. Instead, the signal value will rise rapidly from 0 to c in the vicinity of t_o. Thus a plot of a real world step signal will have a nearly vertical line rising from 0 to c at t_o. We may think of the vertical line shown in the figure above as a reminder that, in the real world, the signal can change rapidly, but cannnot actually have an ideal step change.

[8] Again notice the vertical lines, which are drawn for emphasis, and as a reminder of what happens in the real world.

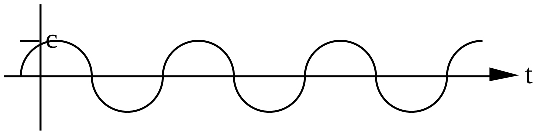
ramp: $x(t) = \begin{cases} 0, & t<t_o \\ c(t-t_o), & t\geq t_o \end{cases}$, increasing if c>0, decreasing if c<0



exponential: $x(t) = \begin{cases} 0, & t<t_o \\ ce^{b(t-t_o)}, & t\geq t_o \end{cases}$, increasing if b>0, decreasing if b<0, constant if b=0



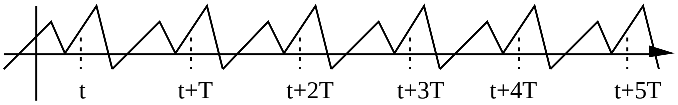
sinusoidal: $x(t) = c \sin(bt+d)$, fluctuating if $b\neq 0$



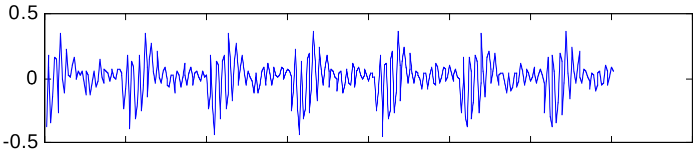
Periodicity: A continuous-time signal x(t) is said to be *periodic with period* T if

$$x(t+T) = x(t) \text{ for all values of } t .$$

It is conventional to require the period T to be a positive number. For example the plot below shows a periodic signal. Its values are marked at a particular time t and also at times t+T, t+2T, ...



Many signals that appear in nature are periodic, or at least nearly so. For example, the following is a segment from a recording of someone speaking the vowel "ee".



Though many signals are *aperiodic*, i.e. not periodic, it turns out that periodic signals can play a key role in their analysis. Several important facts about periodic signals are given next.

Fact 1. A continuous-time signal x(t) with period T is also periodic with period 2T, because for any time t,

$$x(t+2T) = x((t+T)+T) = x(t+T) = x(t), \text{ for all values of } t,$$

where the last two inequalities follow from the definition of "periodic with period T". Indeed it is periodic with period nT for every positive integer n.

Fact 2. Though any periodic signal may be classified as having infinitely many periods, there is always a unique smallest period, which is called the *fundamental period* and which is often denoted T_o. That is, the fundamental period T_o of a signal x(t) is the smallest positive number T such that x(t+T) = x(t) for every value of t.

The reciprocal of T_o is called the *fundamental frequency* f_o of the signal. That is, $f_o = 1/T_o$. It is the number of fundamental periods that occur per unit time. Warning: People often say "period" when they mean "fundamental period". So whenever you hear the word "period", you need to use the context to figure out if they really mean "fundamental period".

Fact 3. If $x(t)$ has fundamental period T_o, then $x(t)$ is periodic with period nT_o for every positive integer n. Conversely, these are the only periods of $x(t)$. That is, if $x(t)$ is periodic with period T, then $T = nT_o$ for some integer n.

Derivation of the converse statement[9]: Suppose $x(t)$ is periodic with fundamental period T_o and is also known to be periodic with period T. We must show that T is an integer multiple of T_o. We use proof by contradiction. Hypothetically suppose that T is not a multiple of T. Then $T = nT_o + r$ where n is the integer part of T/T_o and r is the remainder, $0 < r < T_o$. Since $x(t)$ is periodic with period T_o, it must be that for any time t,

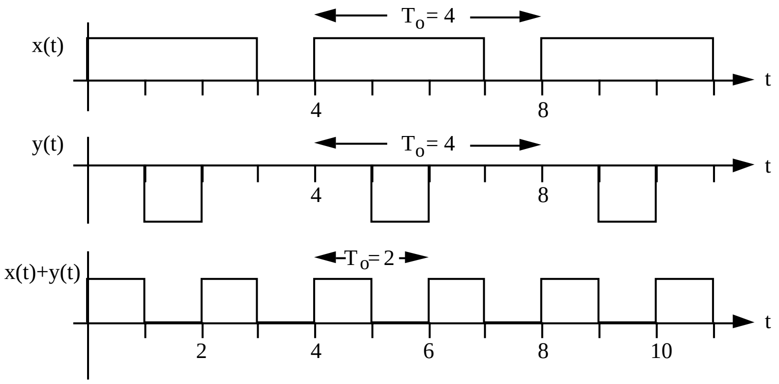
$$\begin{aligned} x(t+r) &= x((t+r)+NT_o)\,, && \text{since } x(t) \text{ is periodic with period } T_o, \text{ we have } x(t+NT_o) = x(t), \text{ which we apply with } t \text{ replaced by } t+r \\ &= x(t+T) && \text{because } T = NT_o+r \\ &= x(t) && \text{because } x(t) \text{ is periodic with period } T \end{aligned}$$

Since $x(t+r) = x(t)$, we deduce that $x(t)$ is periodic with period r. But the fact that $r<T_o$ contradicts the fact that T_o is, by definition, the smallest period of $x(t)$. Therefore, our hypothetical assumption must be false. We conclude that T is a multiple of T_o.

Fact 4. A constant signal, e.g. $x(t) = 3$, is a special case. It satisfies $x(t+T) = x(t)$ for any choice of T. Thus it is periodic with period T for every value of $T > 0$. However, it is conventionally defined to have fundamental period $T_o = \infty$ and fundamental frequency $f_o = 0$. This somewhat arbitrary definition turns out to be more useful than other definitions.

Fact 5. If signals $x(t)$ and $y(t)$ are both periodic with period T, then the sum of these two signals, $z(t) = x(t) + y(t)$ is also periodic with period T. This same property holds when one sums three or more signals. (The derivation of this will be given in class or given as a homework problem.)

Fact 6. The sum of two signals with fundamental period T_o is periodic with period T_o, but its fundamental period might be less than T_o, as the following example illustrates.



[9] This derivation is included for completeness. It is not expected that students can replicate this proof.

Fact 7. The sum of two signals with differing fundamental periods, T_1 and T_2, might or might not be periodic. They will be periodic when and only when the ratio of their fundamental periods equals the ratio of two integers. For example, if T_2/T_1 is 5/3, then the sum will be periodic. However, if $T_2/T_1 = \sqrt{2}$, then the sum will not be periodic.

To see that having an integer ratio makes a difference, consider two signals: x(t) with fundamental period T_1, and y(t) with fundamental period T_2. Suppose that $T_2/T_1 = m/n$, where m and n are integers. Then $nT_2 = mT_1$. Letting $T = nT_2 = mT_1$, we see that

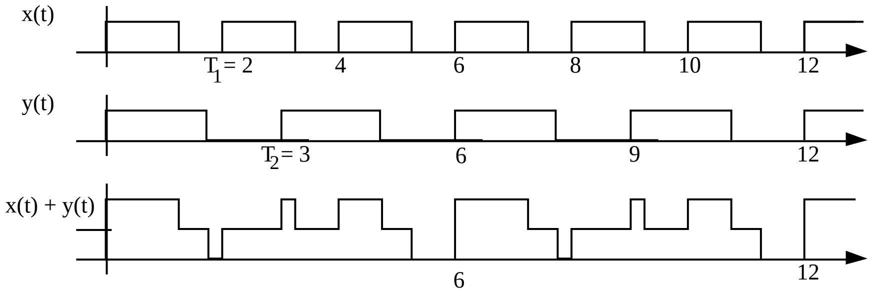
$$\begin{aligned} x(t+T) + y(t+T) &= x(t+mT_1) + y(t+nT_2) \\ &= x(t) + y(t) \end{aligned}$$

because x has period T_1 and y has period T_2. This shows that x(t)+y(t) is periodic with period T.

To complete our discussion, we should also show that if T_2/T_1 is not the ratio of integers, then x(t)+y(t) is not periodic. However, the proof of this is beyond the scope of the course and will not be given here.

In the case where T_2/T_1 is the ratio of two integers, the fundamental period of the sum signal can usually be found by finding the smallest integers m and n such that $nT_2 = mT_1$. In other words, the fundamental period is usually the least common multiple of T_2 and T_1. Correspondingly, the fundamental frequency is usually the greatest common divisor of the fundamental frequencies f_2 and f_1 of the two signals. We say "usually" because there are also examples like that illustrated in Fact 6 where the actual fundamental period is smaller than the least common multiple.

As an example, suppose x(t) and y(t) are the periodic signals shown below with fundamental periods 2 and 3, respectively. Then, their sum x(t) + y(t) is periodic with fundamental period 6.



If instead of summing two periodic signals, we sum M periodic signals, then a discussion similar to the one above shows that the sum is periodic when and only when the ratios of each pair of fundamental periods is a ratio of integers. Moreover, the fundamental period of the sum is usually the least common multiple of the fundamental periods of the individual periodic signals.

Fact 8. The average of a periodic signal with period T over an interval whose length is a multiple of T equals the average over any interval of length T. The same applies to mean-squared valued and energy.

To see why, consider the average over the time interval $[t_1,t_1+mT]$:

$$\begin{aligned} M(x) &= \frac{1}{mT} \int_{t_1}^{t_1+mT} x(t)\ dt \\ &= \frac{1}{mT} \left(\int_{t_1}^{t_1+T} x(t)\ dt + \int_{t_1+T}^{t_1+2T} x(t)\ dt + \ldots + \int_{t_1+(m-1)T}^{t_1+mT} x(t)\ dt \right) \end{aligned}$$

$$= \frac{1}{mT} \Big(\int_{t_1}^{t_1+T} x(t)\ dt + \int_{t_1}^{t_1+T} x(t)\ dt + \ldots + \int_{t_1}^{t_1+T} x^2(t)\ dt \Big)$$

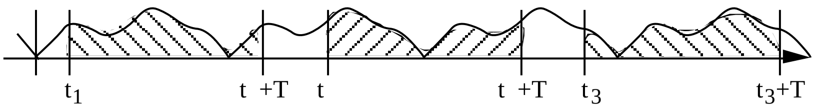
because x(t) is the same in each t second interval

$$= \frac{1}{T} \int_{t_1}^{t_1+T} x(t)\ dt$$

Thus we see that the average over m periods reduces to the average over just one period.

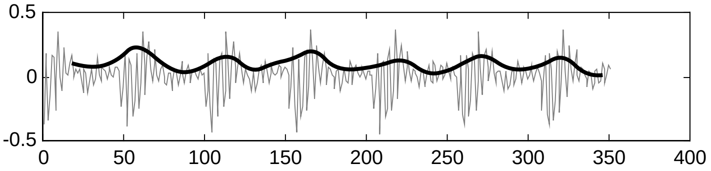
A related "limiting" argument shows that the average over an infinite interval of time reduces to the average over just one period.

Finally, we note that the average is the same over all intervals of length T. This follows from the fact, illustrated below, that the integral of x(t) over any interval of length T is the same, because by periodicity, the same values are being integrated, though perhaps in a different order.



Fact 8 also applies to mean-squared value, because mean-squared value is itself an average. It applies to energy because energy is just mean-squared value multiplied by the interval length.

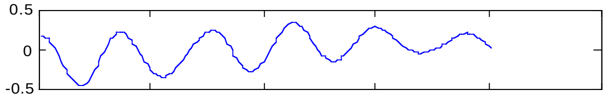
Signal Envelope: This is best introduced with an example. The thick black line overlaying the signal shown below is the *envelope* of the signal. That is, for a rapidly fluctuating signal x(t), the envelope is a smooth curve that approximately follows the positive peaks of the signal. Admittedly this is not a very precise definition, and there is no universally accepted definition that can make it precise. Nevertheless, the envelope is often a useful concept.



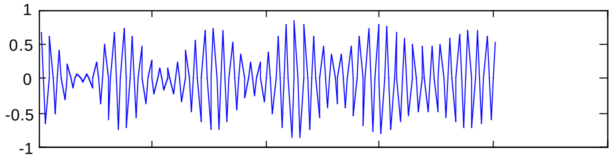
As an example, an AM radio station transmits an audio signal by embedding it in the envelope of a high frequency signal. Specifically, suppose m(t) is the audio signal to be transmitted. Then the radio station assigned to frequency f_o transmits a signal of the form

$$s(t) = (m(t)+c) \cos(2\pi f_o t)$$

where c is a parameter chosen so that $m(t)+c \geq 0$ for all, or at least most, times t. Typically, f_o is a frequency much higher than the rate of fluctuation of m(t). For example, if m(t) is the audio signal shown below,



then the transmitted signal s(t) (m(t)+.5) cos($2\pi f_o t$) is



Can you see the audio signal m(t) embedded in the envelope of the transmitted signal s(t)? Can you think of a way of recovering m(t) from s(t)?

Spectrum

The spectrum of a signal is a terrifically important signal-shape-related characteristic having to do with the "frequency content" of the signal. It is so important that we will not discuss it here. Rather, beginning with Chapter 3, it will be a focus of much of the remainder of the class.

Signal Shape Characteristics of Discrete-Time Signals

Discrete-time signals can have all the same shape characteristics as continuous-time signals. For example, they can be increasing, decreasing or fluctuating. Common signal shapes include all of those mentioned previously: constant, step, rectangular pulse, ramp exponential and sinusoidal. Envelope is again a useful concept, as is periodicity. Because periodicity is such an important concept, we repeat the discussion of it here, this time for discrete-time signals.

Periodicity of discrete-time signals: A discrete-time signal $x[n]$ is said to be *periodic with period* N (an integer) if

$$x[n+N] = x[n] \text{ for all integers } n.$$

This definition is the same as the definition for continuous-time signals, except that instead of the equality holding for all continuous times t, it holds for all integer times n. It is conventionally required that $N > 0$. We now reprise the various facts about periodicity. They are essential identical to the corresponding facts for continuous

Fact 1. A discrete-time signal with period N is also periodic with period mN for any positive integer m.

Fact 2. The *fundamental period*, denoted N_o, is the smallest positive integer N such that $x[n+N] = x[n]$ for all integers n. The reciprocal of N_o is called the *fundamental frequency* f_o of the signal. That is, $f_o = 1/N_o$. It is the number of fundamental periods occurring per sample. (It is always less than or equal to one.) Warning: People often say "period" when they mean "fundamental period".

Fact 3. If $x[n]$ has fundamental period N_o, then $x[n]$ is periodic with period mN_o for every positive integer m. Conversely, these are the only periods of $x[n]$. That is, if $x[n]$ is periodic with period N_o, then $N = mN_o$ for some integer m.

Fact 4. A constant signal, e.g. $x[n] = 3$, is a special case. It satisfies $x[n+N] = x[n]$ for any choice of N. Thus it is periodic with period N for every value of $N > 0$. However, it is conventionally defined to have fundamental period $N_o = \infty$ and fundamental frequency $f_o = 0$. This somewhat arbitrary definition turns out to be more useful than other definitions.

Fact 5. If signals $x[n]$ and $y[n]$ are both periodic with period N, then the sum of these two signals, $z[n] = x[n] + y[n]$ is also periodic with period N. This same property holds when one sums three or more signals.

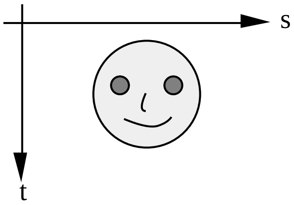
Fact 6. The sum of two signals with fundamental period N_o is periodic with period N_o, but its fundamental period might be less than N_o.

Fact 7. The sum of two signals with differing fundamental periods, N_1 and N_2, is periodic with fundamental equal to the least common multiple of N_1 and N_2 and fundamental frequency equal to the greatest common divisor of their fundamental frequencies f_1 and f_2. Note that unlike continuous-time case, the ratio of the fundamental periods of discrete-time periodic signals is always the ratio of two integers. Therefore, the sum is always periodic. Similarly, the sum of M periodic signals is periodic with fundamental period equal to the least common multiple of the fundamental periods of the individual signals.

Fact 8. The average of a periodic signal with period N over an interval whose length is a multiple of N equals the average over any interval of length N. The same applies mean-squared value and energy.

C. Two-Dimensional Signals

A picture or *image*, as we will usually say, can also be modeled as a signal. However, in this case, it must be modeled as a *two-dimensional* signal x(t,s). That is, instead of single independent parameter t representing time, there are two independent parameters t and s, representing vertical and horizontal position respectively. That is, x(t,s) represents the intensity or brightness of the image at the position specified by horizontal position t and vertical position s, relative to some coordinates. All of the previously mentioned concepts and characteristics can be extended to apply to two-dimensional signals. But we won't discuss them here. However, we do wish to mention that two-dimensional images can be discrete-time as well as continuous-time (discrete-space and continuous-space are better terms). In this case, the signal is x[m,n] where m and n are integers representing vertical and horizontal positions, respectively.



II. Elementary Signal Operations

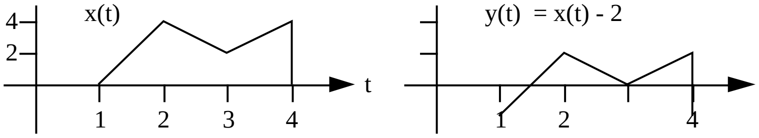
A. Elementary Operations On One Signal.

In our discussions to come of signals and systems, we will routinely use a number of elementary *operations* that, when applied to one signal, result in another closely related signal. In the following we introduce these using continuous-time notation. With one exception to be noted, they apply equally to discrete-time signals, as well.

Adding a constant: This is the operation of adding a constant to the signal. More specifically, there is a number c that is added to the signal value at every time t. If the original signal is $x(t)$, then the result is a new signal

$$y(t) = x(t) + c \ .$$

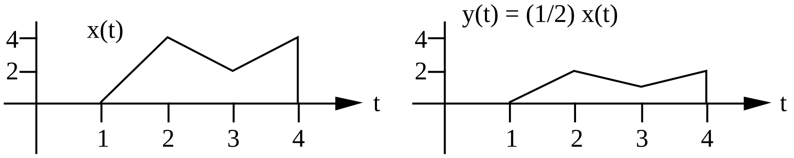
It should be easy to see that this has the effect of increasing the average value of x by c. That is, $M(y) = M(x) + c$.



Amplitude scaling: *Amplitude scaling* is the operation of multiplying a signal by a constant. That is, there is a constant c, called a *scale factor* or *gain*, the value of the signal at every time t is multiplied by c. If the signal being scaled is $x(t)$, then the result of the scaling is

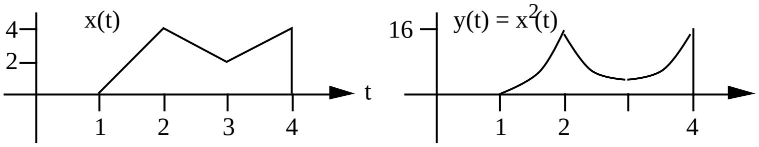
$$y(t) = c\ x(t) \ .$$

This has the effect of scaling both the average and the mean-squared values. Specifically, $M(y) = c\,M(x)$ and $MS(y) = c^2 MS(x)$.



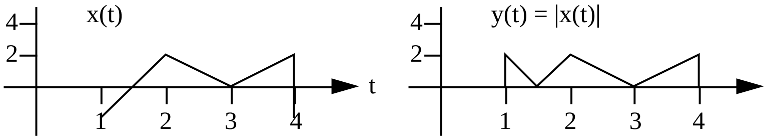
Squaring: Here we simply square the value of the signal at each time, yielding

$$y(t) = x^2(t) \ .$$



Absolute value: As the name suggests,

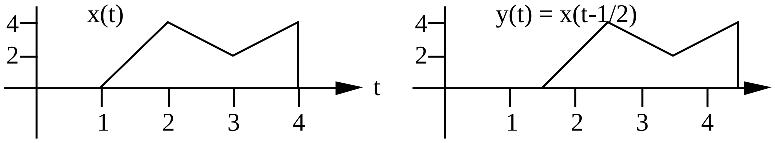
$$y(t) = |x(t)| \ .$$



Time shifting: If $x(t)$ is a signal and T is some number, then the signal

$$y(t) = x(t-T)$$

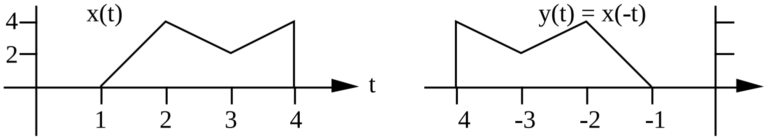
is a *time-shifted* version of x(t). That is, the value of y at time t is precisely the value of x at time t-T. This means that if T > 0, then as illustrated below, anything that "happens" in the signal x also happens in the signal y, but it happens T time units later in y than in x. Similarly, if T < 0, it happens T time units earlier in y. It is useful to remember the rule that a positive value of T leads to a right shift of the plot of x(t) and a negative value of T leads to a left shift.



Time reflection/reversal: The time reflected or time reversed version of a signal x(t) is

$$y(t) = x(-t).$$

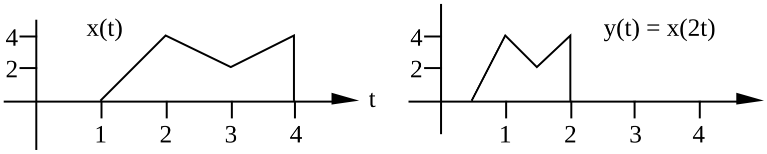
That is, whatever happens in x also happens in y, but at the negative of the time it happens in x.



Time scaling: The operation of *time-scaling* a signal x(t) produces a signal

$$y(t) = x(ct)$$

where c is some positive constant. If c > 1, this has the effect of "speeding up time" in the sense that the value of y at time t is the value of x at time ct, which is a later time. Alternatively, whatever happens in x in the time interval $[t_1,t_2]$ now happens in y in the earlier and shorter time interval $[t_1/c, t_2/c]$.



This is the one property that for which the discrete-time case includes an extra wrinkle. Specifically, in discrete-time, the time values must be an integer. Therefore, if we take

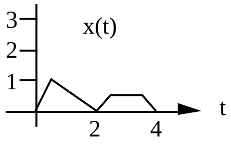
$$y[n] = x[cn] ,$$

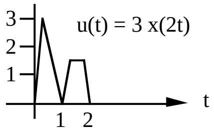
then c needs to be an integer.

Combinations of the above operations: In the future we will occasonally encounter signals obtained by combining several of the operations introduced above, for example,

$$y(t) = 3\ x(-2(t-1)) .$$

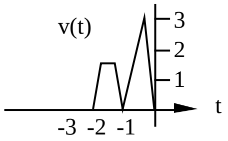
To figure out what signal this is, it is useful to introduce some intermediate signals. For example, in the above, we might start by plotting x(t) and u(t) = 3 x(2t), as shown next.



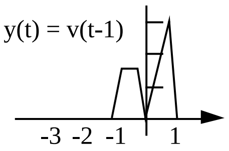


which is an amplitude scaling and time scaling of x(t). Next let's plot

$$v(t) = u(-t) = 3\,x(2(-t)) = 3\,x(-2t).$$



Finally, we plot v(t-1) = 3 x(-2(t-1)) = y(t)



Note that you can also find y(t) by applying the scaling, time shifting and time reversal in some other order, or by applying several operations at a time. But until you are very experienced, it is advisable to apply only one or two at a time.

B. Elementary Operations On Two Or More Signals

Summing: As its name suggests, this is simply the operation of creating a new signal as the sum of two or more signals, as in

$$z(t) = x(t) + y(t) .$$

More specifically, the value of z at time each time t is the sum of x at time t and y at time t.

Linear combining: Linear combining is like summing except that we allow amplitude scaling (i.e. multiply the signals by constants) in addition to summing, as in

$$y(t) = 3\,x_1(t) + 4\,x_2(t) - 2\,x_3(t) .$$

In this case, y(t) is said to be a *linear combination* of $x_1(t)$, $x_2(t)$ and $x_3(t)$. The scale factors multiplying the x(t)'s are often called *coefficients*.

Linear combinations arise in a several ways. As one example, sometimes we are given a collection of signals, say $x_1(t)$, $x_2(t)$ and $x_3(t)$ and are asked to *synthesize* another signal y(t) as a linear combination of the signals in the collection. For example, suppose we need to create the signal y(t), but our hardware can only of produce signals $x_1(t)$, $x_2(t)$ and $x_3(t)$ and perform linear combinations. Often, it is not possible to exactly synthesize y(t) from the given collection and the synthesis must necessarily be approximate.

As another example, sometimes we are given a signal z(t) that is known to be a linear combination of $x_1(t)$, $x_2(t)$ and $x_3(t)$, and we are asked to find the scale factors. This task, which is called *analysis*, happens for example in communications systems, where the scale factors determine the information carried by the signal y(t). It also happens in *Fourier analysis*, to be discussed considerably throughout the course, where we consider a signal y(t) to be the linear combination of sinusoidal signals with different fundamental frequencies.

Multiplying: As its name suggest, this is simply the operation of creating a new signal as the product of two or more signals, as in

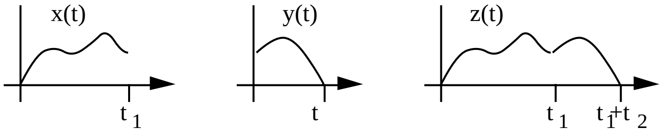
$$z(t) = x(t)\, y(t) .$$

More specifically, the value of z at time each time t is the product of x at time t and y at time t.

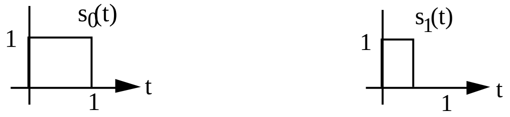
Signal multiplication is a basic operation of most radio transmitters which, as in the example of AM radio described earlier, typically multiply a sinusoidal signal by some information bearing signal.

Concatenating: *Concatenation* is the process of appending one signal to the end of another. For example if $x(t)$ is a signal with support interval $(0,t_1)$ and $y(t)$ is a signal with support $(0,t_2)$, then as illustrated below their concatenation is the signal

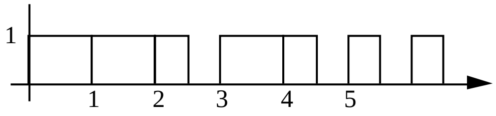
$$z(t) = \begin{cases} x(t), & t \le t_1 \\ y(t-t_1), & t > t_1 \end{cases}$$



Concatenation happens, for example, in digital communications where, for example, to transmit a sequence at the rate of one bit every T seconds, there is a signal $s_0(t)$ with support $(0,T)$ used to send 0's, a signal $s_1(t)$ also with support $(0,T)$ used to send 1's, and the transmitted signal is the concatenation of these. For example, when the signals shown below

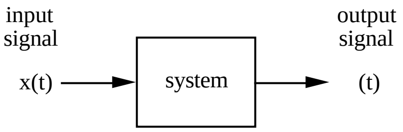


are used to send the binary sequence 0,0,1,0,1,1,1,... the transmitted signal[10] is



Concluding Remarks

The signal operations discussed in this section are elementary operations that are used in a variety of situations. One may view them as basic tools or building blocks. The signal operations considered later in the course (e.g. Chapters 5-8 of the text) are more sophisticated operations, which are developed with some specific task in mind. They can be thought of as *systems* that is, when the operation is applied to a signal $x(t)$, the signal $x(t)$ is viewed as the input to a *system* that performs the operation and produces at its output another signal $y(t)$, which is the result of the operation. In such cases, we often draw a block diagram like the one shown below. Much of the course will be devoted to designing systems to perform the tasks described in the next section.



[10] As usual, vertical lines are shown just emphasize the transitions between transmitted bits, as well as the jumps from 0 to 1 and 1 to 0.

III. Signal Similarity Measures

In many situations, we need a quantitative measure of the similarity of two signals. For example, suppose $x(t)$ is the signal some system should ideally produce, $y(t)$ is the signal the system actually produces. Then, as a measure of how well the system has performed, we need a quantitative measure of how similar $y(t)$ is to $x(t)$. As another example, suppose $r(t)$ is a measured signal that is either the "desired" signal $s_1(t)$ plus some measurement noise, or the "desired" signal $s_2(t)$ plus some measurement noise, and suppose a system must be built that decides which of the two desired signals the measured signal $r(t)$ contains. Such a system needs a signal similarity measure in order to compare $r(t)$ to $s_1(t)$ and $r_2(t)$ to $s_2(t)$.

In summary, signal similarity measures are needed for quantiative performance measures for the systems we design and as an integral piece of certain systems. In the following we introduce and discuss the two most important signal similarity measures.

A. Difference Energy, Mean-Squared Difference and Mean-Squared Error

The difference energy between signals $x(t)$ and $y(t)$ is simply the energy of the difference signal $x(t)-y(t)$. For continuous-time signals, the difference over the time interval (t_1,t_2) is

$$E(x-y) \;=\; \int_{t_1}^{t_2} (x(t)-\hat{x}(t))^2 \, dt \;.$$

Similarly, for discrete-time signals, the difference energy over the time interval $[n_1,n_2]$ is

$$E(x-y) \;=\; \sum_{n_1}^{n_2} (x[n]-\hat{x}[n])^2 \;.$$

A closely related signal similarity measure is the *mean-squared difference* (MSD) between signals $x(t)$ and $y(t)$, which is simply the mean-squared value of the difference signal $x(t)-y(t)$. For continuous-time signals, the MSD over the time interval (t_1,t_2) is

$$MSD(x,y) \;=\; \frac{1}{t_2-t_1} \int_{t_1}^{t_2} (x(t)-\hat{x}(t))^2 \, dt \;=\; \frac{1}{t_2-t_1} \; E(x-y) \;.$$

Similarly, for discrete-time signals, the MSD over the time interval $[n_1,n_2]$ is

$$MSD(x,y) \;=\; \frac{1}{n_2-n_1+1} \sum_{n_1}^{n_2} (x[n]-\hat{x}[n])^2 = \frac{1}{n_2-n_1+1} \; E(x-y) \;.$$

When one of the signals is considered to be the "desired" signal and the other is considered to be an approximation to it, then the difference signal $x(t)-y(t)$ is considered to be an *error signal*, and the mean-squared difference is called the *mean-squared error* and abbreviated MSE(x,y) or simply MSE. MSE is considered a measure of the quality of $y(t)$ as an approximation to $x(t)$, with small MSE indicating good quality.

In many situations, the significance of a particular value of MSE generally depends on the size or strength of the signal $x(t)$. For example, an MSE value of 10 is considered *large* if the squared signal values of the desired signal are mostly smaller than 10, and is considered small if the squared values of the desired signal are much larger than 10. For such reasons, it is common to use *signal-to-noise ratio* as a measure of signal quality, which is defined by

$$SNR(x,y) \;=\; \frac{\sigma^2(x)}{MSE} \;,$$

where $\sigma^2(x)$, which is the variance of $x(t)$, is used as the measure of signal size. Large signal-to-noise ratio indicates good quality.

B. Signal Correlation

Another measure of the similarity of signals $x(t)$ and $y(t)$ is their *correlation*, which is defined

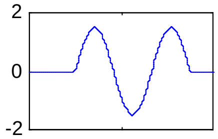
$$C(x,y) \;=\; \int_{t_1}^{t_2} x(t)\,y(t)\;dt\;,$$

where (t_1,t_2) is the time interval of interest. Similarly, the correlation between two discrete-time signals $x[n]$ and $y[n]$ is defined as

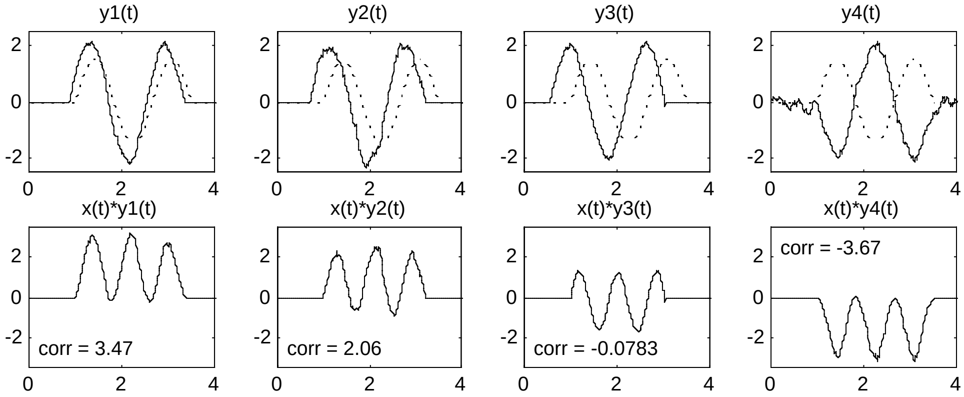
$$C(x,y) \;=\; \sum_{n_1}^{n_2} x[n]\,y[n]\;,$$

where $[n_1,n_2]$ is the time interval of interest. The discussion to follow focuses on continuous-time signals. But everything applies equally to discrete-time signals.

To get a feeling for why correlation is a good measure of signal similarity examine consider the signal $x(t)$ shown below



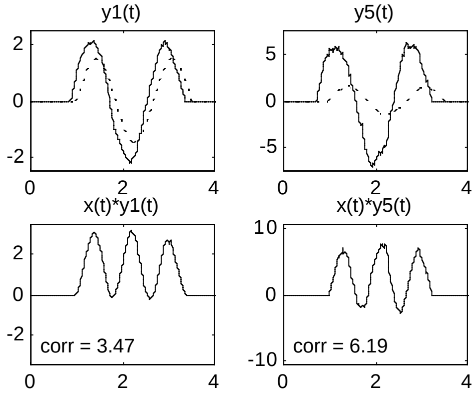
and consider the similarity of each of the signals below, $y_1(t)$, $y_2(t)$, $y_3(t)$, $y_4(t)$, to $x(t)$.


y1(t) y2(t) y3(t) y4(t)
x(t)*y1(t) x(t)*y2(t) x(t)*y3(t) x(t)*y4(t)

As a reference, $x(t)$ is shown with a dotted line in each of the above plots. Also shown below each signal is a plot of the product of $x(t)$ with the signal. The correlation between $x(t)$ and the given signal, which is the area under this plot, is also marked on the plot. Intuitively, we see that $x(t)$ is more like $y_1(t)$ than the other signals, and this is reflected in $C(x,y_1)$ being larger than the other correlations. What is happening is that $y_1(t)$ tends to be positive where $x(t)$ is positive and negative where $x(t)$ is negative. Thus, the product $x(t)\,y_1(t)$ is mostly positive, and the correlation $C(x,y_1)$ is large. The signal $y_2(t)$ has the same sign as $x(t)$ less often. Thus $x(t)\,y_2(t)$ has negative area cancelling some of the positive area, leading to a smaller value of correlation. This is taken to the extreme in $x(t)\,y_3(t)$, for which the positive area is nearlyy completely cancelled by the negative area, causing $C(x,y_3)$ to be zero. The fourth signal, $y_4(t)$, almost always has the oppposite sign of $x(t)$, causing $x(t)\,y_4(t)$ to be almost entirely negative, leading to $C(x,y_3)$ being very negative.

These examples show that $C(x,y)$ tends to be large when $y(t)$ follows the same trends as $x(t)$ -- positive at times t that $x(t)$ is positive, negative at times t that $x(t)$ is negative. This explains why the everyday word "correlation" is taken as the name for the similarity measure $C(x,y)$. We say that $x(t)$ and $y(t)$ are positively or negatively correlated, according to whether $C(x,y)$ is positive or negative. When $C(x,y) = 0$, we say the signals are *uncorrelated*, meaning that they are very different in the sense that the positivity of one at time t gives no clues as to the positivity of the other.

As a next set of examples consider correlating $x(t)$ with $y_1(t)$, shown above, and also with $y_5(t) = 3\,y_2(t)$.



W observe that even though $x(t)$ is intuitively more similar to $y_1(t)$ than to $y_5(t)$, the correlation $C(x,y_1)$ is smaller than the correlation $C(x,y_5)$. What is happening is that correlation is being heavily influenced by the fact that $y_5(t)$ is considerably larger signal than $y_1(t)$, i.e. it has much larger energy. In many situations, it is important to prevent correlation from being influenced by signal size. In such cases, it is customary to use *normalized correlation* as defined by

$$C_N(x,y) \;=\; \frac{C(x,y)}{\sqrt{E(x)}\sqrt{E(y)}} \;=\; \frac{1}{\sqrt{E(x)}\sqrt{E(y)}} \int_{t_1}^{t_2} x(t)\, y(t)\; dt$$

as the signal similarity measure. Here, we have divided $C(x,y)$ by the square root of the energies of both signals. The following lists the values of $C(x,y)$ and $C_N(x,y)$

	$y_1(t)$	$y_2(t)$	$y_3(t)$	$y_4(t)$	$y_5(t)$
$E(y)$	5.42	5.44	4.95	5.42	49.0
$C(x,y)$	3.47	2.06	-0.08	-3.67	6.19
$C_N(x,y)$	0.89	0.53	-0.02	-0.94	0.53

We see now that $C_N(x,y_5) = C_N(x,y_2)$, i.e. that normalized correlation is not affected by the size of the $y_5(t)$.

If, as suggested by the example above, normalized correlation is not affected by the sizes of the signals, then there ought to be some largest value that it can have. The following inequality, called the *Cauchy-Schwarz inequality*, shows that the normalized correlation can never be larger than one, nor less than negative one.

$$\sqrt{E(x)}\,\sqrt{E(y)} \;\le\; C(x,y) \le \sqrt{E(x)}\,\sqrt{E(y)}$$

Equivalently,

$$-1 \;\le\; C_N(x,y) \;\le\; 1\ .$$

The proof of this inequality is beyond the scope of the course[11].

Notice that if $y(t)$ is simply an amplitude scaling of $x(t)$, as in $y(t) = a\, x(t)$ for all t, where $a > 0$, then

[11] One may find a version of the Cauchy-Schwarz inequality in most linear algebra textbooks.

$$E(y) \;=\; E(ax) \;=\; \int_{t_1}^{t_2} (\,a\,x(t)\,)^2 \,dt \;=\; a^2 \int_{t_1}^{t_2} x^2(t)\,dt \;=\; a^2\; E(x)$$

$$C(x,y) \;=\; \int_{t_1}^{t_2} x(t)\;\, a\,x(t)\,dt \;=\; a \int_{t_1}^{t_2} x^2(t)\,dt \;=\; a\;E(x)$$

$$\sqrt{E(x)}\,\sqrt{E(y)} \;=\; \sqrt{E(x)}\,\sqrt{a^2\,E(s)} \;=\; a\;E(x)\;.$$

Thus in the case we see that

$$C(x,y) = \sqrt{E(x)}\,\sqrt{E(y)}$$

or equivalently

$$C_N(x,y) = 1\;,$$

i.e. the Cauchy-Schwarz relation holds with equality. In fact, this is the only way to obtain equality. That is, it can be shown that

$$C(x,y) = \sqrt{E(x)}\,\sqrt{E(y),}\;\; \text{or equivalently,}\;\; C_N(x,y) = 1,$$

when and only when x(t) and y(t) are the same except for a positive multiplicative scaling, i.e. when and only when

$$y(t) \;=\; a\,x(t)\;\; \text{for some}\;\; a > 0\;\; \text{and all}\;\; t\;.$$

Similarly, it can be shown that the only way for $C(x,y)$ to equal $-\sqrt{E(x)}\,\sqrt{E(y),}$ or equivalently for $C_N(x,y)$ to equal -1, is when and only when x(t) and y(t) are the same except for a negative multiplicative scaling, i.e. when and only

$$y(t) \;=\; a\,x(t)\;\; \text{for some}\;\; a < 0\;\; \text{and all}\;\; t\;.$$

A corollary to the Cauchy-Schwarz inequality is the fact that the correlation of a signal with itself equals the signals energy, i.e.

$$C(x,x) \;=\; E(x) \quad \text{for any signal}\;\; x.$$

The relation between correlation and mean-squared difference energy: The relation between mean-squared difference and signal correlation is

$$E(x-y) \;=\; E(x) - 2\,C(x,y) + E(y)\;.$$

Thus, for example, a large positive correlation C(x,y) implies a small difference energy E(x-y). This relation is demonstrated below.

$$E(x-y) \;=\; \int_{t_1}^{t_2} \big(x(t)-y(t)\big)^2\,dt \;\;=\;\; \int_{t_1}^{t_2} \big(x^2(t)\; -2\,x(t)\,y(t)\; +\; y(t)^2\big)\;dt$$

$$=\; \int_{t_1}^{t_2} x^2(t)\; -\; 2 \int_{t_1}^{t_2} x(t)\,y(t)\;dt \;\;+\;\; \int_{t_1}^{t_2} y^2(t)\;\,dt$$

$$=\;\; E(x) - 2\;C(x,y) + E(y)$$

Since difference energy and correlation are closely related, the choice of which to use is a matter of taste, of convenience, or dependent upon other factors. For example, correlation C(x,y) tends to preferred over difference energy in situations where one signal, say x, is much larger than the other, y, is small. In this case $E(x-y) \cong E(x)$, which indicates that E(x-y) depends very weakly on the smaller signal. Thus, it is very sensitive to noise and computational roundoff errors. In contrast, C(x,y) is always greatly influenced by y. For example, when y is much smaller than x, doubling y causes C(x,y) to double, but has little effect on E(x-y). Thus correlation is less sensitive to noise and roundoff errors.

The uses of correlation in EECS 206

Correlation will be used in a couple of the lab assignments as a method for detecting, classifying or recognizing signals. It will also be seen later that one of the principal analysis techniques that we study (Fourier analysis) and the principal kind of systems we study (linear time-invariant filters) are based on correlation. That Fourier analysis is based on correlation relates to the discussion below about "signal components".

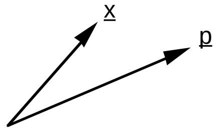
Signal components[12]

The question addressed in this subsection is: What does it mean for one signal to be a component of another? Specifically, suppose we are given signals $x(t)$ and $p(t)$ (or $x[n]$ and $p[n]$ in the discrete-time case).

- Is there a component of $x(t)$ that is like $p(t)$? (or of $x[n]$ that is like $p[n]$?)
- If so, how much $p(t)$ is in $x(t)$? (or $p[n]$ in $x[n]$?)
- How to define "how much of ___ is in ___ "?

For example, is there a component of $x(t)$ that is like $p(t) = \cos(3t)$?

Vector geometry: Such questions are similar to the following traditional questions in vector geometry: Suppose $\underline{x} = (x_1,...,x_N)$ and $\underline{p} = (p_1,...,p_N)$ are N-tuple vectors, illustrated below.

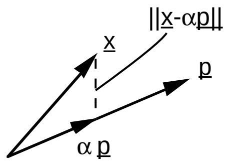


- Is there a component of $\underline{x}$ that is like $\underline{p}$?
- How much of $\underline{p}$ is in $\underline{x}$?

The conventional approach to answering these questions in vector geometry is to find the value α such that $\alpha\underline{p}$ is as close to $\underline{x}$ as possible, i.e. such that $\|\underline{x}-\alpha\underline{p}\|$ is as small as possible, where $\|\underline{u}-\underline{v}\|$ denotes the Euclidean distance between $\underline{u}$ and $\underline{v}$, as defined by

$$\|\underline{u}-\underline{v}\| = \sqrt{\sum_{i=1}^{N} (u_i-v_i)^2}$$

For example, $\alpha\underline{p}$ for one choice of α is illustrated below.



Actually, it's a bit easier to find the value of α that minimizes $\|\underline{x}-\alpha\underline{y}\|^2$, because this avoids the square root. To find the proper α, let's equate to zero the derivative of $\|\underline{x}-\alpha\underline{y}\|^2$ with respect to α, and solve for α. First let's rewrite $\|\underline{x}-\alpha\underline{y}\|^2$:

$$\|\underline{x}-\alpha p\|^2 = \sum_{i=1}^{N} (x_i-\alpha p_i)^2 = \sum_{i=1}^{N} x_i^2 - 2\alpha\sum_{i=1}^{N} x_i p_i + \alpha^2\sum_{i=1}^{N} p_i^2$$

$$= \|\underline{x}\|^2 - 2\alpha(\underline{x} \circ \underline{p}) + \alpha^2\|\underline{p}\|^2$$

[12] This section should be skipped or skimmed. It becomes suggested reading, but not required, when Fourier analsysi is introduced.

where $||\underline{x}||$ and $||\underline{p}||$ are the lengths of $\underline{x}$ and $\underline{p}$, respectively, and $(\underline{x} \circ \underline{p})$ is the dot product defined by

$$(\underline{x} \circ \underline{p}) = \sum_{i=1}^{N} x_i p_i$$

Now differentiating and equating to zero gives

$$0 = \frac{d}{d\alpha} ||\underline{x}\text{-}\alpha p||^2 = \frac{d}{d\alpha} \Big(||\underline{x}||^2 - 2\,\alpha\, (\underline{x} \circ \underline{p}) + \alpha^2\, ||\underline{p}||^2 \Big)$$

$$= -2\, (\underline{x} \circ \underline{p}) + 2\alpha ||\underline{p}||^2 ,$$

which yields

$$\alpha = \frac{(\underline{x} \circ \underline{p})}{||\underline{p}||^2}$$

We conclude that component of $\underline{x}$ that is like $\underline{p}$ is $\frac{(\underline{x} \circ \underline{p})}{||\underline{p}||^2}\, \underline{p}$.

Fact: $\alpha = \frac{(\underline{x} \circ \underline{p})}{||\underline{p}||^2}$ is the unique value of α that makes the residual vector $(\underline{x}\text{-}\alpha\underline{p})$ and $\underline{p}$ orthogonal, where $\underline{u}$ and $\underline{v}$ are said to be orthogonal if $\underline{u} \circ \underline{v} = 0$.

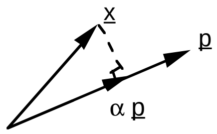
Proof: The dot product of $(\underline{x}\text{-}\alpha\underline{p})$ and $\underline{p}$ is

$$(\underline{x}\text{-}\alpha\underline{p}) \circ \underline{p} = (\underline{x} \circ \underline{p}) - \alpha(\underline{p} \circ \underline{p}) \quad \text{by the linearity of the dot product}$$

$$= (\underline{x} \circ \underline{p}) - \alpha\, ||\underline{p}||^2$$

which is zero when and only when $\alpha = \frac{(\underline{x} \circ \underline{p})}{||\underline{p}||^2}$, i.e. when and only when $(\underline{x}\text{-}\alpha\underline{p})$ and $\underline{p}$ are orthogonal.

With this fact in mind, we see that the component of $\underline{x}$ that is like $\underline{p}$ is the vector in the direction of $\underline{p}$ obtained by projecting $\underline{x}$ onto the direction of $\underline{p}$ as illustrated below.



Back to signals: Let us now return to the original questions for signals: Suppose we are given signals x(t) and p(t).

- Is there a component of x(t) that is like p(t)?
- If so, how much p(t) is in x(t)?
- How to define "how much of ___ is in ___ "?

Our approach will be to find the value α such that the difference energy $E(x(t)-\alpha\, p(t))$ is as small as possible. We will then say that "α p(t) is the component of x(t) that is like p(t)" and "α is the amount of p(t) that is in x(t)". The same approach applies to discrete-time signals.

The idea is that the question we are asking is just like the question for vectors, and we can use the same approach. The only difference is that instead of Euclidean distance as a measure of similarity we use difference energy. Indeed, for discrete-time signals the question is *exactly the same*, because the signals are vectors and difference energy is Euclidean distance squared. Thus in the discrete-time case, we can simply use the answers to the vector question. In doing so, we recognize that what is called "dot product" in the "vector domain", is just what we have called "correlation". Morever, it is easy to check that with "correlation" replacing "dot product", "energy" replacing "length squared", and "uncorrelated" replacing "orthogonal", the answer we found to the vector question applies to continuous-time signals as well as to discrete-time signals. Therefore, we immediately obtain the following:

- The value of α that minimizes the difference energy $E(x(t) - \alpha\, p(t))$ is
$$\alpha = \frac{c(x,p)}{E(p)} .$$
- The amount $p(t)$ that is in $x(t)$ is $\frac{c(x,p)}{E(p)}$.
- The component of $x(t)$ that is like $p(t)$ is $\frac{c(x,p)}{E(p)}\, p(t)$. (++)
- $\alpha = \frac{c(x,p)}{E(p)}$ is the unique value that makes the difference signal $(x(t)\text{-}\alpha p(t))$ and $p(t)$ uncorrelated.
- These answers apply to discrete-time signals as well, with $p[n]$ replacing $p(t)$.
- These answers apply to complex-valued signals, in discrete or continuous time. (Correlation between complex-valued signals is discussed below.)

Comments: Engineers have long recognized the connnections between signals and vectors. As a result, basic ideas from geometry, and more generally from linear algebra, are commonly used in signals and systems analysis. One of the most beneficial transferences is the idea that we can draw geometric pictures that represent signals and their relationships, such as those on the previous pages. For example, uncorrelated signals are drawn at right angles to one another. It often happens that a geometric picture will help one to understand some complex signal situation. It is also true that studying linear algebra will lead to increased understanding of signals and systems. For example, you might wish to learn as much as possible about linear algebra in Math 216 and to take Math 419 as an elective.

III. Basic Signal Processing Tasks

In this section, we describe three broad and nearly ubiquitous tasks that require the processing of signals. That is, there is need to develop systems that perform these tasks. Much of the remainder of the course will be devoted to developing techniques to design and improve such systems.

The first two tasks have a similar flavor. In each, the signal to be processed contains a component that interests us and a component that does not. That is, the signal r(t) to be processed can be modeled as

$$r(t) = s(t) + n(t) ,$$

where s(t) is the component that interests us and n(t) is the component that does not. For example, the component that interests us might be the signal produced by someone speaking into a microphone, and the component that does not might be the signal produced by background noise. In the first task, called *signal recovery* or *noise reduction*, the goal is to recover the signal component s(t) that interests us. For example, we might wish to recover the speech signal without the background noise. In the second task, called *signal detection* or *signal classification* or *signal recognition*, we wish to make a decision about the signal component that interests us. For example, we might wish to decide the identity of the speaker or what the speaker has said. These two tasks will be introduced in the next two subsections.

In each of the tasks, the noise n(t) is not a known signal. If it were known, we could simply subtract it from r(t), and there would be no need for a signal recovery or signal detection system. We also assume that the desired signal s(t), or some aspect of it, is not known. If s(t) were entirely known, we could dispense with r(t), and simply display the signal s(t). On the other hand, there must be something we do know about s(t) and n(t), such as their signal value or signal shape characteristics. Indeed, there must be something we know that is different for s(t) than for n(t). Otherwise, we will have no way to separate one from the other. For example, much of the course will be devoted to developing systems that work when s(t) and n(t) have spectra that differ in known ways, e.g. one contains only low frequencies and the other contains only high frequencies.

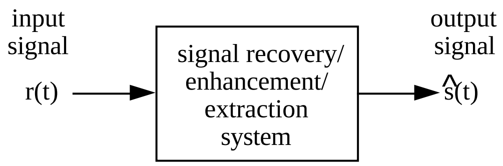
The third task to be discussed is signal digitization. Nowadays, when signals such as audio or pictures or video must be processed, stored or transmitted, it is generally done in digital fashion, i.e. the data is converted to binary. This is done because excellent digital techniques have been found, and because the bits so produced can be processed, transmitted and stored rapidly and reliably.

A. Signal Recovery/Extraction/Enhancement

Suppose we are given a signal r(t) with two components,

$$r(t) = s(t) + n(t) ,$$

and our task is to design a system, such as illustrated below, which processes r(t) in order to produce s(t), or more precisely, an approximation $\hat{s}(t)$ to s(t).

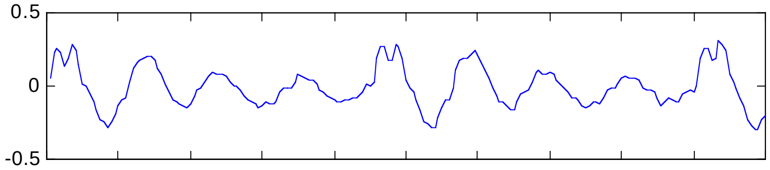


We consider r(t) to be the *original* or *measured* or *received* signal, s(t) to be the *desired signal*, and n(t) to be *noise*. It is sometimes called *signal recovery*, because the system is recovering the signal s(t) from the noise corrupted signal r(t). It is also called *noise reduction* or *noise suppression*, because it attempts to do precisely this.

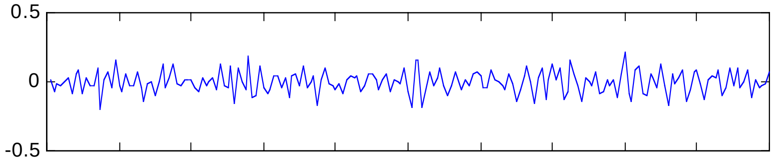
Examples of signals requiring recovery/extraction/enhancement include:

- An audio signal, especially when it is particularly faint, or when the microphone is part of a hearing aid, or when there is much background noise, such as in an automobile or helicopter or crowded cocktail party.
- A photograph or movie or video taken in faint light
- A signal being played back on an analog tape player (video or audio). Magnetic tapes introduce significant amounts of noise due to the granularity of the magnetic media.
- An AM or FM radio signal, or an analog TV signal, as it emerges from the receiving antenna. There is always lots of background noise, much of it due to other radio signals.
- A digital communication signal as it emerges from the receiving wire, antenna or other sensor. This signal must be extracted from background noise and from all other communication signals on the same medium.

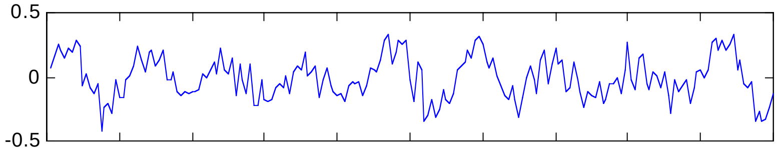
Linear Filters: There are many possible approaches to signal recovery. In this course, we focus mostly on *linear filtering*, which is the most common approach. Let us introduce it with an example. Suppose s(t) is an audio signal, for example the one shown below.



Suppose the measured signal is r(t) = s(t) + n(t), where n(t) looks like the signal below.



Then r(t) is

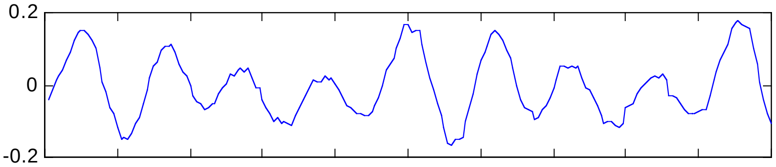


Since the noise signal fluctuates more rapidly than the audio signal[13], a natural approach to reducing the noise is to use a *running-average filter*. That is, we design a system that replaces r(t) by an average of r(t) over an interval up to time t. Specifically, it replaces r(t) with the average over the of r(t) over the time interval (t-T,t), where T is chosen small enough that the audio signal s(t) changes little in the interval and large enough that the noise signal fluctuates a great deal in the interval and, consequently, averages to a small value. In other words, the running-average filter produces the output signal

[13] This is the signal-shape charactertistic that differentiates the s(t) from n(t) in this example.

$$\hat{s}(t) = \frac{1}{T} \int_{t-T}^{t} r(t')\ dt' \ .$$

When such a filter is applied to r(t), it has the effect of smoothing the signal r(t). In our example, it produces the signal shown below, which sounds much more like s(t) than does r(t). Notice that the filtering has not only reduced the noise, but it has also modified the desired signal somewhat.



While the running average filter is fairly common, there are many other linear filters. As a precursor to introducing the full variety of possible linear filters, let us note that by applying the change of variables t" = t'-t to the above integral, we may rewrite the running average filter as producing

$$\hat{s}(t) = \frac{1}{T} \int_{-T}^{0} r(t+t'')\ dt'' \ ,$$

which in turn may be rewritten as

$$\hat{s}(t) = \int_{-\infty}^{\infty} r(t+t'')\ w(t'')\ dt'' \ .$$

where

$$w(t'') = \begin{cases} \frac{1}{T}, & -T \le t'' \le 0 \\ 0, & \text{else} \end{cases} \ .$$

Other linear filters are obtained by replacing the function w(t"), which we call a *weighting function*, by something else. That is, the output is produced by a running average, except that the average is with respect to a weighting function w(t"). We obtain different linear filters by making different choices of w(t"). For example, if we choose

$$w(t'') = \begin{cases} e^{3t''}, & t'' \le 0 \\ 0, & t'' > 0 \end{cases}$$

Then

$$\hat{s}(t) = \int_{-\infty}^{0} r(t+t'')\ e^{3t''}\ dt''$$

In this case, we see that $\hat{s}(t)$ is the average of all past values of r(t). However, in computing the average, past values are multiplied by exponentially decreasing weights.

By careful choice of the weighting function w(t"), one can develop filters that do a better job of extracting a signal from noise than the running average filter. Quite a different sort of weighting function is needed to perform the complex task of extracting a single radio signal from all those at other frequencies. As the course progresses, we will develop better and better techniques for designing filters for recovering signals or suppressing noise.

Actually, in this course, we will focus primarily on discrete-time linear filters for filtering discrete-time signals. (Chapters 5-8 of our text.) Specifically, a discrete-time filter performs the analogous operation

$$\hat{s}[n] \ = \sum_{k=-\infty}^{\infty} r[n+k]\, w[k] \ ,$$

where the w[k]'s are a sequence of weights that distinguish one linear filter from another. For example, if $w[k] = 1/M$, $k = -M+1,...,0$, then we obtain a discrete-time running average filter, which produces

$$\hat{s}[n] \ = \ \frac{1}{M} \sum_{k=n-M+1}^{n} r[k] \ \ .$$

Performance Measure: As engineers, wherever possible we wish to quantify the goodness of the systems that we build. In this course, for the signal recovery task, we will use mean-squared error (MSE) as our measure of goodness. Specifically, if the signal s(t) has support interval (t_1,t_2), then

$$\text{MSE} \ = \ \frac{1}{t_2-t_1} \int_{t_1}^{t_2} (s(t)-\hat{s}(t))^2 \, dt$$

Our goal, then, is is to design a system that makes MSE as small as possible.

One be aware that MSE is sensitive to scale and to time shifts. For example, suppose the signal recovery system has completely eliminated the noise, but has scaled and delayed the somewhat, for example, suppose it prodcues $\hat{s}(t) = 1.2\ s(t-.1)$. Then, even though the system has done well, the measured MSE may be large. In such cases, we may wish to allow $\hat{s}(t)$ to be scaled and time-shifted before measuring MSE.

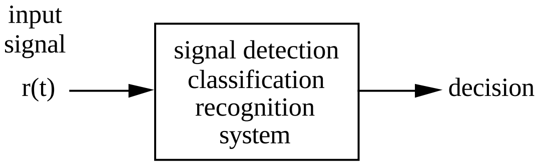
Other Signal Recovery Tasks: There are other situations where the desired signal and noise are not simply added. Rather r(t) depends on the desired signal s(t) in some more complicated way. For example, in AM radio transmission the audio signal we wish to recover is the envelope of the transmitted signal (minus a constant), and it is desired to recover this audio signal from the transmitted signal plus noise. In tomographic imaging (e.g. X-ray, MRI, PET, etc.), the desired signal is a two or three-dimensional image, which must be extracted from a complex set of measurements. The same is true of synthetic aperture radar. These are advanced topics that will not be pursued in this course or in these notes.

B. Signal Detection/Classification/Recognition

Suppose we are given a signal r(t) with two components,

$$r(t) \ = \ s(t) + n(t) \ ,$$

and our task is to design a system, such as illustrated below, which processes r(t) and produces a decision about s(t).



There are three closely related versions of this, introduced below along with examples.

1. **Signal/No Signal?** In this case, $s(t) = 0$ or $s(t) = v(t)$, where v(t) is some known or partially known desired signal. From r(t), the system must decide which of these two possibilities has occurred. This is considered to be a *detection* or *recognition* task because the goal is to *detect* or *recognize* whether or not u(t) has occurred. Some specific examples are given below.

   • Radar: Decide if the signal r(t) from the receive antenna contains a reflected pulse at time t_o. The same issues apply to sonar.

• Dollar bill changer: Decide if the signal r(t) obtained by optically scanning a bill is due to a genuine dollar bill.

• Fingerprint recognition: Decide if the signal r(t) obtained by optically scanning a fingerprint contains the fingerprint of John Smith. Similar tasks include recognition from retinal scans or voice prints.

• Heart monitoring. Decide if an ekg signal r(t) contains a characteristic indicating a heart defect.

2. **Which Signal?** Here, $s(t) = v_1(t)$ or $v_2(t)$ or ... or $v_M(t)$, where M is some finite integer and the $v_i(t)$ are known signals. From r(t) decide which of the $v_i(t)$'s is contained in r(t). This is considered to be a *classification* or *recognition* task because the goal is to *classify* r(t) according to which $v_i(t)$ has occurred, or equivalently to *recognize* which $v_i(t)$ has occurred. Some specific examples are give below.

   • Digital communication receiver: Decide if the received signal r(t) contains the signal representing "zero" or the signal representing "one". That is, the system must decide if the transmitter sent "zero" or "one". In some systems, the transmitter has more than two signals that it might send, and so the receiver must make a multivalued decision.

   • Optical character recognition: Decide if a character printed on paper is a or b or c or ... . This is especially challenging when the characters are handwritten.

   • Spoken word recognition: Decide what spoken word is present in the signal r(t) recorded by a microphone.

   • The "signal/no signal" task may be considered to be a special case of the "which signal task".

3. **Signal? And if So Which Signal?** This is a combination of the two previous subtasks. Suppose s(t) equals 0 or $v_1(t)$ or $v_2(t)$ or ... or $v_M(t)$. From r(t) decide whether or not $s(t) = 0$, and if not, decide which of the $v_i(t)$'s is contained in r(t). Examples:

   • Digital communication receiver: Some digital communication systems operate *asynchronously* in the sense that the receiver does not know when the bits will be transmitted. In this case, the receiver must decide if a bit is present, and if so, is it a zero or a one.

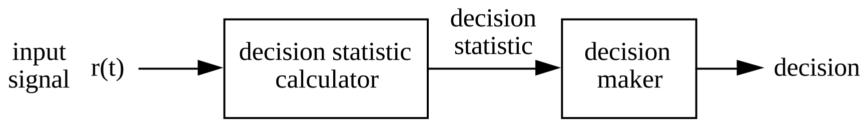
   • Personal identification system: Decide if a thumb has been placed on the electronic thumbpad, and if so, whose thumb.

   • Touch-tone telephone decoder: Decide if the signal from a telephone contains a key press, and if so, which key has been pressed.

   • Spoken word recognition: Decide is a word has been spoken and if so, what word.

For brevity, we will use the term *detection* as a broad term encompassing all of the above.

Detection Systems: As illustrated below, a detection system ordinarily has two subsystems: the first processes the received signal in order to produce a number (or several numbers) from which a decision can be made. The second makes the decision based on the number (or numbers) produced by the first. The number or numbers produced by the first system are called *decision statistics* or *feature values*, and the first subsystem is called a *decision statistic calculator* or a *feature calculator*. The second subsystem is called the *decision maker* or *decision device*. We will discuss two general types of detection systems, corresponding to two types of decision statistic generators -- *energy detectors* and *correlating detectors*.



Quality/Performance Measures: For detection systems, the most commonly used measure of performance is the *error frequency*, which as its name suggests, is simply the frequency with which its decisions are incorrect. We let the symbol f_e denote the error frequency. The typical goal is to design the detection system to minimize f_e.

In some situations, certain types of errors are more significant than others. For example, from the point of view of the owner of a dollar bill recognizer, classifying a counterfeit bill as valid is a more significant error than classifying a genuine dollar bill as invalid. In such cases, one will want to keep track of the frequency of the different types of errors. And one may choose to minimize the total frequency of errors subject to constraints on the frequencies of certain specific types of errors. For example, the owner of a dollar bill recognizer might insist that detector make as few errors as possible, subject to the constraint that it classifiy counterfeit bills as valid no more than one time in a million.

Energy Detectors for Deciding Signal/No Signal: For the "signal/no signal" task, the detector must decide whether $r(t)$ contains signal AND noise, i.e. $r(t) = v(t) + n(t)$, or just noise, i.e. $r(t) = n(t)$. Since it is natural to expect that $r(t)$ will have larger energy in the former case than in the latter, it is natural to choose the energy $E(r)$ of $r(t)$ as the decision statistic. (One would normally measure the energy of $r(t)$ over the support interval of $v(t)$.) The decision maker would then decide that $v(t)$ is present if the energy is sufficiently large, and would decide that $v(t)$ is not present otherwise. To make such a decision, one needs to specify a *threshold*, denoted τ, and the decision rule becomes

$$v(t) \text{ is present if } E(r) \geq \tau, \text{ and } v(t) \text{ is not present if } E(r) < \tau \, .$$

How to choose the threshold? The first thing to note that is that the noise signal $n(t)$ is usually random. That is, it is not known in advance, and it is different every time we measure it. In particular, the energy of the noise will vary from decision to decision. However, based on past experience, it is usually possible to estimate the average value of the noise energy, which we denote $\overline{E}(n)$. Then we can say that when $v(t)$ is not present, the signal $r(t) = n(t)$ has a random energy value, with average $\overline{E}(n)$. On the other hand, when the signal $v(t)$ is present, the energy of $r(t)$, though still random tends to be larger. Specifically, it ordinarily has average energy equal[14] to $E(v) + \overline{E}(n)$. In summary, when the signal $v(t)$ is present, the average energy of $r(t)$ is $E(v) + \overline{E}(n)$, and when $v(t)$ is not present, the average energy of $r(t)$ is $\overline{E}(n)$. It is natural then to choose a threshold that lies half way between these two average energy values. That is, we choose

$$\tau = \frac{1}{2}\big(E(v) + \overline{E}(n)\big) + \frac{1}{2}\overline{E}(n) = \frac{1}{2}E(v) + \overline{E}(n) \, .$$

Energy detectors can also be used for the "which signal" task, provided the signals $v_1(t)$, $v_2(t)$, ... , $v_M(t)$ have sufficiently different energies -- so different that the differences will not be obscured by the noise. In this case, the typical decision maker strategy is to compare $E(r)$ to the average energies $E(v_1)+\overline{E}(n)$, $E(v_2)+\overline{E}(n)$, ..., $E(v_M)+\overline{E}(n)$ that one expects if the various $v_i(t)$'s were present. The decision maker then decides in favor of the signal $v_i(t)$ such that $E(v_i)+\overline{E}(n)$ is closest to $E(r)$.

Correlating Detectors for the "Which Signal Task": For the "which signal" task, an alternate and usually more effective method of detection (than energy detection) is to directly compare $r(t)$ to each of the signals $v_1(t)$, $v_2(t)$, ..., $v_M(t)$.

[14] This is because $v(t)$ and $n(t)$ are usually uncorrelated.

Accordingly, we need a measure of similarity, and we will choose correlation. Specifically, the correlation between two continuous-time signals x(t) and y(t) is defined to be

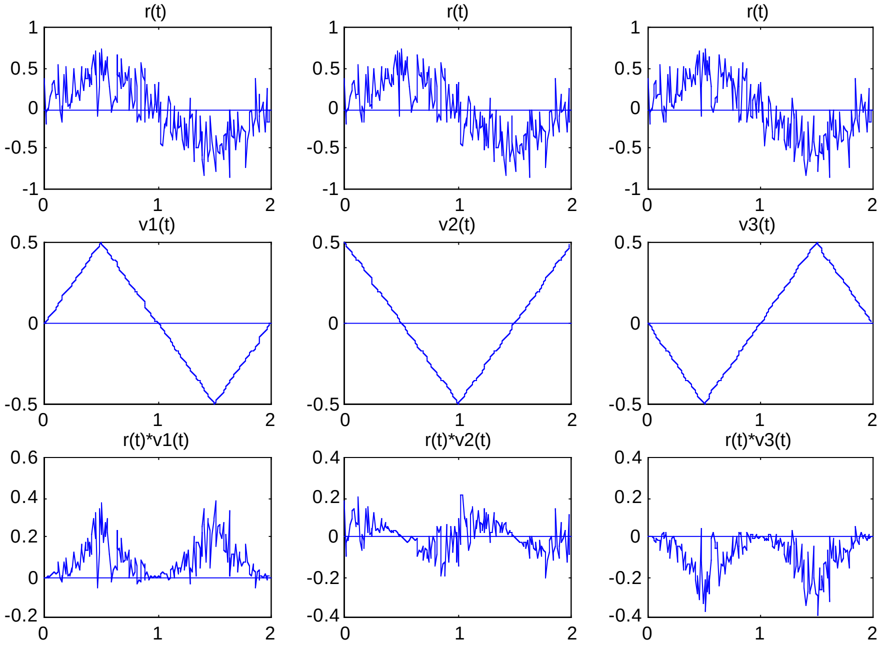
$$C(x,y) \;=\; \int_{t_1}^{t_2} x(t)\,y(t)\;dt\;,$$

where (t_1,t_2) is the time interval of interest. Similarly, the correlation between two discrete-time signals x[n] and y[n] is defined to be

$$C(x,y) \;=\; \sum_{n_1}^{n_2} x[n]\,y[n]\;.$$

For brevity, we will continue the discussion presuming continuous-time signals. To see why correlation is a good measure of similarity to use in detection, consider the signal pairs shown below, in which a signal r(t) is compared to the three possibilities $v_1(t)$, $v_2(t)$ and $v_3(t)$. To aid the comparisons, r(t) is plotted above each signal. One can see that r(t) and $v_1(t)$ are similar in that, roughly speaking, where one is positive, the other is as well; where one is negative the other is as well. Moreover, r(t) roughly follows the shape of $v_1(t)$. On the other hand, the signals r(t) and $v_2(t)$ are rather dissimilar. Where $v_2(t)$ is positive, r(t) is sometimes negative; where $v_2(t)$ is increasing, r(t) is sometimes decreasing. Finally, r(t) and $v_3(t)$ are very dissimilar. Indeed, r(t) is very much like the negative of $v_3(t)$. If one were to make a decision about which of the three signals $v_1(t)$, $v_2(t)$, $v_3(t)$ was contained in r(t) based on visually comparing r(t) to the these signals, one would clearly choose $v_1(t)$. And indeed this is correct, because r(t) was generated by adding noise to $v_1(t)$.

Let us now consider how the same decision could be based on correlation. To do so, let's examine the value of correlation for each pair of signals. The product of each pair of signals is shown below the pair. Correlation is the integral of the product, i.e. the area under the plot of the product signal. For the first pair, the product is almost entirely positive, and the correlation is large. For the second pair, the product is approximately half negative and half positive, and the correlation is small because the positive and negative areas of the product tend to cancel each other. Finally, for the third pair, the product is mostly negative, and the correlation gives a large negative value.



$C(r,v_1) = 16.7$ $C(r,v_2) = 0.82$ $C(r,v_3) = -16.7$

If a detection system had to decide from the three correlation values which of the three signals $v_1(t)$, $v_2(t)$, $v_3(t)$ was contained in $r(t)$, clearly it should choose the one corresponding to the largest correlation, namely, $v_1(t)$.

Though correlation would work well in the example above, consider what would have happened if, for example, $v_2(t)$ were 100 times larger. In this case, it is easy to see that the correlation $C(r,v_2) = 82$, rather than 0.82. Thus even though v_2 has a very different shape than $r(t)$, a decision based solely on the size of the correlation would make the wrong decision. We can remedy this potential shortcoming by normalizing correlation. That is, it is better to make a decision based on normalized correlation, which is defined by

$$C_N(x,y) \;=\; \frac{C(x,y)}{\sqrt{E(x)}\sqrt{E(y)}} \;=\; \frac{1}{\sqrt{E(x)}\sqrt{E(y)}} \int_{t_1}^{t_2} x(t)\,y(t)\;dt$$

where $E(x)$ and $E(y)$ are the energies over the interval (t_1,t_2) of x and y, respectively. If the energies of the $v_i(t)$'s are the same, then signal $v_i(t)$ that has the largest correlation $C(r,v_i)$ also has the largest normalized correlation $C_N(r,v_i)$. However, when the $v_i(t)$'s have different energies, the normalized correlation accounts properly for such and permits the decision to be properly based.

Having discussed correlation, we can now completely describe a typical correlating detector. Suppose we must decide which of the signals $v_1(t)$, $v_2(t)$, ..., $v_M(t)$ is contained in $r(t)$. The decision statistic calculator computes and outputs $C_N(r,v_1)$, $C_N(r,v_2)$, ..., $C_N(r,v_M)$. The decision maker makes finds the largest of these, and outputs the corresponding decision.

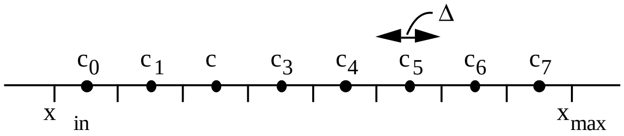
Comparison of Energy and Correlating Detectors: There are some situations where energy detectors cannot be used and some where correlating detectors cannot be used. For example, energy detectors cannot be used for the "which signal" problem when the signals have the same energy, which is often the case in digital communications. On the other hand, correlating detectors cannot be used when the precise shape of the signals is not known. For example, in Marconi's original transatlantic radio transmission, the transmitted signal was generated by a spark, with no known signal shape. Clearly, a correlating detector was out of the question!

In situations where both energy and correlating detectors can be used, it is usually found that the latter performs significantly better than the former, i.e. it makes fewer errors.

C. Signal Digitization for Data Storage and Transmission

In today's world where signal processing is increasingly done by general or special purpose computers, it is necessary to convert signals into digital form. Moreover signal storage and transmission are increasingly done in digital fashion. Again, this necessitates conversion to digital form. Such conversion involves two steps: (1) sampling, and (2) representing each sample as a binary number. Both of these steps generally involve losses, i.e. changes to the signal. Sampling is the topic of Chapter 4 and will be extensively discussed there. Converting to bits will be the subject of one of our lab assignments. However, let us describe here the most elementary method of converting samples to bits, called *uniform scalar quantization*.

With uniform scalar quantization, if we wish to represent a sample value $x[n]$ with b bits, then as illustrated below for the case that $b = 3$, we divide the range of sample values, (x_{min},x_{max}) into 2^b nonoverlapping bins of width $\Delta = (x_{max}-x_{min})/2^b$. These bins are indexed from left to right by the integers $0, 1, 2, ..., 2^b-1$, and each of these integers is represented as a b-bit binary number. For example, if $b = 3$, then $5 \Leftrightarrow 101$. Let $x_i = x_{min} + \Delta/2 + i\Delta$ denote the center of the ith bin. Now, if the sample $x[n]$ to be quantized lies in the ith bin, then we represent it by the binary representation of i, and we consider $x[n]$ to have been *quantized* to the value c_i. Note that when using this binary number in a processing task, we consider it to represent the value c_i, and must act accordingly. Actually, if the processing is done in a general purpose computer, we might convert i to binary using one of the standard conventions that are convenient for doing arithmetic, such as "two's complement".



bin index:	0	1	2	3	4	5	6	7
binary representation:	000	001	10	011	100	101	10	111

A system that does both sampling and uniform scalar quantization is called an *analog-to-digital converter.*

There are more sophisticated methods for converting samples to bits that produce many fewer bits. These are generally called *data compression* methods. Examples include JPEG image compression, MP3 audio compression, and CELP speech compression, which is the system used in digital cellular telephones, digital answering machines, and the like. A simplified version of a JPEG like image compression system is included in one of the lab assignments. Generally speaking, data compression is done in order to reduce the amount of memory needed to store a signal or the amount of time needed to transmit a signal. When the signal actually needs to be processed or played, the compressed representation must ordinarily be changed back into a representation like the one produced by a uniform scalar quantizer. This is called *decompression*.

Concluding Remarks

Having discussed several basic signal processing tasks, it should be mentioned that from now on, we will not focus on them in future lectures. Instead we will focus on developing tools and techniques that enable systems to perform these tasks well. In particular, we will discuss sampling (Chapter 4 of our text), spectra (Chapter 3 and handouts) and linear filters (Chapters 5-8). Although these signal processing tasks will not be the focus of the lectures, from time to time we will discuss how the techniques being developed in lecture apply to them. On the other hand, these basic signal processing tasks will be the focus of a number of the lab assignments in this course.