1

Introduction to Signal Processing

Signal processing is the name given to the procedures used on measured data to reveal the information contained in the measurements. These procedures essentially rely on various transformations that are mathematically based and which are implemented using digital techniques. The wide availability of software to carry out digital signal processing (DSP) with such ease now pervades all areas of science, engineering, medicine, and beyond. This ease can sometimes result in the analyst using the wrong tools – or interpreting results incorrectly because of a lack of appreciation or understanding of the assumptions or limitations of the method employed.

This text is directed at providing a user's guide to linear system identification. In order to reach that end we need to cover the groundwork of Fourier methods, random processes, system response and optimization. Recognizing that there are many excellent texts on this,[1] why should there be yet another? The aim is to present the material from a user's viewpoint. Basic concepts are followed by examples and structured MATLAB® exercises allow the user to 'experiment'. This will not be a story with the punch-line at the end – we actually start in this chapter with the intended end point.

The aim of doing this is to provide reasons and motivation to cover some of the underlying theory. It will also offer a more rapid guide through methodology for practitioners (and others) who may wish to 'skip' some of the more 'tedious' aspects. In essence we are recognizing that it is not always necessary to be fully familiar with every aspect of the theory to be an effective practitioner. But what is important is to be aware of the limitations and scope of one's analysis.

[1] See for example Bendat and Piersol (2000), Brigham (1988), Hsu (1970), Jenkins and Watts (1968), Oppenheim and Schafer (1975), Otnes and Enochson (1978), Papoulis (1977), Randall (1987), etc.

Fundamentals of Signal Processing for Sound and Vibration Engineers
K. Shin and J. K. Hammond.

The Aim of the Book

We are assuming that the reader wishes to understand and use a widely used approach to 'system identification'. By this we mean we wish to be able to characterize a physical process in a quantified way. The object of this quantification is that it reveals information about the process and accounts for its behaviour, and also it allows us to predict its behaviour in future environments.

The 'physical processes' could be anything, e.g. vehicles (land, sea, air), electronic devices, sensors and actuators, biomedical processes, etc., and perhaps less 'physically based' socio-economic processes, and so on. The complexity of such processes is unlimited – and being able to characterize them in a quantified way relies on the use of physical 'laws' or other 'models' usually phrased within the language of mathematics. Most science and engineering degree programmes are full of courses that are aimed at describing processes that relate to the appropriate discipline. We certainly do not want to go there in this book – life is too short! But we still want to characterize these systems – with the minimum of effort and with the maximum effect.

This is where 'system theory' comes to our aid, where we employ descriptions or models – abstractions from the 'real thing' – that nevertheless are able to capture what may be fundamentally common, to large classes of the phenomena described above. In essence what we do is simply to watch what 'a system' does. This is of course totally useless if the system is 'asleep' and so we rely on some form of activation to get it going – in which case it is logical to watch (and measure) the particular activation and measure some characteristic of the behaviour (or response) of the system.

In 'normal' operation there may be many activators and a host of responses. In most situations the activators are not separate discernible processes, but are distributed. An example of such a system might be the acoustic characteristics of a concert hall when responding to an orchestra and singers. The sources of activation in this case are the musical instruments and singers, the system is the auditorium, including the members of the audience, and the responses may be taken as the sounds heard by each member of the audience.

The complexity of such a system immediately leads one to try and conceptualize something simpler. Distributed activation might be made more manageable by 'lumping' things together, e.g. a piano is regarded as several separate activators rather than continuous strings/sounding boards all causing acoustic waves to emanate from each point on their surfaces. We might start to simplify things as in Figure 1.1.

This diagram is a model of a greatly simplified system with several actuators – and the several responses as the sounds heard by individual members of the audience. The arrows indicate a 'cause and effect' relationship – and this also has implications. For example, the figure implies that the 'activators' are unaffected by the 'responses'. This implies that there is no 'feedback' – and this may not be so.

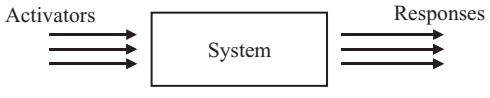


Figure 1.1 Conceptual diagram of a simplified system

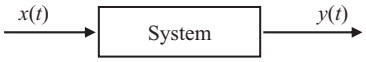


Figure 1.2 A single activator and a single response system

Having got this far let us simplify things even further to a single activator and a single response as shown in Figure 1.2. This may be rather 'distant' from reality but is a widely used model for many processes.

It is now convenient to think of the activator $x(t)$ and the response $y(t)$ as time histories. For example, $x(t)$ may denote a voltage, the system may be a loudspeaker and $y(t)$ the pressure at some point in a room. However, this time history model is just one possible scenario. The activator x may denote the intensity of an image, the system is an optical device and y may be a transformed image. Our emphasis will be on the time history model generally within a sound and vibration context.

The box marked 'System' is a convenient catch-all term for phenomena of great variety and complexity. From the outset, we shall impose major constraints on what the box represents – specifically systems that are **linear**[2] and **time invariant**.[3] Such systems are very usefully described by a particular feature, namely their response to an **ideal impulse**,[4] and their corresponding behaviour is then the **impulse response**.[5] We shall denote this by the symbol $h(t)$.

Because the system is linear this rather 'abstract' notion turns out to be very useful in predicting the response of the system to any arbitrary input. This is expressed by the **convolution**[6] of input $x(t)$ and system $h(t)$ sometimes abbreviated as

$$y(t) = h(t) * x(t) \tag{1.1}$$

where '$*$' denotes the convolution operation. Expressed in this form the system box is filled with the characterization $h(t)$ and the (mathematical) mapping or transformation from the input $x(t)$ to the response $y(t)$ is the convolution integral.

System identification now becomes the problem of measuring $x(t)$ and $y(t)$ and deducing the impulse response function $h(t)$. Since we have three quantitative terms in the relationship (1.1), but (assume that) we know two of them, then, in principle at least, we should be able to find the third. The question is: how?

Unravelling Equation (1.1) as it stands is possible but not easy. Life becomes considerably easier if we apply a transformation that maps the convolution expression to a multiplication. One such transformation is the **Fourier transform**.[7] Taking the **Fourier transform of the convolution**[8] in Equation (1.1) produces

$$Y(f) = H(f)X(f) \tag{1.2}$$

* Words in bold will be discussed or explained at greater length later.
[2] See Chapter 4, Section 4.7.
[3] See Chapter 4, Section 4.7.
[4] See Chapter 3, Section 3.2, and Chapter 4, Section 4.7.
[5] See Chapter 4, Section 4.7.
[6] See Chapter 4, Section 4.7.
[7] See Chapter 4, Sections 4.1 and 4.4.
[8] See Chapter 4, Sections 4.4 and 4.7.

where f denotes frequency, and $X(f)$, $H(f)$ and $Y(f)$ are the transforms of $x(t)$, $h(t)$ and $y(t)$. This achieves the unravelling of the input–output relationship as a straightforward multiplication – in a 'domain' called the **frequency domain**.[9] In this form the system is characterized by the quantity $H(f)$ which is called the system **frequency response function (FRF)**.[10]

The problem of 'system identification' now becomes the calculation of $H(f)$, which seems easy: that is, divide $Y(f)$ by $X(f)$, i.e. divide the Fourier transform of the output by the Fourier transform of the input. As long as $X(f)$ is never zero this seems to be the end of the story – but, of course, it is not. Reality interferes in the form of 'uncertainty'. The measurements $x(t)$ and $y(t)$ are often not measured perfectly – disturbances or 'noise' contaminates them – in which case the result of dividing two transforms of contaminated signals will be of limited and dubious value.

Also, the actual excitation signal $x(t)$ may itself belong to a class of **random**[11] signals – in which case the straightforward transformation (1.2) also needs more attention. It is this 'dual randomness' of the actuating (and hence response) signal and additional contamination that is addressed in this book.

The Effect of Uncertainty

We have referred to randomness or uncertainty with respect to both the actuation and response signal and additional noise on the measurements. So let us redraw Figure 1.2 as in Figure 1.3.

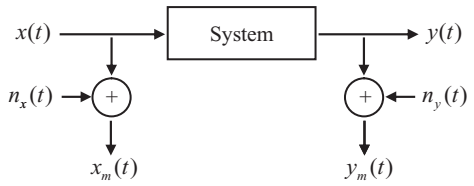


Figure 1.3 A single activator/response model with additive noise on measurements

In Figure 1.3, x and y denote the actuation and response signals as before – which may themselves be random. We also recognize that x and y are usually not directly measurable and we model this by including disturbances written as n_x and n_y which add to x and y – so that the actual measured signals are x_m and y_m. Now we get to the crux of the system identification: that is, on the basis of (noisy) measurements x_m and y_m, what is the system?

We conceptualize this problem pictorially. Imagine plotting y_m against x_m (ignore for now what x_m and y_m might be) as in Figure 1.4.

Each point in this figure is a 'representation' of the measured response y_m corresponding to the measured actuation x_m.

System identification, in this context, becomes one of establishing a relationship between y_m and x_m such that it somehow relates to the relationship between y and x. The noises are a

[9] See Chapter 2, Section 2.1.
[10] See Chapter 4, Section 4.7.
[11] See Chapter 7, Section 7.2.

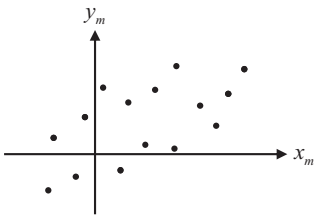


Figure 1.4 A plot of the measured signals y_m versus x_m

nuisance, but we are stuck with them. This is where 'optimization' comes in. We try and find a relationship between x_m and y_m that seeks a 'systematic' link between the data points which suppresses the effects of the unwanted disturbances.

The simplest conceptual idea is to 'fit' a linear relationship between x_m and y_m. Why linear? Because we are restricting our choice to the simplest relationship (we could of course be more ambitious). The procedure we use to obtain this fit is seen in Figure 1.5 where the slope of the straight line is adjusted until the match to the data seems best.

This procedure must be made systematic – so we need a measure of how well we fit the points. This leads to the need for a specific measure of fit and we can choose from an unlimited number. Let us keep it simple and settle for some obvious ones. In Figure 1.5, the closeness of the line to the data is indicated by three measures e_y, e_x and e_T. These are regarded as errors which are measures of the 'failure' to fit the data. The quantity e_y is an error in the y direction (i.e. in the output direction). The quantity e_x is an error in the x direction (i.e. in the input direction). The quantity e_T is orthogonal to the line and combines errors in both x and y directions.

We might now look at ways of adjusting the line to minimize e_y, e_x, e_T or some convenient 'function' of these quantities. This is now phrased as an optimization problem. A most convenient function turns out to be an average of the squared values of these quantities ('convenience' here is used to reflect not only physical meaning but also mathematical 'niceness'). Minimizing these three different measures of closeness of fit results in three correspondingly different slopes for the straight line; let us refer to the slopes as m_y, m_x, m_T. So which one should we use as the best? The choice will be strongly influenced by our prior knowledge of the nature of the measured data – specifically whether we have some idea of the dominant causes of error in the departure from linearity. In other words, some knowledge of the relative magnitudes of the noise on the input and output.

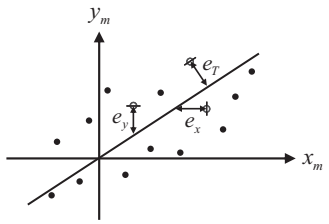


Figure 1.5 A linear fit to measured data

We could look to the figure for a guide:

- m_y seems best when errors occur on y, i.e. errors on output e_y;
- m_x seems best when errors occur on x, i.e. errors on input e_x;
- m_T seems to make an attempt to recognize that errors are on both, i.e. e_T.

We might now ask how these rather simple concepts relate to 'identifying' the system in Figure 1.3. It turns out that they are directly relevant and lead to three different estimators for the system frequency response function $H(f)$. They have come to be referred to in the literature by the notation $\boldsymbol{H_1(f)}$, $\boldsymbol{H_2(f)}$ and $\boldsymbol{H_T(f)}$,[12] and are the analogues of the slopes m_y, m_x, m_T, respectively.

We have now mapped out what the book is essentially about in Chapters 1 to 10. The book ends with a chapter that looks into the implications of multi-input/output systems.

1.1 DESCRIPTIONS OF PHYSICAL DATA (SIGNALS)

Observed data representing a physical phenomenon will be referred to as a time history or a *signal*. Examples of signals are: temperature fluctuations in a room indicated as a function of time, voltage variations from a vibration transducer, pressure changes at a point in an acoustic field, etc. The physical phenomenon under investigation is often translated by a transducer into an electrical equivalent (voltage or current) and if displayed on an oscilloscope it might appear as shown in Figure 1.6. This is an example of a *continuous* (or *analogue*) signal.

In many cases, data are *discrete* owing to some inherent or imposed sampling procedure. In this case the data might be characterized by a sequence of numbers equally spaced in time. The sampled data of the signal in Figure 1.6 are indicated by the crosses on the graph shown in Figure 1.7.

Figure 1.6 A typical continuous signal from a transducer output

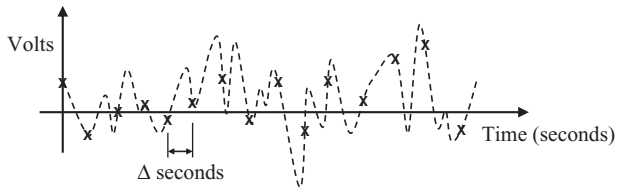


Figure 1.7 A discrete signal sampled at every Δ seconds (marked with ×)

[12] See Chapter 9, Section 9.3.

Figure 1.8 An example of a signal where time is not the natural independent variable

For continuous data we use the notation $x(t)$, $y(t)$, etc., and for discrete data various notations are used, e.g. $x(n\Delta)$, $x(n)$, x_n $(n = 0, 1, 2, \ldots)$.

In certain physical situations, 'time' may not be the natural independent variable; for example, a plot of road roughness as a function of spatial position, i.e. $h(\xi)$ as shown in Figure 1.8. However, for uniformity we shall use time as the independent variable in all our discussions.

1.2 CLASSIFICATION OF DATA

Time histories can be broadly categorized as shown in Figure 1.9 (chaotic signals are added to the classifications given by Bendat and Piersol, 2000). A fundamental difference is whether a signal is *deterministic* or *random*, and the analysis methods are considerably different depending on the 'type' of the signal. Generally, signals are mixed, so the classifications of Figure 1.9 may not be easily applicable, and thus the choice of analysis methods may not be apparent. In many cases some prior knowledge of the system (or the signal) is very helpful for selecting an appropriate method. However, it must be remembered that this prior knowledge (or assumption) may also be a source of misleading the results. Thus it is important to remember the First Principle of Data Reduction (Ables, 1974)

> *The result of any transformation imposed on the experimental data shall incorporate and be consistent with all relevant data and be maximally non-committal with regard to unavailable data.*

It would seem that this statement summarizes what is self-evident. But how often do we contravene it – for example, by 'assuming' that a time history is zero outside the extent of a captured record?

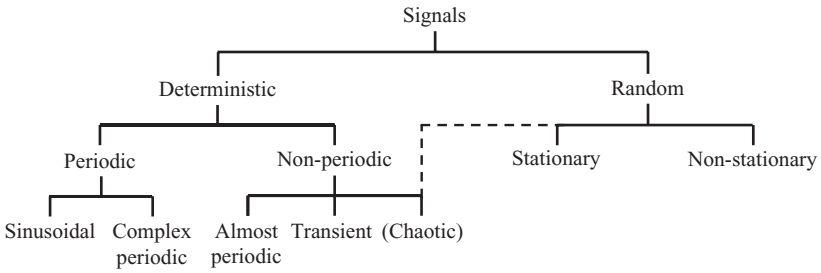


Figure 1.9 Classification of signals

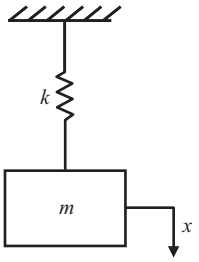


Figure 1.10 A simple mass–spring system

Nonetheless, we need to start somewhere and signals can be broadly classified as being either *deterministic* or *non-deterministic* (*random*). Deterministic signals are those whose behaviour can be predicted exactly. As an example, a mass–spring oscillator is considered in Figure 1.10. The equation of motion is $m\ddot{x} + kx = 0$ (x is displacement and $\ddot{x}$ is acceleration). If the mass is released from rest at a position $x(t) = A$ and at time $t = 0$, then the displacement signal can be written as

$$x(t) = A\cos\left(\sqrt{k/m}\cdot t\right) \quad t \geq 0 \tag{1.3}$$

In this case, the displacement $x(t)$ is known exactly for all time. Various types of deterministic signals will be discussed later. Basic analysis methods for deterministic signals are covered in Part I of this book. Chaotic signals are not considered in this book.

Non-deterministic signals are those whose behaviour cannot be predicted exactly. Some examples are vehicle noise and vibrations on a road, acoustic pressure variations in a wind tunnel, wave heights in a rough sea, temperature records at a weather station, etc. Various terminologies are used to describe these signals, namely *random processes* (*signals*), *stochastic processes*, *time series*, and the study of these signals is called *time series analysis*. Approaches to describe and analyse random signals require probabilistic and statistical methods. These are discussed in Part II of this book.

The classification of data as being deterministic or random might be debatable in many cases and the choice must be made on the basis of knowledge of the physical situation. Often signals may be modelled as being a mixture of both, e.g. a deterministic signal 'embedded' in unwanted random disturbances (noise).

In general, the purpose of signal processing is the extraction of information from a signal, especially when it is difficult to obtain from direct observation. The methodology of extracting information from a signal has three key stages: (i) acquisition, (ii) processing, (iii) interpretation. To a large extent, signal acquisition is concerned with *instrumentation*, and we shall treat some aspects of this, e.g. **analogue-to-digital conversion**.[13] However, in the main, we shall assume that the signal is already acquired, and concentrate on stages (ii) and (iii).

[13] See Chapter 5, Section 5.3.

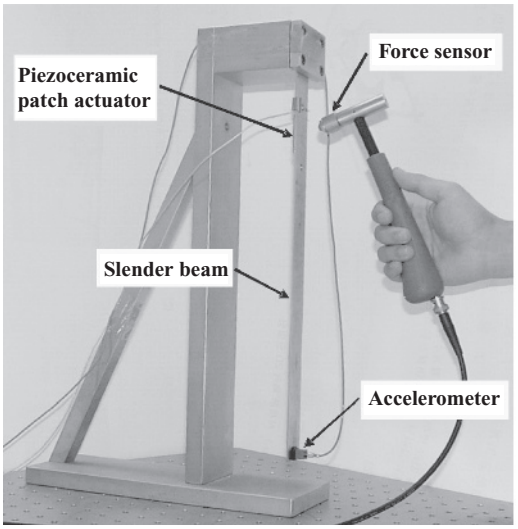


Figure 1.11 A laboratory setup

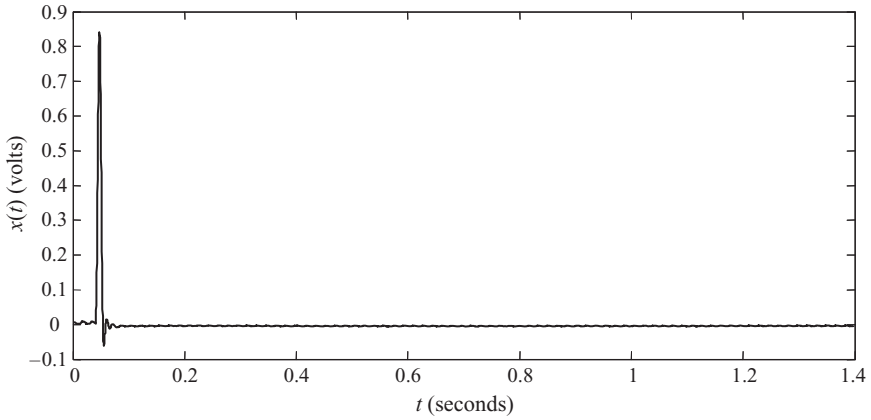
Some 'Real' Data

Let us now look at some signals measured experimentally. We shall attempt to fit the observed time histories to the classifications of Figure 1.9.

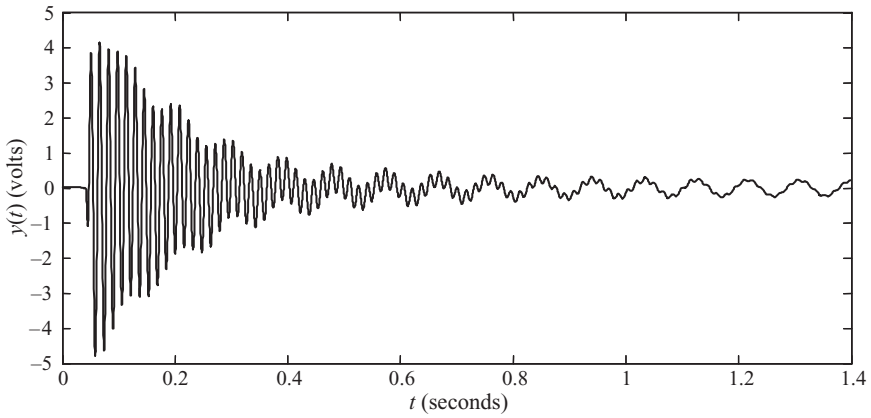
(a) Figure 1.11 shows a laboratory setup in which a slender beam is suspended vertically from a rigid clamp. Two forms of excitation are shown. A small piezoceramic PZT (Piezoelectric Zirconate Titanate) patch is used as an actuator which is bonded on near the clamped end. The instrumented hammer (impact hammer) is also used to excite the structure. An accelerometer is attached to the beam tip to measure the response. We shall assume here that digitization effects (**ADC quantization**, **aliasing**)[14] have been adequately taken care of and can be ignored. A sharp tap from the hammer to the structure results in Figures 1.12(a) and (b). Relating these to the classification scheme, we could reasonably refer to these as deterministic transients. Why might we use the deterministic classification? Because we expect replication of the result for 'identical' impacts. Further, from the figures the signals appear to be essentially noise free. From a systems points of view, Figure 1.12(a) is $x(t)$ and 1.12(b) is $y(t)$ and from these two signals we would aim to deduce the characteristics of the beam.

(b) We now use the PZT actuator, and Figures 1.13(a) and (b) now relate to a random excitation. The source is a **band-limited**,[15] **stationary**,[16] **Gaussian process**,[17] and in the steady state (i.e. after starting transients have died down) the response should also be stationary. However, on the basis of the visual evidence the response is not evidently stationary (or is it?), i.e. it seems modulated in some way. This demonstrates the difficulty in classification. As it

[14] See Chapter 5, Sections 5.1–5.3.
[15] See Chapter 5, Section 5.2, and Chapter 8, Section 8.7.
[16] See Chapter 8, Section 8.3.
[17] See Chapter 7, Section 7.3.



(a) Impact signal measured from the force sensor (impact hammer)



(b) Response signal to the impact measured from the accelerometer

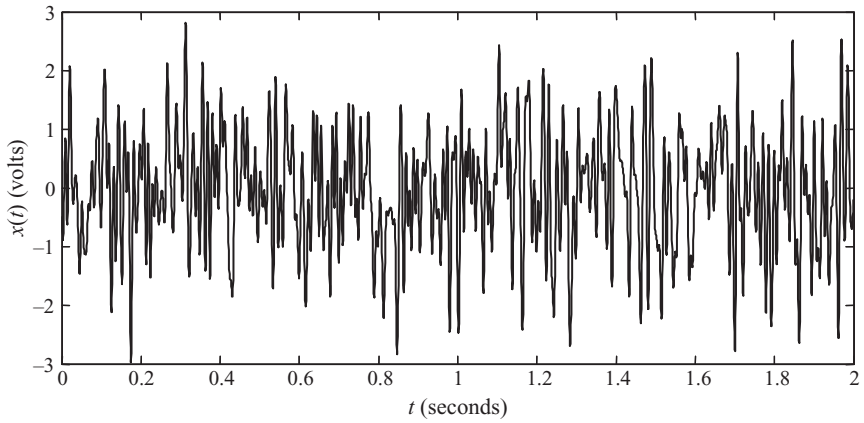
Figure 1.12 Example of deterministic transient signals

happens, the response is a narrow-band stationary random process (due to the filtering action of the beam) which is characterized by an amplitude-modulated appearance.

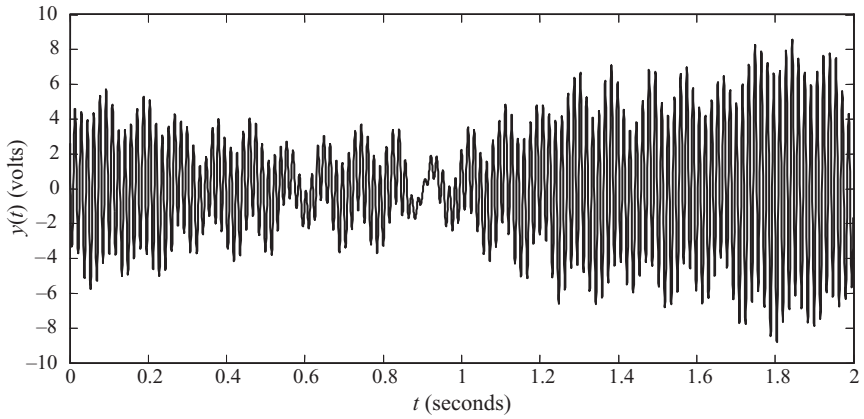
(c) Let us look at a signal from a machine rotating at a constant rate. A tachometer signal is taken from this. As in Figure 1.14(a), this is one that could reasonably be classified as periodic, although there are some discernible differences from period to period – one might ask whether this is simply an additive low-level noise.

(d) Another repetitive signal arises from a telephone tone shown in Figure 1.14(b). The tonality is 'evident' from listening to it and its appearance is 'roughly' periodic; it is tempting to classify these signals as 'almost periodic'!

(e) Figure 1.15(a) represents the signal for a transformer 'hum', which again perceptually has a repetitive but complex structure and visually appears as possibly periodic with additive noise – or (perhaps) narrow-band random.



(a) Input random signal to the PZT (actuator) patch



(b) Response signal to the random excitation measured from the accelerometer

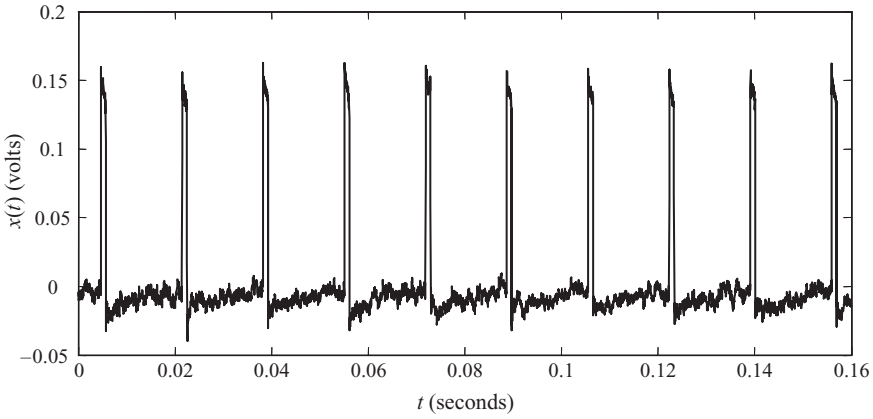
Figure 1.13 Example of stationary random signals

Figure 1.15(b) is a signal created by adding noise (broadband) to the telephone tone signal in Figure 1.14(b). It is not readily apparent that Figure 1.15(b) and Figure 1.15(a) are 'structurally' very different.

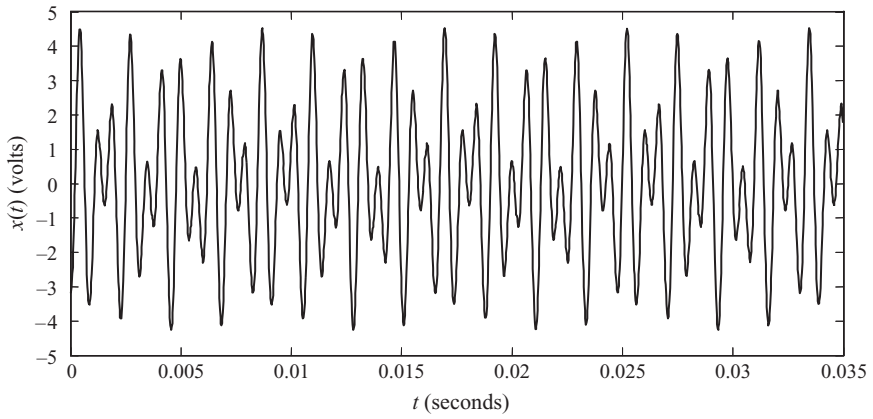
(f) Figure 1.16(a) is an acoustic recording of a helicopter flyover. The non-stationary structure is apparent – specifically, the increase in amplitude with reduction in range. What is not apparent are any other more complex aspects such as frequency modulation due to movement of the source.

(g) The next group of signals relate to practicalities that occur during acquisition that render the data of limited value (in some cases useless!).

The jagged stepwise appearance in Figure 1.17 is due to quantization effects in the ADC – apparent because the signal being measured is very small compared with the voltage range of the ADC.



(a) Tachometer signal from a rotating machine

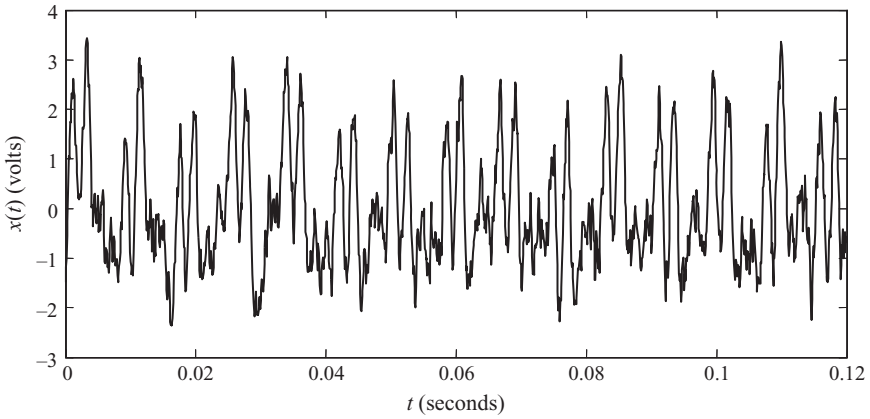


(b) Telephone tone (No. 8) signal

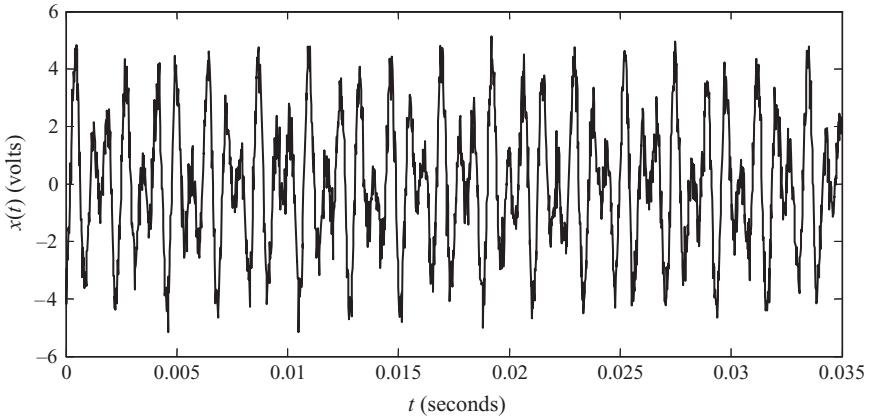
Figure 1.14 Example of periodic (and almost periodic) signals

(h) Figures 1.18(a), (b) and (c) all display flats at the top and bottom (positive and negative) of their ranges. This is characteristic of 'clipping' or saturation. These have been synthesized by clipping the telephone signal in Figure 1.14(b), the band-limited random signal in Figure 1.13(a) and the accelerometer signal in Figure 1.12(b). Clipping is a nonlinear effect which 'creates' spurious frequencies and essentially destroys the credibility of any Fourier transformation results.

(i) Lastly Figures 1.19(a) and (b) show what happens when 'control' of an experiment is not as tight as it might be. Both signals are the free responses of the cantilever beam shown in Figure 1.11. Figure 1.19(a) shows the results of the experiment performed on a vibration-isolated optical table. The signal is virtually noise free. Figure 1.19(b) shows the results of the same experiment, but performed on a normal bench-top table. The signal is now contaminated with noise that may come from various external sources. Note that we may not be able to control our experiments as carefully as in Figure 1.19(a), but, in fact, it is a signal as in



(a) Transformer 'hum' noise



(b) Telephone tone (No. 8) signal with noise

Figure 1.15 Example of periodic signals with additive noise

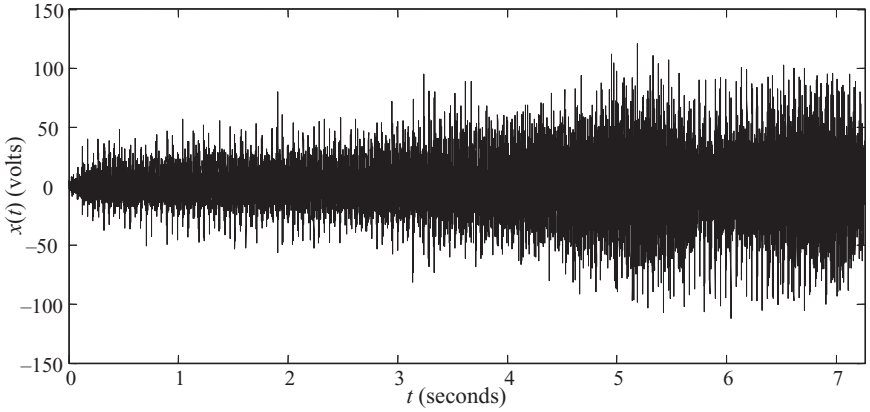


Figure 1.16 Example of a non-stationary signal (helicopter flyover noise)

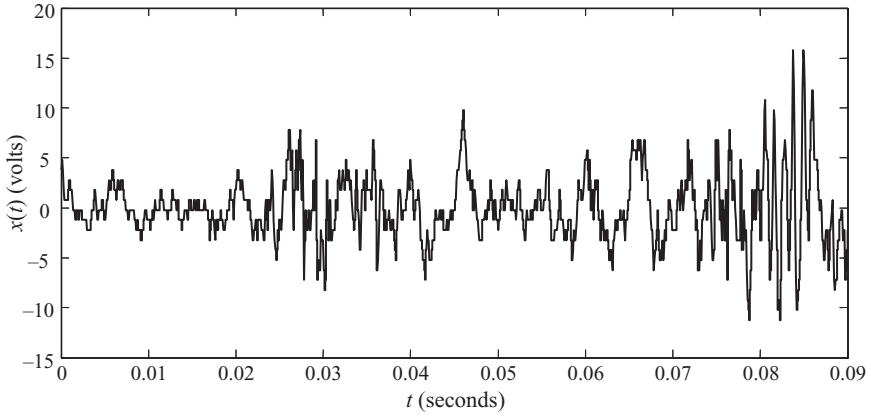


Figure 1.17 Example of low dynamic range

Figure 1.19(b) which we often deal with. Thus, the nature of uncertainty in the measurement process is again emphasized (see Figure 1.3).

The Next Stage

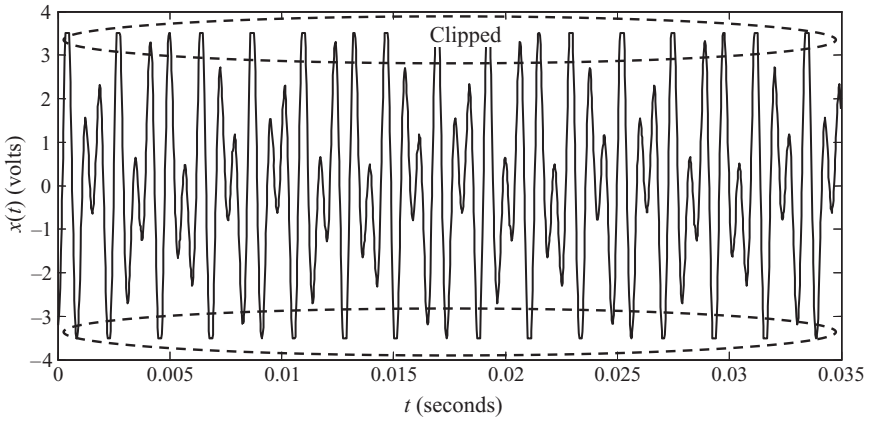
Having introduced various classes of signals we can now turn to the principles and details of how we can model and analyse the signals. We shall use Fourier-based methods – that is, we essentially model the signal as being composed of sine and cosine waves and tailor the processing around this idea. We might argue that we are imposing/assuming some prior information about the signal – namely, that sines and cosines are appropriate descriptors. Whilst this may seem constraining, such a 'prior model' is very effective and covers a wide range of phenomena. This is sometimes referred to as a *non-parametric* approach to signal processing.

So, what might be a 'parametric' approach? This can again be related to modelling. We may have additional 'prior information' as to how the signal has been generated, e.g. a result of filtering another signal. This notion may be extended from the knowledge that this generation process is indeed 'physical' to that of its being 'notional', i.e. another model. Specifically Figure 1.20 depicts this when $s(t)$ is the 'measured' signal, which is conceived to have arisen from the action of a system being driven by a very fundamental signal – in this case so-called **white noise**[18] $w(t)$.

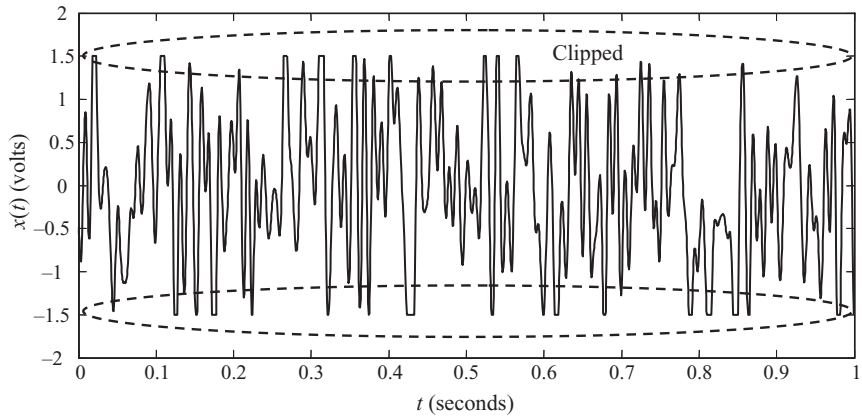
Phrased in this way the analysis of the signal $s(t)$ can now be transformed into a problem of determining the details of the system. The system could be characterized by a set of parameters, e.g. it might be mathematically represented by differential equations and the parameters are the coefficients. Set up like this, the analysis of $s(t)$ becomes one of system parameter estimation – hence this is a parametric approach.

The system could be linear, time varying or nonlinear depending on one's prior knowledge, and could therefore offer advantages over Fourier-based methods. However, we shall not be pursuing this approach in this book and will get on with the Fourier-based methods instead.

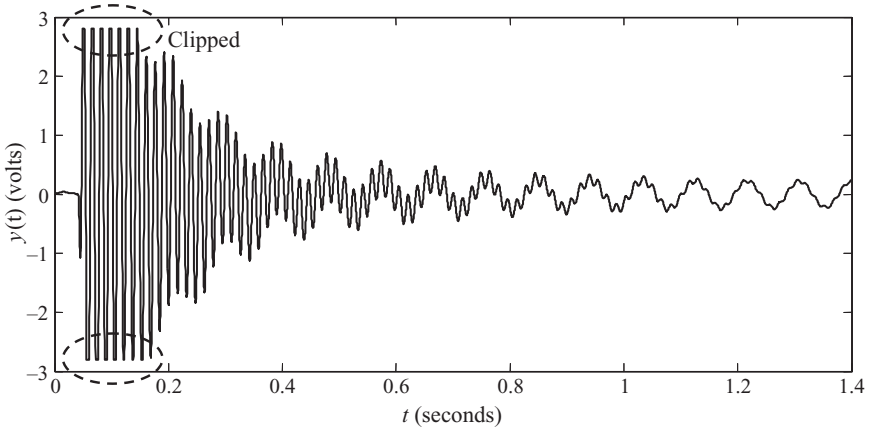
[18] See Chapter 8, Section 8.6.



(a) Clipped (almost) periodic signal

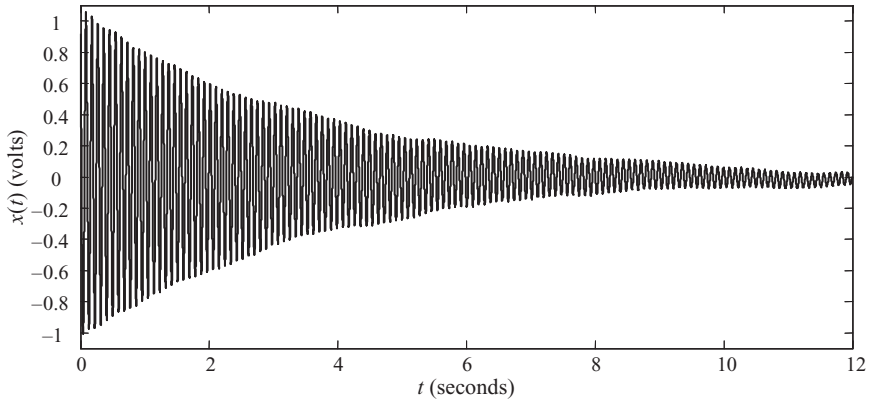


(b) Clipped random signal

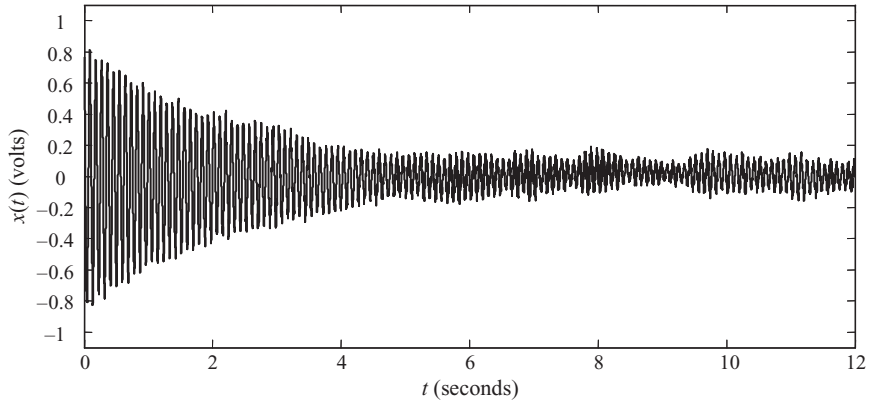


(c) Clipped transient signal

Figure 1.18 Examples of clipped signals



(a) Signal is measured on the optical table (fitted with a vibration isolator)



(b) Signal is measured on the ordinary bench-top table

Figure 1.19 Examples of experimental noise

Figure 1.20 A white-noise-excited system

We have emphasized that this is a book for practitioners and users of signal processing, but note also that there should be sufficient detail for completeness. Accordingly we have chosen to highlight some main points using a light grey background. From Chapter 3 onwards there is a reasonable amount of mathematical content; however, a reader may wish to get to the main points quickly, which can be done by using the highlighted sections. The details supporting these points are in the remainder of the chapter adjacent to these sections and in the appendices. Examples and MATLAB exercises illustrate the concepts. A superscript notation is used to denote the relevant MATLAB example given in the last section of the chapter, e.g. see the superscript ($^{\mathbf{M2.1}}$) in page 21 for MATLAB Example 2.1 given in page 26.