University of Hertfordshire

Sensors & Signal Processing

Diogo Montalvão

PhD CEng MIMechE FHEA

Senior Lecturer in Mechanical Systems

September 2014

www.herts.ac.uk

Summary

This text presents the fundamentals on sensors and signal processing, with emphasises on Mechatronics and applications. Sensors are used to measure signals that present changes in the time domain, e.g. waveforms or digital steps. Different technologies have been developed over the years in order to sense many different physical quantities, such as temperature, flow, force, acceleration, position, sound pressure and intensity of light, among others. Because of their varying nature, all these quantities may be measured under the form of waveforms. However, waveforms - which are analogue signals – are often found difficult to interpret in the time domain and a transformation into the frequency domain is required. The Fourier transform still is the most popular technique used today for converting a time signal into a frequency spectrum. Nevertheless, in signal processing, an analogue-to-digital conversion (ADC) of the time signal is required at some stage, even if the Fourier transform is not used. When proper treatment and filtering approaches are not followed, important features in the signal may be attenuated, and others may be falsely indicated. This text discusses how signals can be measured in order to avoid common pitfalls in signal acquisition and processing. The theoretical background is set in a comprehensive yet practical way.

Table of contents

Chapter 1

Introduction

Sensors exist in every system that interacts with the surrounding world. Any non-arbitrary decision requires a sensor to collect, transmit and process data somewhere. We have all been using sensors since we were born: from our eyes to our nervous system, our body is a complex network of sensors that continuously monitor and send signals to our brain for processing and analysis. The human body has sensors that are capable of acquiring colour images, sound, flavour, odor, texture, and temperature. However, it is not equipped with a sensor that is capable of detecting a magnetic field, like a compass does. On the other hand, the human body sensors have some known limitations as well. In acoustics, a rule of thumb is the rule of 20: roughly speaking, a healthy young human can detect sounds within the $20\ Hz$ to $20\ kHz$ frequency range for pressure waves between $20\ \mu Pa$ and $20\ Pa$ (threshold of pain). Still, a dog can hear sounds up to 50 kHz and a bat is capable of detecting sounds up to 100 kHz [1, 2].

Likewise, applications in mechatronics require sensors to capture light, sound, motion, temperature, force, flow, current, etc. With the development of powerful energy and memory storage units, wireless technology and fast microcomputers, data can be continuously

monitored, processed, stored and made available for immediate use, without the need of further human interaction. However, raw signals are not necessarily useful before being translated into something “readable”. Signal processing is the discipline that allows transforming a signal into something useful without losing relevant information.

This text covers the fundamental aspects on signal processing and sensor technology, including: an introduction to the mathematical representation of harmonic and complex time signals, Fourier analysis and its implementation, an introduction to random signals, sampling and aliasing, windowing and leakage, filtering and smoothing, typical sensors and applications, and representation of signals using logarithmic scales.

Chapter 2

Signals

2.1 Types of Time Signals and Waveforms

A signal can be defined, in abstract, as "a function that conveys information about the behaviour or attributes of some phenomenon" [3]. Generally, their nature is electromagnetic (e.g., the change in current or voltage in an AC signal), mechanical (e.g., the change in velocity of an oscillating pendulum), acoustical (e.g., the change in sound pressure during speech) and visual (e.g., the change in light intensity when looking to a star in the sky), among others.

Time signals all share one thing in common regardless of their nature: they all describe the variation on the amplitude of a quantity with time. The problem, however, is how to extract

from these functions information, in an intelligible and useful way. Signals may be grouped into *deterministic* and *non-deterministic*. Deterministic signals may be represented by analytical equations. Hence, it is possible to predict, with exactitude, how the signal will be in a given moment in the future, as long as no changes are introduced into the system. On the other hand, non-deterministic signals cannot be represented by analytical equations. However, a system that produces non-deterministic signals can usually be treated as stochastic. Forecasts can be done based on probability and statistical methods, but it is not possible to predict, with exactitude, how the signal will be in a given moment in the future. In other words, if the change of the time signal is perfectly known, even if the signal is very irregular, then the signal is said to be *deterministic*. If the change on the time signal is not perfectly known but if it can be considered stochastic, i.e., the subsequent state of the system is determined probabilistically, then the signal is said to be *non-deterministic*.

Non-deterministic signals are also called *random* and can be divided into *stationary* and *non-stationary* signals. A stationary signal is one that is a stochastic process whose joint probability distribution does not change when shifted in time. Consequently, parameters such as the mean and variance, if they are present, also do not change over time and do not follow any trends [4]. One example of a non-deterministic stationary signal is the sound of rain hitting the roof of a car.

Deterministic signals are often divided into two main categories: *periodic* and *non-periodic*. The first ones are those that present the same amplitude, direction and position after a complete cycle in time is completed (*period*). The simplest form of a deterministic stationary signal is the *simple harmonic*, which is characterized by amplitude, frequency and phase (see section 2.2). Examples of predominantly harmonic signals are those generated from, for example, high speed turbines, pumps, electric motors, AC electric currents or the sound produced from a human while whistling. Other systems, like reciprocating engines (as in petrol and diesel cars), still produce deterministic signals but, because many different parts are rotating at different speeds, they produce simultaneous harmonic excitations at various frequencies. Nevertheless, these *complex harmonic* signals can be decomposed and represented by a sum of simple harmonic signals using, for example, a Fourier analysis (see chapter 3).

Non-periodic signals are *transient* signals that can still be described by a deterministic approach (for example, some random non-stationary signals can be composed by a series of transient signals, yet they are not-deterministic because of their random nature). These can be generated during the run-up and coast-down of machines or from periodic impulses, for example from press tools or musical percussion systems. The Dirac is a special case of a non-periodic impulse that still produces a deterministic signal, once the Impulse Response Function

is known (IRF) (see section 4.4), for instance transient bursts observed in some hi-fi systems during turn-on and turn-off.

Figure 1 is a block diagram illustrating the previous discussion on how signals are grouped.

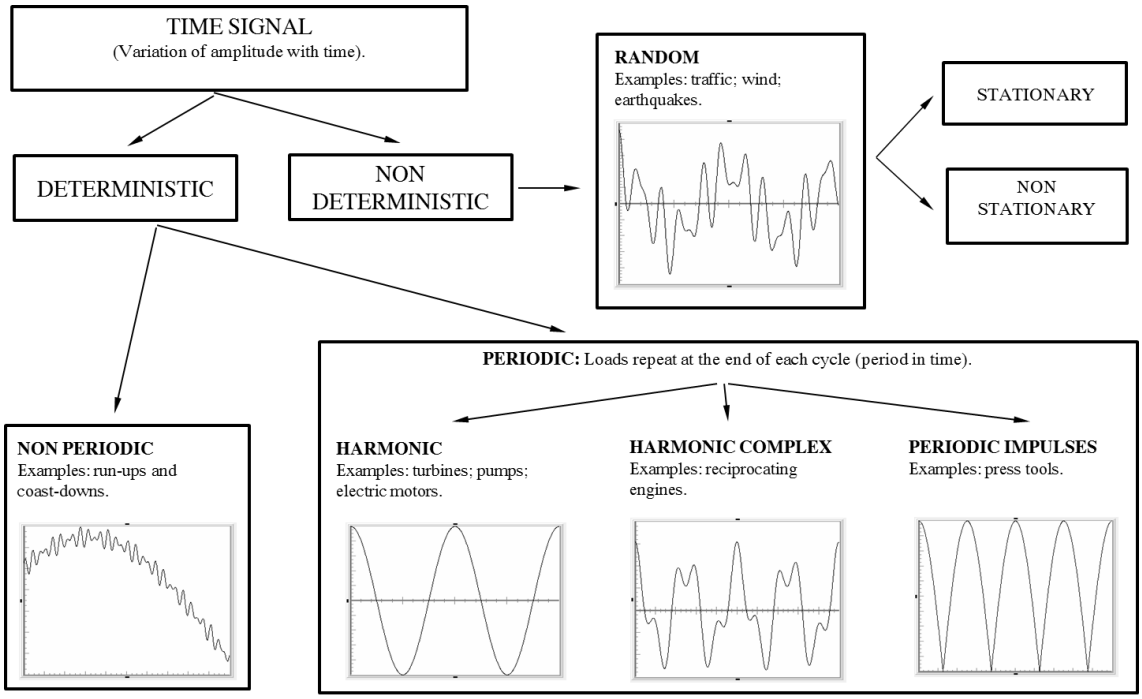


Figure 1 Types of time signals.

2.2 Harmonic Signals

2.2.1 Definition

Simple harmonic (often referred simply to as *harmonic*) signals are used to describe many deterministic systems, even if the functions are, from the point of view of the analyser, apparently rather complex (e.g., the case of complex harmonic waveforms). Harmonic signals are trigonometric circular functions that can be represented either as a function of time (amplitude *vs* time) or as rotating vectors in the Argand plane (real part *vs* imaginary part). This latter representation is going to be discussed in more detail in section 2.2.2.

Before introducing the mathematical equations describing the phenomenon, harmonic motion can be described by means of a simple example. First, let us imagine the simple case in which we have a mass m being hang with a spring with stiffness k, as represented in figure 2 (a). If the mass is disturbed from equilibrium and then suddenly released as shown in figure 2 (b), it will start oscillating up and down about its equilibrium position as illustrated in figure 2 (c). The distance X from the equilibrium position to one of the extremes is the *peak amplitude*

or *magnitude* and the total distance $2X$ travelled from one extreme to the other is the *peak-to-peak amplitude*.

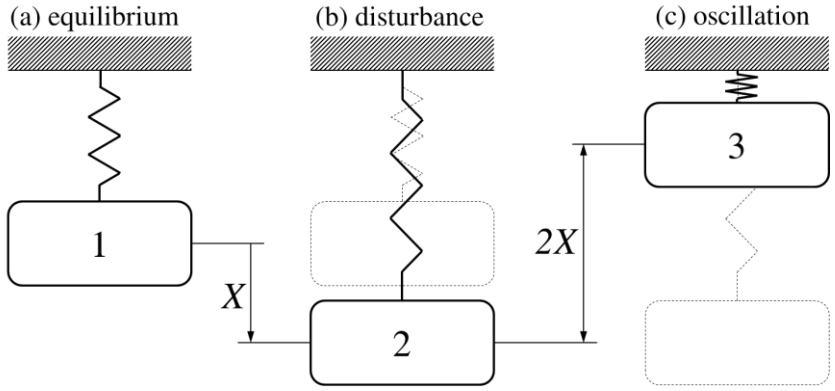


Figure 2 Oscillatory motion in a Single Degree of Freedom (SDOF) mass-spring system.

Thereby, the harmonic motion of this oscillating mass-spring system can be described as a sinusoidal function that depends on time t:

$$x(t) = X\sin(\omega t + \theta) \tag{1}$$

where X is the amplitude, ω is the angular frequency in rad/s and θ is the phase angle in rad. This is represented in figure 3 (b), where T is the period (time required for a complete cycle of oscillation to repeat itself). The period T is related to the angular frequency by:

$$f = \frac{1}{T} = \frac{\omega}{2\pi} \tag{2}$$

with f for frequency in cycles per second or *Hertz* (Hz). The frequency is more often expressed in Hz rather than in rad/s because it is easier to understand, although the rad/s is, in fact, the metric unit.

Figure 3 (a) represents the same signal as figure 3 (b), but in this case this is done in the Argand plane. Instead of the signal being represented by the trace of a point that moves up and down as time goes by, a vector is used. Its size is the amplitude X. This vector rotates about the origin O of a Cartesian system of coordinates with Real (horizontal) and Imaginary (vertical) axis, with constant angular velocity ω. The tip of the vector moves over a circle with radius X and the angle it makes with the horizontal axis at $t = 0$ is defined as the phase angle θ. This form of representation is further discussed in section 4.2.2.2.

The amplitude of the harmonic signal shown in figure 3, X, is also called *peak amplitude* x_{PK}, and the value $2X$ is called the *amplitude peak-to-peak* x_{PK-PK}. This is equally represented in figure 2.

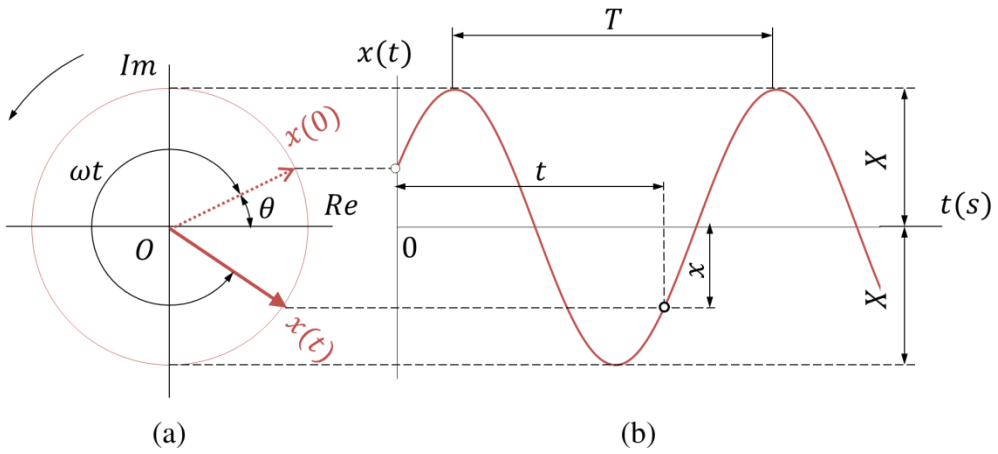


Figure 3 Simple harmonic signal represented in the Argand plane (a) and in an amplitude vs time plot (b).

2.2.2 Harmonic Motion in the Argand Plane

The simple harmonic motion, as written by equation (1), is a trigonometric function that can be alternatively represented as a circular vector rotating in the Argand plane. It will be shown that this is a much more convenient form of representation, since the arithmetics and differentiation of trigonometric functions is much more difficult in comparison to complex functions.

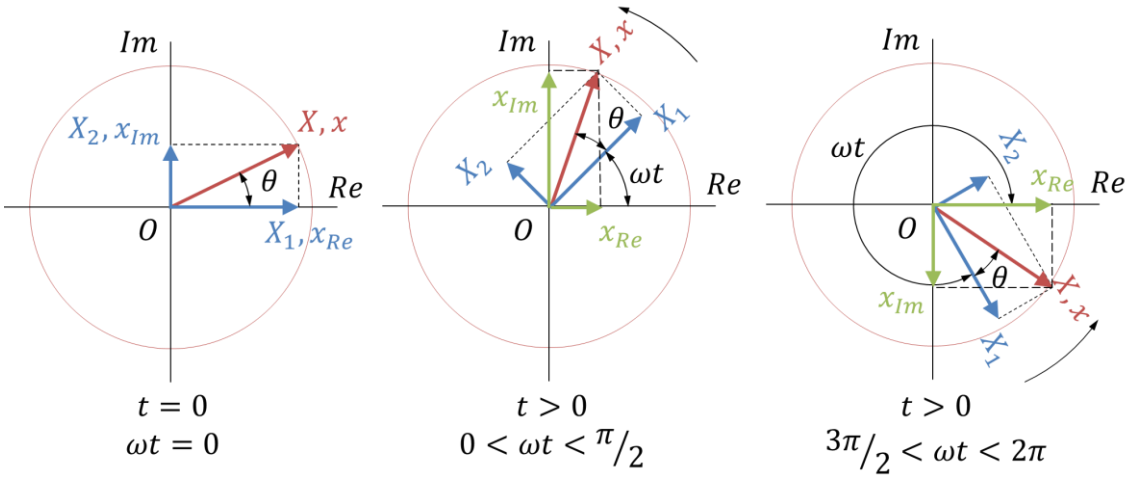


Figure 4 Representation of a circular vector rotating in the Argand Plane.

The Argand plane is composed by two mutually perpendicular orthogonal axes, that we call real (x-axis) and imaginary (y-axis). A period of oscillation is completed when a vector completes a full rotation about the origin O of the Argand plane system of coordinates (red vector in figure 4).

Now, we must recall that a vector is a quantity that is defined by magnitude and direction (which includes line of action and sense) and can be decomposed into coordinates. Two well-known systems of coordinates are the Cartesian and Polar. In the Cartesian system of coordinates, vector $x(t)$ is decomposed into its projections in the horizontal and vertical axes, which, in the Argand Plane, are $x_{Re}(t)$ and $x_{Im}(t)$, respectively:

$$x_{Re}(t) = X\cos(\omega t + \theta) \tag{3}$$

and

$$x_{Im}(t) = X\sin(\omega t + \theta) \tag{4}$$

One interesting aspect is that equations (1) and (4) are exactly the same. Also, equation (4) is equivalent to (3) with a $90°$ ($\pi/2\ rad$) phase shift. These vectors are represented by the colour green in figure 4.

The polar coordinates are, essentially, the magnitude X and the phase angle θ. These can be translated into the Cartesian system of coordinates previously defined by:

$$X = \sqrt{x_{Re}^2(t) + x_{Im}^2(t)} = \sqrt{X_1^2 + X_2^2} \tag{5}$$

and

$$\theta = \tan^{-1}\frac{x_{Im}(t)}{x_{Re}(t)} = \tan^{-1}\frac{X_2}{X_1} \tag{6}$$

Quantities X_1 and X_2 are, respectively, the real and imaginary components of x at $t = 0s$. While $x_{Re}(t)$ and $x_{Im}(t)$ are functions that depend on time (thus, quantities that change with time), X_1 and X_2 are constant quantities.

The quantity X expressed in equation (5) is a real number: it is a scalar. Thus, it only represents the magnitude of $x(t)$, but has no information about the direction. When we incorporate the direction (which is changing with time because the red vector in figure 4 is rotating about O), X becomes $x(t)$:

$$x(t) = X_{Re}(t) + iX_{Im}(t) \tag{7}$$

where $i = \sqrt{-1}$ is the imaginary number. When X is written as a complex number it is, in fact, a vector in the Argand plane. Its real part represents its projection on the horizontal axis of the Argand plane and its imaginary part represents its projection on the vertical axis. If we now replace the quantities in equation (7) with equations (3) and (4), we obtain:

$$x(t) = X\cos(\omega t + \theta) + iX\sin(\omega t + \theta) = X\mathrm{cis}(\omega t + \theta) \tag{8}$$

which is known to be equivalent to:

$$x(t) = Xe^{i(\omega t + \theta)} \tag{9}$$

In this latter formulation, $x(t)$ is, in fact, represented in terms of polar components: the amplitude (or magnitude) X, which is constant, and the angle formed by the vector with the real horizontal axis, $\omega t + \theta$, which changes with time.

2.2.3 Differentiation of Harmonic Signals

One interesting property about harmonic motion is that it can be easily differentiated. This can be useful, for instance, when comparing signals obtained from different sensors. For example, when measuring structural vibrations, the most common sensor used is the accelerometer (see section 5.2). This sensor measures the acceleration of oscillating mechanical systems. However, vibration level limits are many times specified as velocities. So, how can acceleration be translated into velocity? In harmonic motion, this is quite straightforward, as will be shown.

First, let us recall that velocity $\dot{x}$ is the rate of change of position x with time, and acceleration $\ddot{x}$ is the rate of change of velocity $\dot{x}$ with time, i.e.:

$$\dot{x} = \frac{dx}{dt} \tag{10}$$

$$\ddot{x} = \frac{d\dot{x}}{dt} = \frac{d^2x}{dt^2} \tag{11}$$

Thus, if equation 1 is the displacement, then from equation (10) the velocity is:

$$\begin{aligned}\dot{x}(t) &= \omega X \cos(\omega t + \theta) \\ &= \omega X \sin\left[\omega t + \left(\theta + \frac{\pi}{2}\right)\right]\end{aligned} \tag{12}$$

and from equation (11) the acceleration is:

$$\begin{aligned}\ddot{x}(t) &= -\omega^2 X \sin(\omega t + \theta) \\ &= \omega^2 X \sin[\omega t + (\theta + \pi)]\end{aligned} \tag{13}$$

Two important results can be extracted from this. The first one is that the amplitudes of displacement, velocity and acceleration are related to one another by the angular frequency ω. The second one is that there is a phase angle of $90°$ (*quadrature*) between displacement and velocity and between velocity and acceleration, and there is a phase angle of $180°$ (*phase opposition*) between displacement and acceleration.

The amplitudes of velocity and acceleration are, respectively:

$$\begin{aligned}V &= \omega X \\ A &= \omega V = \omega^2 X\end{aligned} \tag{14}$$

and the phase angles of velocity and acceleration are, respectively:

$$\begin{aligned}\vartheta &= \theta + \frac{\pi}{2} \\ \alpha &= \vartheta + \frac{\pi}{2} = \theta + \pi\end{aligned} \tag{15}$$

Thus, harmonic signals can be differentiated (or integrated) by simply multiplying their amplitudes (or dividing them) by the angular frequency and adding (or subtracting) $\pi/2\ rad$ to the phase. This simple process is represented schematically in figures 5 and 6.

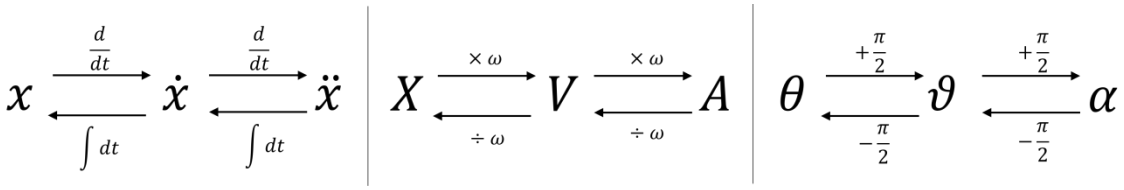


Figure 5: Relationships between displacement, velocity and acceleration in a harmonic signal.

The use of equation (1) is not very practical, however, because the conversion involves two operations: conversion of amplitude *and* conversion of phase. The use of the complex notation (equation (9)) allows converting both the amplitude and phase in a single operation.

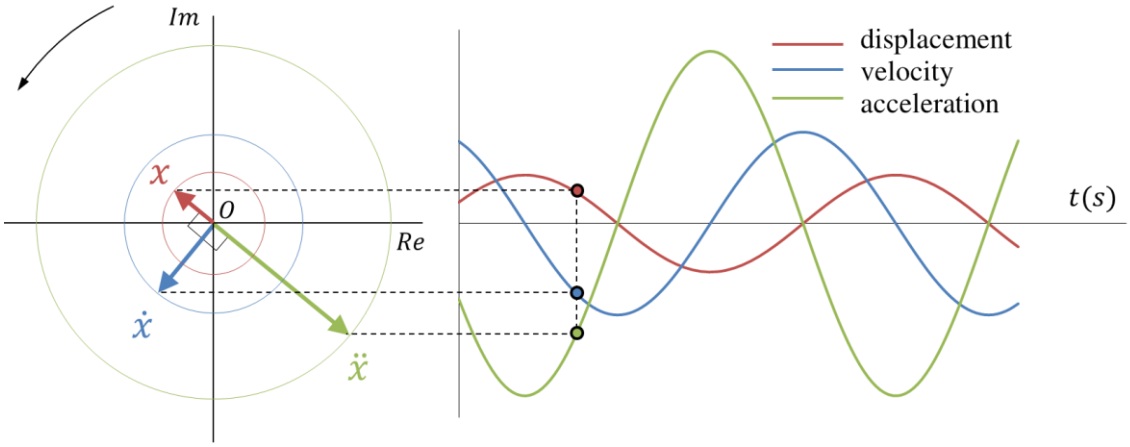


Figure 6 Graphical representation of the relationships between displacement, velocity and acceleration in a harmonic signal.

If we now differentiate equation (9) we obtain the complex forms for the velocity and acceleration:

$$\dot{x}(t) = i\omega X e^{i(\omega t+\theta)} \tag{16}$$

$$\ddot{x}(t) = (i\omega)^2 X e^{i(\omega t+\theta)} = -\omega^2 X e^{i(\omega t+\theta)} \tag{17}$$

The quantity $Xe^{i(\omega t+\theta)}$ in equations (16) and (17) is the displacement $x(t)$ itself as given by equation (9), thus:

$$\dot{x}(t) = i\omega x(t) \tag{18}$$

$$\ddot{x}(t) = i\omega \dot{x}(t) = -\omega^2 x(t) \tag{19}$$

Thereby, complex notation requires only one operation in the differentiation of a harmonic signal: to multiply or to divide by $i\omega$. This is illustrated in figure 7.

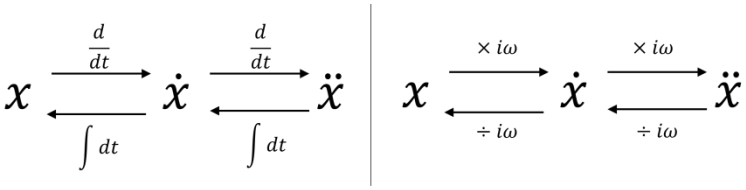


Figure 7 Complex relationships between displacement, velocity and acceleration in a harmonic signal.

In conclusion, one advantage for using complex representation of harmonic signals is that a complex number, by being composed by real and imaginary parts, contains both information on magnitude and phase.

2.3 Quantification of Energy in a Signal - RMS

One of the problems associated to continuous-time signals is related to the quantification of energy within the waveform, whether it is electromagnetic or mechanical. For example, one may need to know the power dissipated by a resistance in an electric circuit, or if the global level of vibration of an engine is above an acceptable limit.

Let us consider two signals, signal A and signal B, as shown in figure 8. One way of quantifying these signals would be by measuring their highest peaks: signal A's peak is 19.6 which is larger than signal B's peak of 16.4. However, signal B is larger than signal A most of the time. Hence, the quantification of a continuous-time signal using just the peak as a measure is not adequate to show how much "energy" the signal contains. A better way of doing it is by using the Root Mean Square (RMS). In this example, signal B's RMS value is 7.8, which is larger than signal A's RMS value of 6.9, contrary to what happens when the peaks are compared.

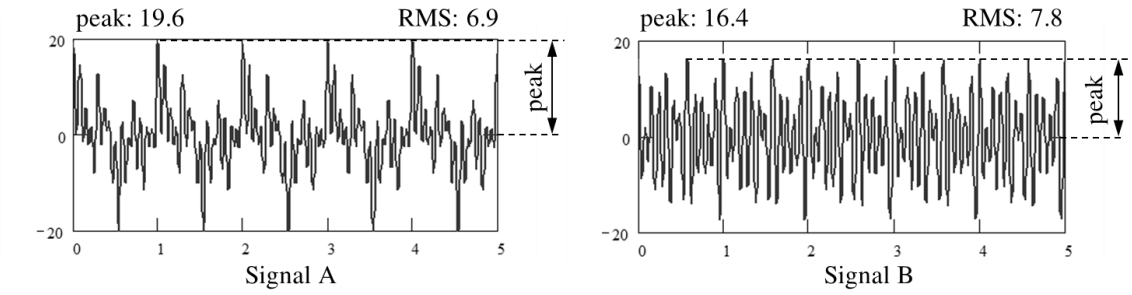


Figure 8: Continous-time signals with different peak and RMS values.

Before defining what the RMS value is, it is important to understand first why this is the preferred method to quantify, globally, a signal. Let us start by defining a harmonic time signal with unitary amplitude and plot the first two periods (figure 9). The plot on the left (a) is a continuous waveform, representing an analogue signal, and the plot on the right (b) is a discrete waveform, representing a digital signal.

(a) Continuous waveform (Analog) (b) Discrete waveform (Digital)

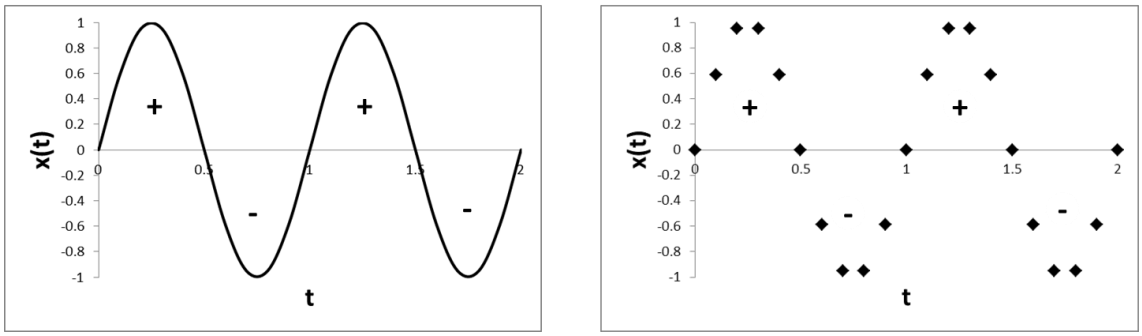


Figure 9 Plot of two periods of a harmonic time signal with amplitude 1 and frequency 1 Hz: (a) continuous (or analogue) waveform and (b) discrete (or digital) waveform.

One way to quantify this signal would be to determine the average. By definition, the average (or arithmetic mean) is:

$$\bar{x} = \frac{1}{T}\int_0^T x(t)dt \tag{20}$$

for continuous functions (analogue signal) and:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{21}$$

for discrete functions (digital signal), where T is the time span and N is the number of points (samples) in the discrete waveform.

However, the average is not useful. Figure 9 also shows that this harmonic signal oscillates about zero with positive and negative values. Thus, when computing the average, it should be obvious that the positive values cancel the negative values. As a consequence, the average is zero for the harmonic time signal shown when an integer number of periods is measured (in this case, two), regardless of the amplitude, frequency and phase.

The immediate suggestion to correct this problem is to determine the average of the absolute values. In this case, the original signal shown in figure 9 takes the form shown in figure 10.

(a) Continuous waveform (Analog) (b) Discrete waveform (Digital)

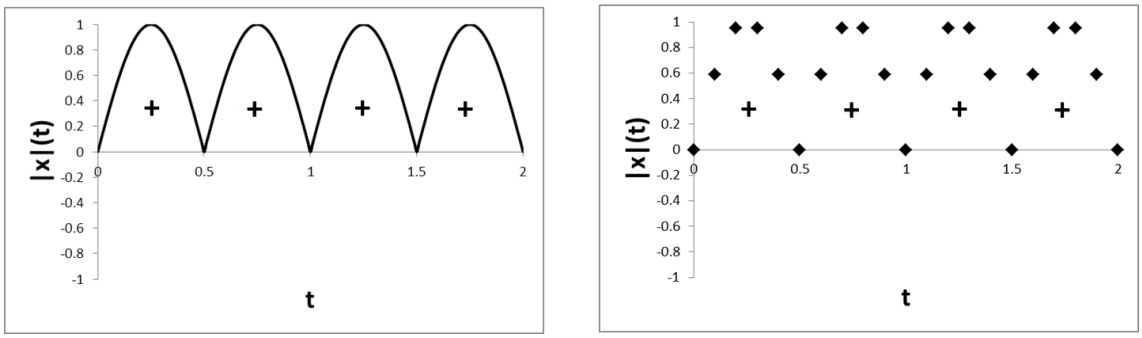


Figure 10 Plot of two periods of the average of the harmonic time signal represented in figure 9: (a) continuous (or analogue) waveform and (b) discrete (or digital) waveform.

The average of the absolute values is determined from:

$$\bar{x}_{abs} = \frac{1}{T}\int_0^T |x(t)|dt \tag{22}$$

for continuous functions (analogue signal) and:

$$x_{abs} = \frac{1}{N}\sum_{i=1}^{N} |x_i| \tag{23}$$

for discrete functions (digital signal).

It is easy to show that the relationship between the peak amplitude and the average of the absolute values, for a harmonic sine wave, is:

$$\bar{x}_{abs} = \frac{2}{\pi}x_{PK} \cong 0.637x_{PK} \tag{24}$$

where x_{PK} is the peak amplitude. It should be noted that equation (24) is not valid for complex waveforms and can only be applied to harmonic sine waves. If the average of the absolute values of a complex time signal is being determined, either one of equations (22) or (23) should be used in place of equation (24).

The average of the absolute values is a first approach to quantify, globally, a time signal. However, it is clear from figure 10 that the derivative is undefined at zero – the function is not continuous, thus it is not differentiable, which may bring some algebra problems.

An alternative consists on determining the Root Mean Square (RMS: the function is raised to the power of two so that all values become positive, it is integrated and then the square root is determined. In this case, the original signal shown in figure 9 takes the form shown in figure 11.

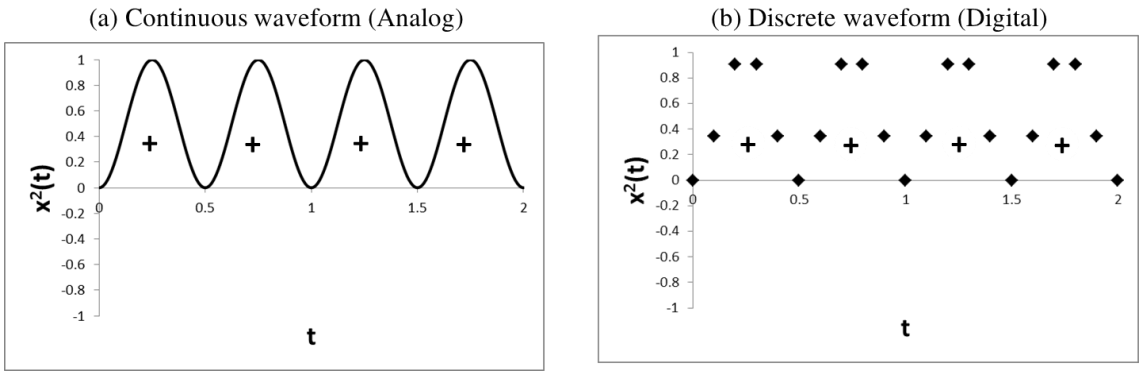


Figure 11 Plot of two periods of the harmonic time signal represented in figure 9 raised to the power of two: (a) continuous (or analogue) waveform and (b) discrete (or digital) waveform.

The RMS is determined from:

$$x_{RMS} = \sqrt{\frac{1}{T}\int_0^T x^2(t)dt} \tag{25}$$

for continuous functions (analogue signal) and:

$$x_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} {x_i}^2} \tag{26}$$

for discrete functions (digital signal).

It can be shown that the relationship between the peak amplitude and RMS value, for a harmonic sine wave, is:

$$x_{RMS} = \frac{\sqrt{2}}{2} x_{PK} \cong 0.707 x_{PK} \tag{27}$$

It should be noted that equation (27) is not valid for complex waveforms and can only be used with harmonic sine waves. If the RMS of a non-harmonic time signal is being determined, either one of equations (25) or (26) should be used in place of (27).

Contrary to figure 10 where the derivative is undefined at zero, figure 11 shows a continuous and differentiable function. This is one important result that encourages the use of the RMS over the use of the average of the absolute values.

Another interesting property is that the RMS is a statistical measure of the magnitude of a varying quantity. The standard deviation of variable x is defined as:

$$\sigma_x = \sqrt{\frac{1}{T}\int_0^T (x(t) - \bar{x})^2 dt} \tag{28}$$

for continuous functions (analogue signal) and:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 dt} \tag{29}$$

for discrete functions (digital signal).

Consequently, for a harmonic signal with zero mean ($\bar{x} = 0$), equations (28) and (29) turn into equations (25) and (26), respectively, and the RMS value is the same as the standard deviation. On the other hand, if the average is not zero, the RMS is related to the standard deviation by the following equation [5]:

$$x_{RMS}^2 = \bar{x}^2 + \sigma_x^2 \tag{30}$$

It is clear that the RMS value of a varying quantity is always greater than the average, in that the RMS includes the standard deviation as well.

In summary, the use of the RMS value to quantify a time-signal has the following advantages:

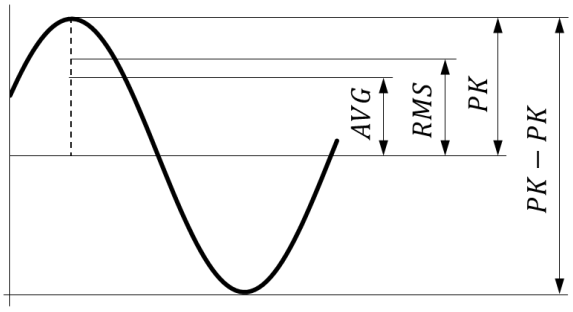
1. Contrary to the average, it can give a measure of the "amount of energy" in a signal[1]. For example, the power dissipated on an electrical resistor depends on the square of the current and the energy stored in a spring depends on the square of the deformation. The use of the square root brings back the units to which we are used to instead of leaving them raised to the power of two.
2. The power of 2 of a harmonic time signal still is a continuous and limited function, which is differentiable, contrary to the average of the absolute values, which is not;
3. The RMS is a statistical measure of the magnitude of a varying quantity. For the particular case of a harmonic signal with zero average, it actually is the standard deviation.

2.4 Useful Relationships and Common Waveforms

Common waveforms, their shapes, equations and RMS values are shown in table 1. Except for the DC and pulse train waveforms, all other waveforms have zero mean, peak amplitude X and peak-to-peak amplitude $2X$.

[1] It is now important to mention that the RMS value is not, in rigour, the same as Energy. Energy is measured in J (*Joule*), while the RMS value has the same units as those in the waveform, e.g., V (*Volt*) or μm (*micrometer*).

For harmonic signals, it is also useful that the relations between RMS, peak and peak-to-peak are known. For a sine wave, this is illustrated in figure 12 (these relations are not valid for non-harmonic signals).



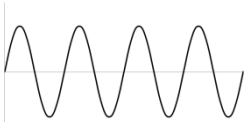
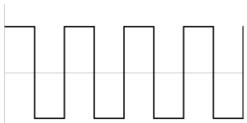
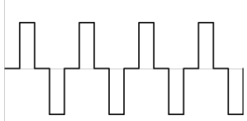
$$x_{PK-PK} = 2x_{PK}$$

$$x_{RMS} = \frac{\sqrt{2}}{2} x_{PK} \cong 0.707 x_{PK}$$

$$\bar{x}_{abs} = \frac{2}{\pi} x_{PK} \cong 0.637 x_{PK}$$

Figure 12 Summary of the relationships between average, peak, peak-to-peak and RMS for harmonic sine waves (these relations are not valid for non-harmonic signals).

Table 1 Common waveforms and their shapes (peak amplitude is X and mean is zero for all waveforms, except for the DC signal and pulse train which average values are non-zero).

Waveform	Shape	Equation	RMS
Sine wave		$x(t) = X\sin(\omega t + \theta)$	$x_{RMS} = \frac{\sqrt{2}}{2}X$
Square wave		$x(t) = \begin{cases} X, & \langle ft \rangle < 0.5 \\ -X, & \langle ft \rangle > 0.5 \end{cases}$	$x_{RMS} = X$
Triangular wave		$x(t) = 2\lvert 2X\langle ft + s\rangle - X\rvert - X$ ($s = 75\%$ is the phase shift for the shape shown)	$x_{RMS} = \frac{\sqrt{3}}{3}X$
Sawtooth wave		$x(t) = 2X\langle ft + s\rangle - X$ ($s = 75\%$ is the phase shift for the shape shown)	$x_{RMS} = \frac{\sqrt{3}}{3}X$
3-level Modified Square wave		$x(t) = \begin{cases} 0, & \langle ft \rangle < 0.25 \\ X, & 0.25 < \langle ft \rangle < 0.5 \\ 0, & 0.5 < \langle ft \rangle < 0.75 \\ -X, & \langle ft \rangle > 0.75 \end{cases}$	$x_{RMS} = \frac{\sqrt{2}}{2}X$

Chapter 3

Fourier Analysis

3.1 Introduction

The Fourier transform is a well-known and widely used mathematical tool nowadays, described in many textbooks, for example [6]. It is not the aim of this section to discuss the many different versions that have been developed over the years, but solely to introduce it and to address a few of the most common problems associated to its practical use.

When the time signal is of the harmonic type, the determination of the frequency is a simple process. However, in many applications, time signals are complex functions that combine many different frequencies together, as seen in the oscillation of bridges, in speech or in modulated radio waves. Usually, complex time signals are very hard to interpret and understand; however, when they are represented in the frequency domain, it is often easier to identify some particular features contained in the signal. The Fourier Transform, introduced in 1822 by Joseph Fourier [7], is a mathematical transformation in which a complex function,

whether continuous or discontinuous, is expanded into a series of sine waves. While this result is not absolutely correct, this became among the most widely used tools for transforming a signal in the time domain to the frequency domain. A practical example is shown in figure 13. Applications of the Fourier transform range from designing filters for noise reduction in audio signals to condition monitoring of rotating machinery.

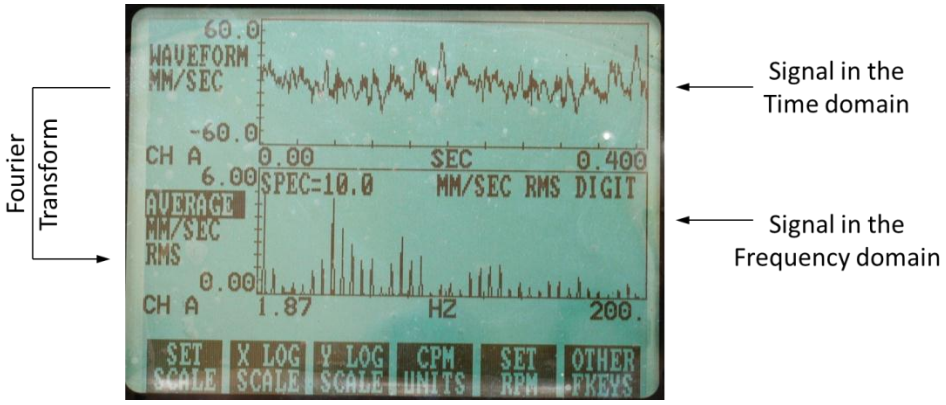


Figure 13 Photo of the screen of an Emerson CSI 2120-2 Machinery Health Analyser during the measurement of structural vibrations.

To transform the signal from the time domain to the frequency domain, the Fourier transform decomposes the complex time signal into a series of harmonic time signals. Each individual harmonic component has its own amplitude and frequency, which are plotted in a xy axis chart with amplitude on the y axis and frequency on the x axis. This process is represented schematically in figure 14. If we take the individual waveforms spaced by the frequency (top plane on a parallelepiped), the time and frequency domains are the orthographic projections on the other two mutually perpendicular faces of the parallelepiped.

This is an important result, because what figure 14 is showing is that each line in the frequency spectrum is a harmonic signal in itself, with amplitude and frequency (and phase as well). As a consequence, equations (14), (15), (18), (19), (24) and (27) and the relations shown in table 1 and figures 5, 7 and 12, which were said to be valid for harmonic time signals only when in the time domain, are also valid for complex signals, whether deterministic or not, as long as when represented in the frequency domain. As an example, these relations cannot be applied directly on the time signal shown on top of figure 13, but they can be applied to the frequency spectrum shown below, because each line alone represents an individual sine wave.

In this section, the basic concepts of Fourier analysis and Fourier transform are firstly presented. The Discrete Fourier Transform (DFT) and Fast Fourier Transform (FFT) will be discussed next. Their limitations and how they can be implemented will also be addressed.

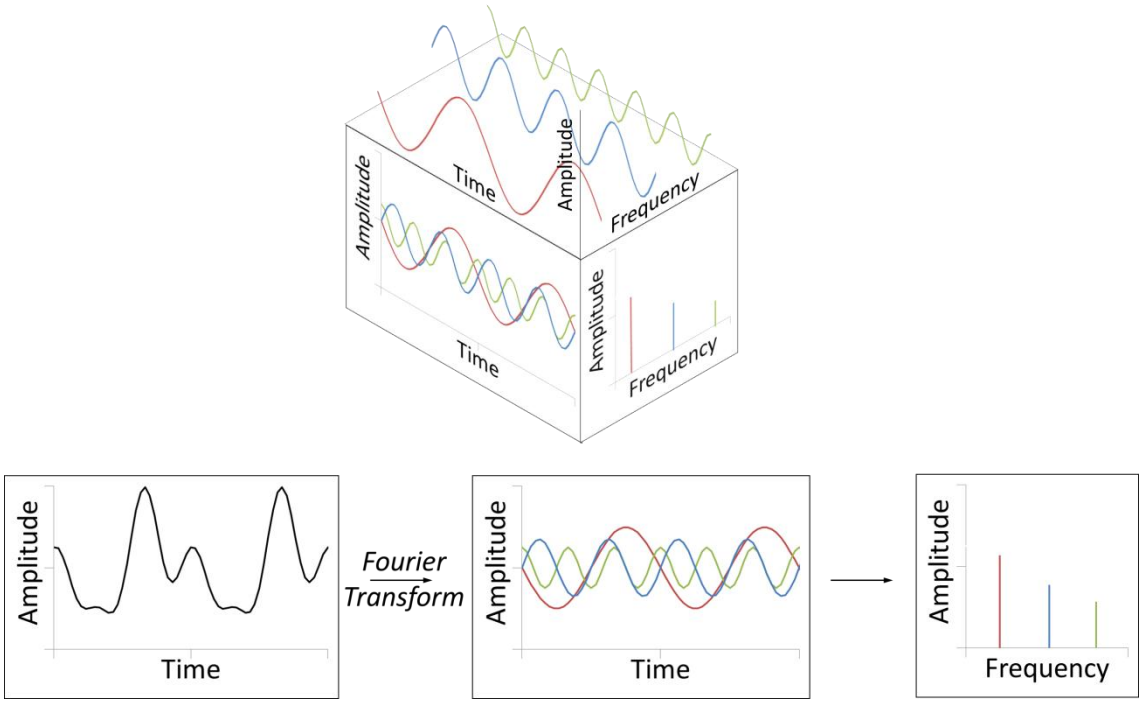


Figure 14 Schematic representation of the Fourier Transform application on a complex time signal and its representation in both the time and frequency domains.

3.2 Fourier Transform

Let us assume the simpler case of a periodic time signal $x(t)$ with period T (deterministic, but not necessarily pure harmonic). First of all, even if the time signal is acquired over a finite period of time, Fourier analysis assumes that the domain of $x(t)$ is infinite, i.e., $t \in]-\infty, +\infty[$. As mentioned earlier, Fourier analysis also assumes that this time signal can be represented as a series of a number $n = 1,2,3, \ldots$ of related sinusoidal waves:

$$x(t) = a_0 + 2\sum_{n=1}^{\infty}\left(a_n \cos\frac{2\pi nt}{T} + b_n \sin\frac{2\pi nt}{T}\right) \tag{31}$$

Since the domain of $x(t)$ is infinite, we may assume that $t \in]-T/2, +T/2[$ during a complete period of time T. Coefficients a_0, a_n and b_n are real constants that can be determined from, respectively:

$$a_0 = \frac{1}{T}\int_{-T/2}^{T/2} x(t)dt \tag{32}$$

$$a_n = \frac{1}{T}\int_{-T/2}^{T/2} x(t)\cos\frac{2\pi nt}{T}dt \tag{33}$$

$$b_n = \frac{1}{T}\int_{-T/2}^{T/2} x(t)\sin\frac{2\pi nt}{T}\,dt \tag{34}$$

in which it can be observed that a_0 simply is the mean value of $x(t)$ over the period T. Coefficients a_n and b_n are called the Fourier, or Spectral, coefficients [8].

If exponential complex functions are used instead of trigonometric ones, i.e.:

$$\cos\frac{2\pi nt}{T} = \frac{1}{2}\left(e^{i\frac{2\pi nt}{T}} + e^{-i\frac{2\pi nt}{T}}\right) \tag{35}$$

$$\sin\frac{2\pi nt}{T} = \frac{1}{2i}\left(e^{i\frac{2\pi nt}{T}} - e^{-i\frac{2\pi nt}{T}}\right) \tag{36}$$

the resulting exponential form for the Fourier series can be written as:

$$x(t) = \sum_{n=-\infty}^{+\infty} c_n e^{i\frac{2\pi nt}{T}} \tag{37}$$

where

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} x(t)e^{-i\frac{2\pi nt}{T}}dt \tag{38}$$

and the following relationship holds between the Fourier coefficients:

$$|c_n| = \sqrt{a_n^2 + b_n^2} \tag{39}$$

It should be noted that, while equation (31) is defined for positive frequencies only, equation (37) is defined for both positive and negative frequencies. This is demonstrated in, for instance, [9].

If $x(t)$ is not a periodic function, say a single impulse or a transient, the Fourier theorem is still valid, because these type of signals can be seen as periodic signals with period $T = \infty$. Substituting (38) into (37) yields:

$$x(t) = \sum_{n=-\infty}^{+\infty} \frac{1}{T}\left(\int_{-T/2}^{T/2} x(t)e^{-i\frac{2\pi nt}{T}}dt\right)e^{i\frac{2\pi nt}{T}} \tag{40}$$

Since $T \to \infty$, $1/T \to 0$ which is the same as saying $1/T = df$. On the other hand, the quantity $n/T = ndf \to f$ and the sum in (40) turns into the following integral:

$$x(t) = \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} x(t) e^{-i2\pi ft} dt \right) e^{i2\pi ft} df \tag{41}$$

This equation can be written in the more convenient form:

$$x(t) = \int_{-\infty}^{+\infty} X(f) e^{i2\pi ft} df \tag{42}$$

with

$$X(f) = \int_{-\infty}^{+\infty} x(t) e^{-i2\pi ft} dt \tag{43}$$

Equations (42) and (43) constitute what is known as the pair of Fourier integrals, defined in from $-\infty$ to $+\infty$. It is this pair that allows transforming a signal from the time domain into the frequency domain, as graphically represented before in figure 14.

It is also interesting to note that the coefficient c_n expressed by equation (38) is equivalent to (43), except it is defined from $-T/2$ to $+T/2$.

3.3 An Example of Application of the Fourier Transform

Common waveforms were previously shown in table Table 1. However, these were defined from 0 to $+\infty$ in the time domain. When applying the Fourier Transform it is more convenient if the functions are defined from $-T/2$ to $+T/2$ instead (table 2).

Table 2 also includes a column about the waveform's symmetry about the Cartesian system of coordinates. One useful aspect to know when applying the Fourier Transform is that some of the Fourier coefficients a_0, a_n and b_n given by equations (32), (33) and (34) can be zero depending if the function is either odd or even. For odd functions like the sine, square or sawtooth waves, a_0 and a_n are zero and do not need to be evaluated, whereas for even functions like the cosine or the triangular waves, it is b_n the one that is zero and does not need to be evaluated.

Let us assume that a switch periodically changes from on to off and that its behaviour can be approximated by the square wave shown in figure 15. We want to evaluate the Fourier transform of this signal up to the 8th component, and plot it both in the time domain and frequency spectrum.

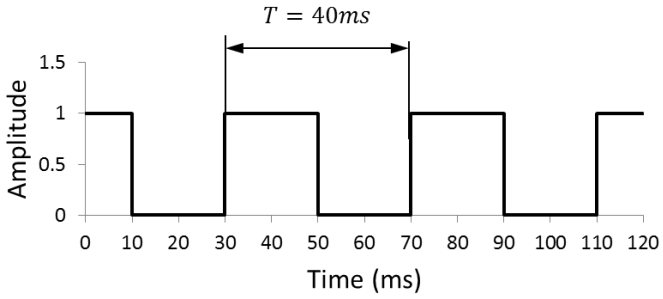


Figure 15 Example of a square wave with amplitude 1 and fundamental frequency $f_0 = 25\ Hz$.

The first step consists in determining the fundamental frequency from the direct measurement of the time period: $f_0 = \frac{1}{T} = \frac{1}{0.04} = 25Hz$.

Now we are in condition of evaluating if the function is odd or even. The most convenient way of graphically representing a period of this wave is in the period ranging from $-T/2$ to $+T/2$. This is done in figure 16. In this case, the square wave symmetry is even, contrary to what is shown in table 2, because one of the Fourier coefficients is a DC signal (a_0), which corresponds to an offset in the y axis.

The equation for the square wave shown in figure 16 is:

$$x(t) = \begin{cases} 0, & -0.02s < t < -0.01s; 0.01s < t < 0.02s \\ 1, & -0.01s \leq t \leq 0.01s \end{cases} \quad (44)$$

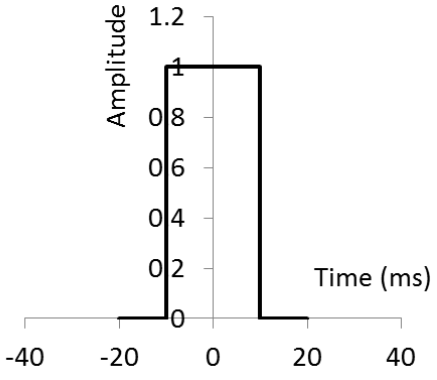
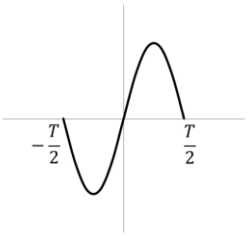
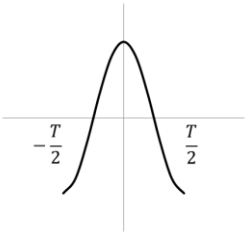
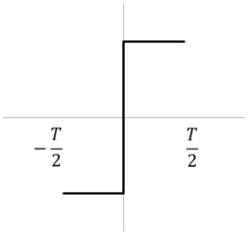
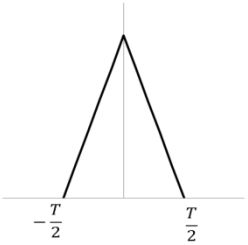
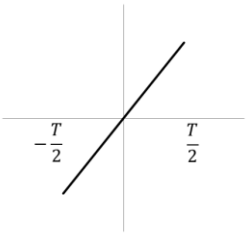


Figure 16 Example of a square wave with amplitude 1 and fundamental frequency $f_0 = 25\ Hz$ represented in the $-T/2$ to $+T/2$ domain.

Table 2 Particular cases of common waveforms and their shapes defined between $-T/2$ and $+T/2$.

Waveform	Shape	Equation	Symmetry
Sine wave (sin)		$x(t) = X\sin(\omega t)$	Odd $(a_0, a_n = 0)$
Sine wave (cos)		$x(t) = X\cos(\omega t)$	Even $(b_n = 0)$
Square wave		$x(t) = \begin{cases} -X, & -T/2 < ft < 0 \\ -X, & 0 < ft < T/2 \end{cases}$	Odd $(a_0, a_n = 0)$
Triangular wave		$x(t) = \lvert 2Xft - X \rvert$	Even $(b_n = 0)$
Sawtooth wave		$x(t) = 2Xft$	Odd $(a_0, a_n = 0)$

Now we are in conditions of determining the Fourier coefficients. First of all, since the function represented in figure 16 is even, we know that the Fourier coefficient $b_n = 0$. The Fourier coefficient a_0 is the mean, which in this case obviously is $a_0 = 0.5$. This is the DC component of the Fourier spectrum and it corresponds to an offset of the time signal in the y-axis.

Evaluation of the Fourier coefficients a_n is made with equation (33), which in this particular case is:

$$a_n = \frac{1}{0.04}\int_{-0.02}^{0.02} x(t)\cos\frac{2\pi nt}{0.04}dt \tag{45}$$

This equation can further be simplified, if we notice that $x(t) = 0$ for $-0.02s < t < -0.01s$ and $0.01s < t < 0.02s$. This means that the interval for integration can be reduced to $-0.01s \leq t \leq 0.01s$ where $x(t) = 1$:

$$a_n = \frac{1}{0.04}\int_{-0.01}^{0.01} \cos\frac{2\pi nt}{0.04}dt \tag{46}$$

Because the evaluation of the Fourier coefficients require integration, sometimes it is practical to have in mind the following anti-derivatives:

$$\int t\cos at\,dt = \frac{1}{a^2}\cos at + \frac{1}{a}t\sin at$$
$$\int \cos at\,dt = \frac{1}{a}\sin at \tag{47}$$

$$\int t\sin at\,dt = \frac{1}{a^2}\sin at - \frac{1}{a}t\cos at$$
$$\int \sin at\,dt = -\frac{1}{a}\cos at \tag{48}$$

The Fourier coefficients a_1 to a_8 are evaluated for $n = 1 \ldots 8$ respectively. As an example, the 3rd coefficient is determined below:

$$\begin{aligned} a_3 &= \frac{1}{0.04}\int_{-0.01}^{0.01}\cos\frac{3\times 2\pi t}{0.04}dt = \frac{1}{0.04}\left[\frac{0.04}{6\pi}\sin\frac{6\pi t}{0.04}\right]_{-0.01}^{0.01} \\ &= \frac{1}{6\pi}\left(\sin\frac{6\pi\times 0.01}{0.04} - \sin\frac{6\pi\times -0.01}{0.04}\right) \\ &= -0.1061 \end{aligned} \tag{49}$$

The frequency associated to coefficient a_3 is $f_3 = 3\times f_0 = 75\ Hz$.

After following the same procedure for all the other seven coefficients, the following results should be obtained: $a_1 = 0.3183$; $a_3 = -0.1061$; $a_5 = 0.0637$; $a_7 = -0.04547$; $a_2, a_4, a_6, a_8 = 0$. Once the Fourier transform has been applied, the time-signal can be written as in equation (31), which in this particular case takes the form:

$$x(t) = 0.5 + 0.637\cos(2\pi \times 25t) - 0.212\cos(2\pi \times 75t) + 0.127\cos(2\pi \times 125t) - 0.0909\cos(2\pi \times 175t) \quad (50)$$

Equation (50) is an approximation of what is shown in figure 15, accomplished by adding up harmonic time signals, each one with amplitude, frequency and phase. A summary of the results is presented in table 3. The graphical representation of this approximated square wave, on both the time and frequency domains, is presented in figures 17 and 18 respectively.

The "SUM" wave shown in figure 17 is the add up of 4 pure harmonic sine waves at related frequencies, amplitudes and phases. If more components would have been considered (i.e., Fourier coefficients), it would be expected that the "SUM" wave would better approximate the original square one.

Nevertheless, where the original signal is a discontinuous function – which is the case of the square wave discussed – the Fourier transform will not be able to perfectly regenerate the signal without some overshoot. This is often called as the Gibbs phenomenon and is depicted in figure 17, at the "corners" of the wave [10].

Table 3 Summary of the results of the Fourier transform applied to the wave in figure 15.

Frequency (Hz)	Amplitude (V)	Phase (°)
0 (DC)	0.5	-
25	0.637	0
75	0.212	180
125	0.127	0
175	0.0909	180

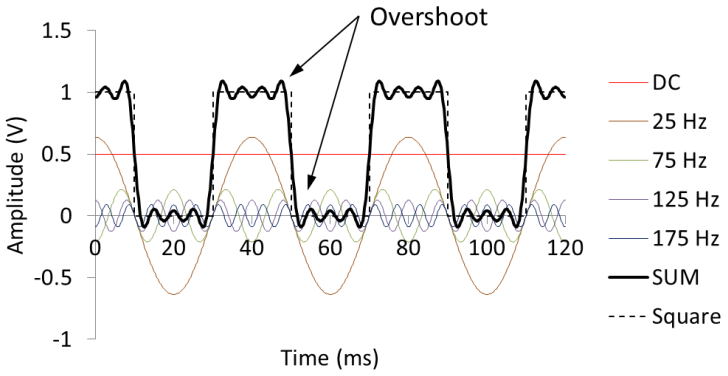


Figure 17 Time domain representation of the square wave shown in figure 15 after applying the Fourier transform with $n = 8$.

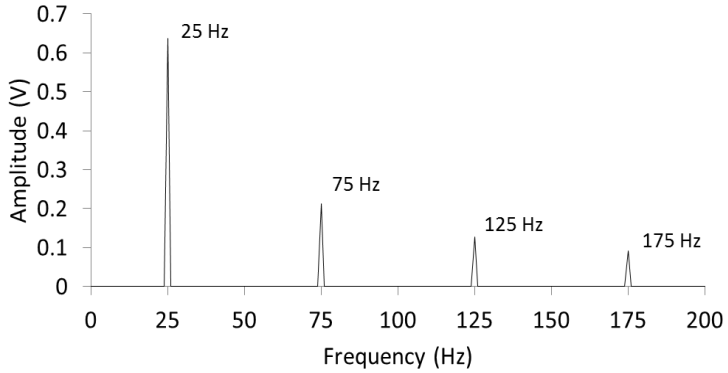


Figure 18 Frequency spectrum of the square wave shown in figure 15 after applying the Fourier transform with $n = 8$.

3.4 Basics of the Discrete and Fast Fourier Transforms

3.4.1 Sampling Frequency

Real-world waveforms are continuous functions of time: signals are analogic, as well as most sensors' output. For example, a pre-amplified piezoelectric accelerometer (see section 5.2.1) produces an output signal that is proportional to the acceleration of the system at the measurement coordinate. However, in signal processing, there always is a process of analogue-to-digital conversion (ADC) at some point, which in this case consists in converting the analogue transducer signal into the digital code used by the processor [6]. Figure 19 shows a signal acquisition module from National Instruments that includes conditioners. The input signals it receives are continuous analogue time signals from transducers, which are converted through an electronic ADC into a discrete time series digital signal.

The ADC converter records the level of the signal at a discrete set of times. Figure 20 illustrates the sampled acquisition of an analogue time signal, where $\Delta = 1/f_s$ is the time

elapsed between each sample and f_s is the sampling frequency. In the example case of the National Instruments acquisition module shown in figure 19, the sampling frequency is 51.2 kS/s (kilo-samples per second) per channel [11]. In practice, what is available is a collection of points with information on the amplitude at discrete and regular intervals of times.

Figure 19 National Instruments four-channel C Series dynamic signal acquisition module (NI 9234) for audio/frequency measurements from integrated electronic piezoelectric (IEPE) and non-IEPE sensors with NI CompactDAQ or CompactRIO systems [11].

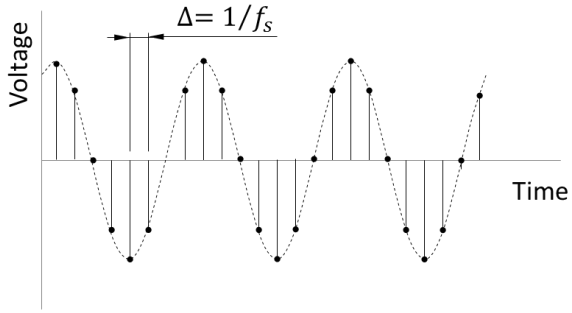


Figure 20 Sampled acquisition of a voltage analogue time signal.

This signal acquisition module shown in figure 19 has separate conditioners for each one of the 4 channels. Less expensive alternatives time-share the signal with multiplexers. Multiplexers will pick one channel of data from a bank of data channels in a sequential manner and connect it to a common input device. Basically, it distributes the sampling frequency through the bank of channels. In a multiplexer, each channel in the bank will have a delay proportional to the inverse of the sampling frequency.

The minimum sampling frequency f_s required in ADC must be at least twice as much the signal's fundamental frequency f_0, i.e., $f_s/f_0 > 2$. This means that with a sampling frequency of 51.2 kS/s, the NI 9234 shown in figure 19 is adequate for waveform measurements up to

25.6 kHz. This is a very important result in signal processing that will be later addressed in section 4.1, when discussing problems associated to signal processing, namely aliasing.

3.4.2 Discrete Fourier Transform

Loosely speaking, when a continuous function is sampled at discrete and regular intervals of time, the major difference in the post-processing mathematics is that integrals are expressed in terms of sums, which is not far from the definition of the Riemann integral. So, if $x(t)$ is sampled at regular intervals of time and represented by the discrete series $\{x(k)\}, k = 0,1,2,\dots,N-1$ where $t = kT/N$, it can be shown that the pair of Fourier integrals (42) and (43) can be replaced by the summations [6]:

$$x(k) = \sum_{j=0}^{N-1} X(j) e^{i2\pi jk/N} \tag{51}$$

with

$$X(j) = \frac{1}{N} \sum_{k=0}^{N-1} x(k) e^{-i2\pi jk/N} \tag{52}$$

and $j = 0,1,2,\dots,N-1$.

These summations constitute the discrete Fourier transform (DFT) pair. In practice, the computation of the Fourier transform by computers and spectrum analysers is done after the ADC, which means that this is made over a discrete time series. However, computation of equation (52) would require nearly N^2 complex operations. For large values of N this can become computationally prohibitive, even for modern and fast processors.

3.4.2.1 Fast Fourier Transform

The Fast Fourier Transform (FFT), as the name suggests, is a much faster algorithm based on the DFT that has become very popular since its development in the 1960's [12,13]. It is not the intention of this book to describe its details, neither the ones from the many derivations subsequently developed. However, an important result on the use of the FFT that affects the sampling frequency will be addressed. This is important, because it gives a better understanding on how spectrum analysers operate.

It has been found that the number of operations in the FFT can be reduced from nearly N^2 to $(N/2)\log_2 N$ multiplications, $(N/2)\log_2 N$ additions and $(N/2)\log_2 N$ subtractions for the more general case of $N = 2^m$ where m is any positive integer [6]. This is illustrated in the plot shown in figure 21, where the computational benefits on the use of the FFT instead of the DFT become quite obvious, especially for larger values of N.

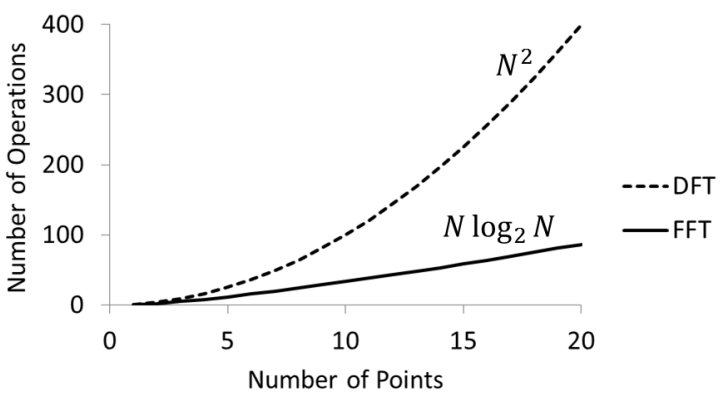


Figure 21 Contrast between the number of operations needed for evaluating the DFT and the FFT.

Since $N = 2^m$, where m is any positive integer, this means that the number of points that are acquired over time can be any from 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 9192, etc. The time T that is required to acquire a certain amount N of points depends on the sampling frequency f_s:

$$N = Tf_s \tag{53}$$

Now, let us suppose the following problem: a time signal is meant to be represented in the frequency domain up to $f_{max} = 400\,Hz$ with a total of $n_{lines} = 1600$ spectral lines. The frequency resolution in the frequency spectrum is determined from:

$$f_{res} = \frac{f_{max}}{n_{lines}} \tag{54}$$

Thus, in this example, $f_{res} = 0.25\,Hz$. The duration for the acquisition of a time period is $T = 1/f_{res} = 4s$. Because the measurement is being made up to 400 Hz, this means that the minimum sampling frequency must be at least $f_s = 800\,Hz$ to avoid aliasing, according to the Nyquist-Shannon theorem that will be later described in section 4.1. So, the number of points needed for the FFT, according to equation (53), is $N = 4 \times 800 = 3200\,points$. However, this number of points is not a power of 2 and cannot be used in the FFT algorithm. The closest powers of 2 to 3200 are $2^{11} = 2048$ and $2^{12} = 4096$. Thus, we must maximize the number of points to 4096, to include all the 3200 points required.

Now it is necessary to check if the Nyquist-Shannon theorem is being respected with this new value $N' = 4096$. Equation (53) can be used once more to determine the new sampling frequency: $f_s' = 4096/4 = 1024Hz$. The ratio between this new 1024 Hz sampling frequency

and the 800 Hz original one of is 2.56, which means that the spectrum analyser is giving information up to $0.39f_s$.

Most spectrum analysers frequently show a number of lines n_{lines} that is based on 25×2^m, namely: 25, 50, 100, 200, 400, 800, 1600, etc. However, even if they do not, it is easy to understand that the sampling frequency in the FFT will always be larger than 2 and thus obey the Nyquist-Shannon theorem, because the number of sampling points in the time domain must be majored by the following power of 2.

Chapter 4

Signal Processing

4.1 Aliasing

On an ADC as the one shown in figure 20, information is lost about the time periods that elapse between each sample of signal, and one assumes what it is in between by extrapolation. As a consequence, if the sampling frequency is not chosen correctly, results may be misleading. Figures 22 and 23 give a few examples of what may happen when a sine wave is sampled at four different sampling frequencies, both in the time and frequency domains.

Let us assume a continuous sinusoidal harmonic signal with frequency $f_0 = 100\,Hz$ and amplitude 1 V is being sampled at a frequency $f_s = 100\,Hz$. The ratio between the sampling frequency and the signal's frequency is $f_s/f_0 = 1$. The resulting data misleads us to think the signal is a DC value. This conclusion is wrong because the original signal is a sine wave. Furthermore, the amplitude of the DC signal so obtained in this particular example is 0.5V, whereas the amplitude of the true sine wave is twice as much.

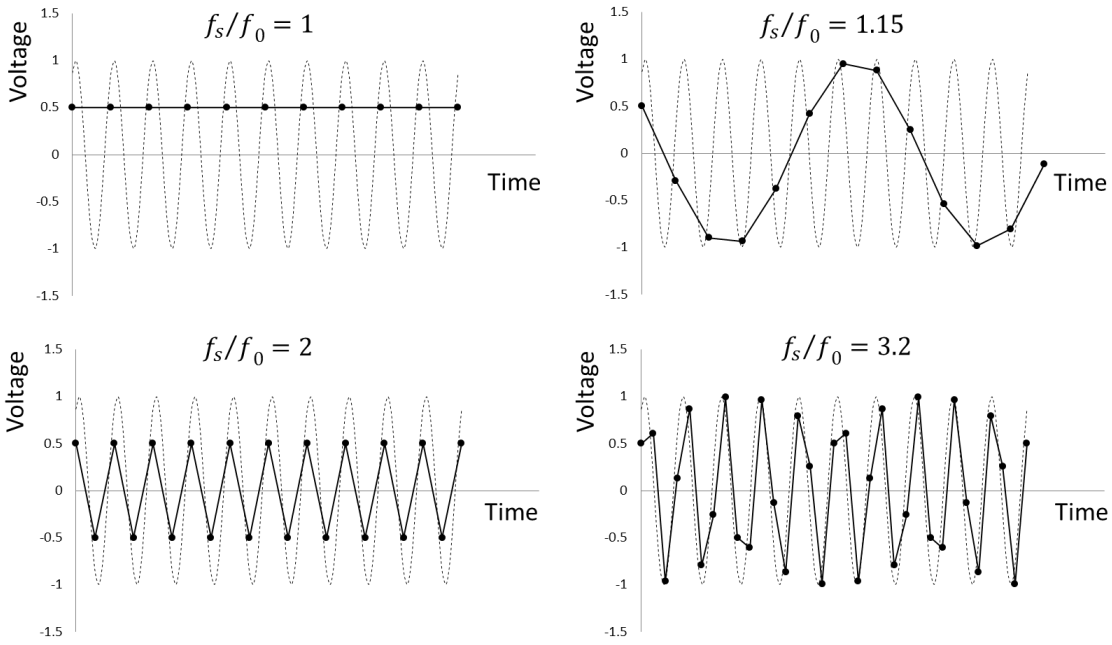


Figure 22: A sine wave is sampled at different ratios between sampling frequency and signal's frequency – representation in the time domain.

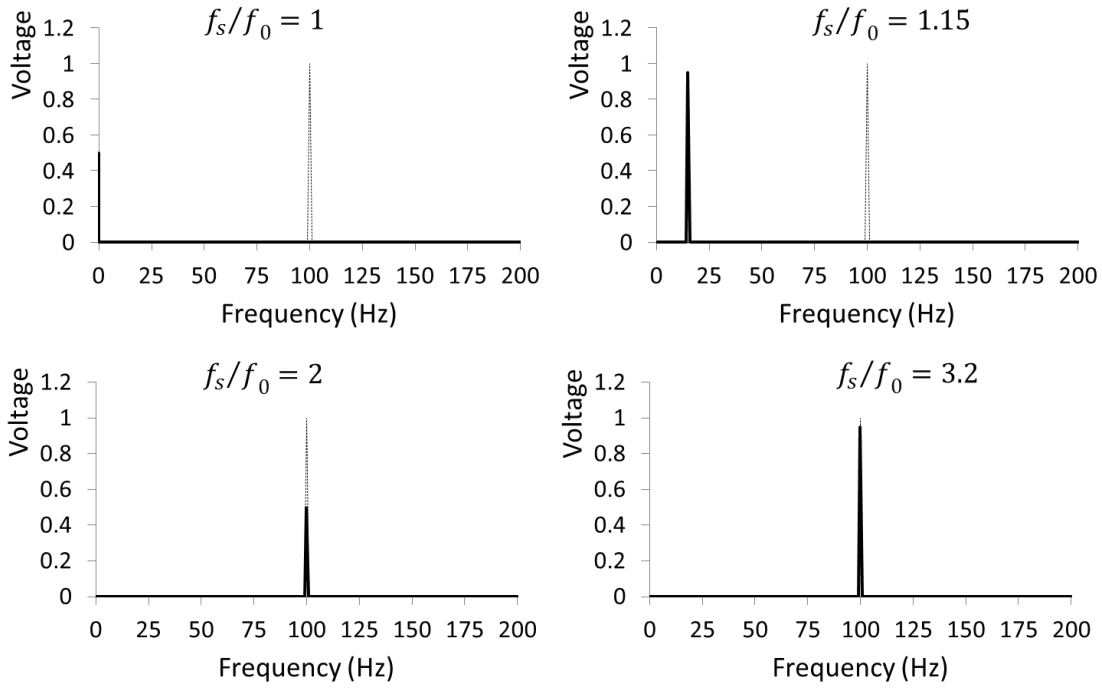


Figure 23 A sine wave is sampled at different ratios between sampling frequency and signal's frequency – representation in the frequency domain.

Let us now assume that the sampling frequency is increased to $f_s = 115\ Hz$ so that the ratio between the sampling frequency and the wave's frequency is $f_s/f_0 = 1.15$. In this case, Figure 22 shows a sine wave with amplitude near to 1 V, but at a much lower frequency than the true signal. Actually, the frequency of this sine wave is the difference between the sampling frequency and the signal's frequency $f_s - f_0 = 15\ Hz$. The misinterpretation of a signal by another one at a lower frequency due to undersampling is called *aliasing*[2].

In the third case shown in figure 22, the sampling frequency is twice as much as the signal's frequency, i.e., $f_s = 200\ Hz$ and $f_s/f_0 = 2$. The sampled data now shows a triangular wave with 100 Hz, although the amplitude is still misleading. The amplitude of the sampled data depends on when the first sample was taken (phase). This occurs whenever the wave is sampled at a ratio that is an integer fraction of its frequency ($f_0/f_s = 1,2,3, \ldots$). Thus, it is reasonable to assume that for $f_s/f_0 > 2$ it will be possible to measure the amplitude. This is known as the Nyquist-Shannon theorem, where it is stated that the minimum sampling rate required for aliasing to be avoided must be at least twice as much the signal's frequency.

Finally, in the case where the ratio between the sampling frequency and the signal's frequency is larger than 2 ($f_s/f_0 = 3.2$), it is possible to measure both the frequency and amplitude, even if the signal is oscillating and only occasionally reaches the peak amplitude of 1 V.

In real measurements, the sampling frequency is set according to the highest value in the frequency range of interest. Let us assume a measurement is being carried out in the 0 to 800 Hz frequency range. The highest frequency in the range is the one that needs faster sampling speed, i.e., it is the one frequency that determines the sampling frequency. Thus, in this case and according to the previous discussion, the sampling frequency would have to be set to at least $f_s = 2 \times 800 = 1600\ Hz$ in order to avoid aliasing.

The question we now ask is: is this enough to prevent aliasing? To answer this question, let us assume that we want to plot the frequency spectrum of a system (initially unknown to us) which is producing a signal with the following components: 300 Hz, 600 Hz, 900 Hz and 1200 Hz. If we now evaluate the ratios between the sampling frequency and these components' frequencies, for $f_s = 1600\ Hz$, we notice that the Nyquist-Shannon theorem is not respected at all the available frequencies: $\frac{1600}{300} \cong 5.3$, $\frac{1600}{600} \cong 2.7$, $\frac{1600}{900} \cong 1.8$ and $\frac{1600}{1200} \cong 1.3$. In this case,

[2] One practical example where aliasing is seen is when watching a film of a moving car: sometimes there is the illusion that the wheels are spinning contrary to the direction of motion; other times the wheels seem to be rotating at a much lower speed than one needed for the car to move at a certain velocity. This is because cameras also have sampling frequency. Actually, a film is a composition of a sequence of photos, which is equivalent to the digital sampling process in a waveform. Typical frame rates for cameras today may range from 24 to 300 fps (frames per second).

the ratios between the 900 and 1200 Hz components of this signal and the sampling frequency are both smaller than 2, which mean that aliasing will occur. This is illustrated in figure 24, where two inexistent signals, called "aliases", will appear on a frequency spectrum analyser's screen at 400 and 700 Hz. The alias peak at 400 Hz corresponds to the 1200 Hz component (1600-1200=400 Hz) and the alias peak at 700 Hz corresponds to the 900 Hz component (1600-900=700 Hz).

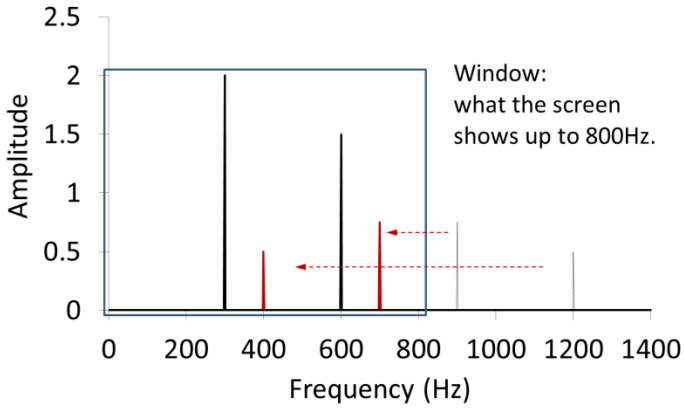


Figure 24 Example of aliasing on the acquisition of a signal with 4 harmonic components (300 Hz, 600 Hz, 900 Hz and 1200 Hz) when measured up to 800 Hz with a sampling frequency of 1600 Hz (no filtering).

This means that, *per se*, the sampling frequency is not enough to avoid aliasing, unless it is known beforehand that there are no higher frequencies in the signal, which, most of the times, is not possible to predict.

Combined with the correct sampling frequency setting, it is necessary to filter the signal. However, it must be stressed out that digital filtering does not avoid aliasing, because if the signal has already been digitized, sampling has already been done. Filtering must be done on the analogue signal, prior to sampling, by the hardware. Most modern acquisition systems and ADCs, like the one shown in figure 19, bring built-in conditioners with low-pass filters with variable cut-off frequency based on the sampling frequency. However, older systems may only offer from a selection of available cut-off frequencies, as is the case of the charge amplifier type 2635 from Brüel & Kjær shown in figure 25.

The charge amplifier is a current integrator that converts the input charge from an electrical source with capacitive nature (say, a piezoelectric charge accelerometer) into a proportional output voltage. The one shown in figure 25 offers a low-pass filter with the following possible cut-off frequencies: 100 Hz, 1000 Hz, 3000 Hz, 10 kHz and 30 kHz. Let us now assume that the frequency spectrum shown in figure 24 is being acquired with a charge accelerometer that is connected to the charge amplifier shown in figure 25. The maximum frequency we want to

measure, 800 Hz, is between the 100 and 1000 Hz cut-off frequencies of the charge amplifier. If we set the cut-off frequency, say, at 100 Hz, we will not be able to measure anything above 100 Hz, so the cut-off frequency must be set to at least 1000 Hz.

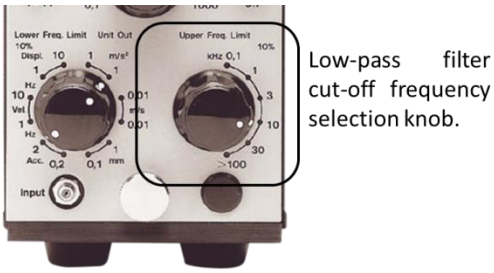


Figure 25 Charge amplifier type 2635 from Brüel & Kjær.

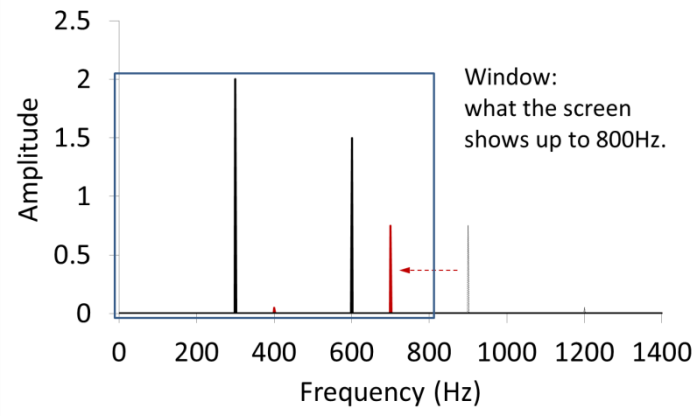


Figure 26 Example of aliasing on the acquisition of a signal with 4 harmonic components (300 Hz, 600 Hz, 900 Hz and 1200 Hz) when measured up to 800 Hz with a sampling frequency of 1600 Hz (low-pass filter with cut-off frequency set to 1000 Hz).

Figure 26 shows the frequency spectrum of the original signal shown in figure 24, but after a low-pass filter with cut-off frequency set to 1000 Hz has been applied to the analogue signal (before digitalization). It is possible to observe that the component at 1200 Hz does not produce an alias at 400 Hz anymore, because the analogue filter eliminated the 1200 Hz component before the signal was sampled (if the filter was not able to eliminate it completely, at least its level has been significantly reduced in comparison to the other components in the spectrum). However, there still exists the alias frequency at 700 Hz. This is because the sampling frequency was not adjusted when the filter was applied. Saying that the filter cut-off frequency is 1000 Hz is equivalent to saying that the ADC will have to deal with a signal up to 1000 Hz, even if the screen only shows up to 800 Hz. This means that the Nyquist-Shannon condition for the sampling frequency must be based on the analogue filter available. Thus, the

sampling frequency should be adjusted to $f_s = 2000\,Hz$, which is twice as much the cut-off frequency of the low-pass analogue filter available (figure 27).

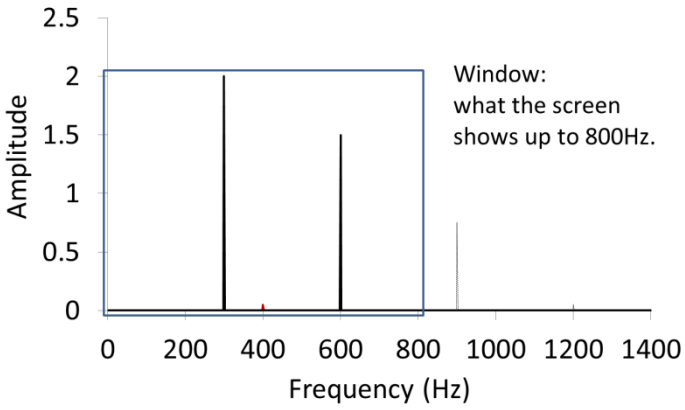


Figure 27 Example of aliasing on the acquisition of a signal with 4 harmonic components (300 Hz, 600 Hz, 900 Hz and 1200 Hz) when measured up to 800 Hz with a sampling frequency of 2000 Hz (twice as much the value of the low-pass cut-off frequency set to 1000 Hz).

In summary, to make sure aliasing is eliminated, two conditions must be met simultaneously:

4. An analogue low-pass filter must be used before the signal is sampled in the ADC;
5. The sampling frequency must be set to a value at least twice as much the value of the cut-off frequency set in the analogue low-pass filter.

4.2 Quantization Errors

When an ADC occurs, there also are a number of possible digital levels for the amplitude, called quantization levels. This means that the level of the signal at the instant it is sampled is rounded to the nearest digital level, as shown in figure 28.

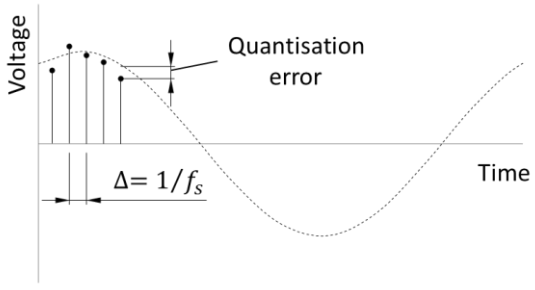


Figure 28 Example of quantization errors during digital sampling of a signal.

The accuracy of the quantization process depends on the number of bits in the converter: 2^{nbit}. For example, the National Instruments signal acquisition module shown in figure 19 has

a 24-bit ADC. The number of peak-to-peak quantization levels is $2^{24} = 16777216$ levels (including zero) and the dynamic range is determined to be $20\log_{10} 2^{24} = 144.5\ dB$ for peak-to-peak measurements[3]. This is equivalent to 138.5 dB for peak measurements and 135.5 dB dynamics range for RMS measurements.

It is recommended that the signal occupies as much of the range of the converter as possible. The loss in the dynamic range is $20\log_{10}\frac{Signal\ level}{Converter\ range}$. This means that if the signal is 10 times smaller than the range in the converter, 20 dB of the measurement range will be lost. If the signal is 100 times smaller than the range in the converter, 40 dB of the measurement range will be lost, and so on. Thus, ideally, signals should be amplified to occupy as much as possible the full range of the converter to minimize quantization errors and improve the signal-to-noise ratio. Another possibility, if existent, is to change the range of the converter. This is possible, for example, when using piezoelectric charge accelerometers that require a charge amplifier. Charge amplifiers often include the possibility to change the dynamic range, besides band-pass filters.

4.3 Leakage and Windowing

The Fourier series given by equations (31) or (37) shows that $x(t)$ can be represented by a series of harmonic components with frequencies $f_1 = f_0,\ f_2 = 2f_0,\ f_3 = 3f_0$, etc., where $f_0 = 1/T_0$ is both the fundamental frequency and the resolution in Hz (spacing of the frequency components).

Section's 4.3.3 example showed that the Fourier analysis works over a function $x(t)$ with period T_0. However, this does not mean that only one period of signal exists; on the contrary, Fourier analysis assumes that signals repeat infinitely in time. If the signals are truly periodic, this is not a problem, since there cannot be any components in the signal at frequencies between those calculated in the Fourier analysis. However, there are many real situations in which this is not the case, either because the signal being acquired has a non-integer number of periods during the measurement period or because the signal is of the random type. Actually, unless there is full control over the source signal and perfect synchronization with the acquisition clock, in most practical situations it is not possible to guarantee an integer number of periods is being acquired.

Let us analyze two contrasting situations, shown in figure 29: in situation (a), n integer periods of a sine wave are measured during an integer period of time nT; in situation (b),

[3] The *dB* is the name of a logarythmic scale that will be explained later on chapter 6.

$n + 0.5$ periods of the same sine wave are measured during a non-integer period of time $(n + 0.5)T$.

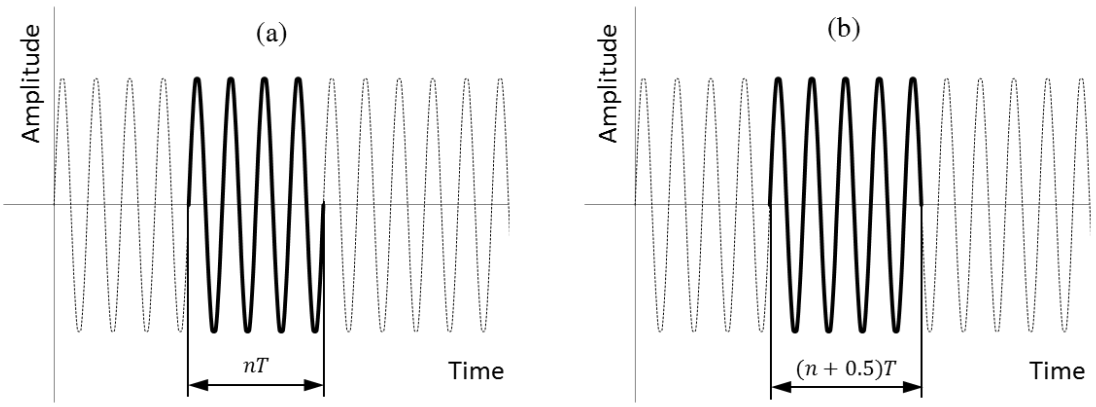


Figure 29 Acquisition of a sine wave during: (a) an integer number of periods; (b) a non-integer number of periods (b).

Since only a finite portion of the signal can be measured and the Fourier transform assumes that time signals are periodic and repeat infinitely in time, what the Fourier Transform will "see" is what is shown in figure 30 instead.

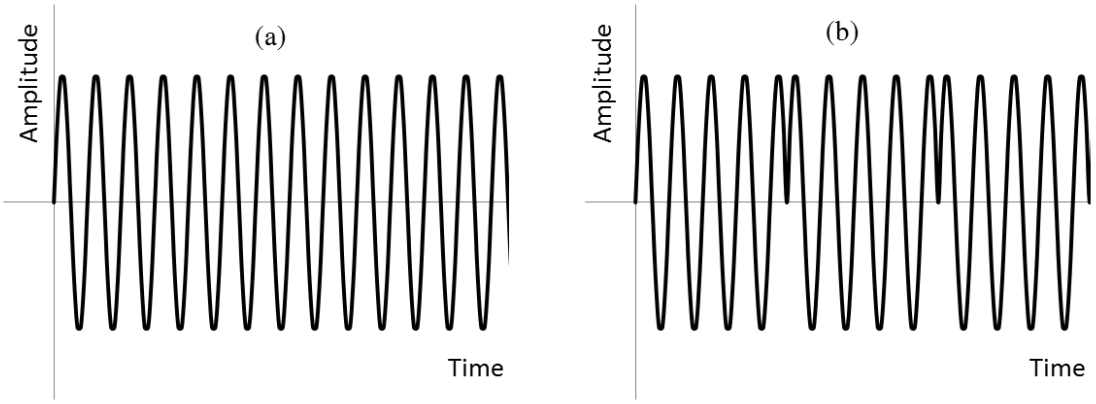


Figure 30 What the Fourier transform "sees" after the acquisition of the sine waves shown in figure 29: (a) integer number of periods; (b) non-integer number of periods.

Figure 30 (b) is no longer a harmonic signal; thus, when applying the Fourier transform, the signal shown in Figure 30 (b) will be regenerated as a sum of sine waves that will be translated into spectral components in the frequency domain. In practice, what happens is that the fundamental frequency is shown with smaller amplitude with spreading to neighbouring frequencies. The phenomenon of spreading of the true spectrum components to other frequencies is called *leakage* and is illustrated in figure 31.

This phenomenon can be seen as a distribution of the energy contained in the fundamental spectral line (which is the same in the particular examples shown in figures 29 and 30) to contiguous frequencies, as if the peak melted and leaked to the sides.

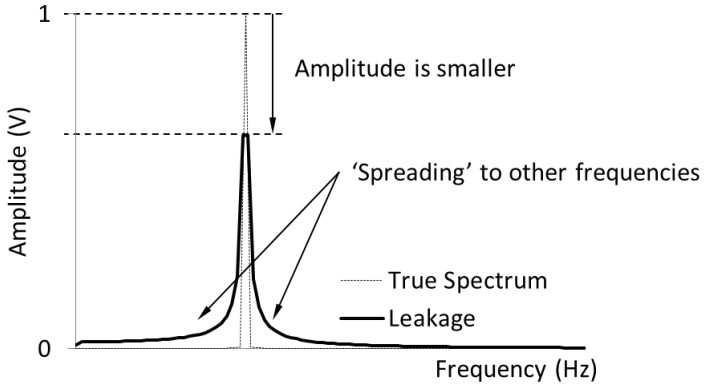


Figure 31 Overlapping of the true power spectra (dashed line) with the one obtained when the number of periods is a half integer (solid line). No window is applied to the time signal. Leakage occurs and is highly visible: the amplitude is different from its true value and 'spreading' of the true spectrum components to other frequencies is accentuated.

In real situations, it is very difficult to guarantee that an integer number of cycles is acquired during a measurement, even if the signal is periodic. Even in a laboratorial environment, where control can be taken over the excitation source, this may be impossible to guarantee if the signals are of the random type. To avoid - or at least minimize - the leakage phenomenon, a function, known as *window,* is multiplied by the time signal before the Fourier Transform is performed. The objective is to obtain a smooth decay to zero at the limits of the recorded time period, so that the resulting signal is continuous and approximates more closely to a periodic one. Figure 32 shows an example of how a window may change time signal before it is transformed into the frequency domain. In this example, a non-integer number of 16.5 cycles has been measured. It can be seen that, at the center, the windowed time signal has a periodic appearance (solid line) at the location where the original time signal presents a discontinuity (dashed line).

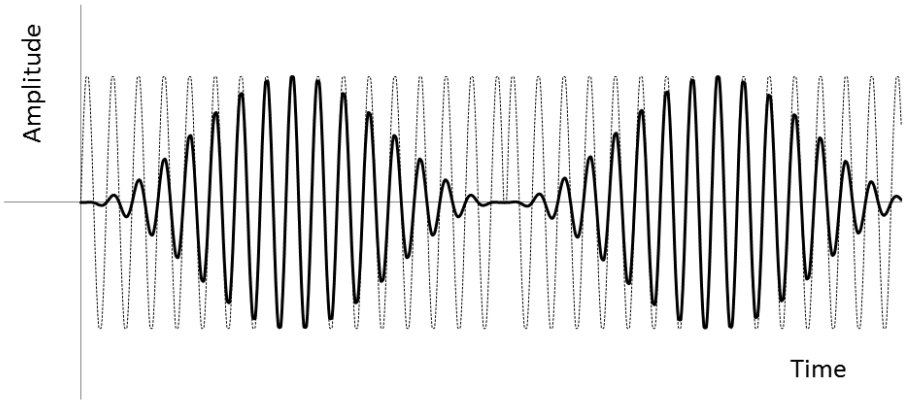


Figure 32 Representation of two measurement periods of a sinusoidal time signal containing a non-integer number of cycles (16.5). The dashed line is the time signal without window and the solid line is the time signal after being multiplied by a window function (Hanning, in the example).

Other applications of the window functions include FIR (Finite Impulse Response) filter design or beamforming, since the window itself can be seen as a type of a digital filter.

Many windows have been developed as an attempt to reduce as much as possible leakage and other problems related to its usage. Just to name a few, there are: Uniform (or rectangular), Barlett (or triangular), B-Spline, Welch, Hanning, Hamming, Blackman, Nutall, Flat top, Tukey, Rife-Vincent, Gaussian, Kaiser-Bessel, Slepian, Parzen, Bohman, Dolph-Chebyshev, Ultraspherical, Exponential (or Poisson), etc. There also exist hybrid windows that result from the combination of two windows, either by multiplication or summation. However, it is not the intent of this text to explore in depth how different windows compare, but rather to give an overview on the use of some windows to prevent leakage in typical applications.

In this book, windows were firstly divided into shapes: rectangular, triangular, bell-shaped and exponential. These shapes and their generating functions are represented in table 5. In this table, the most widely used bell-shaped windows are represented by a generalized cosine function, which coefficients are presented in table 4. However, it should be noted that not all the bell-shaped functions are generalized cosine functions, for example the Kaiser-Bessel. On the other hand, some functions, like the flat top, are not truly bell-shaped, since it is partially negative on the sides. In the particular case of the Hamming window, it is not zero at $\frac{T}{2}$.

The most commonly used window function is the Hanning one, which belongs to the bell-shaped category. This window was named after Julius Von Hann, a meteorologist who applied an equivalent process to meteorological data. When the Hanning window is applied to the signal presented by the solid line in figure 29 (b), the spectrum obtained after performing the Fourier Transform is the one shown in figure 33. When this spectrum is compared to the one obtained in figure 31, it is possible to observe that the spreading of the fundamental frequency

to neighbouring spectral frequencies is much less pronounced. Moreover, the amplitude of the signal was made closer to its true value. As such, the use of a window generally reduces the extent of leakage and the chances of important components of the signal being masked by the leaked components.

Besides the Hanning window, the rectangular and exponential windows are widely used as well. The rectangular or uniform window corresponds to a situation in which no window is used. This can be useful when the signal is truly periodic in time T and sampling is synchronous. The exponential window is useful for transient signals which had faded away within the time record. This can be particularly useful when measuring impulses or impacts.

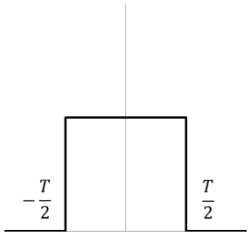
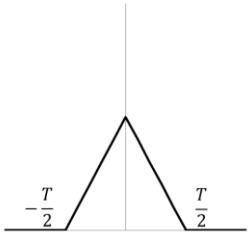
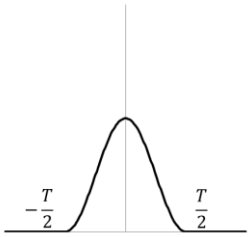
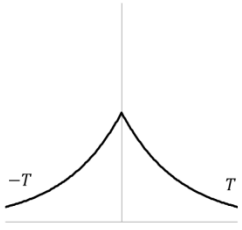
In summary, the choice of the window depends on the application. A few recommendations on their choice are left below:

1. Hanning: for general application (random signals).
2. Blackman: when spectral leakage must be minimized (random signals);
3. Flat Top: for accuracy in amplitude measurements (random signals);
4. Rectangular (or Uniform): for periodic signals with synchronous sampling;
5. Exponential: for transient signals (generated from impulses).

Table 4 Coefficients for typical bell-shaped window functions.

Window	n	a_0	a_1	a_2	a_3	a_4
Hanning	1	0.5	0.5	-	-	-
Hamming	1	0.54	0.46	-	-	-
Blackman	2	0.427	0.497	0.0768	-	-
Nutall	3	0.356	0.487	0.144	0.0126	
Flat Top	4	0.216	0.416	0.278	0.0837	0.00604

Table 5 Typical window functions and their shapes.

Window	Shape	Equation
Rectangular or Uniform **(no window)**		$w(t) = 1, \quad \lvert t\rvert \le \frac{T}{2}$ $w(t) = 0, \quad \lvert t\rvert > \frac{T}{2}$
Triangular **(Barlett)**		$w(t) = 1 - \frac{2\lvert t\rvert}{T}, \quad \lvert t\rvert \le \frac{T}{2}$ $w(t) = 0, \quad \lvert t\rvert > \frac{T}{2}$
Bell-shaped **(Generalised Cosine)**		$w(t) = \sum_{k=0}^{n} a_k \cos\frac{2\pi k t}{T}, \quad \lvert t\rvert \le \frac{T}{2}$ $w(t) = 0, \quad \lvert t\rvert > \frac{T}{2}$
Exponential	 Exponential	$w(t) = e^{-\left\lvert t-\frac{T}{2}\right\rvert\frac{1}{\tau}}$ with $\tau = 10\frac{T}{D}\log e$ and D the target decay in dB over half of the window length.

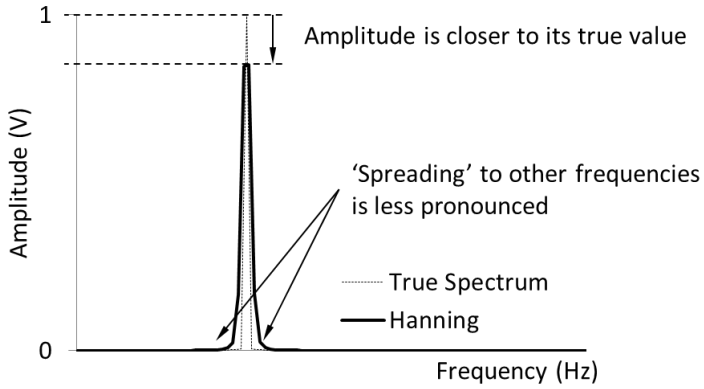


Figure 33 Overlap of the true power spectrum (dashed line) with the one obtained when the number of periods is a half integer (solid line). A Hanning window is applied to the time signal. Leakage still occurs but is less visible (solid line) than in figure 31. The amplitude is closer to its true value and 'spreading' of the true spectrum components to other frequencies is less pronounced.

4.4 Convolution

The convolution of two time signals is regarded by some authors as the most important technique in Digital Signal Processing (DSP) [14]. In convolution, systems are described by an Impulse Response Function (IRF).

Firstly, it is important to define what an impulse is. The simplest form of a non-periodic input function is the unit impulse, or Dirac δ-functon, represented by:

$$f(t) = \delta(t - \tau) \tag{55}$$

which is zero for all values of t except for $t = \tau$, where:

$$\lim_{\Delta t \to 0} \int_{\tau}^{\tau + \Delta t} f(t)dt = 1 \tag{56}$$

The impulse function just defined is represented in figure 34 by a unitary rectangular area of width Δt and height $1/\Delta t$, with $\Delta t \to 0$.

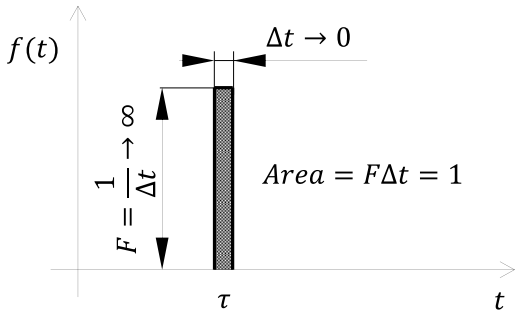


Figure 34 Definition of a unit impulse forcing function represented by a rectangular area of width Δt and height $1/\Delta t$, with $\Delta t \to 0$.

The unit IRF is defined as the response of a system for a unitary impulse of the type represented in figure 34, and is represented as $h(t-\tau)$. In other words, this function is the transfer function of the system for an unitary impulse input (figure 34).

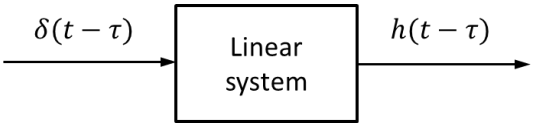


Figure 35 The IRF of a linear system is the output of the system when the input is a Dirac δ-function.

For linear systems, the principle of superposition applies. Linear systems are those in which the response is proportional to the excitation. For example, in SIMO (Single-Input-Multiple-Output) systems, if a given input is doubled, the resulting outputs are doubled as well (considering the system is linear). In MIMO systems, the response due to two simultaneous inputs is equal to the sum of the responses when the inputs are applied one at a time. Thus, for linear systems, the response to an arbitrary input function $f(t)$ may be taken as the superposition (sum, or integral) of the responses to a series of impulses that, when summed up, represent the input function. In this case, the output function can be represented as:

$$x(t) = \sum_{\tau} f(\tau)h(t-\tau)\Delta\tau \tag{57}$$

which, when taken to the limit as $\Delta\tau$ tends to zero, becomes:

$$x(t) = \int_{-\infty}^{t} f(\tau)h(t-\tau)d\tau \tag{58}$$

This integral is called the convolution or Duhamel's integral [6]. Substituting variable τ by $\tau = t - \kappa$, $d\tau$ becomes $-d\kappa$ and the limits of integration change. As a consequence:

$$x(t) = \int_{0}^{+\infty} f(t-\kappa)h(\kappa)d\kappa \tag{59}$$

This expression, in which $h(\kappa)$ plays a role of weighting function (also called memory) [15], is considerably simplified by taking the Fourier transform, which yields:

$$X(\omega) = H(\omega)F(\omega) \tag{60}$$

where $H(\omega)$ is the Fourier transform of the IRF. This is a very important result: firstly, convolutions transform to products [15]; secondly, it is possible to derive the Frequency Response Function (FRF)[4] for a given system just by taking the Fourier transform of the IRF [6].

The convolution can be seen as an operation between two signals, as a sum or a multiplication is. This is often represented with a $*$. In terms of applications, it can be used to solve many mathematical problems, ranging from statistics to differential equations, including DSP [14].

4.5 Random Signals

4.5.1 Auto-Spectrum, Power-Spectrum and Cross-Spectrum

In many real applications, signals are not deterministic. As such, random signals cannot be treated the same way as deterministic signals do, i.e., it is not actually correct to analyze them assuming a periodicity of infinite period [6]. This is because the Dirichlet condition cannot be satisfied for infinite random signals, i.e.:

$$\int_{-\infty}^{+\infty} |x(t)| dt < \infty \tag{61}$$

This condition must be satisfied for the Fourier transform to be used. Thus, the analysis of random signals is made using statistical approaches instead, like the *auto-correlation* function. The auto-correlation of an ergodic[5] signal $x(t)$ is the correlation of that signal with itself when measured after a time lag τ, $x(t+\tau)$:

$$R_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+\tau)dt \tag{62}$$

[4] The FRF – explained later - is a transfer function relating at least one output to one input (for example, the vibration response $x(t)$ to a given random excitation force $f(t)$)

[5] The term ergodic is used to describe a random process which time average of one sample of events is the same as the average of the whole sequence of events.

For a discrete signal, it takes the form:

$$R_{xx}(\delta) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} x_n x_{n+\delta} dt \tag{63}$$

This mathematical formulation allows finding repeating patterns in a signal, like a periodic signal that is invisible when covered away by excessive levels of noise.

To better understand how the auto-correlation function looks like, let us take as an example a 10 Hz harmonic sine wave with added random white noise:

$$x(t) = \sin(20\pi + \pi/3) + random\ noise$$

This signal, as well as its auto-correlation function (which is symmetrical if we consider the time period between $-T/2$ and $T/2$), are represented in figure 36 (for a 'moderate level' of added noise) and in figure 37 (for a 'high' level of added noise).

Regardless of noise, what it can be concluded is that the original time signal has been transformed into a new time signal that tends to zero as $\tau \to \infty$. This means that the Dirichlet condition (61) can now be satisfied and the Fourier transform can be applied on the auto-correlation function instead. When the Fourier transform pair given earlier by equations (42) and (43) is used, the stationary random process is described by the well-known Weiner-Khintchine relationships:

$$R_{xx}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} S_{xx}(\omega) e^{i\omega\tau} d\omega \tag{64}$$

$$S_{xx}(\omega) = \int_{-\infty}^{+\infty} R_{xx}(\tau) e^{-i\omega\tau} d\tau \tag{65}$$

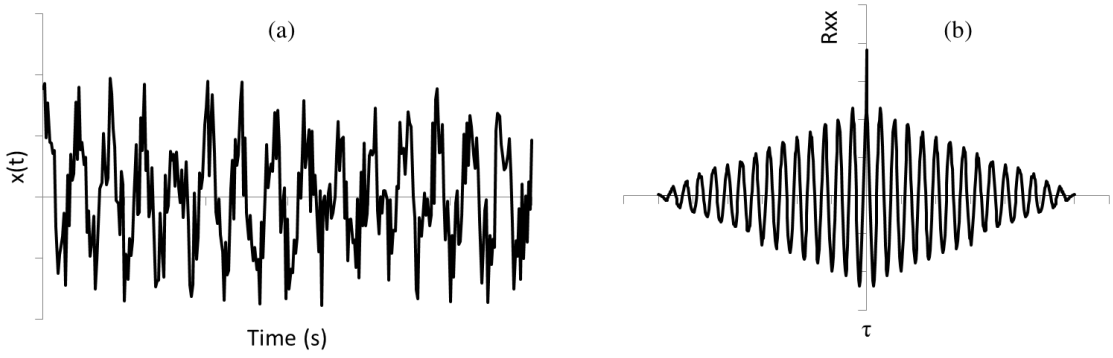


Figure 36 Example of a harmonic time signal with (a) added random white noise and (b) its auto-correlation function (for a 'moderate' level of added noise).

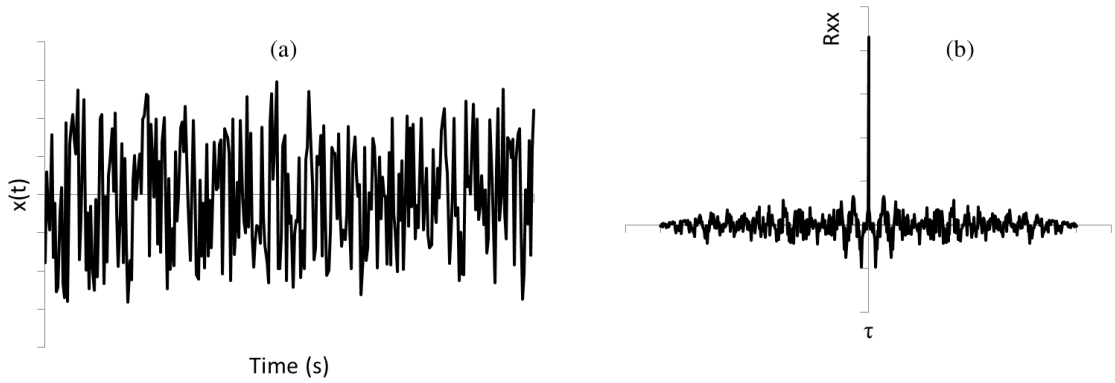


Figure 37 Example of a harmonic time signal with (a) added random white noise and (b) its auto-correlation function (for a 'high' level of added noise).

Equation (65) is also known as the Auto-Spectral Density (ASD) or Power-Spectral Density (PSD), which is a real and even (symmetric) function. The reason why this quantity is called PSD is better understood if we consider $\tau = 0$. In that case, equation (62) combined with (64) yields to:

$$R_{xx}(0) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(t)dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} S_{xx}(\omega)d\omega \qquad (66)$$

in which $x(t)$ is raised to the power of 2, $x^2(t)$.

The Fourier and Power Spectra of the time signals shown in figures 36 and 37 are represented in figures 38 and 39, respectively.

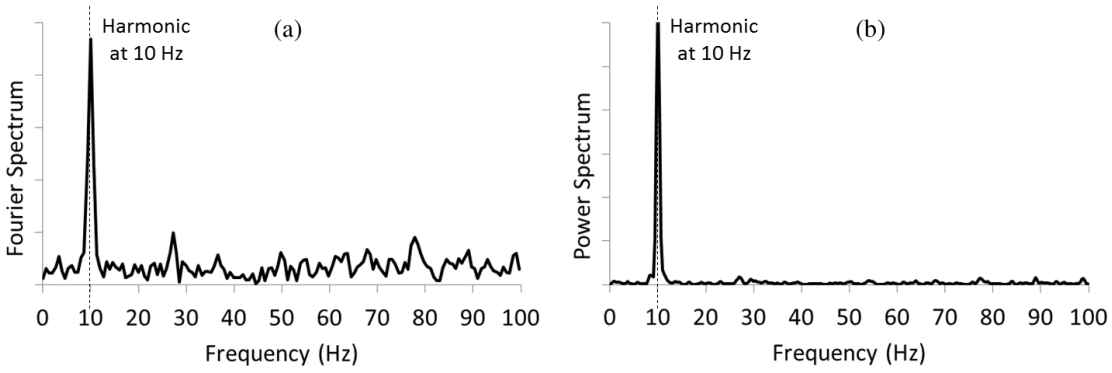


Figure 38 (a) Fourier Spectrum of the time signal shown in figure 36 (a). (b) Fourier Spectrum of the auto-correlation function shown in figure 36 (b).

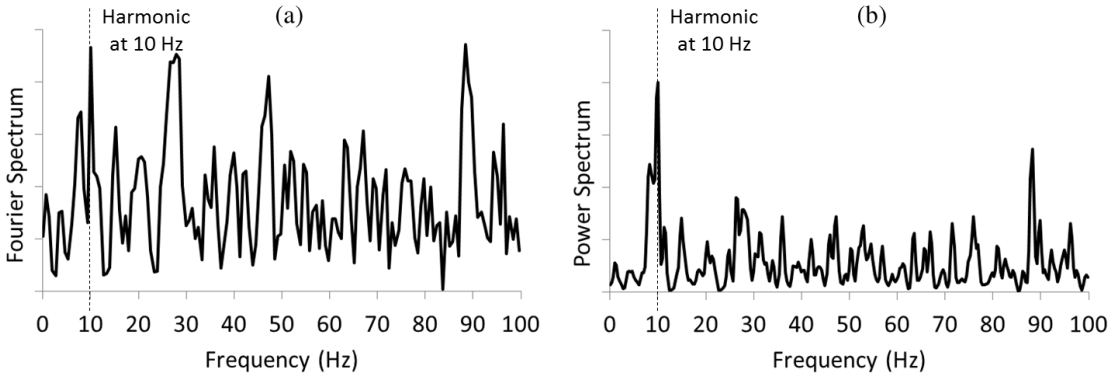


Figure 39 (a) Fourier Spectrum of the time signal shown in figure 37 (a). (b) Fourier Spectrum of the auto-correlation function shown in figure 37 (b).

These concepts can further be extended to correlate two different functions, for example as is the case of the measurement of the so-called Frequency Response Function (FRF). The FRF is a transfer function relating at least one output to one input (for example, the vibration response $x(t)$ to a given random excitation force $f(t)$). In this case, the PSD is called Cross-Spectral Density (CSD) or, more often, cross-correlation, and the Weiner-Khintchine relationships become:

$$\begin{aligned} R_{fx}(\tau) &= \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} f(t)x(t+\tau)dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} S_{fx}(\omega)e^{i\omega\tau}d\omega \end{aligned} \tag{67}$$

$$S_{fx}(f) = \int_{-\infty}^{+\infty} R_{fx}(\tau)e^{-i\omega\tau}d\tau \tag{68}$$

The CSD functions are complex functions (including real and imaginary parts, or magnitude and phase), whereas the PSD is a real function with magnitude only. It can also be shown that [6]:

$$\begin{aligned} R_{fx}(\tau) &= R_{xf}(-\tau) \\ S_{fx}(\omega) &= {S_{xf}}^{*}(\omega) \end{aligned} \tag{69}$$

where the superscript * stands for complex conjugate.

It should be noted that the CSD - equation (67) – is the convolution between $f(t)$ and $x(t+\tau)$, thus illustrating another application for the convolution earlier described in section 4.4.

Although these concepts are just briefly introduced in this text, they are quite important in Mechatronics, as they are used in a wide range of applications, e.g., in condition monitoring, in control, in speech analysis or in electronic communication systems.

Figure 40 Left: a spectrum analyser and signal acquisition unit from Brüel & Kjær. Right: a Dell laptop computer and National Instruments Co. signal acquisition unit. The photos were taken within a 15 year interval.

In the past, spectral analysers, although very powerful and reliable, were bulky and expensive equipment. Today, they have been replaced by powerful computers and smaller data acquisition units. Even everyday consumer products, like mobile phones, are capable of producing spectra from measured time signals (e.g., sounds obtained with the microphone or a mechanical vibration measured with the built-in accelerometer). A comparison between what can be used nowadays and what was the state-of-the-art twenty years ago is shown in figure 40.

4.5.2 Estimators

For a harmonic excitation, the FRF is the relationship between the response $x(t)$ (output) to an the excitation $f(t)$ (input):

$$H(\omega) = \frac{x(t)}{F(t)} \tag{70}$$

which, in terms of block diagram, can be represented as in figure 41.

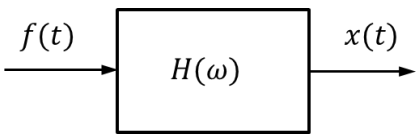


Figure 41 Frequency Response Function (FRF) for a harmonic excitation on a SISO (Single-Input-Single-Output) system.

The function $H(\omega)$ is a transfer function that represents the system. However, equation (70) is not valid unless the signals are harmonic. Thus, the importance of the auto-correlation and further developments explained in the previous section 4.5.1 for non-periodic signals.

The convolution of the response with force can, however, be applied. It is possible to show that [15]:

$$|H(\omega)|^2 = \frac{S_{ff}(\omega)}{S_{xx}(\omega)} \tag{71}$$

and

$$H_1(\omega) = \frac{S_{xf}(\omega)}{S_{xx}(\omega)} \tag{72}$$

Equation (72) represents one of the three versions for the FRF estimators, and that is why an index of 1 was added next to the letter H. This estimator is the 'conventional' FRF estimator and is determined by using the cross input-output spectrum and the input auto-spectrum. It is said to be unbiased with respect to noise on the output.

Another version of the FRF estimator, $H_2(\omega)$, is obtained by dividing the output auto-spectrum by the cross input-output spectrum:

$$H_2(\omega) = \frac{S_{ff}(\omega)}{S_{fx}(\omega)} \tag{73}$$

This estimator is said to be unbiased with respect to noise on the input.

In principle, both estimators $H_1(\omega)$ and $H_2(\omega)$ should yield the same result. However, in practice, this may not happen, due to many reasons:

1. The signals contain noise;
2. The system relating the input and output is not linear;
3. The measured output $x(t)$ is not a consequence of the input $f(t)$ alone, but also from other non-quantified external inputs.

Hence, an indicator of the quality of the analysis can be defined as the ratio of the two estimators:

$$\gamma^2(\omega) = \frac{H_1(\omega)}{H_2(\omega)} = \frac{\left|S_{xf}(\omega)\right|^2}{S_{xx}(\omega)S_{ff}(\omega)} \tag{74}$$

where $\gamma^2(\omega)$ is called the *ordinary coherence function*, which can assume values in the interval $[0, 1]$. Basically, the coherence can be seen as a measure of the 'correlation' between two signals. Here, the term 'correlation' does not have the sense of statistical correlation, but linear relation instead.

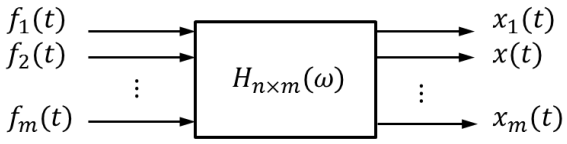


Figure 42 Frequency Response Function (FRF) for a harmonic excitation on a MIMO (Multiple-Input-Multiple-Output) system.

For MIMO systems (figure 42), with m inputs written in the vector form $\{f(t)\}$ and n outputs written in the vector form $\{x(t)\}$, this can be further generalized to obtain, for example:

$$[H(\omega)]^T = [S_{xx}(\omega)]^{-1}\left[S_{fx}(\omega)\right] \tag{75}$$

assuming $[S_{xx}(\omega)]$ is invertible and $[H(\omega)]$ is an $n \times m$ matrix of frequency response functions.

4.5.3 Ensemble Averaging

As mentioned before, the auto-correlation is a function that is used to find patterns in random events. Nevertheless, signals in random processes are inherently noisy by nature. If noise is random, one way to cancel it out (or, at least, reducing it) is by using *ensemble averaging*.

Typically, spectrum analysers bring two ensemble averaging weighting options: linear or exponential. According to the LabVIEW™ help documentation [16] on spectral measurements, linear and exponential averaging operate in the following way:

1. Linear - averages are made over a specified number n of measurement time periods T in a non-weighted manner, i.e., each set of data weighs exactly the same as in an arithmetic mean.
2. Exponential - averages are made over a specified number n of measurement time periods T in a weighted manner. Exponential averaging gives the most recent sets of data more weighting in the average than in older data.

Averaging is performed on spectral quantities. Thus, the following options are generally available as well [16]:

1. Vector averaging - the average of complex FFT spectra quantities is computed directly. Vector averaging eliminates noise from synchronous signals;
2. RMS averaging - averages the energy, or power, of the FFT spectrum of the signal;
3. Peak hold - performs averaging at each frequency line separately, retaining peak levels from one FFT record to the next.

Another method to reduce noise consists in using *overlapping*, in what is also known as the Welch's method. Overlapping is often defined in terms of a % of the time signal. The time signal is split into a set of segments that contained shifted data with finite duration. These segments are then overlapped by a % of the total number of points in one segment. When overlapping is 0%, this process is known as the Barlett's method. Overlapping segments are then windowed. Because most window functions give more importance to the data at the center of the set of data, the overlapping technique compensates for that.

4.6 Butterworth Filter

Filters are used to remove unwanted contents from a signal. For example, there are filters that can selectively remove certain features, like noise or spurious spectral components, to avoid error propagation within an algorithm. Other example might be to simulate how another sensor rather than the one being used would perceive a signal, e.g. filtering a sound obtained with a microphone to represent how the human hearing would have perceived it.

Filters can be linear, non-linear, time-invariant, analogue, digital, passive, active, etc. Filters can also be low-pass, high-pass, band-pass or band-reject. Low-pass filters are designed to cut-off high frequencies, whereas high-pass filters are the exact opposite. Band-pass filters are designed to pass all the signal components within a finite frequency band, blocking away the components outside that band. Band-reject is the opposite to band-pass.

One particular example that most of us are familiar with, is the Dolby Noise Reduction system used in audio. Basically, the Dolby system consists on the use of band-pass filters, both during recording and playback. In essence, the Dolby system consists of an encode/decode system in which the amplitude frequencies in one band is increased during recording and decreased during playback, proportionally. Basically, what this does is to increase the Signal-to-Noise Ratio (SNR) rather than filtering out noise as in subsequent DSP.

Linear continuous-time filters are amongst the most used filters in signal processing, out of which the low-pass Butterworth filter probably is the most popular. Nevertheless, filters like the Chebyshev (types I and II), Elliptical, and others, are quite popular as well. The three aforementioned are time-invariant filters which are based on their analogue counterparts and fit in the IIR (Infinite Impulse Response) category. Filters can also be FIR (Finite Impulse Response), depending on the way they are implemented. IIR filters are the most efficient to implement in DSP. However, IIR filters are not as stable as FIR filters are, because phase-response is not linear and IIR is difficult to implement in hardware while FIR can be efficiently realized on hardware. Ultimately, IIR filters are less accurate than FIR, but faster in a way.

The Butterworth filter is a low-pass filter that is designed to block-off all signal components above a certain frequency limit, while keeping the frequency response's passband as flat as possible (with uniform sensitivity). The magnitude of its transfer function (frequency response function) is represented in figure 43 and given by:

$$|H(i\omega)| = \frac{1}{\sqrt{1 + \left(\frac{i\omega}{i\omega_c}\right)^{2N}}} \tag{76}$$

where N is the order (number of poles) in the filter and ω_c is the cut-off frequency in rad/s.

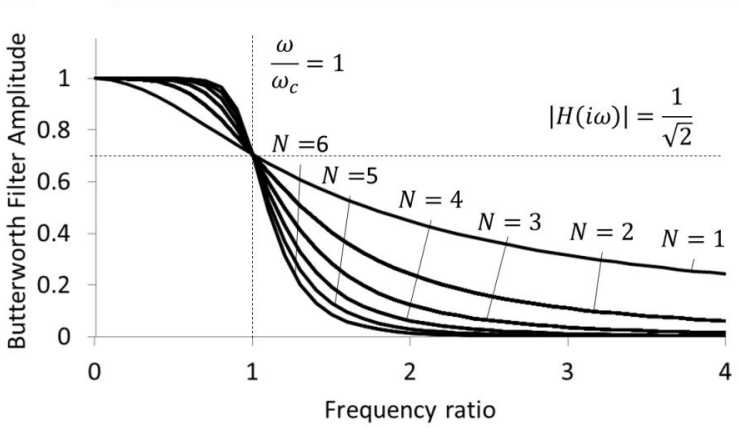


Figure 43 Butterworth low-pass filter gain amplitude for different number of poles.

The Butterworth filter has a slow roll-off (decay) when compared to other filters like the Chebishev or Elliptic (equiripple) filters, although it does not present ripple (a small residual periodic variation). When the graphs in figure 43 are presented in a log-log plot (figure 44), it is possible to have a better understanding on how the filter rolls-off with frequency for different numbers of poles. For example, in a 2nd order Butterworth filter, when the frequency increases by a factor of 10 (i.e., one decade), the magnitude drops 40 dB. In other words, the roll-off is -40 dB/decade.

The order of the filter is quite important, as it defines how the filter rolls-off in the frequency domain. The higher the number of poles, the faster it rolls-off and the higher the attenuation in the stopband. However, this will of course lead to more complicated algorithms and circuits, as Butterworth filters are based on polynomials.

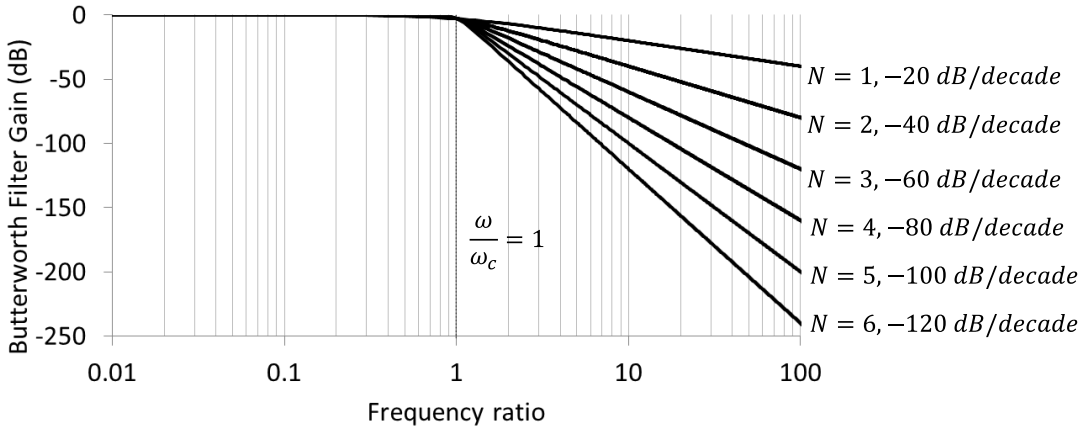


Figure 44 Butterworth low-pass filter gain amplitude and decay for different number of poles (log-log plot).

The design of a digital filter requires mathematical manipulation that is outside the scope of this text. In the case of the Butterworth low-pass filter, the steps can generically be summarized into the following:

1. Define the filter topology (cut-off frequency, ripple, out of band attenuation, etc.);
2. Transform the continuous-time domain filter into a discrete-time domain filter. This can be done, for example, using a bilinear transform;
3. Transform the frequency response function from the frequency domain to the time domain. This can be done by using the z-transform. The z-transform is a special case of the Fourier transform where $e^{i\omega}$ is replaced by the variable z ($= e^{i\omega}$).

Generalizing for any even order Butterworth low-pass filter, the discrete-time system transfer function can be expressed as [17]:

$$H(z) = \frac{\sum_{k=0}^{N} a_k z^{-k}}{1 + \sum_{k=1}^{N} b_k z^{-k}} \tag{77}$$

where:

$$\begin{gathered} a_0 = a_k = \frac{\Omega_c^2}{c(k)}, k = 0, 2, 4, 6, \dots (k\ even) \\ a_k = 2a_{k-1}, k = 1, 3, 5, \dots (k\ odd) \\ b_k = \frac{2}{c_k} - 1, k = 2, 4, 6, \dots (k\ even) \\ b_k = \frac{2(\Omega_c^2 - 1)}{c_k}, k = 1, 3, 5, \dots (k\ odd) \end{gathered} \tag{78}$$

and

$$\begin{gathered} c_k = 1 + 2\cos\left[\frac{\pi(2k+1)}{2N}\right]\Omega_c + \Omega_c^2 \\ \Omega_c = \tan\left(\pi\frac{f_c}{f_s}\right) \\ N = 0, 2, 4, 6, \dots \end{gathered} \tag{79}$$

where f_c is the cut-off frequency in Hz and f_s is the sampling frequency. The filtered time signal is [17]:

$$x'(t) = \sum_{k=0}^{N} a_k x(t-k) - \sum_{k=1}^{N} b_k x'(t-k) \tag{80}$$

The Butterworth filter is a recursive filter: it requires previous filtered time samples (as much as the number of poles N). Because the first filtered time samples are not available, initial guessing is required. One way of doing this consists in, for example, padding (adding as much zeros at the beginning of the signal as the unknown data), or assuming that the first samples of the filtered signal are the same as the unfiltered one.

Let us take as an example the following waveform:

$$x(t) = \sin(14\pi + \pi/3) + random\ noise$$

This is a 7 Hz sine wave with added noise. Let us assume this signal was acquired with a sampling frequency of 100 Hz and that we want to set the digital cut-off frequency at 10 Hz. Also, consider a 40 dB roll-off, i.e., a 2nd order filter. In this case, the Butterworth polynomial coefficients are:

$$a_0 = 0.067455$$

$$a_1 = 0.134911$$

$$a_2 = 0.067455$$

$$b_1 = -1.14298$$

$$b_2 = 0.412801$$

Figures 45 and 46 show the result of the use of this 2nd order Butterworth filter, on the time and frequency domains, respectively. The Butterworth filter was very effective in minimizing high frequency noise (practically eliminated it completely), although it was not as effective in the 10 to 20 Hz frequency range. On the other hand, due to the slow and progressive roll-off of the Butterworth filter, the amplitude of the component at 7 Hz was slightly attenuated by the filter, even if it is below the 10 Hz cut-off frequency.

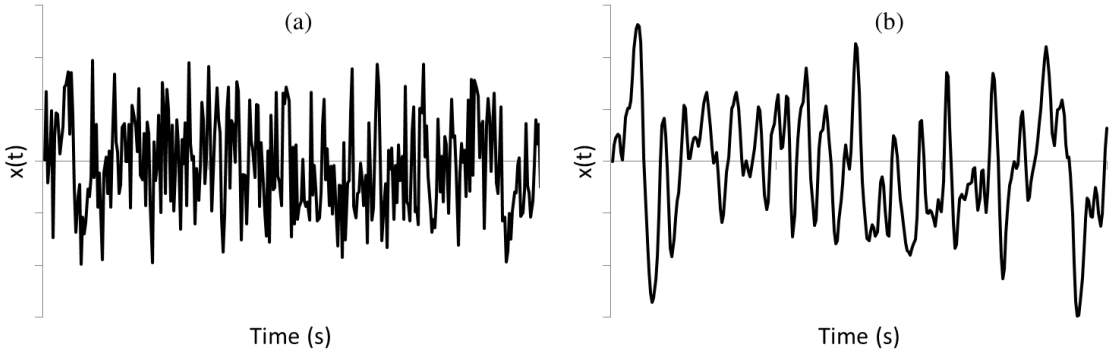


Figure 45 (a) Time waveform of a 7 Hz sine wave with added random noise. (b) The same time waveform after a Butterworth filter is applied with a cut-off frequency at 10 Hz.

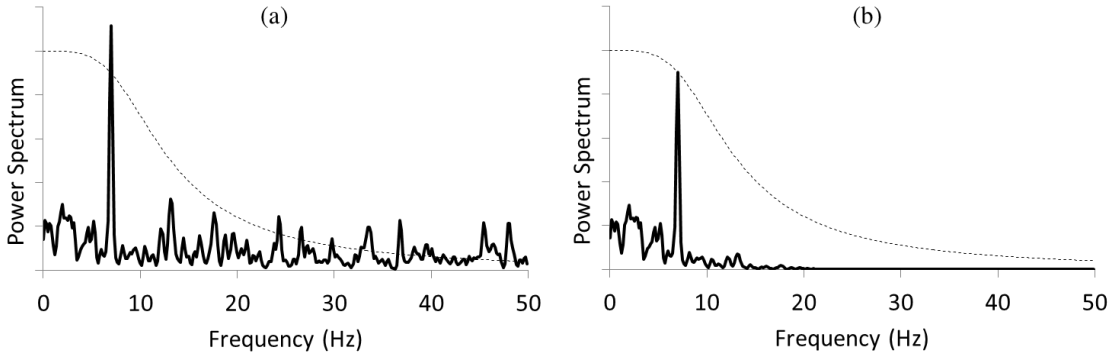


Figure 46 (a) Power spectrum of the time signal shown in figure 45 (a). (b) Power spectrum of the time signal shown in figure 45 (b) (after being filtered with a low-pass Butterworth filter). The dashed line in both pictures represents the Butterworth frequency response function.

The op-amp circuit for the Butterworth 2nd order low-pass filter can take the Sallen-Key topology shown in figure 47 [18,19].

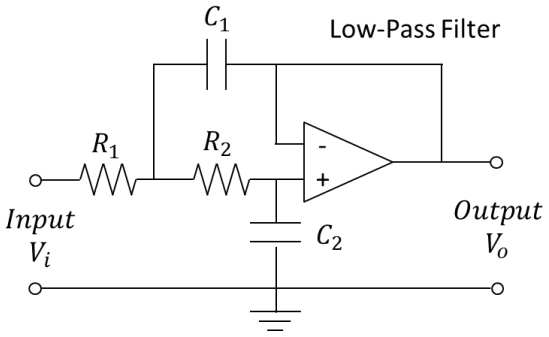


Figure 47 Op-amp circuit of a Butterworth 2nd order low-pass filter.

Defining the constants:

$$\begin{aligned} \tau_1 &= R_1 C_1 \\ \tau_2 &= R_2 C_2 \\ \tau_3 &= R_1 C_2 \end{aligned} \tag{81}$$

the transfer function for the circuit shown in figure 47 is [18]:

$$H(s) = \frac{V_o}{V_i} = \frac{1}{\tau_1 \tau_2 s^2 + (\tau_2 + \tau_3)s + 1} = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n\, s + \omega_n^2} \tag{82}$$

where s is the Laplace variable, ω_n is the undamped natural frequency and ζ is the damping ratio. From (82), the undamped natural frequency and the damping ratio are, respectively:

$$\omega_n = \frac{1}{\sqrt{\tau_1 \tau_2}} \tag{83}$$

$$\zeta = \frac{\tau_2 + \tau_3}{2\sqrt{\tau_1 \tau_2}} \tag{84}$$

For this second-order transfer function to become oscillatory, is is necessary that $\zeta < 1$. Furthermore, ideally, the system should have a zero-resonant frequency $\omega_r = 0$:

$$\omega_r = \sqrt{1 - 2\zeta^2}\omega_n = 0 \tag{85}$$

Thus,

$$\zeta = \frac{1}{\sqrt{2}} \tag{86}$$

and, as a result:

$$(\tau_2 + \tau_3)^2 = 2\tau_1\tau_2 \tag{87}$$

The zero-resonant frequency ω_r must be zero so that the gain is flat and unitary until the cut-off frequency. If, instead of the transfer function, one plots the amplification factor (also called gain), this might be easier to understand (figure 48). The amplification factor of any 2nd order oscillatory system (either Mechanical or Electrical) is given by [6]:

$$Q(\beta) = \frac{1}{\sqrt{(1-\beta^2)^2 + (2\zeta\beta^2)}} \tag{88}$$

where β is the frequency ratio:

$$\beta = \frac{\omega}{\omega_n} \tag{89}$$

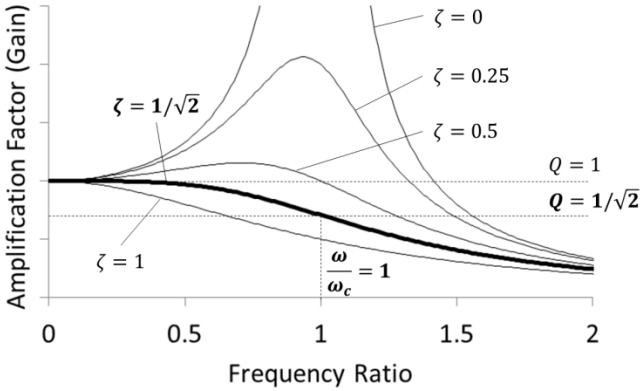


Figure 48 Plots of the amplification factor for different values of the damping ratio.

What figure 48 shows us is that for a value of the damping ratio $\zeta = 1/\sqrt{2}$, the maximum amplification factor (gain) is 1. At a value of the frequency ratio $\beta = 1$, the amplification factor is $Q = 1/\sqrt{2}$. This is equivalent to what is illustrated in figure 43: for a unitary frequency ratio

the filter's gain is $1/\sqrt{2}$. Thus, what this tells us is that ω_n in equation (89) actually is the cut-off frequency ω_c, i.e.:

$$\omega_c = \frac{1}{\sqrt{\tau_1 \tau_2}} = \frac{1}{\sqrt{R_1 R_2 C_1 C_2}} \tag{90}$$

In his original work [20], Butterworth also showed that it is possible to design a high-pass filter from the low-pass counterpart by switching the inductances with the capacitances and vice-versa. Furthermore, he expanded the concept to band-pass filters. The corresponding op-amp circuits for 2nd order filters are illustrated in figure 49.

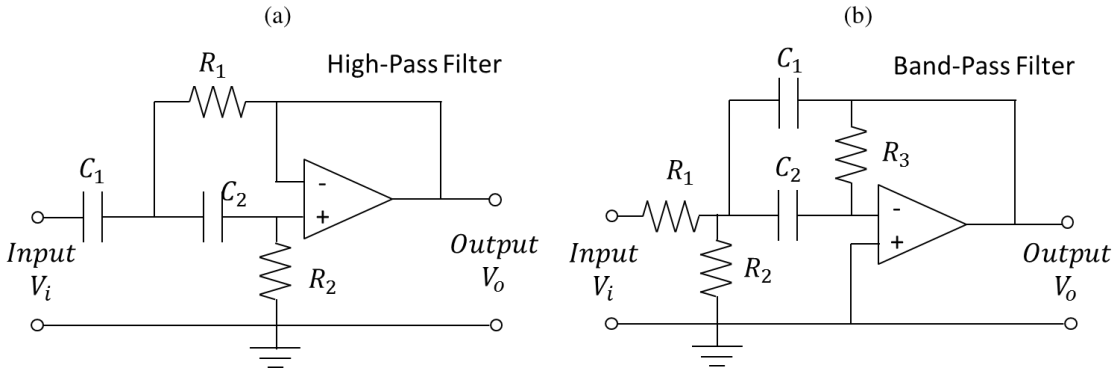


Figure 49 Op-amp circuits of Butterworth 2nd order (a) high-pass and (b) band-pass filters.

4.7 Smoothing Filters

Smoothing is a technique that is used to reduce noise across the whole frequency range of a signal, without filtering out meaningful components. Typically, smoothing consists on constructing an approximating function that attempts to capture important patterns in a signal. However, this must not be confused with curve-fitting methods, as these involve the use of an explicit function form for the result (e.g., with physical meaning), whereas smoothing techniques, strictly speaking, do not.

4.7.1 Moving Average

Many smoothing algorithms have been developed for signal processing. One of the most popular is the *moving average*, where a series of averages of different subsets of the data set are created. This should not be confused with ensemble averaging, as earlier introduced in section 4.5.3. While ensemble averaging operates in successive sets of data, the moving average operates on a single data set exclusively. Ensemble averaging is in itself a particular technique of smoothing, although it is not a filter.

The moving average works by replacing each data point with an averaged value of the neighbouring data points defined within the span. This is similar to a low pass filter (see section 4.6) with the smoothing function being given by [21]:

$$x'_s(t) = \frac{1}{2N+1}\sum_{i=-N}^{N} x(t+i) \tag{91}$$

where $x'_s(t)$ is the smoothed value for the t^{th} data point, N is the number of neighbouring data points on either side of $x'_s(t)$ and $2N+1$ is the span.

As an example, the waveform originally presented in figure 45 (a) and to which a Butterworth low-pass filter was applied resulting in figure 45 (b), has now been *smoothed* using the moving average equation (91). The resulting smoothed waveform, for a value of $N = 3$, is shown in figure 50 (b).

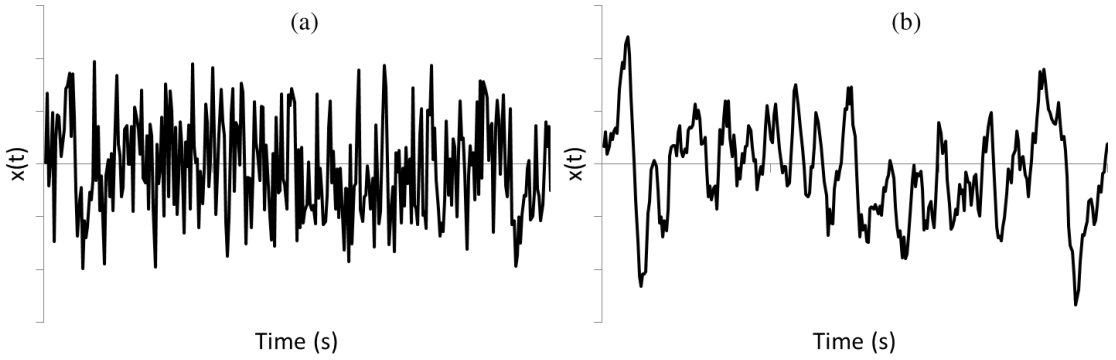


Figure 50 (a) Time waveform of a 7 Hz sine wave with added random noise. (b) The same time waveform after a moving average smoothing filter is applied with $N = 3$.

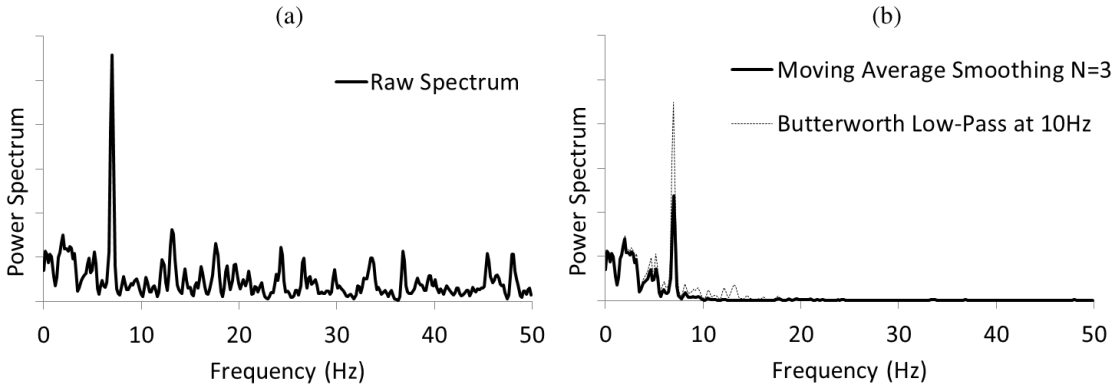


Figure 51 (a) Power Spectrum of the time signal shown in figure 50 (a). (b) Power spectra of the time signals shown in figures 50 (b) (after being smoothed with a moving average with $N = 3$) and 45 (b) (filtered with a low-pass Butterworth filter with cut-off frequency at 10 Hz).

As with the Butterworth filter, it can be observed that high frequency noise has been practically eliminated, although it still is visible in the low frequency range. Nevertheless, the amplitude of the existing harmonic has been highly attenuated as a result, even more than when the Butterworth filter was used.

4.7.2 Savitzky-Golay

The Savitzky-Golay filtering is another popular technique used for smoothing. It can be seen as a generalized weighted moving average. The filter has coefficients that are derived by fitting a polynomial of a chosen degree with least-squares fit. The advantage of the Savitzky-Golay filtering against the simple moving average is that it allows reducing attenuation of data features by increasing the order of the polynomials.

Savitzky and Golay [22] showed that a set of integers c_i could be derived and used as weighting coefficients to carry out the smoothing operation. The use of these weighting coefficients, known as convolution integers, turns out to be exactly equivalent to fitting the data to a polynomial and is computationally more effective and much faster [23]. The smoothed data point $x'_s(t)$ by the Savitzky-Golay algorithm is given by the following equation:

$$x'_s(t) = \frac{\sum_{i=-N}^{N} c_i x(t+i)}{\sum_{i=-N}^{N} c_i} \tag{92}$$

where coefficients c_i are obtained through a least-squares fit and $c_i = c_{-i}$. Typical coefficients for quadratic and cubic Savitzky-Golay filters are shown in table 6. Usually, Savitzky-Golay filters are applied to odd sized windows $(2N+1)$ only, although the concept has further been extended for even sized windows $(2N)$ more recently [24].

A quadratic/cubic Savitzky-Golay smoothing filter with $N = 3$ was used in the same example as earlier presented in both sections 4.6 and 4.7.1, resulting in figures 52 and 53.

In comparison to the moving average smoothing (and even to the Butterworth filter), it is clear that the Savitzky-Golay smoothing filter preserved much better the spectral component at 7 Hz (which is a data feature), even if it was not as effective in noise attenuation.

Table 6: Typical coefficients for quadratic and cubic Savitzky-Golay filters. Note that $c_i = c_{-i}$.

N	$2N+1$	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
2	5	17	12	-3										
3	7	7	6	3	-2									
4	9	59	54	39	14	-21								
5	11	89	84	69	44	9	-36							
6	13	25	24	21	16	9	0	-1						
7	15	167	162	147	122	87	42	-13	-78					
8	17	43	42	39	34	27	18	7	-6	-21				
9	19	269	264	249	224	189	144	89	24	-51	-136			
10	21	329	324	309	284	249	204	149	84	9	-76	-171		
11	23	79	78	75	70	63	54	43	30	15	-2	-21	-42	
12	25	467	462	447	422	387	343	287	222	147	62	-33	-138	-253

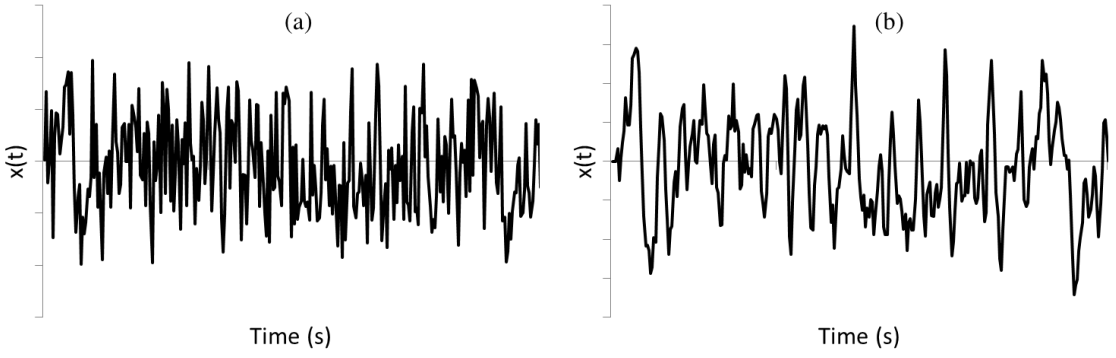


Figure 52 (a) Time waveform of a 7 Hz sine wave with added random noise. (b) The same time waveform after a Savitzky-Golay smoothing filter is applied with $N = 3$.

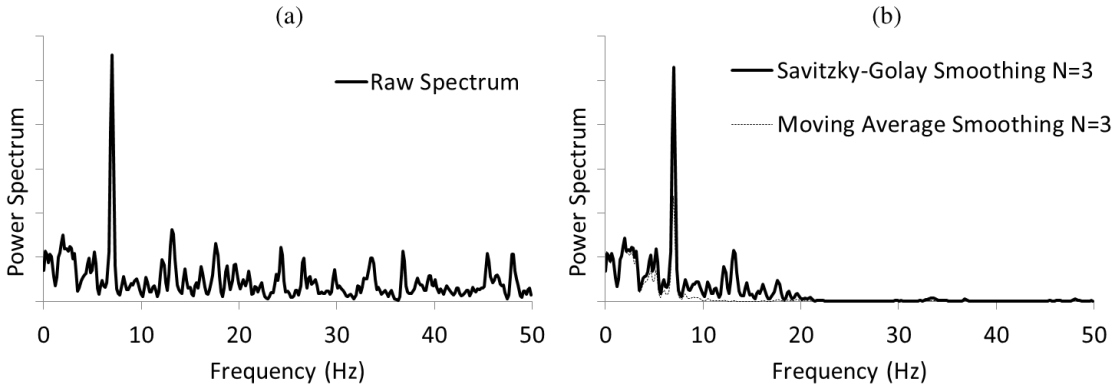


Figure 53 (a) Power spectrum of the time signal shown in figure 52 (a). (b) Power spectra of the time signals shown in figures 52 (b) (after being smoothed with a Savitzky-Golay filter) and 50 (b) (after being smoothed with a moving average with $N = 3$).

Chapter 5

Sensors

5.1 Introduction

Although there may exist other types of sensor classifications, most sensors would fit into active or passive and analogue or digital. Active sensors require external power for their operation. An example is an Integrated Electronic Piezoelectric (IEPE) accelerometer. On the contrary, a passive sensor does not need external power for operation, as is the case of a charge accelerometer.

An analogue sensor produces an output that is proportional to the variable being measured. The relationship between the output and the variable being measured is not necessarily linear, although typically it is within the operating range. This relationship is called the calibration factor (or curve). Typically, the maximum analogue output voltage is 10 V and the maximum analogue current is between 4 to 20 mA.

A digital sensor is one that converts the measuring signal into a digital signal before it is transmitted to the acquisition system. Typically, it has built-in electronics. As an advantage when compared to the analogue sensor, the digital data transmission is not sensitive to cable length, resistance or impedance, and is not affected by electromagnetic noise.

A schematic of a generic sensor setup is shown in figure 54.

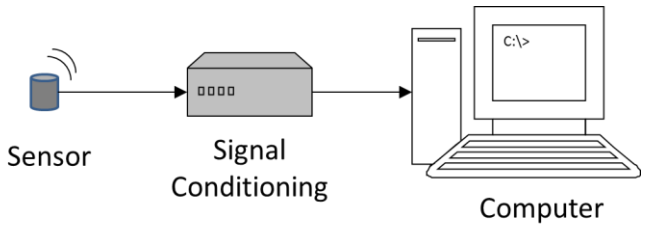


Figure 54 Generic sensor setup.

In this section, an introductory overview of conventional sensors used in Mechatronics and applications is presented. However, in this book, we do not intend to deliver an all-encompassing review on sensor technology and fundamentals. It must be noted that sensors are a vast area that already deserved the publication of entire textbooks on the topic, for example [18].

5.2 Accelerometers

Accelerometers measure motion in the form of acceleration. When used in vibration applications, typical units are ms^{-2} or g ($1\ g = 9.81\ ms^{-2}$).

5.2.1 Piezoelectric Accelerometers

Piezoelectric accelerometers are composed by three main components: a base with a central post, a piezoelectric crystal and a seismic mass (figure 55). The base is attached to the structure via a stud (thread), glue, an adhesive pad, bee's wax or a magnetic base.

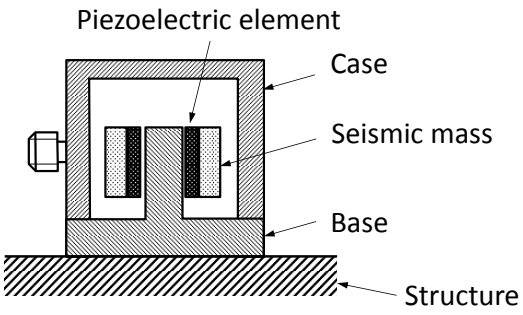


Figure 55 Schematic cross-sectional view of a piezoelectric accelerometer.

The base of the accelerometer moves with the motion of the structure and, because the piezoelectric element and seismic mass are attached to the central post, they will follow motion. It must be noted that the piezoelectric element has some elasticity. Thus, the seismic mass and the piezoelectric element behave as a mass-spring SDOF (Single Degree of Freedom) system to which motion is transmitted. When the piezoelectric crystal is deformed, a charge in pC (pico-Coloumb) is produced. This charge is proportional to acceleration, according to Newton's 2nd law.

These devices typically work in a fairly wide range of frequencies, typically between 1 Hz and 30 kHz. Since the seismic mass and piezoelectric element constitute the elements of a SDOF, the accelerometer itself has a resonant frequency that is usually shown in the calibration chart. A rule of thumb is that an accelerometer is usable up to 1/3 its resonant frequency. In the example shown in figure 56 (Brüel & Kjær charge accelerometer type 4501) the resonant frequency of the accelerometer is 50 kHz, which means that, according to this rule of thumb, the accelerometer is usable (with approximate constant sensitivity) up to 16.7 kHz. This value actually is very close to the one presented by the manufacturer in its calibration chart (16.6 kHz) [25].

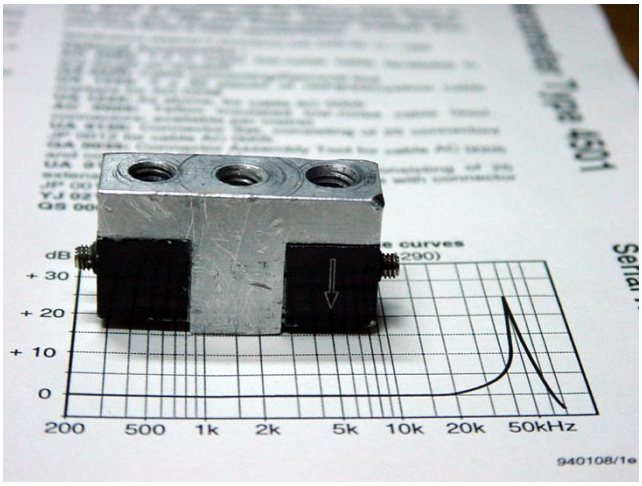


Figure 56 Calibration chart of a Brüel & Kjær charge accelerometer type 4501 showing its frequency curve. The two accelerometers are mounted on a T-shaped block used to measure rotations. It is also possible to see the measurement axis, indicated with an arrow.

The charge accelerometer fits in the passive analogue sensor category. Before the signal is sent to the acquisition unit or computer, it is necessary that the charge in pC is converted to voltage in mV (mili-Volt) in a device named charge amplifier. The setup of a charge accelerometer is shown in figure 57. As a limitation, it requires the use of an expensive low-noise cable between the accelerometer and charge amplifier and the measurement is sensitive to the low-noise cable's length. On the other hand, as an advantage, charge amplifiers like the

one shown in figure 25 allow adjusting the dynamic range, which is made by setting the output value in terms of mV/ms^{-2}.

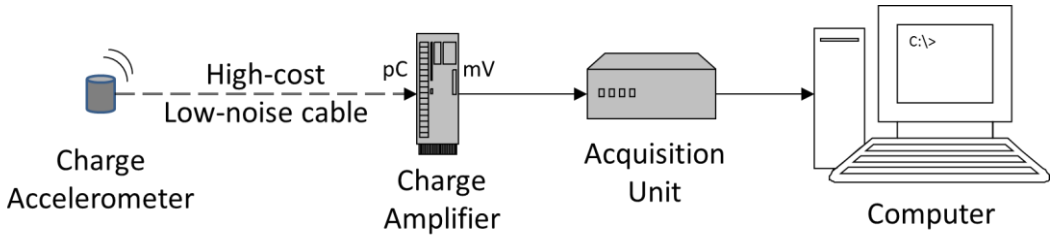


Figure 57 Charge accelerometer setup.

A more modern type of accelerometer is the IEPE, which has got a built-in pre-amplifier. In this case, the charge amplifier is not needed, but the acquisition unit must power the accelerometer (figure 58). It is also recommended that the acquisition unit has an anti-aliasing low-pass filter before the ADC is performed.

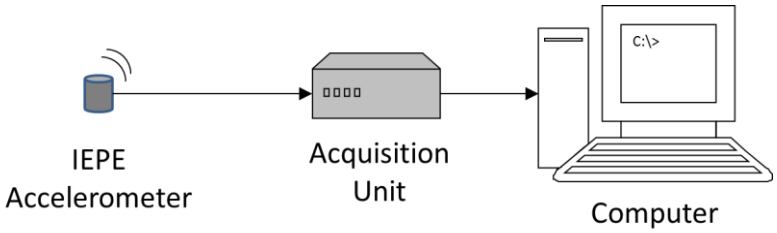


Figure 58 IEPE accelerometer setup.

In comparison to the charge accelerometer, the IEPE has the advantage of having a fixed sensitivity regardless of cable length and quality, being more portable (practically plug-and-play) and to require less expensive conditioners. Nevertheless, as limitations, it has a smaller (and fixed) dynamic range and its operating temperature is limited (maximum of $120\,℃$ against up to $480\,℃$ for the charge accelerometer). Furthermore, if the electronics in the IEPE are damaged due to, for example, mishandling and dropping on a hard floor, the accelerometer will not work, whereas a charge accelerometer may still keep functional, even if the sensitivity may have changed.

Piezoelectric accelerometers are used in many applications, from condition monitoring to Structural Health Monitoring (SHM) or Modal Analysis. An example of application is illustrated in figure 59, in which the vibration levels on a generator's 8 cylinder engine are being measured at different coordinates and directions.

Typically, accelerometers measure linear acceleration in a single direction (although they still have a small transverse sensitivity up to 4%). There also exist triaxial accelerometers that can measure along the three directional axes, but they are more bulky and expensive.

For the measurement of structural rotations, piezoelectric accelerometers have not proven as efficient. In a technique that is many times referenced in the literature [6, 26, 27], two piezoelectric accelerometers are attached at the tips of a solid t-shaped block (usually made from aluminum), as represented in figure 60 (a photo of a t-shaped block was previously shown in figure 56).

Once the accelerations $\ddot{x}_A$ and $\ddot{x}_B$ are known, and as long as the frequency range is small enough to keep the t-shaped block behaving as a rigid body even if the structure is not, the linear and angular accelerations at the pivot point O of the structure can be determined, with reasonable accuracy, from:

$$\ddot{x} = \frac{\ddot{x}_A + \ddot{x}_B}{2} \tag{93}$$

$$\ddot{\theta} = \frac{\ddot{x}_A - \ddot{x}_B}{2s} \tag{94}$$

Figure 59 Two IEPE accelerometers from CSI are being used to monitor the vibration level on a generator's 8 cylinder engine at different coordinates and directions. The accelerometers are attached to the structure by means of two threaded magnets.

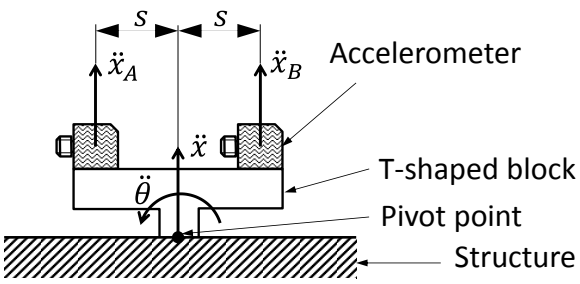


Figure 60 Schematics of the t-shaped block technique to measure rotations with linear (translational) accelerometers.

5.2.2 Piezoresistive and Capacitive Accelerometers

The piezoresistive and capacitive accelerometers work from the principle of deformation of a cantilever-like beam (figure 61). When the structure moves, the seismic mass will produce a force proportional to acceleration (Newton's 2nd law) deflecting the flexible cantilever beam. In the piezoresistive accelerometer, strain gauges are used to measure strain in the flexural element (strain gauges are later discussed in section 5.5). The capacitive accelerometer is based on the principle that, as the seismic mass moves from its equilibrium position, the capacitance between the electrodes is changed proportionally to acceleration.

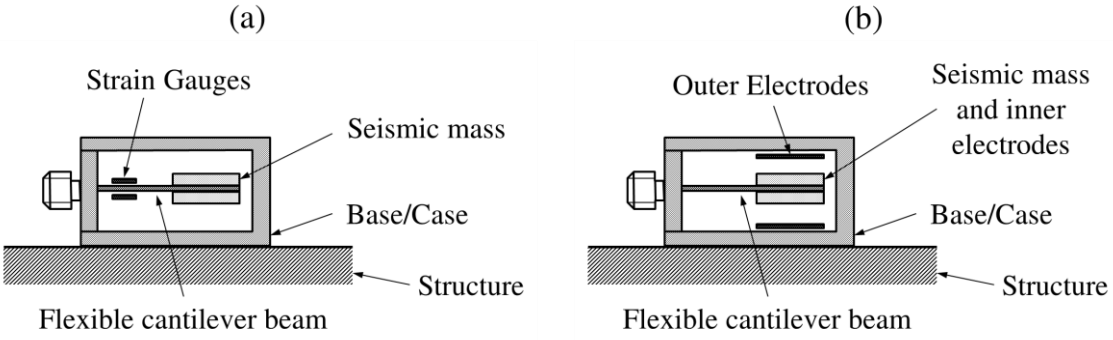


Figure 61 Schematic cross-sectional view of (a) a piezoresistive accelerometer and (b) a capacitive accelerometer.

The greatest advantage on the use of piezoresistive or capacitive accelerometers (when compared to piezoelectric accelerometers) is that they allow DC measurements (0 Hz). As limitations, they are very sensitive to temperature changes and the dynamic range is limited due to their higher low linearity.

5.3 Velocity Transducers

5.3.1 Laser Doppler Velocimeters

Laser Doppler Velocimeters (LDVs) measure velocity of a moving target using the Doppler effect. When used for vibration measurements, typical units are mms^{-1} or ips.

Laser is the English acronym for *Light Amplification by Stimulated Emission of Radiation*. It's functioning is based on the stimulated emission of photons: light is collimated. Two waves are said collimated if they have the same length, a constant relative phase and travel at parallel directions (consequently, collimated light is monochromatic as well).

He-Ne laser vibrometers produce red light in the visible spectrum with a wave length of 632.8 nm. The principle behind a laser vibrometer is interference, i.e., the addition between two collimated waveforms. When two laser beams interfere with one another, there is a phase and wave length relationship between the two waves that can be determined, even if the beams have travelled different distances. In other words, the relationship between the light that is emitted to a target and the light that is reflected from that target contains information about the velocity and position of the moving object. It can be said that motion modulates the phase of the light wave, whereas velocity modulates the optical frequency. However, because a He-Ne laser light has a frequency of $4.74 \times 10^{14}\ Hz$, it is not practical (or possible) to demodulate the signal directly and interferometry techniques must be used [28].

In the modified Mach-Zehnder interferometer (figure 62), the reflected light is combined with a reference beam so that both the emitted and reflected signals interfere with one another. Inside the interferometer, the laser is split into reference and test beams in a polarizing beam splitter PBS1. The reference beam f_0 is transmitted to the photodiode PD passing through a mirror M, Bragg cell BC and beam splitter BS2. The Bragg cell is a device that adds a known frequency f_B shift to the reference beam f_0, i.e., $f_0 + f_B$. This frequency f_B is called the carrier frequency, usually generated by quartz at 40 MHz. The test beam f_0 is sent directly to the target in motion and reflected back. The motion of the target adds a Doppler shift f_D to the reflected light, so it becomes $f_0 + f_D$. The reflected light comes in through the system of lenses P and L and beam splitter BS2, reaching the photodiode PD where the reference and test beams are collected and interfered.

The use of a laser vibrometer has some advantages when compared to the use of accelerometers. One of the greatest advantages is that it is a contactless technology, meaning it does not add mass to the system. The effects of added mass may be relevant in vibration measurements, especially for lightweight systems. Moreover, being contactless means it can measure on moving parts like rotating shafts or turbine blades, it can measure on locations of otherwise difficult access, and it can be used to measure on high-temperature surfaces. In terms of disadvantages, it requires that the test surfaces are somewhat reflective (reflectors can be used), it is a considerably expensive technology and, as any non-seismic sensor, may be affected by its referential, for example, external ground motion.

In terms of advanced applications, LDV suppliers offer scanning and 3D measurement solutions, although the cost may become a hindrance to the use of such technology. Figure 63 shows an experimental setup where a scanning laser vibrometer is used for the development of a SHM technique on composite materials based on the assessment of mode shape changes [29, 30].

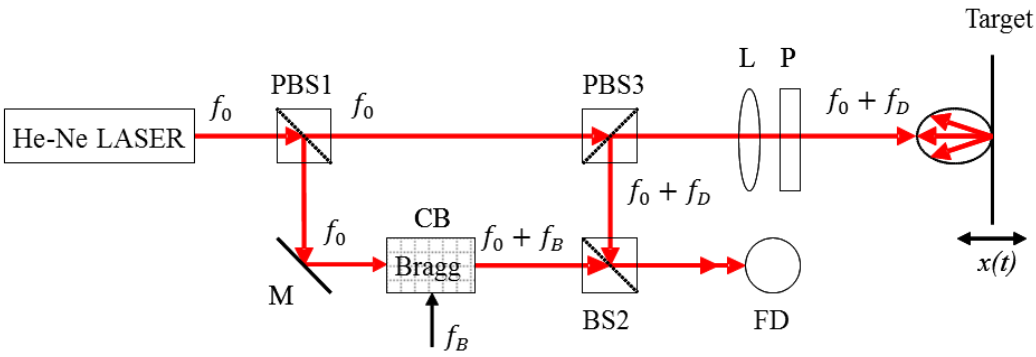


Figure 62 Schematics of the principle of operation of a Mach-Zehnder laser interferometer.

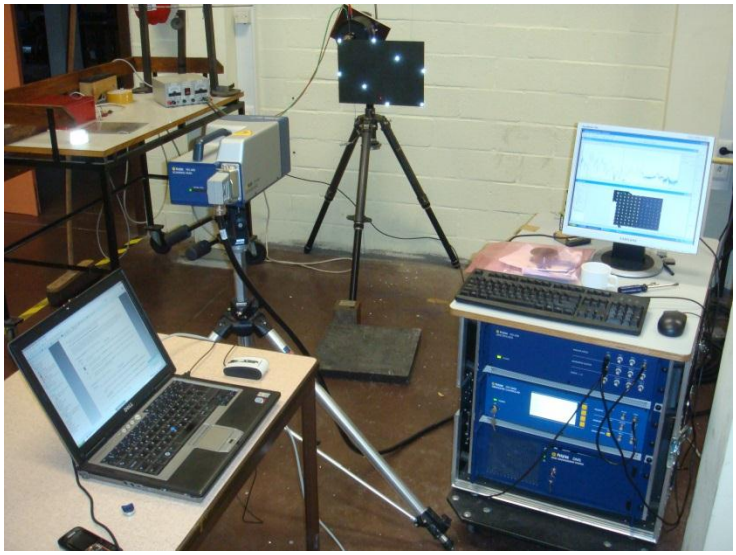


Figure 63 Test setup for the measurement of the vibration modes of a free-free suspended carbon fiber plate with a Polytec scanning LDV [29].

5.3.2 Tachometers

A tachometer is a device that is used to measure velocity on a rotating object, like a shaft, disk or toothed gear. Typical units are rpm (rotations per minute), although this can be easily converted to other more convenient unit, like cps or Hz (cycles per second) or the SI unit rad/s:

$$1\,rpm = \frac{1}{60} Hz = \frac{2\pi}{60} rad/s \tag{95}$$

One example of a tachometer is shown in figure 64. This is a handheld digital tachometer that measures the angular velocity of the rotating disk probe at the tip. The probe's surface is coated with rubber to increase friction. With this particular instrument, it is possible to either

measure the rotation on a disk or shaft or to measure the linear velocity on straight objects like the conveyor belt shown.

Figure 64 A digital tachometer is used to measure the velocity of a conveyor belt.

One type of sensor that is frequently used to measure velocity on rotating shafts is the proximity probe. The main difference between the proximity probe and the tachometer shown in figure 64 is that the proximity probe is a contactless device. Proximity probes can also be used to measure a distance or to just act as a switch (see section 4.4.3.3). The basic principle assumes that the rotating object must have some type of reference, like a reflector, magnets or teeth. In the example case shown in figure 65, the proximity probe used is of the electromagnetic type, operating under the principle of eddy-currents, like many speedometers that are still being used in cars today.

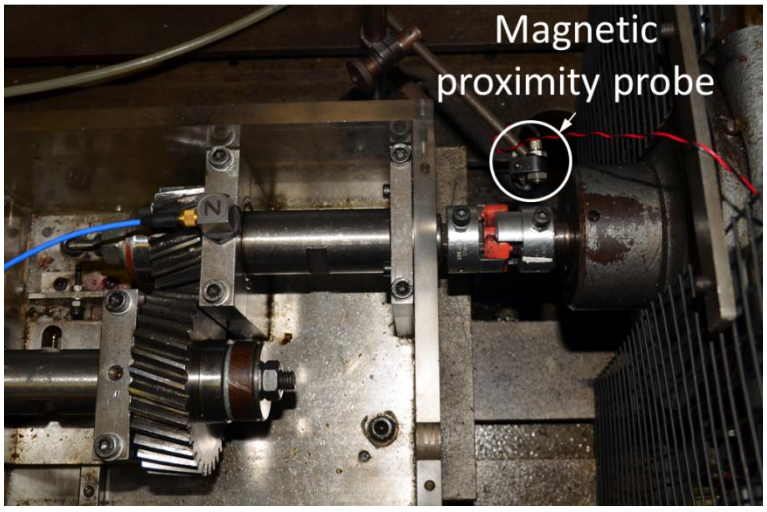


Figure 65 A magnetic proximity probe is used as a tachometer to measure the angular velocity of a driving shaft in a gear box.

Eddy currents are circular induced currents produced within conductors that can generate their own magnetic fields. Any material that is electrically conductive will conduct an induced current when exposed to a varying magnetic field. The stronger the applied magnetic field and the faster the field changes, the greater the currents that are developed.

Electromagnetic proximity probes are extremely sensitive to small displacements; hence, they can be used to measure either a displacement or a velocity. Most of the proximity probe based tachometers work the following way: when the shaft in figure 65 is rotating, the protruding head of the steel bolt (which works as a reference) passes in front of the proximity probe every time the shaft completes a full turn. Every time the head of the bolt passes in front of the probe the magnetic field increases. Based on these changes, a signal processing system counts the amount of times the head of the bolt passes in front of the proximity probe per second.

There are, however, many other technologies available, for example an infrared sensor that uses the reflection of an infrared light in a strip of reflector to measure rotational speed.

5.4 Displacement Transducers

5.4.1 LVDT

Linear Variable (or Voltage) Differential Transformers (LVDTs) have been used extensively for the accurate measurement of translational displacement and for the control of position within closed loop systems [31]. When used for vibration measurements, typical units are mm or μm. Similarly, a RVDT (Rotational Variable Differential Transformer) is used to measure angular displacements (rotations). LVDTs are robust and durable sensors.

The principal of operation is explained using figure 66. The LVDT is composed by a ferromagnetic core that is placed coaxially to three solenoid coils (one primary at the center and two secondary outer coils). The primary coil is connected to an AC power supply at an excitation frequency of 1 to 10 kHz. This produces an alternating electromagnetic field at the center of the transducer. Depending on the position of the core, this will induce a voltage in the secondary coils. When the core changes position, the induced voltages change. Because the coils are connected as shown in figure 66, this will produce a voltage differential at the output, which will be related to the position of the core. When the core is at the center position, this differential should be zero (in principle), due to symmetry. When the core occupies the extreme left or right positions, this differential reaches a maximum value. However, the differential has a signal, given by the phase between the output and the primary

coil current. If the differential is determined from the difference between the left secondary coil and the right secondary coil, in this order, when the core moves to the left the differential will grow from zero and its value is positive, whereas when the core moves to the right the differential will decrease from zero and its value is negative. This way, it is possible to know, exactly, the position of the core inside the transducer.

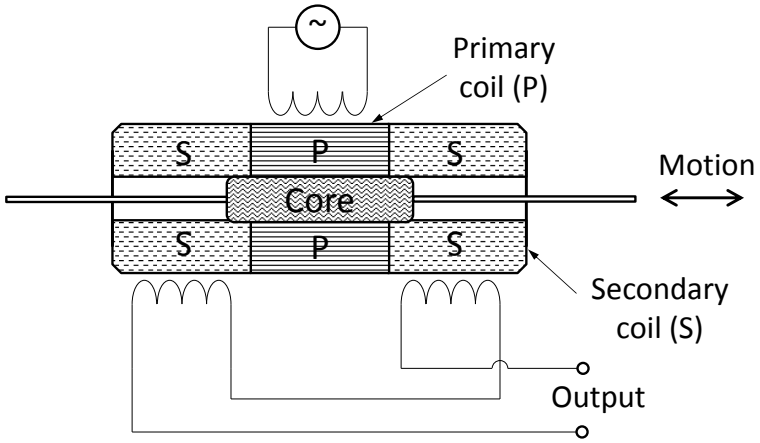


Figure 66 Schematics of the principle of operation of an LVDT.

There is a clearance between the core and the coil so that, when in operation, the core does not touch the coil and thus motion is frictionless. This makes of the LVDT a very reliable and durable sensor with virtually no wear when properly used.

Figure 67 shows an example of application of an LVDT to measure the displacement of a SDOF system when under free vibration.

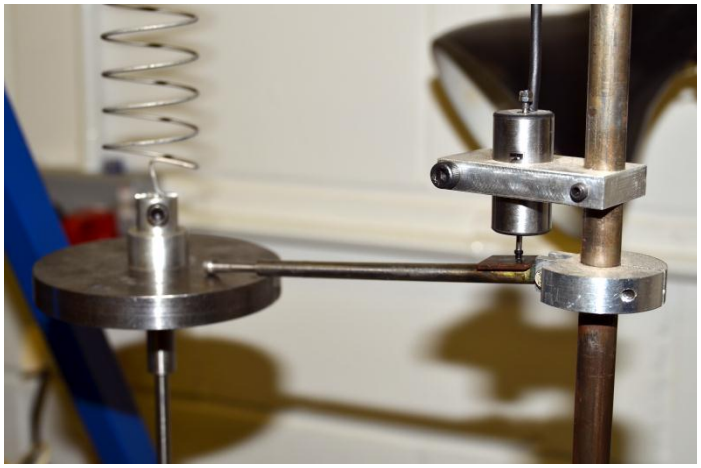


Figure 67 Example of application of an LVDT to measure the displacement of a SDOF free vibrating system. The LVDT is the cylindrical sensor on the right.

5.4.2 Laser

Lasers can also be used for contactless measurement of displacement or position, and not only velocity. In this case, the principle of operation has nothing to do with the Doppler effect as in LDVs. Instead, they are based on the principle of triangulation of light.

The principle of triangulation is illustrated in figure 68: the laser beam reflected by the structure passes through a receiver's lens and hits a CCD (Charge-Couple Device) sensor. The CCD is composed by a grid of photo detectors, like photodiodes, that are able of converting light into voltage.

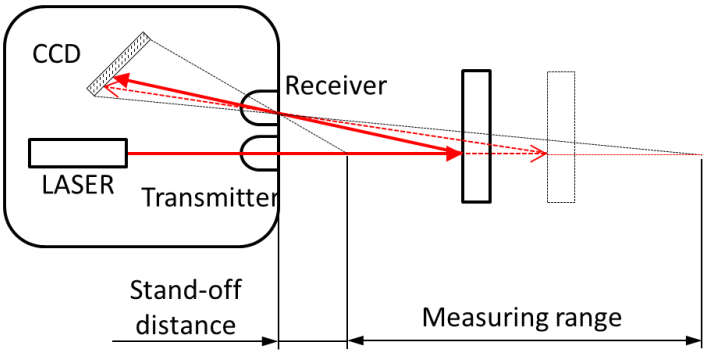


Figure 68 Principle of triangulation of laser light for the measurement of displacement.

These laser devices can be extremely sensitive: some are even used for checking and mapping roughness. Other applications include micro-positioning, semiconductor, silicon wafers, lenses or circuitry production, vibration measurement and differential thickness measurement.

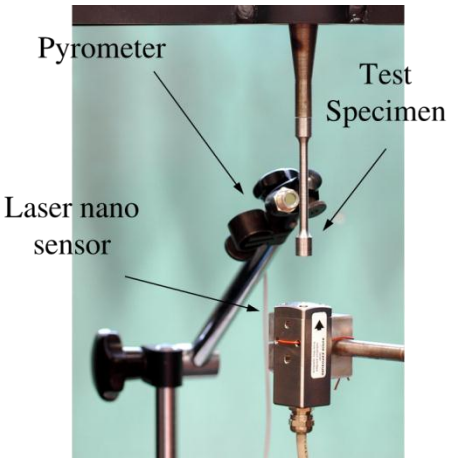


Figure 69 Using a laser nano sensor to measure displacement at the tip of a specimen in a Very High Cycle Fatigue (VHCF) machine.

Figure 69 shows an application in which a FLW laser nano sensor is being used to determine the displacement at the tip of a specimen being tested in a Very High Cycle Fatigue (VHCF)

machine [32]. In a VHCF test the specimen is designed to have the same fundamental axial frequency as the excitation source. In the case shown, the fundamental axial frequency of the specimen is 20 kHz. Because the temperature generated at the center of the specimen will increase dramatically, a pyrometer is used to monitor temperature and trigger a cooling function whenever temperature increases above a certain limit. The abrupt changes in temperature make strain gauges unreliable, because temperature affects their resistance. Also, their life is shorter than the specimen being tested, and will eventually stop functioning before the end of the test. As an alternative, a laser nano sensor is used to measure the displacement at the tip of the specimen, which can be related to the strain in the middle of the specimen, and thus stress, using analytical equations [32].

The laser nano sensor shown in figure 69 has a working frequency up to 200 kHz and has an output response that is approximately linear up to 5 V. The stand-off distance is $1.02\ in\ (26\ mm)$, meaning the distance between the laser and the measurement target must be fixed and precise within a tight tolerance for linear output, contrary to a laser vibrometer where distance to the target is immaterial (as long as focus is possible to get). The maximum measurement range, for linear output results, is $0.13\ mm$, but the resolution is extremely fine ($22.1\ nm$). The major inconveniences with the use of such a sensitive technology is that the measurement range is somewhat limited, the measurement surfaces must be very well polished (mirror like) and the surface must be perpendicular to the laser beam path.

5.4.3 Proximity Probes

Proximity probes can be used to measure not only velocity (as illustrated in section 4.4.2.2), but also a distance or the presence of an object that crosses the line of action of that sensor. This is because, above all, a tachometer of the proximity probe type is a counter. If the sensor is of the electromagnetic type, working under the eddy-currents principle, it will be highly sensitive to a change in the magnetic field, which can be related to the distance between the probe and the object (as long as the object is electrically conductive).

Figure 70 shows an infrared proximity probe that is used to detect the presence of bottles in a conveyor. The proximity probe has both a transmitter and a receiver. It sends a beam of infrared light that, once reflected by an object, is detected by the sensor. As long as the target is within a certain distance, called *nominal distance*, the sensor is capable of detecting the passage of any infra-red reflective object.

In the case of a bottle of glass, an electromagnetic sensor would not be suitable. The sensing technology in which the proximity probe bases its operation depends on some properties of the target as well.

Figure 70 An infrared proximity probe is used to detect the passage of bottles in an industrial bottle washing machine.

5.4.4 MEMS Sensors

MEMS (Micro Electro-Mechanical System) sensors are smaller versions of some of the previous sensors. They have been gaining popularity in the past recent years. They exist in many consumer products today, e.g., smartphones, game consoles, digital cameras, GPS devices, tablets, etc.

MEMS *accelerometers* operate under the same principles as capacitive and piezoresistive accelerometers (see section 4.4.1.2.). As with the capacitive accelerometers, they can measure accelerations down to 0 Hz (statics). This means they can sense the acceleration of gravity, and thus orientation with respect to earth, or a centripetal force in uniform circular motion, which conventional piezoelectric accelerometers used for sensing vibrations cannot (section 4.4.1.1.).

MEMS *gyros* are capable of measuring angular velocities around one or more axis. They are many times confused with gyroscopes, which is a different type of sensor. MEMS gyros are based on the Coriolis effect rather than on a gyroscopic reaction of a rotating body. The Coriolis and gyroscopic effects are different physical phenomena. The Coriolis acceleration appears in bodies that are moving at the radial direction of a rotating body (as it happens on earth when an aircraft travels from South to North and vice versa), whereas the gyroscopic reaction occurs as a consequence of the conservation of the angular momentum when the axis of a rotating body changes direction (as it happens in a bicycle, thus giving the rider balance).

Figure 71 depicts the principle behind a single-axis MEMS gyro of the tuning fork type. Two masses are kept vibrating in opposite directions with velocity v. When a rotation about the pivot point, with angular velocity ω, is produced, the fact that the masses are moving at radial directions with respect to the rotational axis will create a Coriolis acceleration on each one of the two masses. From Newton's second law, it is possible to derive a a force proportional to the mass m as:

$$F = 2m\boldsymbol{\omega} \times \boldsymbol{v} \tag{96}$$

where × denotes the vector cross-product. The Coriolis force generated is enough to displace the masses from their equilibrium position. Because these masses are moving electrodes that are placed between fixed electrodes, a change in the electrical capacitance proportional to displacement is induced.

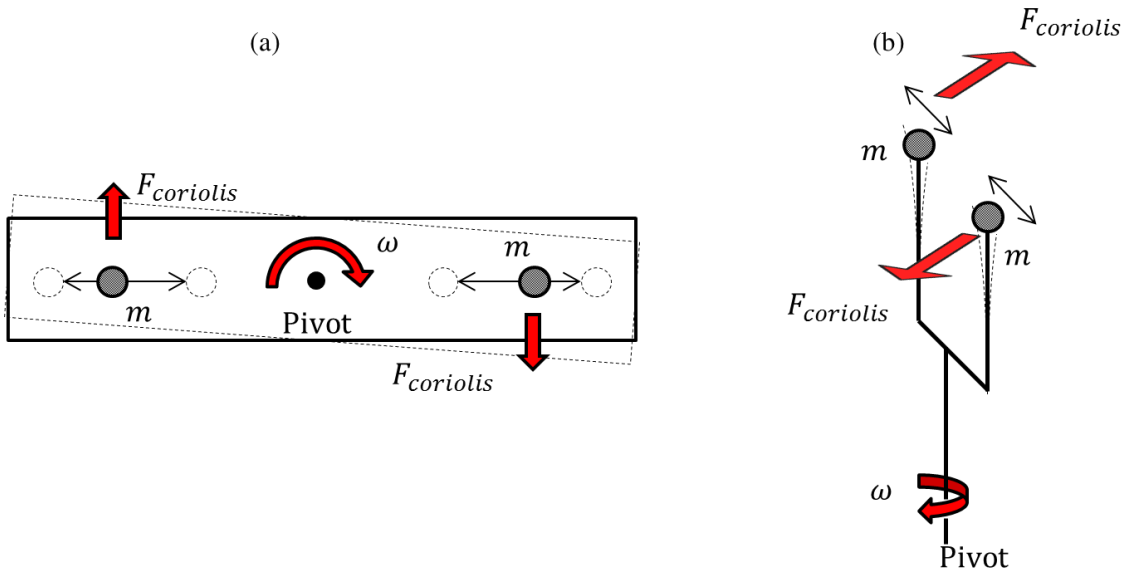


Figure 71 Schematics of the principle of operation of a tuning fork type MEMS gyro: (a) top view and (b) perspective view.

5.5 Strain Gauges

The measurement of strain can be particularly useful in many applications. For example, it can be used to determine critically loaded (stressed) locations in a structure, as illustrated in figure 72.

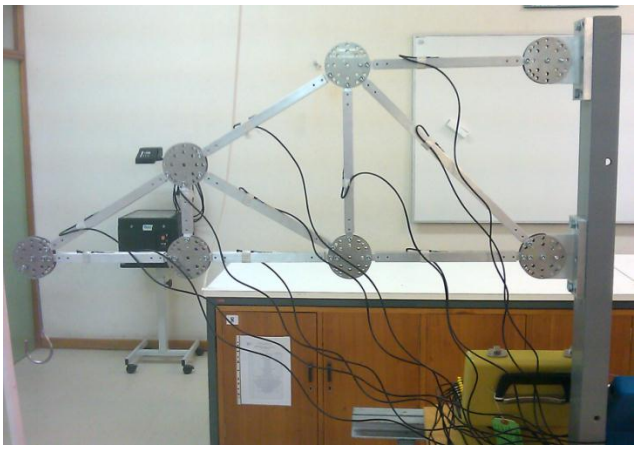


Figure 72 Measuring strains in a scale model of a trussed structure to determine the most critically loaded members.

Another example on the importance of the measurement of strain is in uniaxial tensile testing, where a specimen is stretched (or compressed) along one direction. The measurement of strain provides us with information necessary to determine some of the materials' properties. *Engineering constants* (sometimes known as *technical constants*) are generalized Young's moduli, Poisson's ratios, shear moduli and some other behavioural constants. These constants are measured in simple tests such as uniaxial tensile or pure shear tests [33]. Most simple material characterization tests are performed with a known load or stress. The resulting displacement or strain is then measured.

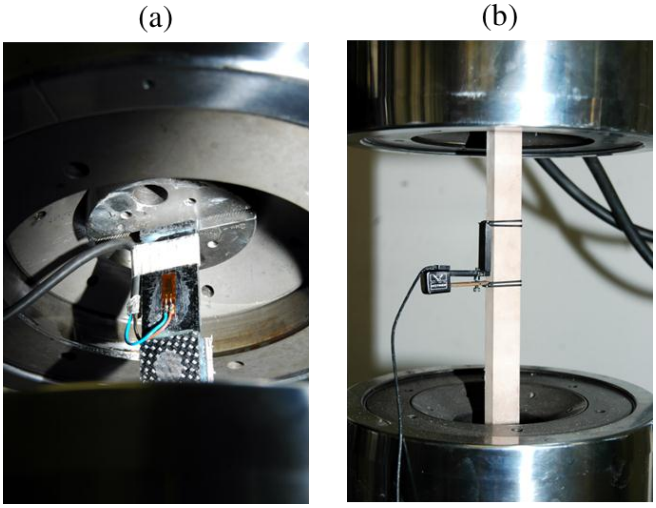


Figure 73 Measuring strain in uniaxial test hydraulic test machines: (a) with strain gauge and (b) with clip-on extensometer.

Strain can be measured directly with a *strain gauge* (figure 73 (a)) or with an *extensometer* (figure 73 (b)). When measuring with an extensometer, strain is determined as the ratio between the displacement $\Delta l = l - l_0$ and initial length l_0 between two marks, i.e.:

$$\varepsilon = \frac{l - l_0}{l_0} = \frac{\Delta l}{l_0} \tag{97}$$

The engineering constants are generally the slope of a stress σ *vs* strain ε curve, i.e., the Young's modulus:

$$\sigma = \frac{F_1}{A} = E\varepsilon \tag{98}$$

or the slope of a strain-strain curve, i.e., the Poisson's ratio:

$$\nu = -\frac{\varepsilon_2}{\varepsilon_1} \tag{99}$$

for $\sigma_1 \neq 0$, all other stresses are zero and $\varepsilon_2 \perp \varepsilon_1$. In these equations, F_1 is the force (measured by a load cell, described in section 5.6, along the specimen's longitudinal direction) and A is the specimen's cross-sectional area at the location where the strain is placed. Figure 74 shows two mutually orthogonal strain gauges, used to determine the Poisson's ratio through equation (99). However, multi-axial strain gauges can be found for this purpose as well: these are called *rosettes*, consisting of two or more strain gauges arranged at different directions (typically $0°$, $45°$ and $90°$). A few examples of strain gauges are shown in figure 75.

Figure 74 A carbon fibre composite specimen, manufactured according to ISO 527-4:1997 (type 2), is instrumented with two orthogonal strain gauges for the measurement of the Young's modulus and Poisson ratio.

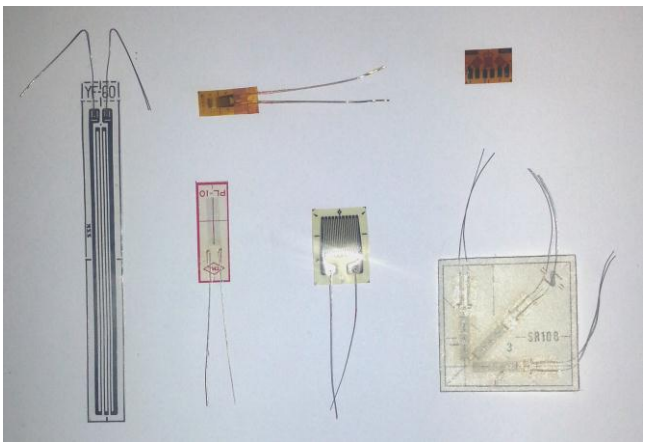


Figure 75 Different strain gauges. The two on the right are called rosettes.

The strain gauge is generally much more accurate than the clip-on extensometer. However, the process of gluing a strain gauge and wiring the leads is very time consuming as it is a work of minutia, requiring expertise. Also, strain gauges respond to temperature changes: in fact, temperature related effects are amongst the major sources of error in strain measurements. This is because most strain gauges are made from metallic alloys, which electrical resistivity

changes with temperature. This important property is in fact taken as an advantage by some temperature sensors (see section 5.7.2).

In terms of principle of operation, a strain gauge is an electrical circuit which resistance changes with strain. The most popular circuit used in strain sensing is called the *Wheatstone* bridge, a network of resistances named after Sir Charles Wheatstone.

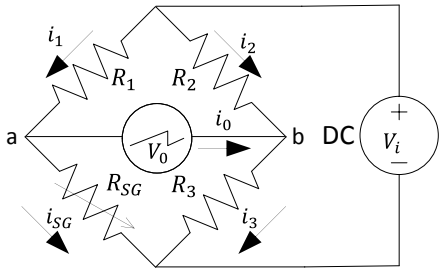


Figure 76 Schematics of the Wheatstone bridge (quarter-bridge).

To better understand how strain gauges work, it is better if this is explained using the quarter-bridge as an example (figure 76), where the resistances R_1 and R_2 are equal and the resistance R_{SG} is variable (representing the strain gauge). The resistance R_3 represents a rheostat, which is usually set at a value equal to the strain gauge resistance when no force is applied. Typically, commercial strain gauges have one from two resistances (when at rest): 350Ω or 120Ω.

When a load is applied, the resistance R_{SG} on the strain gauge changes: it increases when it is stretched and it decreases when it is compressed. With reference to the schematics illustrated in figure 76, it should not be hard to show that the output voltage in a Wheatstone bridge is:

$$V_o = \left(\frac{R_1}{R_{SG} + R_1} - \frac{R_2}{R_3 + R_2}\right) V_i \tag{100}$$

The bridge is is said to be at *balance* when unloaded, i.e., when the current is $i_o = 0$ and the voltage is $V_o = 0$ when measured between points *a* and *b*. Thus, in a balanced bridge:

$$\frac{R_1}{R_{SG}} = \frac{R_2}{R_3} \tag{101}$$

It should be easy to extrapolate these two equations for the half-bridge and full-bridge configurations. In a half-bridge, two legs have variable resistance and in a full bridge all the

legs have variable resistance (figure 77). Opposing legs should be stressed at different directions (when one is being stressed, the opposing one should be being compressed).

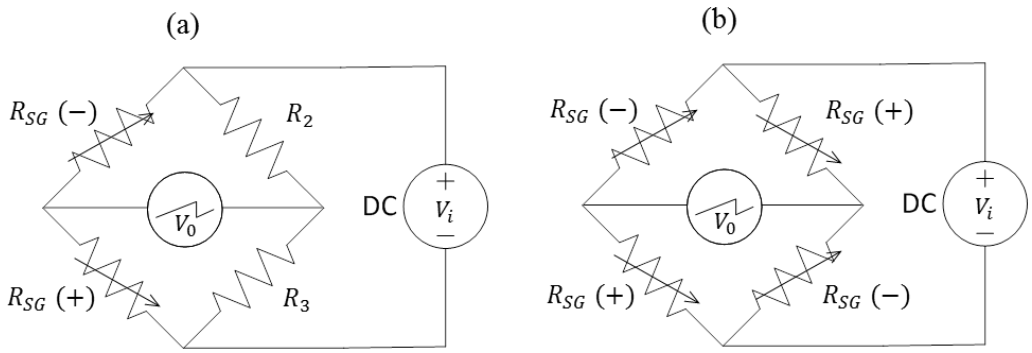


Figure 77 Schematics of the Wheatstone bridge: half-bridge (a) and full-bridge (b). The signals point out that when one member is stretched (+) the other should be compressed (-) for the bridge to be used in strain sensing.

One example on the use of a half-bridge is when measuring a beam subjected to deflection, for example as in the piezoresistive accelerometer shown in figure 61 (section 5.2.2). In a half-bridge configuration, strain gauges are attached at different sides on the beam so that when one side is under traction the opposite side is under compression. This is illustrated in figure 78. With respect to the full-bridge, this is exemplified later on when discussing load cells' design (see section 5.6.2).

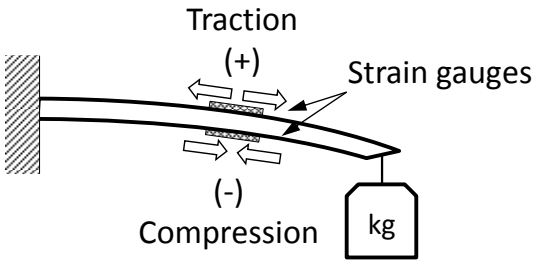


Figure 78 Example of application of the Half-bridge.

Even if both the half-bridge and full-bridge configurations offer better sensitivity, the quarter-bridge is the most used configuration because it is the simplest one in terms of setup. However, the quarter-bridge configuration, as well as the half-bridge, can be highly non-linear, offering a narrower strain measurement range. On the contrary, the full-bridge configuration output voltage V_o is directly proportional to the applied force, as long as the force produces an equal change in resistance in all four legs.

5.6 Load Cells

The same type of technology used in transducers that measure quantities such as displacement, velocity, acceleration, strain, etc., can also be seen in transducers that measure force, i.e., load cells. Available technologies include piezoelectric, capacitance, strain gauge, electromagnetic, tuning fork, etc.

In this section, the two most popular technologies will be discussed: piezoelectric and strain gauge based.

5.6.1 Piezoelectric Force Transducers

Piezoelectric force transducers operate under the same principle as piezoelectric accelerometers, although force transducers do not have a seismic mass. When the piezoelectric crystal is deformed, a charge output, proportional to the rate of change of the force acting on the crystal, is produced. As piezoelectric accelerometers, force transducers can either be of the type charge or IEPE, depending if they bring built-in pre-amplifiers or not. Most piezoelectric force transducers do not measure static (DC) forces.

The way a force transducer is constructed, reminds the way a sandwich is made: a piezoelectric crystal is placed between a base case and a top case. Usually, one side in this 'sandwich' is much lighter than the other. In a conventional setup, the force transducer is placed with the lighter side (called *base side*) towards the structure and the heavy side away from the structure (figure 79 (a)) [6]. This is done in order to avoid as much modification to the structure as possible. The mass of the side that is attached to the structure is called *active* or *live mass*. This is the mass that is 'seen' by the structure at the sensing direction. However, at perpendicular directions, the structure will 'see' the total mass of the force transducer. These masses have been reported to be 3g on the base side for a conventional force transducer weighing a total of 21g, although this varies from model to model [6].

Nevertheless, the force transducer can be attached with the heavier side on the side of the structure (figure 79 (b)). In this case, the discrepancy between the active mass (18g) and the mass 'seen' by the structure in perpendicular directions (21g) will be smaller.

One particular example where force transducers are used is in the measurement of the Frequency Response Function (FRF) of a structure. The FRF contains information about the natural frequencies, modal damping factors and mode shapes of a structure. For a harmonic excitation, the FRF is the relationship between the response and the force. If an accelerometer is used and as long as the excitation is harmonic, the FRF at a given frequency ω is:

$$\alpha(\omega) = \frac{\ddot{x}(t)}{F(t)} \tag{102}$$

where $\ddot{x}(t)$ is the acceleration response to the input force $F(t)$. Since this is a representation of the FRF that makes use of the acceleration, $\alpha(\omega)$ is called *accelerance*. If the FRF was determined from the velocity or displacement responses it would be called *mobility* or *receptance*, respectively.

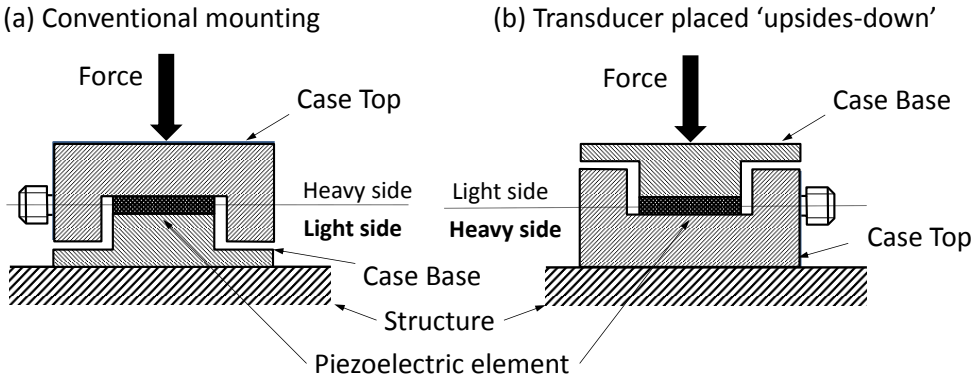


Figure 79 Schematic cross-sectional view of a piezoelectric force transducer: (a) conventional mounting and (b) upsides-down mounting.

Most of the times, it is not practical to measure the FRF, because the excitation force is unknown. However, when in a laboratory environment, it is possible to generate an excitation function, which force is transmitted to the structure using a shaker (figure 80). Typically, these excitation functions are of the type Random, Pseudo-Random, Sweep-Sine, Multi-Sine or Stepped-Sine [34]. When using a synchronous excitation of the type Sweep-Sine, Multi-Sine or Stepped-Sine, a uniform window can be used and no leakage is expected to occur.

Push-rods (also referred in the literature as *stingers* or *drive rods*) are used to apply the excitation force from the shaker to the structure. The objective is to transmit controlled excitation to the structure in a given direction and, at the same time, to impose as little constraint on the structure as possible in all other directions. The locking ball joint fixture allows for simple alignment of the excitation direction, when the structure is not exactly perpendicular to the exciter's axial direction, also minimizing push-rod bending. The whole setup is completely removable and replaceable, thereby avoiding damage to the shaker, structure or transducers, while repositioning is done. One consideration is that it is desirable that the push-rod is as stiff as possible in the longitudinal direction, so that its first mode is above the frequency range of interest, but relatively flexible to lateral and rotational motions between its ends.

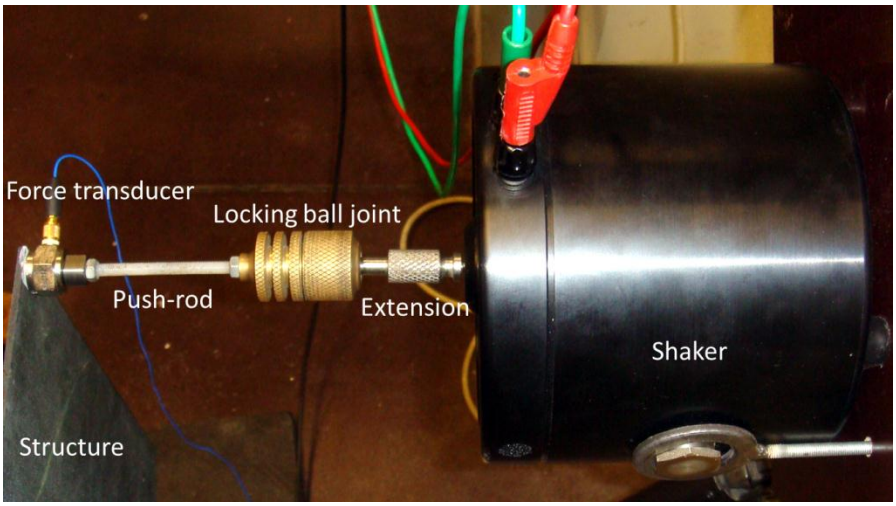


Figure 80 Push-rod connection between a shaker and a force transducer.

When it is not practical (or possible) to use a shaker to excite the structure, a hammer can be used instead. In this case, instead of a FRF, one obtains an IRF, which important results were introduced earlier in section 4.4. The hammer may or may not be instrumented with a force transducer. An example where an instrumented hammer is used to excite a structure and thus measure its natural frequencies is shown in figure 81. The anti-leakage window recommended, in this case, is of the type exponential.

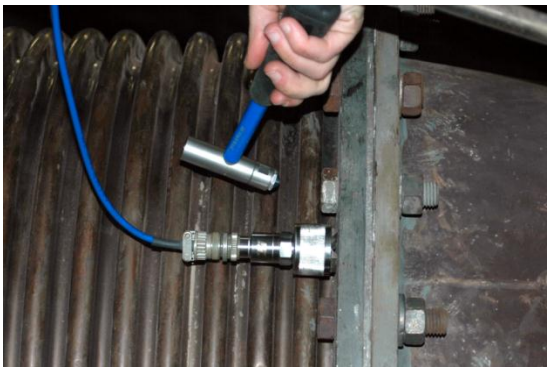


Figure 81 Using a force transducer instrumented hammer and an accelerometer to measure the natural frequencies at the flange of a 700 mm diameter exhaust pipe in a power plant.

5.6.2 Strain Gauge Based Load Cells

Strain gauge based load cells are used in, for example, universal test machines or digital weight scales, among others. Contrary to piezoelectric force transducers, they can usually measure static (DC) forces. Examples of load cells are shown in figure 82.

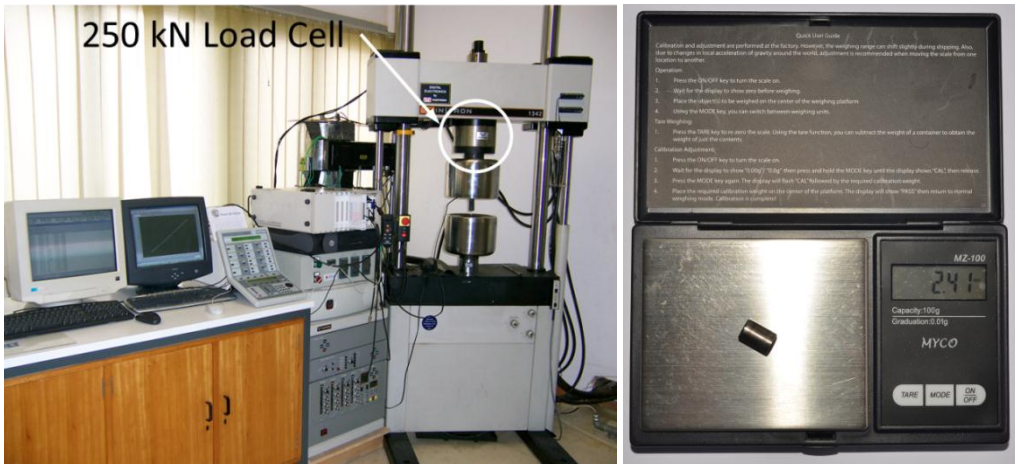


Figure 82 Left: An Instron hydraulic test machine with a 250 kN load cell (photo taken at the Polytechnic Institute of Setubal in Portugal). Right: An inexpensive digital pocket scale with a resolution of 0.01g is being used to measure the weight of a roller bearing cylinder.

Strain gauge based load cells are made from a precision elastic element that has been manufactured with tight tolerances. This element is often made from Aluminum or Stainless Steel. In a strain gauge load cell, four strain gauges are attached at precise locations on the element and connected at a full-bridge Wheatstone configuration (see section 4.4.5 for more information on bridge configurations). A typical binocular type of elastic element used in strain gauge load cells is illustrated in figure 83. In terms of loadings and boundary conditions, it can be modeled as a cantilever beam with an applied load at its tip. A photo of a load cell with a slight different design is shown as an example in figure 84.

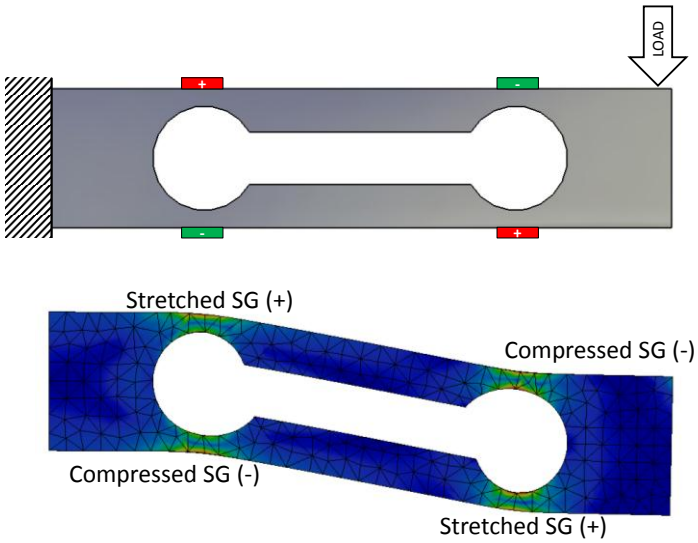


Figure 83 The Finite Element Model above shows that an elastic element used in strain gauge based load cells deforms in such a way it makes it possible to have opposing legs in a full-bridge Wheatstone configuration to be stressed at different directions. In this picture, SG means Strain Gauge.

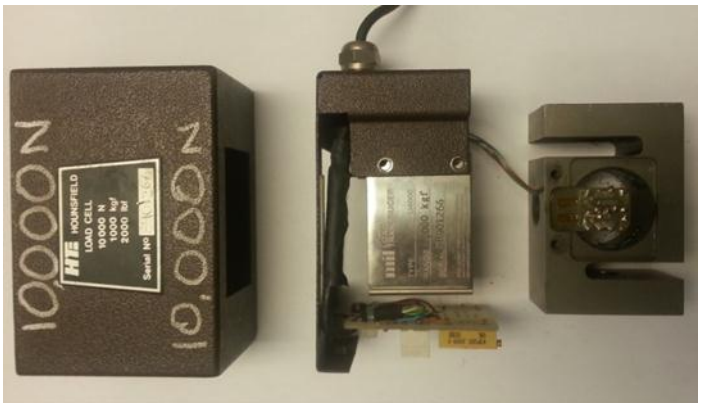


Figure 84 A stripped-down Hounsfield load cell.

5.6.3 Calibration of a Pair Force Transducer *vs* Accelerometer

Let us recall the equation for the frequency response function expressed in terms of accelerance (102). Using the same notation, Newton's second law can be written as:

$$f(t) = m\ddot{x}(t) \tag{103}$$

and the FRF given by equation (102) becomes:

$$H(\omega) = \frac{\ddot{x}(t)}{f(t)} = \frac{1}{m} \tag{104}$$

It is very important to note that this equation (104) is only valid within a limited low-frequency range where the system behaves rigidly. Once the bodies start deforming elastically, it is no longer possible to relate the true value of the mass with the FRF as it is suggested in equation (104). This is also why the inverse of equation (104) is often referred to as *apparent mass*.

Having this into consideration, there is a method that combines the use of an accelerometer with a force transducer (or any other vibration sensor, as long as units are converted to acceleration following the procedure explained in 2.2.3) that allows determining a calibration factor for the pair composed by the force transducer and accelerometer, regardless of their individual calibration values [6]:

$$H_{calibrated}(\omega) = \Theta\frac{\ddot{x}(t)}{f(t)} = \frac{1}{m_{true}} \tag{105}$$

where Θ is the calibration factor to be determined and m_{true} is the total mass of the system composed by a solid block and sensors. The value of m_{true} can be obtained in a weight scale.

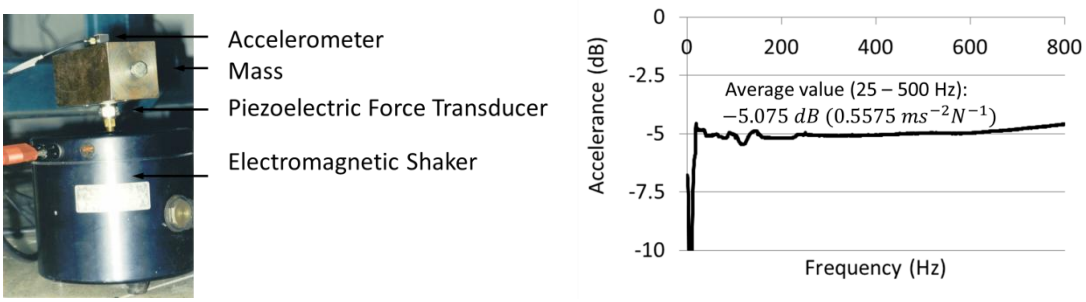


Figure 85 Test-setup to calibrate the pair force transducer vs accelerometer (left) and plot of the accelerance vs frequency so obtained (right).

In the example shown in figure 85, the total mass of the system composed by the accelerometer, force transducer, block, glue and two threaded fixing discs was obtained in a digital weight scale: $m_{true} = 1.8735\ kg$, considering that the force transducer's active mass is 19.73g [26]. The plot of accelerance *vs* frequency shows an approximately flat frequency band ranging from 25 to 500 Hz with an average value of $0.5575\ ms^{-2}N^{-1}$. Application of equation (104) to this result yields $m_{uncalibrated} = 1.7936\ kg$. Thus, the calibration factor is determined to be:

$$\kappa = \frac{m_{uncalibrated}}{m_{true}} = \frac{1.7936}{1.8735} = 0.9574$$

A similar process can be followed to determine the active mass of the force transducer. This has been described in [26].

5.7 Temperature Sensors

Temperature sensors are amongst those which we are most familiar with. These are used in a wide range of different applications: HVAC control, condition and operation monitoring of electronic circuits and machines, food processing, weather forecast, etc.

Besides the conventional thermometer that uses the expansion-contraction of mercury in a glass, there are many other different types of temperature sensors: thermocouples, thermistors, resistance thermometers, bimetallic thermometers and infrared thermometers, among others.

An example of a device that measures both temperature and humidity (a thermo-hygrometer) is shown in figure 86.

Figure 86 Photo of a thermo-hygrometer. The value on top ($16.4\ ℃$*) is the room temperature and the value below (*$7.8\ ℃$*) is the dew point.*

5.7.1 Thermocouple

The thermocouple is composed by a pair of dissimilar metals that perform a junction. When dissimilar metals are connected together, an electrical voltage, proportional to temperature, is produced (figure 87). This thermoelectric effect is called the *Seebeck effect* and is named after Thomas Johann Seebeck.

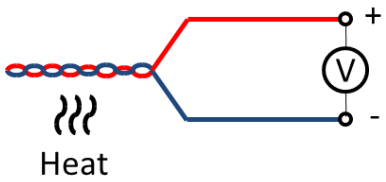


Figure 87 Schematic of the thermocouple principle.

There are different types of thermocouples, made from different alloys. Different alloys determine temperature range, sensitivity and resolution. The most common one is the K-type, a Nickel based thermocouple. In the K thermocouple, one wire is made from Chromel (Nickel-Chromium) and the other wire is made from Alumel (Nickel-Manganese-Aluminum-Silicone). It has a sensitivity of $41\ \mu V/℃$ (approximately) and a working temperature range that generally sits between $-200\ ℃$ and $1250\ ℃$.

Thermocouples are inexpensive sensors that react rapidly to temperature changes. They do not need any source of energy to work. However, the insulation of the wires degrades with time and accuracy is affected. They cannot touch other sources of electricity, because they are electrical conductors.

5.7.2 Thermistors and Resistance Thermometers

Thermistors, as well as resistance thermometers or resistance temperature detectors (RTDs), are resistors which resistance changes with temperature. They can have a positive temperature coefficient or a negative temperature coefficient, depending if the relationship between resistance and temperature is direct or inverse. The name "thermistor" comes from the junction between the words "thermal" and "resistor". The major difference between thermistors and RTDs is the material rather than the principle.

A thermistor is made from a semiconductor, like a polymer, a ceramic or a metallic alloy (e.g., stainless steel). RTDs are made from pure metals (e.g., platinum), which can come in the form of wire or thin-film. RTDs are good for a wider range of temperatures when compared to thermistors, from as low as -200 ℃ to as high as 850 ℃. A thermistor is typically used in more moderate temperature ranges that sit between -90 ℃ and 130 ℃. However, thermistors can often offer better accuracy over their limited working temperature range.

Like a thermocouple, a thermistor is a readily available and inexpensive temperature sensor. However, they do not have such a quick response to temperature changes, although they can deliver more accurate readings.

In a RTD, the change in resistance can be approximated by a linear relationship if the change in temperature is not too large:

$$R = R_0(1 + \alpha T) \tag{106}$$

where R_0 is the initial (or reference) resistance, T is the temperature and α is the temperature coefficient of resistance of the material.

In a thermistor, however, the temperature-resistance relation is far more nonlinear, in part because the temperature coefficient of resistance of the materials used usually is negative. The temperature-resistance relation is given by:

$$R = R_0 e^{\beta\left(\frac{1}{T} - \frac{1}{T_0}\right)} \tag{107}$$

where β, the characteristic temperature (about $4000\ °K$), is temperature dependent itself, thereby adding to the nonlinearity of the device. Hence, proper calibration is essential, especially when measuring in wider ranges of temperatures ($> 50\ °C$).

5.7.3 Bimetallic thermometers

When two thin strips of different metals are bounded together and their temperature change, they will expand according to their coefficients of dilatation. However, because they are coupled together, the shear stresses generated will make the two strips bend together in a curved way (figure 88), so that one strip stretches more than the other.

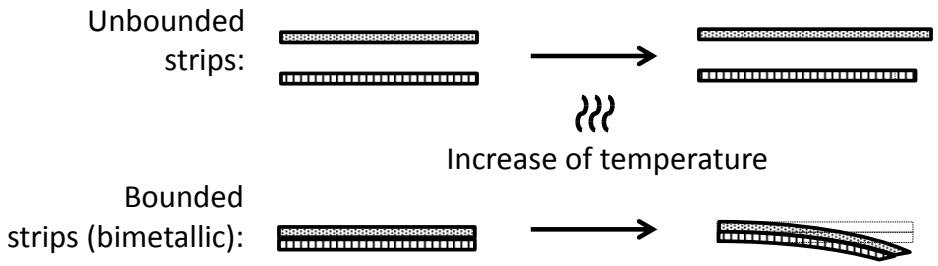


Figure 88 Bimetallic behaviour with temperature increase.

This is the principle behind thermostats to control a room's temperature or dial thermometers used in cooking. In these devices, a bimetallic strip is wrapped into a coil. One end is fixed, while the other end carries a moving electrical contact.

5.7.4 Infrared Sensors

Infrared sensors are spectroscopy contactless sensors that can be used to monitor and map surface temperatures. By being contactless, they do not interfere with the structure, which means they do not change their properties (e.g., stiffness or mass when used on small and lightweight components). They can also be used to measure temperature on moving surfaces, on hazardous locations (e.g., high-voltage) or on locations of difficult access (e.g., that are too high to reach). Their response to temperature changes is very quick, but their sensitivity depends on the emissivity of the materials (the ratio between the energy radiated by the material and the energy radiated by a true black body at the same temperature). They can measure from as low as -70 ℃ to as high as 1000 ℃, although this depends on the device, which usually operate in much narrower bands of temperatures.

Infrared sensors measure the infrared light that is irradiated from the outer surface of an object. Infrared light has longer wavelength (smaller frequency) than visible light and it can range from as low as $0.7\ \mu m$ ($\cong 430\ THz$) to as high as $1\ mm$ ($\cong 300\ GHz$).

ISO 20473:2007 standard defines the following infrared spectral bands:

1. Near-Infrared (NIR): $0.78\ \mu m$ to $3\ \mu m$;
2. Mid-Infrared (MIR): $3\ \mu m$ to $50\ \mu m$;

3. Far-Infrared (FIR): $50\ \mu m$ to $1\ mm$.

The NIR region is also called “reflected infrared” because it requires some source of light (in the same frequency range) to be reflected. Infrared night vision goggles operate in the NIR region. Another example of devices that operate in the NIR region are Passive Infrared (PIR) sensors, like motion detectors used to control an automatic light.

The MIR region is called “thermal infrared”. In this region, a source of illumination is not required. It is within this range that thermography cameras usually operate. Thermography cameras can be used to map temperatures, for example on an engine, on a computer’s CPU, on a fuel cell [35] or in the human’s body (figure 89).

Thermometers that operate at the lower end of the MIR region or in the NIR region are usually called pyrometers, and were originally conceived to detect the temperature of very hot objects (usually incandescent and visible to human eye).

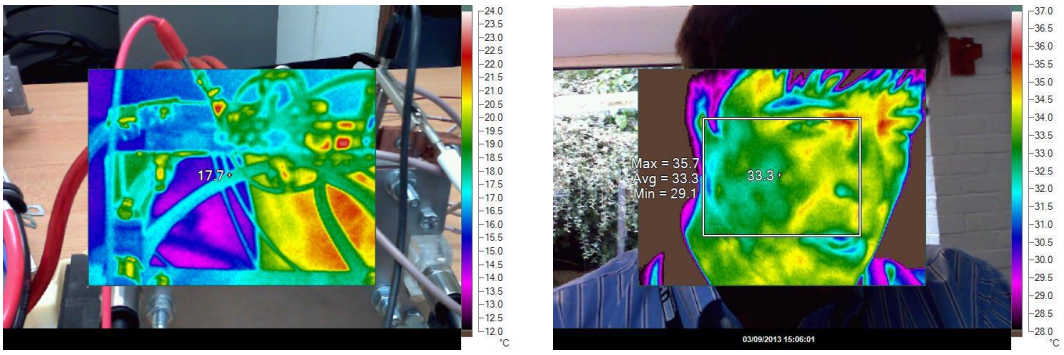


Figure 89 Thermal fluid pattern inside a fuel cell [35] (left) and a picture of this text’s author (right) taken with a Fluke thermography camera Ti25.

The FIR region is used for other applications rather than to measure temperature. Terahertz electromagnetic devices, e.g. FIR lasers or radars, can be used in many applications, for example, in the evaluation of material properties, in NDT (Non-Destructive Testing), in the detection of gas leaks, explosives, chemical or nuclear materials, or in medical imaging.

5.8 Flow Sensors

It can be shown that the flow Q across a constriction of area A obeys the relation [18]:

$$Q = c_d A \sqrt{\frac{2\Delta p}{\rho}} \tag{108}$$

in which c_d is the discharge coefficient for the constriction, Δp is the pressure drop across the constriction and ρ is the density of the fluid. The derived SI units for flow are $m^3 s^{-1}$, as it expresses the volume of moving fluid per unit time.

There are several methods to measure fluid flow. Usually, sensors cannot measure fluid flow directly. Instead, they derive flow from another physical quantity, like pressure or velocity. Methods used to measure fluid flow can be classified into the following categories [18]:

4. Orifice flow meters: pressure is measured across a constriction or opening. Examples include nozzles or Venturi meters;
5. Static-pressure meters: the pressure head is measured, which brings the flow to static conditions. Examples include pitot tubes and rotameters. A Venturi meter can also fit in this category;
6. Flow rate meters: uses a turbine from which a change in the angular momentum is measured;
7. Flow velocity meters: examples include Coriolis meter, LDVs and ultrasonic flow meters.
8. Indirect flow meters: these are meters that measure flow from an effect it creates. Hot wire anemometers or magnetic induction flow meters, are just a few examples.

5.8.1 Venturi tube

The Venturi effect consists in the reduction of fluid pressure that results from flow through a constriction. This effect is exactly the same as is observed in a funnel and is illustrated in figure 90.

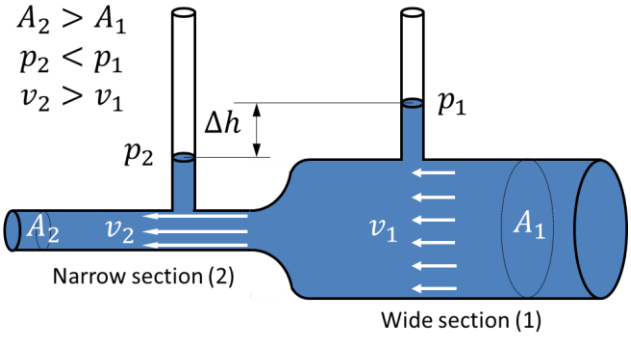


Figure 90 Venturi tube.

Applying Bernoulli's principle, the pressure drop in the Venturi tube is:

$$\Delta p = p_2 - p_1 = \frac{\rho}{2}(v_2^2 - v_1^2) \tag{109}$$

which, once replaced in equation (108), allows measuring the flow rate. The pressure can be measured using the techniques outlined in section 4.4.9.

5.8.2 Pitot Tube

Aircraft use pitot tubes to measure air speed. Pitot tubes only measure flow at a given point of the stream, because the velocity is not uniform across the flow section. The Pitot tube consists of a tube that points directly into the flow at a given location. The moving fluid is brought to rest (stagnates) at the front of the Pitot tube because the tube is already filled in with fluid. This is called the stagnation pressure, total pressure or pitot pressure. In the pitot-static tube (also known as Prandtl tube, shown in figure 91), ports are placed radially to the tube to measure the static pressure.

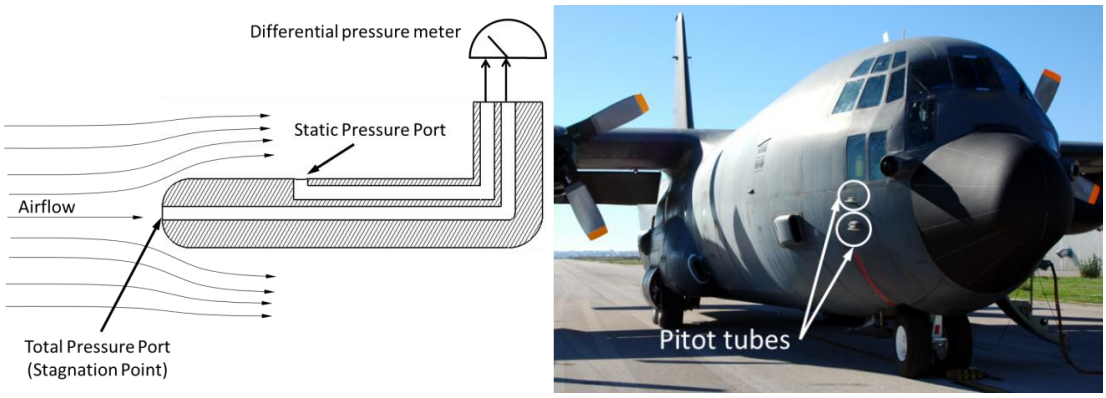


Figure 91 Pitot tube. Left: cross-sectional view illustrating the working principle. Right: pitot tubes on the starboard of a C-130 Lockheed Martin aircraft.

If a fluid is considered to be incompressible (most liquids are incompressible and gases can sometimes be approximated to behave as incompressible as well) and taking into consideration, again, the Bernoulli's principle, the velocity may be determined from:

$$\mathrm{V} = \sqrt{\frac{2(p_T - p_s)}{\rho}} \tag{110}$$

where p_T is the total (stagnation) pressure and p_s is the static pressure.

5.8.3 Anemometers and Angular-Momentum Flow Meters

An anemometer is a device that is used to measure flow speed, rather than flow rate. An example of such a device is shown in figure 92.

Figure 92 A portable anemometer is being used to measure airspeed at the outlet of a dehumidifier.

In an anemometer, the flow passes through a turbine that is made to spin and is supported in low-friction bearings. The flow speed can be determined once the coefficient of power C_p for the turbine is known. As in a wind turbine, the anemometer will produce a a power output P. The flow's velocity v can be determined from:

$$\mathrm{C}_p = \frac{P}{\frac{1}{2}\rho A v^3} \tag{111}$$

where A is the area of the wind turbine.

Another special type of anemometer is the hot-wire (or hot-film) anemometer. This is usually made from a very thin Tungsten or Platinum wire. In the hot-wire anemometer, a conductor carrying a current i is placed in the fluid flow. Thus, the hot-wire is subjected to forced convection. Under steady conditions, the heat loss from the wire into the fluid is exactly balanced by the heat generated by the wire due to its resistance R. Since the coefficient of heat transfer at the boundary of the wire and the moving fluid is known to vary with the square root of the fluid's velocity $\sqrt{v}$, it is possible to relate the velocity of the fluid with the heat balance by:

$$\mathrm{i}^2 R = \left(c_1 + c_2\sqrt{v}\right)\left(T_w - T_f\right) \tag{112}$$

where c_1 and c_2 are constants that can be determined through a least-squares method during calibration (they depend on the geometrical properties of the gauge and physical properties of the medium) [36], T_w is the temperature of the wire and T_f is the temperature of the fluid.

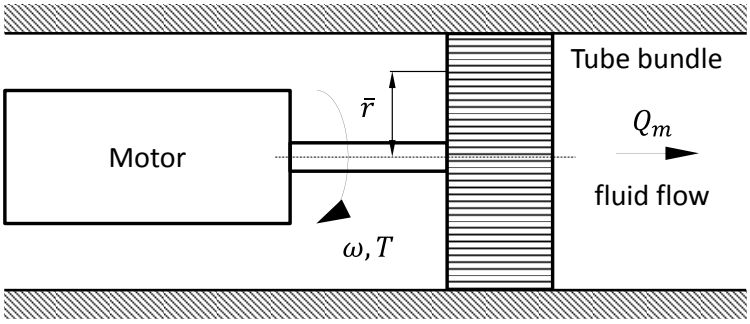


Figure 93 Angular-momentum flow meter.

In the angular-momentum flow rate meter shown in figure 93, a motor is used to govern a tube bundle through which the fluid flows. The motor torque T and angular speed ω are measured. As the fluid mass passes through the tube bundle, it imparts an angular momentum at a rate governed by the mass flow rate Q_m of the fluid. The motor torque provides the torque needed to balance this rate of change of angular momentum. The governing equation is:

$$\mathrm{T} = \omega \bar{r}^2 Q_m \tag{113}$$

where $\bar{r}$ is the radius of gyration of the rotating fluid mass.

5.8.4 Rotameter

One well-known sensor is the rotameter shown in figure 94. Basically, it consists of a cylindrical object floating inside a vertical tube with varying cross section. Typically, both tube and floating object are conic, with the taper growing at the same direction. The weight of the floating object is balanced by the pressure differential on the object. When the flow speed increases, the object rises within the conic tube, thereby allowing more clearance between the object and the tube. The tube is often made from glass and includes a scale, so that the reading can be made directly from the position of the floating object inside the tube.

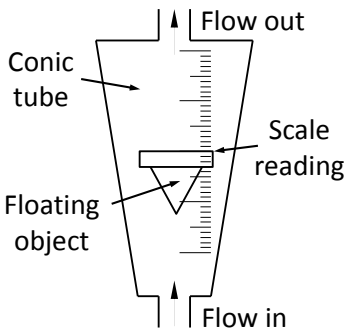


Figure 94 Rotameter.

5.8.5 Other Flow Measurement Sensors

The LDV principle used to measure structural velocity as explained earlier in section 5.3.1 can be applied to fluid flow measurements. Once more, one particular advantage of using laser based technology is that it does not physically interfere with the system, i.e., it does not have an effect on the flow.

Another method of sensing velocity in a flow consists of an ultrasonic burst sent in the direction of the flow. The time of flight, i.e., the time the burst takes to travel through the medium, is measured. The ultrasonic wave propagations' speed is related to the fluid's velocity.

The Coriolis principle can also used to measure flow. In a method, the flow is made to flow through a U-shaped/parabolic tube. Since the parabola does not have constant radius (as a circle does), an acceleration of Coriolis is generated, which can be sensed by displacement sensors from the lateral motion of the tube.

There are many other direct and indirect methods to measure fluid flow. One important thing to note is that, quite often, flow and pressure measurements can be related to one another, so these sensors can sometimes share similar principles of operation.

5.9 Pressure Transducers

As the name suggests, pressure sensors are used to measure pressure, usually exerted by fluids (liquids or gases). They are widely used in aerodynamics, hydraulics and pneumatics applications. The SI unit is Pa (Pascal), which is equal to Nm^{-2} (force per unit area), even though pressure sensor units often come expressed in bar, atm, mmHg (or torr), psi, etc. Some usual pressure units and their conversion to SI units are shown in table 7.

As with thermometers, pressure sensors come in a wide variety of forms and are present in many everyday applications. Depending on the application, pressure sensors can be used not only to measure pressure as an end in itself, but also flow, air speed, water level, depth, altitude or leaks. Another example of application is a binary pressure switch, like a computer mouse button, a door bell or speed detectors in some traffic enforcement cameras.

Almost all pressure sensors are differential, in the sense they measure pressure against a reference value, whether it is vacuum pressure, atmospheric pressure or other reference. An example of a differential pressure gauge is shown in figure 95.

The technology used to measure pressure can be very similar to the one used in force transducers and load cells. Usually, a membrane, a spring, belows, a bourdon tube or a piston are used to measure deflection due to an applied force over a known area. The point is that –

most of the times - the deformation of a sensing element can be converted into a reading of the exerted force, and, once the area is known, into pressure.

Table 7 Pressure units converted to SI units ($1\ Pa = 1\ Nm^{-2} = 1\ kgm^{-1}s^{-2}$*).*

Name	Symbol	Conversion to SI units
Technical Atmosphere	at	$1\ at = 0.98067\ \times 10^5\ Pa$
Standard Atmosphere	atm	$1\ atm\ =\ 1.01325 \times 10^5\ Pa$
Barometric pressure	bar	$1\ bar\ =\ 1 \times 10^5\ Pa\ (\cong 1\ atm)$
Torricelli	Torr	$1\ torr = \frac{1}{760} atm = 133.3\ Pa$
Height of column of Mercury	mmHg	$1\ mmHg \cong 1\ torr\ = 133.3 Pa$
Height of column of Water	mmH_2O	$1\ mmH_2O = 9.8067\ Pa$
	inH_2O	$1\ inH_2O = 248.84\ Pa$
Pounds per square inch	psi	$1\ psi = \frac{1\ lbf}{1\ in^2} = 6894.8\ Pa$
Hectopascal / milibar	hPa / mbar	$1\ hPa = 1\ mbar = 1 \times 10^2\ Pa$
Meter sea water	msw	$1\ msw = 0.1\ bar = 1 \times 10^4\ Pa$
Foot sea water	fsw	$1\ fsw =\ 3.0643 \times 10^3\ Pa$

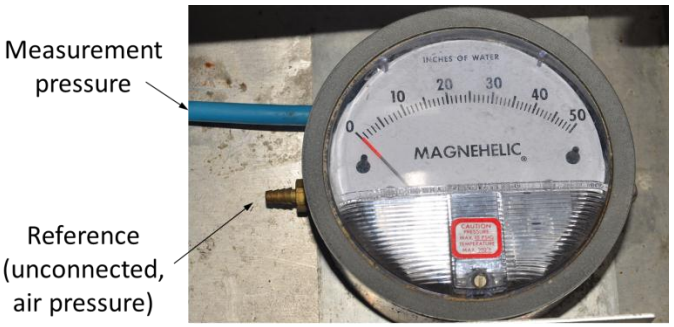


Figure 95 A differential pressure gauge.

Pressure sensors can be piezoelectric, piezoresistive, capacitive, strain sensitive, potentiometric, electromagnetic, inductive, thermal, flow sensitive, among others. Most of these technologies have been discussed in the previous sections, even if they were exemplified under the context of different applications.

For example, it should be already easy to understand that piezoelectric pressure sensors are better suited for dynamic measurements (high speed changes in pressure) as it happens inside the combustion chamber in an engine. On the contrary, inductive pressure sensors are not as suitable to quick changes in pressure, but can be very sensitive and accurate. A flow sensitive pressure sensor (like a Pitot tube) is better suited to measure airspeed in an aircraft.

5.10 Ultrasonic Sensors

Ultrasonic sensors use pressure waves that are transmitted and reflected at frequencies above the human audible range. Ultrasonic sensors are used in many applications, including medical imaging, non-destructive testing, flow measurements, sonars, burglar alarms and car parking sensors. Ultrasonic equipment can use frequencies as low as 20 kHz and can go up to as high as the GHz region ($1\ GHz = 1 \times 10^9\ Hz$).

Ultrasound can be generated according to several principles, for example, from very high frequency oscillations on a piezoelectric element subjected to an electrical potential, or from the use of the magnetostrictive property of ferromagnetic materials when subjected to oscillating magnetic fields.

In essence, ultrasonic sensors project an ultrasonic burst towards a target object, and the time taken for the echoes to be received are clocked. A signal processor, which is calibrated according to the speed of sound in the medium where the sound propagates, determines the position of the echoes with respect to the probe. In other words, position can be determined from:

$$x = \frac{ct}{2} \tag{114}$$

where c is the speed of sound and t is the time of flight (from generator, then to target, and finally to receiver).

As an example, a C-Scan used to map voids or delamination defects throughout the thickness of composite laminates is illustrated in figure 96. In this picture, the times of flight t_1, t_2, t_3 and t_4 are relative to the echoes that happen every time there is a transition: t_1 is the echo at the entrance of the laminate, t_2 is the echo at the defect, t_3 is the echo at the exit of the laminate and t_4 is the echo at the bottom of the tub.

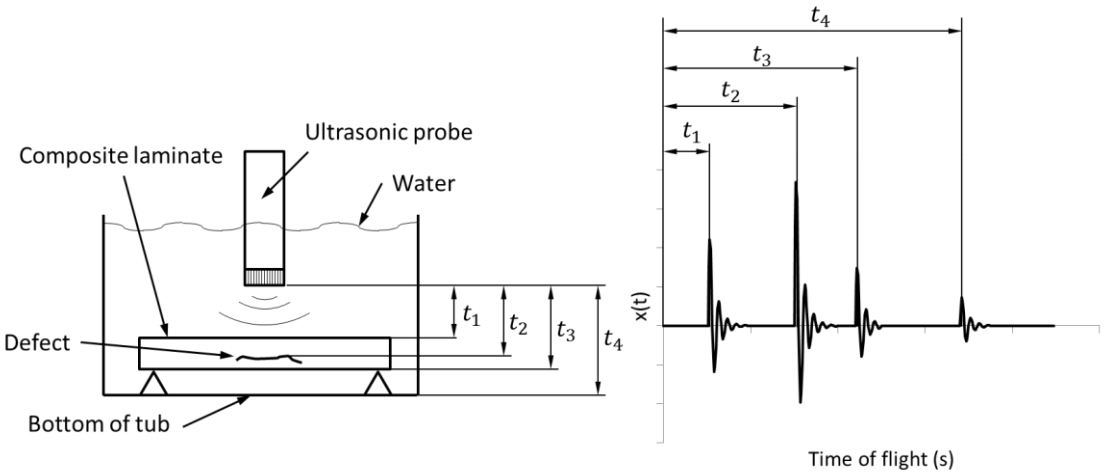


Figure 96 Schematics of the principle of operation of a C-Scan used for mapping damage in composite laminates (left) and echoes seen in the time domain (right).

Alternatively, the velocity of the target (if moving) can be measured using the Doppler effect, a principle that has been described before in section 5.3.1, although using a different sensor as an example.

5.11 Encoders

Encoders are a specific class of digital transducers that can monitor motion or position, from a coded (digital) reading of a measurement. In general, encoders are found in the form of discs that display a code pattern. These can be divided into two classes: incremental encoders (or relative encoders) and absolute encoders.

Encoders are used in a wide variety of applications, because they can be used as feedback devices for speed or position control. Applications include CNC machines, automatic welding in an assembly line, elevators, bar code readers, radars, wind turbines, conveyors, printers, robotics, etc.

5.11.1 Incremental Encoders

The electromagnetic pulse tachometer shown in figure 65 (section 5.3.2) is an example of a simple type of an incremental encoder: when the ferromagnetic screw passes in front of the sensor during shaft rotation, it counts a pulse (1), whereas it does not pick up any signal (0) when the screw is at any other position. When the shaft completes a full rotation since the last pulse, it will count another pulse (1). The time elapsed between these two pulses is the period, which inverse is the frequency in cycles per second (Hz). This can be seen as a type of an incremental encoder.

There are two possible additional configurations for incremental encoders: (1) offset sensor configuration and (2) offset track configuration. When any of these two configurations is used, it is possible to determine both direction (clockwise or anticlockwise) and speed.

Figure 97 (a) shows the first configuration. The disk has a single circular track with identical and equally spaced transparent windows. Two photodiode sensors (pick-offs) are positioned facing the track at a quarter-pitch ($90°$ or quadrature, i.e., half the window length) apart. Assuming the disk is rotating at the anticlockwise direction at the position shown in figure 97 (a), the sensor pick-off 1 is changing state (from 1 to 0) at the edge between the transparent and opaque windows, while sensor pick-off 2 keeps its value unchanged. On the contrary, if the disk is rotating at the clockwise direction, the sensor pick-off 1 will change state from 0 to 1 instead. Encoders must be programmed so that they are able of edge detection, as it is based on the sequence of edge detection that the direction can be determined.

The second possible configuration is shown in figure 97 (b). In this case, the disk has two identical tracks, one offset from the other by a quarter pitch. The two pick-off sensors are aligned along the same radial line unlike the previous configuration. The principle is exactly the same as for the first configuration.

It is not unusual to add an additional track with a single window and associated sensor. This generates a reference pulse (called index pulse) per revolution of the disk, exactly the same way as in the tachometer example shown in figure 65 (section 5.3.2). This index is used as a counter of the complete number of revolutions. When the disk rotates at constant angular speed, the pulse width and pulse-to-pulse period are constant with respect to time. When the disk is accelerating, the pulse width decreases continuously, and vice versa.

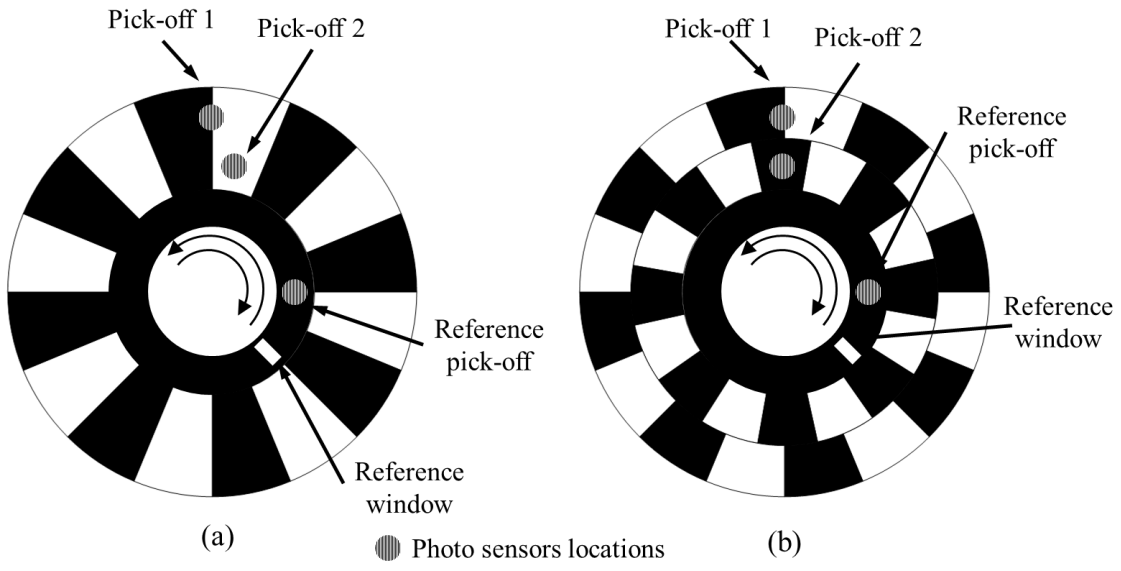


Figure 97 Incremental encoders: (a) one-track wheel with offset sensors and (b) two-track quadrature wheel.

5.11.2 Absolute Encoders

An absolute encoder has many pulse tracks on its transducer disk. The pulses constitute words, written using a number of binary digits (0 or 1) equal to the number of tracks. Typically, the disk has transparent windows for a binary digit of 1 and black opaque windows for a binary digit of 0, exactly the same way as in incremental encoders. A typical encoder uses optical sensors that are capable of differentiating between these windows. Absolute encoders can have different tracks, as well as several pulses per track. When the disk of an absolute encoder rotates, several pulse trains – equal in number to the tracks on the disk – are generated simultaneously. Hence, absolute encoders have a system of coded tracks where no two positions are identical. They also have memory, as they do not lose position after power is switched off.

A simplified code pattern on the disk of an absolute encoder, which uses the direct binary code, is shown in figure 98 (a). The number of tracks n in this case is 4, but in practice n is in the order of 14 and may even be as high as 22 [18]. The disk is divided into 2^n sectors, or bits of data. A set of n pick-off sensors is arranged along a radial line and facing the tracks on one side of the disk, opposite to a light source that illuminates the other side of the disk. As the disk rotates, at a given instant, the pick-off sensors will generate a combination of signals (coded data word) that uniquely determines the position of the disk at that time.

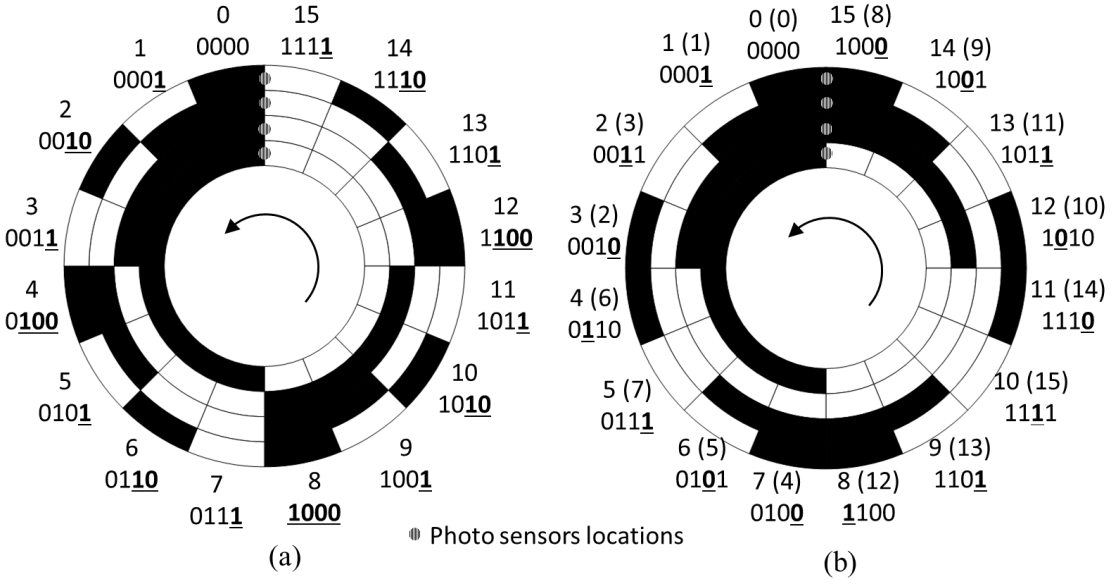


Figure 98 4-bit absolute encoders (16 state): (a) binary code and (b) gray code. Bold underlined digits represent a change from the previous state. The numbers between brackets () on the gray code represent the position of the word on the binary code.

Straight binary code in absolute encoders can lead to a data interpretation problem. Note that, in figure 98 (a), the transition from one sector to an adjacent one may require more than one switching of bits in the binary array of data. This is highlighted with the underlined bold digits that changed from one sector to the other, assuming the disk is rotating at the anticlockwise direction. For example, the transition from sector 7 (0111) to sector 8 (1000) requires four bit switching, the transition from sector 9 (1001) to sector 10 (1010) requires three bit switching, and the transition from sector 11 (1011) to sector 12 (1100) requires two bit switching. If the pick-off sensors are not properly aligned, if the manufacturing tolerances were low, if environmental effects have resulted in irregularities in the sector matrix or if there is a significant increase in vibration level due to the run-out or improper maintenance, then it might happen that bit switching from one reading to the next will not take place simultaneously. This will result in ambiguity in data reading during the transition period.

In order to avoid misinterpretation on the data reading, the sectors in the disk are re-displaced in such a way that only one bit is changed from one reading to the next. This is called a *gray code*, and is shown in figure 98 (b). Nevertheless, for an absolute encoder, the gray code is not absolutely essential, as the sequence of arrays in the matrix is known beforehand. When there is a data switch, it is possible to check it against the two valid possibilities (or a single one, if the direction of rotation is known).

5.12 Other Sensors

It would be very difficult to describe in detail all the existing sensors, technologies and applications. This section serves only to give a flavour on the topic. More in-depth can be found, for instance, in [18], or even by searching for sensor suppliers and manufacturers over the internet. Sensors that we missed include (but not only): potentiometers, effort sensors, torque sensors, tactile sensors, gyroscopic sensors, fiber-optics, microphones, acoustic emission, chemical sensors, camera-based video systems, etc. Several areas were identified where new developments and innovations are made in sensor technology. These include [18]:

1. Micro- and nanominiature (MEMS) sensors;
2. Intelligent sensors, with built-in processing capabilities for decision making;
3. Embedded and distributed sensor networks, e.g., built-in arrays of fiber-Bragg gratings in composite structures [37];
4. Hierarchical sensory architectures, where low-level sensory information is processed to match higher level requirements.

Chapter 6

Logarithmic Scales

6.1 The Decibel

The decibel (dB) is a logarithmic scale that is used in many fields, from electronics to vibration and acoustics. It is used to better represent the ratio between amplitude and a reference value in the *y*-axis that would not be as conveniently expressed as when using linear units.

The decibel owes its name to Alexander Graham Bell, the founder father of telecommunications. As the name suggests, the decibel is a tenth of a bel. One bel represents the ratio between two power quantities of 10:1.

For example, voltage, current, sound pressure and velocity are quantities which square is proportional to power. These quantities are called *root-power* quantities. On the other hand, sound intensity, luminous intensity and energy density are quantities that are directly proportional to power. These quantities are called *power* quantities [38].

An example of an equipment that delivers results in dB is the sound level meter (figure 99). Basically, it consists of a spectrum analyser that uses a microphone as sensor.

6.1.1 Power Quantities

Power quantities (also known as *intensity quantities*) are directly proportional to power. The dB level of the ratio between an intensity I and a reference value I_0 is given by:

$$L_I = 10 \log_{10} \left(\frac{I}{I_0} \right) \tag{115}$$

Rearranging this formula, one can determine the Intensity magnitude from:

$$I = 10^{\frac{L_I}{10}} I_0 \tag{116}$$

The reference value is the value for which the dB is zero. For example, in acoustics, the reference value for the acoustic intensity is $I_0 \cong 10^{-12}\ W/m^2$. This value corresponds to what is considered to be the hearing threshold for an average healthy human adult. When a unit is expressed in dB it should ideally be represented with its value accompanied by the reference value and its units, for example $1\ dB\ re\ 10^{-12}\ W/m^2$, although in many occasions this is omitted. When the reference value is not shown, it is assumed that it is unitary.

Figure 99 A Casella digital sound level meter is used to measure the sound pressure level from an acoustic source inside an anechoic chamber.

To better understand how operations using the dB scale work, let us make a sum between equal power quantities. As an example, let us assume we are making the following sum:

$$70\ dB + 70\ dB = ?$$

We cannot sum dB's as we do with quantities that are expressed in a linear scale. In other words, to say the result is $140\ dB$ would be a wrong answer and quite far from the correct one. Firstly, we must determine the value of the intensity I from equation (116), which, in this case is:

$$I = 10^{\frac{70}{10}} I_0 = 10^7 I_0$$

Now that we have the value of I, we can add them together, obtaining:

$$I_{sum} = I + I = 10^7 I_0 + 10^7 I_0 = 2 \times 10^7 I_0$$

Replacing this value in equation (115), one obtains:

$$L_I = 10 \log_{10} \left(\frac{2 \times 10^7 I_0}{I_0} \right) \cong 73\ dB$$

In conclusion, when adding two equal $70\ dB$ power quantities, one obtains:

$$70\ dB + 70\ dB \cong\ 73\ dB$$

Thus, when adding power or intensity quantities, it is easy to prove that the double of the intensity when expressed in decibel is going to be $3\ dB$ larger. One interesting example is the sum of $0\ dB + 0\ dB$, which adds up to $3\ dB$ for power quantities. This happens because $0\ dB$ corresponds to the reference value. For example, in acoustics, the acoustic intensity at $0\ dB$ is $10^{-12}\ W/m^2$, which means that sound already exists. $0\ dB$ does not mean that no sound is being produced; it is the threshold for human hearing. The sound intensity level of absolute silent would be $L_I \to -\infty\ dB$.

Similarly, a change in the power ratio by a power of 10 is a 10 dB change.

6.1.2 Root-power Quantities

Root-power quantities (also known as *amplitude quantities*), like voltage or sound pressure, are quantities which square is proportional to power. So, it is usual to consider the ratio between the square of the amplitude A and the square of a reference amplitude A_0. The dB level of the ratio between A^2 and a reference value A_0^2 is given by:

$$L_{RP} = 10 \log_{10} \left(\frac{A^2}{A_0^2} \right) = 20 \log_{10} \left(\frac{A}{A_0} \right) \qquad (117)$$

Rearranging this formula, one can determine the value for the amplitude magnitude from:

$$A = 10^{\frac{L_{RP}}{20}} A_0 \tag{118}$$

Let us now make a sum between equal root-power quantities. As an example, let us assume we are making the following sum:

$$70\ dB + 70\ dB = ?$$

Firstly, we must determine the value of the amplitude A from equation (118), which, in this case is:

$$A = 10^{\frac{70}{20}} A_0 = 3162.3 A_0$$

Now that we have the value of A, we can add them together. But, because these are root-power quantities, the sum must be done the following way:

$$A_{sum} = A + A = 3162.3 A_0 + 3162.3 A_0 = 6324.6 A_0$$

Replacing this value in equation (117), one obtains:

$$L_{RP} = 20 \log_{10}\left(\frac{6324.6 A_0}{A_0}\right) \cong 76\ dB$$

In conclusion, when adding two equal $70\ dB$ root-power quantities, one obtains:

$$70\ dB + 70\ dB \cong\ 76\ dB$$

Thus, when adding root-power or amplitude quantities, it is easy to prove that the double of the amplitude when expressed in decibel is going to be $6\ dB$ larger. One interesting example is the sum of $0\ dB + 0\ dB$, which adds up to $6\ dB$ for root-power quantities.

Similarly, a change in the root-power ratio by a power of 10 is a 20 dB change.

6.1.3 Linear *vs* Logarithmic Frequency Plots

The representation of the magnitude of a function *vs* frequency is many times done in dB. For example, let us assume the frequency spectrum of the RMS value of the acceleration of a faulty impeller is plotted. When a linear scale is used, the spectrum will look like the one shown in figure 100 (a). This representation highlights the larger peaks in the spectrum. When a logarithmic (usually expressed in a decibel scale) is used, the spectrum will look like the one shown in figure 100 (b). The dB scale is useful to visualize both the smaller and larger

amplitudes in the same plot. It is clear in this example that there may be some important peaks between 50 and 150 Hz that are hardly visible using a linear scale, whereas using a dB scale they become quite evident. As a disadvantage, the dB scale 'amplifies' noise, making it harder to determine whether a peak is just from noise or if it actually is a pattern on the signal.

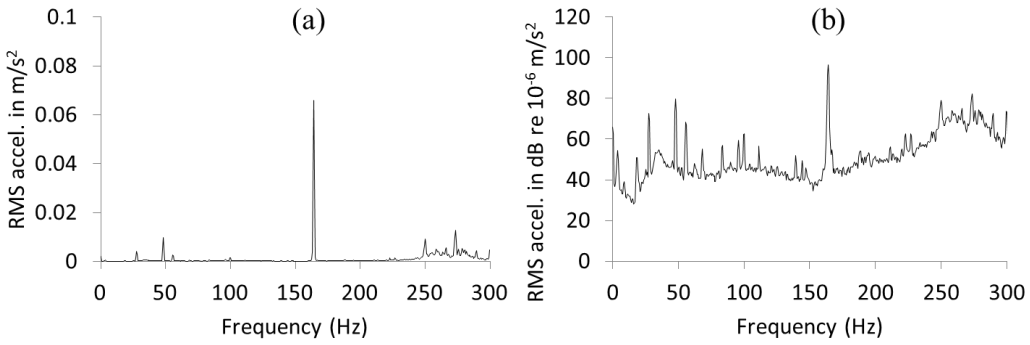


Figure 100 RMS acceleration spectrum obtained on a impeller: (a) linear scale and (b) decibel (logarithmic) scale.

6.1.4 dB Reference Values

Examples of typical reference values for quantities that are represented in a dB scale are shown in table 8, along with their symbols. However, other reference values may be found. Reference levels have an impact in the dB signal: when the measured signal is larger than the reference value, the dB level is positive; when the measured signal is smaller than the reference value, the dB level is negative.

6.1.5 Comparison between the Power and Root-Power dB Scales

A comparison between the dB power and root-power ratio scales is shown in table 9.

Table 8 Typical reference values for quantities that are represented in a dB scale.

Quantity name	**Type**	**Reference value**	**Units**	**Symbol**
Sound Intensity Level	Intensity	10^{-12}	W/m^2	dB SIL
Sound Pressure Level	Amplitude	20×10^{-6}	Pa	dB SPL
Acceleration	Amplitude	10^{-6}	ms^{-2}	dB re $10^{-6}\ ms^{-2}$
Acceleration	Amplitude	1	ms^{-2}	dB ms^{-2}
Velocity	Amplitude	10^{-9}	ms^{-1}	dB re $10^{-9}\ ms^{-1}$
Velocity	Amplitude	1	ms^{-1}	dB ms^{-1}
Displacement	Amplitude	10^{-12}	m	dB re $10^{-12}\ m$
Displacement	Amplitude	1	m	dB m
Power	Intensity	1	W	dBW
Power	Intensity	1×10^{-3}	W	dBm
Voltage	Amplitude	1	V	dBV
Voltage	Amplitude	1×10^{-3}	V	dBmV
Voltage	Amplitude	1×10^{-6}	V	dBuV
Voltage unloaded (Audio)	Amplitude	0.7746	V	dBu
Generic quantity	-	1	$unit$	dB $unit$

Table 9 dB power and root-power ratio scales between 120 and -120 dB.

dB Level	Power quantity (intensity)	Root-power quantity (amplitude)
120	1000000000000 (10^{12})	1000000 (10^{6})
100	10000000000 (10^{10})	100000 (10^{5})
80	100000000 (10^{8})	10000 (10^{4})
60	1000000 (10^{6})	1000 (10^{3})
40	10000 (10^{4})	100 (10^{2})
20	100 (10^{2})	**10**
10	**10**	3.162
6	3.981	**1.995 (~2)**
3	**1.995 (~2)**	1.413
0	1	1
-3	**0.501 (~1/2**)	0.7079
-6	0.251	**0.501 (~1/2)**
-10	**0.1 (10^{-1})**	0.3162
-20	0.01 (10^{-2})	**0.1 (10^{-1})**
-40	0.0001 (10^{-4})	0.01 (10^{-2})
-60	0.000001 (10^{-6})	0.001 (10^{-3})
-80	0.00000001 (10^{-8})	0.0001 (10^{-4})
-100	0.0000000001 (10^{-10})	0.00001 (10^{-5})
-120	0.000000000001 (10^{-12})	0.000001 (10^{-6})

6.2 Octave

The octave is a representation on the frequency *x*-axis widely used in acoustics. One octave is defined as the 2:1 ratio between two frequencies. For example, in music, the middle C (C4) has a frequency of 261 Hz. This means that the upper C (C5), which is said to be “one octave above”, will have a frequency of 522 Hz. This octave bandwidth – the difference between the upper and lower frequencies - is 261 Hz. On the other hand, the lower C (C3), which is said to be “one octave below”, will have a frequency of 131 Hz. In this case, the octave bandwidth is 131 Hz. The scale is not linear, because the difference in frequency between the C5 and C4 is twice as much as the difference in frequency between the C4 and C3.

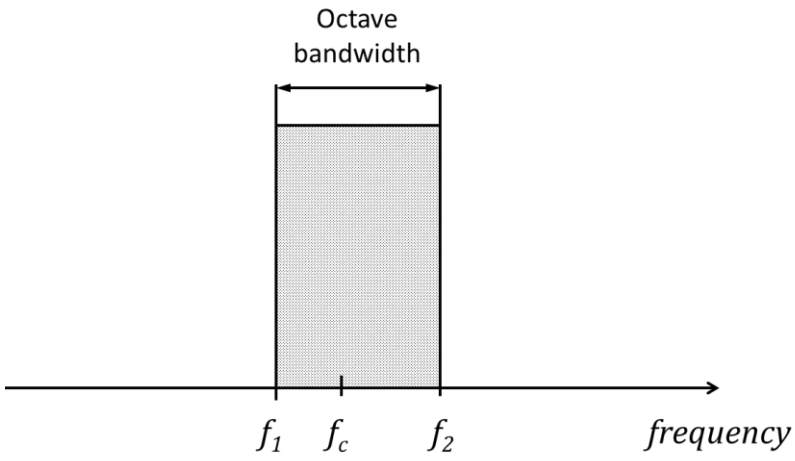


Figure 101 Octave bandwidth, lower frequency edge, upper frequency edge and central frequency.

The octave can also be used in many other applications besides acoustics, for example in filter design. Historically, bandwidth analysis appeared in any frequency analysis application because it was not computationally practical to use the DFT. *Constant percent bandwidth filters* can be used instead of the Fourier Transform. These consist of analogue or digital pass-band filters applied directly over the time signal. However, instead of getting spectral lines, one gets bands, which are frequency intervals.

Figure 101 shows the bandwidth of one octave, where f_1 is the lower frequency edge, f_2 is the upper frequency edge and f_c is the central frequency. The relationships between these quantities are given by:

$$\frac{f_2}{f_1} = 2^n \tag{119}$$

$$f_c = \sqrt{f_2 \cdot f_1} \tag{120}$$

$$\%B = \frac{f_2 - f_1}{f_c} \times 100\% \tag{4.121}$$

where n is the number of octaves and $\%B$ is the percent bandwidth. Typically, the number of octaves is a fraction: 1 octave, 1/3 octave or 1/12 octave, although there are many other possible fractions.

As an example, let us assume we want to determine the high frequency edge of a band 7 octave wide. Equation (119) can be used to determine it immediately:

$$\frac{f_2}{10} = 2^7 \implies f_2 = 1280\ Hz$$

This same result is illustrated in figure 102 for a better understanding. In this picture, it can be seen that each octave's upper frequency is twice as much as the lower frequency and that a total of 7 bands are depicted.

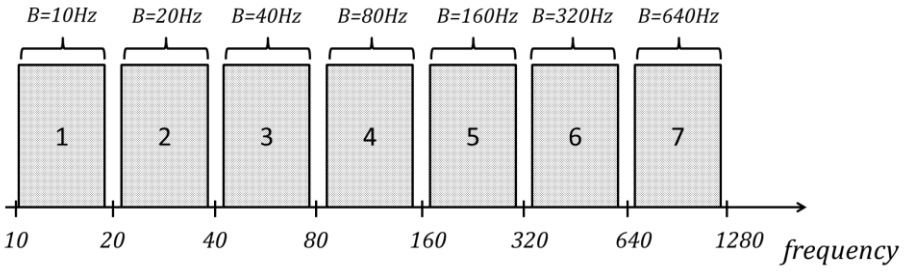


Figure 102: Lower and upper frequency edges of a band 7 octave wide, starting at 10 Hz.

Figure 102 also shows each octave's bandwidth, which increase with frequency. However, when the percent bandwidth is determined from equations (120) and (4.121), the result is found to be constant. One octave has a bandwidth equal to 70.7% of the center frequency, 1/3 octave has a bandwidth equal to 23.2% of the center frequency and 1/12 octave has a bandwidth equal to 5.78% of the center frequency.

For fractional octaves, the analysis gets a bit more complicated. Let us assume we want to determine the frequencies of the upper edges in one octave bandwidth if 125 Hz is the lower frequency of a 1/3 octave band. As it should now be obvious, in one octave band with lower frequency edge at 125 Hz the upper frequency edge must be 250 Hz. To determine the intermediate frequencies, one uses equation (119) consecutively:

$$\frac{f_2}{125} = 2^{\frac{1}{3}} \implies f_2 = 157.5\ Hz$$

$$\frac{f_2}{157.5} = 2^{\frac{1}{3}} \implies f_2 = 198.4\ Hz$$

$$\frac{f_2}{198.4} = 2^{\frac{1}{3}} \implies f_2 = 250\ Hz$$

This same result is illustrated in figure 103 for better understanding.

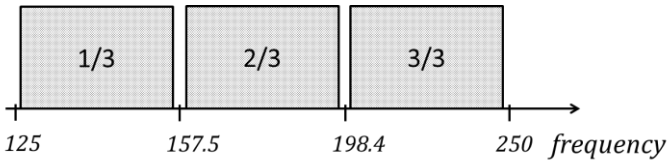


Figure 103 Lower and upper frequency edges of 1/3 octave bands in one octave bandwidth, starting at 125 Hz.

Chapter 7

Final Remarks

The fundamentals on sensors and signal processing were presented in the previous chapters. It is expected that, after reading this text, the Engineer is given the necessary tools that will enable him to implement data acquisition systems avoiding common pitfalls in signal processing. Furthermore, a framework has been set so that it serves as a stepping stone for further research to the more interested reader.

There are many other sensors and signal processing techniques besides those introduced in the previous sections. It is impossible to describe them all, mainly due to space limitations. The ones selected were considered to be those that, besides being widely used in Mechatronics applications, would be capable of illustrating most of the up-to-date techniques used today. For example, the piezoelectric effect, not being exactly novel, is used in a variety of sensing techniques, in the measurement of force, pressure or acceleration, or to generate ultrasonic bursts. In another application, piezoelectric elements are used to harvest energy, a topic that has deserved a growing interest in the past recent years.

References

1. Acoustics, National Physical Laboratory (NPL), Available at: http://www.npl.co.uk/educate-explore/factsheets/acoustics/.
2. Giancoli, D. C., *Physics: Principles with Applications*, Prentice-Hall, Upper Saddle River, New Jersey, USA, 1998.
3. Priemer, R., *Introductory Signal Processing (Advanced Series in Electric and Computer Engineering – Vol. 6)*, World Scientific Publishing, Singapore, 1991.
4. Priestley, M. B., *Non-linear and Non-stationary Time Series Analysis*, Academic Press, London, UK, 1988.
5. Bissel, C. C., and Chapman, D. A., *Digital Signal Transmission*, Cambridge University Press, Cambridge, UK, 1992.
6. Maia, N. M. M, Silva, J. M. M., et al., *Theoretical and Experimental Modal Analysis*, Research Studies Press, Taunton, Somerset, UK, 1997.
7. Fourier, J. B. J., *Théorie Analytique de la Chaleur*, Chez Firmin Didot, père et fils, Paris, France, 1822.
8. Ewins, D. J., *Modal Testing: Theory and Practice*, Research Studies Press, Letchworth, Hertfordshire, UK, 1984.
9. Newland, D. E., *An Introduction to Random Vibrations and Spectral Analysis*, Longman, New York, USA, 1984.
10. Shin, K., and Hammond, J., *Fundamentals of Signal Processing for Sound and Vibration Engineers*, John Wiley & Sons, Chichester, West Sussex, UK, 2008.
11. NI 9234, Available at: http://sine.ni.com/nips/cds/view/p/lang/en/nid/208802.
12. Cooley, J. W., and Tukey, J. W., "An Algorithm for the Machine Calculation of Complex Fourier Series", *Mathematics of Computation*, Vol. 19, pp. 297-301, 1965.
13. Bergland, G. D., "A Guided Tour of the Fast Fourier Transform", *IEEE Spectrum*, Vol. 6, pp. 41-52, 1969.
14. Smith, S. W., *The Scientist and Engineer's Guide to Digital Signal Processing*, California Technical Publishing, San Diego, California, USA, 1997.
15. Hammond, J. K., "Introduction to Signal Processing - Part I: Fundamentals of Signal Processing", *J. M. Silva and N. M. M. Maia (edts.), Modal Analysis and Testing*, Kluwer Academic Publishers, Dordrecht, Netherlands, NATO Science Series E: Applied Sciences, Vol. 387, pp. 35-52, 1999.
16. LabVIEW™ 2012 Help, National Instruments Co., 2012, Available at: http://zone.ni.com/reference/en-XX/help/371361J-01/.
17. Oppenheim, A. V., and Shcafer, R. W., *Discrete-time Signal Processing*, Englewood Cliffs, New Jersey, Prentice Hall, 1989.
18. Silva, C. W., *Sensors and Actuators: Control Systems Instrumentation*, CRC Press, Boca Raton, Florida, USA, 2007.
19. Electronic Circuits and Circuit Design Information, Available at: http://www.radio-electronics.com/info/circuits/.

20. Butterworth, S., "On the Theory of Filter Amplifiers", Experimental Wireless and the Wireless Engineer, Vol. 7, pp. 536-541, 1930.
21. Matlab R2014a documentation, Curve Fitting Toolbox, Available at: http://www.mathworks.co.uk/help/curvefit/smoothing-data.html.
22. Savitzky, A., and Golay, M. J. E., "Smoothing and Differentiation of Data by Simplified Least Squares Procedures", Analytical Chemistry, Vol. 36, pp. 1627-1639, 1964.
23. Signal Smoothing Algorithms, Available at: http://www.chem.uoa.gr/applets/appletsmooth/appl_smooth2.html.
24. Luo, K., Ying, K., and Bai, J., "Savitzky-Golay Smoothing and Differentiation Filter for Even Number Data", *Signal Processing*, Vol. 85, pp. 1429-1434, 2005.
25. Piezoelectric cubic charge accelerometer 4501A, Brüel & Kjær, Available at: http://www.bksv.com/Products/transducers/vibration/accelerometers/accelerometers/4501A.
26. Montalvão, D., *Determination of Rotational Terms of the Dynamic Response by means of Modal Analysis Techniques*, M.Sc. Thesis, Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal, 2003.
27. Montalvão, D., Ribeiro, A. M. R., Maia, N. M. M., and Silva, J. M. M., "Estimation of the Rotational Terms of the Dynamic Response Matrix", *Shock and Vibration*, Vol. 11, pp. 333-350, 2004.
28. Bauer, M., Ritter, F., and Siegmund, G., "High-precision Laser Vibrometers based on Digital Doppler-signal Processing", *Proceedings of the 5th International Conference on Vibration Measurements by Laser Techniques: Advances and Applications*, Ancona, Italy, pp. 50-61, 2002.
29. Montalvão, D., *A Modal-based Contribution to the Damage Location in Laminated Composite Plates*, Ph.D. Thesis, Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal, 2010.
30. Montalvão, D., Ribeiro, A. M. R., and Maia, N. M. M., "Experimental Assessment of a Modal-based Multi-parameter Method for Locating Damage in Composite Laminates", *Experimental Mechanics*, Vol. 51, pp. 1473-1488, 2011.
31. How LVDTs work, Available at: http://www.lvdt.co.uk/how-lvdts-work/.
32. Lage, Y., Reis, L., Montalvão, D., Ribeiro, A. M. R., and Freitas, M., "Automation in Strain and Temperature Control on VHCF with na Ultrasonic Testing Facility", *Journal of ASTM International: Selected Technical Papers of the ASTM 6th Symposium on Automation of Fatigue and Fracture Testing,* in press, 2014.
33. Jones, R. M., *Mechanics of Composite Materials*, Taylor & Francis, Philadelphia, Pennsylvania, USA, 1999.
34. Montalvão, D., Fontul, M., "Harmonica: Stepped-Sine Spectrum Analyser for Transfer Function Measurement and Non-Linear Experimental Assessment", *Proceedings of M2D'2006 - 5th International Conference on Mechanics and Materials in Design*, paper no. A0519.0506, Porto, Portugal, 2006.
35. Scott, P., *Experimental Investigation into a Novel Design Concept for a Modular PEMFC Stack*, Ph.D. Thesis, School of Engineering and Technology, University of Hertfordshire, 2013.
36. Hot-Wire and Hot-Film Anemometers, Thermopedia, Available at: http://www.thermopedia.com/content/853/.

37. Montalvão, D., Ribeiro, A. M. R., and Maia, N. M. M., "A Review on Vibration Based Structural Health Monitoring with Special Emphasis on Composite Materials", *Shock and Vibration Digest*, Vol. 38, pp. 295-324, 2006.

38. IEEE 100, *The Authoritative Dictionary of IEEE Standards Terms*, 7th Edition, Standards Information Network, IEEE Press, New York, USA, 2000.