1. Transducers and Sensors

By the end of this section you should be able to:

- Discuss the definitions/specifications by which sensors are characterised.
- Describe common methods for converting a physical parameter into an electrical quantity and give examples of transducers, including those for measurement of temperature, strain, motion, position and light.
- Explain how to make sensitive measurements using a Wheatstone bridge, including balancing and offset compensation.
- Describe systems for measuring motion, temperature, strain and light intensity.

1.1. Definitions

In this course we will be studying Electrical Measurements, and we will necessarily interplay between techniques and hardware used to **sense** the quantity we wish to measure, techniques and hardware used to process the signal generated by the **sensors** and also algorithms to interpret the final result. We will be, therefore be dealing with **transducers, sensors and actuators.**

Transducers: Devices used to transform one kind of energy to another. When a transducer converts a measurable quantity (sound pressure level, optical intensity, magnetic field, etc) to an electrical voltage or an electrical current we call it a **sensor.** We will see a few examples of sensors shortly.

When the transducer converts an electrical signal into another form of energy, such as sound (which, incidentally, is a pressure **field**), light, mechanical movement, it is called an **actuator.** Actuators are important in instrumentation. They allow the use of feedback at the source of the measurement. However we will pay little attention to them in this course. The study of using actuators and feedback belongs to a course in Control theory.

A sensor can be considered in its bare form, or bundled with some electronics (amplifiers, decoders, filters, and even computers). We will use the word **instrument** to refer to a sensor together with some of its associated electronics. The distinction between a sensor and an instrument is extremely vague, as it is increasingly common to manufacture integrated sensors.

What follows is equally applicable to sensors and/or instruments. The discussion is also applicable to circuits, such as amplifiers, filters, mixers and receivers. Signal processing circuits are, in a sense, instruments. It is not very important that both input and output signals are, for example, voltages.

1.2. The linear model of a Sensor

There is a fair amount of jargon associated with sensors, used to describe the usefulness or quality of a piece of hardware. Sensor specification terms are often used in an erroneous or misleading way, especially in the advertising literature of equipment manufacturers; they tend to manipulate definitions in order to make their product appear better than it is. It is always a good idea to investigate the precise meaning of specifications, before accepting them. Below we attempt the definition of some important specifications from the engineering point of view.

The following discussion refers to an implicit linear model for the sensor. A sensor is assumed to be linear so that its response y to a stimulus x is idealised to have the form:

$$y(x) = Ax \text{ , } 0 \le x \le x_{max} \text{ , } A>0 \tag{1.1}$$

Please note that we have defined the stimulus to be positive. This makes it easier to define quantities such as the threshold, and consequently makes it easier to understand that there may exist gaps in the response of an instrument!

1.2.1. Sensitivity

The constant A in (1.1) is called the **sensitivity** or the **transducer gain** or, simply, the **gain** of the sensor. To simplify the discussion we also take the gain to be positive.

The linear model satisfies the definition of linearity, as it should:

$$y(x+z) = A(x+z) = y(x) + y(z) \tag{1.2}$$

Please note that the response of a sensor defined this way exhibits no time dependence. Such an idealised sensor has no memory and its output instantly tracks the input.

In the more general case we may know the steady state transfer function of the sensor. We can define the sensitivity as the derivative of the output with respect to the input:

$$S = \frac{\partial y}{\partial x}. \tag{1.3}$$

This is a **partial** derivative. As we shall see below, the sensor will exhibit sensitivities to other ambient (e.g. temperature) or operating parameters (e.g. a supply voltage). It is essential to study the sensor with all other (usually unintended) stimuli held constant.

Sensitivity is, in a few words, the ratio of electrical output to signal input (input transducer), or physical output to electrical input (output transducer). e.g., a temperature sensor may be quoted as 50 µV/K. and a loudspeahker as 90dBspl/W. However, the term sensitivity may also be used in its usual electronic sense, i.e. the %change of some property of a device (eg gain) as a result of a % change in some parameter, (eg the ambient temperature). For clarity, we will refer to this as the **cross-sensitivity of x on y.** The sensitivity is also called the **Gain** of the sensor or instrument.

The term sensitivity is occasionally misused to refer to the *minimum detectable signal,* i.e. the sensor's ***detectivity*** or *threshold,* which , incidentally, equals the noise floor of the sensor**.**

1.2.2. Threshold and detectivity

No sensor will respond to arbitrarily small signals. Signals in the range between zero and the sensor **threshold** x_{min} will not cause the output of the sensor to change**.** The existence of a threshold is related to nonlinearity and noise. A stimulus which is too small for the output to exceed the noise floor is considered to be smaller than the threshold. Nonlinearity can play a role as well. Consider an enhancement mode MOSFET as a voltage sensor (MOSFETs are used as very high impedance voltage or charge probes in high end “active” oscilloscope probes). Clearly such an instrument cannot respond to voltages smaller than the MOSFET threshold voltage.

A sensor will also fail to respond to stimuli which are arbitrarily large. A sensor will necessarily have a **range** or a **full scale** x_{max} **.** The full range of a sensor can be limited by **compression** or by clipping. (Note that clipping is an extreme example of compression!) Since both compression and clipping are manifestations of **nonlinearity** we conclude that all sensors are non-linear.

1.2.3. Zero offset

A real sensor will deviate from the idealised linear model. The smallest improvement we can make to the description of an assumed linear sensor is the addition of a constant **zero offset** as follows:

$$y(x) = b_0 + Ax \tag{1.4}$$

This is not a linear form, despite the fact that it is described by a first order polynomial. This is called an **affine** relation. The constant b_0 is called the **zero offset** of the sensor. The **zero offset** can be defined in two ways: The sensor reading when the input is zero, or the value of the stimulus required to make the output zero. The zero offset is simple to correct. By subtracting b_0 from y we recover a linear description of the sensor:

$$y'(x) = y(x) - b_0 = Ax \tag{1.5}$$

1.3. Non-linearity

While still retaining the time independence assumption we can introduce **non-linearity** in the model of the sensor:

$$y'(x) = Ax + b_2x^2 + b_3x^3 + \cdots = Ax + g(x) \tag{1.6}$$

The function $g(x)$ describes how much the sensor response deviates from its linearised description. There are several ways to describe **linearity** or **nonlinearity ,** each one of them described by a different term**.** The terms **linearity** and **nonlinearity** are conjugate, are used interchangeably, and often a value of linearity is quoted as non-linearity, and vice-versa.

Nonlinearity is usually measured in relative units, either as a percentage of the maximum full scale reading of the sensor or the instrument, or locally as a percentage of a reading. Ideally we wish the nonlinearity to vanish, so it must be proportional to $|g(x)|$

We can define the **absolute non-linearity** locally, at x, as:

$$\delta_y(x) = \frac{|g(x)|}{Ax},\ x \in [x_{min}, x_{max}] \tag{1.7}$$

It is more common to compare the maximum of $|g(x)|$ to the range of the sensor:

$$\delta_y = \frac{\max\left|g(x)\right|}{A\left(x_{\max} - x_{\min}\right)} = \frac{\max\left|y(x) - b_0 - Ax\right|}{A\left(x_{\max} - x_{\min}\right)},\ x \in \left[x_{\min}, x_{\max}\right] \tag{1.8}$$

Although this is the correct description of nonlinearity, sometimes the definition is given as:

$$\delta_y = \frac{\max\left|g(x)\right|}{y\left(x_{\max}\right)} \tag{1.9}$$

This form may be used in the presence of large zero offset in order to make the nonlinearity appear smaller than it is!

Nonlinearity results not only in a discrepancy between what the instrument reads and what the linear model describes. It also leads to a gain error, usually referred to as the **differential nonlinearity.**

The differential nonlinearity may be defined as the discrepancy, due to the non-linear character of the sensor's transfer function, of the sensor gain from its modelled gain. The differential non-linearity at a value x of the stimulus is simply:

$$\delta_A(x) = \frac{\left|g'(x)\right|}{A} \tag{1.10}$$

We may, of course, be interested in the maximum value of δ_A over the sensor range:

$$\delta_A = \frac{\max\left|g'(x)\right|}{A} \tag{1.11}$$

1.4. Memory effects

1.4.1. Linear sensors –Laplace transforms and Convolution

The models we have developed so far are not entirely adequate, especially when we are concerned with very fast measurements. In this case we must account for the possibility that the sensor can internally store energy. Its internal energy content can modify the sensor's behaviour. As a result the output of a sensor depends on previous measurements the sensor made, or, equivalently, the sensor exhibits memory. The time dependence of the response of a linear sensor is well known. A sensor can still be linear if its response is described by a linear differential equation:

$$\sum_{n=0}^{N} A_n \frac{\partial^n y}{\partial t^n} = \sum_{k=0}^{K} B_k \frac{\partial^k x}{\partial t^k} \tag{1.12}$$

We can take the Laplace transform of this description to conclude that:

$$y(s, X) = \left(\sum_{k=0}^{K} B_k s^k \Big/ \sum_{n=0}^{N} A_n s^n\right) x = H(s)\,X(s) \tag{1.13}$$

so that, in Laplace transform space, the sensor response is still linear in the stimulus x. The response of a sensor with a transfer function $H(s)$ at time t is the convolution integral between the history of the stimulus x and the inverse Laplace transform $h(t)$ of $H(s)$:

$$y(t)=\int_0^{\infty} h(\tau)x(t-\tau)\,d\tau \qquad (1.14)$$

This model response is also valid for the linearised affine operator, i.e. after any offsets have been subtracted from the sensor response.

Most real sensors behave like low pass filters, and they take some time to respond to their input. Consequently, there is a limit to the maximum stimulus frequency that can be detected. The maximum frequency a sensor can interpret is approximately the inverse of its **response time**.

1.4.2. Non-linear sensors – Volterra integrals

The Taylor series description cannot adequately describe the dynamics of a nonlinear sensor, and neither can the linear transfer function description. The general description of a non-linear response with memory effects requires the use of a **Volterra integral expansion** , which is a generalisation of the convolution integral applicable to nonlinear functions with memory. The reason is that the response function of a non-linear sensor is modified by past input! Formally, the time response of a non-linear network can be written as a sum of integrals:

$$\begin{aligned} y(t) &= h_0+\int_0^{\infty} h_1(\tau)x(t-\tau)\,d\tau+\int_0^{\infty}\int_0^{\infty} h_2(\tau_1,\tau_2)\,x(t-\tau_1)\,x(t-\tau_2)\,d\tau_1 d\tau_2+\cdots+ \\ &+\int\cdots\int h_n(\tau_1\cdots\tau_n)x(t-\tau_1)\cdots x(t-\tau_n)\,d\tau_1\cdots d\tau_n \end{aligned} \qquad (1.15)$$

The functions $h_1(t)$, $h_2(t)$ $\cdots$ $h_n(t)$ are called the Volterra Kernels. The first one, $h_1(t)$ is evidently the linear impulse response of the sensor. The others are not that easy to describe nor to determine. Nonetheless, the Volterra series of a memory-less non-linear function must reduce to the Taylor expansion, and therefore in this case, the Volterra kernels are given by:

$$h_n(t_1,t_2,\cdots,t_n)=\frac{1}{n!}b_n\prod_{k=1}^{n}\delta(t_k) \qquad (1.16)$$

where $\delta(t)$ is the delta function defined by:

$$\begin{aligned} &\delta(t)=0 \quad \forall x\neq 0\ , \\ &\int_0^{\infty}\delta(t)\,dt=1 \end{aligned} \qquad (1.17)$$

We are usually concerned with weakly nonlinear systems whose high order Volterra kernels vanish very quickly with the kernel order, so that only the first and second order kernels $h_1,\ h_2$ are considered. Volterra series are used extensively in high frequency Computer Aided Circuit Design and electromagnetic simulation software.

Example:

Calculate the time dependent step response of the nonlinear sensor given below, to an input step occurring at t=0. The input signal is given by:

$$x(t)=x_0u(t)=\begin{cases}0, t\leq 0\\ x_0, t>0\end{cases}$$

The sensor statically described by $y = a + bx + cx^2$, and its Volterra kernels are:

$$h_0 = a \text{ , } h_1(\tau) = \frac{b}{\xi} e^{-\tau/\xi} \text{ , } h_2(\tau_1, \tau_2) = \frac{c}{\xi_1 \xi_2} e^{-\tau_1/\xi_1} e^{-\tau_2/\xi_2}$$

Solution:

$$y(t) = a + \int_0^\infty \frac{b}{\xi} e^{-\tau/\xi} x_0 u(t-\tau) d\tau + \int_0^\infty \int_0^\infty \frac{c}{\xi_1 \xi_2} e^{-\tau_1/\xi_1} e^{-\tau_2/\xi_2} x_0^2 u(t-\tau_1) u(t-\tau_2) d\tau_1 d\tau_2 =$$

$$= a + \int_0^t b x_0 e^{-\tau/\xi} d\tau + \int_0^t \int_0^t c e^{-\tau_1/\xi_1} e^{-\tau_2/\xi_2} d\tau_1 d\tau_2 = a + \frac{b}{\xi} \xi \left(1 - e^{-t/\xi}\right) + \frac{c}{\xi_1 \xi_2} \xi_1 \xi_2 \left(1 - e^{-t/\xi_1}\right)\left(1 - e^{-t/\xi_2}\right) =$$

$$= a + b x_0 \left(1 - e^{-t/\xi}\right) + c x_0^2 \left(1 - e^{-t/\xi_1}\right)\left(1 - e^{-t/\xi_2}\right)$$

Note that in this rather simple description of the nonlinearity the quadratic term of the transfer function exhibits 3 different time constants in its steady state response!

Example:

It is even more interesting to exercise this sensor with a sinusoid signal starting at t=0:

$$x(t) = \left(x_0 + x_1 \sin(\omega t)\right) u(t) = \left(x_0 + x_1 \operatorname{Im} e^{j\omega t}\right) u(t)$$

Repeating the procedure,

$$y(t) = a + \int_0^\infty \frac{b}{\xi} e^{-\tau/\xi} \left(x_0 + x_1 \operatorname{Im} e^{j\omega(t-\tau)}\right) u(t-\tau) d\tau +$$

$$+ \int_0^\infty \int_0^\infty \frac{c}{\xi_1 \xi_2} e^{-\tau_1/\xi_1} e^{-\tau_2/\xi_2} \left(x_0 + x_1 \operatorname{Im} e^{j\omega(t-\tau_1)}\right) u(t-\tau_1) \left(x_0 + x_1 \operatorname{Im} e^{j\omega(t-\tau_2)}\right) u(t-\tau_2) d\tau_1 d\tau_2 =$$

$$= a + \int_0^t b \left(x_0 + x_1 \operatorname{Im} e^{j\omega(t-\tau)}\right) e^{-\tau/\xi} d\tau + \int_0^t \int_0^t c e^{-\tau_1/\xi_1} e^{-\tau_2/\xi_2} \left(x_0 + x_1 \operatorname{Im} e^{j\omega(t-\tau_1)}\right)\left(x_0 + x_1 \operatorname{Im} e^{j\omega(t-\tau_2)}\right) d\tau_1 d\tau_2 =$$

$$= a + b x_0 \left(1 - e^{-t/\xi}\right) + c x_0^2 \left(1 - e^{-t/\xi_1}\right)\left(1 - e^{-t/\xi_2}\right) + b x_1 \frac{\omega \xi e^{-t/\xi} - \omega \xi \cos \omega t + \sin \omega t}{1 + \omega^2 \xi^2} +$$

$$c x_0^2 \frac{\left(\omega \xi_1 e^{-t/\xi_1} - \omega \xi_1 \cos \omega t + \sin \omega t\right)\left(\omega \xi_2 e^{-t/\xi_2} - \omega \xi_2 \cos \omega t + \sin \omega t\right)}{\left(1 + \omega^2 \xi_1^2\right)\left(1 + \omega^2 \xi_2^2\right)}$$

Note that although this is one of the simplest possible forms of the second order kernel, it has introduced some major complications. We can see, for example, that the quadratic term introduces both a gain correction and a phase shift at the fundamental frequency, plus some frequency broadening (the decaying exponential prefactors) to the single tone signal!

1.4.3. Hysteresis

The existence of non-vanishing Volterra kernels gives rise to the phenomenon of **hysteresis,** i.e. the output evidently depends on previous values of the input. Even a linear sensor with memory will exhibit an apparently hysteretic behaviour if a stimulus of a high enough frequency (typically higher than the first characteristic frequency, pole or zero) is applied to it. True hysteresis, as encountered in magnetics, arises when the second order kernel has a very long time constant in the second order kernel. Hysteresis is not necessarily a complication. It may be used to provide noise immunity in threshold sensing devices. An intentionally hysteretic voltage sensor is the Schmitt Trigger used extensively to combat noise in timer or counter instruments.

1.5. Some more sensor characteristics

1.5.1. Cross-Talk or Cross-sensitivity

So far we have discussed how a sensor can be modelled to increasing degrees of accuracy. We have one final simplification to remove. We have assumed throughout that the coefficients in the model depend only on the stimulus we are measuring. This is usually not the case. We define as cross-sensitivity the gain of the sensor with respect to an unintended stimulus. For a sensor whose output y is intended to depend on a stimulus x, we can define its cross sensitivity on z as:

$$S_{y,z} = \left.\frac{\partial y}{\partial z}\right|_{x=const} \tag{1.18}$$

The existence of cross sensitivities is both a limitation and a blessing on accurate measurements. To make accurate and reproducible measurements the engineer needs to carefully monitor and record environmental parameters. A stable frequency reference, for instance, requires the use of an oven to temperature stabilize a crystal oscillator. On the other hand, a cross sensitivity can be exploited to **amplitude modulate** a measurement, in order to reduce the effect of noise on it. Many sensors/instruments exhibit high cross-sensitivity to electromagnetic fields and to vibration. For this reason precision measurements are performed in acoustically and electromagnetically shielded anechoic chambers.

Example
Consider, for example a thermistor used as a temperature sensor. We may say, for a thermistor, that its conductance is given by:

$$G = ge^{-a/T}$$

Such a sensor can be assumed to be linear, describable by an affine relation for small temperature excursions around a reference temperature T_0 :

$$G\left(T_0+\delta T\right)=G\left(T_0\right)+\frac{d}{dT}G\left(T_0\right)\delta T=G\left(T_0\right)+\frac{a}{T_0^2}G\left(T_0\right)\delta T+\frac{a^2}{T_0^4}G\left(T_0\right)\delta T^2+\cdots$$

However, in order to measure the conductance of this thermistor we need to apply a voltage V and measure the current I. This means that the thermistor will heat up, to a temperature higher than that of the environment. If the thermistor has a thermal resistance to the ambient ρ_T , its temperature, when the ambient temperature is T_A is going to be:

$$T = T_A + \rho_T VI = T_A + \rho_T V^2 G\left(T\right) = T_A + \rho_T V^2 ge^{-a/T}$$

This looks very complicated, and it clearly indicates that the sensor reading depends on the voltage applied. We can get some sense out of this by assuming we have been very careful indeed, and that the temperature rise with respect to the ambient is very small. In this case, the temperature differential due to heating is simply:

$$\delta T_{heat} = \rho_T V^2 G\left(T_A\right)$$

and we can conclude that the conductance reading is going to be approximately:

$$G\left(T_0+\delta T\right)=G\left(T_0\right)+\frac{a}{T_0^2}G\left(T_0\right)\left(\delta T+\rho_T V^2 G\left(T_0\right)\right)-\frac{a^2}{T_0^4}G^2\left(T_0\right)dT\rho_T V^2=$$

$$=G\left(T_0\right)+\frac{a}{T_0^2}\rho_T V^2 G^2\left(T_0\right)+\frac{a}{T_0^2}G\left(T_0\right)\left(1-\frac{a}{T_0^2}\rho_T V^2 G\left(T_0\right)\right)\delta T$$

In this calculation we have also included the linear in δT contribution from the quadratic term of the Taylor expansion of the thermistor's conductance. This sensor's gain is now dependant on the voltage used to measure it.

1.5.2. Resolution

The smallest change of input detectable at the output is called the **resolution**. In analogue systems the resolution is usually limited by noise. In digital systems resolution is 1 LSB (least significant bit). A high resolution does not necessarily imply a high accuracy (a watch may resolve to the nearest second, while it may be a few minutes off). It is very important to realise that the resolution of an analogue device is equivalent to the resolution of a digital device, and that no analogue circuit offers infinite resolution. Indeed, the noise floor setting the resolution of an analogue system results directly into the maximum attainable digital resolution through the Shannon Channel capacity formula:

$$B=\Delta f\ \log_2\left(\frac{S}{N}+1\right) \tag{1.19}$$

Where S is the total signal power and N the total noise power in an analogue channel of bandwidth Δf , and B is the bit rate. This implies that the equivalent number of bits of a noisy sensor (with RMS noise voltage V_N is:

$$N=\log_2\left(\frac{y_{\max}-y_{\min}}{V_N}\right) \tag{1.20}$$

The Shannon capacity formula should make it obvious that resolution can usually be increased at the expense of greater signal power, and consequently at a higher power dissipation.

1.5.3. Dynamic Range

The ratio between the maximum and minimum signals the transducer may handle. The maximum is usually limited by compression or distortion, while the minimum is defined by the threshold. The dynamic range is often measured in units of the RMS noise voltage, and expressed in a number of effective bits, or in decibels. So we will say that the dynamic range of a device is D decibels if:

$$D=20\log_{10}\left(\frac{y_{\max}}{y_{\min}}\right) \tag{1.21}$$

Sometimes we talk about the dynamic range by quoting the number of bits arising from the device dynamic range measured in units of the resolution, i.e. the RMS noise amplitude V_N.

$$N=\log_2\left(\frac{y_{\max}-y_{\min}}{V_N}\right) \tag{1.22}$$

Like resolution, the dynamic range can be increased by increasing the signal power and consequently power dissipation.

1.5.4. Non-monotonicity

We have taken for granted that a sensor's output is monotonic in its input. In other words we assume that the sensitivity never changes sign within the range of the instrument. This does not need to be the case. High resolution (multi-bit) A/D and D/A converters are notorious for being non monotonic.

1.5.5. Accuracy

The difference between the apparent value of the stimulus and the actual value is called the **accuracy**. Accuracy is easy to intuitively understand but somewhat vague to define, and can be undermined by offset, gain error, non-linearity, non-monotonicity and hysteresis.

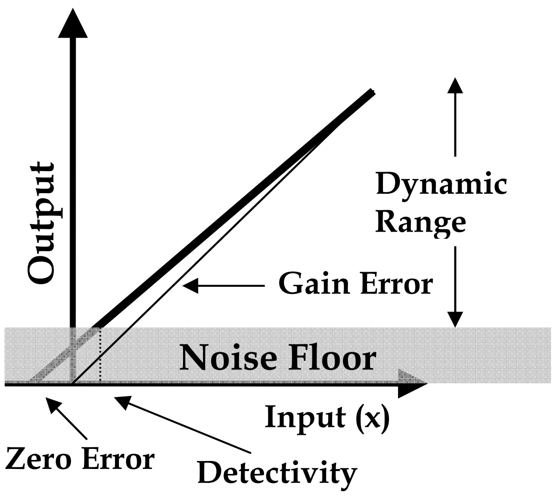

Figure 1.1: Some definitions of sensor properties: Threshold, gain, dynamic range

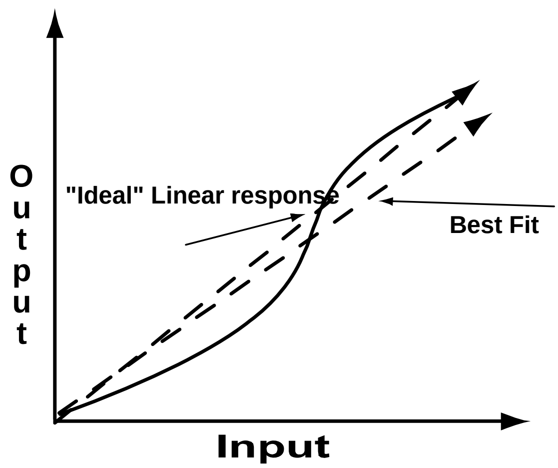

Figure 1.2: Some the concept of linearity and linear approximation to a sensor's response

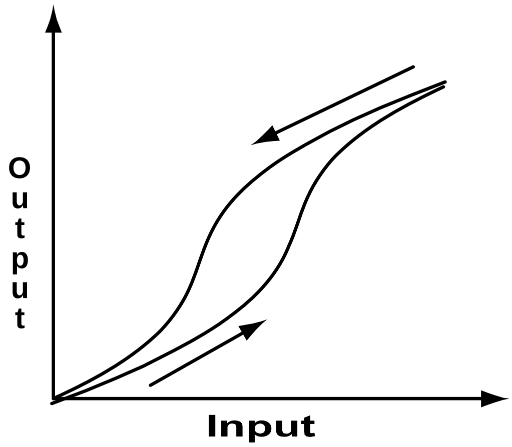

Figure 1.3: the concept of hysteresis in a sensor's response.

1.6. Electrical modelling: Active and passive sensors

Our approach to instrumentation is to treat sensors as circuit elements. We will call a sensor p**assive** when it requires an external power source, i.e. behaves like a passive circuit element such as resistor, capacitor or inductor. We will call a sensor **active when it** derives its power from the stimulus which it is measuring. In the circuit an active sensor appears like a signal source, with a Thevenin or Norton equivalent circuit.

Please note that there is some confusion regarding the use of the terms active and passive in the instrumentation literature, where the opposite definition is often used, i.e. a sensor is called active when we need to supply power to a sensor in order to use it. We prefer to use a term consistent with the notions of activity and passivity in circuit engineering.

1.7. Examples of sensors

1.7.1. Active sensors

These include:
Photovoltaic transducers: e.g. solar cells, portable exposure meters

Piezoelectric transducers generate electric polarisation, which is linearly related to the applied force (stress). Examples include gas igniters, microphones, older record player cartridges, stress/strain gauges. Piezoelectric crystals are used to measure small displacements and also as actuators to implement small (as small as 1 Angstrom!) displacements in scanning tunnelling microscopes (STM) and Atomic force microscopes (AFM).

***Thermoelectri**c* transducers: A thermocouple junction is formed when two dissimilar metals are joined at one end. When the junction is heated, a small voltage appears between the two wires which is monotonically increasing with temperature (the Seebeck effect). By suitably biasing a thermocouple junction we can *cool* a specimen. (the Peltier effect).

Electromagnetic transducers: Lentz's law dictates that a changing magnetic flux through a loop conductor will induce a voltage across its terminals. Electromagnetic sensors include microphones, phonograph pick-ups, metal detectors, and dynamos. Actuators include earphones, loudspeakers and motors, both rotational and linear. Particularly fascinating are the linear motors and associated magnetic levitation . Linear motors were apparently invented by Charles Wheatstone at Kings College in the 1840s and the first working full scale model was developed by Eric Laithwaite at Imperial College in the 1940s. (please read in the web about linear motors, especially the high acceleration types).

1.7.2. Passive sensors

These include:

Variable resistance transducers: The change in resistance of an element can be readily measured. Various components exist whose resistance changes in response to some external parameter, including potentiometers, strain gauges, resistive temperature detectors (RTDs), thermistors, photoconductive devices, and of course, potentiometers. The resistance of most metals and semiconductors depends on magnetic field, but usually in a very minor way. A recent development is that some alloys exhibit Giant Magnetoresistance (GMR). GMR sensors are used in the read heads of many modern hard disk drives.

Other variable resistance devices include:

Photoconductors - photoconductive material drops its resistance when light is shone on it.

Strain gauges - A strain gauge is a piezoresistive element designed to change resistance when a force is applied. A strain gauge is essentially a thin metallic conductor. Stretching (tension) increases the length of the wire while reducing cross-sectional area, thus increasing resistance. Compression has the opposite effect. Strain gauges are generally classified as either bonded or unbonded. An unbonded gauge typically consists of a wire resistance element stretched between two supports. A bonded gauge consists of a thin pattern of conducting foil (e.g. copper-nickel alloy) intimately bonded to a backing material, which is in turn firmly affixed onto a solid object.

Resistive temperature detectors (RTDs) - RTDs are generally constructed from platinum and their resistance increases with increasing temperature (positive temperature coefficient, PTC). The resistance is usually modelled as a polynomial in temperature, and the fitting coefficients are supplied with the sensor:

$$R = R_o\left(1+\alpha_1 T+\alpha_2 T^2+\alpha_3 T^3+........\alpha_n T^n\right) \qquad (1.23)$$

Thermistors (i.e. thermal resistors) are constructed from semiconductors or ceramics which exhibit a strong negative temperature coefficient (“tempco”) (NTC). The temperature characteristic is generally very non-linear. Physically, thermistors come in various shapes and sizes including beads, disks, wafers, rods etc. These are generally encapsulated in glass or resin. Since the conductivity of a piece of semiconductor varies exponentially with temperature,

$$R = R_o \exp\left(B\left(\frac{1}{T}-\frac{1}{T_o}\right)\right) \quad \frac{\partial R}{\partial T} = -\frac{B}{T^2}\exp\left(B\left(\frac{1}{T}-\frac{1}{T_o}\right)\right) \qquad (1.24)$$

the sensitivity is, consequently:

$$\frac{1}{R}\frac{\partial R}{\partial T} = -\frac{B}{T^2 R_o} \qquad (1.25)$$

A related class of passive sensors are a bit more fundamental, appearing pn-junction diodes: The voltage across a biased pn junction is given by

$$V_d = V_T \ln(I_d / I_S + 1) \tag{1.26}$$

where V_T is the thermal voltage kT/q. As long as the bias current greatly exceeds the reverse saturation current, which is typically a few fA,

The reverse saturation current I_S exhibits a strong positive temperature coefficient, and the net effect is that V_d decreases with increasing temperature (typically $-2mV/K$, K is the degree Kelvin), making it difficult to use this measurement for precise temperature observation. If, however, the same diode voltage drop is measured at two different currents, then:

$$\Delta V_d = V(I_1) - V(I_2) = V_T \ln(I_1 / I_2) \tag{1.27}$$

This allows a truly linear measurement of absolute temperature. Alternatively, the two current values may be applied to two identical diodes held at the same temperature, and the voltage difference can be directly measured.

Hall effect sensors: When a current flowing through a rectangular sheet conductor in the x direction is subjected to a magnetic field in the z direction, the electrons experience a force deflecting them sideways and thus producing a voltage across the conductor in the y direction. The Hall voltage measured is proportional to the field strength, and its polarity tells us if the carriers have positive or negative charge (i.e. if they are electrons or holes!). The Hall voltage appears to obey a relation similar to Ohm's law: $V_H = IR_H$, with the Hall resistance R_H given by:

$$R_H = \frac{B}{ne}$$

where n is the sheet carrier density and e the electron charge. At the same time, the ribbon, of sheet carrier density n and electron mobility μ has a resistivity of:

$$\rho = \frac{1}{ne\mu}$$

The resistance in the x direction is then:

$$R_x = \rho \frac{L}{W}$$

and in the y direction a resistance

$$R_y = \rho \frac{W}{L}$$

We can write a vector equation relating I_x , I_y , V_x , V_y , which accounts for finite currents flowing in both the x and y directions:

$$\begin{bmatrix} V_x \\ V_y \end{bmatrix} = \begin{bmatrix} \rho L/W & -R_H \\ R_H & \rho W/L \end{bmatrix} \begin{bmatrix} I_x \\ I_y \end{bmatrix}$$

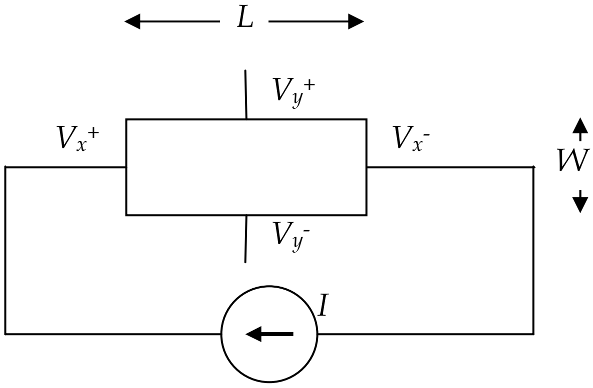

Figure 1.4: A Hall effect sensor with its current bias and terminal voltage definitions

Since the Hall resistance is inversely proportional to the carrier density, Hall probes are usually made as thin strips of low carrier density semiconductors, or even true 2-dimentional sheets of electrons residing at heterostructure interfaces. Examples are the MOSFET inversion layer and the (Quantum well) channel of HEMT (High Electron Mobility Transistors). The Hall Effect is also used to measure carrier densities in materials.

At very high magnetic fields and very low carrier densities and temperatures, a Hall probe may exhibit the *Quantum Hall Effect* in which the Hall resistance is quantised to integral submultiples of a fundamental constant (m is an integer):

$$R_H\left(m\right) = \frac{h}{me^2} = \frac{25.8}{m} k\Omega \tag{1.28}$$

Even more fascinating is the fact that the longitudinal resistance vanishes at the values of the magnetic field where the Hall resistance is quantised. It is superficially surprising that the longitudinal **conductance** also vanishes at these field values. However this surprising result follows by simply inverting the resistance matrix. The Quantum Hall Effect measurement is reproducible between differently constructed samples. It is a reproducible SI standard for the resistance measurement. The Quantum Hall Effect is used as a primary standard in Standards offices for production of secondary calibration standards for resistance meters.

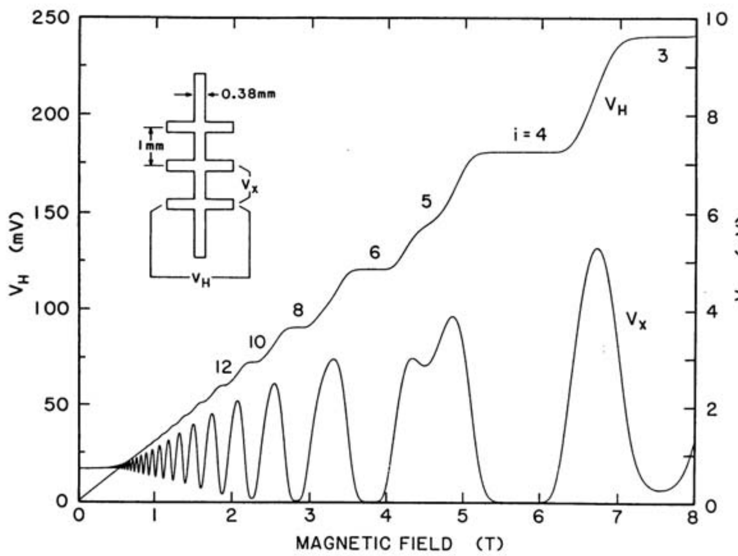

Figure 1.5: The quantum Hall Effect.

Josephson Junctions and SQUID

Two layers of superconductor separated by a thin layer of insulator form a tunnelling diode known as the Josephson junction. A Josephson junction conducts only at zero applied voltage, and then again when the applied voltage exceeds a material dependent characteristic value. If the junction is subjected to a magnetic field, its DC current depends on the applied field, as shown in the figure. When a junction is irradiated with an AC (microwave) field, its IV characteristic is modified, and steps appear, separated in voltage by $\Delta f = \frac{hf}{e}$, (h is Plank's constant) providing a fundamental voltage or frequency measurement. Conversely a junction supports an AC current when a DC voltage is imposed on it, so it can be used to generate microwave radiation.

Josephson junctions can be used as voltage to frequency converters and frequency to voltage converters, sensitive mixers and amplifiers, and also very fast switches.

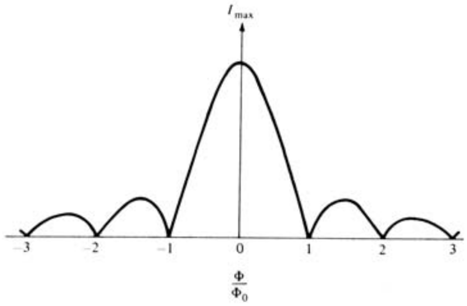

Figure 1.6: The current through a Josephson junction as a function of applied magnetic field. Note its similarity to a single slit diffraction pattern (which is the same as the spectrum of a finite width pulse). Indeed, electron wave diffraction is responsible for this curve, and this is one of the fundamental experimental proofs of the wave nature of electrons.

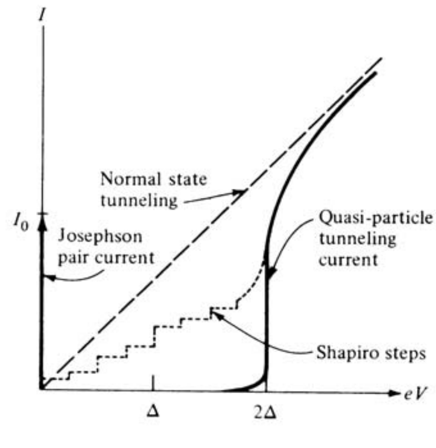

Figure 1.7:The IV characteristic of a Josephon junction without (solid) and with (dotted) an RF field imposed. The dashed line indicates the tunneling IV curve for the equivalent diode with normal (not superconducting) electrodes.

A ring of superconducting material incorporating two Josephson Junctions is called the ***Superconducting Quantum Interference Device (SQUID)***. The SQUID is the most sensitive magnetic field detector in existence, and can also serve as radiofrequency detector. SQUIDs are used in many applications requiring measurement of minute magnetic fields and variations. Perhaps the most exciting application of SQUIDs is in Magneto Encephalography, where we map the neural currents in the brain (and hence brain activity) by measuring the (minute) magnetic field they generate.

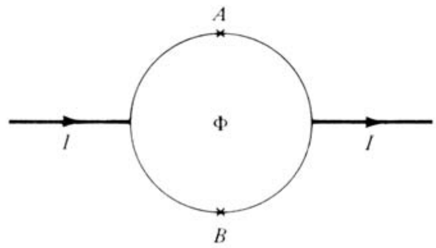

Figure 1.8: SQUID device, consisting of a superconducting ring with 2 Josephson junctions.

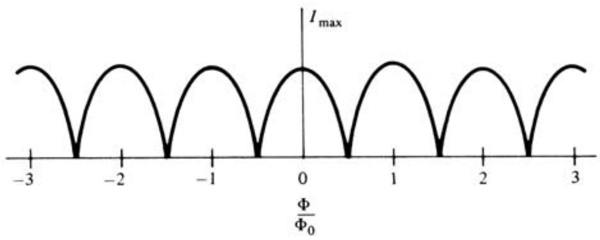

Figure 1.9: Dependence of the current through a SQUID on the magnetic field threading the ring. A squid can detect flux density as small as 10 femptoTesla (!)

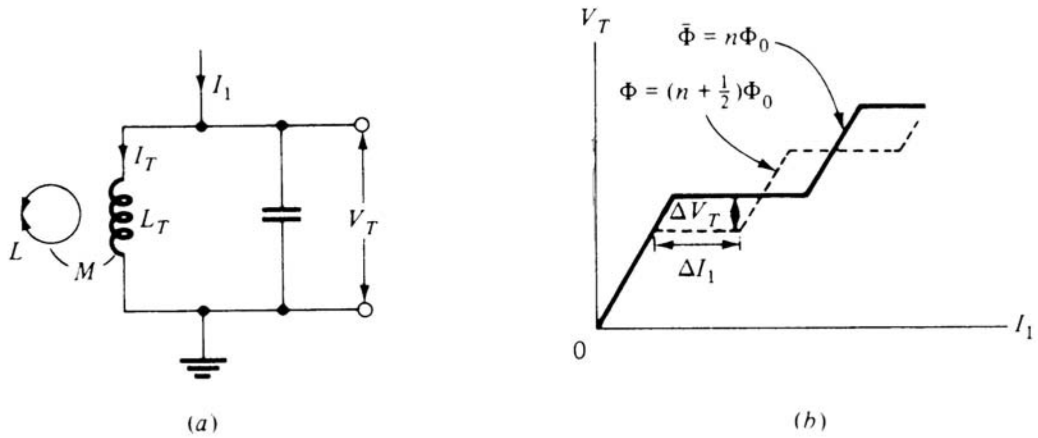

Figure 1.10: Using a SQUID as an RF detector.

Optoelectronic **transducers:** A pn diode will increase its reverse leakage current when it is illuminated. This is due to the creation of extra hole-electron pairs. In phototransistors light is used to generate the base current. Due to charge storage effects, phototransistors are reputed to be relatively slow light detectors. Since LEDs and semiconductor lasers are pn diodes, they can also detect light (but at a slightly different wavelength than they emit). For fast light detection we can use pin, MSM, and avalanche diodes. PIN diodes are pn diodes with an undoped layer between the p and n regions. These are more sensitive than pn diodes since the intrinsic layer acts as a photon absorption volume. They are also faster because of the reduced capacitance due to the intrinsic layer thickness. **Avalanche diodes** are lightly doped pn junctions so that they can be biased into such an extreme reverse bias (100s of volts) that a photoelectron generated in the device has enough energy to multiply by exciting numerous other electrons from the valence to the conduction band of the material. **MSM** (metal semiconductor metal) **diodes** are among the fastest in existence, plus they can be monolithically integrated with high speed electronics. They normally exist in GaAs IC technologies. When the utmost in sensitivity is required there exists no match for the **photomultiplier**, which can detect single photons!

Light (and sound!) detectors can of course be used for distance measurement in a rather trivial fashion by the ***time of flight*** technique where the time it takes light or sound to traverse a distance is measured. High resolution distance measurements are made by ***interferometry,*** i.e. by measuring the interference pattern between two opposite travelling waves. Visible light intereferometry is routinely used in the machine shop to measure dimensions to a fraction of a micron. When by modulating the travel path of one of the interfering beams we can make extremely high-resolution measurements, eg. of time to a few femtoseconds. Note that interferometry is equivalent to heterodyne detection. Interferometry will be discussed in some detail later in the course.

Variable reactance transducers: Can be divided into two main types: variable capacitance and variable inductance.

<u>Variable capacitance</u> - A parallel plate capacitor has a capacitance C = A/d. Any one of these terms can be varied to change the effective capacitance. A capacitive microphone uses acoustic pressure to vary plate spacing, d. A capacitive level indicator (e.g. aircraft fuel detector) varies the effective permittivity as the level of (non-conductive) liquid between the plates varies. The effective permittivity is the weighted average of the permittivities of the air and the liquid. A capacitive displacement transducer operates by varying the overlap area A by displacing the plates.

<u>Variable inductance</u> - Mainly used as displacement transducers. The inductor is wound on a core of high permeability material, and the inductance can then be varied by moving the core relative to the inductor or, by using a saturable core and changing the flux through an auxiliary coil. A variation on variable inductance concept is the variable mutual inductance sensor in which two coils are wound on the same magnetic core.

Variable capacitance or inductance sensors can me utilised either in conjunction with a frequency meter, as the variable capacitance, e.g., can be resonated with a ficed inductor leading to a variable resonance frequency which can then be measured. Variable reactance can also be directly measured.

1.8. Extending sensor usefulness

1.8.1. Calibration

No instrument is intrinsically perfect. **C*alibration techniques*** are used to extend the usefulness of an instrument, correcting for offsets, nonlinearity, hysteresis and other undesired characteristics of an instrument. To calibrate an instrument one needs to measure known quantities, and then devise an ***Error Model,*** i.e. a set of equations that allow the instrument raw reading to be corrected. Error models often involve lookup tables and interpolation, i.e. they are applicable for measurements between the minimum and maximum values of the **Calibration Standards** used. Some calibration scheme is always present in commercial instruments, although it often consists of adjusting a few trimmer potentiometers in the associated electronics. Calibration becomes essential, mathematically complicated and rather tricky to perform at high frequency and/or high precision measurements.

Calibration is related to fitting and interpolation, discussed later in the course. We will discuss calibration issues as we discuss specific measurement techniques.

1.8.2. Bridge Measurements

The sensitivity of a sensor can be increased by incorporating it in a bridge arrangement. Impedance varying sensors are often arranged in DC or AC Wheatstone bridges. In some types of measurement (e.g. in strain measurements) it is possible to perform a differential measurement comparing the outputs of 2 sensors subjected to opposite stimuli (for example, tensile and compression).

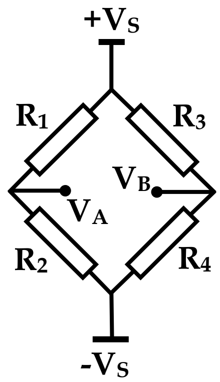

Figure 1.11: Wheatstone bridge

<u>The Wheatstone Bridge Circuit</u>
Current to the bridge flows through the excitation leads. The differential output voltage is measured via the sense leads, generally using an instrumentation amplifier. Frequently more than one transducer is used in the bridge to increase the sensitivity of the measurement. To cancel temperature effects, a 'dummy' transducer is often used on the same side of the bridge as the measuring sensor; this dummy sensor is not subject to the same input signal as the measuring sensor. The response equations of the Wheatstone bridge are well known:

$$V_{sense} = V_{cc}\left(\frac{R_2}{R_1+R_2} - \frac{R_4}{R_3+R_4}\right) \tag{1.29}$$

Equilibrium condition $V_{sense} = 0$ gives $R_1R_4 = R_2R_3$
It is straightforward to compute its sensitivity:

Differential sensitivity:

$$S_{diff\,(R_1)} = \frac{\partial V_{sense}}{\partial R_1} = V_{cc}\frac{R_2}{\left(R_1 + R_2\right)^2} \tag{1.30}$$

Relative sensitivity:

$$S_{(R_1)} = R_1\frac{\partial V_{sense}}{\partial R_1} = V_{cc}\frac{R_1R_2}{\left(R_1 + R_2\right)^2} = V_{cc}\frac{F}{\left(1+F\right)^2} \tag{1.31}$$

where $F=R_2/R_1$. This expression has a maximum when F=1, i.e. when transducer elements are pairwise equal. Increasing Vcc will also increase the sensitivity (at the expense of power dissipation).

Initial Balancing
The tolerance of the four bridge components may be such that there is a non-zero offset output voltage. This offset can be corrected by adjusting a balance current (resistor) to pull the bridge into balance. One method of doing this is demonstrated in Figure 1.12 below.

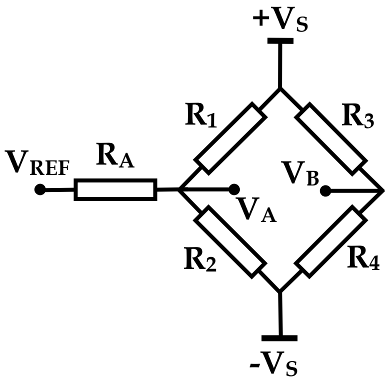

Figure 1.12: Wheatstone bridge with balancing bias

With no signal applied the bridge should be balanced with $V_A = V_B$. If V_A is slightly too high, then a negative voltage V_{ref} is applied to draw current through R_A via R_1, and thus pull down the value of V_A. Similarly if V_A is slightly too low, then a positive voltage is applied at V_{ref}. Generally V_{ref} is made variable over some range (e.g. –V to +V) to allow the bridge to be balanced for various offset values. The resistor R_A is made large, so that the V_{ref}/R_A combination appears to the bridge as a constant current source. Of course, this bias balancing technique provides a neat way to use the bridge in an automated measurement: One can measure the V_{REF} required to bring the bridge in balance. In turn, V_{REF} can easily be under computer control.

AC bridges, diode bridges
By using capacitors and inductors in the place of some or all of R_i the bridge response acquires frequency dependence, and can be used to measure capacitance, inductance or frequency. A bridge of diodes is a mixer (i.e. an analogue multiplier) used in full wave rectification (i.e. homodyne mixing) or heterodyning. Diode bridges are also used as very fast switches.