2

Sensor Characteristics

"O, what men dare do! What men may do!
What men daily do, not knowing what they do."

—Shakespeare, "Much Ado About Nothing"

From the input to the output, a sensor may have several conversion steps before it produces an electrical signal. For instance, pressure inflicted on the fiber-optic sensor first results in strain in the fiber, which, in turn, causes deflection in its refractive index, which, in turn, results in an overall change in optical transmission and modulation of photon density. Finally, photon flux is detected and converted into electric current. In this chapter, we discuss the overall sensor characteristics, regardless of its physical nature or steps required to make a conversion. We regard a sensor as a "black box" where we are concerned only with relationships between its output signal and input stimulus.

2.1 Transfer Function

An *ideal* or *theoretical* output–stimulus relationship exists for every sensor. If the sensor is ideally designed and fabricated with ideal materials by ideal workers using ideal tools, the output of such a sensor would always represent the *true* value of the stimulus. The ideal function may be stated in the form of a table of values, a graph, or a mathematical equation. An ideal (theoretical) output–stimulus relationship is characterized by the so-called *transfer function*. This function establishes dependence between the electrical signal S produced by the sensor and the stimulus $s : S = f(s)$. That function may be a simple linear connection or a nonlinear dependence, (e.g., logarithmic, exponential, or power function). In many cases, the relationship is unidimensional (i.e., the output versus one input stimulus). A unidimensional linear relationship is represented by the equation

$$S = a + bs, \tag{2.1}$$

where a is the intercept (i.e., the output signal at zero input signal) and b is the slope, which is sometimes called *sensitivity*. S is one of the characteristics of the output electric signal used by the data acquisition devices as the sensor's output. It may be amplitude, frequency, or phase, depending on the sensor properties.

Logarithmic function:

$$S = a + b \ln s. \tag{2.2}$$

Exponential function:

$$S = ae^{ks}. \tag{2.3}$$

Power function:

$$S = a_0 + a_1 s^k, \tag{2.4}$$

where k is a constant number.

A sensor may have such a transfer function that none of the above approximations fits sufficiently well. In that case, a higher-order polynomial approximation is often employed.

For a nonlinear transfer function, the sensitivity b is not a fixed number as for the linear relationship [Eq. (2.1)]. At any particular input value, s_0, it can be defined as

$$b = \frac{dS(s_0)}{ds}. \tag{2.5}$$

In many cases, a nonlinear sensor may be considered linear over a limited range. Over the extended range, a nonlinear transfer function may be modeled by several straight lines. This is called a piecewise approximation. To determine whether a function can be represented by a linear model, the incremental variables are introduced for the input while observing the output. A difference between the actual response and a liner model is compared with the specified accuracy limits (see 2.4).

A transfer function may have more than one dimension when the sensor's output is influenced by more than one input stimuli. An example is the transfer function of a thermal radiation (infrared) sensor. The function[1] connects two temperatures (T_b, the absolute temperature of an object of measurement, and T_s, the absolute temperature of the sensor's surface) and the output voltage V:

$$V = G(T_b^4 - T_s^4), \tag{2.6}$$

where G is a constant. Clearly, the relationship between the object's temperature and the output voltage (transfer function) is not only nonlinear (the fourth-order parabola) but also depends on the sensor's surface temperature. To determine the sensitivity of the sensor with respect to the object's temperature, a partial derivative will be calculated as

$$b = \frac{\partial V}{\partial T_b} = 4GT_b^3. \tag{2.7}$$

The graphical representation of a two-dimensional transfer function of Eq. (2.6) is shown in Fig. 2.1. It can be seen that each value of the output voltage can be uniquely

[1] This function is generally known as the Stefan–Boltzmann law.

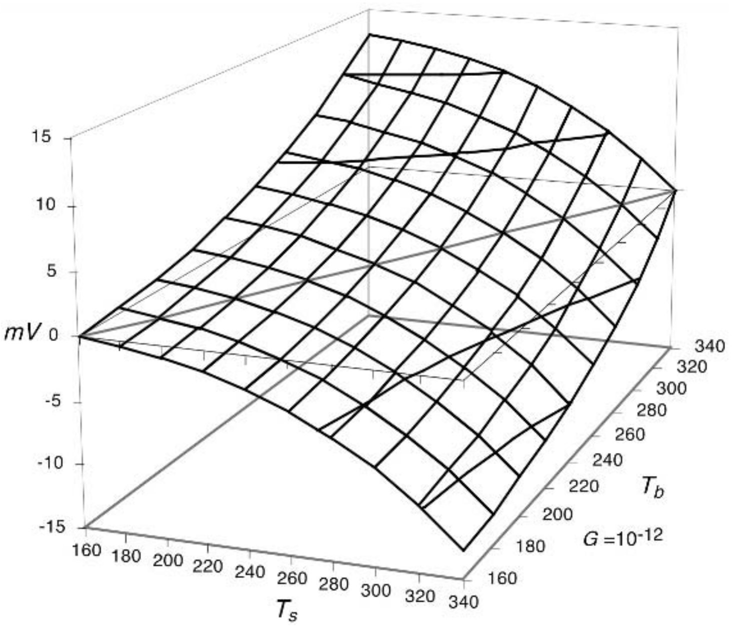

Fig. 2.1. Two-dimensional transfer function of a thermal radiation sensor.

determined from two input temperatures. It should be noted that a transfer function represents the input-to-output relationship. However, when a sensor is used for measuring or detecting a stimulus, an inversed function (output-to-input) needs to be employed. When a transfer function is linear, the inversed function is very easy to compute. When it is nonlinear the task is more complex, and in many cases, the analytical solution may not lend itself to reasonably simple data processing. In these cases, an approximation technique often is the solution.

2.2 Span (Full-Scale Input)

A dynamic range of stimuli which may be converted by a sensor is called a *span* or an *input full scale* (FS). It represents the highest possible input value that can be applied to the sensor without causing an unacceptably large inaccuracy. For the sensors with a very broad and nonlinear response characteristic, a dynamic range of the input stimuli is often expressed in decibels, which is a logarithmic measure of ratios of either power or force (voltage). It should be emphasized that decibels do not measure absolute values, but a ratio of values only. A decibel scale represents signal magnitudes by much smaller numbers, which, in many cases, is far more convenient. Being a nonlinear scale, it may represent low-level signals with high resolution while compressing the high-level numbers. In other words, the logarithmic scale for small objects works as a microscope, and for the large objects, it works as a telescope. By

Table 2.1. Relationship Among Power, Force (Voltage, Current), and Decibels

Power ratio	1.023	1.26	10.0	100	10^3	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}
Force ratio	1.012	1.12	3.16	10.0	31.6	100	316	10^3	3162	10^4	3×10^4	10^5
Decibels	0.1	1.0	10.0	20.0	30.0	40.0	50.0	60.0	70.0	80.0	90.0	100.0

definition, decibels are equal to 10 times the log of the ratio of powers (Table 2.1):

$$1\ \text{dB} = 10 \log \frac{P_2}{P_1}. \tag{2.8}$$

In a similar manner, decibels are equal to 20 times the log of the force, current, or voltage:

$$1\ \text{dB} = 20 \log \frac{S_2}{S_1}. \tag{2.9}$$

2.3 Full-Scale Output

Full-scale output (FSO) is the algebraic difference between the electrical output signals measured with maximum input stimulus and the lowest input stimulus applied. This must include all deviations from the ideal transfer function. For instance, the FSO output in Fig. 2.2A is represented by S_{FS}.

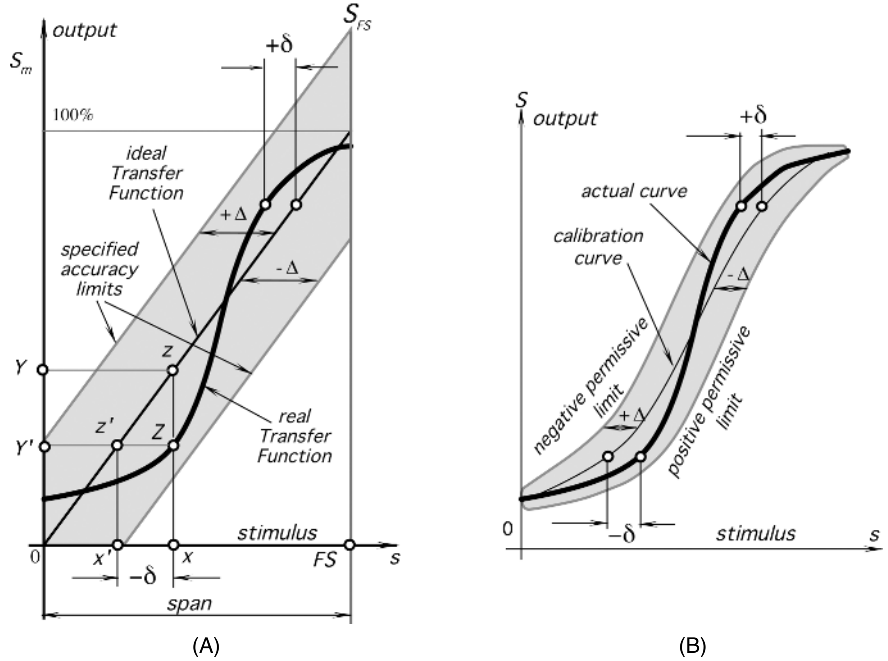

Fig. 2.2. Transfer function (A) and accuracy limits (B). Error is specified in terms of input value.

2.4 Accuracy

A very important characteristic of a sensor is *accuracy* which really means *inaccuracy*. Inaccuracy is measured as a highest deviation of a value represented by the sensor from the ideal or true value at its input. The true value is attributed to the object of measurement and accepted as having a specified uncertainty (see 2.20.)

The deviation can be described as a difference between the value which is computed from the output voltage and the actual input value. For example, a linear displacement sensor ideally should generate 1 mV per 1-mm displacement; that is, its transfer function is linear with a slope (sensitivity) $b = 1$ mV/mm. However, in the experiment, a displacement of $s = 10$ mm produced an output of $S = 10.5$ mV. Converting this number into the displacement value by using the inversed transfer function ($1/b = 1$ mm/mV), we would calculate that the displacement was $s_x = S/b = 10.5$ mm; that is $s_x - s = 0.5$ mm more than the actual. This extra 0.5 mm is an erroneous deviation in the measurement, or error. Therefore, in a 10-mm range, the sensor's absolute inaccuracy is 0.5 mm, or in the relative terms, inaccuracy is $(0.5\text{mm}/10\text{mm}) \times 100\% = 5\%$. If we repeat this experiment over and over again without any random error and every time we observe an error of 0.5 mm, we may say that the sensor has a *systematic* inaccuracy of 0.5 mm over a 10-mm span. Naturally, a random component is always present, so the systematic error may be represented as an average or mean value of multiple errors.

Figure 2.2A shows an ideal or theoretical transfer function. In the real world, any sensor performs with some kind of imperfection. A possible *real* transfer function is represented by a thick line, which generally may be neither linear nor monotonic. A real function rarely coincides with the ideal. Because of material variations, workmanship, design errors, manufacturing tolerances, and other limitations, it is possible to have a large family of real transfer functions, even when sensors are tested under identical conditions. However, all runs of the real transfer functions must fall within the limits of a specified accuracy. These permissive limits differ from the ideal transfer function line by $\pm\Delta$. The real functions deviate from the ideal by $\pm\delta$, where $\delta \leq \Delta$. For example, let us consider a stimulus having value x. Ideally, we would expect this value to correspond to point z on the transfer function, resulting in the output value Y. Instead, the real function will respond at point Z, producing output value Y'. This output value corresponds to point z' on the ideal transfer function, which, in turn, relates to a "would-be" input stimulus x' whose value is smaller than x. Thus, in this example, imperfection in the sensor's transfer function leads to a measurement error of $-\delta$.

The accuracy rating includes a combined effect of part-to-part variations, a hysteresis, a dead band, calibration, and repeatability errors (see later subsections). The specified accuracy limits generally are used in the worst-case analysis to determine the worst possible performance of the system. Figure 2.2B shows that $\pm\Delta$ may more closely follow the real transfer function, meaning better tolerances of the sensor's accuracy. This can be accomplished by a multiple-point calibration. Thus, the specified accuracy limits are established not around the theoretical (ideal) transfer function, but around the calibration curve, which is determined during the actual calibration procedure. Then, the permissive limits become narrower, as they do not embrace part-to-part variations between the sensors and are geared specifically to the calibrated unit. Clearly, this method allows more accurate sensing; however, in some applications, it may be prohibitive because of a higher cost.

The inaccuracy rating may be represented in a number of forms:

1. Directly in terms of measured value (Δ)
2. In percent of input span (full scale)
3. In terms of output signal

For example, a piezoresistive pressure sensor has a 100-kPa input full scale and a 10Ω full-scale output. Its inaccuracy may be specified as ±0.5%, ±500 Pa, or ±0.05Ω.

In modern sensors, specification of accuracy often is replaced by a more comprehensive value of *uncertainty* (see Section 2.20) because uncertainty is comprised of all distorting effects both systematic and random and is not limited to the inaccuracy of a transfer function.

2.5 Calibration

If the sensor's manufacturer's tolerances and tolerances of the interface (signal conditioning) circuit are broader than the required system accuracy, a calibration is required. For example, we need to measure temperature with an accuracy ±0.5°C; however, an available sensor is rated as having an accuracy of ±1°C. Does it mean that the sensor can not be used? No, it can, but that particular sensor needs to be calibrated; that is, its individual transfer function needs to be found during calibration. Calibration means the determination of specific variables that describe the overall transfer function. Overall means of the entire circuit, including the sensor, the interface circuit, and the A/D converter. The mathematical model of the transfer function should be known before calibration. If the model is linear [Eq. (2.1)], then the calibration should determine variables a and b; if it is exponential [Eq. (2.3)], variables a and k should be determined; and so on. Let us consider a simple linear transfer function. Because a minimum of two points are required to define a straight line, at least a two-point calibration is required. For example, if one uses a forward-biased semiconductor p-n junction for temperature measurement, with a high degree of accuracy its transfer function (temperature is the input and voltage is the output) can be considered linear:

$$v = a + bt. \tag{2.10}$$

To determine constants a and b, such a sensor should be subjected to two temperatures (t_1 and t_2) and two corresponding output voltages (v_1 and v_2) will be registered. Then, after substituting these values into Eq. (2.10), we arrive at

$$\begin{aligned} v_1 &= a + bt_1, \qquad (2.11)\\ v_2 &= a + bt_2, \end{aligned}$$

and the constants are computed as

$$b = \frac{v_1 - v_2}{t_1 - t_2} \quad \text{and} \quad a = v_1 - bt_1. \tag{2.12}$$

To compute the temperature from the output voltage, a measured voltage is inserted into an inversed equation

$$t = \frac{v - a}{b}. \tag{2.13}$$

In some fortunate cases, one of the constants may be specified with a sufficient accuracy so that no calibration of that particular constant may be needed. In the same p-n-junction temperature sensor, the slope b is usually a very consistent value for a given lot and type of semiconductor. For example, a value of $b = -0.002268$ V/°C was determined to be consistent for a selected type of the diode, then a single-point calibration is needed to find out a as $a = v_1 + 0.002268t_1$.

For nonlinear functions, more than two points may be required, depending on a mathematical model of the transfer function. Any transfer function may be modeled by a polynomial, and depending on required accuracy, the number of the calibration points should be selected. Because calibration may be a slow process, to reduce production cost in manufacturing, it is very important to minimize the number of calibration points.

Another way to calibrate a nonlinear transfer function is to use a piecewise approximation. As was mentioned earlier, any section of a curvature, when sufficiently small, can be considered linear and modeled by Eq. (2.1). Then, a curvature will be described by a family of linear lines where each has its own constants a and b. During the measurement, one should determine where on the curve a particular output voltage S is situated and select the appropriate set of constants a and b to compute the value of a corresponding stimulus s from an equation identical to Eq. (2.13).

To calibrate sensors, it is essential to have and properly maintain precision and accurate physical standards of the appropriate stimuli. For example, to calibrate contact-temperature sensors, either a temperature-controlled water bath or a "dry-well" cavity is required. To calibrate the infrared sensors, a blackbody cavity would be needed. To calibrate a hygrometer, a series of saturated salt solutions are required to sustain a constant relative humidity in a closed container, and so on. It should be clearly understood that the sensing system accuracy is directly attached to the accuracy of the calibrator. An uncertainty of the calibrating standard must be included in the statement on the overall uncertainty, as explained in 2.20.

2.6 Calibration Error

The *calibration error* is inaccuracy permitted by a manufacturer when a sensor is calibrated in the factory. This error is of a systematic nature, meaning that it is added to all possible real transfer functions. It shifts the accuracy of transduction for each stimulus point by a constant. This error is not necessarily uniform over the range and may change depending on the type of error in the calibration. For example, let us consider a two-point calibration of a real linear transfer function (thick line in Fig. 2.3). To determine the slope and the intercept of the function, two stimuli, s_1 and s_2, are applied to the sensor. The sensor responds with two corresponding output signals A_1 and A_2. The first response was measured absolutely accurately, however,

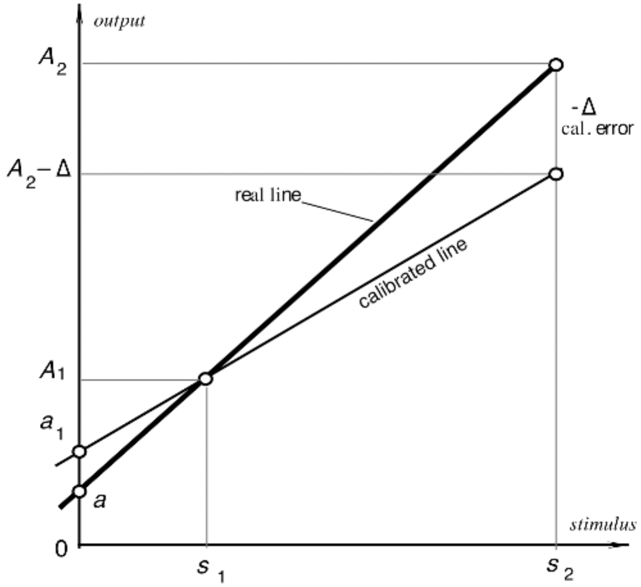

Fig. 2.3. Calibration error.

the higher signal was measured with error $-\Delta$. This results in errors in the slope and intercept calculation. A new intercept, a_1, will differ from the real intercept, a, by

$$\delta_a = a_1 - a = \frac{\Delta}{s_2 - s_1}, \tag{2.14}$$

and the slope will be calculated with error:

$$\delta_b = -\frac{\Delta}{s_2 - s_1}, \tag{2.15}$$

2.7 Hysteresis

A *hysteresis error* is a deviation of the sensor's output at a specified point of the input signal when it is approached from the opposite directions (Fig. 2.4). For example, a displacement sensor when the object moves from left to right at a certain point produces a voltage which differs by 20 mV from that when the object moves from right to left. If the sensitivity of the sensor is 10 mV/mm, the hysteresis error in terms of displacement units is 2 mm. Typical causes for hysteresis are friction and structural changes in the materials.

2.8 Nonlinearity

Nonlinearity error is specified for sensors whose transfer function may be approximated by a straight line [Eq. (2.1)]. A nonlinearity is a maximum deviation (L) of a real transfer function from the approximation straight line. The term "linearity" actually

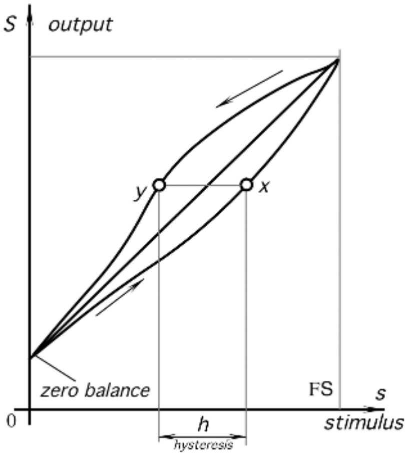

Fig. 2.4. Transfer function with hysteresis.

means "nonlinearity." When more than one calibration run is made, the worst linearity seen during any one calibration cycle should be stated. Usually, it is specified either in percent of span or in terms of measured value (e.g, in kPa or °C). "Linearity," when not accompanied by a statement explaining what sort of straight line it is referring to, is meaningless. There are several ways to specify a nonlinearity, depending how the line is superimposed on the transfer function. One way is to use *terminal* points (Fig. 2.5A); that is, to determine output values at the smallest and highest stimulus values and to draw a straight line through these two points (line 1). Here, near the terminal points, the nonlinearity error is the smallest and it is higher somewhere in between.

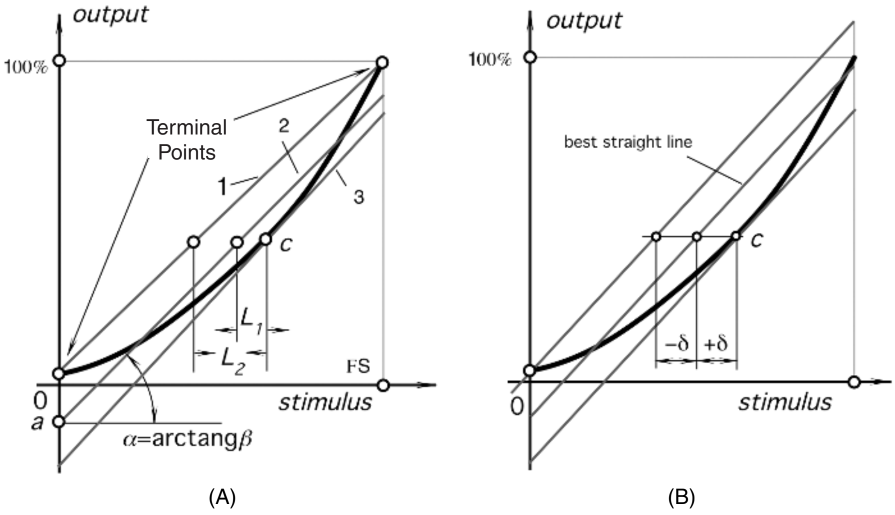

Fig. 2.5. Linear approximations of a nonlinear transfer function (A) and independent linearity (B).

Another way to define the approximation line is to use a method of *least squares* (line 2 in Fig. 2.5A). This can be done in the following manner. Measure several (n) output values S at input values s over a substantially broad range, preferably over an entire full scale. Use the following formulas for linear regression to determine intercept a and slope b of the best-fit straight line:

$$a = \frac{\sum S \sum s^2 - \sum s \sum sS}{n \sum s^2 - (\sum s)^2}, \qquad b = \frac{n \sum sS - \sum s \sum S}{n \sum s^2 - (\sum s)^2}, \tag{2.16}$$

where $\sum$ is the summation of n numbers.

In some applications, a higher accuracy may be desirable in a particular narrower section of the input range. For instance, a medical thermometer should have the best accuracy in a fever definition region which is between 37°C and 38°C. It may have a somewhat lower accuracy beyond these limits. Usually, such a sensor is calibrated in the region where the highest accuracy is desirable. Then, the approximation line may be drawn through the calibration point c (line 3 in Fig. 2.5A). As a result, nonlinearity has the smallest value near the calibration point and it increases toward the ends of the span. In this method, the line is often determined as tangent to the transfer function in point c. If the actual transfer function is known, the slope of the line can be found from Eq. (2.5).

Independent linearity is referred to as the so-called "best straight line" (Fig. 2.5B), which is a line midway between two parallel straight lines closest together and enveloping all output values on a real transfer function.

Depending on the specification method, approximation lines may have different intercepts and slopes. Therefore, nonlinearity measures may differ quite substantially from one another. A user should be aware that manufacturers often publish the smallest possible number to specify nonlinearity, without defining what method was used.

2.9 Saturation

Every sensor has its operating limits. Even if it is considered linear, at some levels of the input stimuli, its output signal no longer will be responsive. A further increase in stimulus does not produce a desirable output. It is said that the sensor exhibits a span-end nonlinearity or saturation (Fig. 2.6).

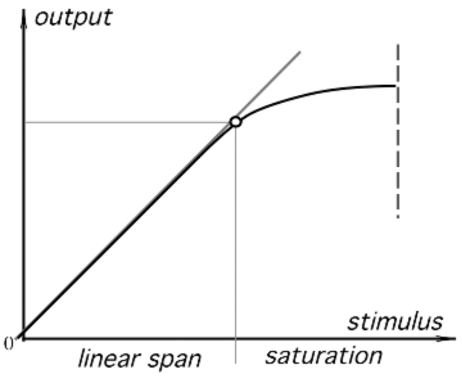

Fig. 2.6. Transfer function with saturation.

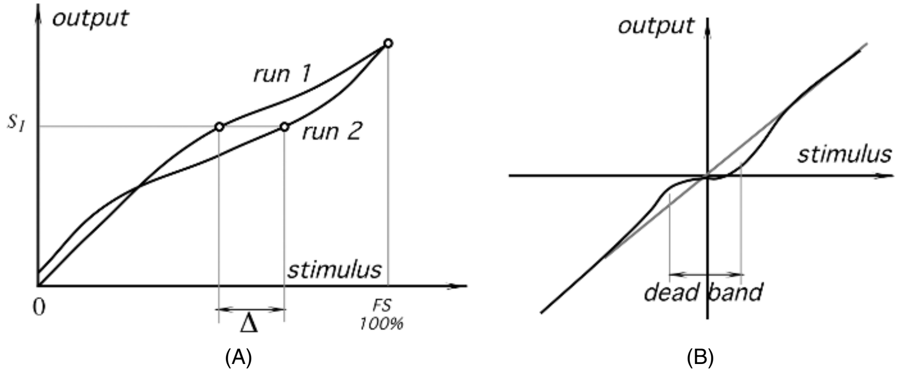

Fig. 2.7. (A) The repeatability error. The same output signal S_1 corresponds to two different input signals. (B) The dead-band zone in a transfer function.

2.10 Repeatability

A *repeatability* (reproducibility) error is caused by the inability of a sensor to represent the same value under identical conditions. It is expressed as the maximum difference between output readings as determined by two calibrating cycles (Fig. 2.7A), unless otherwise specified. It is usually represented as % of FS:

$$\delta_r = \frac{\Delta}{\mathrm{FS}} \times 100\%. \tag{2.17}$$

Possible sources of the repeatability error may be thermal noise, buildup charge, material plasticity, and so forth.

2.11 Dead Band

The *dead band* is the insensitivity of a sensor in a specific range of input signals (Fig. 2.7B). In that range, the output may remain near a certain value (often zero) over an entire dead-band zone.

2.12 Resolution

Resolution describes the smallest increments of stimulus which can be sensed. When a stimulus continuously varies over the range, the output signals of some sensors will not be perfectly smooth, even under the no-noise conditions. The output may change in small steps. This is typical for potentiometric transducers, occupancy infrared detectors with grid masks, and other sensors where the output signal change is enabled only upon a certain degree of stimulus variation. In addition, any signal converted into a digital format is broken into small steps, where a number is assigned to each step. The magnitude of the input variation which results in the output smallest step is specified as resolution under specified conditions (if any). For instance, for the occupancy detector, the resolution may be specified as follows: "resolution—minimum equidistant displacement of the object for 20 cm at 5 m distance." For wire-wound potentiometric angular sensors, resolution may be specified as "a minimum angle of 0.5°." Sometimes, it may be specified as percent of full scale (FS). For instance, for the angular sensor having 270° FS, the 0.5° resolution may be specified as 0.181% of FS. It should be noted that the step size may vary over the range, hence, the resolution may be specified as typical, average, or "worst." The resolution of digital output format sensors is given by the number of bits in the data word. For instance, the resolution may be specified as "8-bit resolution." To make sense, this statement must be accomplished with either the FS value or the value of LSB (least significant bit). When there are no measurable steps in the output signal, it is said that the sensor has *continuous* or *infinitesimal* resolution (sometimes erroneously referred to as "infinite resolution").

2.13 Special Properties

Special input properties may be needed to specify for some sensors. For instance, light detectors are sensitive within a limited optical bandwidth. Therefore, it is appropriate to specify a spectral response for them.

2.14 Output Impedance

The *output impedance* Z_{out} is important to know to better interface a sensor with the electronic circuit. This impedance is connected either in parallel with the input impedance Z_{in} of the circuit (voltage connection) or in series (current connection). Figure 2.8 shows these two connections. The output and input impedances generally should be represented in a complex form, as they may include active and reactive components. To minimize the output signal distortions, a current generating sensor (B) should have an output impedance as high as possible and the circuit's input impedance should be low. For the voltage connection (A), a sensor is preferable with lower Z_{out} and the circuit should have Z_{in} as high as practical.

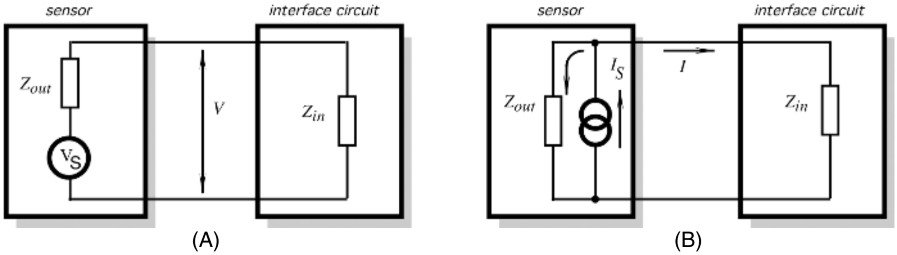

Fig. 2.8. Sensor connection to an interface circuit: (A) sensor has voltage output; (B) sensor has current output.

2.15 Excitation

Excitation is the electrical signal needed for the active sensor operation. Excitation is specified as a range of voltage and/or current. For some sensors, the frequency of the excitation signal and its stability must also be specified. Variations in the excitation may alter the sensor transfer function and cause output errors.

An example of excitation signal specification is as follows:

Maximum current through a thermistor

in still air	50 μA
in water	200 μA

2.16 Dynamic Characteristics

Under static conditions, a sensor is fully described by its transfer function, span, calibration, and so forth. However, when an input stimulus varies, a sensor response generally does not follow with perfect fidelity. The reason is that both the sensor and its coupling with the source of stimulus cannot always respond instantly. In other words, a sensor may be characterized with a *time*-dependent characteristic, which is called a *dynamic characteristic*. If a sensor does not respond instantly, it may indicate values of stimuli which are somewhat different from the real; that is, the sensor responds with a *dynamic error*. A difference between static and dynamic errors is that the latter is always time dependent. If a sensor is a part of a control system which has its own dynamic characteristics, the combination may cause, at best, a delay in representing a true value of a stimulus or, at worst, cause oscillations.

The *warm-up time* is the time between applying electric power to the sensor or excitation signal and the moment when the sensor can operate within its specified accuracy. Many sensors have a negligibly short warm-up time. However, some detectors, especially those that operate in a thermally controlled environment (a thermostat) may require seconds and minutes of warm-up time before they are fully operational within the specified accuracy limits.

In a control system theory, it is common to describe the input–output relationship through a constant-coefficient linear differential equation. Then, the sensor's dynamic (time-dependent) characteristics can be studied by evaluating such an equation. Depending on the sensor design, the differential equation can be of several *orders*.

A *zero-order* sensor is characterized by the relationship which, for a linear transfer function, is a modified Eq. (2.1) where the input and output are functions of time t:

$$S(t) = a + bs(t). \tag{2.18}$$

The value a is called an offset and b is called static sensitivity. Equation (2.18) requires that the sensor does not incorporate any energy storage device, like a capacitor or mass. A zero-order sensor responds instantaneously. In other words, such a sensor does not need any dynamic characteristics.

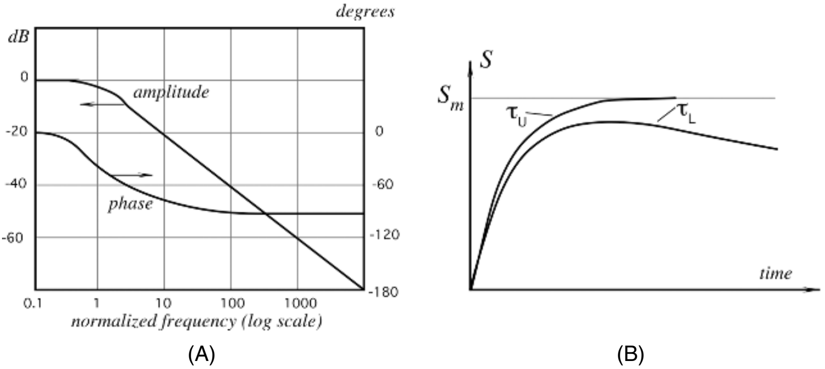

Fig. 2.9. Frequency characteristic (A) and response of a first-order sensor (B) with limited upper and lower cutoff frequencies. τ_u and τ_L are corresponding time constants.

A *first-order* differential equation describes a sensor that incorporates one energy storage component. The relationship between the input $s(t)$ and output $S(t)$ is the differential equation

$$b_1 \frac{dS(t)}{dt} + b_0 S(t) = s(t). \tag{2.19}$$

A typical example of a first-order sensor is a temperature sensor for which the energy storage is thermal capacity. The first-order sensors may be specified by a manufacturer in various ways. Typical is a *frequency response*, which specifies how fast a first-order sensor can react to a change in the input stimulus. The frequency response is expressed in hertz or rads per second to specify the relative reduction in the output signal at a certain frequency (Fig. 2.9A). A commonly used reduction number (frequency limit) is -3 dB. It shows at what frequency the output voltage (or current) drops by about 30%. The frequency response limit f_u is often called the upper cutoff frequency, as it is considered the highest frequency a sensor can process.

The frequency response directly relates to a *speed response*, which is defined in units of input stimulus per unit of time. Which response, frequency or speed, to specify in any particular case depends on the sensor type, its application, and the preference of a designer.

Another way to specify speed response is by time, which is required by the sensor to reach 90% of a steady-state or maximum level upon exposure to a step stimulus. For the first-order response, it is very convenient to use a so-called *time constant*. The time constant, τ, is a measure of the sensor's inertia. In electrical terms, it is equal to the product of electrical capacitance and resistance: $\tau = CR$. In thermal terms, thermal capacity and thermal resistances should be used instead. Practically, the time constant can be easily measured. A first-order system response is

$$S = S_m(1 - e^{-t/\tau}), \tag{2.20}$$

where S_m is steady-state output, t is time, and e is the base of natural logarithm.

Substituting $t = \tau$, we get

$$\frac{S}{S_m} = 1 - \frac{1}{e} = 0.6321. \tag{2.21}$$

In other words, after an elapse of time equal to one time constant, the response reaches about 63% of its steady-state level. Similarly, it can be shown that after two time constants, the height will be 86.5% and after three time constants it will be 95%.

The *cutoff frequency* indicates the lowest or highest frequency of stimulus that the sensor can process. The upper cutoff frequency shows how fast the sensor reacts; the lower cutoff frequency shows how slow the sensor can process changing stimuli. Figure 2.9B depicts the sensor's response when both the upper and lower cutoff frequencies are limited. As a rule of thumb, a simple formula can be used to establish a connection between the cutoff frequency, f_c (either upper and lower), and time constant in a first-order sensor:

$$f_c \approx \frac{0.159}{\tau}, \tag{2.22}$$

The *phase shift* at a specific frequency defines how the output signal lags behind in representing the stimulus change (Fig. 2.9A). The shift is measured in angular degrees or rads and is usually specified for a sensor that processes periodic signals. If a sensor is a part of a feedback control system, it is very important to know its phase characteristic. Phase lag reduces the phase margin of the system and may result in overall instability.

A *second-order* differential equation describes a sensor that incorporates two energy storage components. The relationship between the input $s(t)$ and output $S(t)$ is the differential equation

$$b_2 \frac{d^2 S(t)}{dt^2} + b_1 \frac{dS(t)}{dt} + b_0 S(t) = s(t). \tag{2.23}$$

An example of a second-order sensor is an accelerometer that incorporates a mass and a spring.

A second-order response is specific for a sensor that responds with a periodic signal. Such a periodic response may be very brief and we say that the sensor is damped, or it may be of a prolonged time and even may oscillate continuously. Naturally, for a sensor, such a continuous oscillation is a malfunction and must be avoided. Any second-order sensor may be characterized by a *resonant (natural) frequency*, which is a number expressed in hertz or rads per second. The natural frequency shows where the sensor's output signal increases considerably. Many sensors behave as if a dynamic sensor's output conforms to the standard curve of a second-order response; the manufacturer will state the natural frequency and the damping ratio of the sensor. The resonant frequency may be related to mechanical, thermal, or electrical properties of the detector. Generally, the operating frequency range for the sensor should be selected well below (at least 60%) or above the resonant frequency. However, in some sensors, the resonant frequency is the operating point. For instance, in glass-breakage detectors (used in security systems), the resonant makes the sensor selectively sensitive to a narrow bandwidth, which is specific for the acoustic spectrum produced by shattered glass.

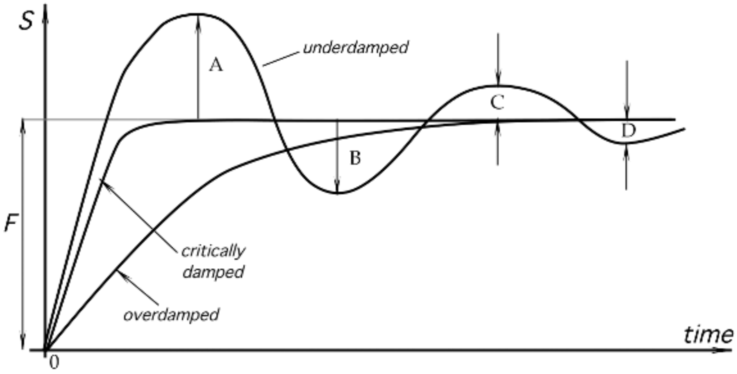

Fig. 2.10. Responses of sensors with different damping characteristics.

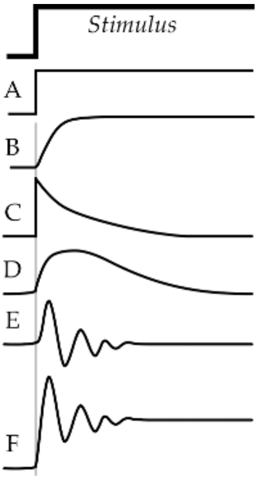

Fig. 2.11. Types of response: (A) unlimited upper and lower frequencies; (B) first-order limited upper cutoff frequency; (C) first-order limited lower cutoff frequency; (D) first-order limited both upper and lower cutoff frequencies; (E) narrow bandwidth response (resonant); (F) wide bandwidth with resonant.

Damping is the progressive reduction or suppression of the oscillation in the sensor having higher than a first-order response. When the sensor's response is as fast as possible without overshoot, the response is said to be critically damped (Fig. 2.10). An underdamped response is when the overshoot occurs and the overdamped response is slower than the critical response. The damping ratio is a number expressing the quotient of the actual damping of a second-order linear transducer by its critical damping.

For an oscillating response, as shown in Fig. 2.10, a *damping factor* is a measure of damping, expressed (without sign) as the quotient of the greater by the lesser of a pair of consecutive swings in opposite directions of the output signal, about an ultimately steady-state value. Hence, the damping factor can be measured as

$$\text{Damping factor} = \frac{F}{A} = \frac{A}{B} = \frac{B}{C} = \text{etc.} \tag{2.24}$$

2.17 Environmental Factors

Storage conditions are nonoperating environmental limits to which a sensor may be subjected during a specified period without permanently altering its performance under normal operating conditions. Usually, storage conditions include the highest and the lowest storage temperatures and maximum relative humidities at these temperatures. The word "noncondensing" may be added to the relative humidity number. Depending on the sensor's nature, some specific limitation for the storage may need to be considered (e.g., maximum pressure, presence of some gases or contaminating fumes, etc.).

Short- and long-term stabilities (drift) are parts of the accuracy specification. The short-term stability is manifested as changes in the sensor's performance within minutes, hours, or even days. The sensor's output signal may increase or decrease, which, in other terms, may be described as ultralow-frequency noise. The long-term stability may be related to *aging* of the sensor materials, which is an irreversible change in the material's electrical, mechanical, chemical, or thermal properties; that is, the long-term drift is usually unidirectional. It happens over a relatively long time span, such as months and years. Long-term stability is one of the most important for sensors used for precision measurements. Aging depends heavily on environmental storage and operating conditions, how well the sensor components are isolated from the environment, and what materials are used for their fabrication. The aging phenomenon is typical for sensors having organic components and, in general, is not an issue for a sensor made with only nonorganic materials. For instance, glass-coated metal-oxide thermistors exhibit much greater long-term stability compared to epoxy-coated thermistors. A powerful way to improve long-term stability is to preage the component at extreme conditions. The extreme conditions may be cycled from the lowest to the highest. For instance, a sensor may be periodically swung from freezing to hot temperatures. Such accelerated aging not only enhances the stability of the sensor's characteristics but also improves the reliability (see Section 2.18), as the preaging process reveals many hidden defects. For instance, epoxy-coated thermistors may be greatly improved if they are maintained at $+150°C$ for 1 month before they are calibrated and installed in a product.

Environmental conditions to which a sensor is subjected do not include variables which the sensor measures. For instance, an air-pressure sensor usually is subjected not just to air pressure but to other influences as well, such as the temperatures of air and surrounding components, humidity, vibration, ionizing radiation, electromagnetic fields, gravitational forces, and so forth. All of these factors may and usually do affect the sensor's performance. Both static and dynamic variations in these conditions should be considered. Some environmental conditions are usually of a multiplicative nature; that is, they alter a transfer function of the sensor (e.g., changing its gain). One example is the resistive strain gauge, whose sensitivity increases with temperature.

Environmental stability is quite broad and usually a very important requirement. Both the sensor designer and the application engineer should consider all possible external factors which may affect the sensor's performance. A piezoelectric accelerometer may generate spurious signals if affected by a sudden change in ambient temperature, electrostatic discharge, formation of electrical charges (triboelectric effect), vibration of a connecting cable, electromagnetic interference (EMI), and so forth. Even if a manufacturer does not specify such effects, an application engineer should simulate them during the prototype phase of the design process. If, indeed, the environmental factors degrade the sensor's performance, additional corrective measures may be required (see Chapter 4) (e.g., placing the sensor in a protective box, using electrical shielding, using a thermal insulation or a thermostat).

Temperature factors are very important for sensor performance; they must be known and taken into account. The operating temperature range is the span of ambient temperatures given by their upper and lower extremes (e.g., −20°C to +100°C) within which the sensor maintains its specified accuracy. Many sensors change with temperature and their transfer functions may shift significantly. Special compensating elements are often incorporated either directly into the sensor or into signal conditioning circuits, to compensate for temperature errors. The simplest way of specifying tolerances of thermal effects is provided by the error-band concept, which is simply the error band that is applicable over the operating temperature band. A temperature band may be divided into sections, whereas the error band is separately specified for each section. For example, a sensor may be specified to have an accuracy of ±1% in the range from 0°C to 50°C, ±2% from −20°C to 0°C and from +50°C to 100°C, and ±3% beyond these ranges within operating limits specified from −40°C to +150°C.

Temperatures will also affect dynamic characteristics, particularly when they employ viscous damping. A relatively fast temperature change may cause the sensor to generate a spurious output signal. For instance, a dual pyroelectric sensor in a motion detector is insensitive to slowly varying ambient temperature. However, when the temperature changes quickly, the sensor will generate an electric current that may be recognized by a processing circuit as a valid response to a stimulus, thus causing a false-positive detection.

A *self-heating error* may be specified when an excitation signal is absorbed by a sensor and changes its temperature by such a degree that it may affect its accuracy. For instance, a thermistor temperature sensor requires passage of electric current, causing heat dissipation within the sensor's body. Depending on its coupling with the environment, the sensors' temperature may increase due to a self-heating effect. This will result in errors in temperature measurement because the thermistor now acts as an additional spurious source of thermal energy. The coupling depends on the media in which the sensor operates—a dry contact, liquid, air, and so forth. A worst coupling may be through still air. For thermistors, manufacturers often specify self-heating errors in air, stirred liquid, or other media.

A sensor's temperature increase above its surroundings may be found from the following formula:

$$\Delta T^\circ = \frac{V^2}{(\xi v c + \alpha) R}, \tag{2.25}$$

where ξ is the sensor's mass density, c is specific heat, v is the volume of the sensor, α is the coefficient of thermal coupling between the sensor and the outside (thermal conductivity), R is the electrical resistance, and V is the effective voltage across the resistance. If a self-heating results in an error, Eq. (2.25) may be used as a design guide. For instance, to increase α, a thermistor detector should be well coupled to the object by increasing the contact area, applying thermally conductive grease or using thermally conductive adhesives. Also, high-resistance sensors and low measurement voltages are preferable.

2.18 Reliability

Reliability is the ability of a sensor to perform a required function under stated conditions for a stated period. It is expressed in statistical terms as a probability that the device will function without failure over a specified time or a number of uses. It should be noted that reliability is not a characteristic of drift or noise stability. It specifies a *failure*, either temporary or permanent, exceeding the limits of a sensor's performance under normal operating conditions.

Reliability is an important requirement; however, it is rarely specified by the sensor manufacturers. Probably, the reason for that is the absence of a commonly accepted measure for the term. In the United States, for many electronic devices, the procedure for predicting in-service reliability is the MTBF (mean time between failure) calculation described in MIL-HDBK-217 standard. Its basic approach is to arrive at a MTBF rate for a device by calculating the individual failure rates of the individual components used and by factoring in the kind of operation the device will see: its temperature, stress, environment, and screening level (measure of quality). Unfortunately, the MTBF reflects reliability only indirectly and it is often hardly applicable to everyday use of the device. The qualification tests on sensors are performed on combinations of the worst possible conditions. One approach (suggested by MIL-STD-883) is 1000 h, loaded at maximum temperature. This test does not qualify for such important impacts as fast temperature changes. The most appropriate method of testing would be accelerated life qualification. It is a procedure that emulates the sensor's operation, providing real-world stresses, but compressing years into weeks. Three goals are behind the test: to establish MTBF; to identify first failure points that can then be strengthened by design changes; and to identify the overall system practical lifetime.

One possible way to compress time is to use the same profile as the actual operating cycle, including maximum loading and power-on, power-off cycles, but expanded environmental highest and lowest ranges (temperature, humidity, and pressure). The highest and lowest limits should be substantially broader than normal operating conditions. Performance characteristics may be outside specifications, but must return to those when the device is brought back to the specified operating range. For example, if a sensor is specified to operate up to 50°C at the highest relative humidity (RH) of 85% at a maximum supply voltage of +15 V, it may be cycled up to 100°C at 99% RH and at +18 V power supply. To estimate number of test cycles (n), the following empirical formula [developed by Sandstrand Aerospace, (Rockford, IL) and Interpoint Corp. (Redmond, WA)] [1] may be useful:

$$n = N\left(\frac{\Delta T_{\max}}{\Delta T_{\text{test}}}\right)^{2.5}, \tag{2.26}$$

where N is the estimated number of cycles per lifetime, $\Delta T_{\max}$ is the maximum specified temperature fluctuation, and ΔT_{test} maximum cycled temperature fluctuation during the test. For instance, if the normal temperature is 25°C, the maximum specified temperature is 50°C, cycling was up to 100°C, and over the lifetime (say, 10 years), the sensor was estimated to be subjected to 20,000 cycles, then the number of test cycles is calculated as

$$n = 20{,}000\left(\frac{50-25}{100-25}\right)^{2.5} = 1283.$$

As a result, the accelerated life test requires about 1300 cycles instead of 20,000. It should be noted, however, that the 2.5 factor was derived from a solder fatigue multiple, because that element is heavily influenced by cycling. Some sensors have no solder connections at all, and some might have even more sensitivity to cycling substances other than solder, (e.g, electrically conductive epoxy). Then, the factor should be selected to be somewhat smaller. As a result of the accelerated life test, the reliability may be expressed as a probability of failure. For instance, if 2 out of 100 sensors (with an estimated lifetime of 10 years) failed the accelerated life test, the reliability is specified as 98% over 10 years.

A sensor, depending on its application, may be subjected to some other environmental effects which potentially can alter its performance or uncover hidden defects. Among such additional tests are:

- High temperature/high humidity while being fully electrically powered. For instance, a sensor may be subjected to its maximum allowable temperature at 85–90% RH and kept under these conditions for 500 h. This test is very useful for detecting contaminations and evaluating packaging integrity. The life of sensors, operating at normal room temperatures, is often accelerated at 85°C and 85% RH, which is sometimes called an "85–85 test."
- Mechanical shocks and vibrations may be used to simulate adverse environmental conditions, especially in the evaluation wire bonds, adhesion of epoxy, and so forth. A sensor may be dropped to generate high-level accelerations (up to 3000g of force). The drops should be made on different axes. Harmonic vibrations should be applied to the sensor over the range which includes its natural frequency. In the United States military standard 750, methods 2016 and 2056 are often used for mechanical tests.
- Extreme storage conditions may be simulated, for instance at +100 and −40°C while maintaining a sensor for at least 1000 h under these conditions. This test simulates storage and shipping conditions and usually is performed on nonoperating devices. The upper and lower temperature limits must be consistent with the sensor's physical nature. For example, TGS pyroelectric sensors manufactured in the past by Philips are characterized by a Curie temperature of +60°C. Approaching and surpassing this temperature results in a permanent destruction of sensitivity. Hence, the temperature of such sensors should never exceed +50°C, which must be clearly specified and marked on its packaging material.

- Thermal shock or temperature cycling (TC) is subjecting a sensor to alternate extreme conditions. For example, it may be dwelled for 30 min at −40°C, then quickly moved to +100°C for 30 min, and then back to cold. The method must specify the total number of cycling, like 100 or 1000. This test helps to uncover die bond, wire bond, epoxy connections, and packaging integrity.
- To simulate sea conditions, sensors may be subjected to a salt spray atmosphere for a specified time, (e.g., 24 h). This helps to uncover its resistance to corrosion and structural defects.

2.19 Application Characteristics

Design, *weight*, and overall *dimensions* are geared to specific areas of applications. *Price* may be a secondary issue when the sensor's reliability and accuracy are of paramount importance. If a sensor is intended for life-support equipment, weapons or spacecraft, a high price tag may be well justified to assure high accuracy and reliability. On the other hand, for a very broad range of consumer applications, the price of a sensor often becomes a cornerstone of a design.

2.20 Uncertainty

Nothing is perfect in this world, at least in the sense that we perceive it. All materials are not exactly as we think they are. Our knowledge of even the purest of the materials is always approximate; machines are not perfect and never produce perfectly identical parts according to drawings. All components experience drifts related to the environment and their aging; external interferences may enter the system and alter its performance and modify the output signal. Workers are not consistent and the human factor is nearly always present. Manufacturers fight an everlasting battle for the uniformity and consistency of the processes, yet the reality is that every part produced is never ideal and carries an uncertainty of its properties. Any measurement system consists of many components, including sensors. Thus, no matter how accurate the measurement is, it is only an approximation or estimate of the true value of the specific quantity subject to measurement, (i.e., the stimulus or measurand). The result of a measurement should be considered complete only when accompanied by a quantitative statement of its uncertainty. We simply never can be 100% sure of the measured value.

When taking individual measurements (samples) under noisy conditions we expect that the stimulus s is represented by the sensor as having a somewhat different value s', so that the error in measurement is expressed as

$$\delta = s' - s, \tag{2.27}$$

The difference between the *error* specified by Eq. (2.27) and *uncertainty* should always be clearly understood. An error can be compensated to a certain degree by correcting its systematic component. The result of such a correction can unknowably be very close to the unknown true value of the stimulus and, thus, it will have a very small error. Yet, in spite of a small error, the uncertainty of measurement may be very large so we cannot really trust that the error is indeed that small. In other words, an error is what we unknowably *get* when we measure, whereas uncertainty is what we *think* how large that error might be.

The International Committee for Weight and Measures *(CIPM)* considers that uncertainty consists of many factors that can be grouped into two classes or types [2,3]:

A: Those evaluated by statistical methods
B: Those evaluated by other means.

This division is not clear-cut and the borderline between Types A and B is somewhat illusive. Generally, Type A components of uncertainty arise from random effects, whereas the Type B components arise from systematic effects.

Type A uncertainty is generally specified by a standard deviation s_i, equal to the positive square root of the statistically estimated variance s_i^2 and the associated number of degrees of freedom v_i. For such a component, the *standard* uncertainty is $u_i = s_i$. Standard uncertainty represents each component of uncertainty that contributes to the uncertainty of the measurement result.

The evaluation of a Type A standard uncertainty may be based on any valid statistical method for treating data. Examples are calculating the standard deviation of the mean of a series of independent observations, using the method of least squares to fit a curve to data in order to estimate the parameters of the curve and their standard deviations. If the measurement situation is especially complicated, one should consider obtaining the guidance of a statistician.

The evaluation of a Type B standard uncertainty is usually based on scientific judgment using all of the relevant information available, which may include the following:

- Previous measurement data
- Experience with or general knowledge of the behavior and property of relevant sensors, materials, and instruments
- Manufacturer's specifications
- Data obtained during calibration and other reports
- Uncertainties assigned to reference data taken from handbooks and manuals

For detailed guidance of assessing and specifying standard uncertainties one should consult specialized texts (e.g., Ref. [4]).

When both Type A and Type B uncertainties are evaluated, they should be combined to represent the *combined standard uncertainty*. This can be done by using a conventional method for combining standard deviations. This method is often called the *law of propagation of uncertainty* and in common parlance is known as "root-sum-of-squares" (square root of the sum-of-the-squares) or RSS method of combining uncertainty components estimated as standard deviations:

$$u_c = \sqrt{u_1^2 + u_2^2 + \cdots + u_i^2 + \cdots + u_n^2}, \tag{2.28}$$

where n is the number of standard uncertainties in the uncertainty budget.

Table 2.2. Uncertainty Budget for Thermistor Thermometer

Source of Uncertainty	Standard uncertainty (°C)	Type
Calibration of sensor	0.03	B
Measured errors		
Repeated observations	0.02	A
Sensor noise	0.01	A
Amplifier noise	0.005	A
Sensor aging	0.025	B
Thermal loss through connecting wires	0.015	A
Dynamic error due to sensor's inertia	0.005	B
Temperature instability of object of measurement	0.04	A
Transmitted noise	0.01	A
Misfit of transfer function	0.02	B
Ambient drifts		
Voltage reference	0.01	A
Bridge resistors	0.01	A
Dielectric absorption in A/D capacitor	0.005	B
Digital resolution	0.01	A
Combined standard uncertainty	**0.068**	

Table 2.2 shows an example of an uncertainty budget for an electronic thermometer with a thermistor sensor which measures the temperature of a water bath. While compiling such a table, one must be very careful not to miss any standard uncertainty, not only in a sensor but also in the interface instrument, experimental setup, and the object of measurement. This must be done for various environmental conditions, which may include temperature, humidity, atmospheric pressure, power supply variations, transmitted noise, aging, and many other factors.

No matter how accurately any individual measurement is made, (i.e., how close the measured temperature is to the true temperature of an object), one never can be sure that it is indeed accurate. The combined standard uncertainty of 0.068°C does not mean that the error of measurement is no greater than 0.068°C. That value is just a standard deviation, and if an observer has enough patience, he may find that individual errors may be much larger. The word "uncertainty" by its very nature implies that the uncertainty of the result of a measurement is an estimate and generally does not have well-defined limits.

References

1. Better reliability via system tests. *Electron. Eng. Times* 40–41, Aug. 19, 1991.
2. CIPM, *BIPM Proc.-Verb. Com. Int. Poids et Mesures* 49, pp. 8–9, No. 26, 1981 (in French).

3. *ISO Guide to the Expression of Uncertainty in Measurements*. International Organization for Standardization, Geneva, 1993.
4. Taylor, B. N. and Kuyatt, C. E. *Guidelines for Evaluation and Expressing the Uncertainty of NIST Measurement Results*. NIST Technical Note 1297, Gaithersburg, 1994.