An Introduction to Modern Sensor Technology

Course No: E05-015
Credit: 5 PDH

Donald Parnell, P.E.

CEDengineering.com
PDH FOR THE PROFESSIONAL

Continuing Education and Development, Inc.

P: (877) 322-5800
info@cedengineering.com

An introduction to Modern Sensor Technology

Intro

This course provides a brief overview of the basic principles on which our modern sensor technology operate. In describing sensors, we will need to also incorporate the basic concepts of physical measurement and the use of measurement instrumentation.

In examining the basics principles of sensor operation, we will cover some of the fundamental effects, phenomena, laws and rules of physics which are used by sensors to measure and quantify.

Though much of the modern sensor technology are based on MEMS (micro electro-mechanical systems), many of the same physical mechanisms which had allowed analog sensors in the past to operate, are still being used today in these micro “devices and machines”.

Table of Contents

Chapter 1: Introduction to Sensing and Measurement

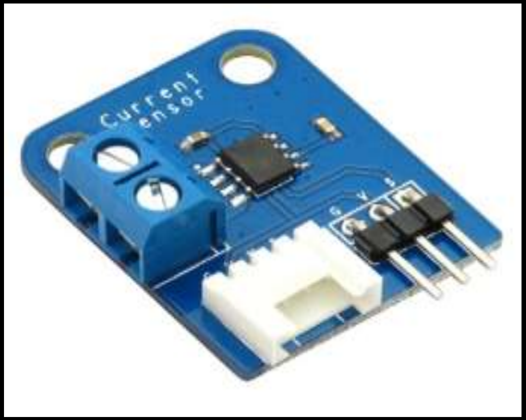
What are sensors?

Sensors are basically devices which "read" a physical stimulus, and then convert that reading into an electrical signal output.

Sensors provide us with an interface between the myriad of analog and digital electronic devices, and the physical events or conditions which occur in the real world.

The simplest definition of a sensor may be….
"a device which detects a measurand or stimulus as input then sends a signal as output to a receiving device."

A sensor is a subcomponent of a "measurement system", which measures an event, condition or change in its environment, and then responds by outputting a "signal" to a receiving device, (such as a meter, computer processor or display.)



The foundations of modern sensor devices

Many of the basic principles used by sensors to generate a measurement are based on fundamental physical conditions which have been observed by scientists since well before the 20th century.

Much of the basis of our modern class of sensor technology originated with principles developed for the analog based transducer devices and concepts which were in use many years ago.

Modern devices such as:

- MEMS) or *microelectromechanical sensors*
- Smart sensors
- Wireless sensors
- Nano-sensors and Nano-generators
- Other revolutionary microtechnologies

Sensors as detectors

The term *detector* is often used to describe a sensor, though in actuality the sensor is merely a subcomponent of a detection system.

A detection system, will detect a presence, indicating the existence of a stimulus without providing a quantifying measurement output. They normally provide an output in the form of alarms, strobes, or other indicating components.

Examples include smoke detectors (image), metal detectors, gas detectors and so on.

Sensors and transducers

The term transducer is often used to describe a sensor as well, though technically a sensor “senses” a stimulus, while a transducer “transduces” (converts one form of energy to another through a process called *transduction*.)

Differences between sensors and transducers

While a sensor and a transducer are basically the same; there are subtle differences between the two:

- A transducer is a device which converts one form of energy into another, such as converting acceleration into an output voltage, while a sensor is a device which can sense and measure a particular physical quantity (known as the measurand or stimulus).
- A sensor can be defined as a subcomponent of the transducer; with the sensor being a part of the initial stage of a transduction process.
- The purpose of a sensor is to observe and acquire information, while the purpose of a transducer is for energy conversion.
- “Transducer” implies that input and output quantities are of dissimilar energies.
- A transducer can be considered as a sensor when it is used to measure a given physical quantity. But the transducer can be an actuator as well, when the electrical input is converted into a mechanical action.
- An actuator is considered to be a transducer, and is basically the opposite of a sensor, (as it converts an electric signal into a nonelectrical form of energy).
- A transducer is a converter of one form of energy into another, whereas the sensor converts any type of energy into an electrical signal, which doesn’t always imply an act of transduction.

Properties Found in a Higher Quality of Sensor Device

Influences on a sensor device's quality

There are certain internal and external properties and conditions of a sensing device which distinguish a good sensor from a lesser quality sensor.

Much of what influences the quality of the sensor internally is based on:

- the manufacturer's level of quality assurance
- the design of the sensor
- manufacturing processes used
- the types of materials used to manufacture the device

In addition, the environment in which a sensor is installed can have an external bearing on the long term quality of the device, as well as the conditions upon which it is used.

Has a high degree of sensitivity

Sensitivity is an indication of how much the output of the device changes with the change in the input quantity to be measured.

Most sensors have a linear transfer function (a mathematical function which theoretically models the device's output for each possible input).

The sensitivity is the slope of this transfer function.

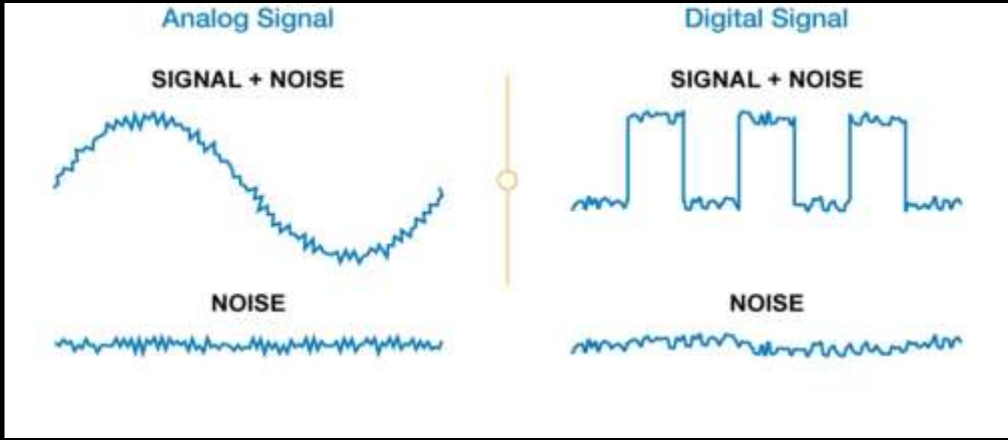
Sensitivity is defined as the ratio between the output signal and measured property.

Has reduced levels of noise and signal interference

The sensor should have a minimal amount of *noise* and *signal interference* (or disturbance) in the signal output.

Noise is the unwanted portion of a transmitted signal which does not pertain to the measured value.

Image Source: predig.com



Has a low power demand

A good sensor should have a minimal requirement of power demand in order to operate.

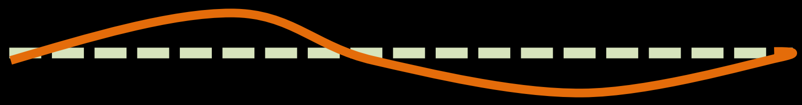
Sensor power demand differs, depending on whether the sensor is:

- Passive - a self-generating type
- Active - requires an external source of excitation to operate

A good sensor has a high resolution (ie. - detects smaller increments of change)

Sensor *resolution* is a device characteristic that refers to the smallest increment of change of input that the sensing device is able to detect.

The smaller the increment of change that can be detected, the better the sensor device will be.



Has good linearity

A sensor's "linearity" is another positive characteristic of a sensor device which describes the amount of deviation (or non-linearity) that the sensor output curve displays, from a theoretical straight line.

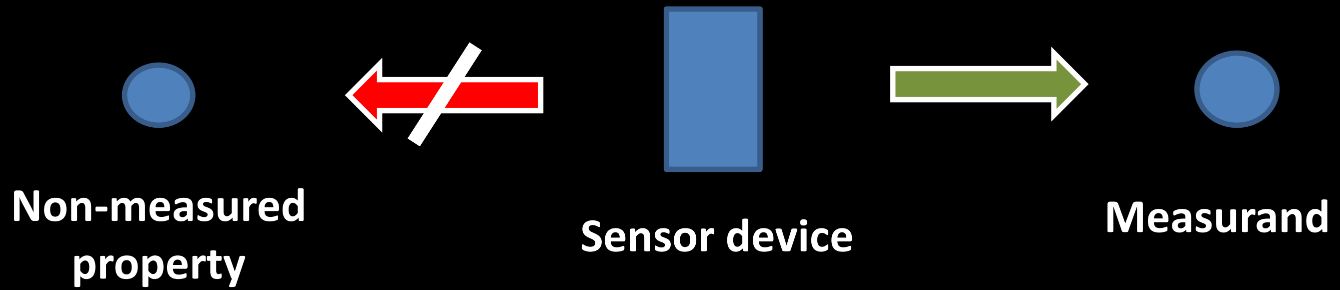
This linearity is analyzed over the extent of the *operating range* of the sensor. The *linearity error value* is normally specified as a percentage of the operating range.

It has a minimal amount of influence on the measurand*

A good sensor will exert a minimal amount of influence on the measurand, meaning that the operating of a sensing device should not change the state of the property, event or condition which is being measured.

Influencing effects from a sensor can include: vibratory, audible or a thermal transfer to the measurand.

**Measurand - an object, physical quantity, or property which is measured.*



Exhibits sensitivity to the measurand only

A good sensor will be strictly sensitive to the measured property, while being insensitive to other non-measured properties.

The Measurement Process

Sensors and the measurement process

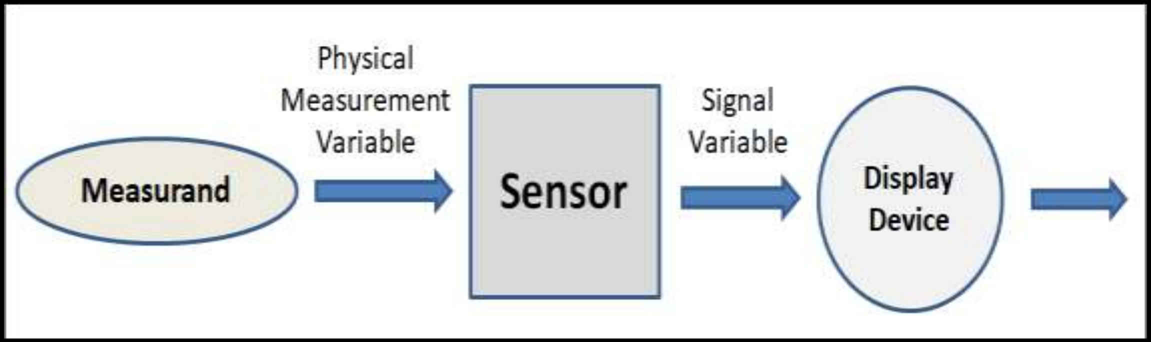
A sensor is the part of a measuring instrument which performs the measurement of a physical stimulus.

The *measuring instrument* is the whole device which is used to measure a physical quantity from the initial input stage to the final output.

A *measurement* is the act of physically quantifying a stimulus condition or event which occurs in the real world.

The *measurement value can be* based on a standardized unit of measurement, or based upon a dimensionless value.

Input



Output

Simple instrumentation model

To explain the fundamental steps in the measurement process using a measurement system and sensors, a simple *instrument model* is helpful for visualizing the process.

The main component of this instrument model is the sensor device.

Measurement process for the instrumentation model

1) Sensor input - the physical value or measurand (X) is observed by the sensor device

2) Sensor output - The sensor generates a signal variable (S) output which is normally electrical

3) Signal conditioning - The signal is transmitted and conditioned if needed (amplified, converted, filtered, etc.)

4) Signal received or rerouted - is received by an intermediate device (controller, processor, router), or output device (HMI, monitor, meter, dial, etc.)

5) Display of measurement – the measurement is then displayed by the output device

Uncertainty in measuring

The *uncertainty* of a measurement represents the random and systematic errors that occur within the measurement process.

This allows for a degree of confidence in the measurement, by evaluating the errors through methodically repeating measurements, to evaluate the accuracy and precision of the measuring system.

The measurement input (measurand vs stimulus)

The purpose of a sensor is to measure a particular stimulus. That measurement which is input into the measuring system is called a *measurand*.

When discussing the measurement process, the terms “stimulus” and “measurand” are often used interchangeably, as a sensor produces a “measurable” signal in response to a physical, chemical, or biological “stimulus”.

However, the subtle difference between the two is that the *stimulus* is the entity (event, property, or condition) being sensed or detected, (and then converted into an electrical transmission signal), while the term *measurand* places more of an emphasis on the quantitative aspect of an entity being measured.

The Signal Output of a Sensor

Signals

Signals are transmissions sent between devices in order to provide or receive data; which may be video, audio, or some other form of encoded data.

Signals are carriers, used to carry or transfer information between an input and output source.

Usually signals are transmitted through wiring, though they can also pass in the air through means such as radio frequency (RF) waves.

Classes of signals

Signals can be classified into the following categories:

- Continuous and discrete time
- Deterministic and non-deterministic
- Even and odd
- Periodic and aperiodic
- Energy and power
- Real and Imaginary

Continuous time and discrete time signals

Continuous signal - A signal is considered to be continuous, when it is defined for all instants of time.

Discrete signal - A signal is considered to be discrete, when it is defined at only discrete instants of time.

Deterministic and non-deterministic signals

Deterministic signal - A signal is considered *deterministic* if there is no uncertainty with regard to its value at any instant of time, or is a signal which can be defined precisely by a mathematical formula.

Non-deterministic (random) signal - A signal is said to be non-deterministic if there's uncertainty with regard to its value at given instant of time.

Non-deterministic signals are random in nature, thus are known as random signals. Random signals cannot be described by a mathematical equation, and are modelled in probabilistic terms.

Even and odd signals

Even signals - A signal is said to be even when it satisfies the condition $x(t) = x(-t)$

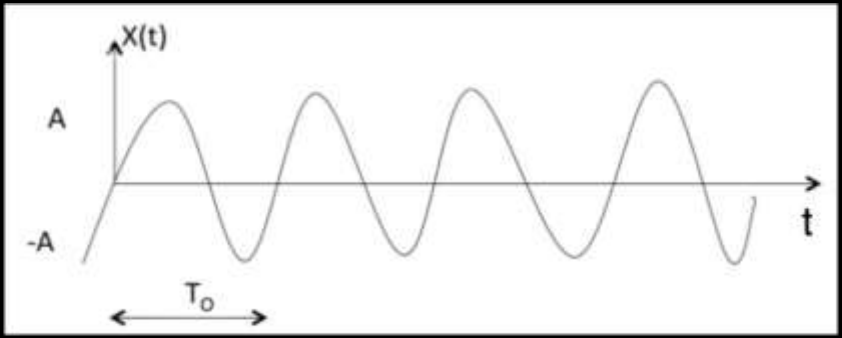
Odd signals - A signal is said to be odd when it satisfies the condition $x(t) = -x(-t)$

Periodic and aperiodic signals

A signal is said to be periodic if it satisfies the condition $x(t) = x(t + T)$ or $x(n) = x(n + N)$.

Where:

T = fundamental time period
$1/T = f$ = fundamental frequency



The signal shown in the above image will repeat for every time interval T_o (therefore it is periodic with the time period T_o .)

Energy and power signals

A signal is said to be an energy signal when it has finite energy.

A signal is said to be a power signal when it has finite power.

A signal cannot be both, energy and power simultaneously, and a signal may be neither energy nor power signal.

Power of energy signal = 0

Energy of power signal = ∞

Real and imaginary signals

A signal is said to be real when it satisfies the condition x(t) = x*(t)

A signal is said to be odd when it satisfies the condition x(t) = -x*(t)

Example:

If x(t)= 3 then x*(t)=3*=3 here x(t) is a real signal

If x(t)= 3j then x*(t)=3j* = -3j = -x(t) hence x(t) is a odd signal

Note: For a real signal, the imaginary part should be zero. Likewise, for an imaginary signal, the real part should be zero.

Types of electric signals and signal variables

Type of electrical signals include: Low level voltage, current, and frequency.

A few of the many examples of signal variables which are output from sensor devices are:

- Force
- Length
- Temperature
- Acceleration
- Velocity
- Pressure
- Frequency
- Capacity
- Resistance
- Time
- Voltage
- Displacement
- Current

Signals in a measurement system

Signals are the transmissions sent between a sensor and the receiving device. Usually signals are transmitted through analog or fiber optic wiring, though they can be transmitted thru wireless means such as radio frequency (RF) waves.

The signal variable can be manipulated during the transmission.

Digital signals are typically sent to a computer processor that can display, store, or transmit the data as output to another system, which will use the measurement.

Means of transmitting signals

There are different means in which a signal may be transmitted.

Types of signal transmission medium include:

- Mechanical
- Harmonics
- Light (fiber optics)
- Thermal
- Magnetic
- Electrical
- Chemical
- Radiation

What is signal modulation?

Modulation is the process through which coding, audio, video, image or text information is added to an electrical or optical *carrier signal* to be transmitted over a communication system.

Modulation enables the transfer of information on an electrical signal to a receiving device that demodulates the signal to extract the blended information.

Types of modulation

Modulation has three different types:

- Amplitude Modulation (AM): Where the amplitude of the carrier is modulated.
- Frequency Modulation (FM): Where the frequency of the carrier is modulated.
- Phase Modulation (PM): Where the phase of the carrier is modulated.

A modem is a common example/implementation of a modulation technique in which the data is modulated with electrical signals and transmitted over telephone lines. It is later demodulated to receive the data.

Use of modulation within the telecom industry

Modulation is primarily used in telecommunication technologies that require the transmission of data via electrical signals.

It is considered the backbone of data communication because it enables the use of electrical and optical signals as *information carriers*.

Modulation is achieved by altering the periodic waveform or the carrier. This includes carrying its amplitude, frequency and phase.

Waveforms

A waveform is a graphical representation of a signal in the form of a wave. It can be both sinusoidal as well as square shaped, depending on the type of wave generating input. The waveform depends on the properties that define the size and shape of the wave.

The input used to create a waveform determines its shape. Square shaped waves represent digital information in the form of a 0 or 1. However, a sinusoidal wave usually shows the exact variation that occurs in the input.

The mathematical function in a waveform

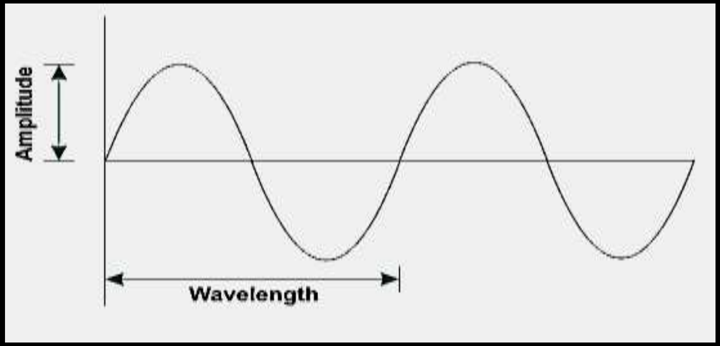
Waveforms follow a mathematical function that defines how they are represented and allowed to be interpreted by the reader.

For example:

Sinusoidal waveform - follows a trigonometric function that allows it to take its current shape.

Square waveform - follows a harmonic function.

Signal Processing

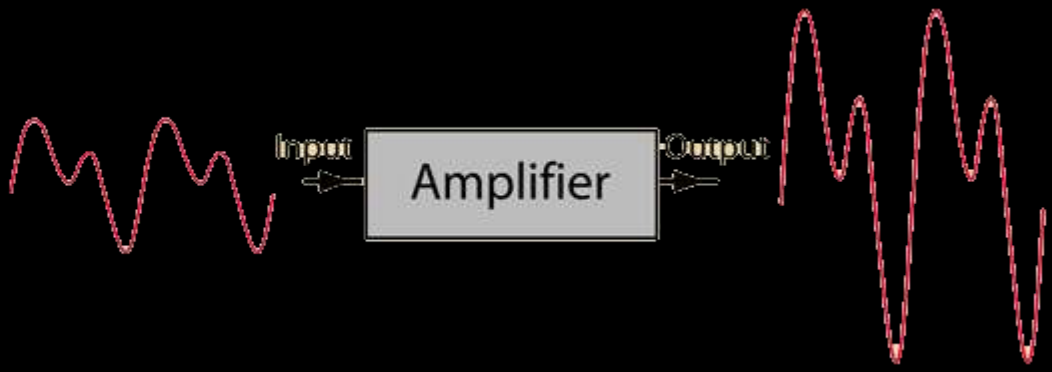


Input signal conditioning and processing

The output signal of a sensor does not typically have desirable characteristics for transmission, or other forms of processing. It may lack the sufficient amplitude, power, level, or bandwidth needed, or it may possess unwanted noise or other interference that distorts or masks the desired data.

Signal conditioning modifies the sensor signal to meet the requirements of the receiver to which it is connected.

The specifications of the signal conditioner used in a measurement system, are designed to meet the needs of the signal as well as the receiver.



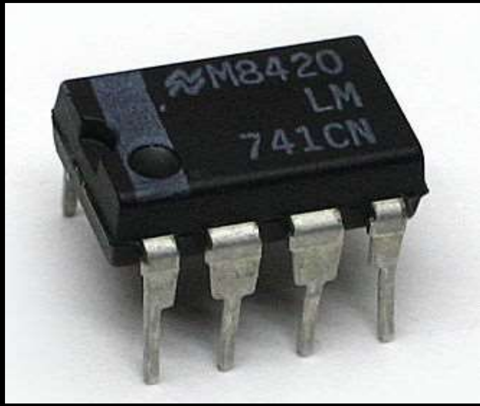
Signal amplification

If the signal output from the sensor is insufficient, it is often necessary to amplify the output signal, using signal amplifiers.

An *amplifier* uses the electric power from a power supply to increase the amplitude of the signal.

The degree of signal amplification provided is measured by its *gain*, which is the ratio of output power to input power.

An amplifier is a circuit that has a power gain which is greater than a value of one.



Op-Amps

Op-amps are one type of signal amplifying unit that is a direct coupled, high gain, electronic voltage amplifier with a differential input, which typically has a single-ended output.

This type of amplifier is popular due to its affordability and versatility, as well as its high gain characteristics, (with a gain which is typically hundreds of thousands of times larger than the potential difference between its input terminals.)

Signal filtering

Signals can be filtered to block or reduce the interfering or modifying inputs.

Some of the types of filtering used are:

- low pass- filters out the high frequencies and “passes” the low frequencies
- high pass - works in reverse, by filtering out low while passing high frequencies
- band pass - by combining low-pass and high-pass filters together, a *band pass filter* can be created that allows signals between two preset oscillation frequencies to pass
- band rejection - the *band reject filter* allows the passage of some frequencies, but rejects others. They are also called *band elimination, band stop, or notch filters*.

Analog to digital conversion

it is often necessary to convert a digital signal output to analog, or vice-versa, in order to properly interface with the receiving device. This signal conversion is performed by an analog to digital converter (or ADC).

There are also digital to analog converters used for the same purpose.

Other signal processing methods

Some of the other signal processing methods used on sensor signals include:

- Compression
- Modulation
- Scaling

Modifying and Interfering Inputs

Modifying and interfering inputs

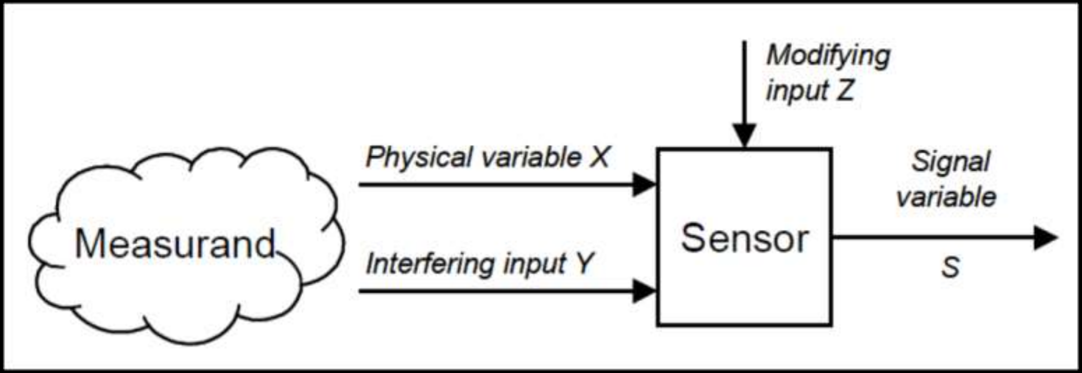
In some cases, a sensor's output is influenced by physical variables other than the intended measurand. Any input other than that of the desired measurand, are considered undesirable inputs or disturbances. These inputs are classified as being either *modifying or interfering*.

Modifying inputs change the behavior of the sensor or measurement system, thereby modifying the input/output relationship and calibration of the device.

A typical example of a modifying input is temperature. Because of this, many devices are calibrated at specific temperatures.

Other commonly occurring examples are:

- Atmospheric pressure
- Humidity
- Magnetic field interferences



Interfering inputs

Interfering Inputs represent an input to which a measuring instrument is unintentionally sensitive.

Methods used to correct a modifying or interfering input

There are several methods used to correct a modifying or interfering input:

The use of opposing inputs
With this method, additional interfering or modifying inputs are introduced to intentionally cancel out the undesirable effects of the other sources of interfering input.

The use of signal filtering
This method introduces elements into the measuring instrument to block out or minimize the interfering or modifying inputs, as mentioned previously.

The use of calculated output corrections
This method measures or estimates the magnitudes of the interfering or modifying inputs, and subtracts from the signal to obtain the desired output.

The use of high gain feedback
This refers to an amplification of the signal in open or closed loop systems.

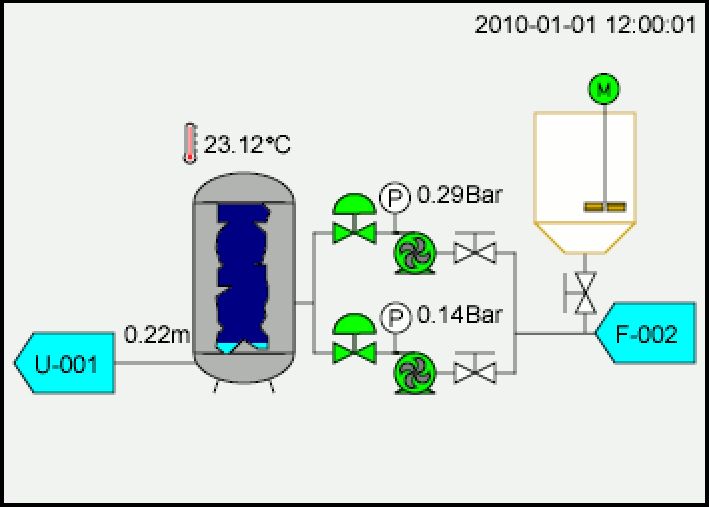
Measurement Output Devices

Signal output devices

The signal from the sensor can be output in a number of ways; as a display, recording, or as a secondary signal to another device or system.

There are many types of devices which are used for displaying an output, from an ordinary dial gage or scale, to a complex computer touchscreen monitor.

The signal can also be redirected to a larger network system such as a SCADA control system (image).



SCADA

SCADA stands for Supervisory Control and Data Acquisition. This is a control system architecture that uses computers, networked data communications and graphical user interfaces for high-level process supervisory management.

It also uses other peripheral devices such as programmable logic controller (PLC) and discrete PID controllers to interface with the process plant or machinery. Sensor devices are at the heart of this interface between machinery and the control network.

Human Machine Interface (HMI)

One type of output display device (or graphical user interface) is the HMI (or Human-Machine Interface).

This is basically a ruggedized monitor, which normally has a touchscreen style of interface. They are typically used to directly display a digital mimic diagram of the system being monitored or controlled.

An HMI monitor is found whenever there is some type of control process being monitored such as a power plant, industrial or automated process, an electric grid, and so on.

Digital readouts

Another type of display is the digital readout or counter. These are usually numerical display devices for outputting measurement quantities.

The display shown in the image to the right, is for outputting Cartesian and polar coordinates from computerized numerical control (CNC) machinery.

Meters

Similar to the readout, is the analog meter (image), digital meter, or smart meter.

Measurement systems can send an output as either an analog and digital type of signal.

Analog meters

One example of an analog meter is the common multimeter or “multitester”. An example is shown in the image to the right.

Though the analog multimeter is an older design, it is still preferred by many electricians and other technicians over a digital type.

The reason being that an analog meter has a higher degree of sensitivity to the subtle changes in the circuit, in comparison to the digital meter.

Digital meters

A digital multimeter samples the quantity being measured and then outputs the signal in discrete incremental steps, compared to the analog multimeter which provides a continuous output value.

Smart meters and AMI

To the right is an example of a digital-readout smart meter, used to monitor residential power consumption. Smart meters typically record energy consumption and then transmit the reading via two-way communication between the meter and the central system. These are used in AMI.

Advanced Metering Infrastructure (AMI) differs from Automatic Meter Reading (AMR) in that it has two-way communication between the meter and the power supplier.

Communications from the meter to the network may be wireless, or via fixed wired connections such as power line carrier (PLC).

Voice activated controllers

An example of an output control device is the voice activated controller; such as Amazon's Echo. These are becoming very popular for controlling a variety of smart home sensor devices.

They can be used as an intermediate means to send signals to and from a number of input/output devices such as thermostats, alarms, lighting, smoke/leak/gas detection, cameras, etc.

Programmable Logic Controllers (PLC)

Another output controller device is the PLC. This is a ruggedized, industrial computer processor which has been adapted to control a variety of manufacturing processes, such as assembly lines, robotic actuators, or other processes that demand a high level of reliability, ruggedness, and relative ease in programming.

Use of Transfer Functions in Measuring Systems

Ideal I/O

A perfectly ideal or theoretical input/output (stimulus/response) relationship exists for any sensing action.

If a measuring system were to have no losses, deviations or errors in the sensing and transmitting of the signal through to the output stage, then the output of such a sensor would always represent the *true value* of the stimulus.

Transfer function

This perfect or ideal input/output relationship can be expressed by use of a table, a graph, a mathematical formula, or as a solution of a mathematical equation.

If the input/output function is time invariant (doesn't change as a function of time) it is called a *transfer function*.

Mathematical representation

The transfer function is a mathematical function that represents the correlation between a stimulus (measurand), and the system response (electrical output signal of the sensor).

This correlation can be expressed as:

$$S = f\,(p)$$

where:

(S) is the electrical output, or signal variable

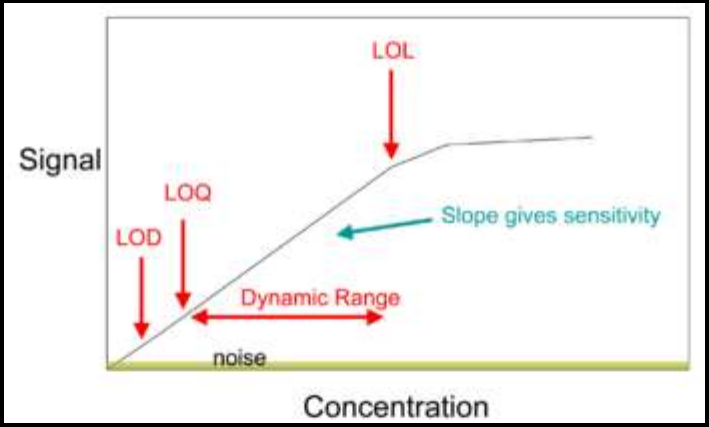
(p) is the stimulus or measurand parameter

Using the inverse function to compute the measurement

Normally, the value of the stimulus (p) is unknown, while the output signal (S) value is known from a sensor measurement.

The inverse function, $f^{-1}(S)$ of the transfer function is used to compute the stimulus (p) from the sensor's response (S).

Calibration of the Sensor

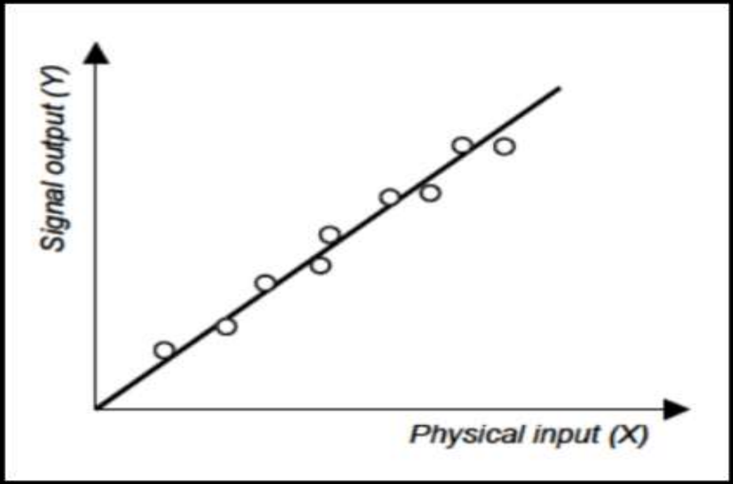


Sensor calibration

The *calibration* of a sensor device is the relationship between the physically measured variable (physical input) and the signal variable (signal output) for a specific sensor or measurement instrument.

A sensor or instrument system is calibrated by providing a known physical input to the system and then recording the output.

The recorded data is then plotted on a calibration curve (image), and the sensitivity of the device is determined by the slope of that curve.



Calibration curve

A calibration curve is a graphed curve or empirical table for a measuring instrument which measures some parameter indirectly, giving values for the desired quantity as a function of values of the sensor output.

The calibration curve is typically used when an instrument uses a sensor whose calibration tends to vary from one measurement to another, or varies with time or use.

If the output of the sensor did not vary, then the instrument would be marked directly from the measured unit.

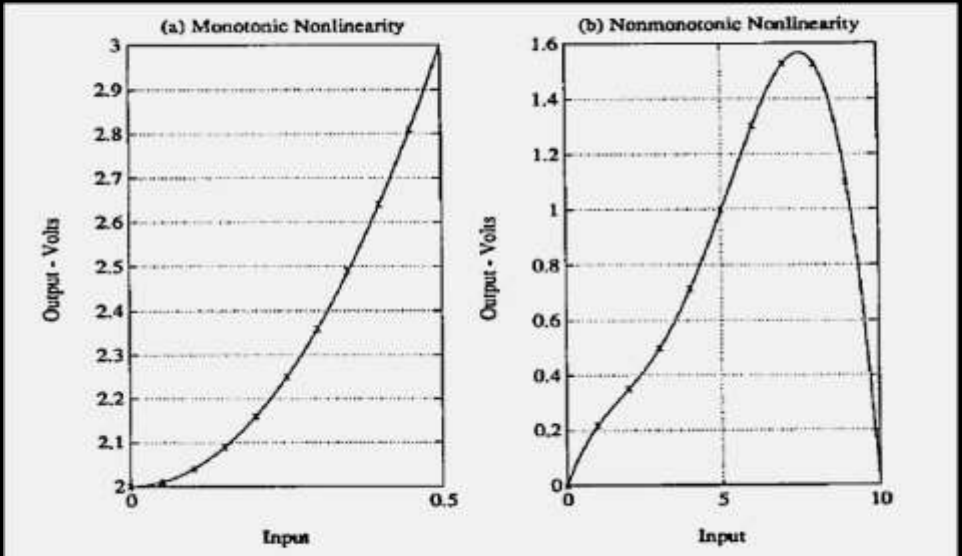


Figure 3: Nonlinear calibration curves
(a) Monotonic within operating range, (b) Nonmonotonic within operating range

Monotonic curve

With this type of curve, the dependent variable in the curve always increases or decreases as the independent variable increases.

The *monotonicit*y of a function's curve tells us if that function is increasing or decreasing. A function is increasing when its curve rises from left to right, and decreasing when it falls from left to right.

A non-monotonic curve consists of both rising and falling sections in the curve, as shown above.

Chapter 2: Classifying of Sensors

Sensor Classifications

Classifying sensor devices

Sensors can be classified by any number of distinguishing criteria, with some of the more common classes being those mentioned in this chapter.

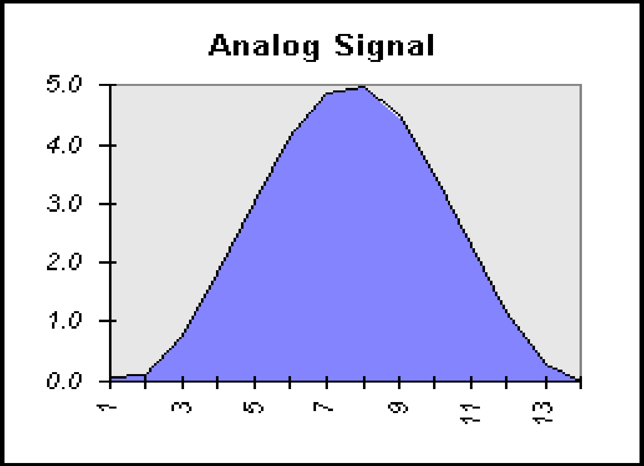
Sensor devices can be classified by the following criteria, and much more:

- signal characteristics
- power source (internal or external)
- referencing system used for locating
- type of stimulus being detected
- operational mode
- internal or external processing
- Cost and level of reliability
- Level of invasiveness
- Scale of the measurand (nano, micro, milli, macro)
- Degree of sensitivity, linearity, response/recovery time, etc.

Classified by signal characteristic: (analog vs digital)

One way to classify a sensor is by the characteristics of its signal output.

- If the signal varies in a continuous pattern, then it’s an *analog* type.
- If the signal varies in incremental steps, it is known as a discrete or *digita*l signal.



Analog sensor

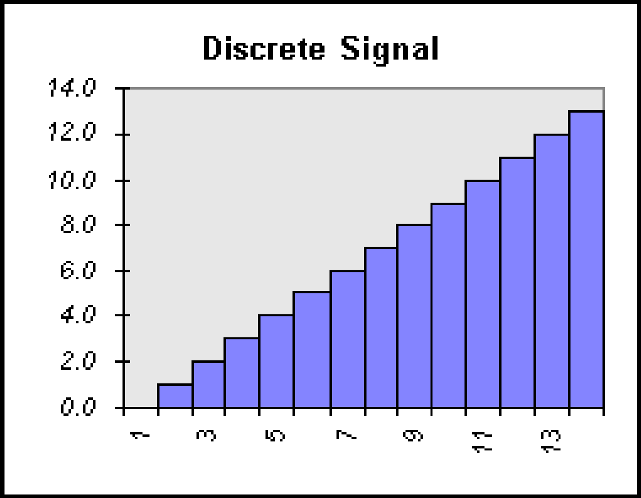
An analog type of sensor produces a non-digital signal, which is continuous over time and is always proportionate to the measured stimulus.

The output produced is in a form which is continuous throughout the measurement range, working in a similar fashion to that of old radio tuners; changing the input continuously instead of in incrementally defined steps.

To use a thermal sensor as an example, if the measuring capabilities of the sensor were between a range of 0 and 100 degrees Fahrenheit, then the measurement input/output could be any value between the minimum of zero degrees, and maximum of 100 degrees, such as 97.60F.

Analog to digital conversion

When using an analog sensor component with a microprocessor-based measuring system, it is necessary to have some type of analog to digital conversion capabilities in the signal conditioning process.



Digital sensor

A digital type of sensor produces a digital signal, with an output which is in predefined, incremental steps throughout the measurement range.

It gives a discrete signal output, in the form of a pulse.

Digital sensors have a signal that is a direct digital representation of the measured stimulus, and is binary (either on or off).

Incremental output

To measure a value in between 0 and 100 degree, a digital thermal sensor would give a rounded off output value in 1 degree increments, 0.1 degree increments, or whichever predefined incremental value, which is specified for that sensing device, by the device's manufacturer.

Classified based on the Mode of Operation (Null or Deflection)

Null or deflection types of sensors

A sensor device can be classified based on the means of physical operation for the measuring instrumentation.

The two modes of operation are:

- the *null* type (using the opposing or balancing force principle)
- the *deflection* type (using a mechanical deflection mechanism)

The "Null" mode of operation

With this method of operation, the measuring instrument exerts an influence to oppose and balance out the effect of the measurand.

The quantity of the influencing effect and the measurand are balanced until they are equal but opposite in value, resulting in a null measurement.

Example: the pan balance

An example of a null type of measuring device is the pan balance weighting scale.

With this scale, a counterweight is placed on one side, with the measurand on the other side.

Either the counterweights or the measured item are adjusted until the weight scale reaches zero or a “null” value.

In the null-type of sensor device, the physical effect caused by the measured quantity is “nullified” by the transducer generating an equal and opposing effect (a zero or null indication).

Pros and cons of null types of sensors

Null sensor measurements are more accurate and sensitive, and don’t interfere with the state of the quantity being measured.

However, the downside is that they have a poor response to dynamic variations and are slow in operation.

The “deflection” mode of operation

In the deflection-type of measuring device, a mechanical “deflection” which is proportional to the magnitude of the measured quantity is produced.

A common type of deflection mechanism is a needle moving on a meter which provides the respective quantity reading.

Examples include metered devices such as voltmeters, or a spring-loaded hanging scale (see image).

Pros and cons of the deflection type

These types are simple in design and operation, and provide a good dynamic response (thus are better suited for dynamic measuring).

However, they can interfere with the state of the measurand; thus are unable to determine highly precise values, states, and conditions of the measured quantity.

Classified by Power Source (Active and Passive)

Passive (or self-generating sensors)

This type of sensor doesn’t require an external power source to function, with the input stimulus energy being directly converted by the sensor into the output signal.

Generating power internally to operate, they are called a “self-generating” type, with the energy to operate coming from the physical quantity which they measure.

An example is the piezoelectric crystal, which generates an electrical charge, (using the piezoelectric effect) when subjected to acceleration.

Thermocouples and photodiodes are other types of passive sensors.

Active (or modulating sensors)

This type of sensor requires an external excitation signal or power source to function. The excitation signal (electrical signal which supplies the power to sensor) is modified by the sensor to produce the output signal.

An example is a piezo-resistive accelerometer.

In addition to piezoresistivity, some of the other types of modulating signal energy sources are: photoconductive, magnetoresistive, thermoresistive, and electrical conductivity.

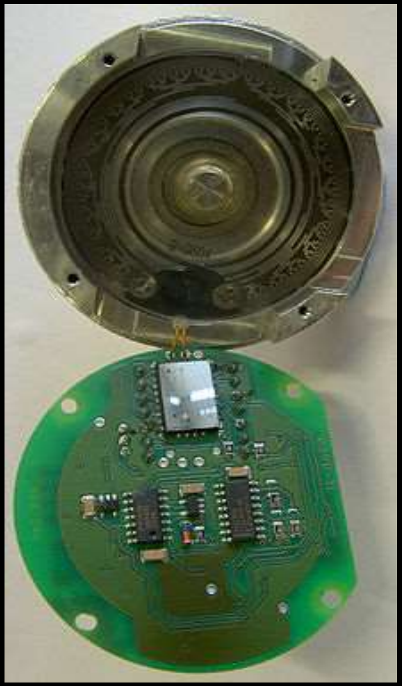
Classified by Referencing Scale (Absolute or Relative)

Positional measuring

Sensors such as the displacement, distance, or positional types need a means of numerically referencing their location in relation to either an origin location or other points of reference.

Referencing can be based on scales like a Cartesian coordinate system (linear motion) or based on polar (rotary motion) coordinates.

One common device which converts the angular position of a shaft or axle into an analog or digital signal, is the rotary encoder (see image to the right).



Absolute sensors

This type of sensor works off of an absolute scale, which is a system of measurement that begins at a minimum, or zero point of reference, and progresses in a single direction.

An absolute scale begins at a minimum value, leaving only one direction in which to progress, and can only be applied to measurements in which a true minimum is known to exist.

Relative or incremental sensors

This type of sensor works off of an arbitrary, or "relative," scale, which begins at some arbitrarily selected point of reference, and can progress in either direction on the scale, relative to that selected reference point.

Classified by the Type of Stimulus Energy

Classification based on stimulus type

Some sensors are classified based on their form of energy.

Examples are:

- Electrical (energy of electron flow)
- Chemical (energy based on chemical reaction)
- Mechanical (kinetic energy of motion)
- Magnetic (energy of repulsion and attraction)
- Thermal (energy of thermal flux)
- Radiant (energy from waves of light, sound, ion radiation, RF, microwave)
- Potential (energy from gravitational position, or elastic tension)

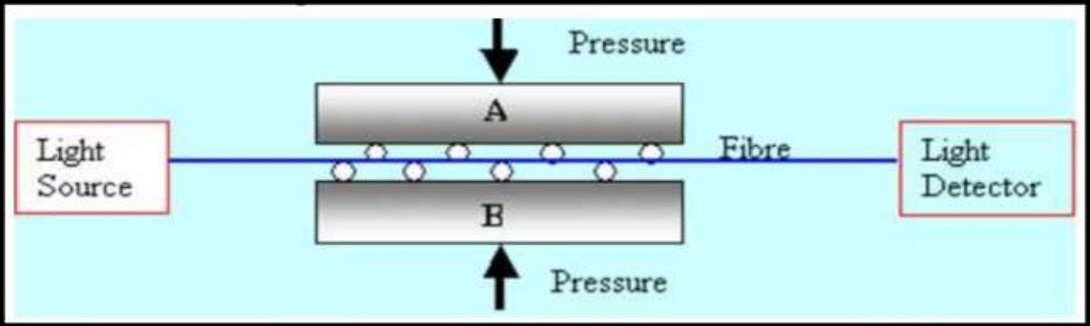
Intrinsic and Extrinsic Types

Fiber optic sensor types

With fiber optic sensors, subtle variations in the light beam are sensed in order to provide the measurement readings.

In this type of sensor, the light rays are manipulated either outside or inside of the fiber optic cable.

Because of this, these sensors would be classified as being either an *intrinsic or extrinsic* type of sensor.

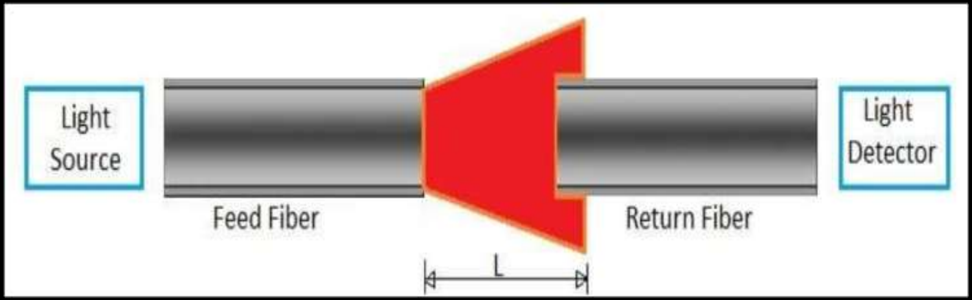


Intrinsic

With this type of sensor, the light beam is manipulated within the cable, prior to reaching the optical detector.

This type is more expensive, easier to connect, and is more sensitive than the extrinsic type.

However, they are more difficult to multiplex (splitting the beam into multiple, simultaneously transmitted signals), and require more complex signal de-modulation (this is the recovering of transmitted data from the modulated carrier wave).



Extrinsic

When the light beam exits the fiber cable and is manipulated prior to reaching the optical detector end, it is known as an extrinsic type of sensor.

The distance L (seen in the image above) can be measured with this type of optical sensor.

This type is more affordable, is more versatile, and easier to multiplex. It does however have issues with connections at entrance and exit points in the fiber cable.

Other Sensor Classifications

Other sensor classifications

There are so many other ways that a sensor can be classified, that it's impossible to list them all.

For example:

Sensors classified based on the environment which it is installed:

- Is it in a climate controlled, extreme hot or extreme cold setting?
- Is it for use in a sterile sanitized clean room, or in a hostile atmosphere?
- Is there a strong presence or dust or corrosive contaminants in the environment?
- How much invisible contamination is there such as radio frequencies, radiation, VOC's, gas and others?

Based on application or industry

Sensors can be classified based on their particular application:
The number of potential applications are endless for sensing devices.

For example - motion, security, proximity, temperature, gas, turbidity, viscosity, density, displacement, speed, distance, and so on.

Sensors can be classified based on their industrial application as well:

- Aerospace
- Smarthome
- Security and Alarm
- Manufacturing
- Water Resources
- HVAC
- Etc.

Based on economy, scale, characteristics, invasiveness

Or sensor selection can be based other classifying parameters such as:

- Cost and level of reliability
- Scale of the measurand (nano, micro, milli, macro)
- Degree of sensitivity, linearity, response/recovery time, etc
- Level of invasiveness (amount of disruption of the measured process)

Chapter 3: Sensor Characteristics

Static and dynamic characteristics

A measurement system consists of two types of the characteristics: *static and dynamic*.

When a measurand is time invariable (doesn't change over time), it is considered to be in a static (stationary and non-changing) state with static characteristics.

When the measurand is time variable (changes over time), it is considered to be in a transient (in motion and constantly changing) or dynamic state, possessing dynamic characteristics.

Static characteristics

These are the properties of the measurement system which do not vary with time.

They’re the characteristics which exist when the dynamic effects cease, and the system reaches a steady state.

Static characteristics

The most important static characteristics include:

- Accuracy and Precision
- Sensitivity
- Linearity (and Non-linearity)
- Drift
- Calibration Curve
- Repeatability, Reproducibility, and Resolution
- Range (Measurement or Operational)
- Minimum Detectable Signal (MDS)
- Selectivity
- Response, Recovery, and Settling Times
- Stability and Specificity
- Hysteresis and Backlash
- Offset and Bias
- Dead Band or Dead Space

Dynamic characteristics

These are properties of a measurement system which do vary with time. Dynamic characteristics describe a sensor's transient (varying) properties and how it responds to changes in the measurand.

The reason for dynamic characteristics is the presence of energy- storing elements, such as masses and springs, or inductive and capacitive elements.

Sensor response to a transient stimulus

The way that a sensing system responds to a measurand in a transient changing state (dynamic state), differs greatly than when it is exposed to a measurand that is in a static, non-changing state.

When quantifying a measurand with transient properties, dynamic characteristics are used to describe the sensing system's transient condition, with a mathematical model (a transfer function) which is used to derive the relationship between the time varying input and output signals.

These mathematical models can be used to analyze the response to variable input waveforms such as: ramp, impulse, step, sinusoidal, or white noise signals.

Effect of energy storing elements* in the system

Dynamic characteristics occur in a signal output due to energy-storing elements in a sensing system.

**These transient properties are caused by:*

- Electronic elements - inductance and capacitance
- Mechanical elements - vibration paths and mass
- Thermal elements – those with heat storing capacity

Zero order systems

These are specified by a time independent transfer function, and doesn't incorporate energy storing elements into the equation.

With a zero order system, the sensor response is instantaneous.

An example is a potentiometer used to measure linear or rotary displacement. This doesn't apply to quick varying displacement situations.

1st order systems

The response of this system is more complicated than that of a zero order system. It incorporates one energy storing element, and one energy dissipating element.

The transfer function is specified by a first order differential equation.

An example of a first order system is a temperature sensor which has a stored thermal capacity.

2nd order systems

The response of a second order system is more complex than that of zero or first order systems. It has two elements in the system which store energy.

The transfer function is specified by a second order differential equation.

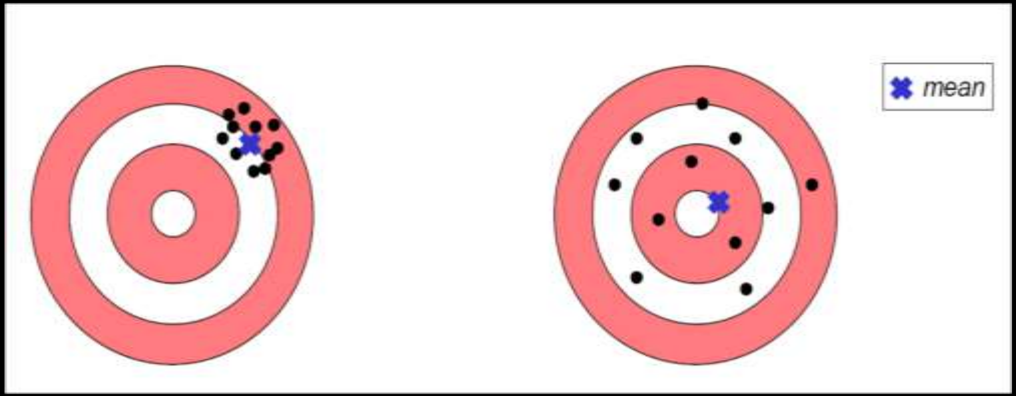
Accuracy and the ideal value

The degree of "inaccuracy" for an instrument is the difference between the true (ideal) value of the measurand and the measured value as indicated by the instrument.

It is a measure of the quality of the data when compared with a recognized standard.

True value is usually defined in reference to some absolute or standardized value; for any measurement there will be some amount of *systematic (bias) and random (noise)* error which will occur which prevent the measurement from being a perfect and ideal value.

Accuracy is how closely a measuring instrument provides results to the "true value" of the measured quantity.



Precision (vs Accuracy)

Precision is the ability of the measuring system to provide the same reading, with repetitive measurements of the identical quantity under identical conditions.

Precision differs from accuracy in that it refers to obtaining similar measured values in a series of successive readings; not how close the values are to the ideal or true value.

The image above illustrates the differences in the precision vs the accuracy of a sensor.

The target's bullseye represents the ideal or true value measurement. The mean value on the left shows a more precise group of measurements, while the mean value on the right is for a more accurate group of measurements.

Improving accuracy with estimation techniques

It is often possible to improve the accuracy of a poor quality sensor device through the use of computerized estimating techniques.

Estimation methods can be simple such as averaging or low-pass filtering to cancel out random fluctuating errors.

Filtering methods

More sophisticated estimating and filtering methods can also be used:

Wiener filtering - computes a statistical estimate of an unknown signal using a related signal as an input and filtering that known signal to produce the estimate as an output.

Kalman filtering - is an algorithm that uses a series of measurements observed over time, containing statistical noise and other inaccuracies, which produces an estimate of unknown variables that tend to be more accurate than those based on a single measurement alone.

Better estimation through better computing!

In many applications, thanks to ever-increasing computing capabilities and the ever-lowering costs of computer devices, it is becoming more desirable to use lower performing sensors, which are enhanced by these sophisticated estimation techniques.

Repeatability

Two sensor characteristics which are closely related to precision are repeatability and reproducibility.

Repeatability - This is the precision of a series of measurements which is taken over a short time interval. It is the ability of a measurement system to repeat its readings under the same set of external parameters.

It is directly related to the sensor characteristic of accuracy. However, a sensor can be inaccurate, though still be repeatable in its observations.

Reproducibility

This is the degree of precision for a series of measurements. However, the measurements are taken over a longer period of time or performed by different operators, instrumentation, or lab locations.

It is the system's ability to produce identical responses after the measurement conditions have been changed; using a different set of external parameters.

Resolution (or Discrimination)

Another characteristic which is closely related to precision is the sensor’s resolution (or discrimination).

The *resolution* of a sensor is the smallest change that can be detected by a given sensor device. The resolution is related to the precision with which the measurement is made, but they are not one and the same.

A sensor's accuracy may much worse than its resolution. Additionally, resolution is limited by any noise occurring in the signal.

(MDS) Minimum Detectable Signal

In a sensing system, when all interfering and modifying factors have been accounted for, the minimum detectable signal (MDS) is the minimum signal increment that can be still be observed.

When this increment is assessed from the value of zero, it is referred to as the threshold or detection limit.

Threshold

This is when the resolution increment is measured from zero. It is the minimal value of input below which the sensor does not have a signal output.

When the increment is gradually increased from the value of zero, initially there is no output.

As the input exceeds the value of the sensor's threshold, a signal output begins. The threshold of a process is the minimum requirement for that process.

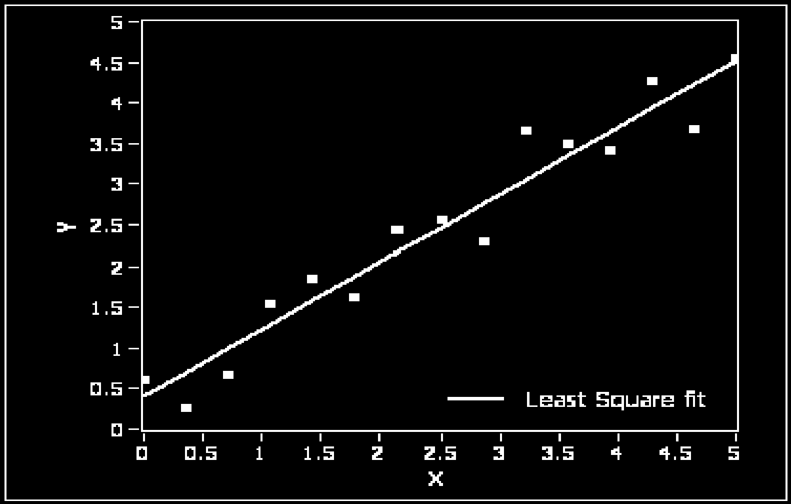
Specificity

In an ideal scenario, this is a sensor's ability to solely identify or target a singularly "specific" type of measurand or analyte (chemical measurand), while completely disregarding all others.

Selectivity

Selectivity is the sensor's ability to measure a "targeted" measurand, even though it is in the presence of other measurand which may interfere with, or influence the reading.

For example, if a carbon monoxide gas sensor were able to provide an accurate and consistent reading of that gas type, even though it is mixed with other gas types such as carbon dioxide or nitrogen oxide, then it would be considered a highly "selective" sensing device.



Linearity

This is an indication of how close the calibration curve is to a specified straight line (based on the least squares fit or best fit of the various measurement readings to the theoretical line).

Linearity is expressed as the percentage of departure from the linear value (or the nonlinearity), meaning the maximum deviation of the output curve from the best fit or “ideal” straight line for a given calibration cycle.

Nonlinearity

Linearity is actually specified as the percentage of nonlinearity, which is derived from the following ratio:

Nonlinearity % = (maximum input deviation) / (maximum, full-scale input)

This static state of nonlinearity is often due to environmental conditions, which include factors such as humidity, temperature, vibration, mechanical hysteresis, electronic amplification, or acoustic noise levels.

A state of ideal linearity is never totally met with a sensor, and these deviations from the ideal are known as *linearity tolerances*.

Dynamic Linearity

Normally, the static curve is used for determining linearity, and may deviate from the dynamic linearity.

Dynamic linearity is an indication of a sensor's ability to follow rapid changes in the input.

Characteristics such as amplitude distortion, phase distortion, and response time are important in determining the dynamic linearity of a sensor.

Sensitivity

This is the ratio of the incremental change of the output signal of the sensor, over the incremental per unit change in the input signal (the entity being measured).

Sensitivity can be derived from this ratio:

Sensitivity = change in output signal / change in input signal

Higher sensitivity means a higher quality reading

Sensitivity is an indication of how large the response of the sensor may be, compared to a small variation in input.

This means that a sensor with a higher sensitivity, will provide a better quality of measurement reading.

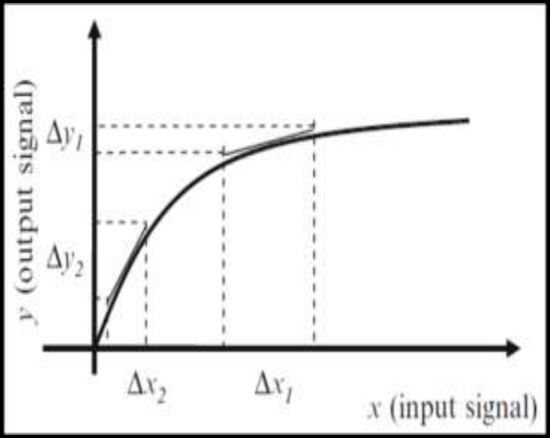
Operating range should be appropriate for the application

For the sensitivity of a sensing device to be high, it should not have an operating range that greatly exceeds the value typically being measured.

For example, a thermostat sensor which measures the temperature of a residential climate should not have a specified operational range from 0F to 212F.

Sensitivity can be calculated based on the slope of the calibration curve.

Also, Sensitivity can either be constant throughout the range of the sensor (when the calibration curve is linear), or it can vary (when nonlinear).



Varying sensitivity

As seen in the image, with nonlinear curves the sensitivity can differ throughout the range, based on the location on the calibration curve.

The ideal sensor would have a high and consistent sensitivity in its operating range.

In the image shown, the sensor eventually reaches the point of saturation, where it no longer responded (provides an output signal) to a change in input.

An example of sensitivity may be that of a temperature sensor.

If the output voltage were to increase by 1V, as the temperature rose by 0.2 F, then the sensitivity will be 5V per degree F (or 5V/1F).

Operating range

This is the minimum to maximum range of values over which a sensor works properly.

Sensors may still perform beyond this range depending on the situation and equipment; however additional calibration may be needed.

However, when trying to operate a sensor beyond this range, it may not perform correctly.

Typical errors might include:

- a constant signal output at the maximum value
- changes in sensitivity
- generating of large inaccuracies
- the irreversibly damaging of the sensor

Usually the measurement range of a sensor device is specified by the manufacturer.

Span and full scale output

Span is a dynamic range of stimuli that may be converted by a sensor. The span or *full scale input*, represents the highest possible input value, which can be applied to the sensor without causing unacceptably large inaccuracy in the output.

The *full-scale output* is the algebraic difference between the output end points, which are typically zero and full scale.

Sensor Adjustment Times

Settling time

This is the amount of time that a sensor requires for reaching a stable output once it's been energized.

In situations such as when a measuring instrument is powered down to conserve energy, an accurate settling time will need to be determined for the sensor, and incorporated into the operational procedures for that sensor device.

Adjustment time (settling, response, recovery)

Sensors require a period of *adjustment time* in order to provide accurate and reliable readings.

They first need a period of time to "power up" (settling), then to respond (response), and then to recover (recovery).

Response time

Response time is defined as the time between when a sensor is first exposed to a measurand, and the time it requires to reach a stable value.

It is typically expressed as the time when the output reaches a certain specified percentage, such as 90% of its final value, in response to a step change of the input.

Recovery time

Recovery time is defined conversely from response time, as the time needed for a sensor to return to its baseline value after the step removal of the measured variable.

This is normally the time required to return to a specified percentage of its final value after step removal of a measured variable.

The typical amount specified as the allowable percentage for recovery time is 10%.

Chapter 4: Sensor Deviations

What are Sensor Deviations?

Sensor measurement deviations

A sensor deviation is a measurement error which is considered as the difference between the actual measured output value which is read by a sensor, and the true value of the measurand.

For example if a temperature being read is 55.0 degrees, and the sensor output is 55.5 degrees, then there is a 0.5 degree deviation.

Sources of deviations

As it is not possible for sensors to provide a perfect representation of a measurand, several types of deviations typically occur which limit the accuracy of the measurement.

These deviations can be caused by a wide range of circumstances such as: noise, hysteresis, drift, dynamic or offset errors, non-linearity, sensitivity errors, etc.

Systematic or random errors

All sensor deviations can be classified as either a systematic error or random error.

Systematic errors - can sometimes be compensated for by means of some form of calibration.

Random errors – such as noise can be reduced by signal processing, such as filtering, usually at the expense of the dynamic behavior of the sensor.

Systematic errors

Systematic errors (or bias) are repeatable errors which originate from the same specified source.

They can be corrected through adjustment, calibration, or compensation. They can be minimized by monitoring measurements, and routinely checking against a standardized measurement.

Systematic errors can often be compensated for, by implementing some form of calibration strategy or by using methods such as feedback or filtering.

Examples of a systematic error would be a scale which reads a weight when it is empty, or a timepiece which is a minute fast.

Random errors

With random errors, when a measurement is repeated, it will generally provide a measured value that differs from the previous value taken.

It is random in that the next measured value cannot be accurately predicted based on previous such values.

Correcting a random error

Random errors such as noise, that can be minimized through signal conditioning processes, such as filtering. However, this filtering usually has a negative effect on the dynamic behavior of the sensor.

Since it is not possible to predict a random error, it is also not possible to make an adjustment for the effect. A random error is an inaccuracy which is non-repeatable, and is caused by an unknown or an uncontrollable influence.

Random errors have an effect on the degree of precision of a measurement, not the accuracy. A random error is a short-term scattering of values around a mean value, which cannot be corrected on an individual measurement basis.

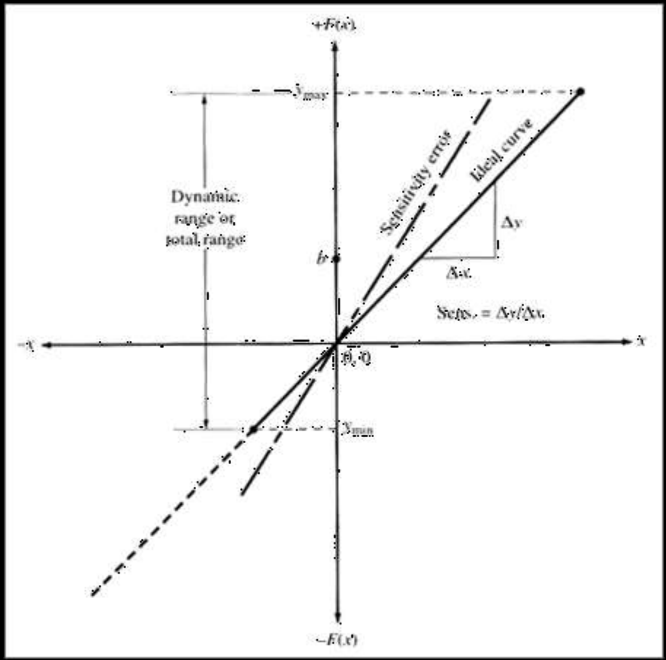
Types of Sensor Deviations

Sensitivity errors

With this error, the sensitivity may differ from the value specified, even though the output is still linear.

Also, the sensor may be sensitive to properties other than that of the measurand.

A good example of this type of deviation is when sensors which are influenced by the temperature of their environment.



Operational range error

Because the operational range of the output signal is always finite, due to limitations from the design of the sensor, the output signal will eventually reach and exceed the minimum or maximum limits of the manufacturer’s specifications, causing reading inaccuracies.

Non-linearity error

This occurs when the sensitivity is not constant over the operational range of the sensor.

This is normally defined as the percentage amount that the output varies from the ideal behavior of the sensor, throughout the full operational range.

Due to dead zones

This is the condition of a non-responsive system within a specific range of input signals.

In this range, the sensor will provide a null response over an entire zone of “dead space”, or dead band. In other words, it is the range of input readings where there is no change in output (no response).

Backlash in gears is a typical cause of this type of dead space error.

Due to saturation

This is when a sensor has reached its limits of operation; thus the sensor cannot be used for measurements beyond this saturation value.

This is when the calibration curve becomes less sensitive, reaching a limiting value for the output signal.

Dynamic error

An error which occurs when there is a deviation that is caused due to a rapid change in the measured property over time.

Approximation error

This is when a sensor has a digital output which is an approximation of the measured property.

This approximation error is also known as a *digitization error*.

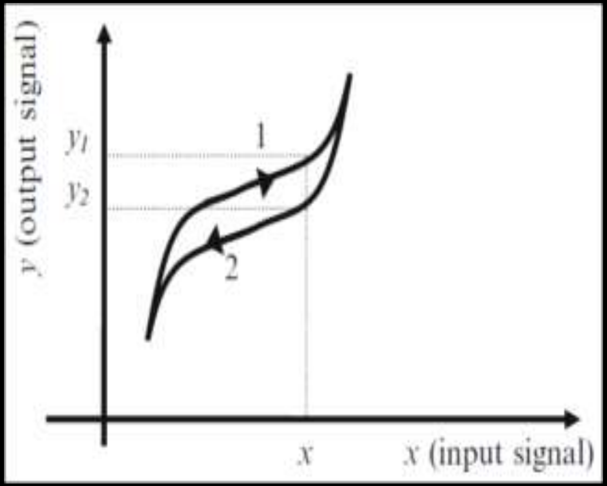
Due to Hysteresis

This error condition occurs when the measured property reverses direction, causing a lag in time before the sensor can respond.

This creates an offset error which differs from one direction to the other.

This is a difference in output values between the loading and unloading (forward or reverse) cycles in a measurement pathway.

It’s the difference between output signals for the same measurand, depending on the trajectory or path followed by the sensor device.

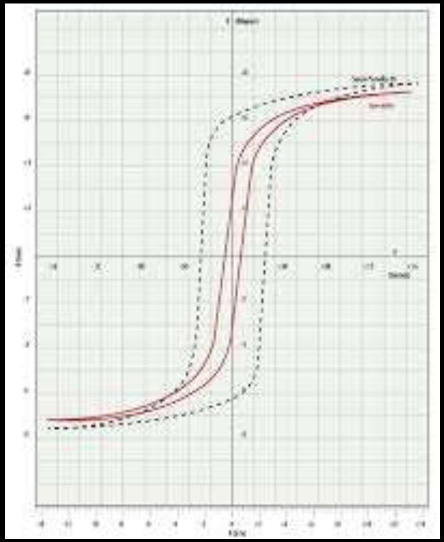


The Hysteresis Loop

The image to the right shows the relation between the output and input of a system which is experiencing hysteresis.

This looping pattern is typical for this type of a system.

Depending on whether the first trajectory or the second is taken, two different outputs for the same input would be displayed.



Hysteresis from magnetism or flex

Hysteresis is normally experienced in ferromagnetic materials such as iron (due to magnetization), or in types of "elastic behavior" materials like rubber, springs etc.

For example, a stretched rubber band would experience differing feedbacks between the stretching and releasing actions, or a spring-loaded device between loading and unloading.

Backlash

Backlash is a particular type of hysteresis in sensor output which is caused by the looseness (the slack or slop) in a mechanical connection.

It is considered the maximum distance or angle that a mechanical system can be moved in one direction without experiencing “slack” in the mechanical part.

Backlash is a mechanical form of a “deadband” condition, as described earlier.

One example of a backlash situation, is experienced when sensor devices are installed in a power train.

The backlash is a result of the gears in the power transmission system which, when reversed, exhibit slack before reengaging.

Pros and cons of backlash

Backlash is an unavoidable condition with most types of reversing mechanical coupling situations, although it's usually possible to mitigate its effects.

Backlash is not always an undesirable effect, as it allows for compensation for the jamming up of meshing parts.

It also aids in lubricating to prevent friction spots, as well as compensation for manufacturing errors, deflection under load, or thermal expansive issues.

Due to sensor drift

Sensor *drift* is a condition where the output signal slowly changes, independently of the measurand value.

Drift is noticed when there is a gradual change in the sensor output, while the measurand continues to remain a constant.

Drift is considered to be a systematic error, which originates from causes such as: mechanical or temperature instabilities, or environmental contaminants.

However the primary reason behind drift is the tendency of a sensor's performance to naturally degrade with time, due to aging of its components.

Drift is inevitable

Nearly all sensors, regardless of the cost and manufacturer's quality standards, are susceptible to sensor drift over time.

Sensor drift from the gradual degradation of the sensor and other components can make readings offset from their original calibrated state.

Factors which contribute to drift

- Expansion and contraction when subjected to cyclic pressure or extreme temperature variations
- The gradual degradation of sensor materials that occur naturally with time
- Being used in extreme environmental conditions, such as high and low temperature and humidity
- Use in environments with high levels of airborne particulate matter
- Exposure to VOC vapors or other highly corrosive agents
- Excessive vibration, impacting or shock over time (cyclic fatiguing)
- Adjacency to other high (or low) temperature equipment and components

Drift and the sensor’s baseline

Baseline is defined as the output value (which is typically zero) of the sensor, when it is not exposed to a stimulus.

Commonly, drift is assessed with respect to the sensor’s baseline. For a sensor which has no drift, the baseline should remain a constant

Manufacturer sensor testing and calibration

Typically, manufacturers will test and calibrate their sensor products in closed environments to achieve the desired specifications and a zero point.

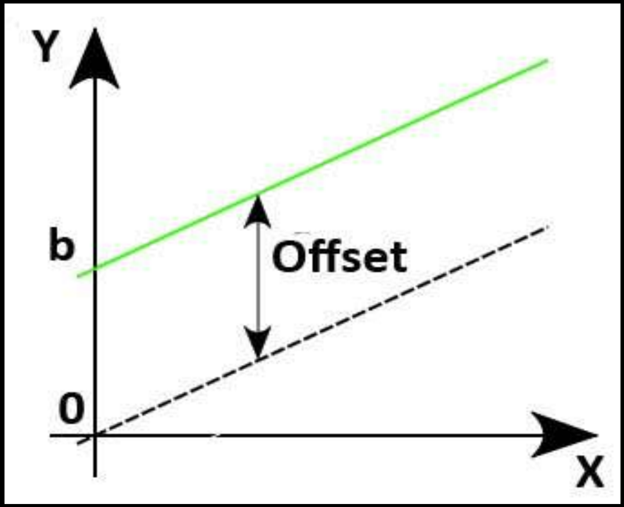
Some manufacturers provide a value for the expected drift or long term stability.

Even so, these specifications are based on the product being used in a highly controlled and stable environment, making the specs irrelevant under normal conditions.

Correcting for Drift - Factory or field re-calibrating can be used to correct for drift, but may not always be required.

Ideal output (No offset)

An ideal output value would mean that the calibration curve intercepts the y-axis (representing the signal output of the sensor) at a value of zero, when the curve intercepts the x-axis (input) at zero, meaning there is zero input occurring from the sensor.



Offset

When there is an offset error, the sensor has an output which differs from zero, even though the quantity being measured is actually zero.

An offset means that the sensor output is higher or lower than the ideal output.

With an offset error condition, the curve has the same slope as the ideal but crosses the Y-axis (output) at b instead of zero (see image).

An offset can have a negative output value as well, with the y-intercept occurring below zero.

Offset due to an external influence

Another form of offset is when there is a difference between the actual output value and the specified output value under a particular set of conditions, such as an offset value due to temperature.

Offset errors are easily corrected by calibrating the sensor.

Noise

Noise is the unwanted fluctuations of the signal output of a sensor, which does not convey information which pertains to the measurand.

When noise exists in the output signal, the readings are inaccurate and misleading.

Although, when an instrument has a high enough signal to noise ratio (SNR), the noise in the output signal will be negligible.

Signal to noise ratio

SNR is defined by the following equation:

Signal to Noise Ratio =
Mean value of the Signal / Standard Deviation of Noise

External noise

Noise can originate from either internal or external sources.

Examples of external noise, which can cause systematic errors, are from electromagnetic signal sources such as those produced by:

- transmission circuitry
- reception circuitry
- power supplies
- mechanical vibrations
- ambient temperature variations

Internal noise

One example of internal noise is that of “electronic noise”.

A common example of electronic noise in electronic instrumentation is caused by the thermal agitation of charge carriers, which is called “thermal noise”.

This thermal agitation causes charge carriers to move in erratic patterns, resulting in random variations in the voltage and/or current.

Thermal noise can still exist even when no current exists.

“Shot” noise

Another type of noise is known as “shot” noise.

Random fluctuations, caused by the signal carriers’ random time of arrival, will produce this “shot noise”.

Shot noise occurs randomly on the quantum level, with signal carriers being electrons, holes, photons, and phonons.

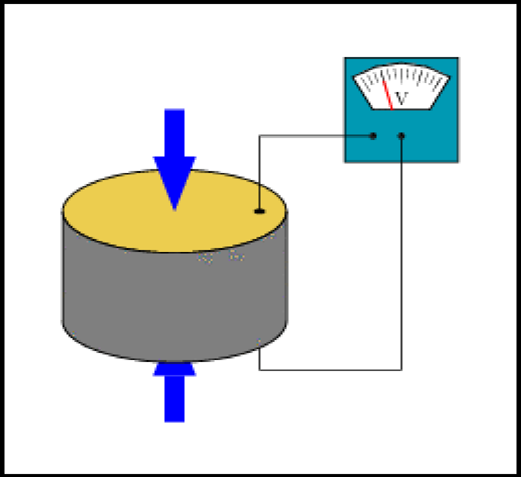
Chapter 5: Physical Phenomena used in Sensing

Physical Phenomena detectable by Sensors

Phenomena used in sensor transduction

There are many physical stimuli that are used by sensor devices. This chapter lists some of the more commonly used physical effects, phenomena, or physical laws.

Some of the most common transduction techniques used to quantify measurand are as follows.



Stress induced (Piezoelectric) transduction

The "Piezoelectric Effect" is a phenomenon which is exhibited in certain materials, with the ability to produce an electrical charge when a mechanical stress is applied to the material.

A piezoelectric material consists of a crystalline structured mineral that is of a latticed configuration of electric dipoles, which when combined, forms a polarized field.

When this polarized field is subjected to a mechanical stress, the polarization of the field is changed, either through a physical reconfiguration of the dipole lattice or by a re-orientation of the dipole moments under stress.

Inverse piezoelectric effect

A unique characteristic of the piezoelectric effect is that it can be reversed.

This means that it's able to generate electricity when stress is applied or conversely, can generate a mechanical response when an electrical field is applied; known as the "inverse piezoelectric effect".

The piezoelectric effect and its inverse

The piezoelectric effect and its inverse are proportional to one another, while the voltage generated by the piezoelectric effect is proportional to the amount of force applied and the type of force applied.

For example, tension and compression produce opposite polarities to one another.

In addition, the mechanical response of the inverse piezoelectric effect is proportional to the field's polarity and strength.

The piezoelectric effect is passive, and requires no additional power other than the mechanical or electric stimulus.

The “Piezoresistivity Effect”

This phenomenon is a change in the electrical resistivity of a semiconductor when strain is applied.

With the strain applied, the atomic spacing becomes altered, affecting the band gap (which is the difference in electron energy between the top of the valence band and bottom of the conduction band).

When the piezoresistivity effect occurs, due to strain on the material, the electrons are raised to the conductive band, altering the resistivity of the material.

Piezo-resistive transduction

Numerous materials exhibit piezo-resistive properties, and it is typically quantified by the *gauge factor*.

Gauge factor - is the change in resistance per given strain per starting resistance.

The image to the right shows an example of a piezo resistive pressure sensor.

Static charge (Triboelectric) transduction

Triboelectric effect (static charging) - is a type of contact electrification on which certain materials become electrically charged after they come into frictional contact with a different material.

An example of this effect is when glass is rubbed with fur, or when a plastic comb is run through hair. The static charge that is observed by standing hair is a result of triboelectrical energy.

The polarity and intensity of triboelectric charges differ, based on the type of material, surface roughness, strain, ambient temperature, and other such properties.

Standing hair is an indication of the presence of triboelectrical energy (image).

The Van Der Graff generator

Another example of the triboelectric effect is observed in the Van de Graff Generator, commonly seen at your local high school science fair.

This is an electrostatic generator which uses a moving belt for accumulating an electric charge on a hollow metal globe atop an insulated column.

This type of generator produces a high voltage direct current (DC) output; however it is safe to the touch because of the low level of current.

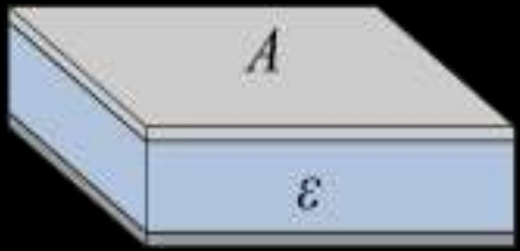
Capacitive transduction

Capacitance is a physical phenomenon which is widely used in the design of sensor devices.

The principles of capacitance can be applied for the measuring of distance, area, volume, pressure, force, the composition of a chemical substance, and much more.

Capacitance is a property that is observed between a set of two conductive surfaces that are within close proximity to one another.

A change in the distance between the two surfaces will create a change in the capacitance. This change in capacitance is the measurand that capacitive sensors quantify to indicate changes in position of a target.



Parallel plate capacitor

The capacitive transducer is passive requiring an external power source for excitation.

It operates on the principle of a change in capacitance from overlapping of plates, or the varying of distance between the plates and a dielectric constant.

A capacitive transducer consists of two parallel metal plates (see image), which are separated by a dielectric medium that is air, material, gas or liquid.

In a typical capacitor, the distance between the plates is at a fixed distance. However, in a capacitive transducer the distance between them varies.

Capacitive style touchscreens

Some touchscreen devices are capacitive sensors, where the panel consists of an insulator, such as glass, coated with a transparent conductor, such as indium tin oxide.

As the human body is also an electrical conductor, touching the surface of the screen results in a distortion of the screen's electrostatic field, which can be measured as a change in capacitance.

Capacitive coupling

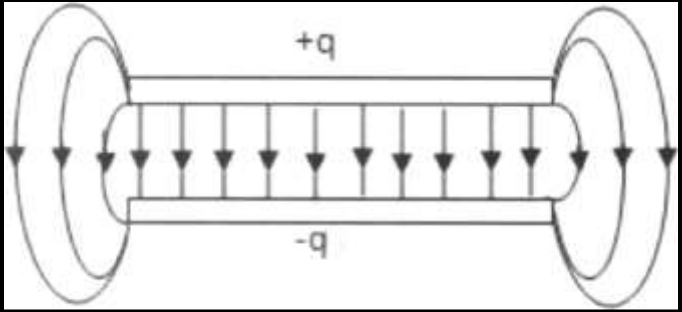
Capacitive sensing is based on capacitive coupling that can detect and measure anything that is conductive or is dielectrically different from air.

Capacitive coupling is the transfer of energy within an electrical network by means of “displacement current” between circuit nodes, induced by the electric field.

Displacement current

This is a type of "virtual current" that is exhibited in electromagnetic devices such as capacitors, the basic rules by which are explained in Maxwell's equations on electromagnetism.

Displacement current has the same units as electrical current. Likewise, it is a source of the magnetic field just as actual current is. However, displacement current is not an electric current of moving charges, but a time-varying electric field.



Fringing effect

A capacitor is made of two conductive objects with a dielectric medium between them. A voltage difference applied between these objects results in an electric field between them.

This electric field (image) not only exists directly between the conductive objects, but extends outward. This is considered the “fringing field”.

Thermoelectric transduction

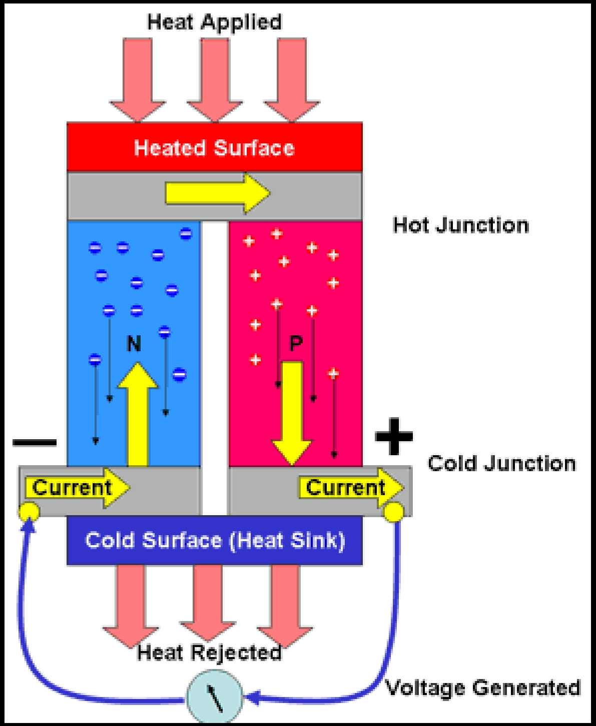
Thermoelectricity is a transduction technique that converts the value of the measurand into a voltage (or electromotive force) generated by the potential difference between the junctions of two selected dissimilar materials due to the Seebeck effect.

Seebeck effect

This is a type of thermoelectric phenomenon, created when one end of a conducting material is located in a cold region, while the other end is in a warm region, thus causing energy to flow from the warmer mass to colder mass.

This energy takes the form of heat, with the intensity of the heat flow being proportionate to the thermal conductivity of the conducting material.

The resulting thermal gradient reflects an electric field within the conductor, which directly relates to the Thomson effect.



Thomson effect

This is a thermoelectric phenomenon caused by the alternating transference of heat (emitting/absorbing) when electric current passes through a circuit made of a homogenous material, which has a temperature differential along its length.

An example being, a solid copper wire which conducts a steady electric current is subjected to an external heat along a short section of the wire, while the remainder stays cool.

Heat is absorbed from the copper as the current approaches the heated section, and heat is transferred to the copper just beyond the hot point.

Joule effect

This is the physical law which expresses the relationship between the heat generated and current flowing through a conductive material.

Gough–Joule effect

This is the tendency of elastomers to contract if heated while they are under tension. For example, if an elastic band is first stretched and then subjected to heating, it will shrink rather than expand.

Peltier effect

This is an effect, where heat is emitted or absorbed when an electrical current is passed across a junction between two materials.

This phenomenon can be useful when a small scale transfer of heat is needed from one medium to another.

Transduction via the pyroelectric effect

This is the ability of certain materials to generate a temporary electric charge when heat is applied by conduction or radiation.

This temperature change alters the positioning of the atoms slightly within the crystal structure, changing the polarization of the material.

Thermoresistive transduction

Thermal resistance is the ability of a material to resist or impede the flow of heat.

The "thermoresistive effect" is a useful transduction principle to use in the design of thermal sensors and devices, due to its simplicity in implementing and its high degree of sensitivity.

Thermoelastic transduction

Thermoelasticity describes the behavior of materials, when thermal energy is added to an elastic material causing it to expand. The “thermoelastic effect” is a principle which is utilized in MEMS resonant sensor technology for the purpose of Thermoelastic damping or (TED).

Thermoelasticity is an important contributor in the modeling of high-quality MEMS resonators or resonant sensor devices (sensors that are configured to have a mechanical resonance frequency).

Ettingshausen effect

This is a thermoelectric (or thermomagnetic) phenomenon that affects the electric current in a conductor when a magnetic field is present.

Transduction using photosensitivity

Photosensitivity is the degree of reaction which an object or material displays upon absorbing photons, especially those within the spectrum of visible light.

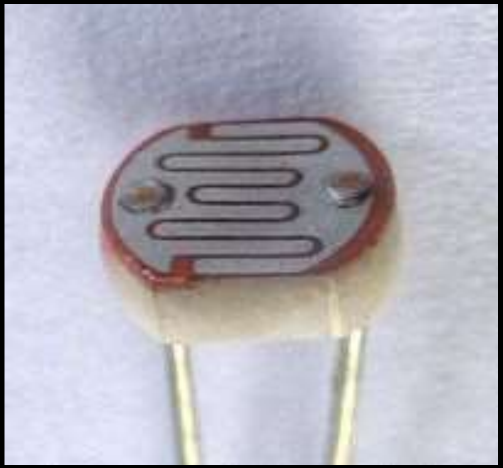
This is a very useful phenomenon in the design of sensors devices; especially those which are used in fiber optic applications.

Transduction using photoconductivity

Photoconductivity is an optical and electrical phenomenon. Photoconductive materials become more electrically conductive when they absorb electromagnetic radiation such as visible light, ultraviolet light, infrared light, or gamma radiation.

When EM radiation is absorbed by a photoconductive material, free electron and electron hole quantities are increased, raising its electrical conductivity.

To create excitation, the light absorbed by the material must have enough energy to raise electrons across the band gap, or to excite the impurities within the band gap.

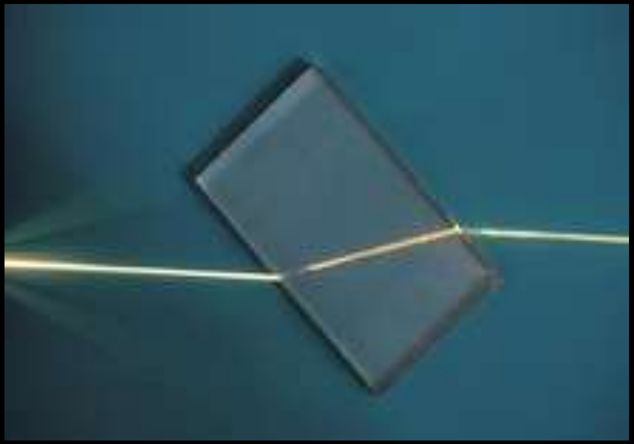


Transduction using photoresistivity

Photoresistance is the resistance in a material that varies as the amount of incident light varies.

A photoresistor or light-dependent resistor (LDR) is a light-controlled variable resistor that exhibits an electrical resistance that decreases with the increase in incident light intensity, thus exhibiting the property of photoconductivity.

A photoresistor (see image) can be used in light-sensitive detectors, and light-activated or dark-activated switching devices.



The Thermo-optic coefficient

The thermo-optic coefficient of a material is the change in the refractive index (RI) as it responses to a temperature variation.

In optics, the refractive index or index of refraction for a material is a dimensionless number describing how light propagates as it passes through that particular medium.

The image shows a ray of light being refracted as it passes through a plastic block.

Photoelectric effect

The photoelectric effect is the emission of electrons or other free carriers when the material absorbs light. Electrons which are emitted due to the photoelectric effect are called photo electrons.

Photovoltaic effect

The photovoltaic effect occurs when voltage and electric current are created in a semi-conductive material when exposed to a source of light.

Photovoltaic cells are used in sensor technology as light sensors, or to power stand-alone sensor networks.

The image shows a photovoltaic (solar) cell made of crystalline silicon.

Photoelectrochemical effect

Photoelectrochemical or PEC cells are solar cells that are able to output electrical energy or hydrogen in a process which is similar to the electrolysis of water.

The PEC effect is being studied for potential future use in biosensing technology.

Electromagnetic phenomenon and the Maxwell’s equations

Maxwell's equations are a set of partial differential equations that, when combined with the Lorentz force law, form the foundation of classical electromagnetism, classical optics, and electric circuits.

These equations are the basis of all electric, optical and radio technologies, including power generation, electric motors, wireless communication, cameras, televisions, computers etc.

Maxwell's equations describe the principles of how electric and magnetic fields are generated by charges, currents, and changes of each other.

These equations help to illustrate how fluctuating electric and magnetic fields propagate at the speed of light.

These waves, which are known as electromagnetic radiation, may occur at various wavelengths to produce a spectrum extending from radio waves to γ-rays.

Lorentz force

In the physics of electromagnetics, the Lorentz force is the combination of electric and magnetic forces on a point charge, due to electromagnetic fields.

Transduction using induction

Electromagnetic or magnetic *induction* is the production of an electromotive force (EMF) across an electrical conductor within a changing magnetic field.

Induction and the induction loop

One transducer device which uses the principles of induction for sensing is the "induction loop".

This is an electromagnetic detection system which uses a moving magnet or an AC current to induce an electric current in a nearby wire. Induction loops are used for transmitting and receiving communication signals.

Other common applications of induction loops are for detecting vehicles at traffic intersections(image), classifying of vehicles, or for use in metal detectors.

Galvanomagnetic effect

This effect is observed by measuring the induced voltage normal to the direction of current flow as with the Hall Effect, but with the magnetic field in the current-voltage plane.

The galvanomagnetic effect is used to measure magnetic quantities, and is utilized by Hall and magneto-resistive types of sensors. The output signals of these sensors correspond to the magnetic quantities of the measurand.

Wiegand effect

This is a non-linear magnetic effect, which is created in specially annealed and hardened wire called “wiegand wire”.

This cold-working process creates a wire which has an outer shell with a greater magnetic coercivity than its inner core.

Wiegand sensors

Wiegand sensors exploit the properties of Wiegand wiring, for use as a pulse generator in a variety of applications. The sensor requires no external power source and has no moving parts.

When a magnetic field changes the magnetic state of the Wiegand wire within the sensor device, an output pulse is produced which can be used as a power source.

Magneto-resistive phenomenon

This is the tendency of a material to alter its electrical resistive value when exposed to an externally-applied magnetic field.

Nernst effect

The Nernst effect is a thermoelectric (or thermomagnetic) phenomenon that is observed when an electrical conductor is subjected to a magnetic field and a temperature gradient which are perpendicular to one another.

The result is an electric field induced perpendicular to both the field and the gradient.

Hall effect

The Hall Effect is an electromagnetic phenomenon that occurs when a potential difference is produced across an electrical conductor, and when a magnetic field is applied in a perpendicular direction to that of the flow of current.

Theory of the Hall effect

The Hall Effect occurs due to the behavior of the current in a electrical conductor. Current is due to the electromigration or movement of charge carriers: typically electrons, holes, and/or ions.

When a magnetic field is present, these charges experience a force, known as the Lorentz force, as mentioned previously.

The Hall coefficient

This is defined as the ratio of the induced electric field to the product of the current density and the applied magnetic field.

It is a characteristic of the material from which the conductor is made, since its value depends on the type, number, and properties of the charge carriers that constitute the current.

Sagnac effect

Also called *Sagnac interference*, is a phenomenon encountered in interferometry* that is elicited by rotation. In an arrangement known as a *Sagnac interferometer*, a beam of light is split and the two beams are made to follow in opposite directions a trajectory that constitutes a ring.

To act as a ring the trajectory must enclose an area. The axis of rotation does not have to be inside the enclosed area.

**Interferometry* is a family of techniques in which waves (usually electromagnetic waves), are superimposed causing the phenomenon of interference in order to extract information.

Sensing of ionizing radiation

Radioactivity is a phenomenon defined by the emissions of neutral or charged particles, or electromagnetic radiations emitting from an unstable atomic nuclei.

Sensor devices are used to detect and measure this “ionizing” radiation. Ionizing radiation is the type of radiation which carries sufficient energy needed to separate electrons from atoms or molecules, (thereby ionizing them).

Ionizing radiation is made up of energetic subatomic particles, ions or atoms moving at high speeds (usually in excess of 1% of the speed of light), and electromagnetic waves on the high-energy end of the electromagnetic spectrum.

Two types of radiation

Ionizing radiation

Gamma rays, X-rays, and the higher UV range of the electromagnetic spectrum are forms of ionizing radiation.

Non-ionizing radiation

This is the lower UV range of the electromagnetic spectrum, and the range beneath the UV spectrum, including visible light, nearly all types of laser light, infrared, microwaves, and radio waves are all considered non-ionizing radiation.

Detecting radiation using scintillation counters

These are instruments which detects and measures levels of ionizing radiation, by using the excitation effect of incident radiation on a scintillator material, to detect the resultant light pulses.

Detecting radiation using Geiger counters

This is another type of instrument used to measure ionizing radiation. Its name comes from the primary component, the Geiger–Müller tube.

Applications for this device include: radiation dosimetry, radiological protection, experimental physics, and the nuclear power industry.

It detects ionizing radiation such as alpha particles, beta particles and gamma rays using the ionization effect produced in the G-M tube (see image).

Electro-acoustical transduction

In the context of electroacoustics, transduction would refer to the conversion of sound energy into electrical energy (or vice versa).

Types of electroacoustic transducers include devices such as loudspeakers, microphones, hydrophones and sonar projectors.

Pressure waves

These devices convert a sound pressure wave to or from an electric signal, with the most common types of transduction principles being: electromagnetism, electrostatics or piezoelectricity.

As an acoustic wave is a mechanical pressure wave, acoustical measuring devices will operate under the same principles as a pressure sensor.

These acoustic sensors consist of a moving diaphragm and a displacement transducer that converts the mechanical motion in the diaphragm into an electrical signal.

Using the Doppler effect for sensing

One type of sensing which utilizes the Doppler Effect is Doppler velocimetry, employed by Doppler radar velocity-measuring devices. This type of sensor is used in speed measuring applications by law enforcement as well as use in baseball to measure pitching speed.

It uses the "Doppler shift" principle, which is the change in frequency or wavelength of a wave which occurs when two bodies (the sound source and the observer) are moving at different speeds relative to one another. An example is the change in pitch which is encountered when an ambulance with a siren, drives by a stationary observer.

Biological transduction

Biosensors, or biological sensors are devices made up of a transducer and a biological element. The bioelement may be an enzyme, a protein, a nucleic acid, or an antibody. Antibody-based biosensors are also known as *immunosensors*.

The bioelement interacts with the biological measurand, or *analyte*, and then that biological response is converted into a signal. Biosensors are useful in measuring a number of different types of analytes such as organic compounds, gases, ions and bacteria.

Depending on the specific application, a biosensors may be called an immunosensor, optrode, resonant mirror, chemical canary, biochip, glucometer, or even biocomputer.

Summary

This concludes our course on modern sensor technologies.

In this course, we covered many basic introductory concepts and principles employed in modern sensor technology.

We explored many of the various ways in which sensors are classified, their characteristics, and assorted applications. Also included were the ways in which sensors experience operational deviations, and the many different types of physical phenomena that can be sensed.

The world of sensors is vast and ever-expanding, with devices available for nearly any possible application.

Bibliography:

- Jacob Fraden, “Handbook of Modern Sensors; Physics, Design, and Applications”, 4th Edition, Springer Press 2010.
- J. G. Webster, “The Measurement, Instrumentation and Sensors Handbook”, CRC/IEEE Press, 1999
- “Sensor Technology Handbook”, edited by Jon Wilson, Newnes Publications, Elsevier, Inc., 2005
- Encoder sensor information and images - anaheimautomation.com
- Handbook of Measuring System Design, John Wiley & Sons, Ltd, 2005
- Prof. Kasim Al-Aubidy, Wright State University, lecture on “Intelligent Sensor Systems”
- Article: “What are Biosensors?”, from news-medical.net, Oct 2014

This concludes our course on “An Introduction to Modern Sensor Technology”.
You may now proceed to the final exam.

Thank you for taking this Flashcard course!

An introduction to Modern Sensor Technology, *2019 Edition*
First Printing: March, 2019