SENSOR SYSTEMS

Luca Colombo | Politecnico di Milano | a.a. 2021/22

DISCLAIMER

These notes cover the arguments of the course ‘Sensor Systems’ held by Professor F.Villa at Politecnico di Milano during the academic year 2021-2022.

Since they have been authored by a student, errors and imprecisions can be present.

These notes don’t aim at being a substitute for the lectures of Professor Villa, but a simple useful tool for any student (life at PoliMi is already hard as it is, cooperating is nothing but the bare minimum).

Please remember that for a complete understanding of the subject there is no better way than directly attending the course (DIY), which is an approach that I personally suggest to anyone. Indeed, the course is really enjoyable and the professor very clear and helpful.

In any case, if you found these notes particularly helpful and want to buy me a coffee for the effort, you’re more than welcome: https://paypal.me/LucaColomboxc

lucacolombo29@gmail.com

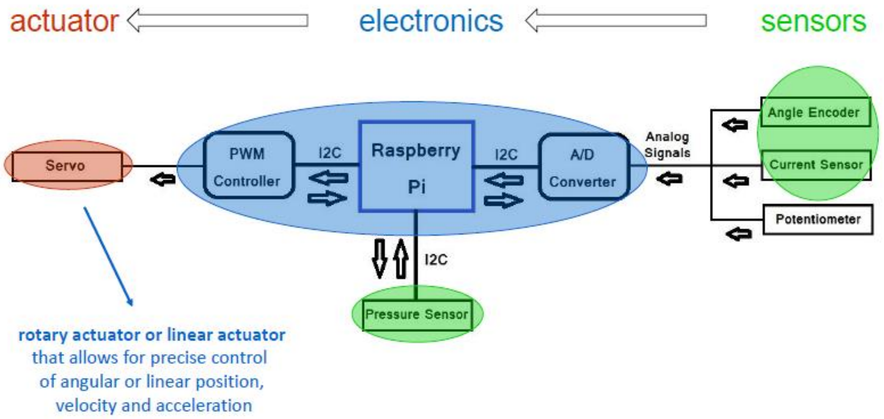
INTRODUCTION

- TRANSDUCER: **device which transforms energy from one type to another, even if both energy types are in the same domain.** Typical energy domains are mechanical, electrical, chemical, magnetic, optical and thermal. Transducer can be further divided into **Sensors** and **Actuators**. Typical domains are electrical, magnetic, optical. A sensor takes for instance a thermal signal and transforms it into al electrical one. Sensors are those transducers that transform a physical parameter in an electrical signal.

- SENSOR: **device which monitor a parameter of a system, hopefully without disturbing that parameter.** The specific input could be light, heat, motion, moisture, pressure, or any one of a great number of other environmental phenomena. The output is generally a signal that is converted to human-readable display at the sensor location or transmitted electronically over a network for reading or further processing.

- ACTUATOR: **component of machines which is responsible for moving or controlling a mechanism or system.** The actuator does the opposite with respect to the sensor. It uses an electrical signal to control some moving parts.

In the bottom we have an example of a sensor-actuator chain.



rotary actuator or linear actuator that allows for precise control of angular or linear position, velocity and acceleration

All the sensors go to a processor that put together all the information and then provides an information to an actuator.

Then we can do 2 main classifications for sensors.

Classification based on physical phenomena

- **Optical**: visible/IR light (photodiode, CCD, CMOS APS, infrared sensor)
- **Thermal**: temperature (RTD, thermistor, thermocouple…)
- **Magnetic**: magnetic field (Hall effect sensor, magneto-resistive sensor)
- **Mechanical**: strain (strain gauge), force (piezo-electric sensor), displacement and distance (capacitive, inductive, acoustic, optical), acceleration and orientation (MEMS), viscosity, pressure, etc.
- **Chemical**: pH (pH-meter)

Classification based on measuring mechanism

- Resistive sensing: the resistance changes its value depending on the parameter we want to measure, e.g. temperature, light.
- Capacitive sensing
- Inductive sensing
- Piezoelectricity
- Hall Effect
- MEMS (Micro-Electro-Mechanical System)

SENSORS’ CHARACTERISTICS

Sensitivity: the ratio between the change in the output signal to a small change in input physical signal. Slope of the input-output fit line.

$$S = \frac{\partial\, out}{\partial\, in}$$

Resolution (LSB): the smallest increment of measure that a device can make (Least Significant Bit). It is referred to the resolution of the ADC. If we have an ADC that converts from 0 to 5V, the FSR of the ADC is 5V. The resolution of the ADC is typically in mV, while in a smart sensor we are not interested in the resolution of the ADC, but it will be expressed in terms of the quantity we want to measure (e.g. mmHg, T…).

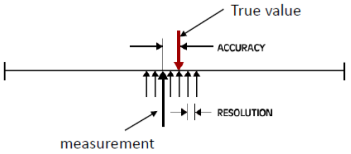
Full Scale Range (FSR): the maximum interval of measure that a device can cover.

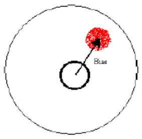
Number of bits (n): in sensors with digital output 2nisthe number of levels in which the FSR is divided.

$$2^n = \frac{FSR}{LSB}$$

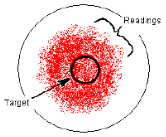
Accuracy: error between the result of a measurement and the true value being measured.

Repeatability/Precision: the ability of the sensor to output the same value for the same input over a number of trials.

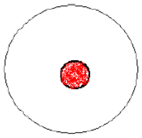




Precision without accuracy



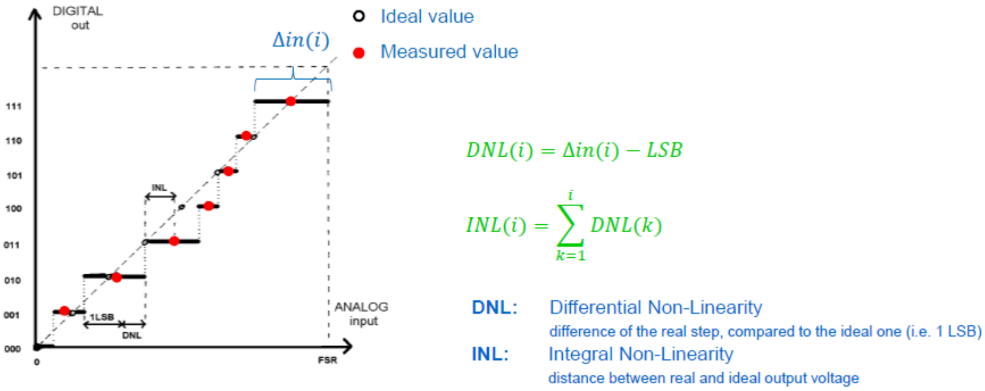
Accuracy without precision



Precision and accuracy

Linearity: the deviation of the output from a best-fit straight line for a given range of the sensor. Differential Non-Linearity (DNL) and Integral Non-Linearity (INL). If we consider the case of a smart sensor, where the output is a digital code, we take the output from the sensor. Theoretically, if we have a very linear sensor, we will follow the blue straight line. However, due to nonlinearities of ADC and sensors, we can move away from the straight line and obtain the red dots.

If we move the FSR, the ADC is not precisely always the same, sometimes is shorter, sometimes longer, so we can go up and down with respect to the ideal values.



$$DNL(i) = \Delta in(i) - LSB$$

$$INL(i) = \sum_{k=1}^{i} DNL(k)$$

DNL: Differential Non-Linearity
difference of the real step, compared to the ideal one (i.e. 1 LSB)

INL: Integral Non-Linearity
distance between real and ideal output voltage

Differential nonlinearity is the incremental difference we have in the real step with respect to the ideal value. It can be either negative or positive.
The INL is the difference between the red spot and the ideal point on the ideal curve. It depends on the previous DNL. If the DNL is for one code negative and for the other positive, we will have red points jumping above and below. If DNL is always positive, will be always lower than the red curve.

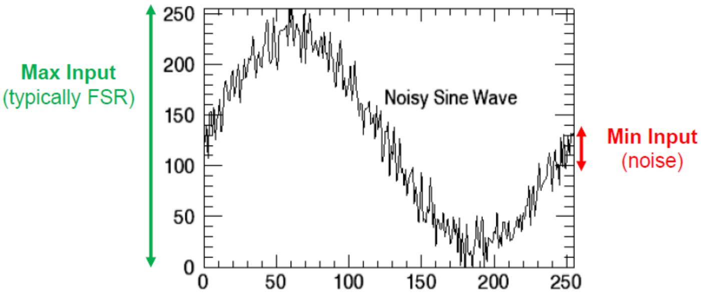
Transfer Function (Frequency Response): the relationship between physical input signal and electrical output signal, which may constitute a complete description of the sensor characteristics as a function of frequency. It depends on the frequency at which we are working.

Bandwidth: the frequency range between the lower and upper cutoff frequencies, within which the sensor transfer function is constant gain or linear. It is the range of frequencies in which the sensor can work.

Noise: random fluctuation in the measured value. It is quantified with its rms (root mean square) value.

Dynamic Range: the ratio of maximum recordable input amplitude to minimum input amplitude. The maximum input amplitude is related to FSR, while the minimum signal we can measure is limited by the noise.

$$DR = 20 \cdot Log\left(\frac{Max\ Input\ Amplitute}{Min\ Input\ Amplitude}\right) \text{ (expressed in dB)}$$



SENSOR READOUT

Analog sensors

- ADC + microcontroller: many microcontrollers have a built-in ADC
- Data Acquisition Cards (DAC): they are boards we can connect directly to the computer and then we have different types of connectors. Then they implement a sort of oscilloscope in the board, and we visualize the voltage on the computer.

Smart sensors

A smart sensor is a sensor with a built-in signal processing and communication (ADC already included) → digital output.

The digital output can be done with:

- Parallel bus
- Serial I/O: the output can be in SPI or I^2C (synchronous) or asynchronous. The data are sent according to a clock.
  - Serial Peripheral Interface (SPI): 1 clock + 1 bidirectional data + 1 chip select/enable
  - Inter Integrated Circuit (I^2C): 1 clock + 1 data

  Asynchronous (no clock): one wire (must match baud rate and bit width, transmission protocol, etc.). They are frequency encoded.

Non-ideal effects

- Offset: nominal output ≠ nominal parameter value, with a fixed difference. It is deterministic.
- Nonlinearity: output not linear with parameter changes. Can be sometimes positive and sometimes negative, we cannot correct it with an equation.
- Cross parameter sensitivity: secondary output variation with other parameters (e.g., temperature drifts)

Calibration

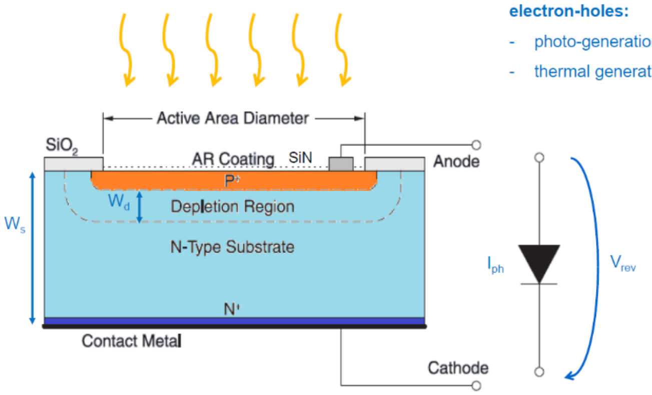
It is the adjustment of the output to match the parameter.

- Analog signal conditioning: for instance for the Strain Gauges we can use some bridge configurations.
- Look-up table, in case of nonlinearities.
- Digital calibration if we know the parameters that correct our value.

LIGHT SENSORS

THE PHOTODIODE

It is a diode with two terminals and typically is sold in package, with an anode and a cathode, where we have the line |. The package has two terminals, one for cathode and one for anode, and then there is glass in front to have light entering. The package is used to protect the golden wires.



electron-holes:

- photo-generation
- thermal generation

As for its cross-section, we see that we have a pn junction like a normal diode, with a highly doped p region on the top and the n doped region. In between the two we have a depleted region, because in p part we have an excess of holes and in the n part an excess of electrons that are free to move within the silicon. When p and n region are put in contact, due to diffusion, holes and e- will move in the opposite region, holes in the n region and electrons in the p region. When they move in the opposite region, they recombine. This happens at the interface between the p and n region, so we have a region that is called depletion region where we have no free carriers.
The depletion region is the region where photons can be absorbed. The wider, the higher the probability to absorb a photon within the depleted region.

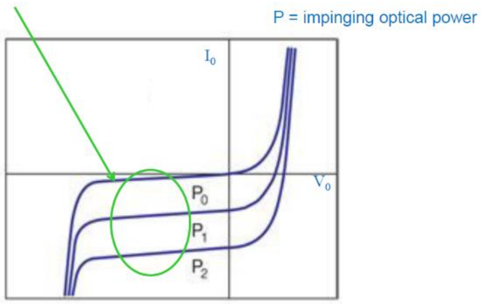
To increase the width of the depletion region, we can reverse bias the photodiode. With this biasing, photon that enters in the depletion region creates e-/h pairs, that can however be also thermally generated. This represents the noise of the detector.

In the characteristic of the photodiode we can see that we have different curves depending on the incident optical power. The highest sensitivity, provided a certain voltage V in the green region, if we increase the power we increase also the reverse current.

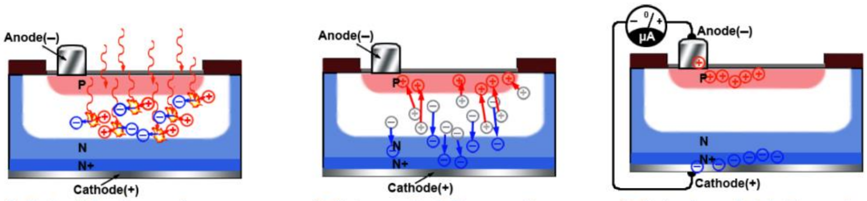
In order to have a PD is important to bias it below the breakdown voltage, so that light is proportional to light intensity (not like in the APD).

Photoconductive linear mode:

- Reverse bias
- Below break-down voltage



Photodiode operation



1 - Electron/Hole pairs generation

2 - Carriers are attracted by reverse bias

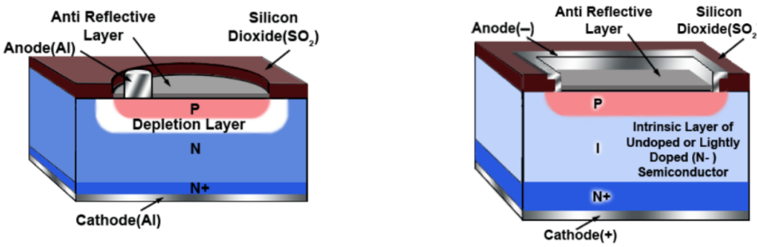
3 - Carriers form a photoelectric current

Light enters in the active area of the photodiode. The photon can generate an e/h pair, being absorbed. Then these electrons and holes are attracted by the reverse bias (reverse bias = Vcathode > Vanode), so we have a separation of charges and a current, called photocurrent. If the photon is absorbed not in the depletion region, due to the large amount of e- in the rest of the n region, the e/h pair recombines.

There are two types of photodiodes:

- Standard PD: seen up to now
- Pin PD: in between we have an intrinsic region, that is not doped or very slightly doped, so it is almost a depleted region (no free electrons or free holes). Still due to diffusion, e- will move towards the p region and holes towards the n region. But electrons and holes won't recombine in the intrinsic region, because there we have very few holes and electrons. So the depletion region will be very wide, from p to n.

  The wider the depletion region, the higher the probability to detect a photon and moreover, having an intrinsic region is good for the readout. Indeed, the PD can be modelled as a capacitor, since it is like to have two plates that separate something that is not conductive, because we have no carriers in the depletion region. The bigger this capacitor, the more difficult is to readout the PD, because with a bigger capacitor the readout circuit won't be stable. For instance, if the PD is readout with a transimpedance amplifier, the higher the capacitor the higher the probability of having instability of the circuit.



p-i-n photodiode has larger depletion region

→ reduced capacitance

→ responsivity in a wider spectrum

  The advantage of the pin is that the capacitance value is reduced because the distance between plates is higher → less issues related to the detector capacitance.

Penetration depth

The responsivity, capability to detect photons, depends on the penetration depth, that is defined as below.

Penetration depth = depth at which the intensity of the radiation inside the material falls to 1/e (about 37%) of its original value at the surface.

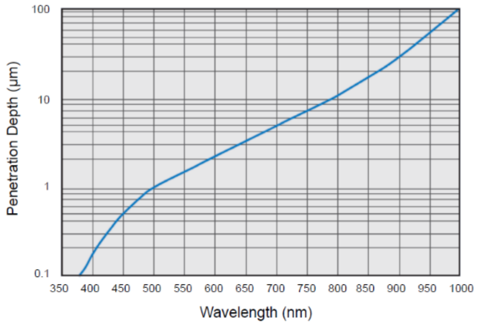


Photo-generation:

Eph > Eg

(Si: Eg = 1.12 eV)

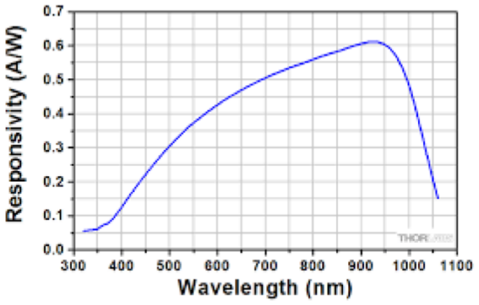
So we have a flux of photons and we want to know the depth at which 67% of the photons are absorbed. The penetration depth depends on the wavelength; for instance, for IR light it is of about 30um. Hence if we increase the depleted region we increase the probability of detecting photons.
The graph refers to the penetration depth in silicon.

The penetration depth depends on the energy of the photon and on the energy gap of the material. in silicon the Eg = 1.12 eV. Moreover, the longer the wavelength, the lower the energy of the photon. This is the region why at shorter wavelength we have a smaller penetration depth.

If we change the material, like indium phosphide, we have a lower energy gap, so they are able to absorb photons also in the near infrared region.

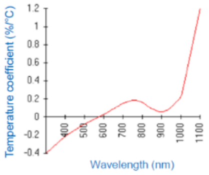
Responsivity

Photodiode Responsivity (sensitivity) =
ratio of the photocurrent I_{ph} to the incident light power P at a given wavelength: $R_\lambda = \frac{I_{ph}}{P}$



R_λ depends on:

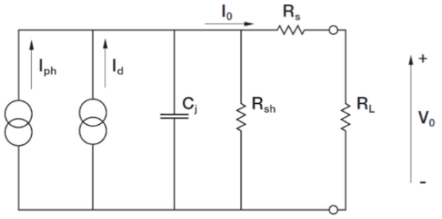
- Incident light wavelength
- Applied V_{rev}
- Temperature



It is the variation of the output current due to the variation of the input power (it is the equivalent of sensitivity). It mainly depends on:

1. Wavelength of incoming light: it typically peaks at 900nm, but then drops due to too high wavelength
2. Reverse bias: if increased, we increase the probability to detect a photon
3. Temperature: depending on the lambda, it is possible that increasing T increases R for wavelength greater than 600nm, while for lambda < 600nm the responsivity decreases. The behaviour is different at different wavelengths because T affects different factors.

Photodiode electrical model



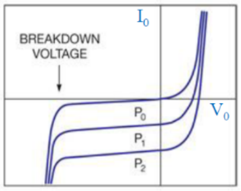
I_{ph} = photo- generated current

I_d = current of the ideal pn-junction (dark current)

C_j = junction capacitance: $C_j = \frac{\varepsilon\cdot A}{W_d}$

R_{SH} = shunt resistance: slope at 0V

R_s = series resistance: $R_s = \frac{(W_s - W_d)\cdot\rho}{A} + R_c$



We can model the PD with a current generator Iph which represents the photocurrent, generated by the photon absorbed, so it is proportional to incoming light. Then we have another generator that represents the dark current, due to thermal promotion (Id).
The passive parameters in the model are the junction capacitance (mentioned before) that depends on the active area of the photodiode (the bigger the higher the Cj) and on the depletion region width.

Then we have a shunt resistor Rsh that is not a physical resistor but a parameter that models the slope we have at 0V. It models the fact that the characteristic of the PD is not perfectly straight but it depends on the reverse voltage we provide. The higher the voltage we provide, the higher the current that flows in the shunt resistor.

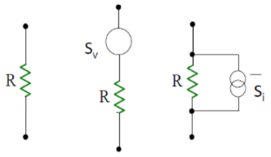
Then we have the series resistance, that is a physical resistance and it is represented by the resistance of the semiconductor (Si) in the region where it is not depleted, so both in n and p region. But also it represents the resistance of the contacts to readout the current. Rs depends on the resistance of the contacts and on the resistance of the non-depleted part (Ws is the width of the entire photodiode minus the Wd depleted region). Then Rs is also reversely proportional to the active area.

NB: active area is the entrance area capable to collect photons, from a cross-sectional point of view.

This model can be now used to do an estimation of the detector noise.

Noise sources

Thermal (Jhonson) noise → Brownian motion of charges

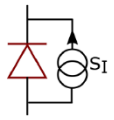


Spectral density

$$S_v = 4KT\cdot R\ \left[\frac{V^2}{Hz}\right] \qquad S_i = \frac{4KT}{R}\ \left[\frac{I^2}{Hz}\right]$$

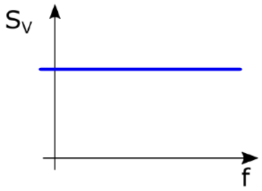
$4kT = 1.66\cdot10^{-20}\ \frac{V^2}{Hz\cdot\Omega}$

Shot noise → "granularity" of charge crossing a junction



Spectral density

$$S_i = 2q\cdot I$$



Noise standard deviation: $V_n = S_v\cdot\Delta f$ and $I_n = S_i\cdot\Delta f$

In the PD we have two sources of noise, thermal noise and shot noise.

- **THERMAL NOISE**: we find it in the resistors and it is due to Brownian motion of charges. If we imagine having a resistor non-biased (0V across), it is like to have for instance Si with free e- and holes. If e- and h are equally distributed, we really readout the across voltage equal to 0, but they can move, and we may have more electrons in one region and more holes in the other → voltage higher or smaller than zero depending on the random motion of carriers.
  So, if we plot the voltage across the resistor, in average is 0, but sometimes is negative and sometimes positive with a certain variance. This is the thermal noise.

  We can represent the thermal noise with a series voltage generator but considering the resistor ideal, or equivalently with a parallel noise current generator. In this latter case, if we have no current externally flowing, we may still have a voltage drop on the resistor due to the injection of current of the noise current generator.

  For both the models we can say that the spectral density of the noise can be represented with the formulas in the image. The spectral density of the noise is the power of noise that we have at a specific frequency. We can see that the spectral density associated to the thermal noise has no dependency on frequency → same noise for each frequency. The unit of measure of Sv is [V^2/Hz]. The same for Si.

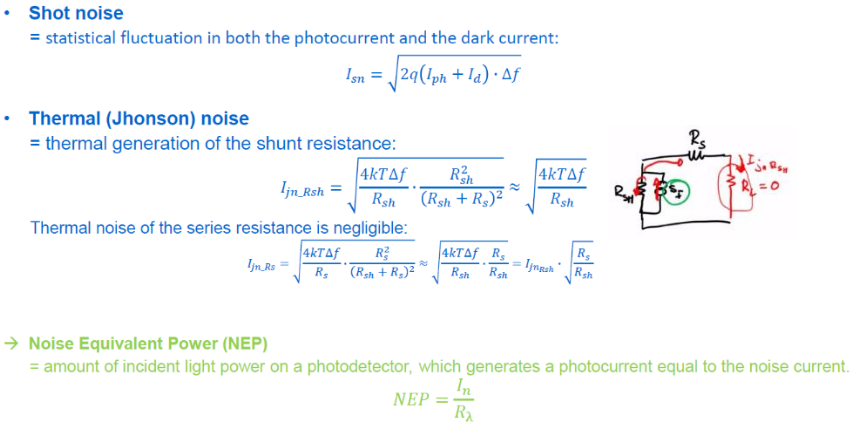
  NB: when we work with spectral densities, so powers, the transfer function must be squared.

- **SHOT NOISE**: noise typically in all the pn junction. Anytime we have a pn junction we have also shot noise. It depends on the current we have crossing the pn junction because it is generated by the granularity of the current that crosses the junction. On average, we have some charges that cross the junction in order to have, for instance 1A. If we consider in 1s we have on average 1C, but on average, sometimes more, sometimes less. So the current that crosses the junction is not exactly 1A, but 1A with some variance, and the variance can be represented with a current generator.
  The spectral density of the current generator depends on the current but not on the frequency.

If we want to have a practical idea of these noises, in terms of amplitudes, I would like to have a quantity expressed for instance in A or V, not divided by Hz → I integrated over the bandwidth.
I would like to express the standard deviation of the noise. To move from spectral density to standard deviation, we have to integrate for the delta f of interest, and then do the square root.

In a very ideal case the delta f of interest is infinite if the noise spectral density is infinite, but in reality we have always something that limits the bandwidth, so the spectral density will be limited in frequency. So the ideal noise generator has an infinite bandwidth, but the real circuit a finite one.

Photodiode noise



- **Shot noise**
  = statistical fluctuation in both the photocurrent and the dark current:

$$I_{sn} = \sqrt{2q\left(I_{ph} + I_d\right) \cdot \Delta f}$$

- **Thermal (Jhonson) noise**
  = thermal generation of the shunt resistance:

$$I_{jn\_Rsh} = \sqrt{\frac{4kT\Delta f}{R_{sh}} \cdot \frac{R_{sh}^2}{(R_{sh} + R_s)^2}} \approx \sqrt{\frac{4kT\Delta f}{R_{sh}}}$$

R_s
R_{SH}
S_I
$I_{j,R_{SH}}$
$R_L = 0$

Thermal noise of the series resistance is negligible:

$$I_{jn\_Rs} = \sqrt{\frac{4kT\Delta f}{R_s} \cdot \frac{R_s^2}{(R_{sh} + R_s)^2}} \approx \sqrt{\frac{4kT\Delta f}{R_{sh}} \cdot \frac{R_s}{R_{sh}}} = I_{jn_{Rsh}} \cdot \sqrt{\frac{R_s}{R_{sh}}}$$

→ **Noise Equivalent Power (NEP)**
= amount of incident light power on a photodetector, which generates a photocurrent equal to the noise current.

$$NEP = \frac{I_n}{R_\lambda}$$

The PD is a pn junction, crossed by a current that is Iph + Id. So we have a shot noise, whose spectral density is 2q(Iph + Id). Then we multiply for delta f and take the square root to have the std.
Then we have also the thermal noise, associated to both Rs and Rsh (Rl is not considered because it is not part of the model of the PD, because it just models the load).
For instance, we can represent the thermal noise with their current spectral densities. To compute the thermal noise in output, we have to do a current divider in between Rs and Rsh in order to find the component of the current that goes in the load resistance.
We are computing the thermal noise also of the shunt resistance because the Brownian motion models also the fact that the slope ad 0V is not flat.

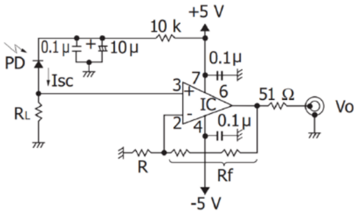
As for the thermal noise of the series resistance Rs, we have to model the current generator for Rs. Again, we have not to consider the component of the current that recirculates in the Rs, but the one that flows in the Rsh, that is the one that will go in our load. Again, we have to perform a current divider.

Since Rs is very small, because it is a parasitic resistance, while the Rsh is a very big resistance (the current doesn't change much if we change the voltage, because it is an almost flat characteristic). So Rs << Rsh. We can hence simplify the equation at the denominator, where Rs is negligible. By splitting Rsh, we can see that one part is identical to the thermal noise of Rsh, but since Rs << Rsh, their ratio is smaller than 1, hence the overall contribution of the thermal noise of Rs can be neglected with respect to the thermal noise associated to Rs.

Typically, these sources of noise are not provided separately, but the datasheet provides the noise equivalent power, NEP. It represents the **amount of incident light we need to have a signal that is the same of the noise**. It is the optical power that I need to make Iph = Inoise (not just Id).

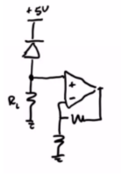
In this regard, the NEP = In/R, where R is the responsivity, R = Iph/P. From this equation of R, P = Iph/R, and by placing NEP in place of P and In in place of Iph we get the NEP formula.

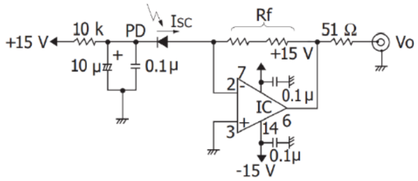
Basic readout circuit



Non-inverting configuration

$$V_o = I_{SC} \cdot R_L \cdot \left(1 + \frac{R_f}{R}\right)$$

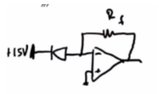




Transimpedance Amplifier

$$V_o = I_{SC} \cdot R_f$$

Typically, must be compensated!



Non-inverting configuration

The readout circuits are take from the datasheet of the photodiode. In the datasheet are introduced also capacitors to keep everything stable, and storage capacitor if we need a high delivery of charge. Moreover, the capacitors must be placed very close to the diode to prevent stray resistances, and we also introduce a 10k resistance to perform a low pass filter of the power supply.

We have some capacitors also close to the power supply of the amplifier, that are decoupling capacitors. Then the feedback resistance Rf is represented with two resistors because very often we cannot buy just a resistor with the value we want, but we need to use more resistors.
In the academic representation, we have the ideal power supply, the PD and some current that will go in the load resistor Rl, generating a voltage multiplied by the non-inverting configuration gain.

The main problem of this circuit is that the reverse bias provided to the PD depends on Iph itself. If Iph increases, V+ increases and we reduce the reverse bias, so we have a sort of negative feedback. So if light increases, we decrease the reverse bias, so we reduce the responsivity and we go to a sort of saturation, because increasing the current we decrease the responsivity of the PD.

To avoid this issue, we can use the second architecture.

Transimpedance amplifier

Still we have a low pass filter for the power supply, capacitors to provide stable voltage also to the supply of the amplifier. We have also a 51 Ohm resistance and the symbol of a coaxial cable, that has an equivalent resistance of 50Ohm, so we have to put 51Ohm in our circuit to avoid reflections.

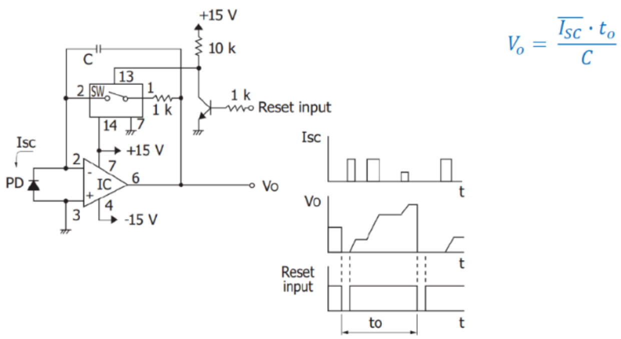
The PD is still reverse biased, so Iph will go in Rf and generate the output voltage (a minus is missing in the formula).
The advantage with respect to the previous one is that the anode of the PD is always at 0V, and it doesn't depend on current.

Light integration

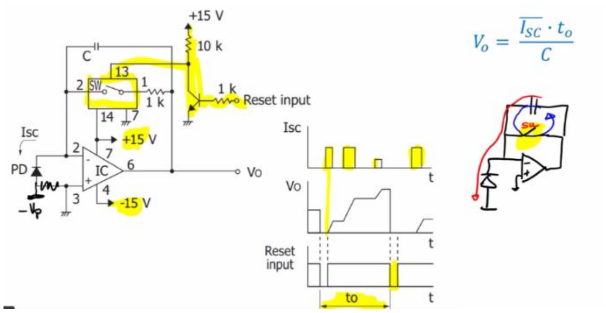
This is another possible readout system. The circuit on the left can be redrawn in another way (the PD must be connected at -Vp to be reverse bias). So we have a reverse bias PD and a switch that can be ideal or a series resistance with a capacitor in parallel to it. When the switch is open, all the current of the PD will go in the capacitor, so we will have an integration of all the charge coming from the PD in the capacitor.

Of course, after a certain period the circuit must be reset, otherwise due to bias current or leakage current the output of the amplifier may saturate. This is the reason why we have the switch; after a certain period it is closed to discharge the capacitor.
If we look more in detail in the schematic from the datasheet we have also power supply of the opamp and more sophisticated models for the switch. Looking at the waveform, we suppose to have a PD current like pulses, and we have an integration of charge in a period called t0. Each t0 we have a reset period in which we close the switch and reset the capacitor. When the Vo is constant we are not integrating charges.

Then the output is provided by the classical relationship we have across a capacitor when we are considering a constant current integrating over it. In this case we consider the average current.



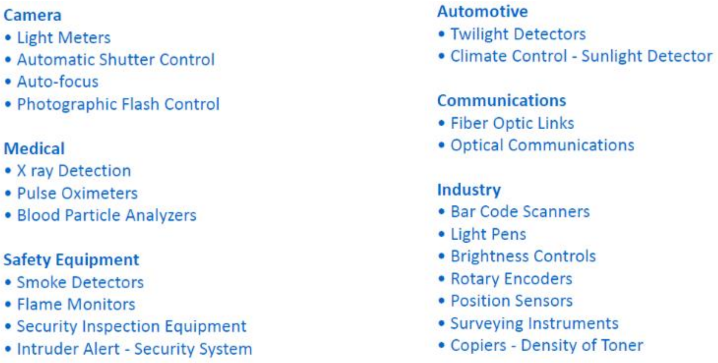
$$V_o = \frac{\overline{I_{SC}} \cdot t_o}{C}$$



$$V_o = \frac{\overline{I_{SC}} \cdot t_o}{C}$$

Photodiode applications

They can be used in very different applications, like for some features of the camera, medical applications, and a lot in the optical communication field.



Camera
- Light Meters
- Automatic Shutter Control
- Auto-focus
- Photographic Flash Control

Medical
- X ray Detection
- Pulse Oximeters
- Blood Particle Analyzers

Safety Equipment
- Smoke Detectors
- Flame Monitors
- Security Inspection Equipment
- Intruder Alert - Security System

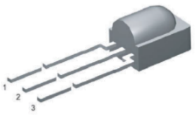
Automotive
- Twilight Detectors
- Climate Control - Sunlight Detector

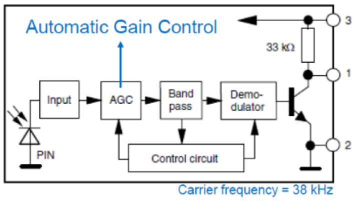
Communications
- Fiber Optic Links
- Optical Communications

Industry
- Bar Code Scanners
- Light Pens
- Brightness Controls
- Rotary Encoders
- Position Sensors
- Surveying Instruments
- Copiers - Density of Toner

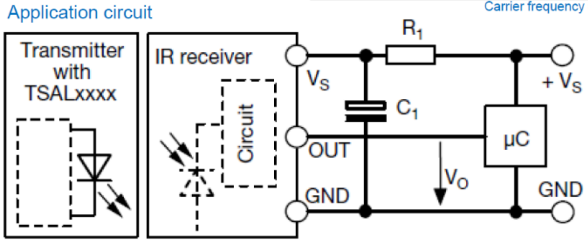
An Example – IR receiver module

It is a module constituted by a PD and an IR emitter. It is used for communication.
It is an IR LED transmitter and a receiver.





Application circuit

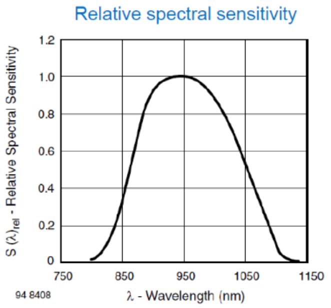


The receiver is not only composed by PD, but PD and electronics → it can be called smart sensor, because it embeds dome front-end electronics. In particular, we have an amplifier, an AGC block to adjust the gain depending on the incoming light, a bandpass filter and demodulator. We need them because communication happens with modulated code. For instance, if we have to transmit a stream of data 0 1 1 0, we cannot simply switch on and of the LED, otherwise we would be very prone to ground illumination of the surrounding behaviour. Hence in order to make the circuit less sensitive to the background, we use modulation. Hence the 1 will be expressed as a series of pulses at a certain frequency for example. The demodulator converts the modulated signal in the original one.

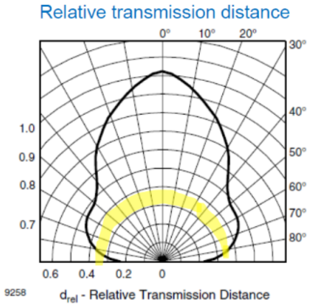
We have 3 pins: power supply, signal and ground. Some external circuits must be connected to these 3 pins, and the output may for example go to a microcontroller.

If we look at the datasheet, we have the graphs below. One is the spectral sensitivity, that is the sensitivity of the PD. In this case the peak is at 950 nm, so in the near IR. To have a PD sensitive to near IR, some filters are placed before the PD to filter out visible light.

The second graph represents the relative transmission distance. Indeed, the sensitivity of the PD is dependent on the direction from which the light arrives. For instance if the angle is 0, the sensitivity is maximal, while if light comes from 80°, the sensitivity that we read on the horizontal axis is about 0.35 with respect to the maximum (35% of the maximum).



The peak sensitivity is at 950 nm
→ Near Infra-Red (NIR)

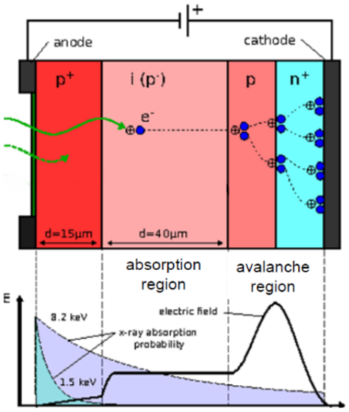


The sensitivity for light beam
with 0° angle is the highest

The same sensitivity for a light beam
with 80° angle is obtain at 0.35 relative distance.

AVALANCHE PHOTODIODE

The structure is similar to a PIN PD, so we have p+, n+, intrinsic region in the middle but then we have also a p region close to the n+ one.



- P-i-N junction
- Bias: below but close to breakdown
- Analog output proportional to incoming light
- Internal multiplication (gain: M ≈ 100)

Basic structure:

- absorption region = separate photo-generated holes and electrons
- multiplication region = high electric field to provide internal photo-current gain by impact ionization

P and n+ create a pn junction and we have a very high electric field in this region. This electric field allows to accelerate the carriers, in particular electrons, to create other e/h pairs. Let's imagine to have incoming light that reaches from left the depletion region. If the photon is absorbed, it will generate e/h pair. The electron will go towards the cathode, and hence it will arrive in the region of the very high electric field, and here it is very much accelerated, so it may generate another e/h due to impact ionization → we will create more and more free carriers. So we have an increase of the charge generated, even if we absorb just one photon.

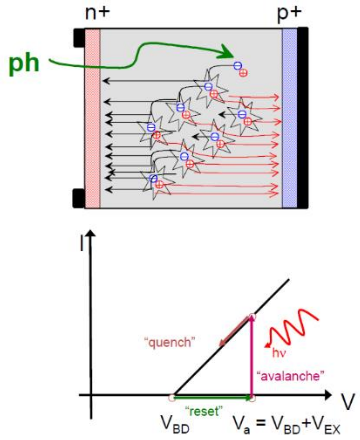
It is important to have a high enough electrical field in the avalanche region, but not to high in order not to go in breakdown, otherwise also holes will be able to be accelerated and able to generate e/h pairs. In the case of the classical APD only e- generate e/h pairs, and this because e- and holes have a different coefficient of ionization.

Typically the gain is in the order of 100, and this structure has a separate absorption and multiplication region, because the absorption one is in the depleted intrinsic silicon, while the multiplication is in the pn junction region. APD is used because if we want a current from the sensor to be higher than the noise of the electronics so that it can be easily read, we need a multiplication. With APD we are still not able to detect single photons.

The disadvantage of APD is that it introduces an additional source of noise, because on average the multiplication is of 100, but obviously this is a statistical number → another source of uncertainty comes from the uncertainty of the gain.

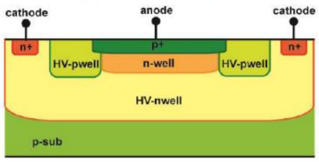
SPAD

To detect also single photons, we need to use a different structure. The structure is similar to a normal PIN PD, but it is biased above the breakdown voltage.

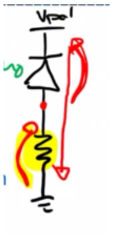


- P-N junction
- Bias: well *ABOVE* breakdown
- **digital output**
- single photon sensitivity
- no read-out noise
- output: >10mA current for every photon

solid-state device:
can be integrated together
with electronics
(counter & TDC)

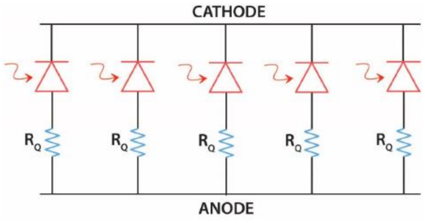


We bias the photodetector above Vbreak, where in theory we should have a macroscopic current. However, SPAD can stay in a metastable stat in which we are above Vbd but with no current. When a photon is detected, a current is generated (in the order of mA even if we detect one single photon). We generate this huge current because we have a positive loop. Indeed, the electrical field in the junction is so high that impact ionization is not only due to electrons, but also holes, and this creates a positive loop. Hence this mechanism will never stop by itself, and to quench the avalanche we can use proper quenching circuits. A very simple one can be implemented with a simple resistor. In fact, if we detect a photon and we generate our avalanche current, the current will generate a voltage drop across the resistor and so we increase the anode voltage → decrease of the bias voltage, so we will go below Vbd.



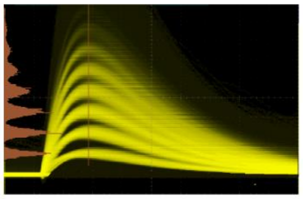
Based on this elementary cell, we may have other photodetectors that are SiPMs.

SiPMs



Hundreds / thousands **SPADs**
+ **quenching** resistance in parallel

- ✓ Single photon sensitivity
- ✓ Photon number resolving capability
- ✓ Large active area



Made by several SPADs connected in parallel. It is used to overcome the main limitation of SPAD. This limitation is that we are sensitive to one signal photon, but either if we detect 1 photon or more than one photon at the same time, we will have always the same signal in output, so we cannot distinguish how many photons arrives simultaneously on the SPAD cell. With sipm, if we get 3 photons simultaneously, we will have 3 times the current of a single SPAD in output.

If we look at the yellow plot, we can see that the bottom line is with one photon, then all the other current pulses are proportional to the number of incoming photons.

LIGHT DEPENDENT RESISTOR (LDR)

It is a resistor that changes its resistivity depending on the incoming light. The light will generate an excess of free carriers within the resistor, so if we have more free carriers, the resistivity decreases. The main issue of this sensor is that the light to current dependency is not linear.

In the dark → R very high (up to 1MΩ)
Exposed to light → R drops dramatically (few ohms)

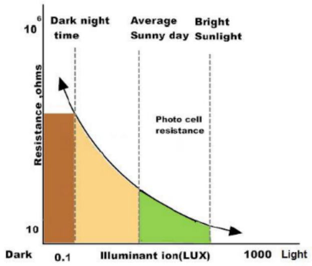
Working principle:

Photons excite electrons from the valence band to the conduction band

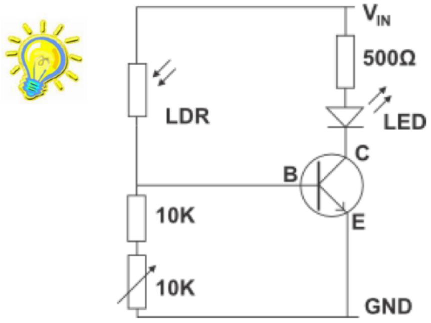
→ more free electrons in the material

→ lower resistance

But…
- low sensitivity
- non-linear characteristic
- sensitive to temperature changes



It is hard to precisely recognize the amount of light we have, so typically in the dark region the resistance is of MOhms, and the resistor is typically used for on-off detections, to check if the light is on or off (if on the value of the resistor is very low). The readout circuit is very simple, and it is a voltage divider. If we are in dark conditions, the base of BJT is almost at 0V, while if we have light the LDR will be small, and the base will be high. For instance, it can be used to switch on a BJT, that is on if we have 0.7V between base and emitter.



WITHOUT LIGHT

High LDR
$V_{BE} < 0.7$ → BJT off
→ LED off

WITH LIGHT

Low LDR
$V_{BE} = 0.7$ → BJT on
→ LED on

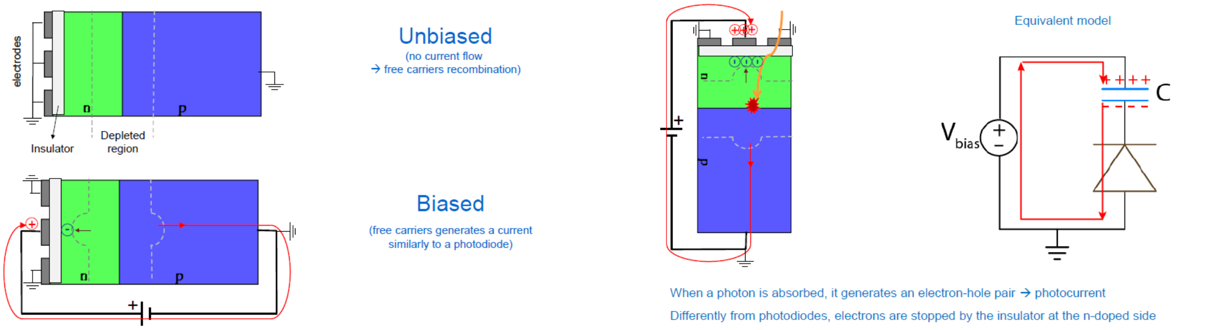
IMAGE SENSORS

An image sensor is a sensor made by many pixels, and thanks to the pixelated structure is able to create an image. There exist two main technologies:

1. CCD (charge couple devices)
2. CMOS active pixel image sensors

CCD are more used for scientific applications, while CMOS for consuming electronics.

CCD



When a photon is absorbed, it generates an electron-hole pair → photocurrent

Differently from photodiodes, electrons are stopped by the insulator at the n-doped side

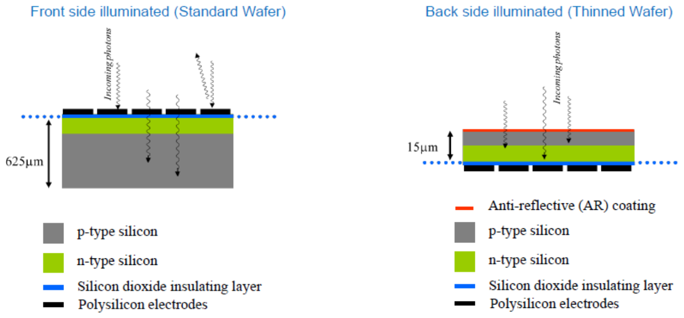
Is built to cope with a microscope, so we don't have lenses on the camera. The working principle of each pixel of a CCD is similar to a PD, but different. If we look at the cross section, we have a p region in contact with a n region. So we have a pn junction with a depleted region in the middle where we absorb photons and generate e/h pair. But then we can contact directly connect the p part to ground, but we cannot contact directly the n region, because we have an insulator, and electrodes are placed on an insulator. So we have electrodes an n region with insulators in the middle → similar to a capacitance, hence the pixel of a CCD can be modelled as a capacitor in series with a diode. Typically the insulator is Si02, silicon dioxide.

So we cannot bias directly the n region, but we bias it like in the FET. We have this structure because if we want to separate pixels one adjacent to the other, we can use the electrodes. For instance, we get a photon in the middle region corresponding to the middle electrode, so e- generated in the pixel will be collected below the central electrode. So we can generate e- wherever in the device, but they will be collected below the middle electrode, and in this way we are prone to separate charges from different pixels.

So we bias with 0V the p region and the two external electrodes, while the central one is biased with positive voltage → when charge is generated it is collected below the central electrode. This means that automatically within the CCD pixel we have an integration of charges in the middle electrode. For the readout, we will transfer and readout the charge collected on the capacitor of the pixel.

Front/back side illumination

There are 2 structures of CCD detectors, front side illuminated and back side illuminated. In FSI we have the silicon structure (e.g. p type substrate n doped), then we deposit the layer of insulator and on top of that the electrodes. If light comes from the top of the structure, it is important that electrodes are not metallic, otherwise light is back reflected → electrodes made of polysilicon, that is almost transparent, so some e- will still be backscattered and reflected. Typically the width of this FSI is of 625um.


Front side illuminated (Standard Wafer)
Incoming photons
625µm
p-type silicon
n-type silicon
Silicon dioxide insulating layer
Polysilicon electrodes
Back side illuminated (Thinned Wafer)
Incoming photons
15µm
Anti-reflective (AR) coating
p-type silicon
n-type silicon
Silicon dioxide insulating layer
Polysilicon electrodes

→ 3D vertical integration
anti-reflective coating

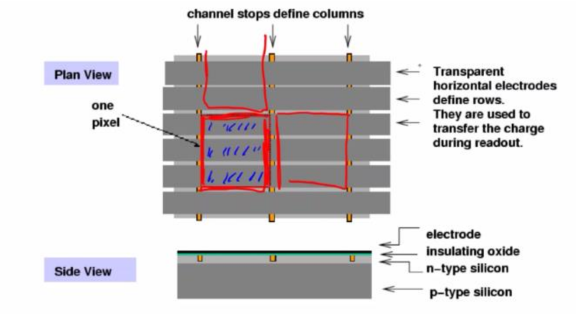
If we use the other structure, BSI, the Si wafer is similar to FSI, but then we flip the structure and we remove all the silicon we have on the bulk (p part) until reaching a width of 10 to 20 um. This structure is preferrable because light doesn't meet the electrodes, because it comes from the bottom. Furthermore, the back side is perfectly flatten, because we flattened it with a process, while in FSI we have electrodes. Since it is flatten, we can put an antireflecting coating, that helps to match the reflective index of air an Si, to reduce the backscattered light.
Moreover, it is important to remove bulk Si to prevent that photons are absorbed before reaching the depletion region. In this case the width of the system is too small, so it may break, so we need to add another wafer below the electrodes for mechanical stability. However, in this wafer we can ut the readout electronics for instance to connect the electrodes and drive them.

Both in FSI and BSI the process used for manufacturing the CCD is not the standard one for diodes, but it is an optimized process for imaging.

CCD structure

The one in red is a pixel. We recognize the 3 electrodes. Pixels must be kept separated, and vertically charges are separated by unbiased electrodes, grounded electrodes.

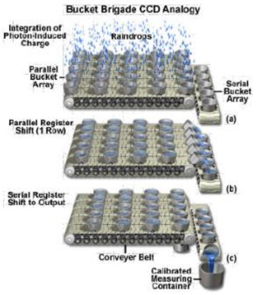
To prevent charge outflow horizontally, we use channel stops which are depleted region used to separate the charges. It is a hardware separation, while the vertical one can be changed by changing the voltage on the electrodes, and this change in voltages is used to readout the charge.

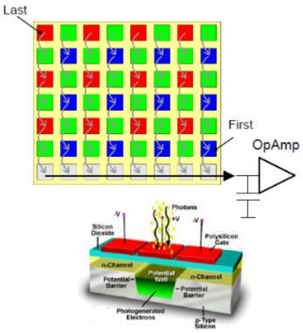

channel stops define columns
Plan View
one pixel
Transparent horizontal electrodes define rows. They are used to transfer the charge during readout.
Side View
electrode
insulating oxide
n-type silicon
p-type silicon

CCD readout

The charge is collected below the central electrode, and this phase is called **detection phase**. Charge is collected in a potential well below the central electrode. Then in the last bottom row we imagine to have charge, and we want to readout it. What we can do is to transfer the charge on the last pixel on a capacitor and then readout the voltage with an amplifier. When we transfer the charge on the last pixel on the capacitor, we also transfer the charge of the previous pixel in the last one and so on → chain of transfers of charges. This is the horizontal readout.

When the horizontal readout is concluded, we can vertically transfer the charge from one row to the bottom other.

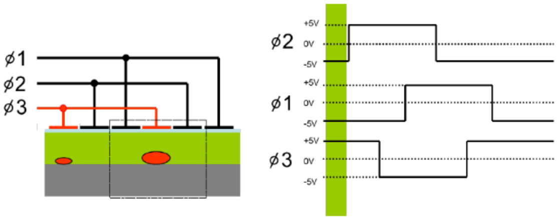

Bucket Brigade CCD Analogy
Integration of Photon-Induced Charge
Raindrops
Parallel Bucket Array
Serial Bucket Array
(a)
Parallel Register Shift (1 Row)
(b)
Serial Register Shift to Output
Conveyer Belt
(c)
Calibrated Measuring Container

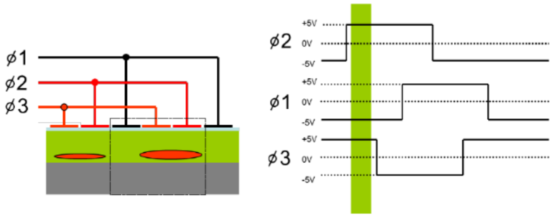


Readout steps

In step 1 we have charge integration of the CCD, and charge stored is proportional to the incoming light. In this case we are biasing only the central electrode of the pixel.

In step 2 we bias also the 2nd electrode, so charge spreads and is shared between electrode 2 and 3. In step 3 we unbias electrode 3, so the charge is only below electrode 2 → we have done a shift of one electrode. We still have electrode 1 unbiased, so the charge between adjacent pixels doesn't mix up. Then I have to bias electrode 1 in order to share charge between electrode 2 and electrode 1 and so on.





In this way I can move charge between electrodes always keeping one electrode unbiased to keep charges separated.

CCD performance

We analyze them in terms of ability to generate charge: we will discuss about QE (ability to generate charge) and low ability to generate dark current.

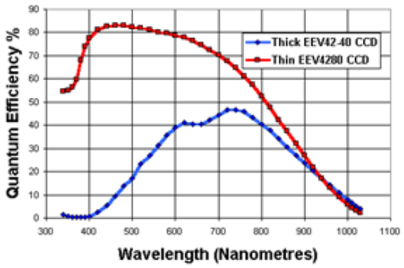
As for charge collection, this refers to the ability of collecting charges below the central electrode. Another important parameter is the ability of the CCD to transfer charge for the readout. Finally we will analyze the readout itself.

Charge generation - QE

QE corresponds to the responsivity of a PD, but in PD is defined relative to the incident optical power, while here in terms of number of incoming photons. QE is defined as the ratio between the number of detected photons (so the ones that generate e/h pairs) and incident photons.

QE depends on the wavelength, and we are sensible mainly in the visible range or near infrared. Depending on the structure (dopants level and biasing) of the CCD, the QE can change a lot.

Obviously, **QE depends on reflection**, **absorption** (probability to absorb the photon in the depleted region and not before or after it) and **transmission** (passing through the detector without being detected). Photons with very low power have high probability to passe through without being absorbed because the energy is not sufficient to generate e/h pair.



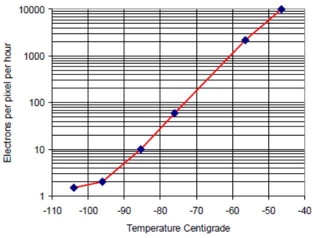
$$QE = \frac{detected\ photons}{incident\ photons}$$

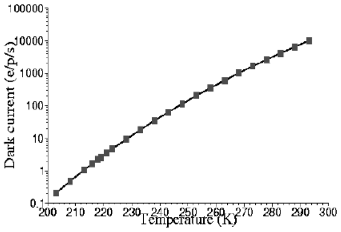
Probability that
a photoelectron will be released
for each incident photon

Mechanisms which hamper photon collection:

- Reflection
- Absorption
- Transmission

Charge generation – Dark current





Thermal effects cause an electron to move from the valence band to the conduction band.

The majority of dark current is created near the interface between the Si and the SiO_2, where interface states at energy between the valence and conduction bands act as a stepping stone for electrons.

Other parameter regarding charge generation. When we put in dark a CCD we expect no charge in the potential well, but still we can collect some charge in the potential well due to thermal generation. Some e/h pair can be generated not due to photon absorption but just due to Brownian motion and thermal generation of electron hole pairs.

In the graphs we can see that one is at very low temperatures (left) and one at temperatures close to the ambient temperature. We can see that dark current strongly depends on T. If CCD is used for scientific application, we can cool down the temperature, for instance up to –100°C, and if we do so, we can have an electron per pixel per hour generation that is down to about 1 electrons. So due to thermal generation at this temperature we generate only 1 electron per hour.
If instead we use them at ambient temperature, for instance 300K, generation of dark electrons increases a lot, 10'000 e- per second.
Hence we have a minimum value of incoming light that we can measure, because it must be bigger than the dark current.

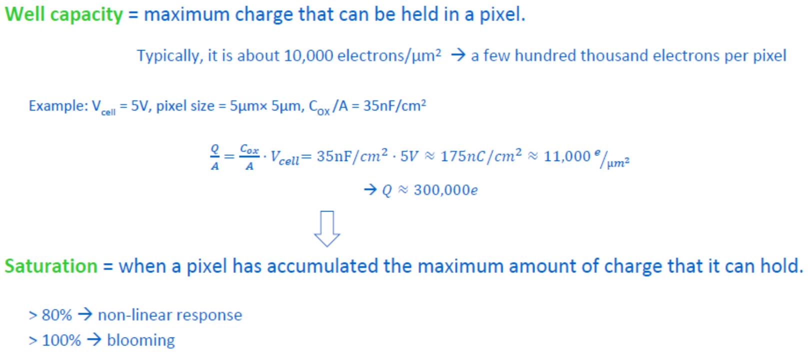
The majority of dark current is due to the interface between the SiO2 and Si, because we have more defects because we have two different materials put in contact.

Charge collection – Well capacity

The well capacity is the maximum charge that a pixel can collect. A typical value is about 10'000 e-/um^2.
If we imagine to have a pixel of a CCD biased at 5V (on the central electrode). If the pixel is of 5x5 um^2 (but there exist also smaller pitches, but to increase the pixel capacity is better to have bigger dimensions). Cox/A = 35nF/cm^2.

The charge that can be collected on a pixel is given by capacitance multiplied by voltage. In the end we obtain 11’000 e/um^2.

Then the total charge we can collect in out cell is obtained by multiplying this value for the area. So in each pixel we can collect up to 300’000 photons (each photons corresponds to an e/h pair). If we arrive close to this maximum value of charges we are in saturation. Already at 80% we have a nonlinear response, because the baising of the junction start to decrease, because we are collecting charges.



Well capacity = maximum charge that can be held in a pixel.

Typically, it is about 10,000 electrons/µm² → a few hundred thousand electrons per pixel

Example: V_{cell} = 5V, pixel size = 5µm× 5µm, C_{ox}/A = 35nF/cm²

$$\frac{Q}{A} = \frac{C_{ox}}{A} \cdot V_{cell} = 35\text{nF}/cm^2 \cdot 5V \approx 175nC/cm^2 \approx 11{,}000\, {}^{e}/_{\mu m^2}$$

$$\rightarrow Q \approx 300{,}000e$$

Saturation = when a pixel has accumulated the maximum amount of charge that it can hold.

> 80% → non-linear response

> 100% → blooming

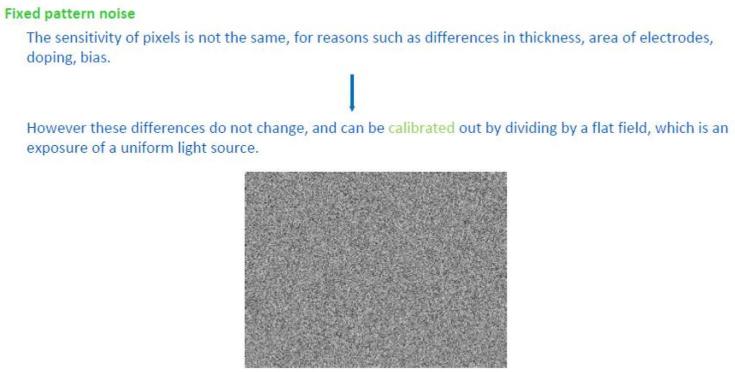
At 100% we arrive at a state called blooming. Blooming is the fact that our charge cannot stay in the well capacitance, because since we reach the maximal well capacity, charge starts to move also below the other electrodes → we can mix up charges between adjacent pixels → we get a bloomed star image.

It happens only vertically and not horizontally, because channel stops separate better than unbiased electrodes.

Charge collection – pixel uniformity

When we illuminate with a constant flat light the entire camera, we should expect to have a flat image, in which each pixel provide the same intensity, but this is not the case. In fact, in this case we would obtained a raw data (not processed) in which pixels have different colors one with respect to the other, and this is due to differences between pixels, like differences in the thickness of electrodes, biasing, doping level and so on, and these differences cause a different sensitivity of the pixel (e.g. the more doped the more sensitive). However, all these parameters are fixed.

Hence when CCD are manufactured, a sort of calibration is done by the manufacturer; they expose the CCD to a constant illumination, they acquire an image (down below) and the image is store in an internal memory that can be either in the chip of the CCD or at the system level in the camera and



Fixed pattern noise

The sensitivity of pixels is not the same, for reasons such as differences in thickness, area of electrodes, doping, bias.

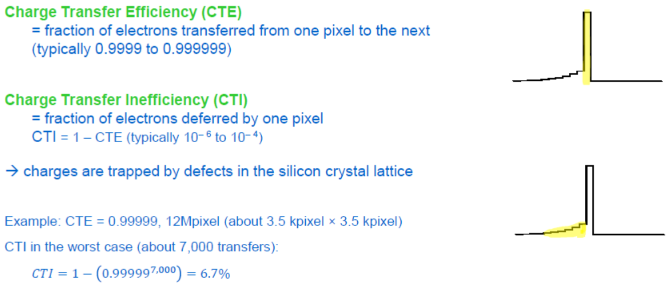
However these differences do not change, and can be calibrated out by dividing by a flat field, which is an exposure of a uniform light source.

this image is used to do a calibration of the image. It is not a simple subtraction, but with this image we can compensate the sensitivity variations. This process is known as flat fielding.

Charge transfer efficiency (CTE) and inefficiency (CTI)

Every time we transfer the charge, part of the charge is left in the previous pixel. Typically CTE is very high, almost 100%, but this is associated to just 1 transfer. If we have a large camera, we have a lot of transfers, hence we may loose a huge amount of charge.
For simplicity, it is better to consider CTI = 1 – CTE.

If we consider the case of a 12Mpixel camera and a CTE = 0.99999. Which is the worst case? The worst case is the top left pixel, if we assume to have the readout circuit in the bottom right (worst case is the opposite pixel with respect to the readout system). This charge will be transferred 3,5 k times vertically and 3,5 k times horizontally.
The overall charge lost is given by 1 – (CTE)^(n° of transfers) and we lost more or less 6,7%.
To increase CTE the only way is to slow down the readout. Indeed, charge is lost because maybe the transfer between two electrodes is not complete.



Charge Transfer Efficiency (CTE)
= fraction of electrons transferred from one pixel to the next
(typically 0.9999 to 0.999999)

Charge Transfer Inefficiency (CTI)
= fraction of electrons deferred by one pixel
CTI = 1 – CTE (typically 10^{-6} to 10^{-4})

→ charges are trapped by defects in the silicon crystal lattice

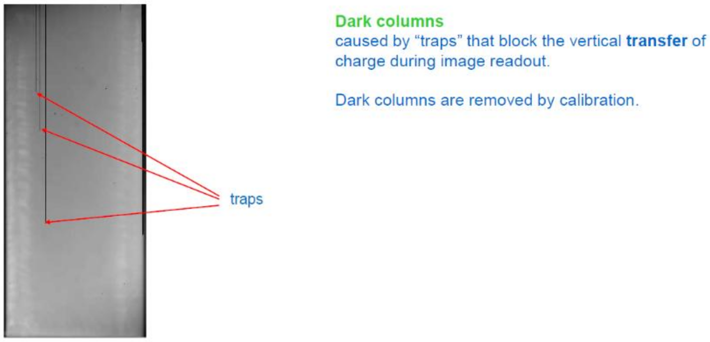
Example: CTE = 0.99999, 12Mpixel (about 3.5 kpixel × 3.5 kpixel)
CTI in the worst case (about 7,000 transfers):

$CTI = 1 - \left(0.99999^{7,000}\right) = 6.7\%$

If we have a camera with good CTE we have a high contrast, because the charge of one pixel doesn't influence the charge of the adjacent one. If we have bad CTE, part of the charge remains in the previous pixel.

Dark columns

Another aspect to be considered are the possible defects in the Si structure. These defects can trap electrons and holes that are moving, so we have dark columns. When we try to transfer the charge of pixels before the trap, all charge is trapped, so we cannot see the charge collected by pixels before the trap.
Typically dark columns are removed by post processing just with a sort of calibration, by averaging or interpolating the adjacent pixels.

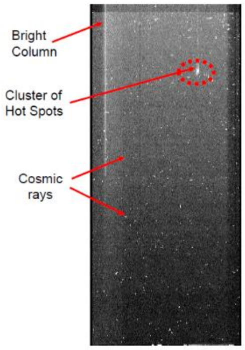


Dark columns
caused by "traps" that block the vertical **transfer** of charge during image readout.

Dark columns are removed by calibration.

Other defects

- Bright columns: we have a defect able to generate a lot of charges. Hence during transfer these pixels are readout as very bright. Again this error can be removed by interpolation.
- Hot spots: due to pixel with higher dark current. If the dark current tis higher than the normal one, we have bright pixels. Sometimes the dark current is so high that generates secondary photons, and hence a sort of aloe around the pixel. The secondary photons are indeed eventually reabsorbed by the neighbour pixels.
- Cosmic rays: bright spots but not always in the same position in each image, so not due to traps or high dark current, but due to cosmic rays. In fact, CCD are able to detect cosmic rays. They can be detected in different positions in every different image, they are not due to defects in the camera itself.



Bright columns
Caused by traps. Electrons contained in such traps can leak out during **transfer** causing a vertical streak.

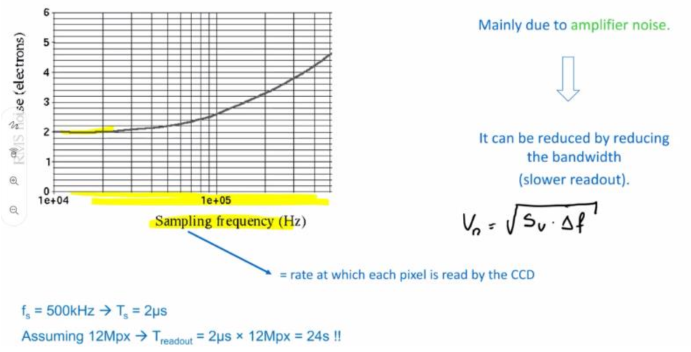
Hot Spots
Pixels with higher than normal **dark current**. Their brightness increases linearly with exposure times.

Somewhat rarer are light-emitting defects which are hot spots that act as tiny LEDs and cause a halo of light on the chip.

NOISE SOURCES

Readout noise

It is mainly due to the amplifier, and it can be reduced if we reduce the bandwidth of the amplifier. If we reduce delta-f, so the bandwidth of the amplifier, we reduce the noise. To reduce the bandwidth, we also make the amplifier slower, so we also need to readout the CCD camera slowly.
If we increase the sampling frequency, we also increase the noise due to the readout. For instance, if we want to have a noise corresponding to the last point of the plot, we need to work at a frequency of 500kHz. If we assume to have 12Mpx, to readout all the pixels we need 2us per pixel, and hence 24s for all the camera. So it is a value too high. This is the reason why CCD camera are used for scientific applications.



Mainly due to amplifier noise.

It can be reduced by reducing the bandwidth (slower readout).

$V_n = \sqrt{S_v \cdot \Delta f}$

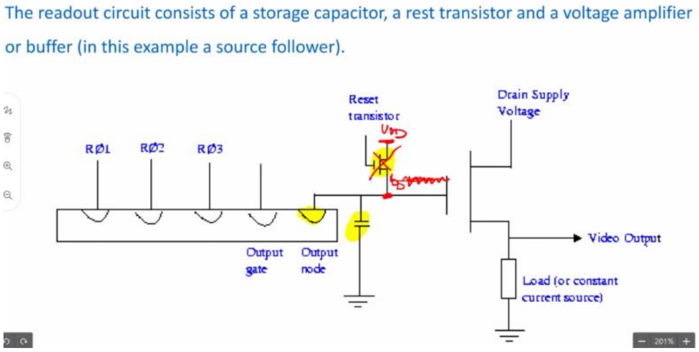
= rate at which each pixel is read by the CCD

f_s = 500kHz → T_s = 2μs
Assuming 12Mpx → $T_{readout}$ = 2μs × 12Mpx = 24s !!

Noise associated to the capacitor (reset noise)

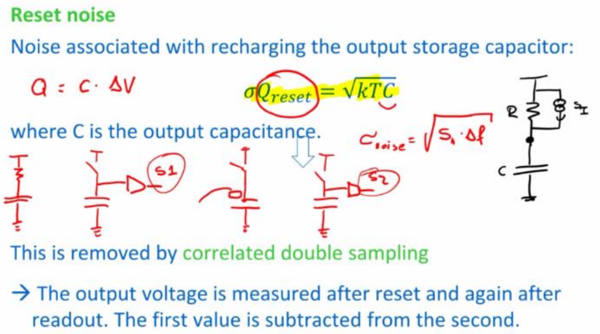
Apart from the noise related to the amplifier, we have another source of noise related to the fact that we store the charge on a capacitor. Apparently it is strange the fact that storing charge on a capacitor introduces noise, because typically capacitors don't introduce noise as resistors and pn junction do.

However, the capacitor needs to be reset before every readout. Anytime we reset the capacitor, we switch on the reset transistor, which works in the ohmic region and so it can be considered as a resistor → introduces a thermal noise. This means that during the reset, we don't provide exactly Vdd to the top plate of the capacitor, but Vdd+noise. When we switch off the transistor, the capacitor will sample a specific value that can be different from Vdd. So every time we close the transistor, we have different value that is noise that sums up to well charges.



The readout circuit consists of a storage capacitor, a rest transistor and a voltage amplifier or buffer (in this example a source follower).

Computations



Reset noise

Noise associated with recharging the output storage capacitor:

$$\sigma Q_{reset} = \sqrt{kTC}$$

where C is the output capacitance.

This is removed by correlated double sampling

→ The output voltage is measured after reset and again after readout. The first value is subtracted from the second.

If we try to compute the std, considering also the pole of the circuit that is limited by the RC, the std is equal to sqrt(kTC). So the noise is due to the resistor, but the value of the noise depends on the capacitor. So the capacitor is not able to introduce noise, but noise depends on it due to computations, because sigma of the noise is sqrt(S*delta-f), and delta-f depends on the capacitor. Moreover S and delta-f depend on the resistor value, so they simplify.

There is a simple way to eliminate this source of noise (reset noise) and it is the correlated double sampling, that consists in sampling two times the voltage across the capacitor.

We start resetting the capacitor, then we open the transistor, sample the value of the noise before transferring the charge, then we transfer the charge and sample again. In this way we can subtract sampling 1 and sampling 2 to subtract the noise. However, this further decreases the readout speed. In fact, for the next pixel we need to replicate these steps.

So the resistor is the source of noise, but the noise depends on the capacitance.

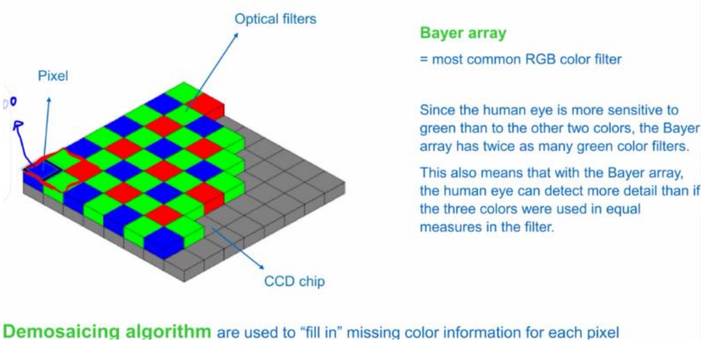
COLOR FILTERING

How to distinguish colors with a CCD camera?

We have different efficiency depending on the wavelength but all the wavelength can eb detected by a CCD and the result is the same for all them, they generate e/h pair.

Color images can be generate thanks to optical filters placed on the top of the CCD camera. Typically we use Bayer array.

Let’s consider an elementary cell made out of 4 pixels; in it we have 3 different filters, one lets only blue light pass, then red and green → we are doing an RGB image. So one pixel is able to detect blue, one red and two green. Two pixels for green because our eyes are more sensible to green color, so to match the way in which human eye sees images, more pixels are dedicated to the green.
So a 12Mpx camera is not a 12Mpx camera for each color, but overall. Then the result for each color can be obtained by interpolation.



Bayer array

= most common RGB color filter

Since the human eye is more sensitive to green than to the other two colors, the Bayer array has twice as many green color filters.

This also means that with the Bayer array, the human eye can detect more detail than if the three colors were used in equal measures in the filter.

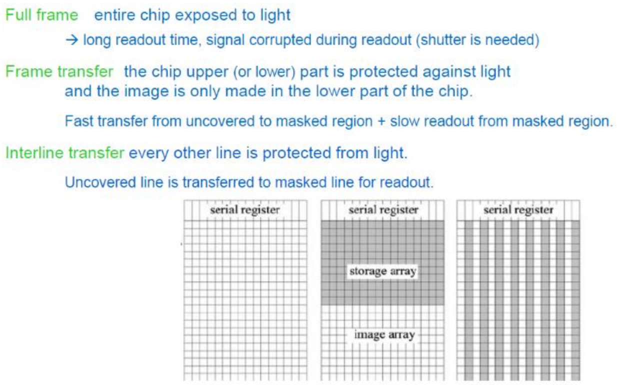
Demosaicing algorithm are used to “fill in” missing color information for each pixel

Using some interpolation, we can find the value of green, blue and red for each pixel and, combining the RGB component we can find the final value for each pixel. This procedure is called demosaicing algorithm, that is used to combine together the 3 RGB colors, so interpolating different pixel to find the value.

Types of CCD

So in a CCD in a first phase we collect charges, then we have the readout. During the readout in principle we don’t want the camera to detect other photons, to prevent mixing up the information from different pixels. In order to avoid this issue, we have 3 possible types of CCD cameras:

- Full frame: the entire chip is exposed to light, so all the camera can detect photons. In this case we need a mechanical shutter, that is something mechanical that is able to cover the entire CCD mechanically, preventing photons to reach the CCD. Typically is not very used.
- Frame transfer: some pixels are used just for the transfer. Part of the pixels are exposed to light, while the other half is covered, for instance with a metallic layer. The image array is always exposed to light, then we do a fast vertical transfer from the image array to the storage array during the readout phase. So we transfer the charge only vertically. We do this transfer as fast as possible, of course taking into account the charge transfer efficiency. Then during the next acquisition time we can read slowly the storage array.
- Interline transfer: we have every other column able to detect light, and intermediate ones covered by metal lines. We do a very fast horizontal transfer when we want to readout the pixels. In this case the transfer time is just one clock, much faster than the frame transfer topology, and then during the next exposure time of the white columns we can readout slowly the black columns. However, in this case we are reducing the resolution of our camera, because the pixel pitch is increased horizontally, because a pixel every two is not used for collecting charge. Again, we can use interpolation to find a reasonable value for the black pixels.



Full frame entire chip exposed to light

→ long readout time, signal corrupted during readout (shutter is needed)

Frame transfer the chip upper (or lower) part is protected against light and the image is only made in the lower part of the chip.

Fast transfer from uncovered to masked region + slow readout from masked region.

Interline transfer every other line is protected from light.

Uncovered line is transferred to masked line for readout.

CCD applications

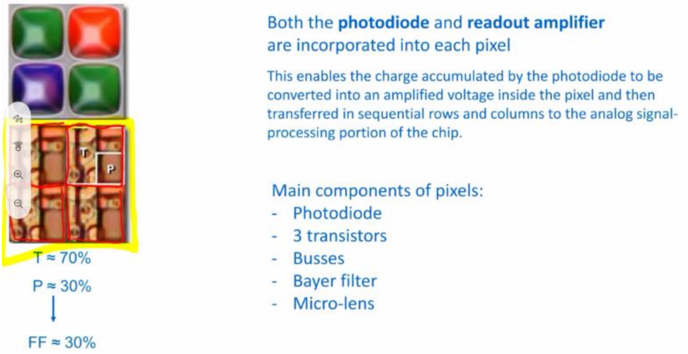
They are mainly used for scientific purposes.

CMOS ACTIVE PIXEL SENSOR

Characterized by the fact that is fabricated in CMOS technology, that is the technology used for the 99% of the electronics. In this technology we are able to integrate both a standard photodiode and some transistors in order to perform the readout. The mean advantage is that in the readout we don't have to transfer the charge, but we have a conversion from charge to voltage within the pixel because we can integrate transistors in the pixel.

ACTIVE PIXEL SENSOR (APS)

The red portion is one pixel. Within the pixel we have the PD (P) but also transistors, that are used to convert the charge into a voltage. Typically we have about 70% of the area dedicated to transistors and 30% dedicated to PD. This means that a typical fill factor of an APS is of 30% (sensitive area divided by overall area of the pixel). A lower FF means a lower probability to detect photons.



Both the **photodiode** and **readout amplifier** are incorporated into each pixel

This enables the charge accumulated by the photodiode to be converted into an amplified voltage inside the pixel and then transferred in sequential rows and columns to the analog signal-processing portion of the chip.

Main components of pixels:

- Photodiode
- 3 transistors
- Busses
- Bayer filter
- Micro-lens

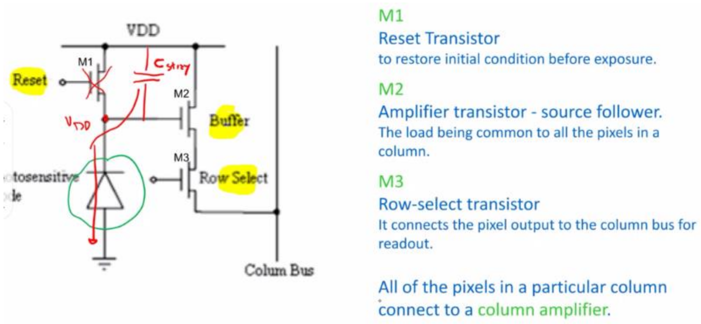
However, we are lucky, because we can mount some microlens arrays on the top of the pixel. The microlenses array are able to focus all the light coming to the pixel towards the photodiode. These microlenses are shaped like a dome, and they have also color filters.

In standard configuration, the transistors are 3. Then we also have busses, metal line to drive the signals.

3 transistors APS

We have the PD that is reverse biased, so the anode voltage is higher than the cathode voltage. The intensity of the photocurrent will depend on intensity of light. Then we have 3 transistors.

We have a reset transistor, a buffer transistor and a row selector transistor. Before integration time, so acquiring an image, we need to reset the anode voltage of the PD, in order to start with an equal condition for all the pixels. So the reset transistor brings the PD to Vdd. Once we are catching light, the reset transistor is turned off. So PD current will go in the stray capacitance related to M2. So we have a configuration that recalls the structure of a CCD, where we have a pn junction that is modelled with a PD, and an insulator that creates a capacitance.



M1
Reset Transistor
to restore initial condition before exposure.

M2
Amplifier transistor - source follower.
The load being common to all the pixels in a column.

M3
Row-select transistor
It connects the pixel output to the column bus for readout.

All of the pixels in a particular column connect to a column amplifier.

During the acquisition time, the anode node is discharged by an amount proportional to the current flowing into the stray capacitance. Then the M2 determines a charge to voltage conversion. Than the row selector is needed because the column bus is shared among all the pixels of the same column. We cannot readout simultaneously the voltages of all the pixels, but we need to switch them on one at a time. We have to select one row at a time and then on the column bus we readout the voltage of the column we have selected. We are not transferring charges from one pixel to another, but we are just enabling a transistor → much faster process because we don't have the problems related to charge transfer (e.g. inefficiency in the transfer).

NOISE SOURCES

- Fixed pattern noise: we had it also in CCD camera, and it was due to differences between pixels, like in the dopants levels, thickness of contacts, etc. also in CMOS we can have differences between pixels, but also mismatches between the transistor. So typically the FPN of CMOS camera is higher than the one we have in CCD. However, also in the case of CMOS we can compensate it with a calibration that is the flat fielding like in CCD.
- Reset noise: we still have the same noise contribution of the CCD (in the slide is expressed in voltage, no more in charge as in CCD). It is a noise due to the resistance of the reset transistor, so it is a thermal noise, but in the equation appears only the capacitoance because the capacitance is in the bandwidth. It is expressed in voltage because the output of CMOS is in voltage, while in CCD the output was in charge.
- Amplifier noise: it is the 1/f noise, that is typical of MOS transistor. It has a spectral density of noise that is not flat in frequency, but it depends on 1/f, it is a slope higher at low frequency. In CMOS technology, the 3 transistors must be MOS transistors because they can be manufactured as very small. In CCD this noise was not mentioned because we have 1 transistor for the overall CCD, we don't have to integrate 3 mosfets for each pixel, and moreover we can use JFETs, that have a smaller noise in the case of CCD, while in CMOS we are forced to use MOSFETs due to dimensions constrains.
- Photodiode shot noise: proportional to the current intensity we have in the PD and due to the granularity of charge.

- Fixed Pattern Noise (FPN) caused by:
  - Variation in photodiode efficiency and noise
  - Mismatches in source follower transistor performance

  FPN is constant and reproducible
  → flat field correction (= on chip subtraction of an image with constant illumination)

- Reset noise (kTC noise)

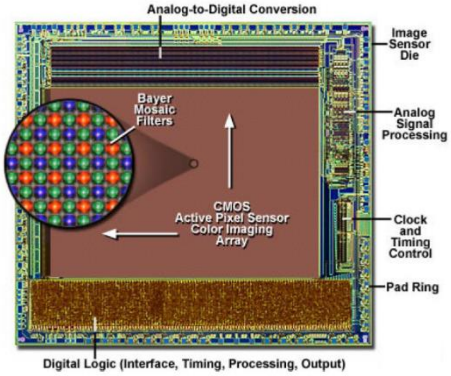
$$v_n = \sqrt{\frac{KT}{C}}$$

- Amplifier noise (1/f noise)

- Photodiode shot noise

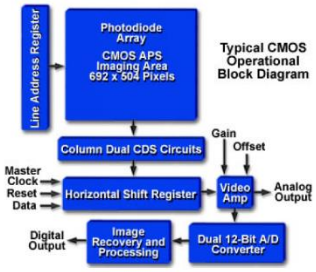
CMOS SENSOR GENERAL STRUCTURE

In the CMOS chip we have a central part in which we have the arrays of pixels (brown one). Then in the peripheral part of the chip we have the processing electronics, like the column buffers in the upper part, ADC converter to convert the analog voltage in a digital code directly on the chip.
We can also have some preprocessing, like filtering, and clock and timing control to provide reset and row selection.
Then in the border of the array we have the pad, to wire bond the chip to the external PCB.


Analog-to-Digital Conversion
Image Sensor Die
Bayer Mosaic Filters
Analog Signal Processing
CMOS Active Pixel Sensor Color Imaging Array
Clock and Timing Control
Pad Ring
Digital Logic (Interface, Timing, Processing, Output)

FUNCTIONAL BLOCKS

- VGA resolution (640x480)
  + shielded pixels for black level compensation
- Line address register for readout
- Correlated Double Sampling circuit
  to reduce the pixel reset noise
- Horizontal Shift Register
- Video Amplifier with adjustable gain and offset
- ADC converter
- Image processing
  (interpolation, smoothing, white and black balancing…)


Line Address Register
Photodiode Array
CMOS APS Imaging Area 692 x 504 Pixels
Typical CMOS Operational Block Diagram
Column Dual CDS Circuits
Gain
Offset
Master Clock
Reset
Data
Horizontal Shift Register
Video Amp
Analog Output
Digital Output
Image Recovery and Processing
Dual 12-Bit A/D Converter

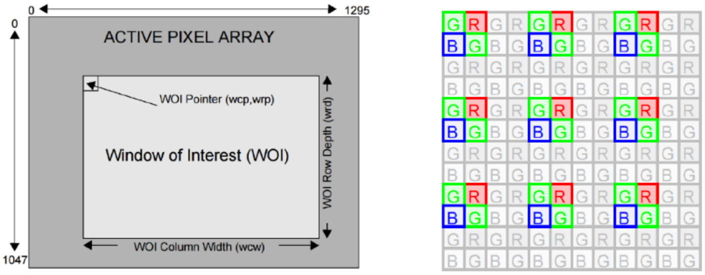
Alternative:
One Video Amp + ADC for each column
(high frame rate)

We have the PD array, so the array of pixels, then for instance a VGA resolution (640x480) plus some pixels covered with metal to do some compensation for instance to control dark current. Then the line address register is a sort of row selector, then we have the circuit for correlated double sampling, to do a readout after the reset and after the acquisition to do a compensation of the noise due to the reset transistor.
Then we have shift register to provide the voltage of each pixel to the video amplifier and so on. In modern sensor we can have many amplifier and many ADC to speed up the readout.

ACTIVE WINDOW

Another advantage of CMOS with respect to CCD, is that we are not forced to readout the entire array, because we can decide to select only some row and readout only some columns. So we can read only the window of interest (WOI). In this way we can have a high frame rate.

Another way to speed up the readout is to do a window subsampling, so instead of reading out all the pixel we readout every other pixel. Typically a group of 4 pixels is skipped and then interpolating.



Window of interest

Window subsampling

→ speed up the readout

Nowadays, interpolation is not just a mathematical operation, but we can also use other techniques to do interpolation for instance by neural networks (not just a mathematical average between two pixels).

READOUT MODES

Let's imagine we are reading out the camera, and we reset all the pixel at the same time instant. Then we integrate the charge for all the pixels and then we readout the pixel after a period delta-t. After delta-t we readout the first row, and then all the other. So the first row will have a certain integration time, but the integration time progressively increases if we readout the row later. So this will create differences in pixels. This error can be mathematically compensated, but it can also be internally compensated so that the transistors reset at different times, with the same time delay we have in the readout.

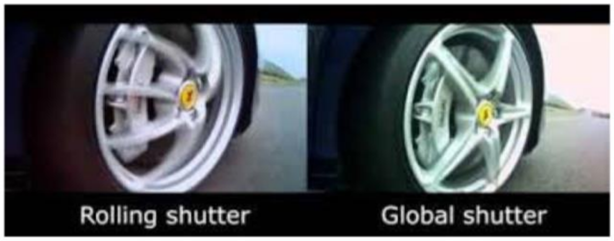
This way of operating is called rolling shutter: all the pixel have the same integration time because the reset is shifted. If we are acquiring a static image we don't have problems, but if the image is dynamic, we have some artifacts, because different pixels are looking at the image in different time intervals.

Rolling shutter

each frame is captured by scanning rows across the scene rapidly, so not all parts of the image of the scene are recorded at the same instant. This produces predictable distortions of fast-moving objects or rapid flashes of light.

Global shutter

the entire frame is captured at the same instant.



To avoid so, we can use mechanical shutters, and in this case we have a global shutter architecture. We reset all the pixel together at the same time instant, then we close a mechanical shutter so that all the pixels integrate light for a certain amount of time. Integration time is stopped at the same time for all the pixels

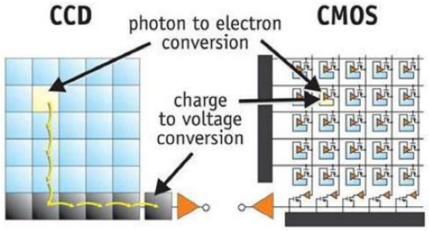
Having mechanical shutter is not the best idea because they can degradate easily after several repetitions. For this reason, in some pixels we have an architecture with 4 transistor, and the 4th transistor is used to store the charge at a specific time. So the integration time is the same, but now we can do a slower readout.

CMOS APS APPLICATIONS

- High Volume Imagers for Consumer Applications (e.g., camera-phones)
- WebCams
- Imagers for Machine Vision (area and line scan)
- High Speed Motion Capture Cameras
- Digital Radiography
- Endoscopy (pill) Cameras
- DSLRs (Digital Single-Lens Reflex) camera

COMPARISON

In CCD we integrate the charge and then transfer it and only at the boarder we have the conversion into voltage, while in CMOS the conversion is directly in the pixel.



CCD advantages:

- Uniformity (less structured noise)
- Higher Fill-Factor
- Almost all chip area devoted to light capture
- Higher sensitivity in the Near Infra-Red (NIR)

CMOS advantages:

- Faster readout (one charge to voltage converter per pixel)
- Selectable active window (for higher frame rate)
- Digital output
- Less electronics outside sensor (internal Amplifier, ADC, processing…)
- Less power (→ lower temperature)

Cost considerations

- Custom imager → CCD
- Large scale economy → CMOS

Moreover CCDs have higher fill factor because we don’t have transistors in the pixel, and moreover they have higher sensitivity because they are built for a specific application. Conversely, CMOS have a less power because we don’t have to transfer charges.
Moreover, if we need a small replica of the sensor, CCD is better, instead in a market scale CMOS is better.

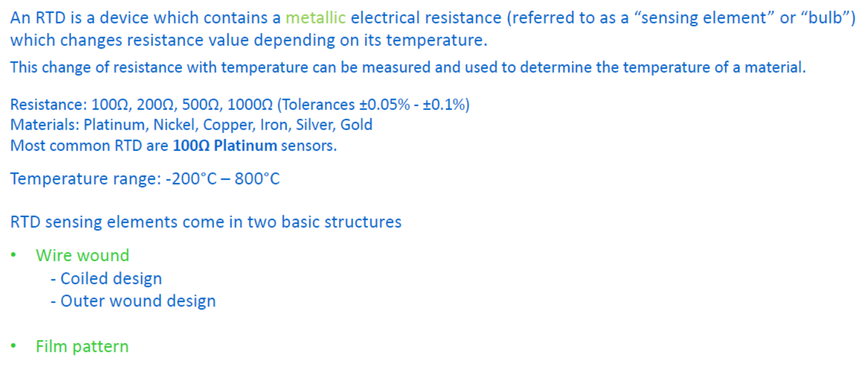
TEMPERATURE SENSORS

- Resistance temperature detectors: mechanic resistors
- Thermistors
- Thermocouple
- Diode and bandgap temperature sensors: they can be integrated in CMOS technology
- Infrared thermometer

RESISTANCE TEMPERATURE DETECTORS (RTD)

RTDs are metallic resistances and typically they are made out of Platinum (or Nickel, Copper, Iron, Silver, Gold). The parameter that is specified is the resistance at ambient temperature, and the classical sensor used is the PT100, mad out of Pt with a resistance of 100Ohm at room temperature.
The value of the resistance changes as the temperature changes. Typically RTD are positive temperature coefficient, PTC, so if we increase the temperature we increase also the resistance.

The resistance increases because in a metal we have a lot of free e- and since the CB is already full of free e-, if we increase T, we increase the motion of the e- and so the probability of the electrons to interact one with the other, so one electron can stop the other one that is in motion. This is different from NTC, where resistance decreases with temperature increase.



An RTD is a device which contains a metallic electrical resistance (referred to as a “sensing element” or “bulb”) which changes resistance value depending on its temperature.

This change of resistance with temperature can be measured and used to determine the temperature of a material.

Resistance: 100Ω, 200Ω, 500Ω, 1000Ω (Tolerances ±0.05% - ±0.1%)
Materials: Platinum, Nickel, Copper, Iron, Silver, Gold
Most common RTD are **100Ω Platinum** sensors.

Temperature range: -200°C – 800°C

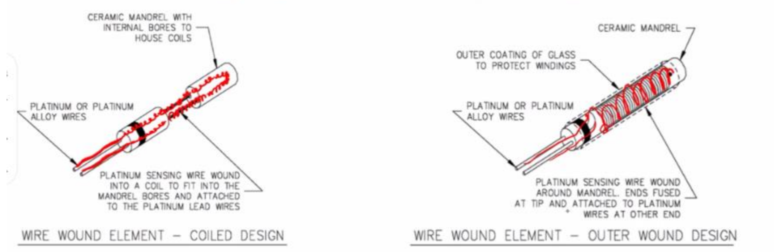
RTD sensing elements come in two basic structures

- Wire wound
  - Coiled design
  - Outer wound design
- Film pattern

Typical temperature range for RTD is from -200°C to 800°C.
Mainly we have RTD with 2 possible structures: wire wound and film pattern.

Wire wound

Wire wound elements contain a very small diameter metallic wire (typically, 0.5 – 1.5 mil diameter, 1mil = 0.0254 mm).


CERAMIC MANDREL WITH INTERNAL BORES TO HOUSE COILS
PLATINUM OR PLATINUM ALLOY WIRES
PLATINUM SENSING WIRE WOUND INTO A COIL TO FIT INTO THE MANDREL BORES AND ATTACHED TO THE PLATINUM LEAD WIRES
WIRE WOUND ELEMENT – COILED DESIGN
CERAMIC MANDREL
OUTER COATING OF GLASS TO PROTECT WINDINGS
PLATINUM OR PLATINUM ALLOY WIRES
PLATINUM SENSING WIRE WOUND AROUND MANDREL. ENDS FUSED AT TIP AND ATTACHED TO PLATINUM WIRES AT OTHER END
WIRE WOUND ELEMENT – OUTER WOUND DESIGN

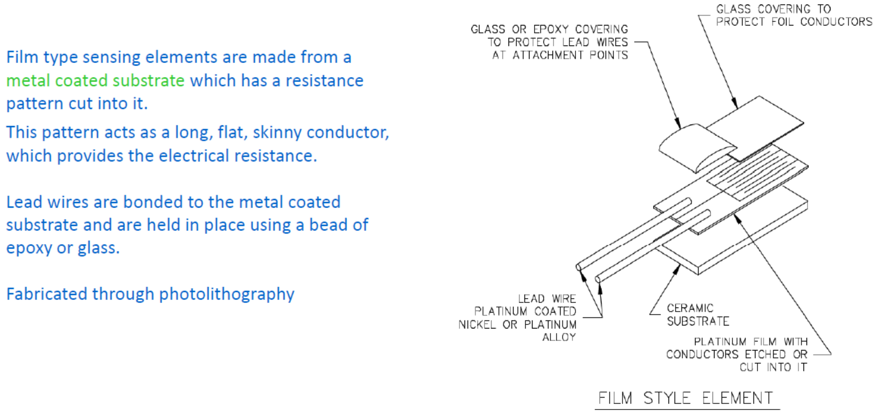
wound into a coil and packaged inside a ceramic mandrel

wound around the outside of a ceramic housing and coated with an insulating material

In wire wound we can have 2 different designs. In one case the metallic wire that represents the resistance is wounded within a ceramic mandrel. So we have the ceramic mandrel outside and inside it we have the wire wounded on itself. The other possible design is an outer design in which the metal is wounded around the ceramic mandrel. Then in order to protect the wire there is another insulator on the top of the wire wounded.

Film pattern

We have a pattern of metal which is impressed with a photolithographic process on a substrate.



Film type sensing elements are made from a metal coated substrate which has a resistance pattern cut into it.

This pattern acts as a long, flat, skinny conductor, which provides the electrical resistance.

Lead wires are bonded to the metal coated substrate and are held in place using a bead of epoxy or glass.

Fabricated through photolithography

GLASS OR EPOXY COVERING TO PROTECT LEAD WIRES AT ATTACHMENT POINTS
GLASS COVERING TO PROTECT FOIL CONDUCTORS
LEAD WIRE PLATINUM COATED NICKEL OR PLATINUM ALLOY
CERAMIC SUBSTRATE
PLATINUM FILM WITH CONDUCTORS ETCHED OR CUT INTO IT

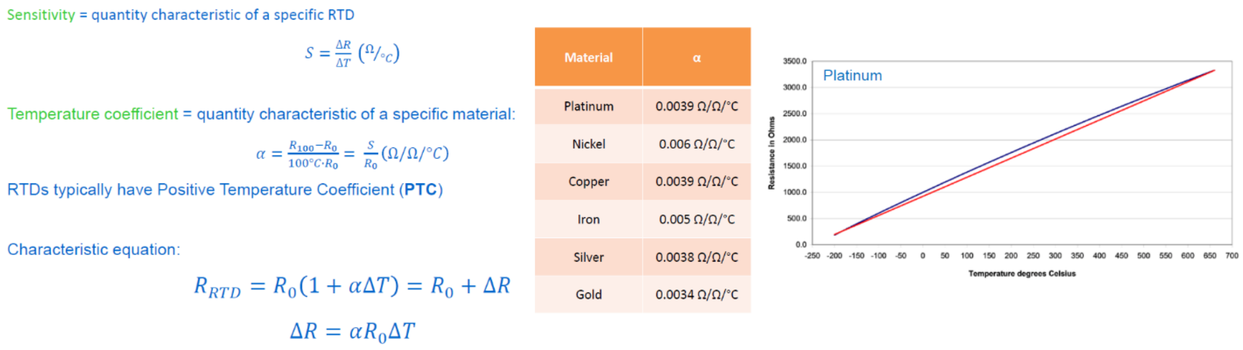
FILM STYLE ELEMENT

In this case we use a photolithographic process; it consists in depositing a metal substrate on the carrier. So we fill all the carrier with a metal substrate and then we deposit a photoresists with a specific pattern and then we illuminate the photoresist. So the photoresist is deposited with a specific pattern and then eliminated. After it has been eliminated, the metal is attached where the photoresists is no more present. So in this way we have the metal on the carrier only in the positions we want → we create a snake shaper on the carrier. Then we glue the leads to connect the resistance.

RTD parameters

- Sensitivity: defined as the variation of the output, so the resistance in our case, divided by the variation of the input, that is the temperature. Sensitivity depends on the material we use to build the resistance and on the geometrical characteristics of the resistance.
- Temperature coefficient: it is the sensitivity divided by R0, where R0 is the resistance at room temperature (e.g. for Pt100 the R0 is 100 Ohm). The temperature coefficient can be expressed also as the difference between the resistance at a certain temperature minus the nominal resistance R0 divided by the delta-T multiplied by R0. What is important is that we use the same delta R for the defined delta-T.
  So we have a ratio between resistances, and hence the dependency on the geometrical characteristics can be eliminated, and only hence alpha depends only on the type of material.

Typically, RTD are PTC with a linear characteristic, so the temperature coefficient doesn't change if we change the interval of delta-T → we can express the equation of an RTD as a linear equation.
This is very good because we don't need a look up table or exponential equation, but it is a very simple linear equation.



Sensitivity = quantity characteristic of a specific RTD

$$S = \frac{\Delta R}{\Delta T}\left(\Omega/_{°C}\right)$$

Temperature coefficient = quantity characteristic of a specific material:

$$\alpha = \frac{R_{100} - R_0}{100°C \cdot R_0} = \frac{S}{R_0}(\Omega/\Omega/°C)$$

RTDs typically have Positive Temperature Coefficient (**PTC**)

Characteristic equation:

$$R_{RTD} = R_0(1 + \alpha\Delta T) = R_0 + \Delta R$$

$$\Delta R = \alpha R_0 \Delta T$$

Material	α
Platinum	0.0039 Ω/Ω/°C
Nickel	0.006 Ω/Ω/°C
Copper	0.0039 Ω/Ω/°C
Iron	0.005 Ω/Ω/°C
Silver	0.0038 Ω/Ω/°C
Gold	0.0034 Ω/Ω/°C

The temperature coefficient for different material is typically very low, which means that the sensitivity is not very high, for huge variation of temperature we have a very small variation of resistance, but the advantage is that we have a very linear characteristic.

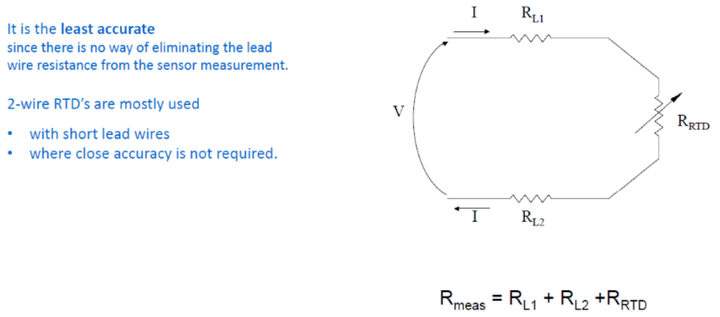
MEASURING CIRCUITS

This architecture can be used for all the resistive sensors, not only for temperature resistive sensors. We have two possible architectures. The 2 lead wires setup that is simply, we inject a current and measure the potential across the resistance. The other one, the Wheatstone bridge, measures the variation of temperature with respect to the nominal one.

2 lead wires setup

There is however a non-ideality made by the wire resistance. So we have a circuit that injects a current, but the sensor is far away from the circuit → long wire with parasitic resistances. Also the parasitic resistances can change their value with temperature.

So what happens is that we inject the current and we measure the differential voltage across the sensor with a normal amplifier. We don't measure just the R of the RTD, but also the one of the wires.



It is the **least accurate** since there is no way of eliminating the lead wire resistance from the sensor measurement.

2-wire RTD's are mostly used

- with short lead wires
- where close accuracy is not required.

$R_{meas} = R_{L1} + R_{L2} + R_{RTD}$

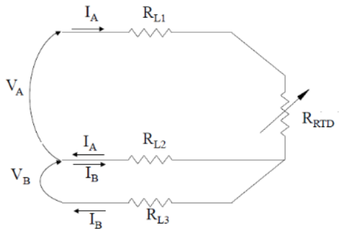
A possible solution to limit the problem is to use the 3 lead wires setup. We add a floor wire that allows us to limit the problem of stray resistances of the wire. In a first measurement we connect a current generator to the first wire an put to ground the second wire and we measure the voltage difference Va between the two wires. If we make the ratio between Va and Ia we get the resistance Ra. Then, by means of a multiplexer, we disconnect the current generator from the first wire and we connect it to the second wire and we connect to ground the third wire. In this case we measure the voltage Vb (the other one not connected is left floating, so that all the current goes in Rl2 and Rl3). Now we get the resistance Rb.

In the end we can compute the difference between the two resistances we have measured. If we compute it and we assume that the lead resistance is the same for all the leads, the Rl2 simplifies, and also Rl1 with Rl2. So we get a resistance that depends only on Rrtd.

This assumption is good in almost all the setup, because the stray resistances depend on the material and geometry, so if the wire is the same, we can assume to have the same stray resistances on the cables.

It is the most used in **industrial applications**
where the third wire provides a method for removing the average lead wire resistance from the sensor measurement.

When long distances exist between the sensor and measurement/control instrument, significant savings can be made in using a three-wire cable instead of a four-wire cable.



$R_A = R_{L1} + R_{L2} + R_{RTD}$
$R_B = R_{L2} + R_{L3}$
$R_{meas} = R_A - R_B = R_{RTD}$

Assuming $R_{L1}=R_{L2}=R_{L3}$

I4 lead wires setup

We don’t need to do any assumption. We inject the current between lead 1 and lead 4. So the current will go in Rl1 and Rl4. We won’t have any current in Rl2 and Rl3 if we perform a readout with high impedance, so if we read the voltage with an INA, that is a differential opamp.
An important property of the INA is that the two inputs are high impedance inputs, which means we have zero current entering or going out from the amplifier, because we have the gate of a transistor in the inputs.
If we have no current, we have also no current in the lead resistances Rl2 and Rl3. Hence the Vdiff on the sensor is the same differential voltage measured by the INA → we measure exactly the RTD value.
But **it is important to readout the voltage with high impedance**, otherwise I impair the measurement.

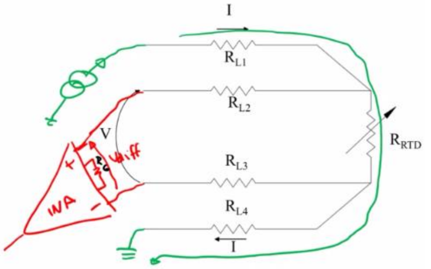
This is a setup used in laboratory application where we need high accuracy, while in industrial application we use the 3 leads approach because it uses one wire less.

It is used primarily in **laboratory** where close accuracy is required.

In a 4 wire RTD the actual resistance of the lead wires can be removed from the sensor measurement.

This method compensate for any differences in lead wire resistances.

The input impedance of the voltage measurement circuitry must be high enough to prevent any significant current flow in the voltage leads.



$V_{OUT} = V_{diff} \cdot G_{INA}$ $R_{meas} = R_{RTD}$
$G_{INA} = 1 + 2\frac{R_F}{R_G}$

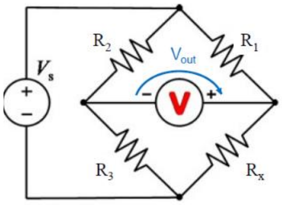
If we imagine having an RTD with a high value of R0 (nominal resistance) and we want to measure small variation of resistance (1 Ohm) around the nominal value. In this circuit we have a big voltage associated to the nominal value, and a small one to the differential value.

With the WB we can amplify the small difference, we amplify the voltage due to the variation of resistance.

Wheatstone bridge

We have a voltage divider between resistance R2 and R3 on one branch, and on the other branch it is between R1 and Rx, where Rx is our RTD. In this case we need to provide an external voltage power

supply. We can obtain the two voltages at the two intermediate nodes and then we connect the amplifier between V+ and V-.
Our aim is to have Vout = 0V when we are at the nominal temperature → we consider that when Rx = R0, we need that R2/R3 = R1/Rx0. So R1 = R2 and R3 = Rx0. In this case we have a perfectly balanced bridge with 0 value at its output.



R_{x0} nominal value of R_x → $R_x = R_{xo} + \Delta R$

In order to have V_{out}=0 when $R_x = R_{x0}$:

$$\frac{R_2}{R_3} = \frac{R_1}{R_{x0}}$$

Considering $R_1 = R_2$ and $R_3 = R_{x0}$:

$$V_{out} = V_s \cdot \left[\frac{R_{xo} + \Delta R}{R_1 + R_{x0} + \Delta R} - \frac{R_{xo}}{R_1 + R_{x0}}\right] = V_s \cdot \frac{R_1}{(R_1 + R_{x0})^2} \cdot \frac{\Delta R}{1 + \frac{\Delta R}{R_1 + R_{x0}}} \quad x$$

$$V_{out} \approx V_s \cdot \frac{R_1}{(R_1 + R_{x0})^2} \cdot \Delta R = V_s \cdot \frac{R_1}{(R_1 + R_{x0})^2} \cdot \alpha R_{x0} \Delta T$$

Choosing $R_1 = R_2 = R_3 = R_{x0}$ provides the maximum $\frac{V_{out}}{\Delta T}$ → $\frac{V_{out}}{V_s} = \frac{1}{4} \cdot \frac{\Delta R}{R_{xo}}$

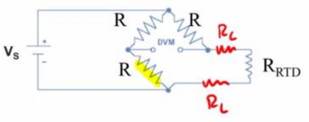
Now we move from the nominal temperature, so we have a variation of T. the resistance won't be anymore Rx, but Rx0 + deltaR. Considering now this value, Vout can be easily computed by means of a voltage divider and V+ - V- (R1 = R2 and Rx0 = R3).
If we do some computations, we obtain the expression x. We can simplify it considering that deltaR/(R1 + Rx0) can be neglected with respect to 1 since deltaR is very small.

In the end we can express deltaR with the temperature coefficient so that we can easily find the relation between Vout and deltaT.

If we want to maximize the sensitivity, so the Vout/deltaT, we can try to derivate the expression with respect to deltaT; the maximum of the expression is found when the 3 resistances are equal to the nominal value of the RTD. If so, Vout/Vs will be the last formula in the bottom right, and this is the maximum sensitivity we can obtain.

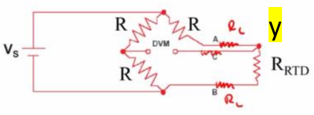
2 and 3 wires bridge

2-wire



x Ideally (R_L = 0): $V_{out} = V_S\left(\frac{R+\Delta R}{2R+\Delta R} - \frac{1}{2}\right)$

3-wire



Considering R_{L1}=R_{L2}=R_L

In 2-wire bridge:

$$V_{out-2w} = V_S\left(\frac{R+\Delta R+2R_L}{2R+\Delta R+2R_L} - \frac{1}{2}\right)$$

$$\varepsilon = V_{out-2w} - V_{out} =$$

$$V_S \frac{2R_L \cdot R}{(2R+\Delta R)^2\left(1+\frac{2R_L}{2R+\Delta R}\right)} \approx V_S \frac{R_L}{2(R+R_L)}$$

In 3-wire bridge:

$$V_{out-3w} = V_S\left(\frac{R+\Delta R+R_L}{2R+\Delta R+2R_L} - \frac{1}{2}\right)$$

$$\varepsilon = V_{out-3w} - V_{out} =$$

$$V_S \frac{-R_L \cdot \Delta R}{(2R+\Delta R)^2\left(1+\frac{2R_L}{2R+\Delta R}\right)} \approx -V_S \frac{R_L}{2(R+R_L)} \cdot \frac{\Delta R}{2R}$$

The object can be far away from the power supply, so also here we can have stray resistances. If we use the previous approach for WB, we have an error introduced by Rl, so we measure also the variation of Rl. For this reason we can introduce an approach with 3 wires. Also in this case we have lead resistances, but the differential voltage is readout with another wire and high impedance.

If Rl = 0 as in the ideal case, we get the formula x (considering zero stray resistances).
Let's now consider the stray resistances to compute the error of the differential voltage computed in the 2 and 3 wires case with respect to the ideal case.
In the case of the **2 wires approach**, the output will include also the Rl. Now we can define the error epsilon as the difference between the voltage we actually measure minus the ideal voltage. If we compute this difference, we obtain the bottom green expression. In this expression we can neglect the deltaR when it is summed to the R, because we assume that the variations of deltaR are much smaller than R, so we can omit the terms deltaR in the computations of the error.

In the case of the **3 wires approach**, we do the same computations. What changes is the computation of the V+, because in the V+ we have to consider that we are reading out the node y, then we have a very high input impedance. So I have a voltage divider in which I have to consider the voltage drop or Rl + RTD and the Rl is the one on the bottom with respect to the RTD (this is the numerator). Then at the denominator I have to consider all the resistances crossed by a current, that are R, 2Rl and RTD (that is R + deltaR).
Again, we can compute the error epsilon. If we compute this error, we obtain the last red equation (always neglecting the deltaR terms).
Looking more in detail, we have a part that is equal to the error we have in the 2 wire approach, but then in the case of the 3 wires we multiply by a factor deltaR/2R, that is a term much smaller than 1 → the error in the 3 wires approach is smaller than the 2 wires approach.
We still have an error, so we cannot get a zero error, but it is significantly reduces.

Does it exist an approach with 4 wires also for WB?
No, not practically. The 4 wire approach was used for a differential readout. Also the WB exploits a differential readout, so making a differential readout over a differential readout is not practical.

SELF HEATING

If I inject current in a resistance, I will have some power in the resistance, so the resistance will warm up just because of Joule effect. This phenomenon is called self-heating.
We cannot completely avoid self-heating, but we can keep it at reasonable values, and in order to do so we limit the current in the resistance and we provide packages for the sensor that are able to dissipate the power. Typically, wire wounds are better if we want to dissipate the heat.
As for the current, typically is limited to 1mA to prevent self-heating.

However, in some applications, self-heating is something that is exploited, because for instance we want to have a resistance that increases its value if current increases (a sort of negative feedback).

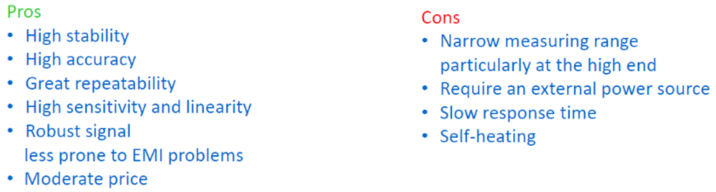
Since RTDs are resistors, they will produce heat when a current is passed through them.
The normal current limit for industrial RTDs is **1 mA**.

- Thin film RTDs are more susceptible to self-heating so 1 mA should not be exceeded.
- Wire wound RTDs can dissipate more heat so they can withstand more than 1 mA.

The larger the sheath or the more insulation there is the higher the error caused by self-heating. In some applications self-heating is exploited as an advantage.

PROS AND CONS

The main advantage is the very high linearity, and moreover the price is moderate. The main disadvantages are related to the small ranges of temperatures we can measure, they always require an external power supply and then there is self-heating.



Pros
- High stability
- High accuracy
- Great repeatability
- High sensitivity and linearity
- Robust signal
  less prone to EMI problems
- Moderate price

Cons
- Narrow measuring range
  particularly at the high end
- Require an external power source
- Slow response time
- Self-heating

THERMISTORS

Resistances that change their resistance depending on temperature. There are two big families:

- PTC: will increase the resistance with increasing temperature
- NTC: will decrease resistance increasing the temperature

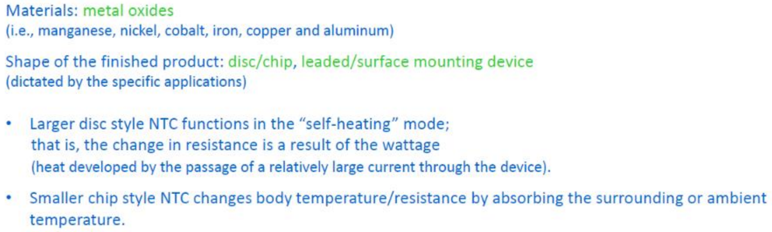
The main differences with respect to RTD are two:

1. The material used for fabrication: RTD are made out of metallic materials, while thermistor, in particular NTC, are made out of ceramic semiconductors or metal oxides.
2. The main property of RTD is to be very linear, whereas both NTC and PTC have a strongly nonlinear behaviour. For this reason, **RTD are used to measure the absolute temperature of an object, while NTC as switching sensors, to check if the temperature exceeds a given value**.

An expectation is that, recently, it has been developed a linear PTC, so a thermistor with a linear characteristic that can be used in the same applications of RTD.

NTC THERMISTOR

They are made with metal oxides and provided with different shapes; we have the disc, bead/chip or surface mounting device. The first two are 'through hole' so they must pass through the board, while the latter is directly soldered on the pcb.



Materials: metal oxides
(i.e., manganese, nickel, cobalt, iron, copper and aluminum)

Shape of the finished product: disc/chip, leaded/surface mounting device
(dictated by the specific applications)

- Larger disc style NTC functions in the "self-heating" mode;
  that is, the change in resistance is a result of the wattage
  (heat developed by the passage of a relatively large current through the device).
- Smaller chip style NTC changes body temperature/resistance by absorbing the surrounding or ambient temperature.

disc

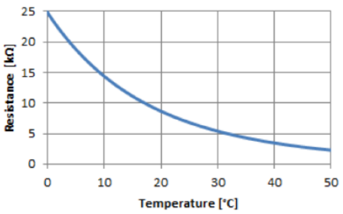
chip

smd

Mainly, the chip and smd products are used when we want to measure the temperature of an object, because they have a very small package, so the thermal capacitance is low and they can rapidly measure the change in temperature. Conversely, disc NTC have a high thermal capacitance, so they are more used to exploit self-heating phenomena, because they hardly change their temperature due to external changes.

NTC CHARACTERISTIC

It is strongly nonlinear. We can see that we have an exponential behaviour, so the variation of resistance depends on the power of the temperature.



$$R(T) = R_{25} \cdot e^{\beta\left(\frac{1}{T}-\frac{1}{T_{25}}\right)}$$

R_{25} resistance at 25°

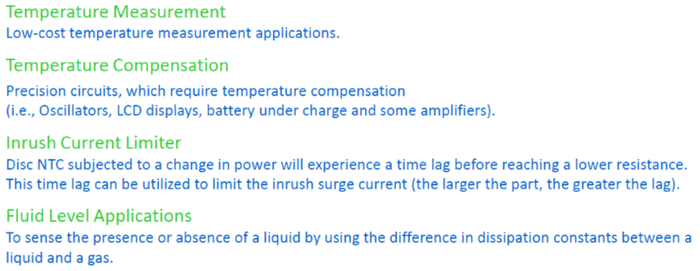
β thermistor constant (depends on material)

T temperature expressed in Kelvin

T_{25} = 298.15 K

Applications

They are used for inrush current limiter, that is something able to limit the current at the power on of my device. In fact, when we switch on the circuit, we expect a low temperature, so we have a high value of resistance that can limit the current. Then, when the circuit warms up, the value of the resistance decreases and the current is increased → way to perform a soft start of the circuit.



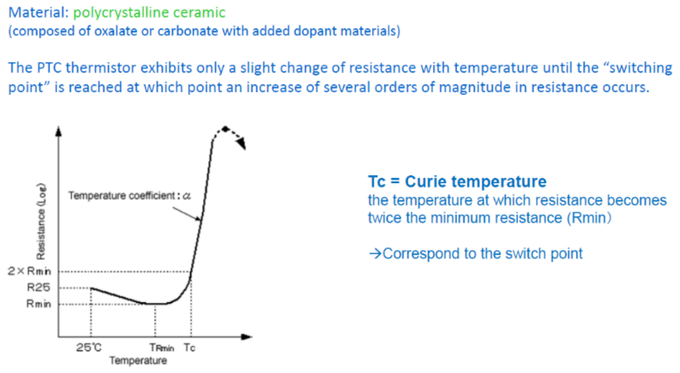
Temperature Measurement
Low-cost temperature measurement applications.

Temperature Compensation
Precision circuits, which require temperature compensation
(i.e., Oscillators, LCD displays, battery under charge and some amplifiers).

Inrush Current Limiter
Disc NTC subjected to a change in power will experience a time lag before reaching a lower resistance.
This time lag can be utilized to limit the inrush surge current (the larger the part, the greater the lag).

Fluid Level Applications
To sense the presence or absence of a liquid by using the difference in dissipation constants between a liquid and a gas.

PTC THERMISTOR



Material: polycrystalline ceramic
(composed of oxalate or carbonate with added dopant materials)

The PTC thermistor exhibits only a slight change of resistance with temperature until the “switching point” is reached at which point an increase of several orders of magnitude in resistance occurs.

Tc = Curie temperature
the temperature at which resistance becomes twice the minimum resistance (Rmin)

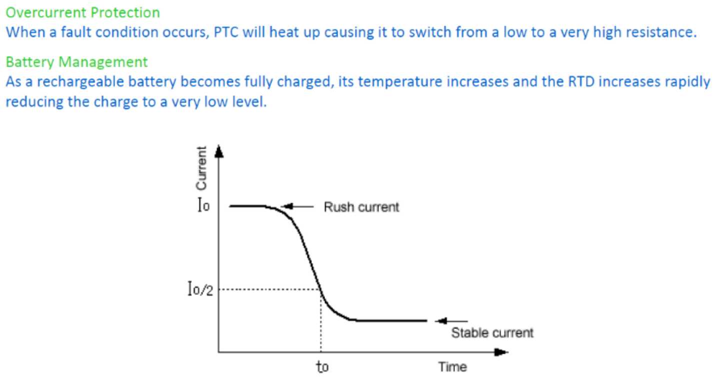
→Correspond to the switch point

Made out of polycrystalline ceramic material with a high nonlinear and non-monochromatic behaviour. At low T, the behaviour is similar to the one of an NTC, then the resistance starts to increase. We define Tc the Curie temperature, the temperature at which the PTC starts to increase its resistance very rapidly (the characteristic is almost vertical, so we cannot measure the temperature with this kind of sensors).

For this reason, PTC are used as switch detector, to check if the temperature exceeds the Tc.

Applications

We don’t measure temperature, but we have an overcurrent protection, that is similar to the role covered by a fuse (a fuse assures that if the current exceeds a given high value it opens up and high current doesn’t reach the device. But the fuse breaks, so it must be removed and changed). In the case of PTC instead of a fuse, we can have a sort of fuse that doesn’t break but that increases the resistance if the current increases the



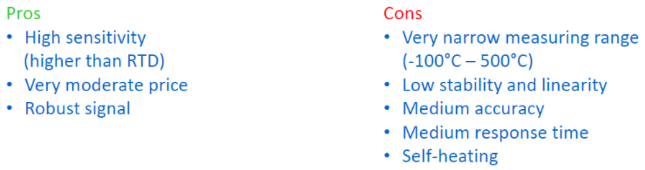
Overcurrent Protection
When a fault condition occurs, PTC will heat up causing it to switch from a low to a very high resistance.

Battery Management
As a rechargeable battery becomes fully charged, its temperature increases and the RTD increases rapidly reducing the charge to a very low level.

temperature of the sensor, and the resistance is so high that becomes a sort of open circuit. It can be regarded as a reversable fuse.

They were used also for battery management. At the beginning of the recharging, the battery has low T, so we inject a high current. As the recharge proceeds, T increases and so we want to reduce the current provided. However, they are no more used for battery management nowadays.

THERMISTORS PROS AND CONS



Pros

- High sensitivity (higher than RTD)
- Very moderate price
- Robust signal

Cons

- Very narrow measuring range (-100°C – 500°C)
- Low stability and linearity
- Medium accuracy
- Medium response time
- Self-heating

They have a very very moderate price. The main disadvantage is the limited range and they are not linear.

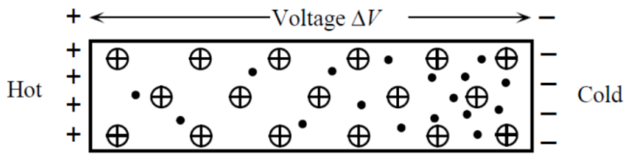
THERMOCOUPLE

It is a temperature sensor based on the Seebeck effect.

SEEBECK EFFECT

If we have a gradient of T across a material, we will experience also a voltage difference across the material. let’s suppose to have a hot side and a cold side. At the hot side, the thermal motion of electrons (we are considering metals) increases a lot, so e- diffuses more towards the cold side, because on the hot side the energy is higher and particles tends to go towards lower energy side. So we have a flow of electrons from one side to the other. But this motion causes a voltage difference across the material, we will have a positive voltage on the hot side with respect to the cold one.

A temperature difference between two points in a conductor or semiconductor results in a voltage difference between these two points.



The electrons in the hot region are more energetic (higher energy levels)

→ net diffusion of electrons from hot to cold

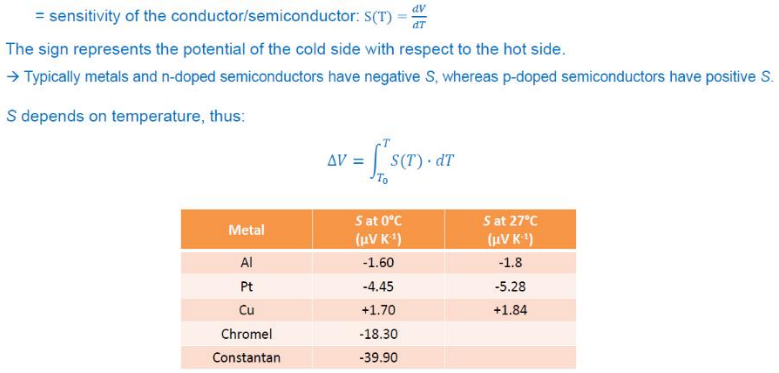
→ voltage

When the diffused current, so the current due to electrons that move towards minimal energy regions, and the current due to the effect of the deltaV across the material are equal, so the diffuse current is equal to the one due to the electric fiels, we reach a steady state, where we have no more current.

Seebeck coefficient – Sensitivity

The sensitivity of such detector is obtained as dV/dT, but it strongly depends over temperature (so I cannot write deltaV/deltaT), because S is not constant and it changes with temperature. Hence the deltaV must be computed with an integral.
For metals the sensitivity is typically negative, because the delta is between the cold and hot side, and on the cold side we have a negative voltage. An exception is represented by Copper.

So we cannot define a general sensitivity, but the sensitivity at a specific temperature.



= sensitivity of the conductor/semiconductor: $S(T) = \frac{dV}{dT}$

The sign represents the potential of the cold side with respect to the hot side.

→ Typically metals and n-doped semiconductors have negative *S*, whereas p-doped semiconductors have positive *S*.

S depends on temperature, thus:

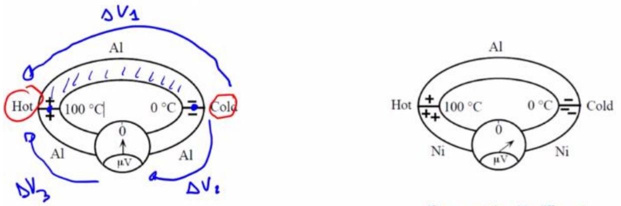
$$\Delta V = \int_{T_0}^{T} S(T) \cdot dT$$

Metal	*S* at 0°C (μV K⁻¹)	*S* at 27°C (μV K⁻¹)
Al	-1.60	-1.8
Pt	-4.45	-5.28
Cu	+1.70	+1.84
Chromel	-18.30	
Constantan	-39.90	

THERMOCOUPLE PRINCIPLE

We want to exploit the Seebeck effect to measure the temperature difference between two points. Let’s suppose to use an Al wire to measure the differential temperature. However, using only one material is not possible.

Only the net Seebeck voltage difference between different metals can be measured.



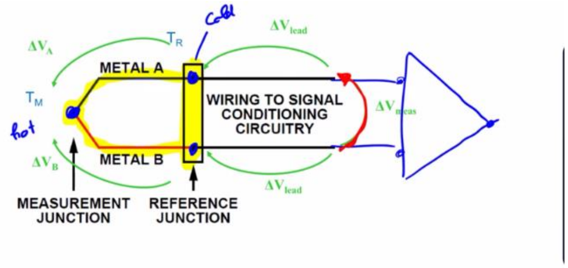
If one metal is used the net voltage is 0V.

If two metals with different Seebeck coefficient are used there is a net voltage.

So we have a piece of Al and a voltage difference (deltaV1) across it due to the difference in temperature. Then we connect the two hot and cold points using other Al wires. But since again we have Al between different points with different temperature, we have deltaV3 and deltaV2. But since we have the same temperature at the two extreme points, deltaV1 = deltaV2 + deltaV3.

So now the voltage that is measured by the voltmeter will be equal to 0, because Vout = deltaV3 – deltaV1 + deltaV2. So if we use only one metal we cannot measure any voltage difference.

Let’s now use different materials, so an Al wire between hot and cold side that experiences a certain deltaV1. Then we use Nikel, that has a different Seebeck coefficient, so the voltage we readout is no more equal to zero → we always need to use two different materials when we build a thermocouple.



$$\Delta V_{meas} = \Delta V_{lead} + \Delta V_B - \Delta V_A - \Delta V_{lead} = \Delta V_B - \Delta V_A \quad \rightarrow \quad \Delta V_{meas} = \int_{T_R}^{T_M} (S_B(T) - S_A(T)) \cdot dT = \int_{T_R}^{T_M} S_{AB}(T) \cdot dT$$

$$S_{AB}(T) = S_B(T) - S_A(T)$$

So we have a tip where we measure the temperature, and in the tip we have the junction between two different material, and this part is called reference junction.
So with a thermocouple we cannot measure an absolute temperature, but only delta temperatures.
Then we connect the thermocouple to the readout circuit (NB: the sensor are the cables in which we have the two different metals), so we will have wires (e.g. made out of Copper) that connect the terminals of the thermocouple with the readout circuit. But we have to be careful because we can experience a deltaV also on the Copper lines.
The red one is the deltaV we measure and it is given by different contributions: the voltage difference due to the leads because of different temperature between the reference junction and the leads, then we have the deltaV due to the metal A and then metal B and then the deltaV due to the second lead. Fortunately, the two deltaV due to the leads are the same, so we can simplify them.

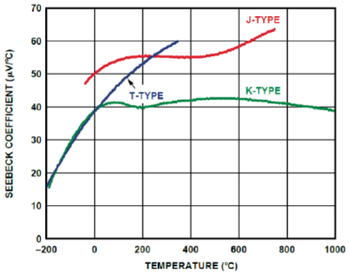
In the end we have an output voltage that is proportional only to the difference in deltaV due to the thermocouples.
The output, computed as an integral, will depend on the difference between the sensitivity of material B and material A. For this reason, in a thermocouple the parameter that is defined is the Sab, that is already the difference between the sensitivity of the two metals.

In order to have a good thermocouple we need hence to have two materials with different Seebeck coefficients (i.e. sensitivities) so that Sab is maximized.

All the thermocouples are characterized by a very wide temperature range, so they are very robust at high temperature, with respect for instance to RTD. This is their main advantage.

Thermocouple characteristic

In the plot we have Sab, that can be constant or changing with temperature.



→ The voltage signal is not linear; a calibration is required.

REFERENCE JUNCTION

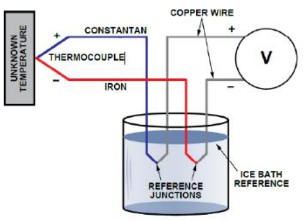
What is the reference temperature in a thermocouple?

It is a well-known temperature, and one way to have a reference temperature is to consider a reference junction that is a bath of ice and water, so that we are sure that the temperature is 0. But this solution is not practical at all, so we simply use another temperature sensor that is able to measure the temperature of the reference junction. So inside the instrument we have a temperature sensor (e.g. RTD) that measures the reference temperature. If we know the temperature of the reference junction and the deltaT measured by the thermocouple we can get the final temperature of the object.

$T_{obj} = T_{ref} + deltaT_{therm}$

Why not to use just the temperature sensor?

We need a thermocouple because it is difficult to find detectors that work at very high temperatures (e.g. 2000°C). In fact, typically used metals for RTD melts at these temperatures.



NOT PRACTICAL !!

Reference-junction compensation

= the reference junction temperature is measured with another temperature sensitive device:

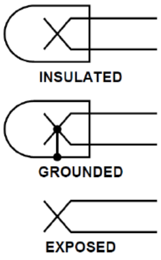
- RDT
- thermistor
- thermal diode
- integrated temperature sensor

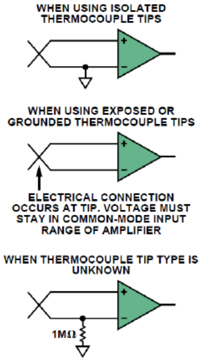
Another advantage of thermocouple is that they are small. It is made by just two wires of different materials. The tip is the central point that is put in contact with the object of which we want to measure the temperature. The tip can be:

- Insulated: we have a cap that aelectrically insulates the thermocouple from the object. This insulation introduces a slightly higher thermal resistance
- Grounded: we have a cap but the cap is electrically connected to the tip. Hence it is very similar to have an exposed thermocouple with the tip directly in contact with the object
- Exposed

In the case of grounded and exposed thermocouple we must be careful in designing the readout circuits. The output is a differential amplifier typically.

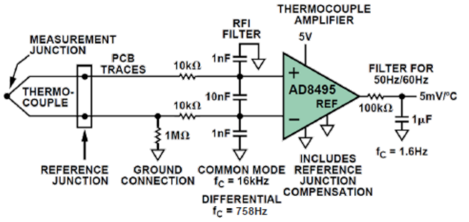
- Insulated: we have a differential voltage but the common mode is not defined by anything, and this is the reason why I have to put one of the two wires at ground, and the CM is the average between ground and the voltage.
- Grounded or exposed: the CM is given by the voltage of the object we are measuring. The object voltage must be between the input range of the amplifier. If the object is outside the range, the amplifier won't be able to read it out correctly, because the CM is outside the voltage range of the amplifier.
- If we want to make a general purpose readout system, not knowing the nature of the tip, we an use the third solution, where one wire is grounded through a 1MOhm resistance. Through this ground we provide a reference voltage if the tip is insulated, while if it is grounded or exposed we don't provide ground, but we use the resistance to limit the current that passes from the object to the thermocouple when using a non-insulated solution.




WHEN USING ISOLATED THERMOCOUPLE TIPS
+
−
WHEN USING EXPOSED OR GROUNDED THERMOCOUPLE TIPS
+
−
ELECTRICAL CONNECTION OCCURS AT TIP. VOLTAGE MUST STAY IN COMMON-MODE INPUT RANGE OF AMPLIFIER
WHEN THERMOCOUPLE TIP TYPE IS UNKNOWN
+
−
1MΩ

PRACTICAL READOUT CIRCUIT

The thermocouple is connected to an INA differential amplifier with an internal gain of 122. In particular, we also added some capacitors and resistors to make some filtering. We have the 1MOhm resistor to ground to use both the possible solutions. Then we have some capacitors and resistances before the input in order to do some filtering before the input. We need this filtering because if we have the two cables of the thermocouple, they can be very long. This may create a loop that can couple electromagnetic interference. This electromagnetic interference can create RF disturbances, so we add the filters for RF interferences. We want to introduce some poles both for the CM and for the differential mode. The RF must be filtered before the INA because it can change the CM and move the input outside the range of the INA.



INA

Gain = 122

Ref Temperature compensation

Non-linearity correction for k-thermocouples

Noise/disturbances reduction:

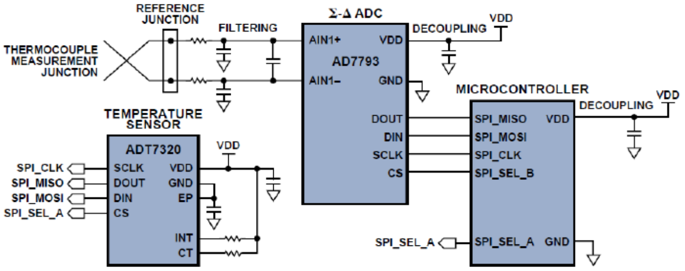
- Differential input (eliminates common mode disturbances)
- Radio Frequency Interference filter (low pass filter)
- Filter for 50/60Hz

Then, after the amplification we have another filter, a LP filter that is the one for the 50Hz or 60Hz of the power lines.

DIGITAL READOUT

We still have the two wires, a filtering before the input with the same purposes as before, and then we bring the two inputs to an ADC. In particular, in the example we have a sigma-delta ADC, that is a very fast ADC with a single wire output. Internally to the ADC is integrated also an INA. The value converted by the ADC can then be sent to a uC using an SPI communication protocol.

Then we have a third component that is another temperature sensor to measure the temperature of the reference junction. Also this sensor communicates with the uC with an SPI bus.



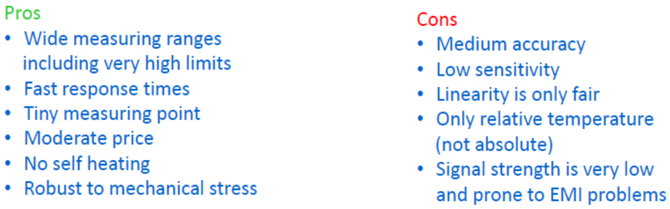
Flexible (J-, K-, T-type thermocouples)

Optimized for accuracy

INA included in the ADC

PROS AND CONS

The most significant advantage is the wide measuring range and the fast response due to the small wire. Also the non-self-heating is important, because the resistance of the wires is almost null → we are measuring the actual temperature of the object, while it was not the case for RTD and thermistors



Pros

- Wide measuring ranges including very high limits
- Fast response times
- Tiny measuring point
- Moderate price
- No self heating
- Robust to mechanical stress

Cons

- Medium accuracy
- Low sensitivity
- Linearity is only fair
- Only relative temperature (not absolute)
- Signal strength is very low and prone to EMI problems

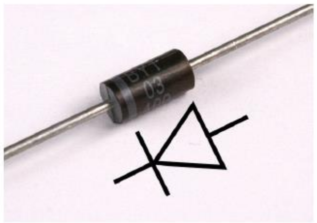
One big disadvantage is that the signal strength is low (mV), and they are prone to electromagnetic interferences problem.

DIODE AND BANDGAP TEMPERATURE SENSORS

They main advantage is that they can be integrated in CMOS technology. If the y can be integrated in CMOS are also called ‘solid state’.

THERMAL DIODES

The current of the diode can be expressed by the formula x. Is is the reverse current and it is a value multiplied by an exponential where we have V, voltage across the thermal diode and T, the temperature. So the current on the diode exponentially depends on T once we have given a voltage.



x $I = I_s \cdot (e^{\frac{qV}{mkT}} - 1)$

I_s reverse saturation current

m technology parameter

$$V = m\frac{kT}{q} \cdot ln\left(\frac{I}{I_S} + 1\right)$$

$$\downarrow$$

$$\frac{\Delta V}{\Delta T} = m\frac{k}{q} \cdot ln\left(\frac{I}{I_S} + 1\right)$$

Not linear since I_s depends on temperature

In this case it is convenient to have a fixed current and measure voltage because the relationship between voltage and current it linear, while between current and T is exponential. We can also compute the sensitivity, which seems to be constant with temperature, since k, q, m are constant and the ln(I/Is + 1) has I that is constant, while Is depends on temperature. Hence is it not a constant sensitivity. So the characteristic is not perfectly linear.

So we inject a current in the thermal diode and measure the voltage.

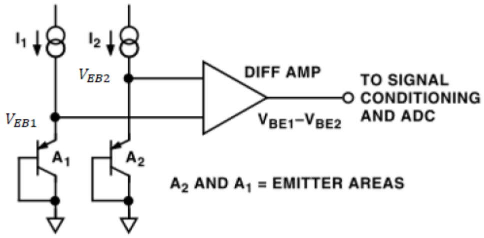
BANDGAP TEMPERATURE SENSORS

In this sensor we use a differential approach. We have two BJT transistors which are made with two junctions. In the case of the image they are pnp transistor, the base is doped n and emitter and collector p. In this case we are shorting the base with the collector, so we are shorting one junction together → this BJT configuration is very similar to a normal diode. It is called transdiode.

The system principle is the same of the diode, but now we use a differential readout approach. So the emitter-base voltage can be computed as if the BJTs were pure diodes, and I neglect the +1 in the parenthesis assuming I >> Is. In the equations, saturation current Is is expressed in terms of current density, so Js*A where A is the area.
We can now compute the emitter base voltage for the two transdiodes. The areas and the currents of the BJTs can be different, while the Js is the same if we use the same technology.

If we use a differential approach, we have then a differential amplifier and we can subtract Veb1 – Veb2.
Now, let's imagine to inject the same current, and to have the same area → ln = 0 and we won't measure any voltage different. Hence we have to consider either different currents or different areas. In the image the currents are the same, while the areas are different and their ratio is called r.

In this case the output voltage is perfectly linear with output voltage, because we have simplified the dependency with the saturation current.

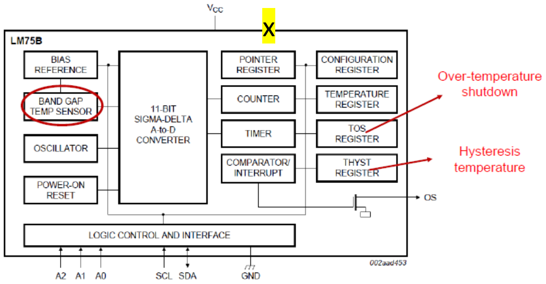


$$V_{EB1} \approx \frac{kT}{q} \cdot ln\left(\frac{I_1}{A_1 J_s}\right) \quad V_{EB2} \approx \frac{kT}{q} \cdot ln\left(\frac{I_2}{A_2 J_s}\right) \quad \rightarrow \quad V_{EB1} - V_{EB2} = \frac{kT}{q} \cdot ln\left(\frac{I_1 A_2}{I_2 A_1}\right)$$

If $I_1 = I_2$ and $\frac{A_2}{A_1} = r$ → $V_{EB1} - V_{EB2} = \frac{kT}{q} \cdot \ln(r)$

$T_{measured} = const \cdot (V_{EB1} - V_{EB2})$ with $const = \frac{q}{k \cdot \ln(r)}$

Digital temperature sensor



I²C interface

Block diagram from a datasheet. We have a smart sensor, so we don't have only the temperature sensor, but also other electronics to provide us an output that is a digital output. Serial clock and serial data output (SDA) are used to communicate with a I2C protocol. In the middle between the temperature sensor and the I2C interface we need some electronics; an ADC (sigma-delta also in this case) in order to convert the output of the differential amplifier in a digital value, a customized uC. We have a part x where we have some registers similar to the ones we have in a uC.

This smart sensor is designed to make a temperature control above a threshold temperature. If the threshold is exceeded, the OS pin goes to 0 → we don't need to read the I2C interface, but just to see if the threshold has been exceeded.

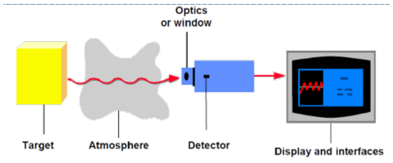
INFRARED THERMOMETER

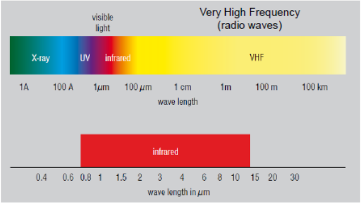
Infrared thermometers are a different class of temperature sensors. We have always considered them as in contact with the object to measured. These infrared thermometers can measure the temperature at a certain distance, because they are not measuring the actual temperature but the infrared emission. They can also be made out of arrays of sensors so that we have a temperature image.

INTRODUCTION TO IR SYSTEMS

We can divide the emission spectra in different parts. The visible range is a very small range between IR and UV. The visible light is measured with optical sensors like CCD, while when we want to measure the temperature we are interested in measuring the IR range, because all the bodies emit IR light and the amount of light emitted depends on the temperature of the body. In particular, the IR range can be divided in 3 main bands:

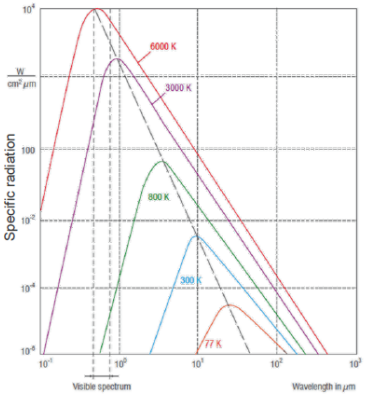
- Near wave infrared up to 2um
- Medium wave infrared
- High wave infrared





BLACK BODY EMISSION

The black body is an ideal body in which the absorption is always equal to the emission. So we have no reflection and no transmission. In general, if a body is at temperature stability, emission must be equal to absorption. In a black body they are both equal to 1 → total incoming power can either pass through (transmission), reflected (reflection) or absorbed. The sum of these three phenomena must be equal to the incoming power, but for a black body we don't have transmission and reflection. If the black body is at a certain temperature, we reemit the same light absorbed.



Black-body:

A = ε = 1 R =T = 0

A absorption
ε emissivity
T transmissivity
R reflection

Stefan-Boltzmann law:

$$\frac{P}{A} = \varepsilon \cdot \sigma \cdot T^4$$

$\sigma = 5.67 \cdot 10^{-8}$ W/(m²K⁴)
Stefan-Boltzmann constant

Wien displacement law:

$\lambda_{max} \cdot T = 2898$ µm · K

The emission spectra for a black body at a certain T is provided by the Stefan-Boltzmann law. The specific power (power divided by area) is proportional to the 4th power of the temperature.
Not only the total power changes with temperature, and increasing T we increase the power, but also the peak of the wavelength changes with T; at higher T we have a lower wavelength. This is very intuitive; a very hot object is so hot that emits something that is also visible, because in the red range of emission.

The Wien law expresses the wavelength at which we have maximal emission.

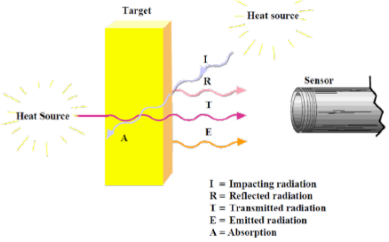
GRAY BODY

In this case the absorption is always equal to the emission at the steady state, but smaller than 1 because we have some reflection and transmission. In solid grey body the transmission is equal to 0.

If we measure the emission, is not straight forward the relation between T and emitted radiation, because the emission coefficient is smaller than 1, so to measure the T we need also to know the epsilon of the material.

Non-metallic materials have high emissivity epsilon, while metallic objects have low emissivity and so strong reflectivity, so the emission is quite slow.

Emissivity can change a lot depending on the object of which we are measuring the temperature, so we need a calibration to measure temperature.



Generic grey-body:
$A = \varepsilon < 1$
$R + T + \varepsilon = 1$

Solid grey-body:
$A = \varepsilon < 1$
$R + \varepsilon = 1 \quad T = 0$

Non-metallic materials (wood, plastic, rubber, organic materials, rock)
→ low reflectivity, high emissivity (0.8 - 0.95)
Metals (especially those with polished or shiny surfaces)
→ low emissivity (around 0.1)

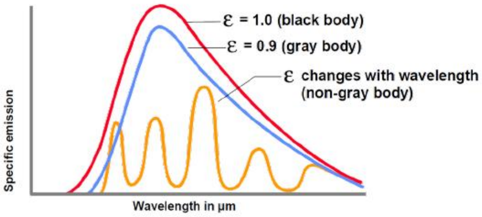
→ Need of a calibration to estimate ε

NON-GRAY BODY

They are objects in which the shape of the emission is not the same shape of the black body. In a gray body the shape is the same of the black body but with a scaling factor because epsilon < 1. In non-gray body it is different because I can have different molecules that absorb and emit radiation in different ways depending on the wavelengths.

If we consider for instance the integral of the emitted power, it is not a good parameter of merit, because the overall area depends on the shape and different object have different shapes.

So it is better to have the IR sensor that considers a small ranges of wavelengths. For a specific wavelength we can then do a calibration for each specific object.



Non gray-body: non-oxidized metals, glass, plastic films

→ Better to consider one single wavelength

The indicated epsilon values hold for all the wavelengths. So black body emits what absorb for all the wavelengths, the gray body has a reduction of 10% at each lambda with respect to black body, while for a non-gray body we have a reduction with respect to black body emission in terms of epsilon that depends on the wavelength, the distance between curves changes with lambda.

Determining emissivity

It can be determined by having a a-priori knowledge of the material we are measuring the temperature. If we don't have this, we can perform a sort of calibration. The first thing we can do is to use an IR sensor and a contact temperature sensor, e.g. a thermocouple. It is done just one time at the beginning, then the thermocouple can be discarded and use only the UR sensor.

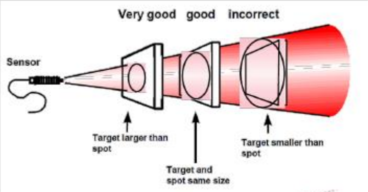
Another way to do the calibration is to attach to the object we measure the temperature a plastic sticker of a known epsilon. If we don't want to put a sticker, we can only use it to perform a calibration. We put the sticker on a portion of the object, we measure the temperature of the sticker and we know the emissivity of the sticker, then we measure the temperature in the same portion of the object without the sticker and now we can compute the emissivity with calibration and remove the sticker.

Measurement spot

In an IR temperature sensor we have to be careful, because they have a field of view. The temperature we measure is the average of a certain area, and the dimension of the area increases with distance → field of view increases with distance. If the FOV is bigger than the object itself it is a problem, because the average of the temperature depends also on the ambient.

IR temperature sensors can erroneously called LASER temperature sensor because often they integrate a laser pointer to point at the direction where I'm measuring the temperature, but I'm not using the laser to measure the temperature, it is just a pointer to see which object we are measuring.

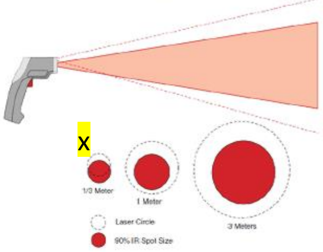
The field of view of the sensor is slightly different than the laser pointer field of view due to the parallax error. If we are far away from the sensor, the FOV is smaller than the point of the laser pointer, while if we are close to the sensor we can have a difference given by the parallax error x.



Single spot sensor

or

cameras



Laser pointer

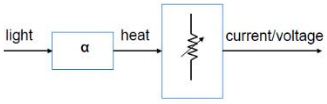
for exact spot size making

(even very precise
coaxial pointers)

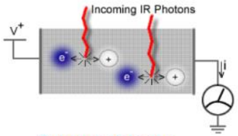
IR DETECTORS

We have two types:

- Thermal detectors: we have IR radiation incoming in our detector and then we have a first element which is called absorber. The absorber is a material with very high absorption coefficient, so able to absorb most of the incoming radiation. If it absorbs radiation, it changes its temperature, and then its temperature is measured by a classical temperature sensor, and then we have a conversion to current or voltage. So it is like a non-contact temperature sensor → we measure the radiation from the object, we absorb the radiation with an absorber and then we measure the temperature of the absorber.
- Quantum detectors: their working principle is the same of photodiodes, we have a conversion of the incoming radiation in e/h and then we readout the current. We cannot use Si because it has a high energy gap, so we need a lot of energy to generate an e/h pair, while in IR we have a longer wavelength with respect to visible range, so we need a material with lower energy gap.



Thermal detector



Quantum detector

THERMAL DETECTORS

Thermopile detector.

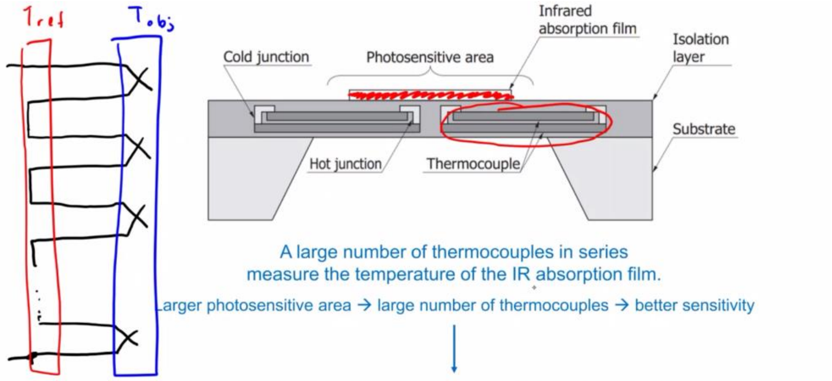
We need to have an absorber that is the red material, which corresponds to the photon sensitive area abel to absorb IR radiation. Then we want to measure the temperature of this absorber, and in this case it is done with a thermocouple.

However, one disadvantage of the thermocouple is that the generated voltage is very small, so to measure small variations of T, we cannot use a single TC, because the output voltage would be smaller than the electronic noise → we put many TCs in series to obtain a thermopile.

We connect in series the TCs by connecting the wires. All the TCs are experiencing the same temperature difference between reference junction and object temperature, so they will have the same delta-V.

Obviously, the larger the sensitive area the larger the number of TCs we can put in series and so higher the sensitivity of the instrument.

This detector has very high sensitivity but it is not possible to do large arrays of thermopiles, so thermal images cannot be acquired with thermopiles.



A large number of thermocouples in series measure the temperature of the IR absorption film.

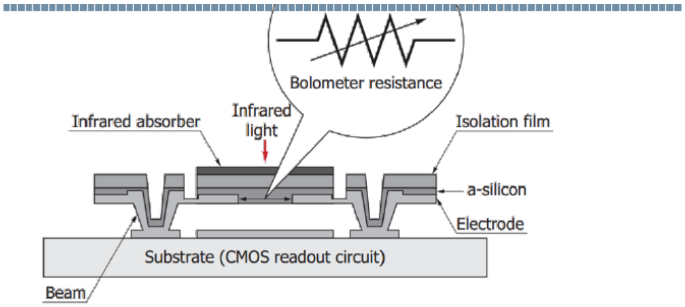
arger photosensitive area → large number of thermocouples → better sensitivity

Typically single-element detector or array with limited number of elements

Bolometer detector

Better if we want to have thermal images. Also in this case we have an IR absorber and then we measure its temperature by means of an RTD or thermistor.

The advantage of this approach is that resistance can be easily integrated in CMOS technology, so this system can be integrated with a CMOS readout circuit → we can make many pixels.



Boolometer resistance = metallic or semiconductor PTC resistor.

+

Read Out Integrated Circuit to sense the resistance variations.

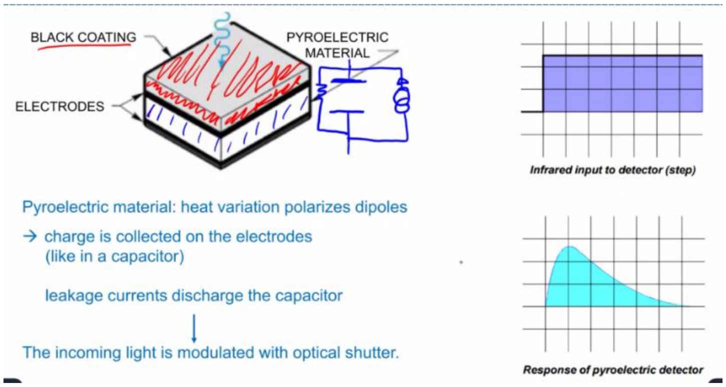
Thermal isolation between boolometer resistance and CMOS ROIC

Commonly used for multi-pixel cameras.

It is very important to create e thermal insulation between the CMOS readout circuit and the absorber. ROIC (readout integrated circuit) will be at a different temperature than the absorber, but we want to measure the temperature of the absorber → contact between resistance and readout circuit must be thin to increase the thermal resistance in order to read the real temperature of the absorber not influenced by the one of the readout circuit.

Pyroelectric detectors

Similar to piezoelectric effect, but pyroelectric materials are sensible to temperature. We have the absorber, called black coating. Then we put in contact the absorber with the pyroelectric material, that is a material that changes its atomic properties, in particular the polarization of its dipole, depending on the temperature. Then at the extremities of the pyroelectric material we place two electrodes; the overall structure can be modelled as a capacitor, because we have two electrodes isolated by a non-conductive element. So the sensor can be modelled as a capacitor and then, every time that T changes, we have a modification of the polarization of the dipole. If we modify the polarization of a dipole it is similar to inject a current in the capacitor, so we can imagine that every time that we have a variation of T we are injecting a current (like a delta of current). But then, this polarization, in some way, after the fast variation of T, relaxes to the steady state → **the detector is sensible to variation of T, not to constant temperatures**. Hence is it similar to put in parallel to the capacitor a resistor. Hence if we have a fast injection of current, at the beginning the current will integrate on the capacitor, but then the capacitor will discharge on the resistor.



Infrared input to detector (step)

Pyroelectric material: heat variation polarizes dipoles

→ charge is collected on the electrodes (like in a capacitor)

leakage currents discharge the capacitor

The incoming light is modulated with optical shutter.

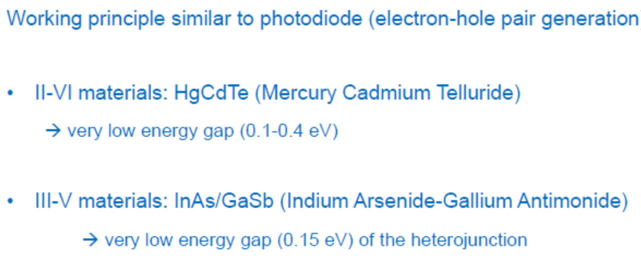
Response of pyroelectric detector

If we imagine to have a step in temperature, the output of the sensor has an increase of the voltage across the two electrodes but then a relaxation. It is a very useful sensor if we want to measure variation of temperature. If we want to monitor an environment and see if there is a cold object moving in the environment, we can check the variations of temperature.

If instead we want to measure constant temperature with these kind of detectors we need an optical shutter to be added in front of the detector and in some period of time it prevents the radiation to arrive. So sometimes we block the radiation with the optical shutter, then we open the shutter so that light can reach the absorber and then we can measure the output.

QUANTUM DETECTORS

They work like the normal PD. We cannot create them with Silicon, but we can use from II to VI materials in the periodic table combined together in one molecule, or we can combine together in a junction, we put in contact two different materials, like InAs/GaSb.



Working principle similar to photodiode (electron-hole pair generation).

- II-VI materials: HgCdTe (Mercury Cadmium Telluride)
  → very low energy gap (0.1-0.4 eV)

- III-V materials: InAs/GaSb (Indium Arsenide-Gallium Antimonide)
  → very low energy gap (0.15 eV) of the heterojunction

Spectral responsivity

For both thermal detectors and quantum detectors. It is almost the same of sensitivity, variation of the output depending on the variation of the input, and the output is different depending on the implementation we are using, it can be delta-V, delta-I in the case of quantum detectors etc.. Also the delta input variation is different, because in thermal we measure the power of incoming light (delta-P), which will determine the heating up of the absorber, while in the quantum detector we have the delta intensity of the light, more related to the number of IR photons that reach the detector.

Thermal detectors absorb all the different wavelengths, that will heat up the detector, so spectral responsivity is independent from lambda but on the overall power, while in the case of quantum detectors we have different efficiency in the absorption depending on the lambda.

Detector efficiency in the case of quantum detectors, we can say that spectral responsivity depends on lambda. It is defined as the number of e/h pairs generated depending on the number of incoming photons in the detector. The sensitivity can be computed as the photocurrent generated divided by the incoming power of light. Then current can be seen as number of e/h pairs generated multiplied by the charge of the electron, while the incoming power can be seen as the number of incoming photons multiplied by the energy of the photons, and the energy of the photon depends on Plank constant and frequency. Hence in the end there is a dependency on lambda.

Ne/Np that is the number of e/h pair generated divided by the number of photons is the efficiency of the detector and then we have a constant term lambda/1.24.
Hence in this case the sensitivity depends on wavelength both because it appears in the formula, but also because efficiency is itself a function of wavelength.

Spectral responsivity: $S = \frac{\Delta_{out}}{\Delta_{in}}$

For **thermal detectors** it is independent from λ.
Depends on the efficiency of the absorber and the sensitivity of the temperature sensor.

For **quantum detectors** ($S = \frac{I_{out}}{P_{in}}$) it depends on λ, in fact:

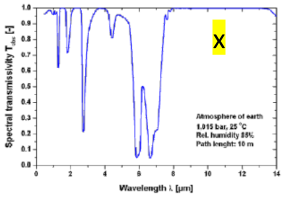
Detector efficiency: $\eta = \frac{n_e}{n_p}$ (n_e electrons generation rate, n_p incoming photons rate)

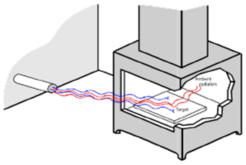
$$S = \frac{I}{P} = \frac{n_e \cdot q}{n_p \cdot h\nu} = \frac{n_e \cdot q}{n_p \cdot \frac{hc}{\lambda}} = \frac{n_e}{n_p} \cdot \frac{\lambda}{\frac{hc}{q}} = \eta \cdot \frac{\lambda}{1.24}$$

NON-IDEALITIES

Measurement with IR sensors is difficult due to non-idealities:

- Spectral transmissivity of air: not all the emitted radiation will reach the sensor because in the middle we have air that will absorb radiation. If we look at the plot, we have the spectral transmissivity of air. We have some bandwidth in which we are more able to transmit radiation (x), that is in the case of long wavelength infrared range. This is the reason why most of the IR detectors work in this range. If we try to work at 6um we have a peak of absorption of air → almost all the radiation is absorbed by air and won't reach the detector.
- Ambient radiations: imagine we want to measure the temperature of my target that is within an oven. The oven itself generate radiation that goes to the target and the target can reflect radiation from the oven to the detector. Hence we are also measuring the reflected radiation from the oven. To correct it we need to do some compensations or to perform a shielding (that however is not practical).
- Dust and particles in the ambient: they can absorb radiation, so we need to have a cleaning system for the lenses (autocleaning of the sensor) so that the measurement is not impaired.





APPLICATIONS

- Maintenance and service in industrial applications
  (moving machines, medium and high voltage facilities…)
- Measuring temperature of electronic circuits
- Over-heating of electric bus contacts
- Checking of transformers
- Localization of defective cables

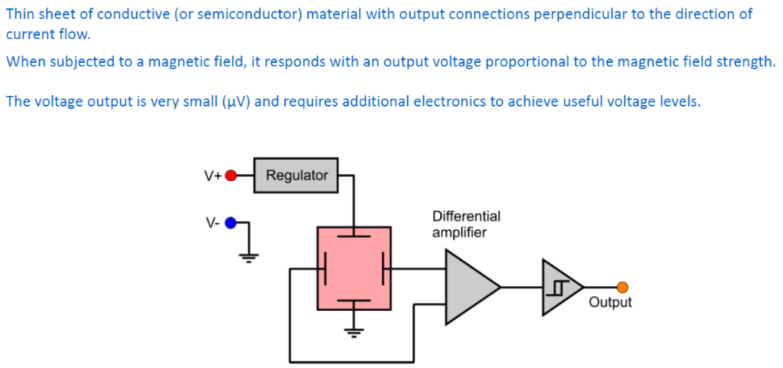
MAGNETIC FIELD SENSORS

Mainly, we have 2 types of sensors: Hall sensors and Magnetoresistive sensors.

HALL SENSORS

They have the advantage to be very simple sensors, with only 3 pins. They are made by a piece of material which can conduct current. They are very cheap sensors.

This sensor is biased with an input voltage (represented by the regulator) and the opposite side is connected to ground. Hence we will have a current thanks to the biasing voltage. We measure a voltage in the opposite side that will be proportional to the magnetic field to the Hall effect. In this case we have 4 connections, but we can have also 3 connections if one of the external differential voltage is connected to ground.



Thin sheet of conductive (or semiconductor) material with output connections perpendicular to the direction of current flow.

When subjected to a magnetic field, it responds with an output voltage proportional to the magnetic field strength.

The voltage output is very small (µV) and requires additional electronics to achieve useful voltage levels.

The main disadvantage is that the output voltage is very small, typically uV, so we need additional electronics to amplify this small voltage → typically together with the sensor we have a differential amplifier and an output stage. In the image we have a trigger because typically the sensor is not used to measure an output proportional to the magnetic field but as a trigger to detect the presence of a magnetic field. To measure the field, we will use magnetoresistive sensors.

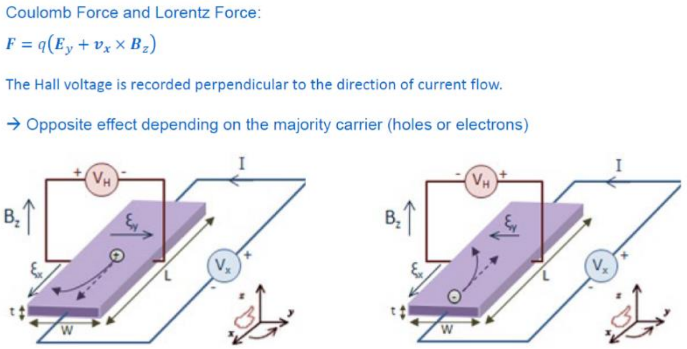
Hall Principle

We have the piece of material and the biasing voltage that makes a current flow withing the piece of material. Then, on the positive charges that represent the carriers, we will have two forces: one is related to the electric field that we are providing thanks to the voltage generator (Coulomb force), but then, if we have a magnetic field supposed orthogonal to the direction of the current, we will have also the Lorentz force. By applying the right hand rule, we can see that the Lorentz force causes the holes to drift laterally.

If we have deflection of positive charges towards the left, on average we will have a higher voltage on the left with respect to the right, so we will measure a differential voltage that will be proportional to the magnetic field.

If we consider now a negative charge (image on the right), the current is still in the same direction, but charges will move in the opposite direction. If we apply the right hand rule, we still have the electrons drifting left (because we have a negative charge, so even if the force points right, we have to invert it), but know we will have a generated voltage with opposite sign with respect to the left situation.

So it is important to know the free carriers moving in the sensor, because depending on them we will have a different direction of voltage. This information is typically provided in the datasheet as the Hall constant, that it will be positive in case of holes and negative in case of electrons.



Coulomb Force and Lorentz Force:

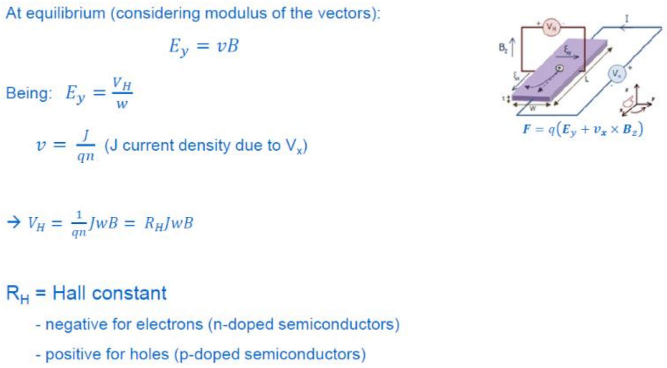
$\boldsymbol{F} = q(\boldsymbol{E}_y + \boldsymbol{v}_x \times \boldsymbol{B}_z)$

The Hall voltage is recorded perpendicular to the direction of current flow.

→ Opposite effect depending on the majority carrier (holes or electrons)

Magnetic field computation

We have the expression of the force that is the overall force acting on the free carriers. At the equilibrium, we will have a force on the charge in the y direction equal to 0. I'm considering only the force in the y direction by now, because the force on the x direction doesn't contribute to the voltage we cause in the x direction.



At equilibrium (considering modulus of the vectors):

$$E_y = vB$$

Being: $E_y = \frac{V_H}{w}$

$v = \frac{J}{qn}$ (J current density due to V_x)

→ $V_H = \frac{1}{qn} JwB = R_H JwB$

R_H = Hall constant

- negative for electrons (n-doped semiconductors)
- positive for holes (p-doped semiconductors)

$\boldsymbol{F} = q(\boldsymbol{E}_y + \boldsymbol{v}_x \times \boldsymbol{B}_z)$

Since we want the force equal to zero, we can consider when E, so the electric field in the y direction equals v*B.

First of all, what we measure in the y direction is not directly the magnetic field B, but the voltage Vh, voltage due to Hall effect. We know that magnetic filed is equal to the voltage divided by w, where w id the horizontal dimension of the material.

Then we can rewrite the velocity v of our particle, that is related to the current density, to the charge of the free carrier and to the density of free carriers.

Now, if we substitute, we can get the expression of the voltage in the y direction with respect to the magnetic field.

1/qn depends only on the material we are using, so it is a constant number and it takes the name of Hall constant. J will depend on the input provided voltage and w on the dimensions of the sensor.

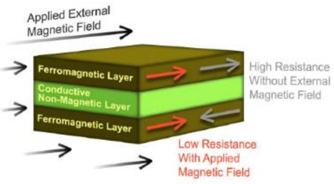
Hence the sign of the Hall constant depends on q, so then Vh will change direction depending on the charge of the free carriers.

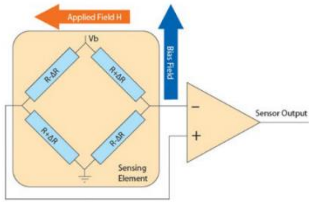
Hall sensors application

Used where we don’t want to measure precisely the B, but to use them as a switch:

- Proximity sensors or displacement sensor: if the magnet is closed to the Hall sensor the output will be triggered
- Magnetic switches
- Doors interlock: systems used to protect very dangerous environments.
- Magnetic encoders
- Current measurements
- Compass

MAGNETO-RESISTIVE SENSORS

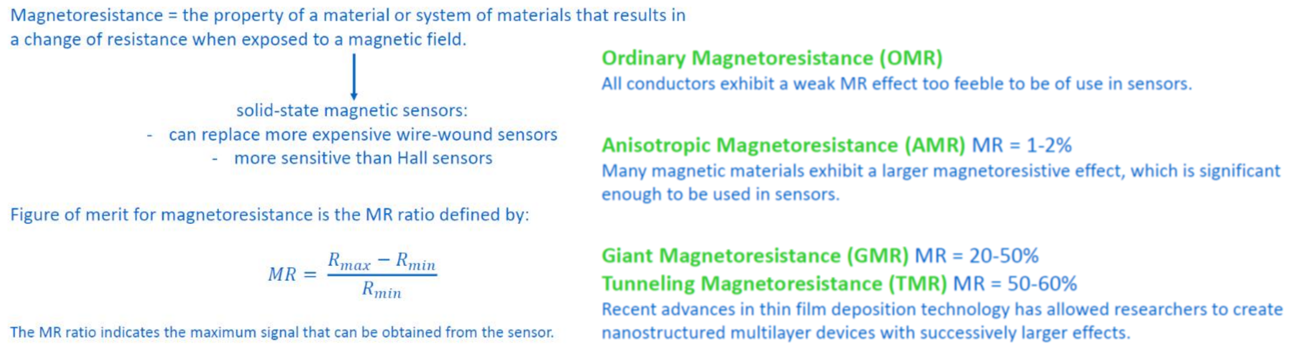

Applied External Magnetic Field
Ferromagnetic Layer
Conductive Non-Magnetic Layer
Ferromagnetic Layer
High Resistance Without External Magnetic Field
Low Resistance With Applied Magnetic Field



They are sensor based on a resistance that changes its value depending on the value of the external magnetic field. With respect to the Hall sensors, they are more sensitive to B; in particular, MR nowadays used are solid-state resistors, which, with respect to wire wound resistors (previously used), are less expensive. So we will focus only on solid state technology.

The main figure of merit of the sensor is the MR ratio, that is the maximum variation I can have of the resistance. It depends on the material and its structure.
In ordinary material, so not tailored to be MR, the MR ratio is practically negligible. Then we have the anisotropic MR (AMR), in which MR ratio is in the order of 1-2% with respect to the minimum resistance. It seems not so much but they can eb used in practical application, the variation is significant enough to use them as sensors.

Then we have giant MR effect (GMR) with a ratio of 20-50% or tunneling MR (TMR) with 50-60%. The possibility of building this material is possible thanks to the recent technology that is able to deposit very thin layers of material.



Magnetoresistance = the property of a material or system of materials that results in a change of resistance when exposed to a magnetic field.

solid-state magnetic sensors:
- can replace more expensive wire-wound sensors
- more sensitive than Hall sensors

Figure of merit for magnetoresistance is the MR ratio defined by:

$$MR = \frac{R_{max} - R_{min}}{R_{min}}$$

The MR ratio indicates the maximum signal that can be obtained from the sensor.

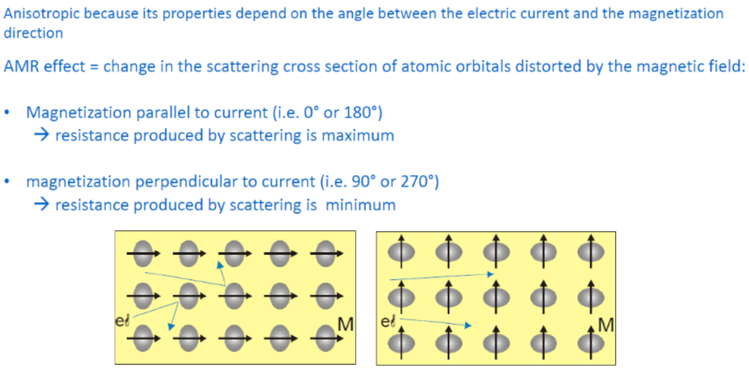
Ordinary Magnetoresistance (OMR)
All conductors exhibit a weak MR effect too feeble to be of use in sensors.

Anisotropic Magnetoresistance (AMR) MR = 1-2%
Many magnetic materials exhibit a larger magnetoresistive effect, which is significant enough to be used in sensors.

Giant Magnetoresistance (GMR) MR = 20-50%
Tunneling Magnetoresistance (TMR) MR = 50-60%
Recent advances in thin film deposition technology has allowed researchers to create nanostructured multilayer devices with successively larger effects.

AMR

Anisotropic because this resistance presents different resistance depending on the angle between the current flowing in the resistor and the magnetic field. In particular, if we look at the two images, imagine that we have a current flowing from right to left, made of electrons. The electrons in their flow can interact with the orbitals of the atoms that constitute the material in with e- are flowing. If the cross section of the atomic orbital is not circular but elliptical, depending on the orientation of the orbitals, the probability of having this interaction changes. In the case of the image, if orbitals are in the vertical direction, we increase the probability of interaction and electrons can be scattered, hence the resistance is higher. If instead the orientation of the orbitals is in the horizontal direction, the probability of interaction is smaller, so current flow is easier and resistance decreases.



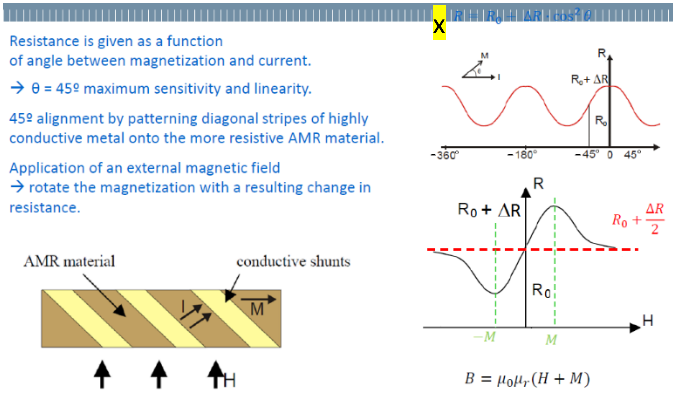
Anisotropic because its properties depend on the angle between the electric current and the magnetization direction

AMR effect = change in the scattering cross section of atomic orbitals distorted by the magnetic field:

- Magnetization parallel to current (i.e. 0° or 180°)
  → resistance produced by scattering is maximum
- magnetization perpendicular to current (i.e. 90° or 270°)
  → resistance produced by scattering is minimum

AMR working principle

The value of the resistance of AMR can be expressed by equation x. So we have a minimum resistance R0 that is the Rmin we have in the equation of the MR ratio; then we have a variation delta-R and this delta-R (that is the maximal delta-R we have) is multiplied by cos(theta), where theta is the angle between the current flow and the overall magnetization H.



X $R = R_0 + \Delta R \cdot \cos^2\theta$

Resistance is given as a function of angle between magnetization and current.

→ θ = 45º maximum sensitivity and linearity.

45º alignment by patterning diagonal stripes of highly conductive metal onto the more resistive AMR material.

Application of an external magnetic field
→ rotate the magnetization with a resulting change in resistance.

$B = \mu_0\mu_r(H + M)$

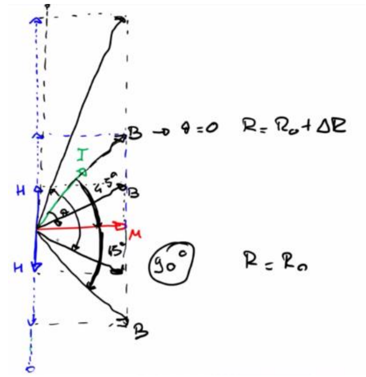
As we see, we can draw the R equation depending on R. So minimum resistance is R0, maximal is R0 + delta-R. The maximal sensitivity is close to +-45°, where we have the maximal slope and higher linearity. This is the reason why we will organize AMR to work around this point.

Hence in our AMR material we have an intrinsic magnetization of the material called M, independent from the external magnetic field H. Then we make the material in a way so that current flows in a 45° with respect to the M. To do so, we will provide an external voltage to the material; we add some **conductive shunts** in the material (yellow slices) at 45° which alternates to the AMR material. Hence the current will go in a way to cross the minimum distance withing the AMR material, whose resistance is much higher than conductive shunts → current will flow at 45°.

So without any external field we have the intrinsic magnetization M (red) and current (green) flowing at 45°. Hence theta is -45°, that is the operating point we want to have. If the external magnetic field is 0, then we will start with a resistance R that is the same that we have at -45°, that is R0 + delta-R/2. Now, let's imagine to add also an external magnetic filed H (blue).
The overall magnetic field is given by M + H (vectorial sum), and we can see that the angle between B and current reduces → we move from -45° towards 0°, so we are reducing the value of the resistance. So if we increase the H we will increase the resistance.

At a certain point we will have the maximum of the resistance, in particular when H = M. In this case B will be exactly at 45°.

When H (external magnetic field) equals M we have reached the maximal resistance. Let's now try to increase H again. If we do so, it happens that theta is now a positive angle and, at the limit if H is increase to infinite, theta will be equal to 45°. For H = inf, B will be perfectly vertical and theta = +45°, so we are coming back to R = R0 + delta-R/2.

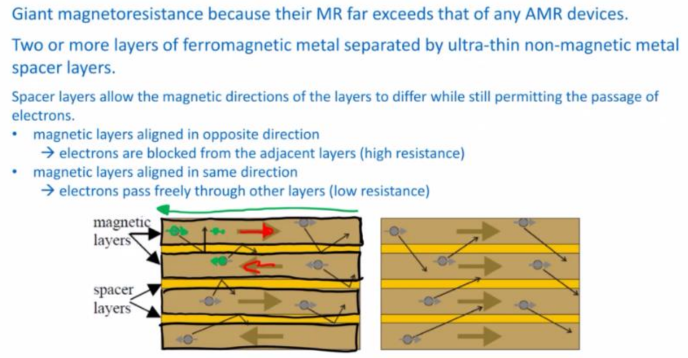


The exact same reasoning can be done with negative external fields, with theta that is not becoming positive, but more negative. If H is in the opposite direction and equal to M, B will form an angle theta (angle between current and overall magnetic field, by definition) is -90°. In this case R = R0, so if H = -M, we reach the minimum value of resistance.

In the end, we discover that the resistance value doesn't vary linearly with the external field, so better to use the sensor between -M and M, where the variation of resistance is quite linear, especially if we work around 0.

GMR working principle

GMR are used for instance in the head of hard drives. The name 'giant' is because they have a MR ration higher than AMR. In this case we have at least two, but typically more layers of ferromagnetic materials separated by ultrathin non-magnetic spacer layers (orange). The ferromagnetic layers have their own magnetization represented by the brown arrows, and then in proximity we put another ferromagnetic material separated by the orange layer that is a non-magnetic metal, which means that it makes a sort of magnetic isolation. Hence if below we have another magnet, the magnetization will be opposite (north attracts the south).


Giant magnetoresistance because their MR far exceeds that of any AMR devices.

Two or more layers of ferromagnetic metal separated by ultra-thin non-magnetic metal spacer layers.

Spacer layers allow the magnetic directions of the layers to differ while still permitting the passage of electrons.

- magnetic layers aligned in opposite direction
  → electrons are blocked from the adjacent layers (high resistance)
- magnetic layers aligned in same direction
  → electrons pass freely through other layers (low resistance)

If a current flows within this material horizontally, the free carriers are moving within the material. If the free carriers are electrons, they will move from left to right. Electrons within a ferromagnetic material will have all the same spin (the grey arrow represents the spin of the electron). Instead, the spin in the other layer will be opposite, but the direction of motion is the same in the two layers.

If one electron tries to go in the other layer, the transition is not possible with high probability because it should change its spin to flow in the other layer. So mainly all the electrons are confined in their layers → if they try to change layer they are scattered back in the same layers.

If we have an external magnetic field and the intrinsic magnetization in the ferromagnetic layers is oriented in the same direction, now electrons can change layer. All electrons will now have the same spin (right image) and so they can move easily from one layer to the other one → less scatters in the material and so the resistance will be lower.

Considerations

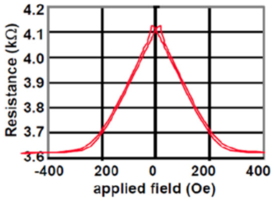
GMR are fabricated thanks to the development of very thin layer allowed by current technology. Indeed, the non-magnetic layers must be very thin in order to have a strong exchange coupling between the two magnetic layers. Strong exchange coupling means that one layer must still sense the magnetization of the other in order to have opposite magnetization when no external field is applied. So the opposite magnetization is granted only if the orange layer is thin enough to have the bottom ferromagnetic material to sense the top ferromagnetic material orientation of the magnetization.

If we now imagine to add, starting from the initial condition where we have an antiparallel alignment of the magnetization, a magnetic field, adding a H in the positive or negative direction doesn't change anything because the condition of the material is perfectly symmetric. So in the case of positive applied H I will align all the M in the same direction. For this reason, the characteristic of the resistance vs applied filed is an even function, so symmetric with respect to no applied field.

Moreover, the resistance always decreases because in the initial starting condition the M is antiparallel. The external H aligns the M until we reach a value at which almost all the layers have parallel magnetization, so we cannot observe anymore a variation of resistance.

With thin spacer layers (few atoms)
→ strong "exchange coupling" favors antiparallel alignment
(high resistance)

When an external magnetic field overcomes the interlayer coupling
→ force all of the layers to align
(low resistance)



Magnetic field in either direction
causes alignment of the magnetizations

→ R vs. H curve = even function

The only disadvantage of GMRs with respect to AMR is that we are not able to distinguish if the external field is positive or negative, because we have an even function.

GMR sensing

We can have two configurations.

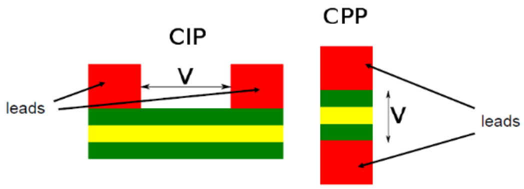
Current-in-plane of the films (CIP)

In this configuration we have a current that is crossing the material horizontally, from one red connection to the other. Depending on the orientation of the two layers the resistance will increase ore decrease.

Current-perpendicular-to-plane (CPP)

In this case we have the two terinals at the opposite sides of the material so current will flow vertically from one connector to the other. In this case we will have more sensitivity than before because in CIP it is not a problem for an electron to flow staying in the same layer, while in this case the electrons to flow from top to bottom have to pass from one layer to the other, but this transfer from one layer to the other will depend on the magnetization of the materials. Hence the sensitivity is higher but the value of the resistance is very low, because overall V in the case of CPP is very small, few mm, so the R0 is very low.

- current-in-plane of the films (CIP)
  MR is reduced because of current shunting through the layers
- current-perpendicular-to-plane (CPP)
  resistance that is too low for practical circuit applications

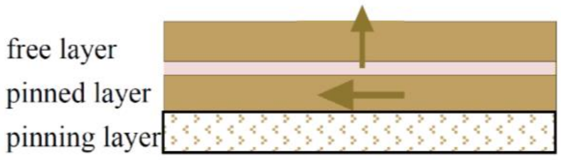


Tunneling magnetoresistance (TMR)

They exploit the *tunneling effect*, that is an effect related to quantum mechanics. In classical mechanics, if we have two levels at lower and higher energy, we cannot move from lower to higher energy without injecting energy in the system. Instead, in quantum physics, this transition can happen but with a very low probability. In theory tunneling effect cannot happen in classical mechanics, while it is explained by quantum physics.

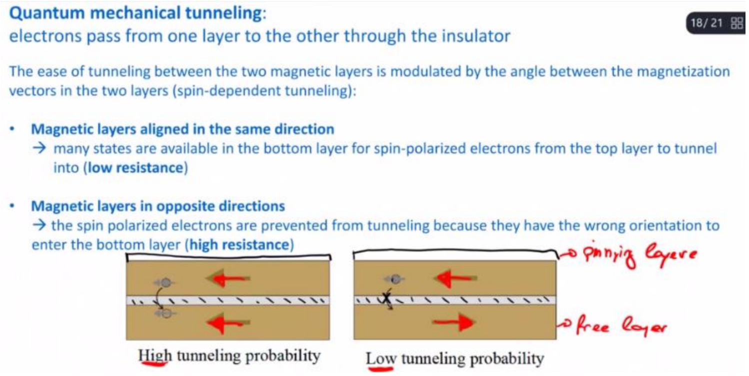
TMR uses two magnetic layers with an ultra-thin insulating layer in the middle:

- bottom layer deposited on top of an antiferromagnetic "pinning" layer
  (no net magnetization, but hold the magnetization of the adjacent ferromagnetic layer fixed in one direction)
- top layer is free to rotate its magnetic field in response to an external field
  (its rest position is made to be perpendicular to the pinned layer)



In particular, we can see the cross section of this resistance. We have a bottom layer, the pinning layer, that is a layer that has no magnetization, it is not a ferromagnetic material, but it is made in a way that is able to keep constant the magnetization of the layer on top of it, that is the pinned layer. So the magnetization of the pinned layer cannot change even if we apply a strong external magnetic field. Then we have a thin layer that isolates the two ferromagnetic materials (ultra-this isolator) and then we have another ferromagnetic layer (free layer) in which we have an intrinsic magnetization orthogonal to the layer, but the overall magnetization can be modified by the external magnetic field.

So at rest in the free layer I have a magnetization orthogonal to the pinning layer.
Since we have the insulator between the two ferromagnetic layers, theoretically (hence in classical physics) e- cannot move from one layer to the other. But for quantum mechanics this transition is possile but with very low probability. The probability is ultra-low if the two magnetization are in the opposite directions, because we have the insulator and moreover the electrons have to change their spin. If instead the two magnetization are in the same direction we have only the barrier of the insulator.
Using quantum mechanics, the probability in this latter case is higher, still low but higher with respect to the previous case.

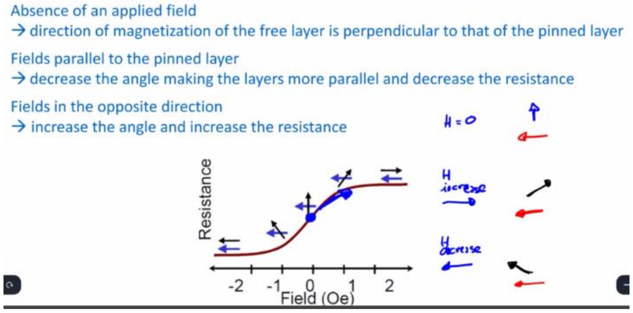


Quantum mechanical tunneling:
electrons pass from one layer to the other through the insulator

The ease of tunneling between the two magnetic layers is modulated by the angle between the magnetization vectors in the two layers (spin-dependent tunneling):

- **Magnetic layers aligned in the same direction**
  → many states are available in the bottom layer for spin-polarized electrons from the top layer to tunnel into (**low resistance**)
- **Magnetic layers in opposite directions**
  → the spin polarized electrons are prevented from tunneling because they have the wrong orientation to enter the bottom layer (**high resistance**)

At the beginning, with no external field, we are in the middle condition with respect to the two in the image, but then if I applied a field H and H is directed from left to right, the pinning layer has the same magnetization, while the one of the pinned layer will become opposite.



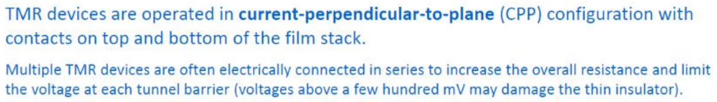
Absence of an applied field
→ direction of magnetization of the free layer is perpendicular to that of the pinned layer

Fields parallel to the pinned layer
→ decrease the angle making the layers more parallel and decrease the resistance

Fields in the opposite direction
→ increase the angle and increase the resistance

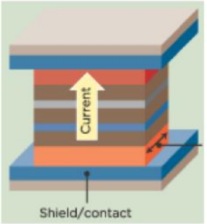
TMR sensing

In this case, only the CPP configuration is possible, otherwise the electrons will never cross the insulator and we won't see any tunneling effect. Sometimes, to increase the sensitivity and the R0 of the material we put many layers, like many resistances put in series.



TMR devices are operated in **current-perpendicular-to-plane** (CPP) configuration with contacts on top and bottom of the film stack.

Multiple TMR devices are often electrically connected in series to increase the overall resistance and limit the voltage at each tunnel barrier (voltages above a few hundred mV may damage the thin insulator).



MR applications

Used when we want to sense a magnetic field and measure it, not just use them as switches.

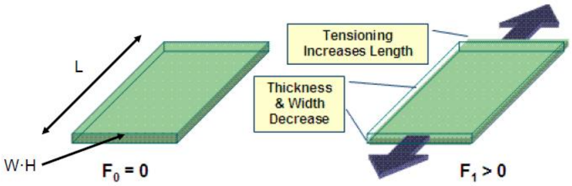
- Eddy current sensing: to detect cracks or flaws into materials
- Stray field sensing: to detect defects and non uniformities in magnetic materials
- Remote monitoring of stresses in embedded steel reinforcement and fasteners
- Displacement encoders

RESISTIVE FORCE SENSORS

STRAIN GAUGES

They are resistive materials, very similar to RTD. The technology to manufacture them is the same of RTD, but what changes is the substrate, that is no more rigid (ceramic) but now it must be elongated → plastic substrate.

BASIC OF MECHANICS



Stress: $\sigma = \frac{F}{WH} \left(Pa = \frac{N}{m^2}\right)$

Strain: $\varepsilon = \frac{\Delta L}{L}$ $(a.u.\ \ typically\ \mu\varepsilon)$

Young modulus: $E = \frac{\sigma}{\varepsilon}$ $\left(Pa = \frac{N}{m^2}\right)$

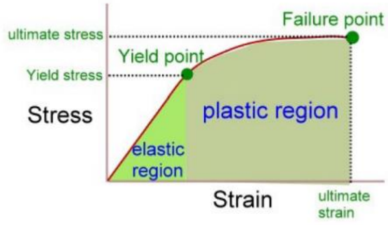
Poisson ratio: $\nu = -\frac{dW/W}{dL/L} = -\frac{dH/H}{dL/L}$ $(a.u.)$

Every time we apply a stress on a material, the stress is defined as the force divided by the cross-section area of the material, so it has the same units of pressure. Every time we apply a force we will cause an elongation of the material, expressed in terms of strain.
Stress and strain are then related thanks to the Young modulus E. The unit of E is still Pascal because strain is adimensional, even if we sometimes we find u-strain as unit of measure.

If we apply a force, we will have an elongation, but since the total volume of the material cannot change, we will also reduce the cross-section of the material, which will become thinner. So we both modify w and h. The reduction of w and h is quantified through the Poisson ratio.

Stress vs strain curve

The Young modulus is not a constant value, it is constant just in the elastic region of the material. If we enter in the plastic region we start to have a saturation until we arrive to the failure point where we break the material. We have always to work in the elastic region to have linear relationship and also not to permanently modify the material.



In the elastic region:

$\sigma = E \cdot \varepsilon$ → E slope of the curve

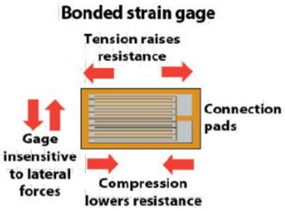
RESISTIVE SG

A resistive SG is a resistance whose overall resistance depends on the tension or compression we have on the material we are measuring the force.
We have a resistor fabricated on a flexible substrate that can be elongated or compressed. If we apply a tension on the material, the length increases and area decreases, so the resistance will change. For extension, since we increase length and decrease A we increase the resistance, while for compression we decrease it. The only direction I which the SG is sensible is the longitudinal one, because on the other side we don't see any significant variation. So resistance can change its value due to geometrical aspects.

Moreover, there are some materials also with **piezoresistive effect**, which means that also the resistivity of the material ro changes with the applied stress, and the amount of stress is expressed in terms of Young modulus.

$R = \rho \cdot \frac{L}{A} = \rho \cdot \frac{L}{WH}$

- Tension:
  L increases
  A decreases (necking)
  → R increases
- Compression:
  L decreases
  A increases
  → R decreases


Bonded strain gage
Tension raises resistance
Connection pads
Gage insensitive to lateral forces
Compression lowers resistance

Materials with piezo-resistive effect: $\frac{d\rho}{\rho} = \beta \cdot \sigma = \beta \cdot E \cdot \varepsilon = \beta \cdot E \cdot \frac{dL}{L}$

Gauge factor

The sensitivity of the SG is expressed in terms of Gauge factor. It is defined as the variation of the output, so resistance, with respect to the variation of strain in the material. If we do the partial derivative of the resistance, we can see that we have the partial derivative of all the components.

Sensibility of the strain gauge: $G = \frac{dR/R}{\varepsilon} = \frac{dR/R}{dL/L}$

$$R = \rho \cdot \frac{L}{WH}$$

$$\rightarrow \quad \frac{dR}{R} = \frac{dL}{L} - \frac{dW}{W} - \frac{dH}{H} + \frac{d\rho}{\rho}$$

In piezo-resistive materials: $\frac{dR}{R} = \frac{dL}{L} + \nu\frac{dL}{L} + \nu\frac{dL}{L} + \beta E\frac{dL}{L}$

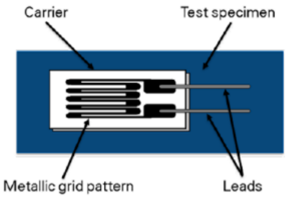
$$\frac{dR}{R} = (1 + 2\nu + \beta E)\frac{dL}{L}$$

$\rightarrow G = 1 + 2\nu + \beta E$

Thanks to the Poisson ratio, we can compute dw/w.

In case of piezoresistive materials, we can put also the beta part in the expression of the dR/R. In conclusion, in all the components of dR/R we have the dependance on the strain. So we collect the strain and we can see that the strain is related to a constant value. In the end we get the Gauge factor.

Gauge factor depends on the material, and they are typically from 1 to 100 for semiconductors. The fabrication process is the same for RTD, but with a different substrate, that must be flexible.

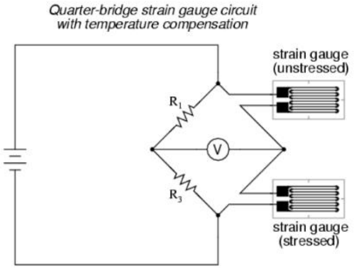


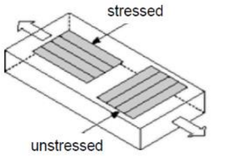
Carrier:
- flexible
- electrical insulation

Fabrication → photolithography
- photoresist deposition (positive or negative)
- exposure to light
- etching
- photoresist removal

TEMPERATURE COMPENSATION

SGs, since fabricated with the same technology of an RTD; will change resistance also with the temperature, so we need to correct temperature variations → we put to SGs in orthogonal positions. One is longitudinal and will experience variation in the same direction of the stress, the other one in a direction orthogonal to the stress, so it won't experience any change in resistance but it experiences the same temperature. Using a WB, we can put both the SGs on the same side of the WB and if we have a temperature variation we increase both the resistances but they are matched, so compensated.

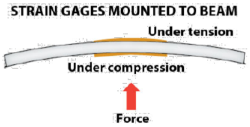

Quarter-bridge strain gauge circuit with temperature compensation
strain gauge (unstressed)
R_1
V
R_3
strain gauge (stressed)

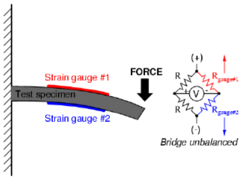


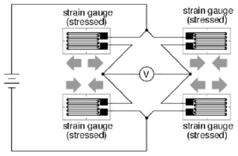
Two orthogonal strain gauges are used to compensate the temperature variations

BENDING

We want to measure if a material is bended. In particular, one face is under tension, the other under compression, so that we are trying to bend the material. I can put both the SGs on the longitudinal side of the material I want to measure, but one will experience tension, and the other compression. In this case one will increase resistance (under tension) and the other under compression will decrease it → imbalance in the output of WB. To double the sensitivity, we can double the SGs and put on the opposite sides of the WB the other two SGs, in antiparallel configuration.





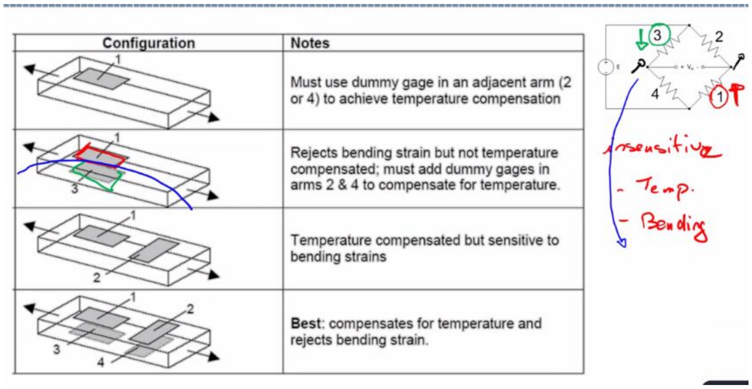


SGs CONFIGURATIONS

There are some configurations if we want to compensate temperature variations and also to compensate bending and measure only if we have a compression or tension. I want something insensitive both to temperature variations and bending, so able to measure only tension or compression.

In configuration 1 we put only 1 SG, but it is not used because sensible both to temperature and bending. The other possibility is to place the SGs in two different places, one on the top of the material and the other in the bottom and we place them in antiparallel in the WB. If we have bending, one resistor increases resistance, the other decreases it. If the other resistances in the WB are fixed, the voltage on the left will increase but also the voltage on the right. So overall I'm increasing the voltage on both sides, so overall I won't see bending, I'm compensating it (about 0V at the output will be seen,

about because the equation is not perfectly symmetric due to component mismatch). However it doesn't compensate for temperature variations.
Then the third configuration has the SGs on the same side but orthogonal one with respect to the other. It compensates temperature variations but not for bending.



Configuration	Notes
1	Must use dummy gage in an adjacent arm (2 or 4) to achieve temperature compensation
1 3	Rejects bending strain but not temperature compensated; must add dummy gages in arms 2 & 4 to compensate for temperature.
1 2	Temperature compensated but sensitive to bending strains
1 2 3 4	**Best**: compensates for temperature and rejects bending strain.

3
2
E
V_o
4
1
Insensitive
- Temp.
- Bending

The last configuration compensates both for bending and temperature.

APPLICATIONS

- Railroad maintenance, to check if there are any stretches in the longitudinal side
- Smart bridges
- Wind tunnels
- High precision robotic medical applications
- Bathroom scales to measure the bodyweight of a person

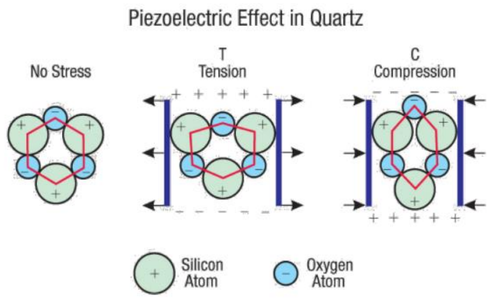
PIEZOELECTRIC FORCE SENSORS

They are used to directly measure force. SGs measure stress, so force indirectly, because they measure the strain of an object.

PIEZOELECTRIC EFFECT

Physical effect that we have in materials like quartz in which we have a reticle made with molecules with different charges. Some part of the reticule is positively charged and some parts are negatively charged. Normally, these charges are distributed in a balanced way, and in the end we don't have a macroscopic separation of charges and the material is neutral. If we apply tension or compression we have an initial distortion of the reticule, so we have a redistribution of charges. If we put tension, for instance, positive charges are located more on the top, and on the bottom more negative charges → we have created a separation of charges. The same for compression.

This modification of the reticule structure is <u>instantaneous</u>, but then it happens a relaxation of the reticle. Even if we are modifying it, the reticle structure will relax. It is similar to the pyroelectric effect seen for IR sensors. Also in this case the variation of voltage depends on the variation of force, on its derivative, if we apply a constant force we don't see any voltage.



= electric charge that accumulates in certain solid materials (e.g. crystals, ceramics, and biological matter) in response to applied mechanical stress.

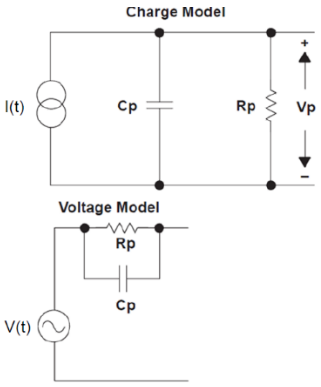
Or opposite phenomena: voltage → displacement

SENSOR MODEL

We can use both the charge and voltage model. In the charge model we have a current generator, in the other a voltage generator and they are like the Norton and Thevenin equivalents.

As for the charge model, we can model our material, our quarts, with a capacitor (the material can be considered an insulator because without carriers inside) and when we apply a stress we are like charging or discharging the capacitor.

So the capacitor models our quartz that can be charged or discharged by a current generator that is able to create pulses of current. Then we put also a resistor Rp in the model that is not a physical resistor, but it models the fact that the material returns by itself to the non-stress condition with a certain time constant.



$$I(t) = \frac{dQ(t)}{dt}$$

$$I(s) = s \cdot Q(s)$$

$$V(t) = \frac{dQ(t)}{dt} \cdot R_p \cdot \left(1 - e^{-t/(R_p C_p)}\right)$$

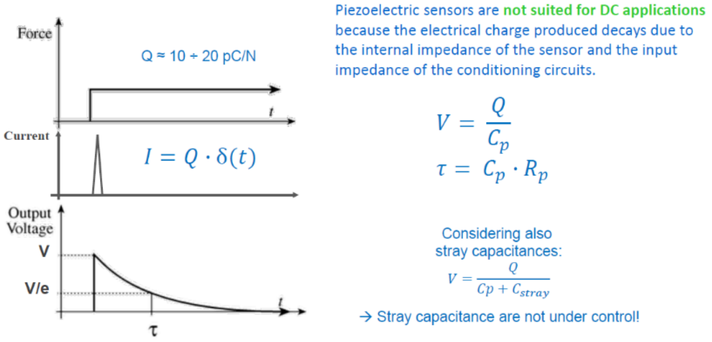
$$V(s) = s \cdot Q(s) \cdot \frac{Rp}{1 + sC_pR_p}$$

SENSOR RESPONSE

Imagine we apply a step force, our current generator that models the separation of charges will create a spike of current, a Dirac delta with an area equal to Q, where Q is the equivalent charge we put at the two plates of the capacitor.

At the beginning all the charge will be stored on the capacitor because in the delta the frequency is infinite, so the impedance of the C is very low, and all the current will go in the capacitor and charge it. Then, after this first period in which we are charging the capacitor, the current generator goes back to 0 and the capacitor starts to discharge on the resistor. So at the beginning we have a peak of voltage that can be easily computed as Q/C. This is the maximum voltage we can have across the capacitor, then we will have the discharge of the capacitor by a time constant given by R*C.

This is in the ideal case of no parasitic capacitances, but we will always have some cables connecting the sensors and hence the coupling capacitances related to the cables. The stray capacitances are always sum to Cp because they are in parallel with respect to Cp. The main issue is that Cstray cannot be controlled precisely, it cannot be adjusted. We want the output to be as much independent as possible from Cstray → we improve the readout circuits.



Piezoelectric sensors are not suited for DC applications because the electrical charge produced decays due to the internal impedance of the sensor and the input impedance of the conditioning circuits.

$$V = \frac{Q}{C_p}$$

$$\tau = C_p \cdot R_p$$

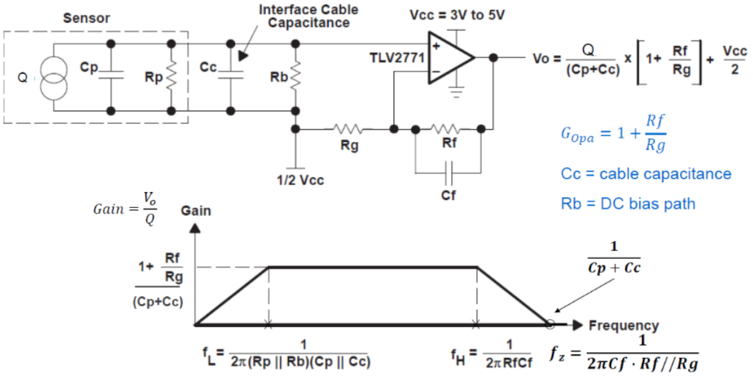
Considering also stray capacitances:

$$V = \frac{Q}{Cp + C_{stray}}$$

→ Stray capacitance are not under control!

READOUT CIRCUIT – VOLTAGE MODE AMPLIFIER

Vout = Q/(Cp+Cs) where Cs is the stray capacitance multiplied by a gain and some DC voltage. So the output is influenced by Cc. In this case it is important to check if Cc << Cp to discard it possibly, and Cp is quoted in the datasheet of the sensor.



$$Vo = \frac{Q}{(Cp+Cc)} \times \left[1 + \frac{Rf}{Rg}\right] + \frac{Vcc}{2}$$

$$G_{opa} = 1 + \frac{Rf}{Rg}$$

Cc = cable capacitance

Rb = DC bias path

$$Gain = \frac{V_o}{Q}$$

$$\frac{1 + \frac{Rf}{Rg}}{(Cp+Cc)}$$

$$\frac{1}{Cp + Cc}$$

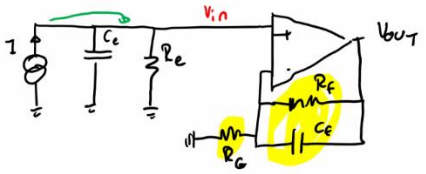
$$f_L = \frac{1}{2\pi(Rp \,||\, Rb)(Cp \,||\, Cc)} \quad f_H = \frac{1}{2\pi RfCf} \quad f_z = \frac{1}{2\pi Cf \cdot Rf//Rg}$$

On the left we have the model of our sensor. Then we have Cc that represent the stray capacitances. Rb is the biasing resistor. We need it because the power supply of the operational amplifier is a single-ended power supply, we can only provide a voltage from ground to a positive one, not a negative one. Since it is a single power supply opamp, also the output can be positive. This means that if we expect to have both tension and compression, the charge from the current generator is both positive or negative, so we may have both positive and negative voltages at the input of the opamp. So what we can do is to shift this voltage thanks to a biasing.

So instead of connecting the sensor to ground, we can connect it to 1/2Vcc. It means that our capacitor will be charged or discharged but with respect to 1/2Vcc and hence the input of the opamp is either smaller or bigger than 1/2Vcc. The resistor Rb is used to create a DC path for the voltage to reach the node, because in DC the capacitor are open circuit and we cannot consider Cp and Cc, and Rp is not a physical resistance. Hence we need a resistor Rb to reach the (+) terminal of the opamp. Then the configuration of the opamp is a non-inverting one.

We can now consider the following circuit in which we have the equivalent capacitor and resistor. This goes to the input of the amplifier on which we have Rf and Cf and then we have Rg. Our main purpose is to compute the relationship that exists between Vout and the charge Q. Because in the datasheet we found the relationship between the Q generated and the force. So we have to obtain the relationship between Vout and Q and so we can get the relationship between Vout and force.

Before this, we can find the relationship between Vout and the current. Since I = s*Q, we can consider the circuit divided in two parts. A first one is made in order to find Vin, and then Vout = Vin*(1+Rf//(1/sCf)/Rg).



$$\frac{V_{OUT}}{Q} \qquad \frac{V_{OUT}}{I}$$

$$I = s \cdot Q$$

$$V_{in} = I \cdot R_e /\!/ \frac{1}{sC_e} = I \frac{R_e}{1 + sC_eR_e} \qquad \begin{cases} R_e = R_p /\!/ R_b \\ C_e = C_p + C_c \end{cases}$$

$$V_{OUT} = V_{in} \cdot \underbrace{\left(1 + \frac{R_f /\!/ \frac{1}{sC_f}}{R_G}\right)}_{\text{non-invertig stage}} = V_{in}\left(1 + \frac{R_f}{1 + sC_fR_f} \cdot \frac{1}{R_G}\right)$$

Vin is given by the current that goes in the parallel between Ce and Re. As for Vout, we have to consider the non-inverting configuration gain. Then we can do some very simple computations and we will obtain Vout.

Since Vin = I*Re//(1/sCe), we can substitute and we can the relationship existing between Vout and I. Now, I want the relationship between Vout and the charge: I = sQ.

$$V_{OUT} = V_{in} \cdot \left(1 + \frac{R_f}{R_1}\right) \frac{1 + sC_fR_f /\!/ R_G}{1 + sC_f \cdot R_f}$$

$$V_{OUT} = I \cdot \frac{R_p /\!/ R_b}{1 + s(C_p + C_c) \cdot R_p /\!/ R_b} \cdot \left(1 + \frac{R_f}{R_G}\right) \frac{1 + sC_fR_f /\!/ R_G}{1 + sC_fR_f}$$

$$V_{OUT} = s \cdot Q \frac{R_p /\!/ R_b}{1 + s(C_p + C_c) \cdot R_p /\!/ R_b} \cdot \left(1 + \frac{R_f}{R_G}\right) \frac{1 + sC_fR_f /\!/ R_G}{1 + sC_fR_f}$$

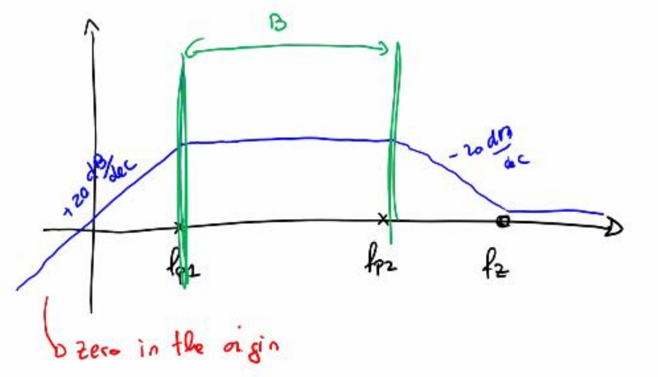
Now I divide by Q to find the relationship, grouping all the constant parts.

$$\boxed{\frac{V_{OUT}}{Q} = R_p /\!/ R_b \cdot \left(1 + \frac{R_f}{R_G}\right) \frac{s \cdot (1 + sC_fR_f /\!/ R_G)}{(1 + s(C_p + C_c) \cdot R_p /\!/ R_b)(1 + sC_fR_f)}}$$

I want now to plot this t.f. we know that s = j*omega, and we have one zero in the origin and another zero with a tau given by CfRf//Rg. Then we have two poles at the denominator. Typically, the first pole is given by the characteristics of the sensor itself, and typically Cp is big (but this is an assumption I'm doing), and hence it is a LF pole. For sure, we want to have the first pole at lower frequency with respect to the other one depending on the Rf and Cf. We want tau-p-1 > tau-p-2.

Then if we try to compare tau-p-2 with the tau of the zero, the capacitance is the same, but the resistance of the zero is smaller, because it is Rf in parallel to something, so smaller than Rf itself. Hence tau-p-2 is bigger than tau related to the zero.

If we try to plot the whole thing, we have the zero in the origin, then a pole fp1 due to the input, then a second pole fp2 and the zero at fz. Since we have one zero in the origin, we start with a slope of +20 dB/dec. Then we have the pole and we loose 20dB/dec.
Obviously, we need to use this circuit when the gain is flat, so we are interested in using the circuit in the green bandwidth of interest.



Which is the gain in this flat zone?
We are after the first pole, but before the second pole and the zero. Being after a pole means that we are at an omega (or frequency, is the same apart of a factor 2*pi) where the omega is larger than the omega of the first pole but smaller than the one of the second pole and the zero in the flat band. If we look at the t.f., if omega is much higher than 1/tau, it means that omega*tau is much bigger than 1 and s = 1+j*w*tau can be considered as j*w*tau.
If instead we are lower than the pole, w << 1/tau, it means that we can neglect the second part and consider only the 1.

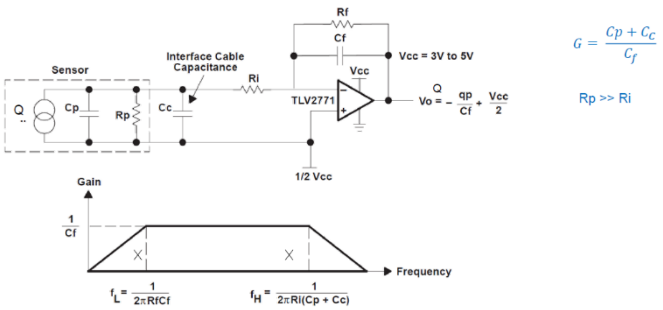
$$1 + s\tau = 1 + j\omega\tau \qquad \omega \gg \frac{1}{\tau} \rightarrow \cancel{1} + j\boxed{\omega\tau} \qquad \omega \ll \frac{1}{\tau} \Rightarrow 1 + \cancel{j\omega\tau}_{\ll 1}$$

If we now consider the t.f., if I'm after the first pole, it means that omega is much bigger than the 1/tau of the first pole, so I can neglect the 1 for that pole. It is much bigger because we are consider a decade after the pole. For the second pole and the zero I can do the opposite reasoning. If I neglect the 1, I can simplify the s with the s at the numerator, and Rp//Rb.
In the end I obtain that Vout/Q in the bandwidth of interest is (1+Rf/Rg)/(Cp+Cc). We can see that this gain is the one expected, and unfortunately it depends on the cable capacitances.

READOUT CIRCUIT – CHARGE MODE AMPLIFIER

It is based on a transimpedance amplifier. If we consider to have a shortcircuit, all the current generated by the charge generator will go in the feedback, because we see a very low impedance. This is very good because the output will depend on the parallel between Rf and Cf, so independent on Cc. In the ideal case where Ri = 0 of a perfect virtual ground, Vout = I*(Rf // 1/sCf).

Now, if we go on, we can write as I = sQ, so Vout/Q = s*(Rf/(1+sCfRf). So mainly, without the Ri input resistance, we will have a t.f. with a zero in the origin and then a pole given by the tau = CfRf. (HP filter t.f.). However, in general the t.f. has not a gain different from zero at infinite frequency, in the sense that the bandwidth cannot be infinite, so after a certain point we will have a HF pole for instance given by the HF pole of the amplifier that limits the bandwidth. However, the HF bandwidth limitation we don't want to have it limited by something 'not really under control' like the internal poles of the amplifier, but we want to establish the pole also at HF. So we can add another pole. This is the reason why we put the resistance Ri. This Ri introduces a new pole at the frequency given by Ri*(Cp+Cc).



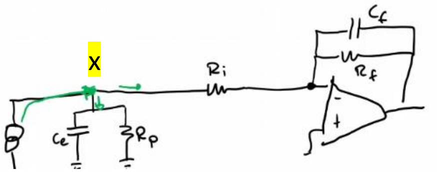
$G = \frac{Cp + C_C}{C_f}$

Rp >> Ri

V_{out} depends only on the small and well controlled C_f

Working frequencies: $f_L < f < f_R$ (f_R = resonance frequency)

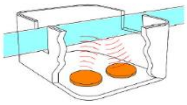
To compute the pole, we need to do a current divider at node x. We will discover that if Ri is small, we have a LF pole that is the same already computed given by Rf*Cf, and the HF pole given by Ri.

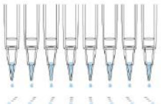


In the end, the gain depends only on the feedback capacitance. This second solution is a little bit more critical in terms of stability, we need to select Ri correctly to keep the circuit stable, otherwise the output will start to oscillate.

Applications

- Force measurements
  (e.g., force platform for rehabilitation or sport)
- Energy harvesting
  (converting the available energy from the environment from sources such as ambient temperature, vibration or air flow)
- Ultrasonic technology
  (both actuator and sensor,
  e.g. proximity, detection of gas bubbles)
- Actuators
  (pumping and dosing, production of homogeneous aerosol)





Piezoelectric effect happens in both the senses, so if we apply a voltage then we will create a modification of the material. Hence the piezoelectric material can be used also as an actuator, so we can generate ultrasound with the material. It can be used either as a sensor or an actuator.

DISPLACEMENT AND DISTANCE SENSOR

Displacement and distances can be measured with a lot of technologies:

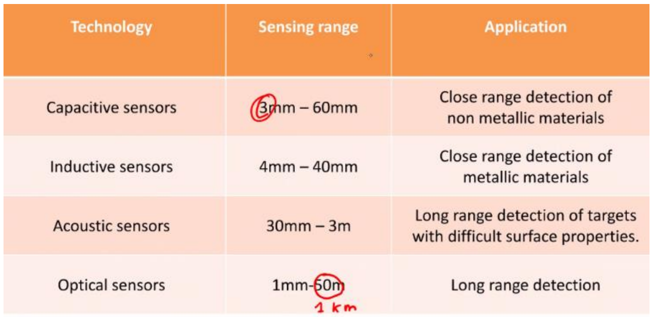
- Potentiometers
- Capacitive sensors
- Inductive sensors
- Acoustic sensors
- Optical sensors
- Magnetic sensors

With these technologies we can measure:

- Linear displacement
- Angular displacement
- Proximity (digital detectors, presence or not present)/distance

PROXIMITY/DISTANCE SENSORS COMPARISON

To measure distance we can use different technologies, and they differentiate for the sensing range.

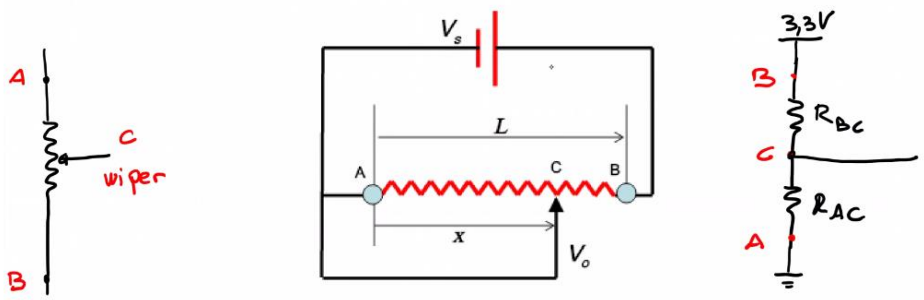


Technology	Sensing range	Application
Capacitive sensors	3mm – 60mm	Close range detection of non metallic materials
Inductive sensors	4mm – 40mm	Close range detection of metallic materials
Acoustic sensors	30mm – 3m	Long range detection of targets with difficult surface properties.
Optical sensors	1mm-60m 1 km	Long range detection

POTENTIOMETERS

They are the easiest detectors we can have. They can be used both for measuring rotation or to measure linear displacement. The working principle is simple; we have a resistor and two terminals, A and B and a third terminal which is C and typically it can move. So we can move up and down the C terminal (called wiper). So we are like performing a voltage divider AC – CB. In the intermediate node C we readout the voltage.

Resistive potentiometer = resistance element with movable contact (slider).



$$V_0 = V_s \cdot \frac{R_{AC}}{R_{AC} + R_{CB}} = V_s \cdot \frac{R_{AC}}{R_{AB}}$$

V_0 α R_{AC} α position

Potentiometers’ constructions

They can be either linear or rotary, if the resistance is in a circular configuration. Even linear potentiometer can measure rotary movements by means of moving rods to perform the transformation from rotation to translation and vice versa.

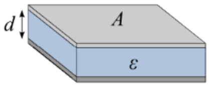
CAPACITIVE SENSORS

Used both for linear displacement and proximity. In a parallel plate capacitor, the capacitance can be computed as the dielectric constant in vacuum, relative to the dielectric, area and distance. The quantity of interest are epsilon-r, area and distance. We can have also cylindrical capacitor, and still the equation depends on epsilon-r, the length and the two radius of the parallel cylinder. We use capacitors that can be considered thin cylinders, which means that the two radius of the internal and external cylinders are very similar → R1 = R2 = R, and R1 – R2 = W ~ 0. Hence we can rewrite the equation in a more simple way.

The final equation is more similar to a parallel plate capacitor equation, where 2*pi*R*l can be considered the area and W the distance.

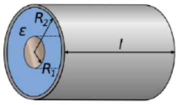
Parallel plate capacitor:

$$C = \varepsilon_r \varepsilon_0 \frac{A}{d}$$



Cylindrical capacitor:

$$C = \frac{2\pi\varepsilon_r\varepsilon_0 l}{ln\left({R_2}/{R_1}\right)}$$



Thin cylindrical capacitor

($R_1 \approx R_2 = R$ $R_1 - R_2 = W$ → $C = \frac{2\pi\varepsilon_r\varepsilon_0 l}{ln(1+W/R)}$)

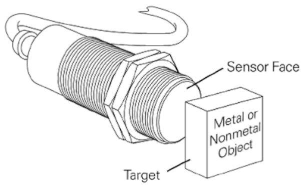
$$C = \varepsilon_r \varepsilon_0 \frac{2\pi R l}{W}$$

Working principle

In general, we compute the change of capacitance that are due to the chance in epsilon-r, area or distance.

NB: capacitive sensors can be used both for metallic and non-metallic object, and this is different from the inductive sensors.

Capacitive sensors produce an electrostatic field, the capacitance is modified by the presence/absence or the displacement of a metallic or non-metallic object at small distance from the sensor.



Capacitance variations:

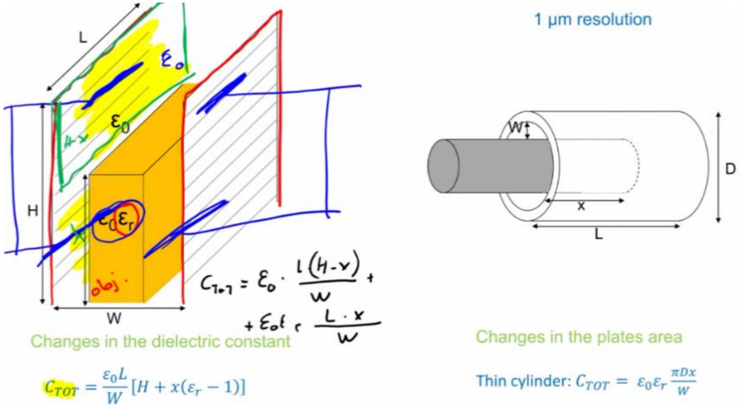
- Changes in the dielectric constant
- Changes in the distance between two plates
- Changes in the plates area

Displacement capacitive sensors

On the left we have a parallel plate capacitor, and the orange object is moving. It is the object which we want to measure the displacement. In practical cases the orange object is mechanically connected

with the object of which we want to measure the displacement. The orange material has an epsilon-r much different than 1, because we want to create a difference between the upper area where epsilon-r is the one of the void (1), and the other part (the bottom one) that sees the epsilon-r of the material. So to compute the overall capacitance we can consider to have two capacitors in parallel.
Hence the total capacitance is given by the sum of the two contributions. In the formula, H is the overall height of the structure.

The total capacitance depends on the displacement of our object in a linear way, so we have a constant part plus one linear dependent on x → we are changing the epsilon of one part of the object. The issue is that this kind of solution depends on epsilon0, that we don't want to have because in the upper part we don't have exactly void, but air, and epsilon of the air depends on temperature, pressure etc. → total displacement depends also on temperature and pressure.



$$C_{TOT} = \varepsilon_0 \cdot \frac{L(H-x)}{W} + \varepsilon_0\varepsilon_r \frac{L \cdot x}{W}$$

Changes in the dielectric constant

$$C_{TOT} = \frac{\varepsilon_0 L}{W}[H + x(\varepsilon_r - 1)]$$

Changes in the plates area

Thin cylinder: $C_{TOT} = \varepsilon_0\varepsilon_r \frac{\pi D x}{W}$

The second possibility is with a cylindrical configuration. The gray object is moving inside out with respect to an external cylinder. In this case we will modify the area of the capacitor, that is given only by the portion in which the two cylinders are facing. We can do some simple computation in which we compute the overall capacitance. We will use the approximated equation for thin cylinders; as a total length of the cylinder we will consider the portion of length in which the two cylinders are concentrical.
Also in this case we have a total capacitance proportional to x (displacement), we don't have a constant part that we have in the parallel plate case, but we still have the dependency on epsilon0.

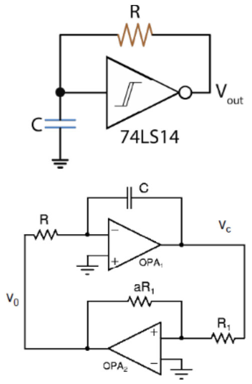
Square wave oscillators

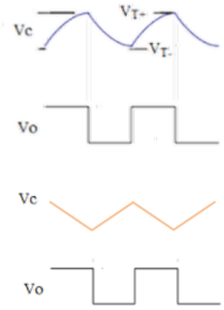
How can we measure the value of a capacitance we have to relate it to the displacement of the object?
The most used way is to use oscillators. We can build some oscillator which frequency of oscillation depends on the capacitance value. In the image we have two examples of square wave oscillators, where the to capacitors C are the one we want to measure (variable capacitors).
In the first case we have a digital inverter, in which if the input voltage exceeds the voltage threshiold the output goes to 0, while if we are below the logic threshold the output is 1 (it implements a digital NOR). It is important that the trigger is a Schmidt trigger, which has an hysteresis, so that the thresholds for a value from low to high is different from high to low.
If we suppose at the beginning the output is high, the input will charge to a high value, and as soon as the input reaches the threshold, the output will go low, and the capacitor will discharge until we reach the second threshold that provides the commutation from low to high and then the cycle begins again.

We have a RC network where the tau depends on the value of the C, so by knowing the R we can get the capacitance value.

The second circuit includes an integrator in the upper branch (negative feedback). We will integrate the current injected in the capacitor, while the second part has a positive feedback that implements a Schmidt trigger, and it implements an hysteresis. So we don't have a single threshold but two different thresholds. The output of the trigger is always saturated at the power supply, so the output can be either the low or high power supply, and it remains constant in one part of the period, hence we are injecting a constant current to the capacitor, that hence it is charging and discharging with a constant slope.

When the output has a commutation we change the current direction and we charge or discharge the capacitor. What matters is that the slope that we have in the triangular wave depends on the value of the capacitor, since I = C*delta-V/delta-t. We use deltas and not derivative because we have a constant current.
In general, if C increases the oscillation frequency decreases.

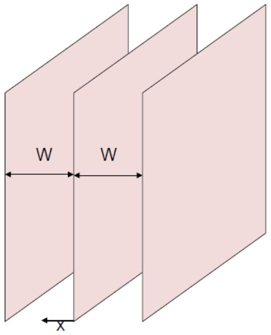




With an object in proximity C increase
→ oscillation frequency decreases

Differential displacement sensors

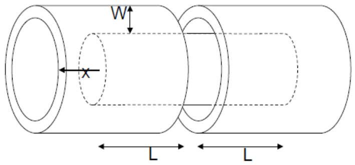
To cope with the issue of the dependency on epsilon-r.



$C_1 = \varepsilon_0 \cdot \frac{A}{W-x}$ $C_2 = \varepsilon_0 \cdot \frac{A}{W+x}$

With Wheatstone bridge:

$$V_{out} = V_A \cdot \frac{x}{2W}$$



$C_1 = \frac{\varepsilon_0 \pi D}{W} \cdot (L+x)$ $C_2 = \frac{\varepsilon_0 \pi D}{W} \cdot (L-x)$

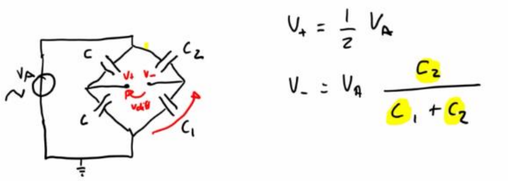
With Wheatstone bridge:

$$V_{out} = V_A \cdot \frac{x}{2L}$$

Parallel plate case

We have two capacitors in series with one plate in common; the central plate is the movable one. We can now compute the value of the two capacitances, supposing we have void in between the plates of the capacitors. To readout this circuit we use a Wheatstone bridge.

We have a Vsource that must be a sinusoidal source, otherwise in DC capacitor is a very high impedance, while we want a finite impedance. Then on one branch of the WB we place constant capacitors, and on the other branch the two variable capacitors. Then we will readout the differential voltage.
Instead of C2 and C1 we place the two equations depending on x and we obtain the final equation.



$$V_+ = \frac{1}{2} V_A$$

$$V_- = V_A \frac{C_2}{C_1 + C_2}$$

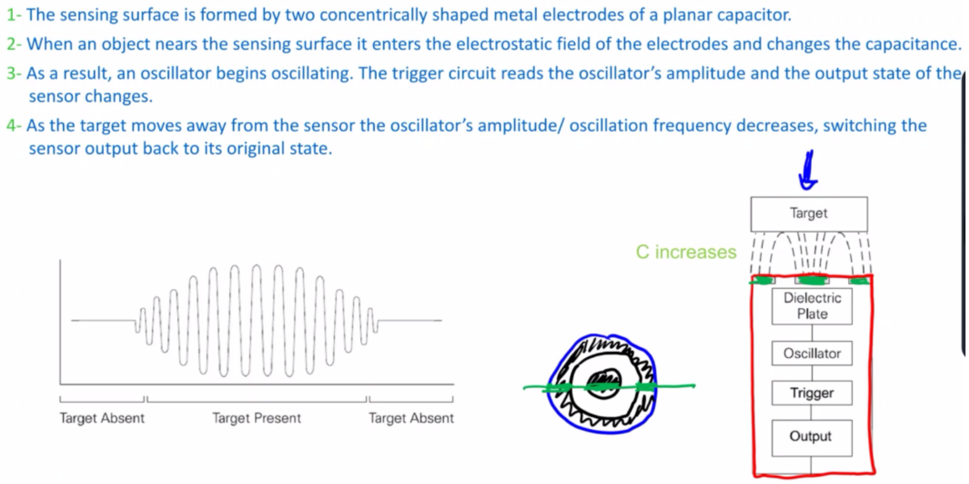
With this differential configuration we have an output directly proportional on x and we remove the dependency on epsilon-r.

Cylindrical geometry

We have two external fixed plates, and the inner cylinder can move within the other two cylinders. Also in this case we can consider to have two capacitors that are in series. We can compute the capacitance value considering the system in void. Also in this case the two capacitors can be connected within a Wheatstone bridge.

Proximity capacitive sensors

In our sensor we have a concentrical capacitor, and in the cylinder we have two plates; on is the external plate and the other is the internal plate. With this sensor we will generate an electric field with an oscillator, and the electric field will be influenced by what we have in proximity of the detector.



1- The sensing surface is formed by two concentrically shaped metal electrodes of a planar capacitor.

2- When an object nears the sensing surface it enters the electrostatic field of the electrodes and changes the capacitance.

3- As a result, an oscillator begins oscillating. The trigger circuit reads the oscillator's amplitude and the output state of the sensor changes.

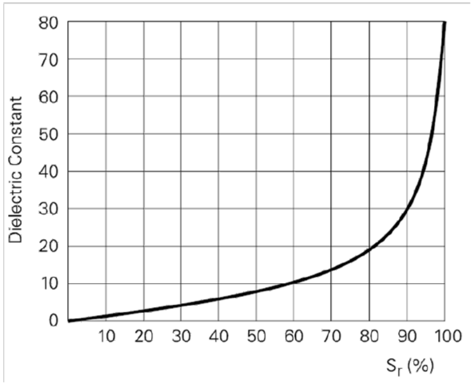
4- As the target moves away from the sensor the oscillator's amplitude/ oscillation frequency decreases, switching the sensor output back to its original state.

In particular, if we have no objects, the epsilon is the one of the air, while if we have an object, we see the epsilon of the object. Typically what happens is that if we don't have the target the oscillator is not able to oscillate, while if we have a target in proximity we start to have oscillations. This because oscillating circuits depend on the value of capacitance and resistance and so oscillation can happen or not depending on the damping.

Different materials have different dielectric constants epsilon-r, very close to the vacuum or distant (like water, where it is 80, it is the higher one).

If we look at the datasheet of proximity sensors, we see typically the graph in the next page. It represents the distance at which we can sense a material but depending on the epsilon of the material itself. For instance, since water has a high epsilon, it can be sensed even if far away, while Teflon requires to be very close to the sensor.

So typically, in the datasheet of proximity sensor the distance at which we find an object is for instance 10mm, but it is true for water, that has the higher dielectric constant. Then we have a plot that shows how the distance changes with respect to water distance depending on the dielectric constant. Since for water Sr is 100%, we have indeed 10mm. If we look for alcohol (epsilon-r = 26), Sr = 88%, so we will see it at 8.8mm → it must be closer to water to sense it.



In capacitive sensors the sensing distance is rated for water (ε_r=80).

E.g., capacitive sensor rated for 10 mm, with alcohol (ε_r=25.8) the effective sensing distance is 8.5 mm.

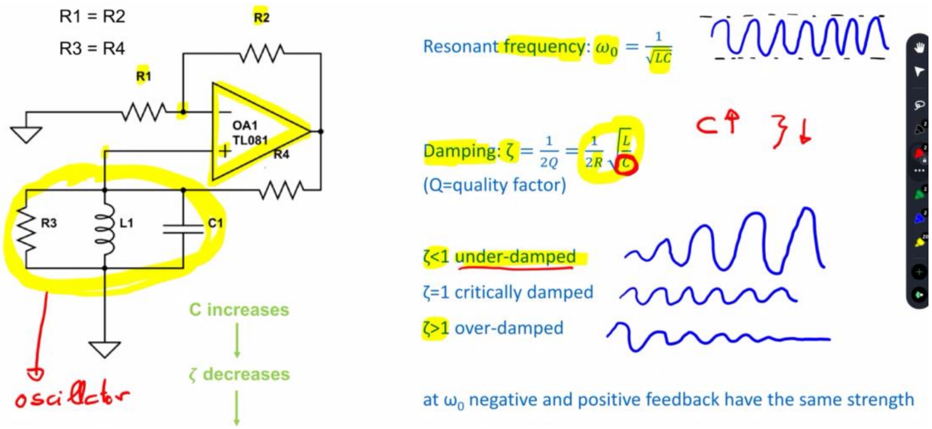
Sinusoidal oscillator

We want to implement the oscillator for capacitor proximity sensors. At a glance, we can see we have an OpAmp with both a negative feedback through R2 and R2, but through R4 and the RLC parallel we have a positive feedback. This configuration with both feedbacks is typical for all the oscillators (also in the oscillators seen in the previous square wave case).
The positive feedback sustains the oscillations, while the negative one prevents the oscillations to reach saturation. We want the oscillations to continuously go with the same amplitude.

Looking more in detail to the RLC parallel, this is an oscillator, so it is a circuit that oscillates mainly because we have some energy in the capacitance and inductance that will start to go from the capacitor to the inductor and viceversa. Then we also have the resistance R3 that is the dispersive element that consumes power due to the Joule effect, so we need the OpAmp to continuously provide the energy to the system that is lost due to the resistor.

The resonant frequency is given by 1/sqrt(LC). However, another important parameter is the damping csi, that depends in R, L and C. it tells us if an oscillator will oscillate or not. If csi is smaller than 1, the circuit is underdamped, so the oscillator will continuously increase. If equal to 1 it is critically damped, so the oscillation is stable. If higher than one the circuit is overdamped, so oscillations will die.

If we modify C, we can also modify the possibility of the oscillator to oscillate or not, because we change the damping. If C increases, since it is at the denominator, csi decreases, so the system can become underdamped, and starting to oscillate. Hence if we put a target in proximity of our sensor, we increase for sure the epsilon, because epsilon-r is always greater than 1, and so we decrease the damping and the system starts to oscillate.

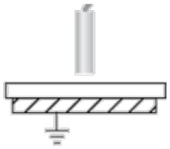

R1 = R2
R3 = R4
R2
R1
OA1
TL081
R4
R3
L1
C1
oscillator
C increases
ζ decreases
Resonant frequency: $\omega_0 = \frac{1}{\sqrt{LC}}$
Damping: $\zeta = \frac{1}{2Q} = \frac{1}{2R}\sqrt{\frac{L}{C}}$
(Q=quality factor)
C↑ ζ↓
ζ<1 under-damped
ζ=1 critically damped
ζ>1 over-damped
at ω_0 negative and positive feedback have the same strength

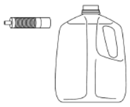
In the end, in order to keep this oscillations stable, so not diverging even if we are underdamped, we use the negative feedback, whose strength must be the same of the positive one to prevent increase of the oscillations. To have so, R1 = R2 and R3 = R4.
It can be demonstrated that at w0, so the resonance frequency, the impedance of the RLC parallel is equal to R3. So in order to make the positive feedback equal to the positive one it is sufficient to have R1/R3 = R2/R4.

In pSpice, to model the target I can use a capacitor in parallel to C1 with some switches so that they can be closed or open. We have both the switches close together in between 2ms and 5ms, the rest of the time they are open. In the results, before 2ms we see only C1, so we don't have oscillations, then between 2 and 5ms we have C2 in parallel so I increase the capacitance and we start to have oscillations.
In reality, we don't have exactly the two ratio equals to have a system not oscillating without the target, R3 should slightly be higher because we have that the OpAmp is not ideal.

Applications

- Position Measurement
  (Automation requiring precise location, Precision stage positioning)
- Dynamic Motion
  (Vibration measurements)
- Nonconductive Thickness
  (Label counting, Sensing water-based fluids applied to materials)
- Assembly testing
  (Capacitive sensors have a much higher sensitivity to conductors than to nonconductors. Therefore, they can be used to detect the presence/absence of metallic subassemblies in completed assemblies.)
- Detection through barriers





INDUCTIVE SENSORS

They can be used both to measure linear displacement (LVDT) and proximity, but for proximity the capacitive sensing is preferred, because with inductive sensors we can measure the proximity only of conductive objects.

Electromagnetic concepts

Faraday-Neumann-Lenz law:

$$V_i = -n \cdot \frac{d\Phi\left(\vec{B}\right)}{dt}$$

$$\Phi\left(\vec{B}\right) = \vec{B} \cdot \vec{A} = \mathrm{BA} \cdot \cos\alpha$$

With $\vec{B}$ normal to A:

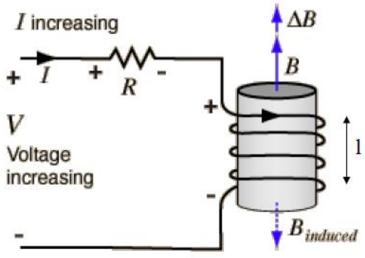
$$V_i = -n \cdot A \cdot \frac{dB}{dt}$$

For a solenoid:

$$B = \mu_0 \mu_r \cdot \frac{n}{l} \cdot I$$

thus:

$$V_i = -\frac{\mu_0 \mu_r \cdot n^2 \cdot A}{l} \cdot \frac{dI}{dt} = -L \cdot \frac{dI}{dt}$$



Self-inductance: $L = \frac{\mu_0 \mu_r \cdot n^2 \cdot A}{l}$

For the Faraday-Neuman-Lens law, if we have a variable flux of magnetic field, we induce a voltage across the solenoid (composed by n coils). The minus refers to the fact that the voltage induced is opposed to the flux of the magnetic field.

Then a second equation tells us that if we have a current tin a solenoid, the solenoid will generate a magnetic field. Then if we put together the two equations, we can see that if we have a current in the solenoid that is not constant but e.g. sinusoidal, the current will generate a magnetic field, but the B will concatenate with the solenoid itself and so it will generate a voltage.

In the end V = -L*dI/dt, where L is the **self-inductance**.

A similar phenomenon occurs also if we have two coils one close the other. Let's imagine to have a current in one of the two solenoid → we will generate a magnetic field that can concatenate with the second coil (of courses with some losses, not all the B is concatenated). So we will have a parameter K<1 that tells us the percentage of B1 that concatenates with the second coil, so to get B2. If I1 is variable, B1 and B2 are variable magnetic fields, so they will generate an induced voltage.

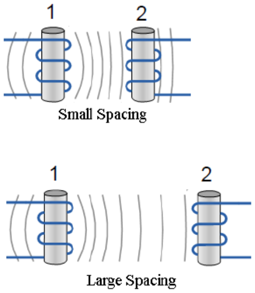
I in (1) induce $\vec{B}$ in (2):

$$B_2 = k \cdot B_1 = k \cdot \mu_0 \mu_r \cdot \frac{n_1}{l} \cdot I_1$$

thus:

$$V_{i2} = -n_2 \cdot A \cdot \frac{dB_2}{dt} = -\frac{k \cdot \mu_0 \mu_r \cdot n_1 n_2 \cdot A}{l} \cdot \frac{dI_1}{dt}$$

$$V_{i2} = -M \cdot \frac{dI_1}{dt}$$



Small Spacing

Large Spacing

$l_1 = l_2 = l$

$A_1 = A_2 = A$

Mutual-inductance: $M = \frac{k \cdot \mu_0 \mu_r \cdot n_1 n_2 \cdot A}{l} = k\sqrt{L_1 L_2}$

k = coupling coefficient (0<k<1)

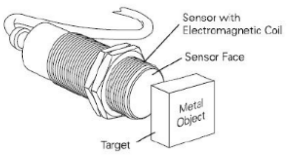
Using the previous equations, we have now to distinguish B1 and B2, and we get a formula depending on M, a constant term that is the **mutual inductance**.

Working principle

In general, inductive sensors exploit variations of the inductance L or of the mutual inductance M. If we exploit M, we in reality exploit variations in the K coefficient. The damping can be provided by some losses that we will have.

In this case <u>we can only measure the presence or displacement of metal objects</u>.

Inductive sensors produce an electromagnetic field, the inductance is modified by the presence/absence or the displacement of a metallic object at small distance from the sensor.



Inductance variations:

- Changes in coupling coefficient
- Dumping

Linear Variable Differential Transformer (LVDT)

It is a sensor used to measure the displacement of an object. We have a primary coil that is the coil A and a secondary coil B that is separated in two coils. We excite the primary coil with a variable current and so it will generate a variable primary magnetic field, and it will concatenate with the two coils of B, but the coupling factor K on the two coils od B will be different depending on the displacement we have of the inner material.

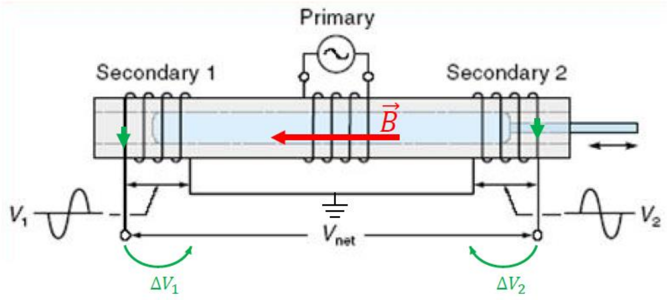
We must have a ferromagnetic core that moves inside, because ferromagnetic materials can create a low impedance path for the magnetic field.

Core in its center position

→ secondary coils are equally coupled

Core displaced from its center position

→ one of the secondary coils is more strongly coupled



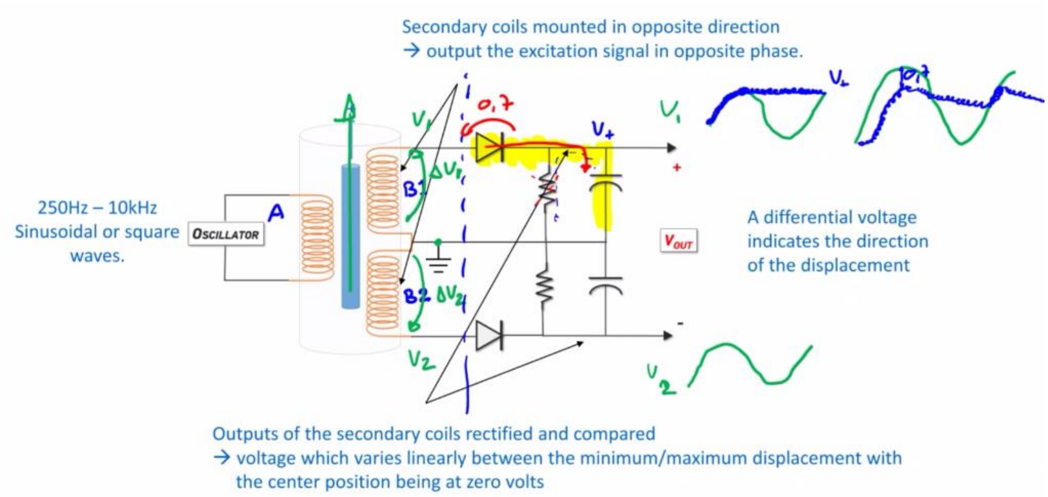
So we have our primary coil wired around the ferromagnetic material that can move horizontally. The primary coil has an alternate current that generates and alternate magnetic field. Then this magnetic field, in the resting position with the ferromagnetic material in the center, will concatenate in the same way with the two coils of B. The two secondary coils are wounded in opposite directions, because we want to create an induced voltage in opposite directions on the two coils.
In this way, we typically use just two terminals, one of B1 and the other of B2 to readout the external voltage. At rest, we will read 0V at the output. If we create

some difference in the coupling, we will have a higher or smaller voltage with respect to the resting situation.

LVDT readout circuit

We have the A coil connected to an oscillator to generate an alternate current, and then we have the two coils B1 and B2.

The other part of the circuit is a rectifier, a peak stretcher. V1 will be a sinusoidal, and V2 is perfectly in phase with V2. If we use a rectifier, what we see is that V+ follows V1 until V1 increases, but as soon as V1 decreases, since we cannot decrease the voltage on the capacitor because we cannot have a reverse current in the diode, the voltage is kept constant on the capacitor. So we are starting from and AC voltage and getting a DC voltage in output proportional to the amplitude.



Secondary coils mounted in opposite direction
→ output the excitation signal in opposite phase.

250Hz – 10kHz
Sinusoidal or square waves.

A differential voltage indicates the direction of the displacement

Outputs of the secondary coils rectified and compared
→ voltage which varies linearly between the minimum/maximum displacement with the center position being at zero volts

However, we have some non-idealities:

- **Diode** that has some losses, because it is not ideal → V+ will be always 0.7V smaller than V1.
- **Resistor**, where the capacitor can discharge, so we have a slight exponential discharge. However, the resistance is needed because if the object moves, the amplitude of V1 can increase or decrease, so we cannot have something insensible to this variation → we want something that discharges C. Hence R must be high enough not to loose to much between to cycles, but not too high to limit the response of the sensor to variations. The same reasoning can be done on V2.

In output we have V+ and V- that correspond to the amplitude of the sinusoids. If we are at rest, the output is 0, while if the object is in the top, V+>V- and we will see in output an almost DC voltage proportional to the displacement of the object.

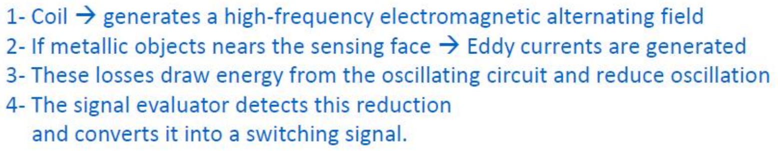
Inductive proximity sensor

With inductive sensors it is the opposite of capacitive counterpart. Normally without any target the oscillator oscillates. Then if an object arrives in proximity of the detector the oscillation stops.

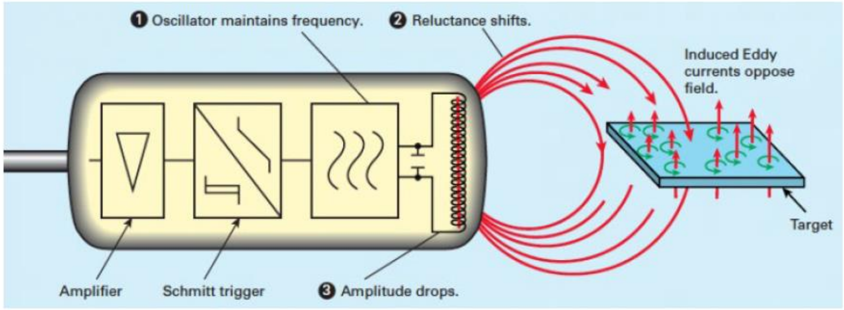
The physical principle is the presence of eddy currents in metallic materials. They are induced currents that can be generated in **conductive materials** when exposed to a magnetic field. They are typically due to the Faraday-Neumann-Lens; indeed, with a variable flux of B we induce a voltage in a material and hence also a current.

So if we have a conductive material that experiences a variable flux of B, also in this case microcurrents will be generated and they will oppose to the variation of the flux of B.
Now, if we imagine to have an oscillation circuit that provides an AC voltage to an inductor. The inductor crossed by a current generates a magnetic field, so since the voltage is variable also the B generated will be variable.
The magnetic field can be for instance increasing if we increase the voltage, and we will generate an induced current to generate a magnetic field that is induced in the opposite direction that opposes to the increase of the external field.

Then, if we have eddy currents, we will have also power dissipation, and this power most come from somewhere, and typically is stored in the oscillator itself → the oscillator experiences a loss of power due to the eddy currents.
Hence eddy currents can be modelled with a parallel resistor that steals power form the oscillator. Hence the system that was oscillating will stop to oscillate, it has no power enough to oscillate.



1- Coil → generates a high-frequency electromagnetic alternating field
2- If metallic objects nears the sensing face → Eddy currents are generated
3- These losses draw energy from the oscillating circuit and reduce oscillation
4- The signal evaluator detects this reduction
and converts it into a switching signal.


❶ Oscillator maintains frequency.
❷ Reluctance shifts.
Induced Eddy currents oppose field.
Target
Amplifier
Schmitt trigger
❸ Amplitude drops.

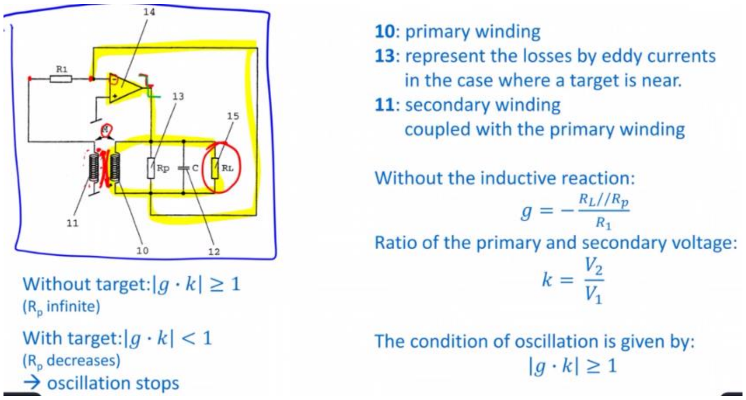
In a typical oscillator like the one in the capacitive case, the elements that were causing losses were the resistances. So also in this case we can model the presence of an object with a resistor, that steals current and dissipates power.

Oscillator schematic

It is a real circuit that can be really implemented. First of all, we have an OpAmp which has a negative feedback through 10, 13, 12, 15. We also have a positive feedback because the have the transformer M that is coupled with opposite directions. Indeed, the dots indicate the direction of the voltage. If we imagine that we are decreasing the output of the OpAmp, we will have a voltage drop across the inductor that indicates that the output is decreasing. This means that we are providing a positive input to the inverting input of the OpAmp. If we do so at the inverting node, we further decrease the output.

Hence we have both a positive and negative feedback. Every time we want to design an oscillator we need to have both.
In order to oscillate, the condition |g*k|>=1 must be respected.

g is the gain of the amplifier without considering the positive feedback, while k is the coupling factor between the two inductors of the transformers, so the voltage across the secondary coil divided by the one in the primary coil. Since V2 < V1 always, 0 < k < 1 always.



10: primary winding
13: represent the losses by eddy currents in the case where a target is near.
11: secondary winding coupled with the primary winding

Without the inductive reaction:

$$g = -\frac{R_L//R_p}{R_1}$$

Ratio of the primary and secondary voltage:

$$k = \frac{V_2}{V_1}$$

The condition of oscillation is given by:

$$|g \cdot k| \geq 1$$

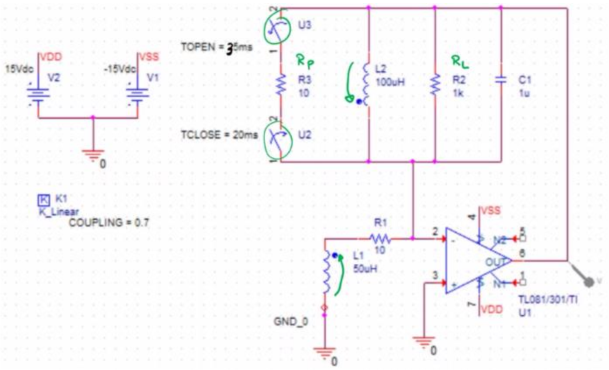
Without target: $|g \cdot k| \geq 1$
(R_p infinite)

With target: $|g \cdot k| < 1$
(R_p decreases)
→ oscillation stops

|g*k| expresses the strength of the positive feedback, because the higher the coupling between V1 and V2, the higher is the voltage I provide at the input and higher the gain the lower the voltage I will provide to the output. If we imagine to have a 0 gain, even if we have a very high voltage in V2 the voltage won't arrive to the output, so the overall positive feedback won't be effective. So for the positive feedback is important to have both coupling and a gain so that the input influences the output. If we have g*k high enough, then we have a strong enough positive feedback and so we can sustain the oscillations of the circuit.
Indeed, in an oscillator, both positive and negative feedbacks are important because the positive one will be unstable and provide the oscillation that increase too much if the negative feedback is too low, and that are too damped if the negative feedback is too high.

Rl is mainly the load resistance of the circuit, so it is a physical resistor we put in the oscillator. Then Rp represents the losses due to the eddy currents. If we consider Rp, with no object in proximity it is infinite, while with an object close to the detector it tends to a finite value. If Rp decreases, we decrease the Rp||Rl, so we decrease g. If then we decrease g we decease g*k, so at a certain point with an object in proximity |g*k| < 1, so the oscillations will stop.

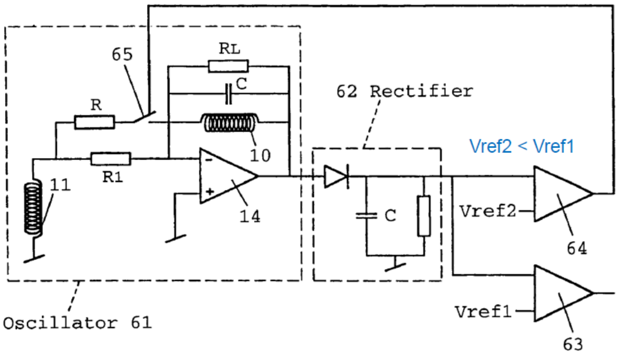
We can simulate this circuit in pSpice. If we suppose Rp to be infinite, the two switches are like open, so we have a k = 0.7 and |g| = 1k/10 = 100 → g*k = 70 >1 → I expect the circuit to oscillate.



Then the switches are closed at 20ms and Rp is connected. In this case g = Rp || R_L = 1 and hence g*k = 0.7 < 1 → circuit stops to oscillate.

Then to restart the oscillation we need some time because the circuit needs to regenerate its power, it is not instantaneous.

If we want to reduce the delay to restart the oscillations, we can use the configuration below. The basic circuit is the one already seen, with 10 coupled with 11 and R_1 is the input resistance. Then we add the parallel resistor R and the rectifiers and the two comparators.
The rectifier is used to have a DC voltage proportional to the amplitude of the oscillations, via an RC discharge and a diode to catch the envelope. The output of the rectifier is a DC rectified voltage.
Then we can compare this DC voltage with a reference voltage with normal comparators. If we focus on 63, it compares the Vrect with the reference voltage Vref1. If Vrect > Vref1, then the output will be high → no object in proximity. Then if Vrect < Vref1, the output of the comparator will go low, so it means that there will be an object in proximity.



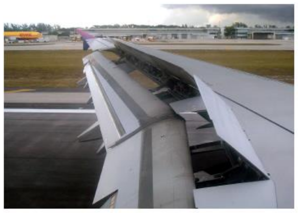
Comparator **64** closes switch **65**, increasing the gain of amplifier **14**, thus restarting the oscillation.

The second comparator is made with a lower threshold, Vref2 < Vref1. When we go below even this second threshold, the circuit will close the switch 65, meaning that to the input resistance R1 we put the resistance R → we are trying to increase the factor g, given by the feedback resistor divided by the input resistor, that will become from simply R1 to R | | R1 → we are decreasing the input impedance so increasing g. This is useful because if we have an object in proximity, the feedback is still strong enough to prevent oscillation, but also, as soon as the Vrect is smaller than the second threshold, we are faster in restarting the oscillations, because we are boosting the g, g is strong enough to start the oscillations faster, because we restart with a big g.

In this case we have hence a second amplifier Vref2 that activates a switch in order to add a resistor that increase the gain g to make easier the restart of the oscillations when we remove the object.

Applications

LVDT:
resolution down to 1 mm
→ Industrial, military, and aerospace applications
- aircraft wing flaps
- off-axis rotational movement of wheels



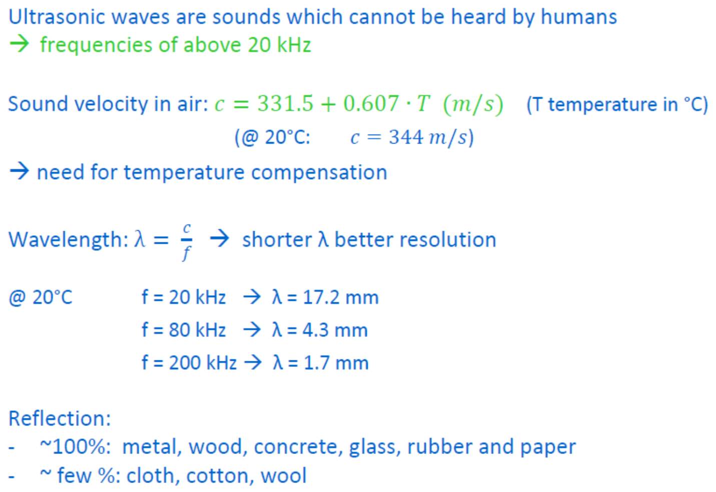
Proximity:
Safety and warning systems
- parking sensors
- ground proximity warning system

ACUSTIC SENSORS

They can be used both to measure proximity and distance. They are mainly made of two pieces of piezoelectric material, that can be used both as a sensor or as an actuator. In this case we us one of the piezoelectric material as an actuator, so we provide an AC voltage to it to start the oscillations (ultrasounds in acoustic sensors); then the generated sound interacts with the object, it is backscattered and sensed by another piezo used as a sensor.

Ultrasonic waves basics

Ultrasounds are sounds with a frequency above 20kHz. The sound velocity in air is, at 20°C, about of 344 m/s, but the speed of the sound depends on temperature in the general equation → some sensors will need for temperature compensation to account for the variation of the velocity of the sound. Another important parameter is the wavelength of the sound. In general, the shorter the lambda, the better the resolution because, as a rule of thumb, we can detect an object only if the object is larger or in the same order of magnitude of the wavelength.



Ultrasonic waves are sounds which cannot be heard by humans

→ frequencies of above 20 kHz

Sound velocity in air: $c = 331.5 + 0.607 \cdot T \ (m/s)$ (T temperature in °C)

(@ 20°C: $c = 344 \ m/s$)

→ need for temperature compensation

Wavelength: $\lambda = \frac{c}{f}$ → shorter λ better resolution

@ 20°C f = 20 kHz → λ = 17.2 mm

f = 80 kHz → λ = 4.3 mm

f = 200 kHz → λ = 1.7 mm

Reflection:

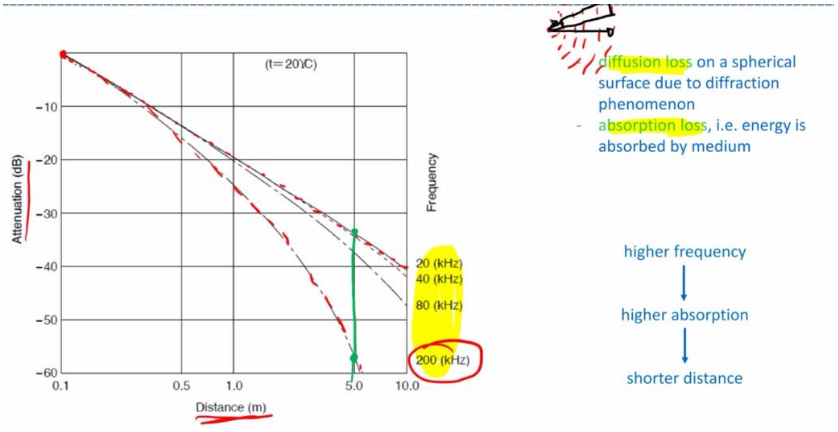
- ~100%: metal, wood, concrete, glass, rubber and paper
- ~ few %: cloth, cotton, wool

Let's consider now the c = 344 m/s. If we use a 20kHz frequency, the lambda = 17mm → we can detect an object with a size more or less of 20mm, so that the probability of interaction with the ultrasound wave is high enough. If the object is of 1mm, the frequency is not high enough to detect the presence of the object.

Not only the size is important, but also the kind of material. For instance, if we have metal, wood, glass etc. we reflect about 100% of the energy of the incoming sound. With other materials, they will absorb the sound, so the reflected energy is small compared to the incoming one. Hence the capability of detecting an object depends on the size of the object and on the material.

So increasing the frequency we improve the resolution. Hence why not going to extreme frequencies?

We have an upper limitation given by the attenuation we have in air. The graph represents the attenuation of air depending on the distance. So we have the source that generates the sound. If we are close we have almost no attenuations, then the intensity of the sound decreases far away from the source. But attenuation depends also on the frequency of the sound. With 20kHz we have an almost linear attenuation, while with 200kHz it is non-linear.
This happens because the attenuation in air depends on two factors:



- diffusion loss on a spherical surface due to diffraction phenomenon
- absorption loss, i.e. energy is absorbed by medium

higher frequency ↓ higher absorption ↓ shorter distance

- Diffusion loss: it can be called also geometrical attenuation. If we imagine to have the source of the sound, the sound propagates with 'spherical waves'. If we are close to the source, all the energy is concentrated in a small diameter, while far from the source the sphere in which the sound is diffused is larger, so the energy per unit angle is smaller. **This phenomenon doesn't depend on the wavelength**.
- Absorption loss: it is a source of losses caused by absorption because the sound propagates in a medium and it can be absorbed by the medium, and the higher the frequency the higher the probability for the sound to be absorbed, because very small particles are enough to absorb the sound.

So if we want to have high resolution we have to work with high frequencies, but we cannot go too high, we must be close to the source of sound. To detect far away, we can use lower frequencies but we can detect only bigger objects → trade off.

SPL and Sensitivity

Sound pressure level (SPL) = volume of sound, expressed by:

$$SPL = 20 \cdot Log\frac{P}{P_0}\ (dB)$$

P = sensor sound pressure
P_0= reference sound pressure (20µPa)

Sensitivity in linear scale (V/Pa) is expressed by:

$$S_{lin} = \frac{V_{out}}{P_{in}}\ \left({}^{V}/_{Pa}\right)$$

V_{out} = output voltage
P_{in} = input sound pressure

Typically sensitivity is expressed in dB:

$$S_{dB} = 20 \cdot Log\frac{S_{lin}}{S_{ref}}\ (dB)$$

with S_{ref} = 1V/Pa

We can use as a parameter the sound pressure level (SPL) that expresses the volume of the sound, so it is the sensor sound pressure divided by the reference pressure, expressed in dB.

Then we can also quantify the sensitivity of the detector, defined as the ration Vout/Pin, where Pin is the sound pressure in input. It can be expressed in linear scale or in dB with respect to a reference sensitivity of 1V/Pa. Hence if Sdb = 0, it means that S = 1V/Pa.

Open structure

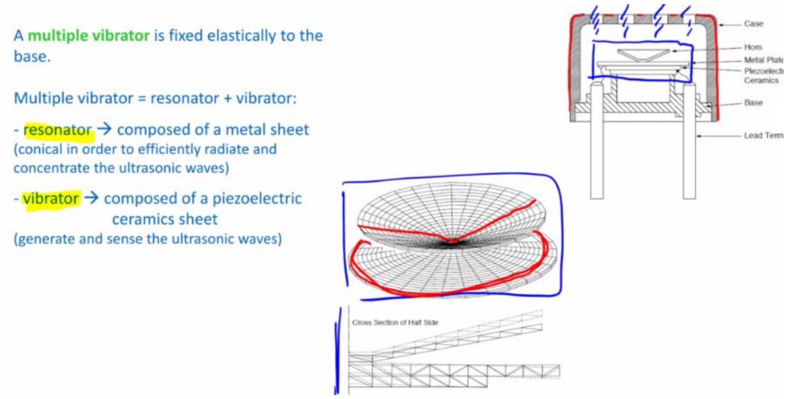
The question is: how is the sensor fabricated?

We can have different structures, starting from the open structure. In the open structure we have a case in which we put the sensor where we have some aperture where the sound can get inside. Then we have the sensitive part that is the blue one, where we have both a resonator and a vibrator.

- Resonator: it is the conical red part that is used to concentrate the sound on the vibrator, that is the bottom part.
- Vibrator: sensor itself, made out of piezoelectric material.

Then we have the contacts to readout the voltage.

The main limit of this structure is that it is not protected against dust, rain, water and since a typical application of this kind of sensors is in automotive like parking sensors, we cannot use this open structure, so other structures are preferred for outdoor applications.



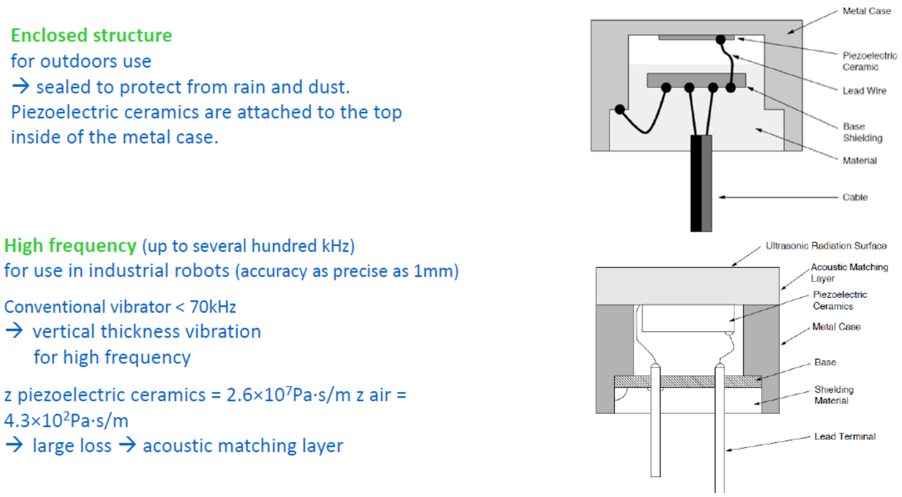
A multiple vibrator is fixed elastically to the base.

Multiple vibrator = resonator + vibrator:

- resonator → composed of a metal sheet (conical in order to efficiently radiate and concentrate the ultrasonic waves)

- vibrator → composed of a piezoelectric ceramics sheet (generate and sense the ultrasonic waves)

Enclosed structure



Enclosed structure
for outdoors use
→ sealed to protect from rain and dust.
Piezoelectric ceramics are attached to the top inside of the metal case.

High frequency (up to several hundred kHz)
for use in industrial robots (accuracy as precise as 1mm)

Conventional vibrator < 70kHz
→ vertical thickness vibration for high frequency

z piezoelectric ceramics = 2.6×10^7Pa·s/m z air = 4.3×10^2Pa·s/m
→ large loss → acoustic matching layer

We have no openings in the metal case and for this reason we put in vibration the entire metal case. Then the piezoelectric ceramic is directly connected to the metal case, so the vibration and pressure are transmitted directly from the metal case to the piezoelectric material.

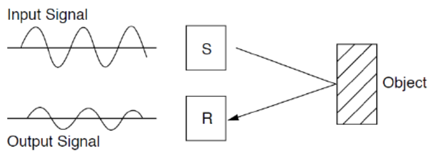
High frequency

Particular structure used if we want to measure high frequencies, because we need a sensor that is very sensitive because the pressure that reaches the sensor is very low due to the attenuation of air.
We simply add an acoustic matching layer that is like the antireflective coating on optical sensors. Let's imagine we have the sound that arrives on the sensor. The impedance of the air is very low, in the order of 10^2 Ps*s/m. This means that the sound can easily propagate in the air, while the impedance of a piezo material is in the order of 10^7 → the material will most probably absorb the material very easily, the sound will not easily propagate through it. This means that the sound is backscattered, because the sound prefers to stay in air.

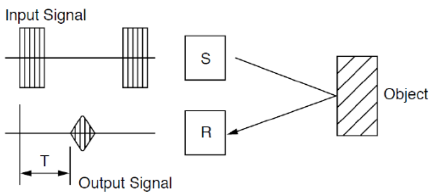
So if the sound propagating in air arrives in a material with high impedance, it will probably backscattered. If instead we put the acoustic matching layer, it is a layer with an intermediate impedance Z, so we make easier for the sound to get inside and then to reach also the piezoelectric ceramic. So the acoustic matching layer will be a material with an intermediate impedance → typically we use plastics, so soft materials.

Operation modes

We can operate acoustic sensors in continuous waves. In this case the source and the receiver are separated and the source will generate continuous waves. In this case we generate the ultrasounds and they will interact with the object and then they will backscattered. Depending on the distance of the object, the signal coming back will have a lower amplitude if the path is longer. The longer the path the sound has to cover, the smaller the amplitude of the signal we receive back.
This kind of detection is used for proximity sensor. If we have an object in proximity, the received amplitude is very similar to the sent one.



Proximity sensor
Detect the amplitude of the return signal

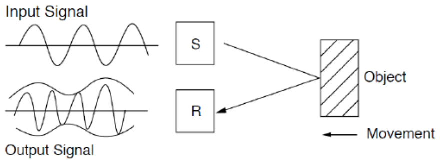


Distance meter
Measure the time of flight of the return signal

The other possibility is to use them as distance meters. We have the source that generates a pulse, the spike will propagate, interact with the object, backscattered and it will arrive to the receiver with a certain delay. In this case we can have source and receiver separated or together. Now the information is not about amplitude, but travel time. If we use them as a distance meter we need a circuit to measure the time T.

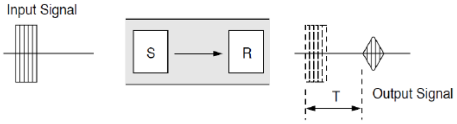
In other operational modes we can use them also as displacement sensor, so not to measure the actual distance but to measure the displacement of the object, so the variation in position with respect to a resting position, so for instance to check if an object in front of me is vibrating. In this case we can exploit the **Doppler effect**; we have the source that is generating an ultrasound at a fixed frequency. Then if we have an object moving with respect to the source, the backscattered sound will be affected by the Doppler effect, so the receiver will receive back a sound with a frequency that is changing depending on the variation of position of the object. If the object is going far away we are decreasing the frequency. So the modulation in frequency depends on the motion of the object. It is typically used to measure vibrations.

The last operational mode is the direct propagation time and exploits the changing in velocity depending on the medium.
Let's imagine having a medium between the source ad the receiver, and we want to monitory the density of the medium. We can generate a pulse from the source, like in the distance meter, and then measure the travel time for the pulse to reach the receiver. The travel time will depend on the medium itself. Of course, in this case also temperature will influence the speed of the sound, so we need to compensate for temperature variations.



Displacement sensor

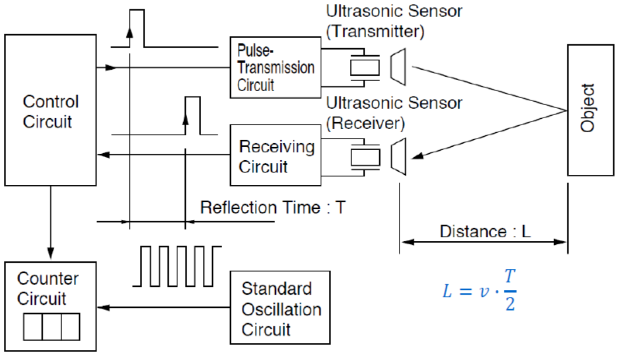
Exploits Doppler effect to measure the reciprocal movements between sensor and object



Direct propagation time

Utilizes the change of sound velocity to measure density of a medium

Distance measurement complete system


Ultrasonic Sensor (Transmitter)
Pulse-Transmission Circuit
Control Circuit
Ultrasonic Sensor (Receiver)
Receiving Circuit
Object
Reflection Time : T
Distance : L
Counter Circuit
Standard Oscillation Circuit

$$L = v \cdot \frac{T}{2}$$

TDC = Time-to-Digital Converter (10 ns → ~2µm)

In this case the system has two different piezoelectric material, one used as a transmitter and the other one used as a receiver.

The transmitter is connected to the pulse-transmission circuit that it the circuit that will provide the modulated voltage to the piezo to generate the sound. Then, once the sound spiking is generated, the sound travels to the object and it is backscattered to the receiver, and the receiver is connected to a receiving circuit that reads the voltage across the material. So we have two pulses, one that corresponds to the generation of the sound and the other one that corresponds to the detection of the sound. We want to measure the travel time between emission and detection.

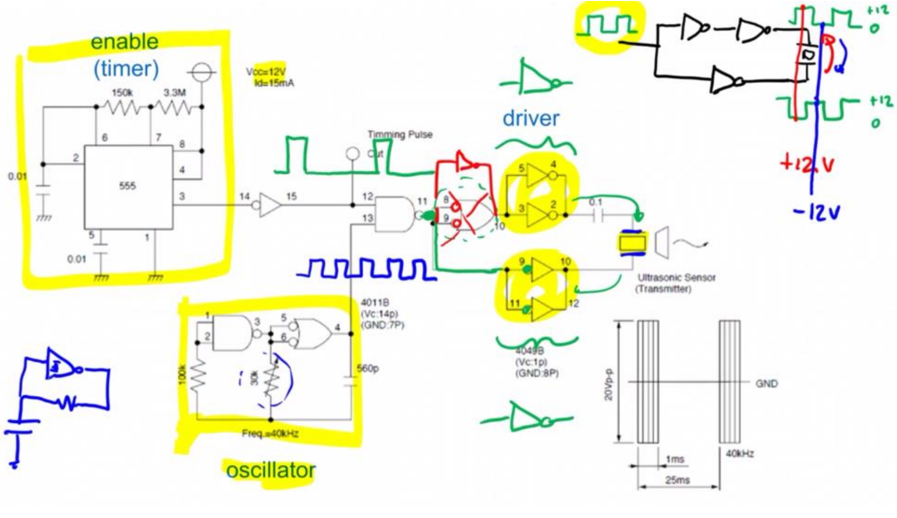
To measure it, we can use the counter of the microcontroller, counting the clock cycles between the start and stop signals. In fact, we don't need a very high time resolution, because the speed of the sound (340 m/s), a timing resolution of 10ns for 100MHz clock, it corresponds to 2um in the space resolution.

In the case of optical sensors, the timing resolution is much more critical because the speed of light is much greater than the speed of sound, so we cannot use a simple timer of a microcontroller.

Transmitting circuits

We have the piezoelectric material and the two electrodes to which we want to provide the oscillating voltage on the right hand. The voltage is provided as a square wave modulation; we have the two inverters 4049B in parallel to increase the maximal current we can provide to the load, but they can be considered as a normal unique inverter. They are the drivers, and provide the current to charge and discharge the capacitance of the piezo. Then the signal that arrives to the two inverters is not the same, because in one case we have as input a NOR with two inputs shortened. Also this is a normal inverter. So we have that in the lower branch the signal goes directly in the inverter, while in the upper branch it passes through two inverters → the voltage we have on the top is the opposite of the one we have in the bottom. Hence this system is used to double the voltage we have across the piezoelectric material. This also increases the intensity of the sound.

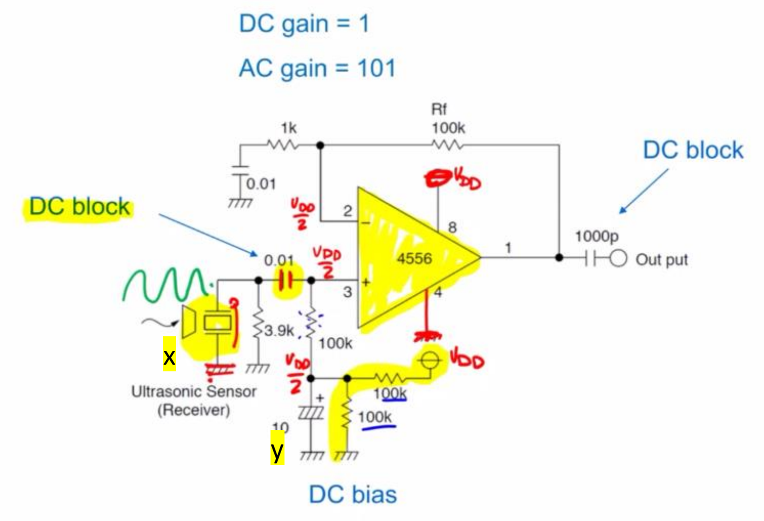
Then the other part of the circuit is used to generate the oscillating voltage. Hence we have an oscillator that is almost equivalent to the oscillator based on a R-C structure with a Schmidt trigger inverter. In this case the trigger is not used, so the hysteresis is obtained with the variable resistance.
Then the oscillator will generate constant oscillations, it will always oscillate, but in the distance meter we need to generate only a short pulse. So we also need a circuit, the enable, that is able to generate



and enable window; typically it is a simple timer. Then we use an AND in which we put the output of the oscillator and the timer output.

Receiving circuits

We have the receiver at the bottom x that generates a voltage that depends on the ultrasound amplitude. Then we have an amplifier to amplify the signal and send it to the output. The capacitor at the input represent DC block to filter out the continuous voltage. The frequency of the voltage will be the same of the ultrasound, above the 20kHz. Then we have a circuit to provide the CD bias, so that the amplifier can be used both with positive and negative input. The power supply of the amplifier is a single power supply, we have a positive voltage but not a negative. Since the input can be either positive or negative, we need a circuit that puts at the steady state the voltage in input to the amplifier in the middle of its dynamic range → the voltage divider provides us a signal reported to the + input. In DC the capacitor is like an open circuit, so the 100k resistor cannot be considered. In DC the output is Vdd/2, in the middle of the dynamic. Then if we add a positive input it can increase up to Vdd, with a negative one down to 0. The capacitor is used in DC bias to keep the node y as much stable as possible at Vdd/2. The capacitor of 0.01 instead allows us to have different gains between the DC and AC, and in AC the gain is given by the non-inverting configuration, so 1 + Rf/Rk, and so 101.

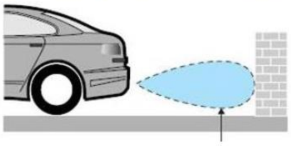


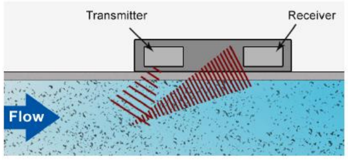
In output then we have also the DC block to see only the variations and not the Vdd/2 in output. So in DC we see 0V in output. So the variations in output will be centered around Vdd/2 before the output capacitor, while the oscillations will be around 0 after the 1000p capacitor. This DC block is not mandatory, it depends on what we have after.

Applications

We can use them as proximity sensors, in parking meters to check if the slots is free or not. Then they can be used as distance meter, as back sonars for car → not only to measure proximity but also the actual distance to give an idea.

- Proximity sensor:
  - Counting instruments
  - Access switches
  - Parking meters
- Distance meter
  - Automatic doors
  - Level gauges
  - Back sonars of automobiles
- Displacement sensor
  - Intruder alarm systems
  - Flowmeters
- Direct propagation time
  - Densitometers
  - Flowmeters





OPTICAL SENSORS

The difference with respect to the sound sensors is the speed of the signal, while the working principle is the same. They can be used to measure proximity or distance or in displacement encoder.

Optical proximity sensors

They can be used mainly in 2 configurations, **transmission** or **reflection**.

Transmission

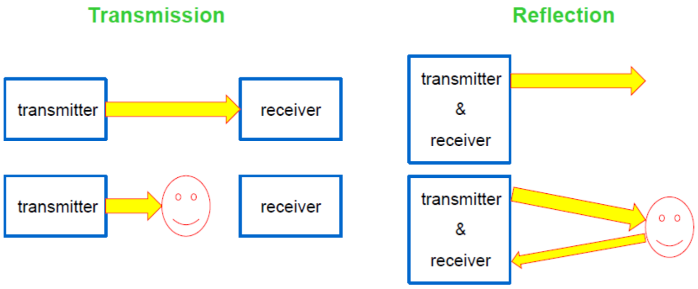
We have in two different positions a transmitter and a receiver in opposite position. With no object in between the receiver will receive the light emitted by the transmitter. If we have an object in the middle the flux of light will be stopped → we can detect the obstacle in between the transmitter and the receiver. It is a simple setup but sometimes not very practical, because we need to have the obstacle exactly in between the transmitter and the receiver, so we can use it in a very controlled environment, while in automotive cannot be used for sensing obstacles, so reflection is preferred.

Infrared **transmitter** (LED or laser) + **receiver** (photodiode):

- immunity to electro-magnetic field
- immunity to ambient light
- long range

Reflection

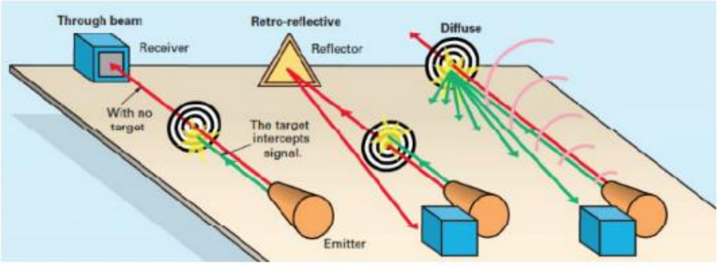
Transmitter and receiver are in the same position, if we have no obstacle no light comes back, while if we have the obstacle, then the light will be backscattered. This sensors are used in the smartphones to detect if we have the face close to the smartphone to switch off the screen. The problem is that the proximity sensor based on the amplitude of the signal strongly depends on the reflectivity of the target. If we have e.g. a black target, in principle it absorbs all the radiations, so we cannot detect it → this sensors are not so much reliable.



Sensing set-ups

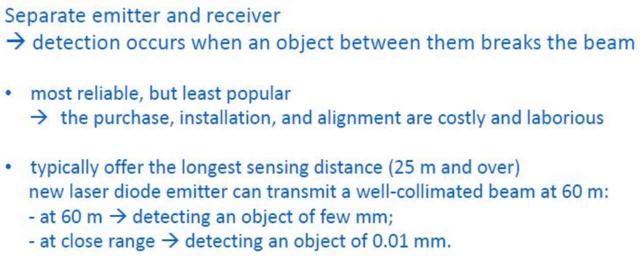
We have 3 possible setups:

1. Two beams: we have the emitter and the receiver on the opposite sides. It is the transmission setup.
2. Retro-reflective: we can put emitter and receiver on the same side but we need also a retroreflective element able to backscatter the light. It is a configuration in the middle between transmission and reflection. Like in the reflection setup source and receiver are on the same side, but we can detect light only if we have no target, like in the transmission.
3. Diffuse: it is the reflective configuration.



Typically, set-ups based on transmission can reach longer distances because light travels only one time the distance. With the retroreflective setup the distance is more or less 10m compared with the 25 meters of the transmission.
Moreover, with the diffuse setup we cannot provide a maximum distance because it strongly depends on the object itself, on how much the object is black or white. If it is white it can be further away.

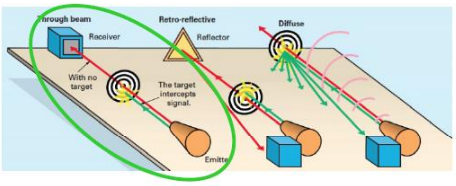
However, having a variable sensing distance is a common feature for proximity sensors, also in the case of capacitive sensors.



Separate emitter and receiver
→ detection occurs when an object between them breaks the beam

- most reliable, but least popular
  → the purchase, installation, and alignment are costly and laborious

- typically offer the longest sensing distance (25 m and over)
  new laser diode emitter can transmit a well-collimated beam at 60 m:
  - at 60 m → detecting an object of few mm;
  - at close range → detecting an object of 0.01 mm.

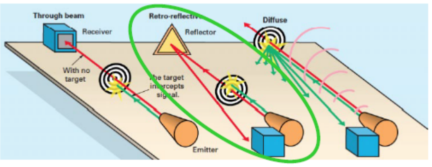


The emitter produces light beam towards a specially designed reflector, which then deflects the beam back to the receiver.
→ detection occurs when the light path is broken by an object

sensing distance up to 10 m

Emitter and receiver are both located in the same housing
→ **one wiring location**, the opposing side only requires reflector mounting.

Very shiny or reflective objects (like mirrors) sometimes reflect enough light to trick the receiver into thinking the beam was not interrupted, causing erroneous outputs.
→ polarization filtering (allows detection only from specially designed reflectors)

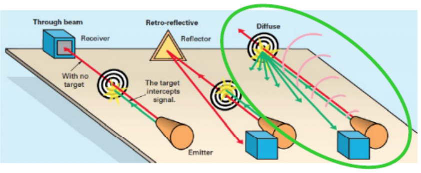


The emitter sends out a beam of light and the target diffuses it in all directions, and part of the backscattered beam reaches the receiver.
→ detection when light reflected from object.

A non-reflective target will have a significantly decreased sensing range as compared to a bright white target → color dependent.
Certain versions are suitable for distinguishing dark and light targets in applications that require sorting or quality control by contrast.

False triggers caused by reflective background → diffuse sensors with focus:

- 2 receivers focused on background and target (compares the intensities)
- Triangulation



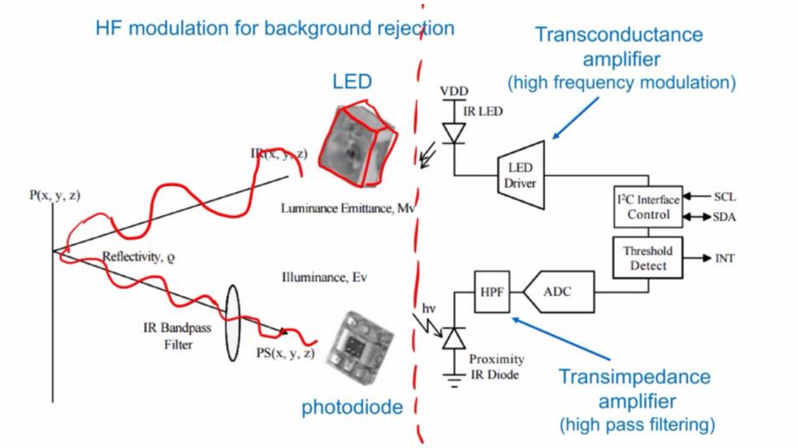
PROXIMITY SENSOR SYSTEMS

The problem is that we can have also background illumination, that is like noise. If we are in the case of reflection with no object in proximity, I expect to receive no signal, but this is not true because the background illumination (sun or lamps) can be detected by the receiver. So how can the receive distinguish where the light detected comes from?

It is thanks to the modulation of the light. Indeed, solar illumination is a continuous illumination with no particular pattern, while the LED that emits the light detected by the PD can emit modulated light (but not of 50 – 60 – 100 Hz, because it is the modulation for old lamps made out of neon). If we do so, we can demodulate the signal received by the PD and check if it is indeed the light emitted by the LED.

On the left side of the image we have the working principle; the LED emits modulated light (continuous, not pulsed) that, if there is the object in proximity, reaches it and it is backscattered and detected by the PD. Of course the amplitude will be smaller because the reflection won't be of 100%.

As for the right part of the image, we still have the LED, that is still a diode and so we need to have a LED driver. Typically the LED driver is a transconductance amplifier, that is an amplifier in which the input is a voltage and the output is a current → LED is driven with a fixed current and depending on the current the emitted intensity will be different. Typically the modulation is in the order of kHz. There exist LED driver with a digital interface so that we can directly program them or analog ones in which we have to provide a certain voltage. Typically in the case of the digital ones the communication is based on a I2C protocol.



Then we have a standard PD and a HP filtering that is used because we don't want to see the contribution due to the all the components in DC (e.g. solar components). Then after the HP filter we have the sinusoid only due to the signal and with the ADC we can convert the signal. Often, before the ACD we can put a peak stretch to obtain only the amplitude and so sample with a very slow rate of the ADC, not needing a fast ADC. Then we have a threshold and we check if the signal exceeds it. changing the threshold we will change the distance at which we will sense objects.

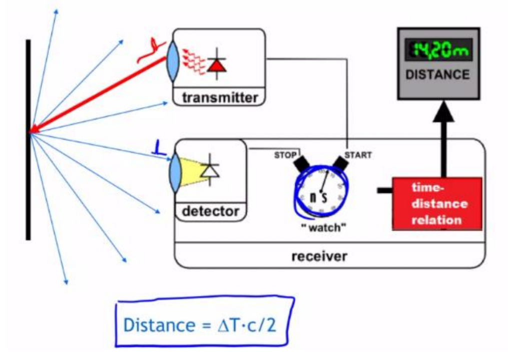
The main limitation is that the backscattered light strongly depends on the reflective index of the object.

To overcome this issue, we can use a system based on TOF. Instead of detecting the amplitude, we can measure the actual distance of the target measuring the time of flight of the emitted light. In this case it is important that the emitted light by the transmitter is a very short pulse of light.

Time of flight

In this setup doesn't matter the amplitude of the light, but the time difference between the emission and the detection → we need a circuit able to detect the time interval. It is the same principle of acoustic sensors with Rx and Tx and distance measurement.

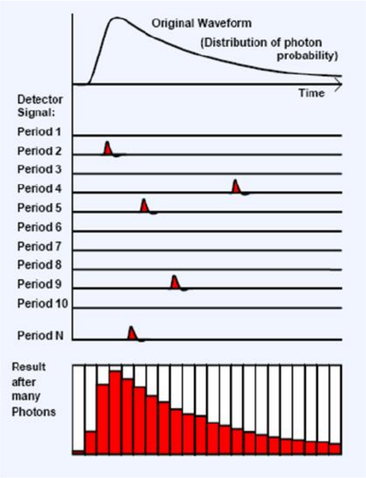
The difference is now the speed of the signal, because in optical sensors the equation is still the same but the velocity is 3*10^8 m/s, so we need a very better resolution.



Now it is not enough to use the timer of the uC, which has a resolution of 10ns circa, but we need a much better resolution. Indeed, 300ps translates in a 5cm resolution with light. So to have a very fine resolution TOF optical sensors are not indicated, but they can reach longer distances with respect to sound sensors, but ultrasound sensors are more precise.

For the transmitter implementation, typically we use LED or laser, because with the laser we can provide higher optical power and keep the light more focused on the target.
As for the detector we can use:

- **Linear detectors**: PD or better APD. In this case with linear sensors we get an output current whose amplitude and shape is proportional to the amplitude and shape of the detected light. The detected pulse corresponds to the optical pulse generated by the laser. In this case we need a much higher optical pulse.
- **Single photons detectors**: SPADS. Spads are similar to APDs because they have an internal multiplication but in this case with just one photon we generate a very high current but that is indistinguishable from the one generated by two or more photons. So if we use this sensors, we can reduce a lot the power intensity of the laser emitter because we don't need to reconstruct the entire shape like in the case of linear detectors, but getting just one photon is enough. The problem is that we cannot distinguish if the photon is due to LED light or background emission → we need to repeat the measurement several times and then build a histogram of the time arrivals. After some repetitions, we have the histogram that represent the probability of detecting a photon in each time instant, and it reflects the intensity of the emitted light. It can be demonstrated that if the signal emitted has the shape like the on in the image, also the histogram will have the same shape → **this technique is called time correlated single photon counting (TCSPC)**.



Linear sensor (photodiode):

- Single shot measurement
- High optical energy per pulse

Single-photon sensor (SPAD)

- Many repeated measurements
- Low optical energy per pulse

↓

Time Correlated
Single Photon Counting
(TCSPC)

So with linear sensors we do a single shot measurement but with a very high energy per pulse. In the case of single photons sensors we need to repeat many time the measurement but with lower optical energy per pulse, because we can detect only one photon. If we want a faster measurement, then probably linear sensors are better, but if we want to see target at very long distance but keeping the system safe, so not shining a very high energy laser, SPS are better because they can reach long distances but with low power.

LiDAR

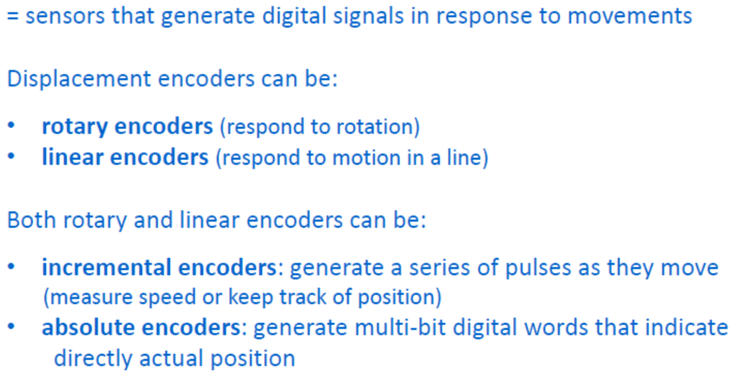
The technique of measuring distance with light is called LiDAR, light detection and ranging. It is possible to reconstruct the 3d shape of the object.

TOF sensors

Now they are often used in smartphones, especially for the autofocusing. We measure the distance of the object we are taking in the image and we regulate the focus of the camera. They are typically based on spads.

DISPLACEMENT ENCODERS

They are sensors that generate digital signals in response to movements. The output is not proportional to the distance of the object but to the movements. The displacement encoders can be **rotary** encoders or **linear** encoder if we want to respond to a linear motion.



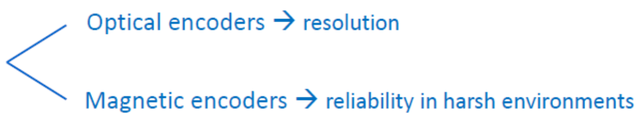
= sensors that generate digital signals in response to movements

Displacement encoders can be:

- **rotary encoders** (respond to rotation)
- **linear encoders** (respond to motion in a line)

Both rotary and linear encoders can be:

- **incremental encoders**: generate a series of pulses as they move (measure speed or keep track of position)
- **absolute encoders**: generate multi-bit digital words that indicate directly actual position



Optical encoders → resolution

Magnetic encoders → reliability in harsh environments

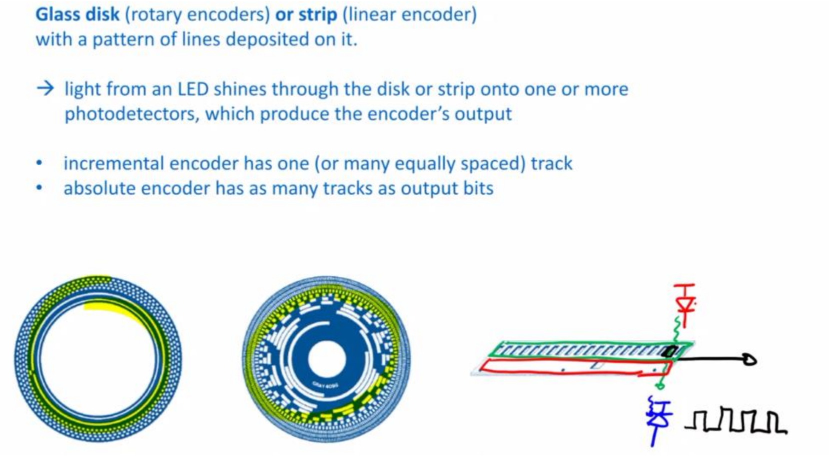
Both of them can be either **incremental** or **absolute** encoders. Incremental ones give us a pulse everytime the rotating wheel makes a rotation of a certain angle, but we don't know the precise position of the disc, while the absolute ones provide also the precise position.

The displacement encoders can be made with an optical working principle or with a magnetic sensing.

OPTICAL ENCODERS

Optical sensing can be done with glass discs or strips, used to measure rotation or linear displacement. If we consider the strip, for instance, we have a LED that emits light and then on the opposite side we have the PD which is able to detect the light. Mainly, every time that we move the strip, then the output signal of the PD will provide us an alternation of current and no current, depending on the position of the strip. Depending on the frequency of the output signal we can get the velocity with which we are moving the strip.

Sometimes we have also a secondary hole because we can have also a second LED focused on another part of the strip (red one). We have also a secondary PD, and the signal will be 0 until we reach the hole, and this gives us an idea regarding the absolute position.



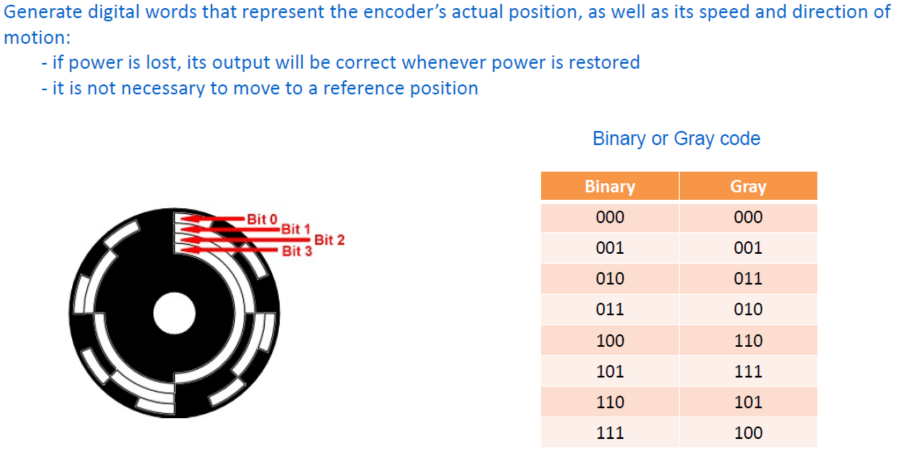
Glass disk (rotary encoders) **or strip** (linear encoder)
with a pattern of lines deposited on it.

→ light from an LED shines through the disk or strip onto one or more photodetectors, which produce the encoder’s output

- incremental encoder has one (or many equally spaced) track
- absolute encoder has as many tracks as output bits

The same can be done with rotary encoders. So all the other holes provide an information regarding the velocity, the other region the absolute position.

The one in the middle is an absolute rotary encoder, and there are some configurations of holes that provide us in a binary way an encoding corresponding to the position. It is like writing with binary numbers the positions. It uses a grayscale, it is not a binary code. The gray is made in a way so that one bit changes every time, not more than one like in the binary.



Generate digital words that represent the encoder’s actual position, as well as its speed and direction of motion:

- if power is lost, its output will be correct whenever power is restored
- it is not necessary to move to a reference position

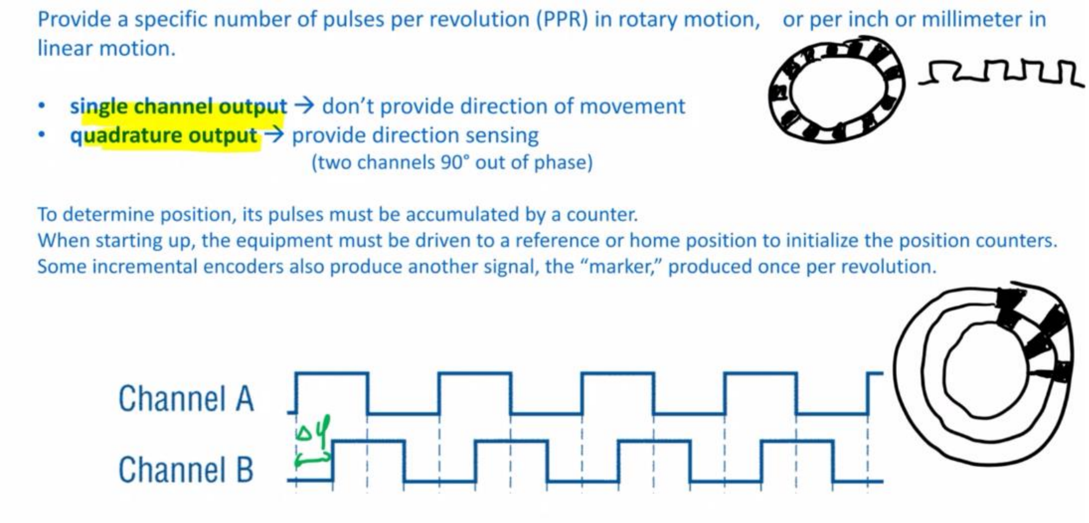
Binary or Gray code

Binary	Gray
000	000
001	001
010	011
011	010
100	110
101	111
110	101
111	100

Incremental encoders

We can have 2 kinds of them:

- Single channel output: we have a wheel with alternating spaces where light is blocked or can pass through. It doesn't provide the direction of the movement.
- Quadrature output: we have 2 sensing channels, two concentric wheels in which we have the blocking which is in quadrature. So the output will be as in the image. We will have a channel A that generates some pulses and B that generates pulses with a phase shift of ¼ of the period. With this solution we can measure if the object is rotating clockwise or counterclockwise. If A arrives before B we are rotating in one direction.



Provide a specific number of pulses per revolution (PPR) in rotary motion, or per inch or millimeter in linear motion.

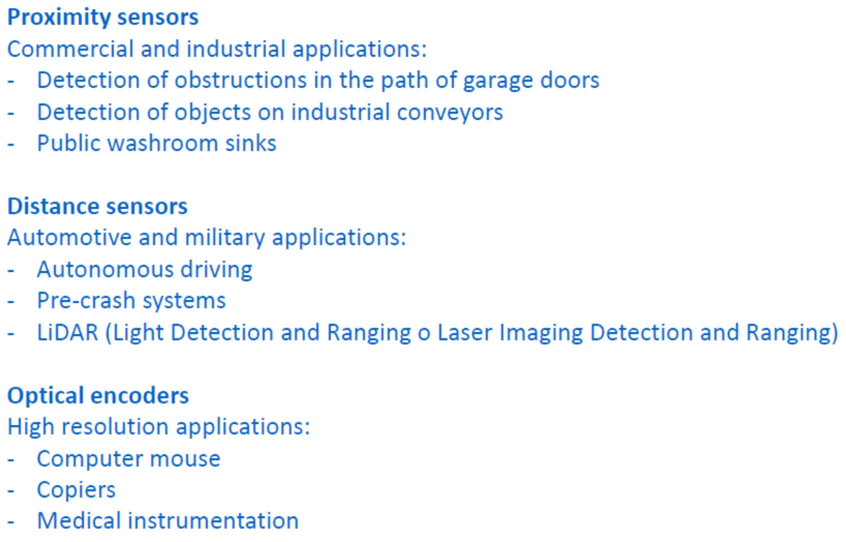
- **single channel output** → don't provide direction of movement
- **quadrature output** → provide direction sensing (two channels 90° out of phase)

To determine position, its pulses must be accumulated by a counter.
When starting up, the equipment must be driven to a reference or home position to initialize the position counters.
Some incremental encoders also produce another signal, the "marker," produced once per revolution.

Absolute encoders

We have a certain number of bits that encode the precise position.

Applications



Proximity sensors

Commercial and industrial applications:

- Detection of obstructions in the path of garage doors
- Detection of objects on industrial conveyors
- Public washroom sinks

Distance sensors

Automotive and military applications:

- Autonomous driving
- Pre-crash systems
- LiDAR (Light Detection and Ranging o Laser Imaging Detection and Ranging)

Optical encoders

High resolution applications:

- Computer mouse
- Copiers
- Medical instrumentation

VARIABLE RELUCTANCE SENSORS

Encoders can be created also based on magnetic sensors. They can be made either with optical sensors or magnetic sensors. As for the latter case, they can be based on a variable reluctance sensor. In this case they can detect rotations of only ferromagnetic targets. So we have a rotating wheel made out of ferromagnetic material; we put the sensor with a certain air gap in between and sense the variation of the reluctance due to the field of the rotating field.

Magnetic reluctance *R* of a magnetic circuit ⟺ Resistance R of an electrical circuit

$$R = \frac{F}{\Phi(\vec{B})}$$

R = magnetic reluctance

F = magneto-motive force: $F = \oint H \cdot dl$

$\Phi(\vec{B})$ = magnetic flux: $\Phi\left(\vec{B}\right) = \oiint \vec{B} \cdot \vec{n}\ dS,$

for fixed surface normal to B: $\Phi\left(\vec{B}\right) = B \cdot A = \mu_0 \mu_r H\ A$

The magnetic reluctance is the equivalent of the resistance in electrical circuits. In magnetic circuits, R is the reluctance, and since R = V/I, we have in place of V the magnetomotive force F and in place if I the magnetic flux of B. if we change the reluctance, given a fixed F, we will experience a variable flux of B.

Working principle

We have a rotating wheel and the presence of the dents increases or decreases the reluctance. If we are very close to the ferromagnetic material, the reluctance reduces, because the ferromagnetic materials are good conductors of the magnetic field, while if we are close to the part without tooth, the reluctance increases.. If the wheel is rotating, the reluctance varies over time. If we use a permanent magnet to generate the magnetic field H, F is fixed due to the fixed magnet, R varies because the wheel rotates and so the flux of B varies. But if so, due to the Faraday-Neumann-Lenz law, every tiem we have a variable flux of B we can induce a voltage across a coil. So we have a coil whose voltage will change depending on the derivative of the flux of B. Then the changes in the flux of B determine a change in the magnetic reluctance.

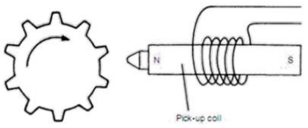
Permanent magnet: establishes a fixed magnetic field (H)

Ferrous metal target: changes the magnetic reluctance → changes $\Phi(\vec{B})$

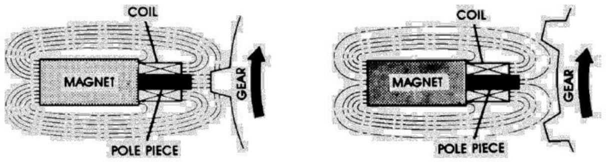
Coil winding: experiences an induced voltage $V_i = -n \cdot \frac{d\Phi(\vec{B})}{dt}$

So if we measure the frequency of the induced signal we can obtain the speed of the rotation. However, in this case we are insensitive to the direction of the rotation.

Instead of using a permanent magnet, we can exploit the semi-coil used to detect the



Pick-up sensor

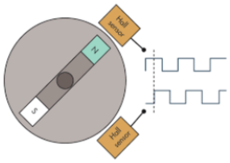


variation of flux of B also the generate the magnetic field. In fact, a coil crossed by a current, even a DC current, will generate a magnetic field.
So we bias the coil in DC with a voltage generator that provides a DC current to the inductor to generate a magnetomotive force F, but then across the inductance we also have a Vout given by a constant component plus a small signal that is due to the induced voltage.

Magnetic wheel encoders

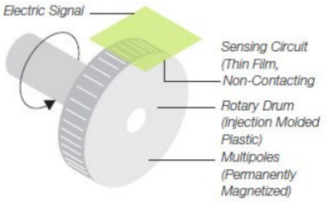
If we don't have a rotating ferromagnetic wheel, we still can use a magnetic sensor just putting on the rotating object a permanent magnet. Then we can use Hall sensors to measure the magnetic field. We can use just one if we want to have a single channel output like in the optical rotary incremental encoders, or we can put two of them in quadrature if we want to measure the direction of rotation, like in the optical incremental quadrature encoder.

Rotation frequency and direction



- One permanent magnet in the wheel
- Two magnetic field sensors
  (Hall sensors or magneto resistance)

Rotation speed


Electric Signal
Sensing Circuit (Thin Film, Non-Contacting
Rotary Drum (Injection Molded Plastic)
Multipoles (Permanently Magnetized)

- Many permanent magnets in the wheel
- One magnetic field sensor
  (Hall sensors or magneto resistance)

With the circuit on the left we can measure the frequency of the rotation and also the direction. To measure the speed of rotation, we need to put many Hall sensors, but this is not convenient, so we use only one sensing circuit but many magnets.

Magnetic encoders applications

Non contact sensor, reliable in harsh environments

- Rotation speed
  - motors feedback
  - ABS control
- Rotating gears position
  - high precision positioning
  - human machine interface
  - machine to machine interface

MICROPHONES

They are sensors used to sense the sound pressure.

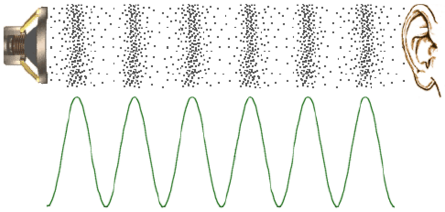
SOUND FIELD

sound = vibration that propagates as a typically audible mechanical wave of pressure and displacement, through a medium.

Sound is transmitted as longitudinal waves through gases and liquids

⇩

waves of alternating pressure deviations from the equilibrium pressure, causing local regions of compression and rarefaction



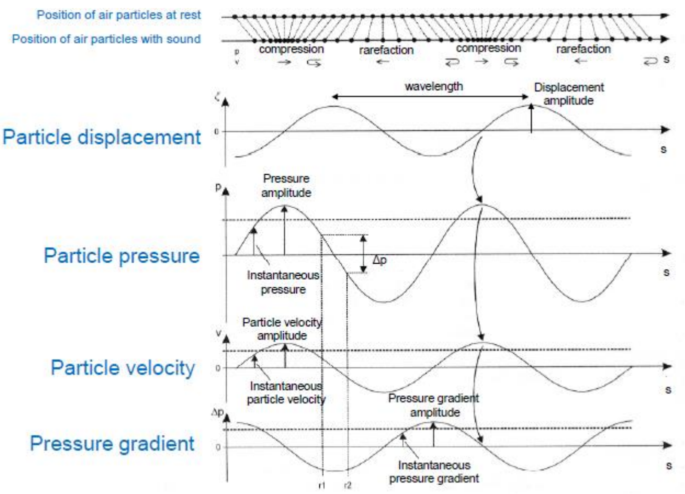
Sound is a vibration that propagates within a medium, it is a vibration of the particles of the medium itself. Typically, these vibrations can be sensed by the human ear, s they are audible. The sound is transferred as a longitudinal wave, that is a wave in which the propagation speed is on the same direction of the wave itself. Hence the propagation speed is perpendicular to the wavefront.

SOUND FIELD QUANTITIES

If we imagine to have an ideal medium with no sound propagation in which all the particles are all at the same distance, it there is a sound propagation in the medium, the we will have some area in which we have compression and some where we have rarefaction. This means that the particles of this medium are moving around the resting position, and this movement is a sinusoidal movement. The movement represents the particle displacement. To this displacement also corresponds a variation of pressure in the medium, where we have rarefaction we have the minimum pressure, where we have compression the maximal. Pressure is an important parameter because it is typical measured by the microphones.

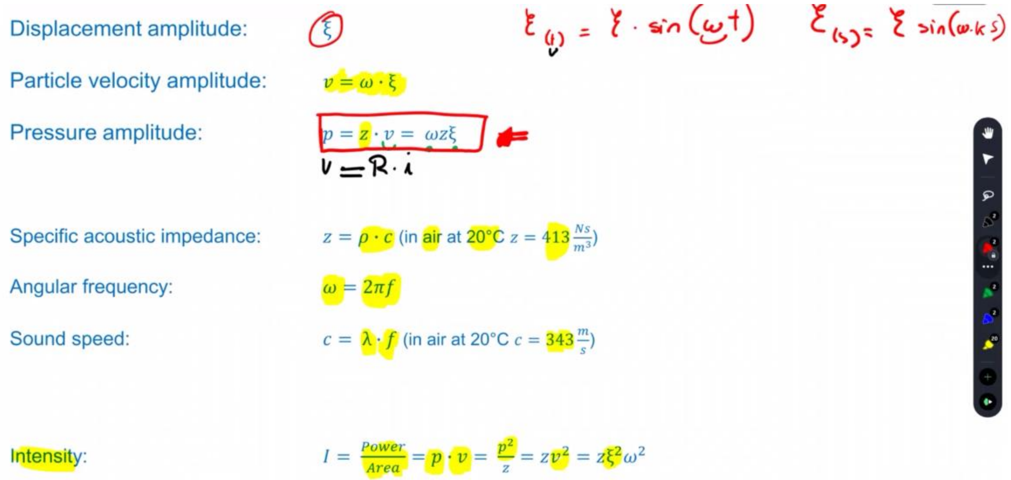
Another parameter is the particle velocity, that is different from the sound velocity, that is the velocity with which the waves propagate in the medium, while the particle velocity is the speed with each particle oscillates around its equilibrium position. It is simply the derivative of the displacement.

Then we introduce other parameters, like the pressure gradient. It is the gradient of pressure; if we consider the pressure in two points, the difference in pressure we have is the pressure gradient. It is the derivative of the pressure.
This parameter is important because some mics are sensible to particle pressure, others to pressure gradient, so pressure derivative.



Relationship among these quantities

All the previous quantities have a sinusoidal behaviour, but which is the amplitude of the sinusoids? As for displacement, we have csi. So if we express the displacement over time, it is the amplitude multiplied by the sinusoid. We can write the dependance either on t or on the space. It is the same because we can analyze a single point and we can see that over time on that point we have a sinusoidal behaviour (over time) or, over space, fixing the time, we have still a sinusoidal behaviour of the displacement.



Displacement amplitude: ξ $\quad \xi_{(t)} = \xi \cdot \sin(\omega t) \qquad \xi_{(s)} = \xi \sin(\omega \cdot k s)$

Particle velocity amplitude: $v = \omega \cdot \xi$

Pressure amplitude: $\boxed{p = z \cdot v = \omega z \xi}$

$V = R \cdot i$

Specific acoustic impedance: $z = \rho \cdot c$ (in air at 20°C $z = 413 \frac{Ns}{m^3}$)

Angular frequency: $\omega = 2\pi f$

Sound speed: $c = \lambda \cdot f$ (in air at 20°C $c = 343 \frac{m}{s}$)

Intensity: $I = \frac{Power}{Area} = p \cdot v = \frac{p^2}{z} = zv^2 = z\xi^2\omega^2$

As for the amplitude of the velocity, the velocity is the derivative of the displacement, so the amplitude is given by omega*amplitude of the displacement.

As for the pressure amplitude, it is defined as z, that is the acoustic impedance, multiplied by the velocity. It is like an electrical analog in which V = R*I.
To find relationship between pressure and displacement we substitute the formula for the velocity.

The specific acoustic impedance is a property of the medium in which the sound is propagating. It is given by the density ro multiplied by c, where c is the speed of sound, that is not a particle velocity but the speed at which the sound propagates in the medium. Z depends on temperature because the velocity of the sound propagation depends on temperature and in air at 20°C is 413 Ns/m^2.

As for the angular frequency, omega is related to the frequency of the sound with the usual relationship between omega and f.

As for the sound speed, it is given by the wavelength lambda per the frequency. In air, c is more or less 343 m/s.

An important quantity for the sound is the intensity, that is what our ear can perceive. Intensity is defined as power over area, and it can be computed as pressure multiplied by the velocity. If we substitute to pressure and velocity all the previous formulas we can find various dependencies.

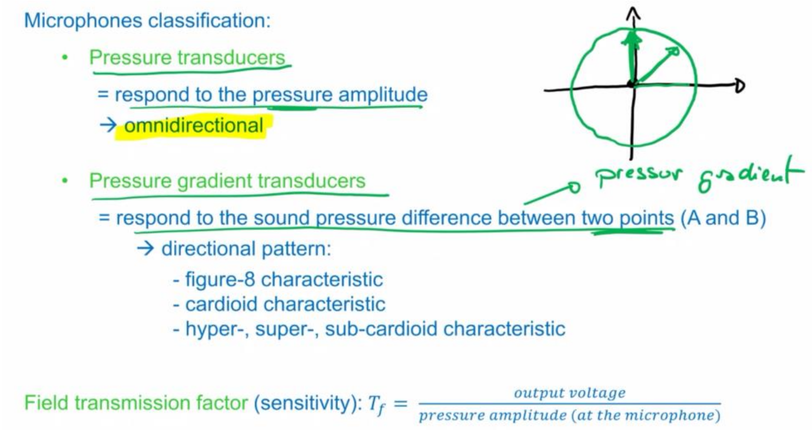
Particularly important are the relations between the pressure and the velocity or the displacement. Indeed, some capsules, that are the sensible parts of the mics, measure pressure measuring the velocity of the particle, while others exploit the displacement of the particle.

MICROPHONES FUNDAMENTALS

Mics can be classified in two big families:

- Pressure transducers: they measure the particle pressure, they respond to pressure amplitude. They have an **omnidirectional** characteristic. This means that this kind of mics have the same sensitivity independently from the direction of the sound. The sensitivity is called **transmission factor**, and the green plot represents it.
- Pressure gradient transducers: they respond to the sound pressure difference between two points (they respond to pressure gradient). They have a **directional pattern**. This means that the sensitivity of the mic is different depending on the direction from which the sound arrives. This is important because if we imagine to have a mic for a singer, we want to amplify just the voice of the person that is singing, not the one of the public, so amplifying only the sound that arrives orthogonal. In this case we use a pressure gradient transducer.

  Typical directional patterns are:
  - Figure-8 characteristic
  - Cardioid characteristic
  - Hyper-, super-, sub-cardioid characteristic



Microphones classification:

- Pressure transducers
  - = respond to the pressure amplitude
  - → omnidirectional
- Pressure gradient transducers
  - = respond to the sound pressure difference between two points (A and B)
  - → directional pattern:
    - - figure-8 characteristic
    - - cardioid characteristic
    - - hyper-, super-, sub-cardioid characteristic

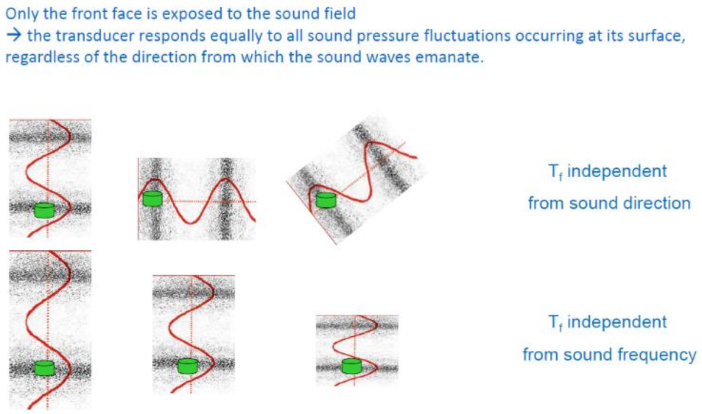
Field transmission factor (sensitivity): $T_f = \frac{\textit{output voltage}}{\textit{pressure amplitude (at the microphone)}}$

As said, the sensitivity of the mic is the transmission factor, defined as the output voltage of the mic divided by the pressure amplitude (classical output over input).

PRESSURE MICROPHONES

In pressure mics we have only one sensitive capture. It is the sensitive element, that is sensible. So the green element is the sensible one and the pressure measured by the capture will change sinusoidally. The pressured measured by the capture is independent from the direction in which the sound propagates.

Also the sensitivity is independent from the sound direction and also from the sound frequency. If we change the frequency, we will measure a pressure varying with a different frequency but with always the same amplitude.



Only the front face is exposed to the sound field
→ the transducer responds equally to all sound pressure fluctuations occurring at its surface, regardless of the direction from which the sound waves emanate.

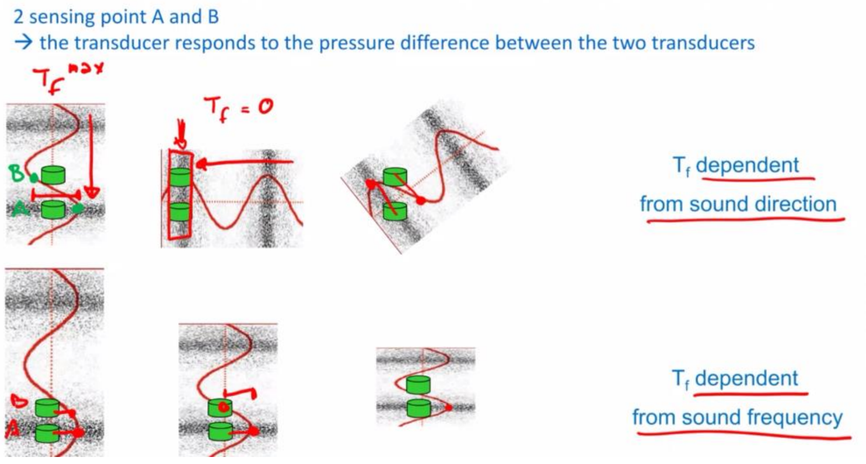
T_f independent from sound direction

T_f independent from sound frequency

PRESSURE GRADIENT MICROPHONES

In pressure gradient mic we have to do the derivative of the pressure, and to do so we can simply measure the pressure in two different points, and then the output will be the difference between the pressure measured in these two points.

Let's imagine to have two different separates capsules. Then these two capsules are omnidirectional, able to measure the pressure in one point. Capsule A measures the pressure in point A, and capture B in point B. Then the output will be proportional to the difference measured by A and B. Now there is a strong dependence of the sensitivity of the mic with the propagation direction. If the sound arrives orthogonally to the two capsules of the mic, then we will have the maximum sensitivity, because we will see the maximal difference in pressure between A and B. Instead, if the propagation of the waveform is parallel to the capsules, then in this case the two capsules will always measure the same pressure. In this case Tf = 0.



2 sensing point A and B
→ the transducer responds to the pressure difference between the two transducers

T_f dependent from sound direction

T_f dependent from sound frequency

Hence Tf depends on the sound direction, but Tf depends also on the sound frequency. If we imagine to have a sound at very low frequency, the two capsule measure almost the same value. If we increase the frequency of the incoming waveform we also increase the difference of the pressure we are measuring (the difference in pressure is given by the horizontal line). If we increase the frequency we will arrive at a certain point in which we are at Tf max, that is when the distance between the two capsules is half of the lambda of the sound → **Tf depends on frequency, and the sensitivity is higher for higher frequencies**.

Directional characteristic

There are 2 possible characteristics.

Figure-8

The name is given because the graph that represents the transmission factor is an eight (8). We have capture A and B in the vertical direction as in the graph. If the sound is perfectly perpendicular to the mic, so at 0°, we have the maximal transmission factor (Tf), while if the sound comes from 90° the Tf will be equal to 0. This is something seen in the previous image.

If the sound is arriving from a different angle, we will have an intermediate transmission factor. The dependence of Tf with the incoming angle of the sound can we expressed by the green equation, in which the Tf is the Tf at 0° multiplied by the cosine of the angle. We consider 0 the orthogonal (vertical) direction.
So for instance if theta is 90°, Tf = 0. Tf from 180° is a negative value, and this means that the gradient that is measured is not positive but negative, but since we are dealing with a sinusoidal pressure, we are simply measuring a sound with opposite phase.

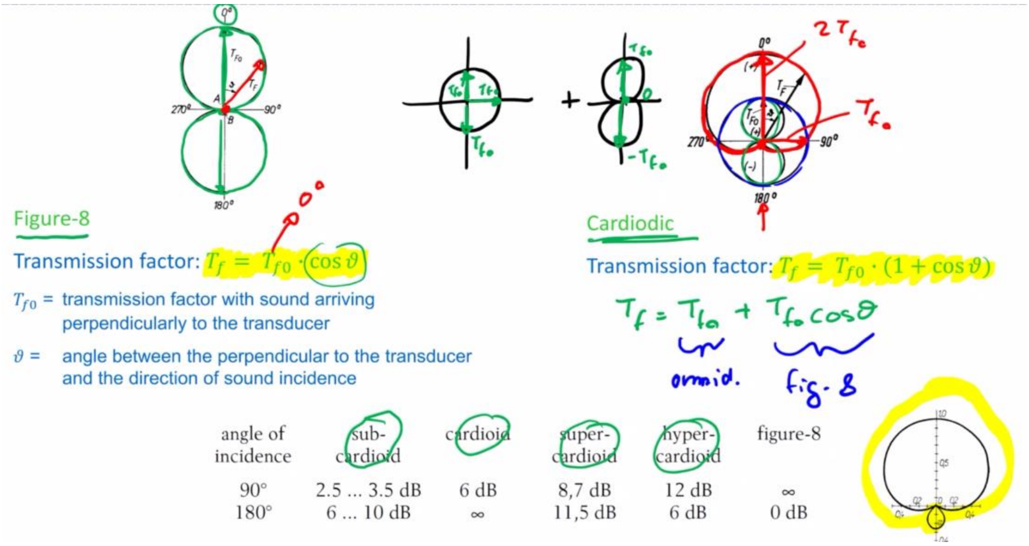


Figure-8

Transmission factor: $T_f = T_{f0} \cdot \cos\vartheta$ 0°

T_{f0} = transmission factor with sound arriving perpendicularly to the transducer

ϑ = angle between the perpendicular to the transducer and the direction of sound incidence

Cardiodic

Transmission factor: $T_f = T_{f0} \cdot (1 + \cos\vartheta)$

$$T_f = \underbrace{T_{f0}}_{\text{omnid.}} + \underbrace{T_{f0}\cos\theta}_{\text{fig. 8}}$$

angle of incidence	sub-cardioid	cardioid	super-cardioid	hyper-cardioid	figure-8
90°	2.5 ... 3.5 dB	6 dB	8,7 dB	12 dB	∞
180°	6 ... 10 dB	∞	11,5 dB	6 dB	0 dB

Cardiodic (or cardioid)

It is called in this way because the characteristic has a shape similar to a heart. It can be seen as the sum of an omnidirectional characteristic plus a figure-8 characteristic. At 0° we have Tf0 of the omnidirectional and of the figure8. So if we sum the two characteristic, we end up with a Tf that is two dimes Tf0.

Then if we change angle, like 180°, we still have tf0 for the omnidirectional and -Tf0 for the figure8 → Tf for the cardioid is 0, so the sensitivity is 0. At 90° we have to perform the same reasoning, and we will have a transmission equal to Tf0 because Tf0 of the figure8 is 0.

This **cardioid shape can be described as the Tf of the omnidirectional plus the one of the figure8**.

We can have also intermediate cases described as sub-cardiodic, super-cardiodic and so on. They depend on the weight that we use to sum the omnidirectional and the figure8.

So far we have studied the dependency on the direction, but pressure gradient mics (PGM) shows also a dependency on frequency.

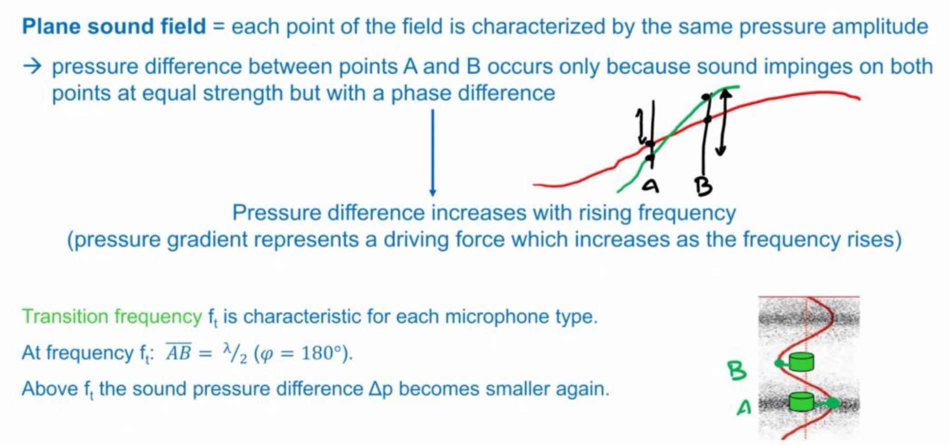
FREQUENCY DEPENDENCY FOR PGM

Plane sound field

Let's start with the case in which we have a plane sound field, that is a field in which the sound intensity can be considered everywhere the same. If we imagine to be **in a plane sound field, the sound is not attenuating in the space**, so we have always the same intensity. In this case, if we have our two capsules, the difference in pressure between the two capsules depends only by the factor that we are measuring a different phase of the pressure. So the difference between A and B depends only by the fact that they are in different places and so they measure only a different phase of the waveform, even if the waveform has always a constant amplitude.

If I increase the frequency, I increase the difference between the two points. If the frequency is low, then the difference we measure in pressure will be small. With higher frequency, then we will measure a higher difference between A and B → *increasing frequency we also increase the gradient of pressure, given a constant amplitude of the pressure*.

However, there is a maximum frequency for which our mic increases is Tf depending on frequency. Indeed, if we try to increase too much the frequency, then the difference between A and B will start to decrease again, because we will start to measure a smaller difference.



Plane sound field = each point of the field is characterized by the same pressure amplitude

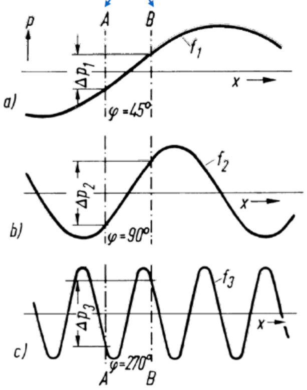
→ pressure difference between points A and B occurs only because sound impinges on both points at equal strength but with a phase difference

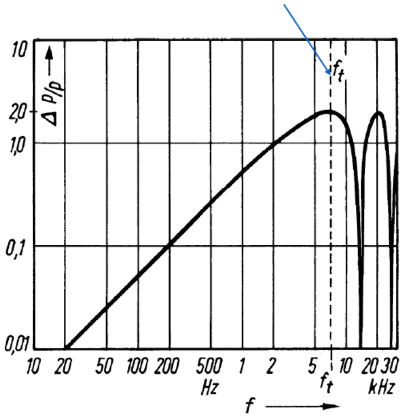
Pressure difference increases with rising frequency
(pressure gradient represents a driving force which increases as the frequency rises)

Transition frequency f_t is characteristic for each microphone type.
At frequency f_t: $\overline{AB} = \lambda/2$ $(\varphi = 180°)$.
Above f_t the sound pressure difference Δp becomes smaller again.

So the frequency for which we have the maximal sensitivity is called transition frequency. This transition frequency is the frequency that corresponds to a wavelength that is double the distance between A and B. when the distance between A and B is equal to lambda/2, we have the maximum sensitivity. Above this frequency, deltaP starts to decrease again.

If we plot it in a graph, we have the sensitivity Tf that increases with frequency up to the transition frequency and then starts to decrease again, falling to 0. When the distance between A and B is exactly equal to lambda, the Tf falls to 0. In this case we are in a case where the two capsule will always measure the same pressure. Every time the distance between A and B is a multiple of the lambda we have a Tf = 0. However, our mic works for frequencies that are lower than the transition frequency → operating bandwidth is just up to the transition frequency.

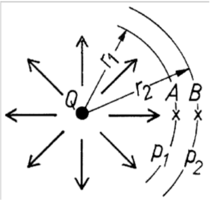

Pressure gradient transducer


Transition frequency

Spherical sound field

It is more similar to real sound fields. There is a source of sound which propagates in all the directions, but during the propagation it will also attenuate. So **the difference in pressure A and B** doesn't **depend** only on frequency, but **on frequency and on the distance from the source of the sound**.

Before, in the plane sound field, we have said that if we put the two capsules exactly at the distance corresponding to the wavelength, then we will have an output equal to 0, while in this case if the two capsules A and B are at the distance corresponding to the wavelength; if the sound field is plane the pressure on capsule A is equal to the one of B. If the sound filed is spherical, since we have an attenuation of the sound, our sound is also attenuating, so we will not measure exactly the same pressure, but a pressure different from 0. This to say that we have to consider both the effects due to the phase shift that is proportional to frequency and the effect of the attenuation of the sound in space.



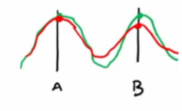
When a point source of sound is approached,
the sound pressure rises at a ratio of 1/r (r = distance)

2 contributions to the output signal:

- phase-related Δp → frequency dependent
- distance-related Δp → frequency independent

most noticeable in the low frequency

→ pressure gradient microphones tend to boost low-frequency components when held close to the mouth (i.e. when the distance r from the sound source is approximately equal to the length of the sound wave).



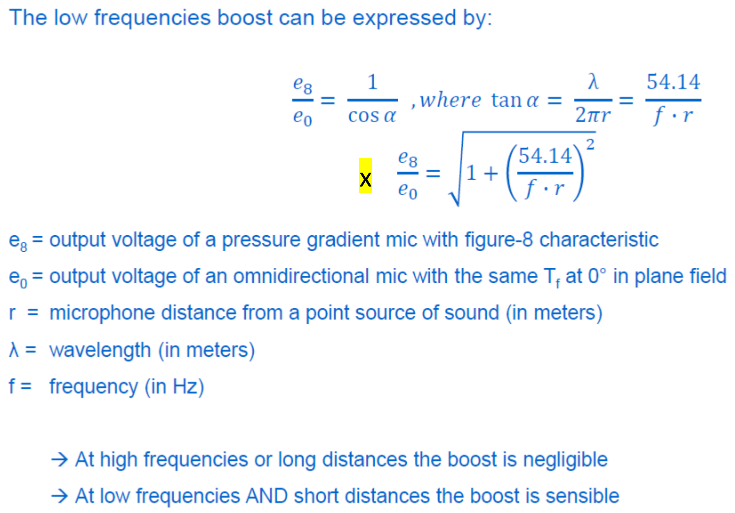
So we have two contributions:

- Phase related deltaP: depends on frequency, as seen in the plane field
- Distance related deltaP: the fact that the sound is attenuating in space. It is frequency independent.

<u>The distance related contribution is most noticeable at low frequency</u>. Indeed, at low frequency, the sensitivity due to the phase shift is very low, due to the phase shift, because the deltaP we have due to the phase shift is small. This means that the phase related deltaP at low frequency is quite low, so the other contribution becomes the dominant one. Instead, at HF, the dominant contribution is given by the phase related deltaP, because we have a high transmission factor.

This is the reason why we have a boost of the low frequency when we put the mic close to the mouth, producing a sort of distortion. This simply because the low frequencies are more related to the distance related deltaP.

LF boost in spherical field



The low frequencies boost can be expressed by:

$$\frac{e_8}{e_0} = \frac{1}{\cos\alpha} \;, where \;\tan\alpha = \frac{\lambda}{2\pi r} = \frac{54.14}{f \cdot r}$$

$$x \quad \frac{e_8}{e_0} = \sqrt{1 + \left(\frac{54.14}{f \cdot r}\right)^2}$$

e_8 = output voltage of a pressure gradient mic with figure-8 characteristic

e_0 = output voltage of an omnidirectional mic with the same T_f at 0° in plane field

r = microphone distance from a point source of sound (in meters)

λ = wavelength (in meters)

f = frequency (in Hz)

→ At high frequencies or long distances the boost is negligible

→ At low frequencies AND short distances the boost is sensible

These equations express which is the output of a pressure gradient microphone (e8) compared to an omnidirectional one, so to a pressure microphone (e0).
We alpha that is not an angle but it is expressing the relationship between lambda and r. If we look at the equation x, we see that the output of a figure8 mic with respect to the omnidirectional has a boost, it is 1 plus something.

This something, if the frequency is high or the distance is high is almost 0. So the output of the e8 is similar to e0. If instead we decrease the frequency and the distance, which means at low frequencies and close to the mouth, then we experience the boost, so the output of figure8 will relatively increase.

Hence the equation x states that only if we are close to the mouth and the frequency is low we have that output.

To summarize, e0 can be implemented with only one capsules, which measures the pressure, while the figure8 by two capsules.

CARDIOID TRANSDUCER

There are 3 ways to get a cardioid characteristic:

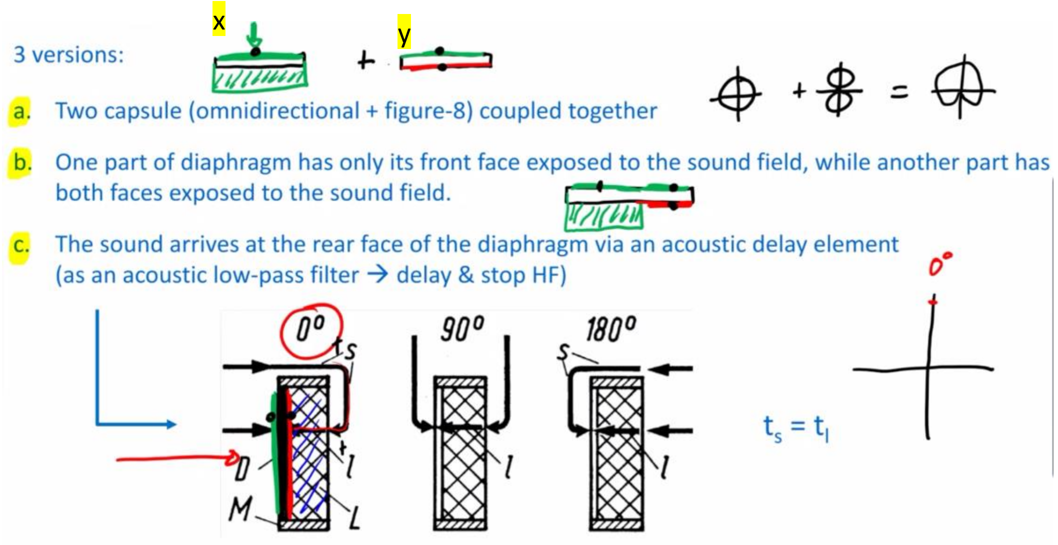
1. Summing the omnidirectional and the figure8. So we use two capsules, one is omnidirectional and the other one is the figure8 capture and then we sum the two outputs. We have two mics in one mic.

   Typically, to create an omnidirectional mic, we have a diaphragm and we make it in a way so that only the sound coming from the top direction are sensed (x). So we measure the pressure locally on the top of the diaphragm (the bottom part is isolated). To make a figure8, we make the diaphragm so that the sound can reach both the top part of the diaphragm but also the bottom part. At the end in the figure8 we measure the difference between the pressure that we have at the top and the one we have at the bottom.
2. We use a single diaphragm in which a portion of it is exposed on both the faces, while the other portion only in the front. In the end the diaphragm will experience part of its deformation only due to top pressure, and part due to top and bottom.
3. We implement a sound delayer. It is a material that is able to delay the sound; it is a material in which the sound propagates with lower velocity with respect to air. Using a material different from air we can slow down the sound. In this mic we have the diaphragm (between green and red lines) and the top of the diaphragm (green) is directly exposed to the sound; the rear of the diaphragm (red) is not directly exposed because the sound has to travel to reach the back and pass through the material (blue) that makes a sort of delay. If the sound (that propagates in the red arrow direction) comes from a 0° angle, we have the maximum phase shift between the sound that reaches the top and then has to travel and go to the back to reach the bottom. So we have a phase shift between the sound that reaches the front face and the one that reaches the rear face. In particular, the time needed to reach the rear face is given by ts + tl. Ts is the time to propagate around the microphone, while tl is the time introduced by the delayer, that can be finely regulated.

   At 0° we have the maximum phase shift between the front and the rear, so we will have also the maximum transmission factor Tf.

   If the sound comes from 90°, the sound at the front arrives directly, while the one that needs to reach the rear face has to travel within the delayer, so it will arrive with a delay equal to tl. So the phase shift is not given by ts + tl but only tl. If we create the mic so that tf = tf, in this case we will have half of the phase shift of the 0° case.

   For the 180°, the sound, in order to reach the front phase, has to travel a certain distance and it will take the time ts. Instead, to reach the rear phase, the time needed is tl. If ts = tl, it means that the sound reaches exactly in phase the front and rear phase → the diaphragm will not vibrate and the Tf is equal to 0.



3 versions:

a. Two capsule (omnidirectional + figure-8) coupled together

b. One part of diaphragm has only its front face exposed to the sound field, while another part has both faces exposed to the sound field.

c. The sound arrives at the rear face of the diaphragm via an acoustic delay element (as an acoustic low-pass filter → delay & stop HF)

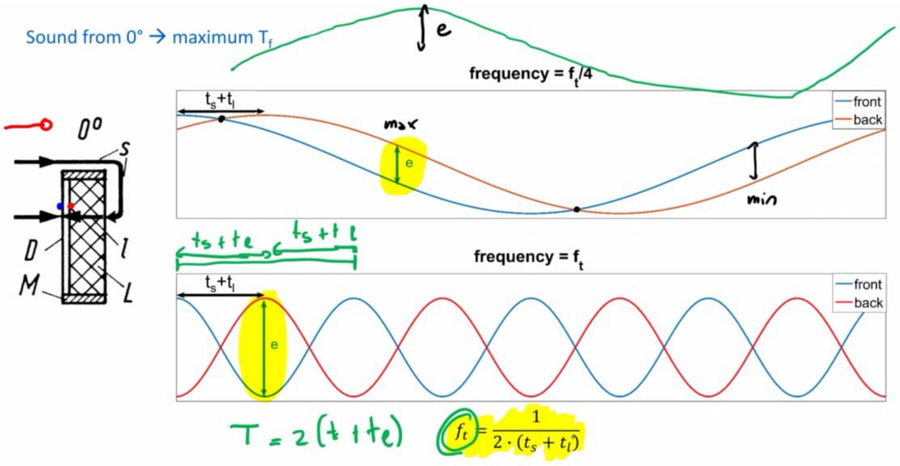
Configuration 3

Which is the maximum frequency at which the mic in configuration 3 can operate to maximize its output?

Pressure gradient transducers have a dependency with frequency.

If we imagine to have a sound coming from a 0°, that is the angle at which we have the maximal sensitivity, we have two different waveforms: the one that reaches the front face (blue) and, with a certain delay, the sound reaching the back of the diaphragm. This delay for the 0° angle is ts + tl. The diaphragm will measure the difference between the blue and red waveforms. This difference will have a sinusoidal shape.

We are not interested in the phase of this 'difference sinusoid', but in its amplitude e. Given a generic wavelength, we have this amplitude e. Now, let's imagine to have a lambda that matches the delay ts + tl, and in particular in the case in which ts + tl is lambda/2. In this case, the wavelength that reaches the back is in opposition of phase with respect to the one coming from the frontal direction.



Sound from 0° → maximum T_f

$T = 2(t + te)$ $f_t = \frac{1}{2 \cdot (t_s + t_l)}$

We are in the case of the bottom plot, in which the two curves are in opposition of phase because the ts + tl matches exactly half of the wavelength. In this case we have the maximal sensitivity that we can have.

If we increase more the frequency, then the output, so the difference between the two, starts to decrease.
So we can say that the transition frequency is ½(ts+tl). The period of the sinusoidal with maximal transition factor is (ts+tl)*2. If I have a wave with this period, the Tf will increase up to this transition frequency.

MICROPHONE DIMENSIONS

Depending on the frequency of the sound we want to measure, the mic, in order to have the ideal characteristic, so for instance the omnidirectional characteristic, we need to have a mic that is small enough not to disturb the sound propagation. If the mic is an ideal point, this point will not discharge the propagation of the sound, having hence all the ideal characteristics shown up to now.
To be considered 'small', the mic must be smaller than the wavelength. Object indeed can interact with sound only if the size of the object is similar to the wavelength.

Microphones with dimensions similar to or greater than the wavelengths being picked up present an obstacle for the sound waves

Frequency (Hz)	Wavelength (cm)
32	1050
320	105
3200	10.5
16,000	2.1

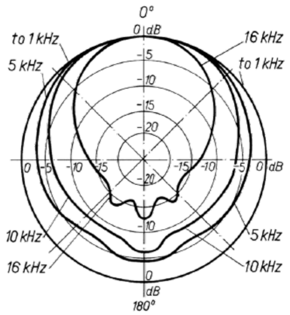
All effects caused by the dimensions of the microphone are frequency-dependent.

→ Mic no larger than 6 mm in all three dimensions for a limit frequency of 16 kHz

The wavelength at 16kHz is 2.1cm and as a rule of thumb, the mic should be no larger than 1/3 of the wavelength in order to be able to detect frequency as high as 16kHz without any distortion in the characteristic of the mic.

If the mic is bigger, than we will have the ideal characteristic only for low frequencies (circular one), while if we increase the frequency the characteristic starts to becomes more similar to an hyper-cardiodic characteristic.

- sound arriving perpendicularly to the diaphragm exerts more force on the diaphragm as the result of pressure build-up
- sound waves impinging diagonally do not strike all parts of the diaphragm simultaneously, giving rise to interference cancellations that are dependent on both direction and frequency (interference transducers)



at high frequencies omnidirectional characteristic gradually changes to a unidirectional polar pattern

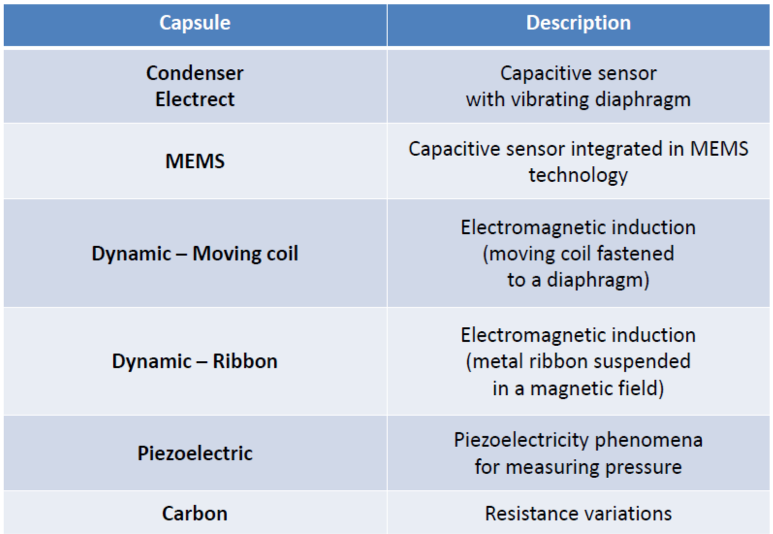
↓

distortions in diffuse field

If the sound comes from 0° nothing changes independently from the frequency, because the sound directly arrives to the diaphragm. If instead the sound arrives from 180° the size is important, because before arriving to the diaphragm the sound interacts with the mic itself.

CAPSULES

They are the sensitive elements. They can be done with different technologies. MEMS are like condenser microphones, based on capacitive effect, but at the micrometric scale. Instead, dynamic capsules use the inductive law, the Faraday-Neumann-Lenz law.



Capsule	Description
Condenser Electrect	Capacitive sensor with vibrating diaphragm
MEMS	Capacitive sensor integrated in MEMS technology
Dynamic – Moving coil	Electromagnetic induction (moving coil fastened to a diaphragm)
Dynamic – Ribbon	Electromagnetic induction (metal ribbon suspended in a magnetic field)
Piezoelectric	Piezoelectricity phenomena for measuring pressure
Carbon	Resistance variations

When we analyze the output of a mic, the output is influenced by 3 parts. If we analyze the transfer function of a mic, it is influenced by 3 parts:

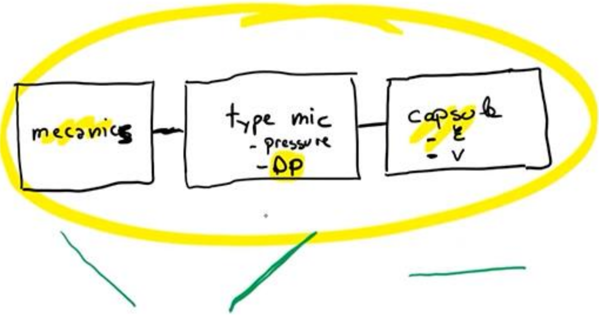
1. Mechanics: the mic has a mechanical part, that is the diaphragm, and it is put in vibration by the sound. So the first part of the transfer function is how the sound puts in vibration the object, and it is a pure mechanical part.
2. Type of microphone: in the transfer function it is important to consider if the mic is a pressure or a pressure gradient, because if it measures pressure the output will be independent on frequency and so on.
3. Capsule-dependent part: it depends on how I measure the pressure or the delta pressure, if I measure the velocity or the displacement of the particle. There are some capsules that measure the displacement and others that measure the velocity. Condenser are mics that measure the displacement, while dynamics measure the velocity.

   Condenser are based on a capacitive effect, so the diaphragm is one plate of the capacitor. Hence the plate of the capacitor will start to vibrate together with the particle displacement. The amplitude of the vibration will depend on the displacement of the particle.

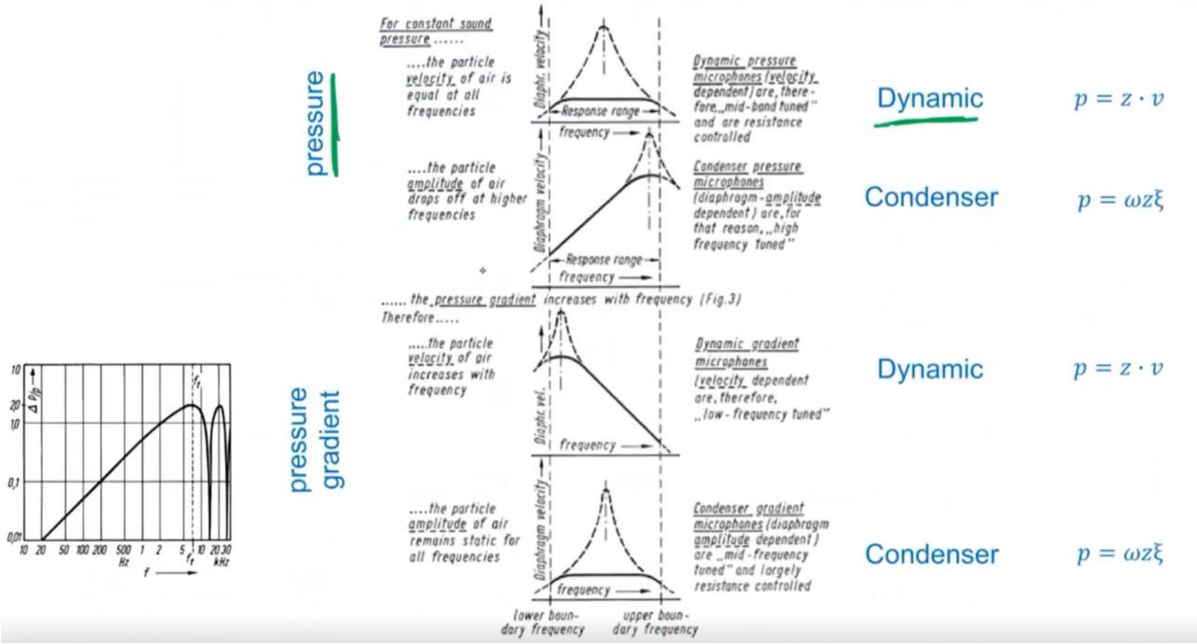
   Instead, dynamics mics, since based on the Faraday-Neumann law have an output that will depend on the derivative of the flux of v, and the derivative of the flux of v will depend on the velocity at which we move a moving coil.

Overall, **we want to have a transfer function the more insensitive to frequency as much as we can**, we don't want to amplify a lot the HF but not the LF. So for instance, if we use a pressure gradient microphone (that have a characteristic that depends on frequency), we have to try to compensate the dependence with frequency either with mechanics or a proper capsule which has a transfer function that is opposite.

Since we have a pressure gradient mic, the tf (transfer function) increases with frequency and we can compensate with a mechanics whose tf decreases with frequency. Then we have for instance a tf for the capsule independent from frequency.
In the end, thus the tf amplifies all the wavelengths in the same way.



Mechanics


pressure
pressure gradient
For constant sound pressure ......
....the particle velocity of air is equal at all frequencies
....the particle amplitude of air drops off at higher frequencies
...... the pressure gradient increases with frequency (Fig.3)
Therefore.....
.....the particle velocity of air increases with frequency
....the particle amplitude of air remains static for all frequencies
Diaphr. velocity
Diaphragm velocity
Diaphr. vel.
Response range
frequency
lower boundary frequency
upper boundary frequency
Dynamic pressure microphones (velocity dependent) are, therefore, "mid-band tuned" and are resistance controlled
Condenser pressure microphones (diaphragm-amplitude dependent) are, for that reason, "high frequency tuned"
Dynamic gradient microphones (velocity dependent are, therefore, "low-frequency tuned"
Condenser gradient microphones (diaphragm amplitude dependent) are "mid-frequency tuned" and largely resistance controlled
Dynamic
Condenser
Dynamic
Condenser
$p = z \cdot v$
$p = \omega z \xi$
$p = z \cdot v$
$p = \omega z \xi$
Δp
f_r
10
20
50
100
200
500
Hz
1
2
5
10
20
30
kHz
0,1
0,01
f

In general, pressure mics are independent on frequency, so they are flat in frequency. Instead, a pressure gradient mic has a tf that increases with frequency.
If we have a dynamic mic, they are sensible to the velocity of the particle. Pressure and velocity are related by z that is independent from frequency. This means that, in general, dynamic mic are independent from frequency (both for pressure and pressure gradient). This means that the frequency response is flat.

Instead, for condenser mics, the relation between the displacement and the pressure depends on omega, so on frequency. In particular, if we increase the frequency we decrease the displacement.

Hence the tf will have a displacement that decreases with frequency.

Now, we can reason on how we want to have the mechanics. If we imagine to have a pressure transducer implemented with a dynamic capsule, since both pressure mic and dynamic capsules are independent on frequency, I want also the mechanical response to be independent from frequency, because I don't have to compensate anything.

If instead I have a pressure transducer implemented with a condenser capsule, pressure is independent on frequency, but the capsule will have a tf that decreases with frequency, so I will try to compensate this decrease with a mechanical tf that increases with frequency its response → **I'm tuning the response**.

When the response must be flat we are mid-band tuned, while when we work with a mechanical response that increase with frequency we are high frequency tuned.

HF tuning - pressure

Typically, all the mechanical components that vibrate have their resonance frequency. We build the mechanics in a way that the resonance frequency is at the highest limit of frequency that we want to measure. If we are at a frequency smaller than the resonance frequency we are working in a region where the tf is increasing.

Mid-band tuning - pressure

If we want to be mid-band tuned, we want the resonance frequency in the middle of the range we want to measure, but then we try to damp the resonance frequency. Instead, in the case of the condenser we don't want to damp this effect, because we want to have the increase in frequency.

Pressure gradient

Its tf increases with frequency. If we implement it with a dynamic mic, the dynamic mic has a flat tf with frequency, so we have to compensate the increase due to the pressure gradient using a mechanical part with a tf whose output decrease with frequency. So we are LF tuned. We make the mechanics in a way that the LF that we don't want to damp is around the resonance frequency.

If instead the mic is implemented with a condenser mic, the two tf almost compensate each other, so the mechanics of the mic should be mid-band tuned, with the resonant frequency in the middle of the range.

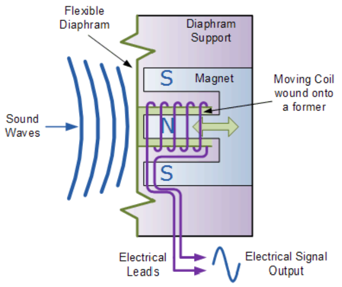
DYNAMIC MICROPHONES – MOVING COIL

Dynamic microphones are mics whose output is proportional to the velocity of the particle.
In this case we have a diaphragm that is mechanically connected to a moving coil; if the diaphragm vibrates, also the coil. The coil is wounded around a permanent magnet; if we move a coil in a magnetic field we will have a variable flux of B, so we will induce a voltage across the coil.

In particular, the output will be proportional to the derivative of the movement, so to the velocity.
With this dynamic mic we can implement both pressure transducers when only the front part of the diaphragm can sense the sound, or we can do pressure gradient transducers in which both the front and the back of the diaphragm are exposed to the sound.

Moving coil microphones have a small self-supporting coil that is fastened to a lightweight plastic diaphragm and moves in the air gap of a powerful permanent magnet.

→ A voltage proportional to the sound particle speed is induced at the coil terminals.



particle velocity transducer:

$$v = \frac{p}{z}$$

2 types of moving coils:

- pressure transducers
  (only front diaphragm)
- pressure gradient transducers
  (front and back diaphragm)

Pressure moving coil mic

If we want to do pressure transducers based on a dynamic characteristic, we want a mechanics to be mid-band tuned, so the mechanical output should be as flat as possible in the range of frequency that we have.

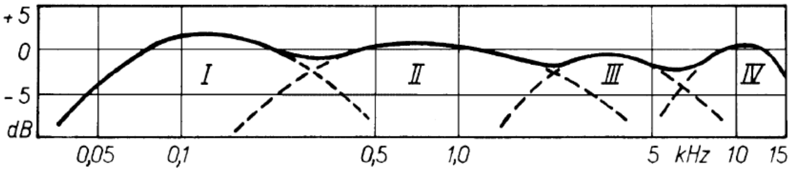
With moving coil, the diaphragm is connected to the moving coil, and the moving coil has a mass that is much bigger than the diaphragm. So when the mass that we put in movement is very big it is very hard to damp the resonance.

For this reason, they make many cavities in the mic in order to have many different resonant frequencies. In the end, if we sum all the characteristics of the resonant frequencies, we end up with a quite flat band.

For constant sound pressure the particle velocity is equal at all frequencies

→ «mid band tuned» with dumped resonance

mass of diaphragm + coil
many times greater than diaphragm mass
↓
hard to dump resonance
↓
additional resonances
with air cavities



DYNAMIC MICROPHONES – RIBBON MIC

In this case we don’t have a very heavy coil but we only have a very light ribbon. This metallic ribbon is the diaphragm itself, so it is able to vibrate together with the sound vibration. This ribbon experiences the magnetic field due to the permanent magnets.

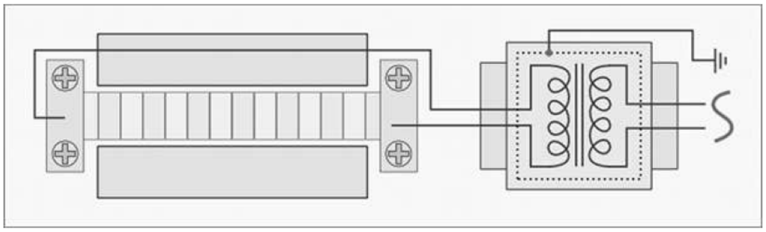
We have like a strip of aluminum moving within a magnetic field, so also in this case the strip will experience a variable magnetic field because it is moving into it, and so we induce a voltage across the coil. But this time we don’t have a solenoid with many windings, we just have one coil, one winding. So the voltage induced ad the output will be very low → mic characterized by a very low sensitivity.

So the advantage with respect to the previous one is that the coil is very light, **so it doesn’t suffer from resonant issues**, but the sensitivity is so low that we cannot use normal amplifiers, because amplifiers based on opamp have a too high noise. Hence typically we use the step-up transformer.

Step-up transformer

The ribbon is connected to the primary of the transformer that is coupled with the secondary of the transformer, and the ration is conventionally 1:37 → we have an amplification of 37 of the signal.
This amplification is without noise, because inductances have no noise related. However, the disadvantage is that they cannot be integrated in CMOS circuitry.
The step-up transformer is like an amplifier but with no noise, just amplification.

- flat and resonance-free frequency response
- very low sensitivity → step-up transformer

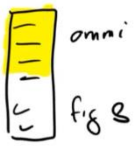


Very low noise amplifier
Standard ratio → 1:37

Ribbon mic characteristics

With ribbon mic we can implement either figure8, omnidirectional or cardioid mics, depending if we expose only one side or both the sides of the ribbon to the sound. For the cardioid we usually expose one part only to front face, the other to both front and back face.

- Figure-8
  → both front and rear sides of the ribbon exposed to sound
- Omnidirectional
  → only front side of the ribbon exposed to sound
  (rear portion terminated with acoustically absorptive material)
- Cardioid
  → only a portion of the ribbon is terminated at the rear,
  the remaining portion is exposed o sound on both sides



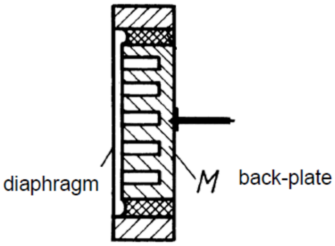
CONDENSER MICROPHONES

They are mics that are capacitive sensors. One plate of the capacitor is the diaphragm itself.
In this case the amplitude of the vibration, so the variation of the capacitance, depends on the displacement of the particle.

Basic construction:

- diaphragm with a thickness of 1÷10 µm made of metal or metallized plastic
- perforated, electrically conductive oppositely-charged electrode (backplate)

→ Impinging sound waves move the diaphragm and change its distance from the back-plate and thus the capacitance of the air-dielectric capacitor formed by the diaphragm and the back-plate.



particle displacement transducer:

$$\xi = \frac{v}{\omega} = \frac{p}{\omega \cdot z}$$

Also in condenser mics, depending on the fact that we expose to the sound only the front face or both the faces, we can get omnidirectional, figure8 or cardioid characteristics.

- Omnidirectional
  → only front side exposed to sound (pressure transducer)
  "high-tuned": diaphragm resonance at mic's upper cut-off frequency
  (in order to operate in the rising portion of the resonance curve)

- Figure-8
  → both front and rear sides exposed to sound (pressure gradient)
  (back-plate is drilled all the way through)
  "mid-band tuned": pressure gradient and displacement relationship with pressure compensate each other

- Cardioid
  → a- only a portion of the back-plate is provided with through holes
  b- back-plate introduces time delay and low pass filter

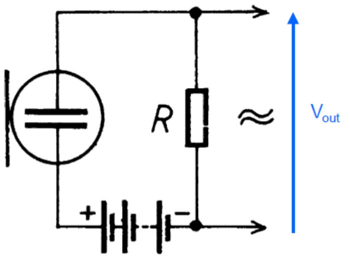
DC bias

How to readout these circuits for condenser mics?

Considering the capacitor that represent the capsule of the microphone, this capacitance will change over time → C(t). then we can measure the output voltage Vout that will depend on the variation of the capacitor.
To see an output, we need a biasing voltage on the capacitor, and this is the main difference between condenser mics and dynamic mics. Indeed, in dynamic mics we need a permanent magnet but we don't need any electrical biasing, because the voltage is induced due to the Faraday law.
Here we need to bias the circuit, and we can consider the charge across the capsule Q to be almost constant. This is true if we work at frequencies higher than the pole of the circuit that is given by C*R.

If we are at higher frequencies than the pole, we are modifying so fast the direction of the current that the current is not able to charge or discharge in a sensible way the capacitance, so we can say that Q is almost constant.

We will make the circuit in a way that the pole C0*R is lower than the audible range. If e.g. we use 20Hz that is in the audible range, we need to size R > 100MOhm, assuming that C0 = 20 – 1000pF.



E_0 = 40 – 200 V

Capsule biased with constant Q in the frequency band:

$$f > \frac{1}{2\pi C_0 R}$$

C_0 = capsule capacitance (20pF-100pF)

R = typically > 100 MΩ (to have f^{min}=20Hz)

With sound pressure:

$$C = C_0 + c(t)$$

$$V_{out}(t) = E_0 \frac{c(t)}{C_0}$$ x

indeed: $Q_0 = C_0 E_0$ $\quad Q = C(E_0 - V_{out}(t))$

$Q_0 = Q \rightarrow C_0 E_0 = (C_0 + c(t))(E_0 - V_{out}(t))$

We can say that the capacitance value C will have a constant part C0 plus a small signal c(t) that considers the variations due to the pressure.
Without any sound pressure, we have that the charge stored by the capacitor is Q0 = C0*E0, so simply charge multiplied by the voltage.

Instead, when we have the sound pressure, Q = C*(E0 – Vout(t)), where E0 - Vout(t) is the voltage across the capacitor. Since we are in the assumption that the Q is almost constant, after the pole, we can equalize the two equations. By doing so we will get the final equation x.

In order to obtain this equation x **I've considered that Vout is much smaller than E0**. This is the reason why E0 must be a very big value, in the order 40 – 200V → condenser mics must be biased with a very high voltage.

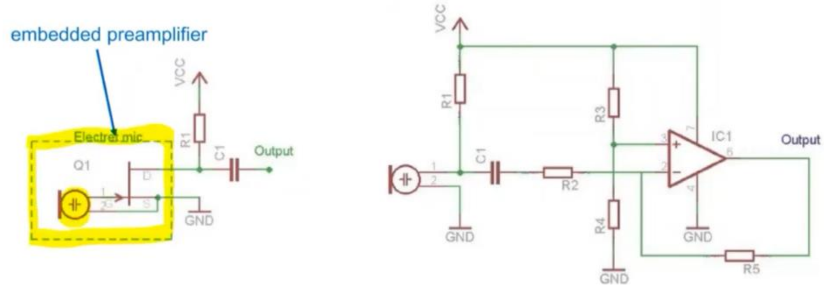
Electret mic – DC bias

In order to overcome the previous issue, electret mics can be used. They are mics in which the back plate is made by a permanently polarized material. This means that the material that makes the plate of the capacitor is always charged and able to keep this charge.

Permanently polarized foil membranes, using materials that can accept and maintain electrical charges (e.g., Teflon).

To incorporate the negative charge carriers, the film is subjected to electron bombardment.

The foil is mounted on the surface of the back electrode, and the diaphragm can thus be realized using the standard materials

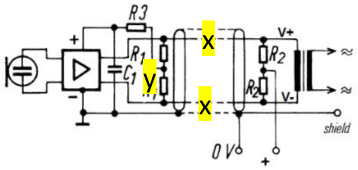


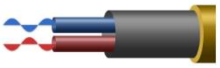
In this case we will have the diaphragm that will oscillate and the other plate permanently polarized. In order to charge these materials, typically they are electron bombarded (negative polarized).
The typical capsule of an electret mic is the yellow one in which we have the capacitor that now doesn't need a DC biasing (it has always its charge Q) and then a preamplifier that is simply a MOS transistor in which the source is connected to ground and the drain to the external circuit. The gate is connected to the output of the capacitor. On the right we have a more complex implementation for the yellow capsule.

Phantom power

- used to bias condenser microphones
- it doesn't disturbs dynamic microphones
- typical voltages: 12V, 24V, 48V
- balanced cables

XLR connectors





Balanced cables:
identical input and output impedances

XLR are connectors typically used for mics both to bias condenser mics but also for dynamic mics. They are made in a way that if we change the type of mic they won't disturb the dynamic mics. They use two differential cables and one that is the ground. They are also able to provide the power supply in case of condenser mics.

X are the two differential output signals. This signal is always centered around the biasing voltage we want to provide (intermediate signal y). In this way we can use it also to bias the condenser mic.

x
x
x