Compendium of Medical Physics, Medical Technology and Biophysics

for students, physicians and researchers

By Nico A.M. Schellart

Department of Biomedical Engineering and Physics
Academic Medical Center
University of Amsterdam
Amsterdam

Version of Decembre 2009, 434 pages.

Introduction

This compendium describes from the physical point of view in simple concepts the principle of working of some 120 methods, techniques, apparatus and equipment used in medical practice. These subjects, divided in about 200 chapters, grouped in 12 sections (such as ‘’Transport”, “Vision”), have been written for especially students in medicine, medical biology, medical information and other medical disciplines. However, the text is also readable for clinicians and other academic physicians, and moreover for paramedical scientists, engineers and technicians with a higher education.
This compendium gives for some subjects more explanation than one would expect in a typical compendium. For extensive explanation, especially when it respects derivation of equations the reader is directed to textbooks and other publications (journals, the internet).

The subject descriptions comprise three parts.
The first one, “**Principle**” gives the basic principles.
‘’**Application**” is devoted to medical applications, but often also gives a very comprehensive description of applications in science, technology and daily live. In many cases “Application” does not literally consider a method, but the physical behavior of a biomaterial, organ or tissue governed by the described working principle.
A more detailed description of “Principle” can be found in “**More Info**”. In general, this last part is also readable for students in the above mentioned disciplines, but in other cases it gives details and occasionally equations that require more knowledge of physics and mathematics.

The compendium has not yet been finished (various subject descriptions are still under construction).
Underlined blue text fragments refer to chapters (some are in preparation).

To write the compendium innumerous sources, especially in the internet, papers in biomedical journals and textbooks are scrutinized. Important ones are given at the end of the subject descriptions. Many text fragments, figures and tables of the Wikipedia Free Encyclopedia were used. I greatly acknowledge all the concerning institutions and their authors for the free use of their material.

This compendium (a slightly shorter version with a different lay-out) can also be found at other websites:

http://onderwijs1.amc.nl/medfysica/compendium.htm (normal size)
http://onderwijs1.amc.nl/medfysica/compendium2.htm (PDA format)

I hoop that this electronic book meets a need, since comprehensive books in this field accessible for non-physicists are scarce.
I would very much like to obtain comments about obscurities, errors, etc. in order to improve this publication (n.a.schellart@amc.uva.nl, telefax +31-20-6917233, dept. of Biomedical Engineeing and Physics, PO Box 22660, 1100 DD, Amsterdam).

What is new in the last 2009 version?

Particle-X-ray interaction

Many rewordings and other improvements.

Nico A.M. Schellart, Decembre 2009.

Contents

SI units and derivatives, abbreviations, symbols, constants, variables

SI units and derivatives

In this manuscript the basic SI (Système Internationale) units (or their derivatives) are preferably used.
For literas symbol L is used instead of the official symbol l, to avoid confusion with 1 (one).

The seven basic SI units

Measure	*Symbol*	*Name*
Length	m	meter
Time	s	second
Mass	kg	kilogram
Current	A	ampère
Temperature	K	Kelvin
Quantity of matter	mole	mole*
Light intensity	cd	candela

*1 mole gas at 1 bar and 0 ^{o}C comprises $6.02 \cdot 10^{24}$ particles and has a volume of 22.7 L

Additional SI units

Measure	*Symbol*	*Name*
Planimatric angle	rad	radian
Spatial angle	sr	steradian

SI-units with own name and symbol

Measure	*Symbol*	*Name and/or Dimension*
area		m^2
volume		m^3
volume	L	$\equiv 0.001\ m^3$
frequency	Hz	1/s
force	N	newton $\equiv$ $kg \cdot m \cdot s^{-2}$)
energy, work	J	joule = N·m)
pressure	Pa	pascal $\equiv N/m^2$, 1 bar $\equiv 10^5$ Pa
power	W	watt $\equiv$ J/s)
electric charge	C	coulomb $\equiv$ A.s
electric voltage	V	W/A
electric capacity	F	farad $\equiv$ C/V
electric resistance	Ω	V/A
electric conductance	S	siemens $\equiv 1/\Omega$
magnetic flux	Wb	weber $\equiv$ V/s
magnetic inuction	T	Wb/m^2
flow of light	lm	lumen $\equiv$ cd.sr
luminous strength	lx	lux = lm/m^2
radioactivity	Bq	becquerel $\equiv$ 1/s
absorbed dose	Gy	gray $\equiv m^2/s^2$

Decimal prefixes

prefix	symbol	multiplication factor
exa	E	10^{18}
peta	P	10^{15}
tera	T	10^{12}
giga	G	10^{9}
mega	M	10^{6}
kilo	k	10^{3}
hector	h	10^{2}
deca	da	10
deci	d	10^{-1}
centi	c	10^{-2}

milli	m	10^{-3}
micro	μ	10^{-6}
nano	n	10^{-9}
pico	p	10^{-12}
femto	f	10^{-15}
atto	a	10^{-18}

notice that $1.\ 10^{-6}$ kg is not written as 1 μkg but as 1 mg

Abbreviations, units, constants and variables (Some very occasionally applied ones are not listed. The numerical value of many constants of nature are approximations.)

a	length (m)
a	acceleration ($m \cdot s^{-2}$)
atm	1 physical atmosphere = 101325 Pa; often 1 atm is rounded to 1 bar
A	age (year), area (m^2), absorption or extinction
A-mode	amplitude mode
A_N	numeric aperture
α	Wien's displacement constant for frequency (≈ 2.821), heat convection coefficient $\alpha = 1.35\ (\Delta T/H)^{1/4}$, thermal diffusivity = $\lambda/(\rho \cdot c_p)$, Womersley number
b	Wiens displacement constant for wavelength ($2.898 \cdot 10^{-6}$ nm·K)
bar	≡ 10^5 Pa. The pressure of 10 m H_2O at 293 K is 0.9807 bar
B	magnetic field strength ($N \cdot Amp^{-1} \cdot m^{-1}$)
B	angle
BEE	basal energy expenditure (J/h)
BF	body fat (%)
BMI	body mass index, mostly W/H^2 (kg/m^2)
BOLD	blood oxygenation level dependent, fMRI signal
Bq	one Bq is one radioactive decay of an atomic nucleus per second
BTPS	body temperature pressure saturated, gas volume standardization
β	compressibility, thermal expansion coefficient (1/K)
c	specific heat coefficient (at 0 °C), speed of light and sound (m/s), capacitance (F)
c	concentration (mol/L)
c_p	specific heat coefficient at constant pressure (J/(kg·K)
C	heat transport by convection (W), degree Celsius, Coulomb (= 1 A·s or 1 F·V), capacitance (F)
C	compliance ≡ 1/*E* (1/Pa), (for hollow subjects L/Pa)
$\boldsymbol{C}_{L,dyn}$	dynamic lung compliance
$\mathbf{C}_V$	volume of dissolved gas at 1bar per volume of liquid (L_n/L)
Cal	calorie (1 Cal ≡ 4.185 J)
CCD	chrage-coupled device
CEU	contrast enhanced ultrasound
γ	c_p/c_v ratio (ratio of specific coefficients of heat capacity with constant p and constant V), surface tension (N/m)
d	diameter (m), distance (m)
dB	decibel
dyn	dynamic (subscript)
D	diameter (m), distance (m), diffusion constant ($m^2 \cdot s^{-1} \cdot bar^{-1}$)
DC	direct current
Δ	difference
Δp_{tp}	transpulmonal pressure difference (Pa)
ΔH_x	specific evaporation heat of fluid x (kJ/kg)
Δp	pressure difference over the airways system from mouth to alveoli (Pa)
Δp'	Δp plus pressure difference over mouth piece and demand valve (Pa)
e	elementary charge, being $1.602 \cdot 10^{-19}$ C
el	elastic (subscript)
E	energy, work (J), heat transport by evaporation (W), electrical potential
E	Young's modulus or modulus of elasticity or stiffness (the reciprocal of compliance) (Pa), electric field strength (N/C or V/m)
E	elastance, the elasticity modulus for hollow subjects (Pa/L)
ECG	electrocardiogram
EPR	electron paramagnetic resonance
ERV	expiratory reserve volume (L)
ESR	electron spin resonance
ϵ	(relative) dielectric constant, emittance factor of radiation, tensile strain (= relative extension), remaining error (statistics)

ϵ_0 dielectric constant of vacuum, $\epsilon_0 = 8.854 \cdot 10^{12}$ F/m
η dynamic viscosity coefficient (Pa·s or Poise; 1 Poise = 0.1 Pa·s), sometimes as μ
η_{air} dynamic viscosity coefficient: $17.1 \cdot 10^{-6}$ (Pa·s)
Θ angle
Φ_m total metabolic power (W)
f friction coefficient (N·s/m), focal length (m)
fMRI functional MRI
F force (N)
F Faraday's constant, $F = 96.4853$ C·mol^{-1}
FE specific fraction of expired air
FEF forced expiratory flow
FEV1 forced expiratory volume in 1 s (L/s)
FCIP fluorescent calcium indicator proteins
FI specific fraction of inspired air
FRC functional residual capacity (L)
g Earth gravitational acceleration, 9.8067 m/s² , electrical conductance (mho, Ω^{-1})
G shear modulus or modulus of rigidity (Pa)
GPF green fluorescent protein
Gr Grashof number
Gy One Gy (gray) is the absorption of 1 J of radiation energy by 1 kg of matter: 1 Gy = 1 J/kg = 1 m^2·s^{-2}
h Planck's constant ($6.626 \cdot 10^{-34}$ Js), height (m), hour
H height (m)
H enthalpy (J)
HBO hyperbaric oxygen
HPLC high performance liquid chromatography
HR heart rate (beat/min)
Hz frequency in Hertz (s^{-1})
i electric current (A), numerical index, $\sqrt{-1}$
I spectral radiance (J·s^{-1}·m^{-2}·sr^{-1}·Hz^{-1})
I intensity of radiation
IR infra red
J sound intensity (W/m^2)
k constant, k is Boltzmann's constant, $1.3807 \cdot 10^{-23}$ J/K or $8.617\ 3 \cdot 10^{-5}$ eV/K
k extinction coefficient, heat conductivity constant
K Kelvin, bulk modulus (Pa), constant
l length, distance (m)
L length (m), self inductance (Henry), lung (subscript)
LED light emitting diode
λ wavelength (m), heat conduction coefficient (W/(m·K), length constant (1/e decrement)
m mean (arithmetic), molecular mass (relative mass with respect to the mass of one proton)
m mass (kg)
max maximal (subscript)
min minute
M magnification (in angle or size)
MI magnetic resonance imaging
MR magnetic resonance imaging
MRI magnetic resonance imaging
MVV maximal voluntary ventilation (L)
MVV_{30} MVV with 30 breaths/min (L)
μ Poisson ratio $\equiv -\Delta d \cdot d^{-1} / \Delta L \cdot L^{-1}$
n integer number, exponent, particle concentration (m^{-3}), refractive index
n number of kmoles
N number of particles
N_A number of Avogadro $\equiv$ number of particles in 1 mole: $6.0225 \cdot 10^{23}$ mol^{-1}
ν kinematic viscosity $\equiv \eta/\rho$ (is dynamic viscosity/density) (m^2/s), frequency (Hz)
O area (m^2)
p pressure (bar, Pa)
p_{tp} transpulmonary pressure $\equiv \Delta p + \mathbf{E} \cdot V_{tidal}$ (Pa, cmH_2O)
P pressure (bar, Pa), power (W)
P poise, a unit of the dynamic viscosity (= 0.1 Pa·s)
Pa Pascal, 1 Pa $\equiv$ 1 N/m^2
PET positron emission tomography
ppb parts per billion
ppm parts per million
Pr Prandtl number
P_{ST} pressure due to surface tension

q	charge of particle (single charge is $1.60 \cdot 10^{-19}$ C), Bunsen's absorption coefficient (L_{gas}/L_{liquid} or g_{gas}/L_{liquid})
Q	electric charge (C), unit of flow (e.g. heat)
r	radius (m), electric resistance (Ω)
rad	radian
R	universal gas constant (= 8315 J/(kmol·K)), electric resistance (Ω)
R	heat transport by radiation (W), reflection coefficient
R	flow resistance
Ra	Rayleigh number
R_{aw}	resistance of airways ($Pa \cdot L^{-1} \cdot s^{1}$)
Re	Reynolds number
RMV	respiratory minute volume (L/min)
RV	residual (lung) volume (L)
ρ	specific density at 0 °C (kg/m^3)
ρ_{air}	specific density of air: 1.29 kg/m^3 at 0 °C
radian	$1/2\pi$
s	Laplace operator (=iω)
S	energy state of electron orbital
SaO_2	arterial Hb bounded with SaO_2
SPL	sound pressure level, 1 dB SPL = 20 μPa
stat	static (subscript)
STPD	standard temperature and pressure dry, gas volume standardization
Sv	the sievert measures the biological effect of radiation: 1 Sv = 1 J/kg = 1 $m^2 \cdot s^{-2}$
σ	constant of Stefan-Boltzmann; $5.6704 \cdot 10^{-8}$ $W/(K^4.m^2)$, tensile stress (Pa)
t	time, temperature in °C
T	temperature in Kelvin (K); 273.15 K ≡ 0 °C
TLC	total lung capacity (L)
τ	time constant (s)
Φ_m	total metabolic power (W)
U	electrical potential (V), internal energy (J)
UV	ultra violet
v	velocity (m/s)
vis	viscous (subscript)
ν_{max}	peak frequency (Hz) of radiation
V	volume (m^3 or L), voltage (V)
V_{tidal}	volume of inspiration (L)
$\dot{V}$	volume flow (m^3/s or L/s)
VC	vital capacity (L)
VEGF	vascular endothelial growth factor
VO_2	aerobic capacity (mL/kg or mL/(min.kg))
VOR	vestibule-ocular reflex
W	weight (kg), wall (subscript)
ω	angular frequency (radians/s)
Ω	ohm
x	distance (m)
Z	impedance, e.g. electric (V·s/A), acoustic (Pa·s/m)
ζ	resistance coefficient of flow through curvature or bend

Systems and basic concepts

Autoregulation

Principle

In biomedicine, autoregulation is the process of changing the principal system parameter to cope with a change in conditions. It is most known for the circulatory systems, but macromolecular systems can also have autoregulation.

Application

The principle is applied in for instance the cerebral, systemic, muscular and renal circulation and the physiology of the heart. Cerebral autoregulation will be discussed here,

Cerebral circulation
The cerebral flow of blood is autoregulated by altering cerebral vessel diameters. Proper cerebral perfusion pressure (CPP) is achieved by dilatation (widening) of arterioles that lowers pressure and creates more room for the blood, or constriction to raise pressure and lower cerebral blood volume. Thus, changes in the body's overall blood pressure (see also Blood pressure: measurement) do not normally alter cerebral perfusion pressure drastically, since the vessels constrict when systemic blood pressure is raised and dilate when it is lowered. Arterioles also constrict and dilate in response to different chemical concentrations. For example, they dilate in response to higher levels of CO_2 in the blood.
Systemic circulation
The capillary bed of an organ usually carries no more than 25% of the maximal amount of blood it could contain, although this amount can be increased through autoregulation by inducing relaxation of smooth muscle. The capillaries do not possess this smooth muscle in their own walls, and so any change in their diameter is passive. Any signaling molecules they release (such as endothelin for constriction and NO for dilation) act on the smooth muscle cells in the walls of nearby, larger vessels, e.g. arterioles.
Muscular circulation
Myogenic autoregulation is a form of homeostasis (see Bioregulation, homeostasis). It is the way arteries and arterioles react to an increase or decrease of blood pressure to keep the blood pressure within the blood vessel constant. The smooth muscle of the blood vessels will react to the stretching of the muscle by opening ion channels which cause the muscle to depolarize leading to muscle contraction.
The effect of this is to reduce Blood flow through the blood vessel as the volume of blood able to pass through the lumen is significantly reduced. Alternatively, when the smooth muscle in the blood vessel relaxes due to a lack of stretch, the ion channels will close resulting in vasodilatation of the blood vessel. This increases the rate of flow through the lumen as its radius is greater.
Renal circulation
This system is especially significant in the kidneys, where the glomerular filtration rate is particularly sensitive to changes in blood pressure. However, with the aid of the myogenic mechanism it allows the glomerular filtration rate to remain very insensitive to changes in human blood pressure.

More Info

Cerebral autoregulation
When pressures are outside the range of 50 to 150 mmHg, the blood vessels' ability to autoregulate pressure is lost, and cerebral perfusion is determined by blood pressure alone, a situation called pressure-passive flow. Thus, hypotension (inadequate blood pressure) can result in severe cerebral ischemia in patients with conditions like brain injury, leading to damage (ischemic cascade).
Other factors that can cause loss of autoregulation include free radical damage, nervous stimulation, and alterations of the partial CO_2 and O_2 pressure. Even in the absence of autoregulation a high pCO_2 can dilate blood vessels up to 3.5 times their normal size, lowering CPP, while high levels of oxygen constrict them. Blood vessels also dilate in response to low pH. Thus, when activity in a given region of the brain is heightened, the increase in CO_2 and H^+ concentrations causes cerebral blood vessels to dilate and deliver more blood to the area to meet the increased demand. In addition, stimulation of the sympathetic nervous system raises blood pressure and blocking lowers pressure.

Colloid

Principle

Colloids are all types of, generally, liquid-like mixtures existing of a solvent and a (semi-)macro-molecular substance. More precisely, a colloidal solution or colloidal dispersion is a type of mixture intermediate between a *homogeneous mixture (*particles < 1 nm, also called a solution*)* and a *heterogeneous mixture (*particles > 1 µm). Also its properties are intermediate between the two.
Typical membranes restrict the passage of dispersed colloidal particles more than for ions or dissolved molecules. Many familiar substances, including butter, milk, cream, aerosols (fog, smog, and smoke), asphalt, inks, paints, glues, and sea foam are colloids. In a colloid, the size of dispersed phase particles range from 1 nm to 1µm. Dispersions where the particle size is in this range are referred to as colloidal aerosols, colloidal emulsions, colloidal foams, or colloidal suspensions or dispersions. Colloids may be colored or translucent because of the Tyndall effect (see Light: scattering), which is the scattering of light by particles in the colloid.
When the dispersed substance is a liquid or solid, the mixture is a Suspension. When both phases are liquid, than the mixture is an Emulsion.
Since there are 3 basic aggregation states one would expect 9 kinds of solids but since gas is always soluble in another gas, 8 types remain. Table 1 gives these combinations.

Table 1 Types of colloids

		Dispersed Medium		
		Gas	Liquid	Solid
Continuous Medium	Gas		**Liquid Aerosol** Examples: fog, mist	**Solid Aerosol** Examples: smoke, dust
	Liquid	**Foam** Examples: whipped cream, nose spray	**Emulsion** Examples: milk, hand cream, salve	**Sol** Examples: paint, pigmented ink, blood
	Solid	**Solid Foam** Examples: aerogel, styrofoam, pumice	**Gel** Examples: gelatin, jelly, cheese, opal	**Solid Sol** Examples: cranberry glass, ruby glass

Applications

They are numerous, as Table 1 indicates, also in medicine.

More info

Interaction between colloid particles
The following forces play an important role in the interaction of colloid particles:
- Excluded Volume Repulsion This refers to the impossibility of any volumetric “overlap” between hard particles.
- Electrostatic interaction Colloidal particles often carry an electrical charge and therefore attract or repel each other. The charge and the mobility of both the continuous and the dispersed phase are factors affecting this interaction.
- Van der Waals forces This is due to interaction between two dipoles (permanent or induced). The van der Waals force is always present, is short range and is attractive.
- Entropic forces: According to the second law of thermodynamics, a system progresses to a state in which entropy (see Thermodynamics: entropy; irreversible increase of disorder at microscopic scale) is maximized. This can result in effective forces even between hard spheres.
- Steric forces between polymer-covered surfaces or in solutions containing non-adsorbing polymer can modulate interparticle forces, producing an additional repulsive steric stabilization force or attractive depletion force between them.

Stabilization of colloid suspensions
Stabilization serves to prevent colloids from aggregating. Steric and electrostatic stabilization are the two main mechanisms for colloid stabilization. Electrostatic stabilization is based on the mutual repulsion of like electrical charges leading to very large charge double-layers of the continuous phase around the particles. In this way the specific density differences are so small that buoyancy or gravity forces are too little to overcome the electrostatic repulsion between charged layers of the dispersing phase.
The charge on the dispersed particles can be observed by applying an electric field: all particles migrate to the same electrode and therefore must all have the same sign charge.

Destabilizing a colloidal suspension

Unstable colloidal suspensions form flocks as the particles aggregate due to interparticle attractions. This can be accomplished by a number of different methods:

- Removal of the electrostatic barrier that prevents aggregation of the particles (by addition of salt to a suspension or changing the pH). This removes the repulsive forces that keep colloidal particles separate and allows for coagulation due to van der Waals forces.
- Addition of a charged polymer flocculant (bridging of individual colloidal particles by attractive electrostatic interactions).
- Addition of non-adsorbed polymers called depletants that cause aggregation due to entropic effects.
- Unstable colloidal suspensions of low volume fraction form clustered liquid suspensions wherein individual clusters of particles fall to the bottom or float to the top, since Brownian motion become too small to keep the particles in suspension. Colloidal suspensions of higher volume fraction can form colloidal gels with viscoelastic properties. These gels (e.g. toothpaste) flow like liquids under shear but maintain their shape when shear is removed. It is for this reason that toothpaste stays on the toothbrush after squeezing out.

Colloidal particles are large enough to be observed by Confocal microscopy.
Just as a solution, a colloid has an osmotic effect (see Osmosis).

References

Wikipedia/Colloid (main source of knowledge).

Compliance (hollow organs)

Principle

Elastance is a measure of the tendency of a hollow organ to recoil toward its original dimensions upon removal of a distending or compressing force. It is the reciprocal of *compliance*.

Compliance of a hollow organ is calculated using the following equation:
$C = \Delta V/\Delta P$,
where ΔV is the change in volume, and ΔP is the change in pressure. In SI-units, its dimension is dm^3/Pa.
It should not be mixed up with the one-dimensional compliance, or better the reciprocal, i.e. modulus of elasticity as defined in the theory of strength of materials, see Elasticity and Hooke's law, Elasticity 1: elastic or Young's modulus and Tensile strength.

Application

Cardiovascular system
The terms elastance and compliance are of particular significance in cardiovascular physiology. Specifically, the tendency of the arteries and veins to stretch in response to pressure has a large effect on perfusion and blood pressure.
Veins have a much higher compliance than arteries (largely due to their thinner walls). Veins, which are abnormally compliant, can be associated with edema. Pressure stockings are sometimes used to externally reduce compliance, and thus keep blood from pooling in the legs. An extreme application is the use of pressure trousers or suites by astronauts.
Lungs
Compliance of the lungs is an important measurement in pulmology. Fibrosis is associated with a *decrease* in pulmonary compliance. Emphysema is associated with an *increase* in pulmonary compliance.

More Info

For the cardiovascular system, see Elasticity of the aorta, Blood pressure: models and Windkessel model.
For the lungs, see Lung gas transport 2, pressure, volume and flow and Lung gas transport 2, resistance and compliance.

Electrical resistance, capacitance and self-inductance

Principle

When a current is flowing through an electrical resistance, capacitance or inductance, the relation between current and voltage may be changed by these elements. Their action can be described as follows.

Resistance $R=V/I$ or $V=IR$, (1)

This is Ohms law, where V is voltage (V), I is current (Amp) and R is resistance (Ohm). In a circuit with only resistors, all current and voltages are in phase. Halving the resistance means doubling the current when the voltage over the resistance remains the same. A daily analogue is taking a second water hose to sprinkle the garden. Then total flow of water is doubled since the pressure of the water supplying system is constant.
Resisters dissipate heat, caused by electrical “friction”. This can be expressed as power P (in Watt):

$$P = IV = I^2R = V^2/R \text{ (W)}. \quad (2)$$

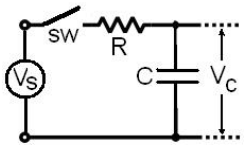


Fig. 1 Circuit to charge the capacitance C of the capacitor with source voltage V_S. sw is switch.

Capacitance $dV/dt = (1/C)dQ/dt = I/C$, (3)

where C is the capacitance (F, farad), Q the instantaneous charge (in coulombs, Q) of the capacitor and t time. When Q is constant in time (a charged capacitor, also a battery) it holds that $C = Q/V$.
For a circuit with a constant voltage source (so called DC voltage from direct current) and comprising of a resistor and capacitor in series (Fig. 1), the voltage across the capacitor cannot exceed the voltage of the source. Suppose that before the start the circuit is switched off and the charge of the capacitor is zero. Next, the circuit is switched on (closed). Then a current provided by the voltage source is charging the capacitor until an equilibrium is reached: the voltage across the capacitor V_C finally becomes the source voltage V_S and the charging current becomes zero. For this reason, it is commonly said that capacitors block direct current (DC). At the start the current is maximal and at the end of the charging process minimal, whereas the voltage then reaches the highest values. So, one could say that the voltage over the capacitor lags the current. The current diminishes exponentially with a time constant (see Halftime and time constant and Linear first order system) $\tau = RC$.
When the voltage of the source changes periodically the current will also change periodically, called alternating current (AC). Then the current through a capacitor reverses its direction periodically. That is, the alternating current alternately charges the plates: first in one direction and then in the other. With the exception of the instant that the current changes direction, the capacitor current is non-zero at all times during a cycle. For this reason, it is commonly said that capacitors "pass" AC. However, at no time do electrons actually cross between the plates, unless the dielectric breaks down. Such a situation would involve physical damage to the capacitor.
With sine waves as input voltage, the voltage across the capacitor V lags the current 90^o, since the voltage is proportional to the integral of the current (see (3)). That is, the voltage and current are 'out-of-phase' by a quarter cycle. So, the phase is -90^o, the minus sign denotes *lag* (notice $\sin(x - 90) = \cos x$).
The amplitude of the capacitor voltage depends on the amplitude of the current divided by the product of the frequency f of the current with the capacitance, C, so $V \sim I/fC$.
For a more formal description we need the quantity impedance (Z), which reflects in a network the action on amplitude and phase. It is the ratio of the voltage (of unit amplitude and phase zero) across a circuit element over the current through that element. It comprises a real part, a resistance, and imaginary part, due to capacitance (or inductance, the action of a coil).

$$Z = R + jX. \quad (4)$$

For a capacitor, the impedance is given by:

$$Z_c = V_c / I_c = Vj/(2\pi fC) = Vj/(\omega C) = -jX_c, \quad (5)$$

where f is the frequency (Hz) of a sinusoidal voltage, ω (= 2πf) is the angular frequency and X_c (= –1/(ωC) is called the *capacitive reactance*, the quantity denoting the imaginary part of the impedance. j is the imaginary unit (j = √–1, in mathematics and often in physics denoted by i). It denotes the phase *lag* of 90° between voltage and current. Since a pure capacitance has no resistance, the real part of the impedance is not present in (5).
Ideal capacitors dissipate no energy.

Inductance $\int V dt = LI,$ (6)

where L is self inductance. The action of an inductance can be called reciprocal of that of a capacitance with respect to amplitude: the *inductive reactance* is $X_L = \omega L$. Consequently, an inductance (in hardware a coil) blocks AC (high frequency) due to its virtual high resistance. It passes DC, since then $X_c = 0$. With respect to phase, its action causes sign reversal. So, there is a *phase lead* of 90°. Equation (6) shows that directly since V is proportional to the derivative of I.
Ideal inductances dissipate no energy, but the intrinsic resistance of coils (the winding of the wire) of coarse do, the reason why operating coils are always warm.

The formal definition of inductance is:

$$L = \Phi / I. \quad (7)$$

where Φ is magnetic flux (in H, henry = weber/ampere). In linear systems (6) is sufficient, but when electromagnetic fields are generated the theory of electromagnetic fields are of importance and so (7). The description of this theory is beyond the scope of this compendium, with exception of the small contribution Lorentz force.

In conclusion:

$$Z_R = V_R / I_R; \quad (8a)$$
$$Z_C = V_C / I_C = -1/j\omega C; \quad (8b)$$
$$Z_L = V_L / I_L = j\omega L. \quad (8c)$$

In the complex Z-plane it can be visualized as in Fig. 2.

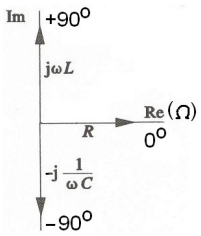


Fig. 2. Impedances in the complex impedance plane.

Application

It needs no further explanation that in all electric instruments etc. resistors and capacitors are applied, as individual components or incorporated in integrated circuits.
Coils can be applied in for instance specific electronic filters (see Linear first order system, Linear second order system and System analysis), in dynamo's to generate current, in transformers, for spark ignition, in mechanical current and voltage meters and in liquid current meters. Further, they are applied in (medical) equipment to refract or change electromagnetic radiation (Mass spectrography and MRI machines, particle accelerators needed for PET, Electron Spin Resonance machines).

More info

In hardware the resistor is the circuit element to provide the action of resistance. A resistance always gives some inductance: current flowing through a cable produces a small magnetic field. It has also some capacitance. Both are very dependent on the design (material, configuration) of the resistor.
Similar considerations hold for the capacitor, the physical realization of the capacitance. It has some resistance (the dielectricum does not isolate perfect) and there is some self inductance.
In general these imperfections can be avoided by choosing the right type of resistor and capacitor (their numerical value as well as technical design). However, these imperfections generally only play a role with extreme high frequencies.
The coil is the realization of the *self inductance*. It has always a resistance, and generally this is so large that one should consider it in circuit designs and calculations. It also has some capacitance.

Resistors
Resistors, say R_1, R_2 etc., in a parallel configuration have the same potential difference (voltage).

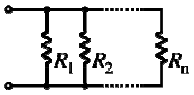


Fig. 2 Resistors in parallel.

Their total equivalent resistance (R_{eq}) is:

$$1/R_{eq} = 1/R_1 + 1/R_2 + .. 1/R_n. \quad (9a)$$

The parallel property can be represented in equations by two vertical lines "||" (as in geometry) to simplify equations. For two resistors,

$$1/R_{eq} = R_1 \,||\, R_2 = R_1R_2/(R_1 + R_2). \quad (9b)$$

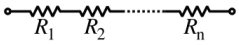


Fig. 3 Resistors in series.

The current through resistors in series stays the same, but the voltage across each resistor can be different. The sum of the individual voltage differences is equal to the total voltage. Their total resistance is:

$$R_{eq} = R_1 + R_2 + .. R_n. \quad (9c)$$

Capacitors
For capacitors in parallel hold:

$$C_{eq} = C_1 + C_2 + .. C_n. \quad (10a)$$

And in series:

$$1/C_{eq} = 1/C_1 + 1/C_2 + .. 1/C_n. \quad (10b)$$

Coils
For reasons given by the electromagnetic field theory, in electric circuits mimicking linear systems, coils are generally not put in series or parallel. The above simple rules about parallel and serial generally do not hold.

Analogues of R, C and L

Mechanical: R_M (Ns/m), mechanical resistance, due to friction;
C_M (m/N), mechanical compliance;
mass (kg).

Fluid dynamics: R_F (Ns/m^5), fluid resistance;
C_F (m^5/N), compliance;
M_F Ns/m^4), inertance.

Acoustical : R_A (Ns/m^5), acoustic resistance;
C_A (m^5/N), acoustic compliance ;
M_A Ns/m^4), acoustic inertance.

Emulsion

Principle

Emulsions are part of a more general class of two-phase systems of matter, called *colloids* (see Colloid). Although the terms colloid and emulsion are sometimes used interchangeably, in emulsion *both the dispersed and the continuous phase are liquid*. Emulsions are also a subclass of Suspension.
An emulsion is a mixture of two immiscible (unblendable) substances. One substance (the dispersed phase) is dispersed in the other (the continuous phase).
Emulsions have a cloudy appearance, because the many phase interfaces (the boundary between two phases) scatter light (see Light: scattering).
Emulsions are unstable and thus do not form spontaneously. Energy input through stirring, etc., or spray processes are needed to form it. Over time, emulsions tend to revert to the stable state of for instance oil separated from water. Surface active substances (surfactants, see Surface tension) can increase the kinetic stability of emulsions greatly so that, once formed, the emulsion does not change significantly over years of storage.
Emulsification is the process by which emulsions are prepared.

Application

In medicine
Emulsions are frequently used in drugs.
In daily life
A large field of application is food and cosmetic industry. Examples of emulsions include oil in water and butter. In butter and margarine (are also suspensions), a continuous lipid phase surrounds droplets of water (water-in-oil emulsion).

More Info

There are three types of emulsion instabilities:
- flocculation, where the particles form clumps;
- creaming, where the particles concentrate towards the surface (by buoyancy or by e.g. centrifugation) of the mixture while staying separated;
- breaking, where the particles coalesce (recombination to form bigger ones) due to lack of shaking and form a layer of liquid.

Emulsifier
An emulsifier, also known as an emulgent or surfactant, is a substance which stabilizes an emulsion. An example of food emulsifiers is egg yolk (where the main emulsifying chemical is the phospholipid lecithin). Proteins and low-molecular weight emulsifiers are common as well.
Detergents, another class of surfactant, chemically interact with both oil and water, thus stabilizing the interface between oil or water droplets in suspension. This principle is exploited in soap to remove grease for the purpose of cleaning. A wide variety of emulsifiers are used to prepare emulsions such as creams and lotions.

Whether an emulsion turns into a water-in-oil emulsion or an oil-in-water emulsion depends on the volume fraction of both phases and on the type of emulsifier. Generally, the so-called Bancroft rule applies: emulsifiers and emulsifying particles tend to promote dispersion of the phase in which they do not dissolve very well. For example, proteins dissolve better in water than in oil and so tend to form oil-in-water emulsions (that is they promote the dispersion of oil droplets throughout a continuous phase of water).

Fourier analysis

Principle

In medicine, in all kinds of ways and for many aims signals are measured from the human body. Signals can be represented as a function of area or space (2-D or 3-D), for example with CT and MRI scans and as a function of the time, as with ECG, EEG (Electroencephalography) and Magnetoencephalography (MEG), but also as function of time and space as with fMRI (see MRI: functional MRI).
We will limit our description to analogue, ongoing time signals. These can be periodic (for example a cosine or ECG) and non-periodic. Non-periodic signals are for example a once occurring pulse or a noisy signal, such as the EEG. This description is limited to the real notation. For the description in complex notation see Fourier transforms.

Often it is important to know which frequencies occur in the signal, e.g. in communication technology but also in medicine. Then the signal is no longer represented in the time domain but in the so-called frequency domain. The theory on which this is based, is the theory of Fourier, which says that each (random) signal can be described by the sum of a series of sine or cosine functions, everyone with its own amplitude and phase (or as the sum of a series of sinus *and* cosine functions, everyone with its own amplitude).
The frequencies of the sine and cosine waves have the ratio 1, 2, 3, 4 etc. and they are called the harmonics. The first harmonic (with ratio 1) is the ground or fundamental frequency. Since any signal comprises this harmonic series of frequencies Fourier analysis is also called frequency analysis.
The term harmonic comes from the term harmonic oscillator. An example is the harmonic movement performed by a weight suspended from a spring. The excursion of the weight as a function of time is described by a sine wave.
Fig. 1 shows an analysis of the blood flow signal.

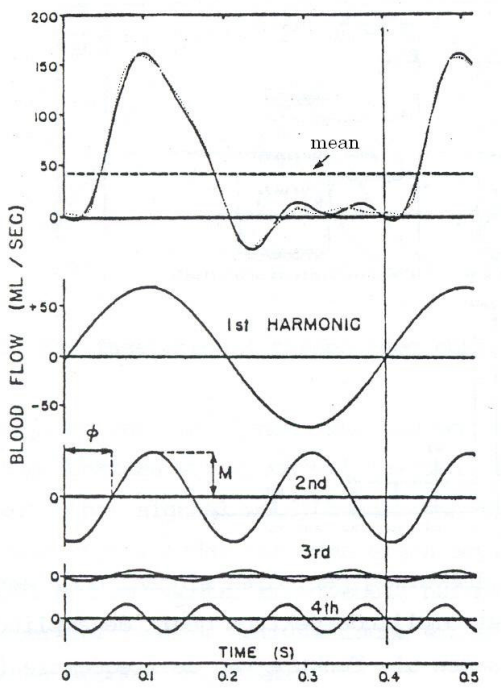


Fig. Fourier analysis of blood flow in a dog lung artery. The upper panel gives the measured signal, indicated by a dotted line and the drawn curve is the mean signal plus the first 4 harmonics.

Fourier analysis and synthesis
Like already said, an important description method is the approach of a signal by the sum of a number of sine and cosine functions. A single sine function is described by its amplitude C, angular frequency ω (ω = 2πf with f the frequency in Hz) and its phase φ (in radians) with regard to another signal of the same frequency (or position in time):

$$x(t) = C\sin(2\pi ft + \varphi). \qquad (1)$$

In Fig. 2 is demonstrated how a number of harmonic sine waves can describe some signal, here a square wave. The fluently drawn line in the four panels is the sum of the solid curve in the previous panel (e.g. in panel a) plus the newly added dotted harmonic (in *b* the third and in *c* the fifth). When this procedure is continued, then a sum signal with the shape of a square wave arises, such as indicated in panel d. The more harmonics are used, the better the Fourier synthesis approaches the square signal. The square wave is a special case of the general function:

$$x(t) = \sum_{k=1}^{\infty} C_k \sin(2\pi kft + \varphi_k)\,, \tag{2}$$

where x(t) is the sum of an infinite number of harmonics, each with its own amplitude C, frequency kf and phase angle φ. The amplitudes C_k together constitute the *amplitude spectrum* (with amplitude versus k) and similarly φ_k yields the *phase spectrum*.
The above makes clear that any signal can be developed in a series of Fourier terms and that a signal can be synthesized with such a series: *Fourier synthesis*.

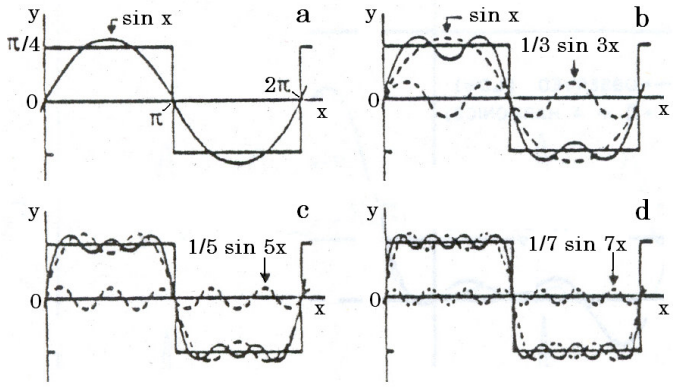


Fig. 2 Fourier analysis and synthesis of a square wave. a. signal fitted with the 1st harmonic with amplitude 1. b. With the first 3 harmonics. The third contributes with amplitude 1/3. The sum signal is the drawn curve, dashed curve around the horizontal axis is the 3rd harmonic and the other dashed curve the sum of the previous panel. c. The same for the first 5 harmonics. d. For the first 7 harmonics. The even harmonics are all zero.

The average value and the Fourier components

The average value of a signal x(t) with a duration T is:

$$x_{mean} = T^{-1}\int_0^T x(t)dt \tag{3}$$

Averaged over one period, sine and cosine waves have an average value of zero. This means that the average of a signal described with equation (2) has also as an average value of zero. (The average of a sum = the sum of the averages). For a random signal the average value x_{mean} is generally not zero but has some value, C_0. This C_0 must be added to (2), yielding:

$$x(t) = C_0 + \sum_{1}^{\infty} C_k \sin(2\pi kft + \varphi_k)\,. \tag{4}$$

Equation 4 describes any periodic signal or function. More Info describes how C_k is calculated.

Application

Fourier analysis, in its digital version, is performed by many PC software packages (MatLab etc.) and spread sheets.

More info

By defining that $\varphi_0 = \pi/2$ (4) becomes the general equation:

$$x(t) = \sum_{0}^{\infty} C_k \sin(2\pi kft + \varphi_k)\ \ , \tag{5a}$$

$$x(t) = \sum_0^\infty C_k \sin(2\pi kft)\cos(\varphi_k) + \sum_0^\infty C_k \cos(2\pi kft)\sin(\varphi_k) \quad (5b),$$

(with the goniometric rule $\sin(a+b) = \sin a \cdot \cos b + \cos a \cdot \sin b$) where C_0 adopts the character of the amplitude of a sinusoid with frequency zero.
Now, the sine of (5a) with phase φ and amplitude C has been written as the sum of a cosine and sin of 2πkft. With Csinφ = A and Ccosφ = B, (5b) becomes:

$$x(t) = \sum_0^\infty A_k \cos(2\pi kft) + \sum_0^\infty B_k \sin(2\pi kft)\,. \quad (6)$$

C_k (*the* amplitude) and φ_k of equation (5) are calculated from:

$C_k = \sqrt{(A_k^2 + B_k^2)}$ and (7a)
$\varphi_k = \arctan B_k/A_k$. (7b)

It is general practice to present A_k along the x-axis and B_k along the y-axis. After applying the theorem of Pythagoras the resulting C_k makes the angle of φ_k with the x-axis. With a vector representation, C_k is absolute value and φ_k the argument.
The coefficients A_k and B_k are calculated as follows:

$$A_k = (2/T)\int_0^T x(t)\cdot\cos(2\pi kft)dt \text{ , and}$$

$$B_k = (2/T)\int_0^T x(t)\cdot\sin(2\pi kft)dt \quad (8)$$

The integral only gives an output A_k or $B_k > 0$ when x(t) comprises a cosine or sine of the same frequency kf. A_k and B_k are independent since the integral of sin*j*acosk*b* gives zero (j and k are the harmonic rank numbers). This property is a condition for so called orthogonality of two functions, here sine and cosine.

The elegance of the Fourier development is that its coefficients agree exactly with the least square fit of the signal and this applies even to each harmonic separately.
When x(t) is written as a cosine series, then formula (5) becomes:

$$x(t) = \sum_0^\infty C_k \cos(2\pi kft + \varphi_k - \pi/2) \quad (9)$$

If $B_k = 0$ for each k, then we call the signal even because the signal can be mirrored with respect to the y-as (amplitude) or for the line parallel with the y-as on t = kT. As $A_k = 0$ (apart from A_0) for each k, then the signal is called odd because there is a so-called point-symmetry at the origin. This holds also for the time instants kT on the x-as. One speaks more commonly of even when f(t) = f(–t), and odd functions when f(t) = –f (t). Polynomials with exclusively terms with even exponents are even functions. Likewise polynomials with exclusively odd exponents are odd functions. However, an arbitrarily signal or function is neither even nor odd. A sine is odd and a cosine even. Also these signals can be written as a as polynomials (with infinite terms). The step function (see Linear first order system) on t=0 is odd (sum of sinuses) and the delta function (sum of cosines on t = 0) is an even function. (The delta function is everywhere 0, except on t = 0 the amplitude is ∞, see Linear first order system).
Looking back once more to the synthesis of the block-signal (Fig. 2), then we see that the number of maxima is equal to the rank number of the highest term that is used for the synthesis. Additionally, the ripples at the sides are larger than those in the middle of the plateau, the phenomenon of Gibbs.

Fourier transform

Principle

Fourier analysis and synthesis are called Fourier (forward) transform and Fourier backward transform respectively when complex notation is used.

Forward transform
The continuous Fourier transform (FT) X(jω) of signal x(t) is defined as:

$$X(j\omega) = \int_{-\infty}^{+\infty} x(t) e^{-j\omega t} dt\,, \qquad (1)$$

where ω = 2πf and f the frequency.
(It can also be considered as a bilateral Laplace transform with s = jω +α, but reduced to s = jω, in order to prevent integrals becoming infinite, so they should be physically realistic.
The modulus (absolute value) is the amplitude spectrum A(ω):

$$A(j\omega) = ((\mathrm{Re}\{ X(j\omega)\})^2 + (\mathrm{Im}\{ X(j\omega)\})^2)^{0.5}\,, \qquad (2)$$

and the argument is the phase characteristic φ(ω) (in radials):

$$A(j\omega) = \mathrm{Im}\{ X(j\omega)/\mathrm{Re}\{ X(j\omega)\} \qquad (3)$$

From A(ω) the power spectral density function S(ω) can be calculated:

$$S(\omega) = \tfrac{1}{2} A(\omega)^2. \qquad (2a)$$

Backward transform
This transform from the frequency to the time domain is:

$$x(t) = (2\pi)^{-1} \int_{-\infty}^{+\infty} X(j\omega) e^{j\omega t} d\omega. \qquad (4)$$

Notice that the backward transform also considers the negative frequencies. This implies that the complex notation is also very appropriate to describe non-periodical and non-deterministic (so noisy or stochastic) signals in a concise notation.

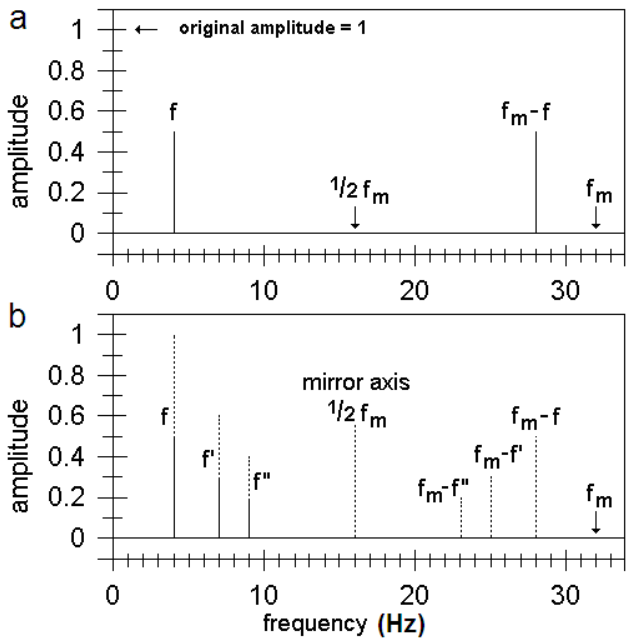


Fig. 1. a. Result of the discrete Fourier transform of a single frequency, here 4 Hz. The outcome at –4 Hz is shifted to 28 Hz. See further the text. b. The same, but now for a signal comprising 3 frequencies, here 4, 7 and 9 Hz. The negative frequencies shifted to the positive side result in amplitudes indicated

by the dashed bars on he right. After folding over half the sample frequency (½f_m) these amplitudes are added directly to those at the right (f, f' or f''), since the phases are the same.

The relation between the complex and real notation (see Fourier analysis) can be seen easily since $e^{j\omega t}$ is constituted of the cosine and sine terms ($e^{j\omega t}$ = cosωt + jsinωt). With the cosine term plotted along the horizontal (real) and the sine term along the vertical (imaginary) axis, a vector representation is obtained. With the frequency going from 0 to infinite the polar diagram of the signal is obtained. This way of presentation is very similar to the polar diagram of a linear system, as discussed in Linear first order system.
The difference with the real notation is that the frequency of the outcome is not discrete but continuous and that the transform goes from – to + infinite for time and for frequencies. After sampling of the signal(with frequency fm) and performing a discrete FT (see **More Info**) the transform yields also an amplitude spectrum for negative frequencies that is the mirror around f = 0. This negative frequencies have the same phase as the positive ones.
With the notion that physically negative frequencies do not exist, this negative frequency is shifted to the positive side over a distance of f_m. Fig. 1a illustrates what happens. It shows the outcome of the FT of a single frequency, which is found at 4 and at 32 – 4 = 28 Hz, both with amplitude 0.5. The sample frequency, f_m is 32 Hz (subscript m comes from the German and Dutch word for sample). Now, the contribution at 28 Hz is folded backward over the frequency 0.5 f_m = 16 Hz and the both amplitudes of 0.5 Hz add up to 1, the amplitude of the original signal. Fig. 1b illustrates what happens with a signal comprising the frequencies f, f' and f''. After mirroring the amplitudes of frequencies f_m –f, f_m – f' and f_m – f'' are added to those of f, f' and f'' respectively, yielding the original amplitudes.
Self evident, for practical applications, also with the complex notation, the signals or spectra should be digitized.

Application

Fourier theory in real notation has only educational significance, since (after sampling) the algorithm is time consuming. The number of mathematical operations is basically N^2 with N the number a samples. The Fourier transform, performed with the Fast Fourier Transform (FFT) algorithm to make a Discrete Fourier Transform (DFT) is usually applied with the number of samples being an integer power of the number two. However, this is not actually necessary, any integer can be used, even primes. The number of mathematical operations is basically NlogN.
Applications are innumerous in science and technology and consequently in medical apparatus and processing of medical recordings as function of time and/or space.

More Info

In mathematics, with DFT a sequence of N complex numbers x_0, ..., x_{N-1} is transformed into the sequence of N complex numbers X_0, ..., X_{N-1} according to:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N} \quad \text{k = 0, 1,...., N-1} \qquad (5)$$

The sequence x is generally a signal in the time or 1D-(space-)domain, and its imaginary part is zero. The sequence X is in the frequency domain (temporal or spatial frequency). As outcome it has generally a real and imaginary part. From both, in the frequency domain, the amplitude (from the absolute value) and the phase (from the argument) of the complex X's can be calculated, similar as equations (2) and (3). In other words X_k comprises the amplitude and phase spectrum of the real input signal x_n.
When N goes to infinite, (5) can be rewritten as (1). The integer k is the equivalent of t of the continuous FT and k the equivalent of frequency f.
The inverse DFT (IDFT) is given by:

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j2\pi kn/N} \quad \text{n = 0, 1, ..., N-1.} \qquad (6)$$

The forward FT is sometimes denoted by the symbol $\mathcal{F}$, e.g. as X=$\mathcal{F}$(x) and the backward transform by $\mathcal{F}^{-1}$.
As **Principle** describes, after a Fourier transform the amplitude spectrum can be obtained. However, with a DFT the pertinent condition is that the signal lacks any harmonic contribution at frequencies above 0.5 f_m . Now the question arises what happens when the actual signal comprises frequencies above 0.5 f_m. This is addressed in Fourier transform and Aliasing.

Fourier transform and aliasing

Principle

In statistics, signal processing, computer graphics and related disciplines, aliasing refers to an effect that causes different continuous signals to become indistinguishable (or aliases of one another) when sampled. It also refers to the distortion or artifact that results when a signal is sampled and reconstructed as an alias of the original signal.

Suppose we have a sine with frequency f = 0.1 Hz, depicted in blue in Fig. 1, that is sampled with frequency f_m = 1 Hz. To our surprise we see that the samples also represent a sine with a much higher frequency, the red sine with exactly the frequency f_m - f = 0.9 Hz.

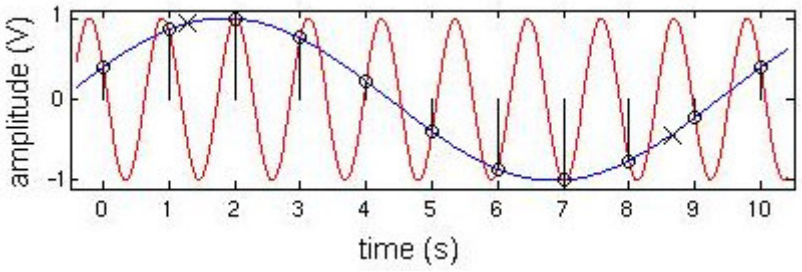


Fig. 1 Two different sinusoids that fit the same set of samples

This effect is often called aliasing or backfolding. It basically always happens with a discrete Fourier transform (DFT), also when signals which are sampled with such a high frequency that the signal seems to be well presented. However, *aliasing should preferably only be used when a signal is undersampled*. Therefore, Fig. 1 is not a good illustration of the aliasing effect. Undersampling means that there are less than 2 sample point per period; the crosses in Fig. 1 cannot represent the sine of 0.1 Hz. Many sine frequencies can underlie the crosses. More precisely, it implies that any signal comprising harmonics (see for harmonics Fourier analysis) which exceed the Nyquist (from the Nyquist–Shannon sampling theorem) or folding frequency $0.5f_m$, are folded backward over this frequency. Folding means: "alias frequency" = $f_{harmonic}$–$0.5f_m$ (where $0.5f_m < f_{harmonic} < f_m$).

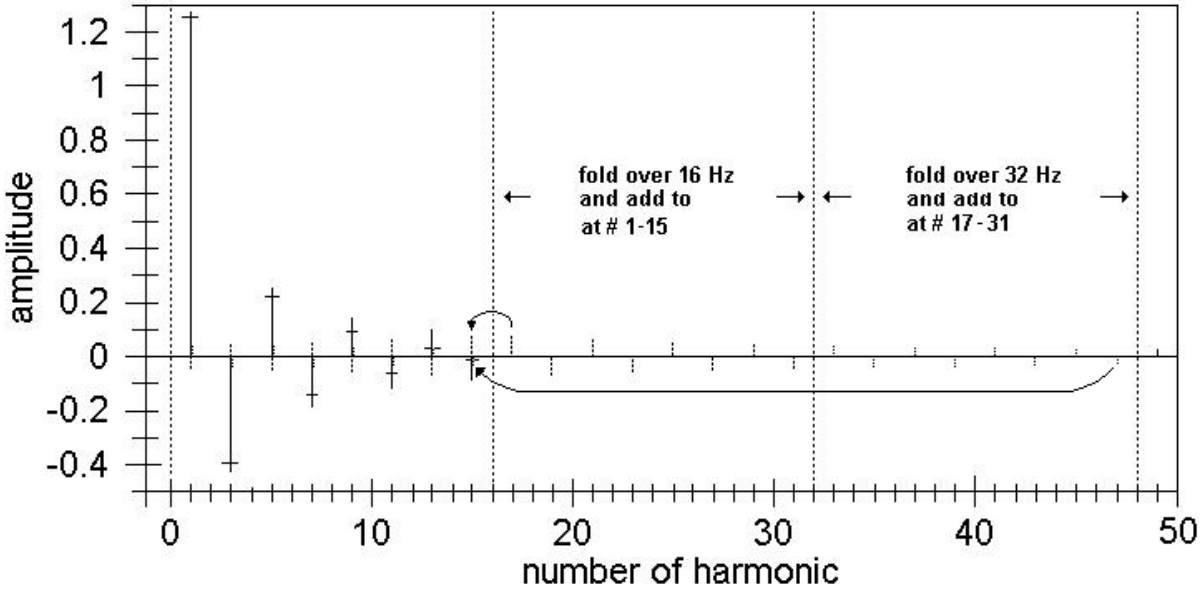


Fig. 2 Amplitude spectrum of a pure (unfiltered) symmetric square signal of 1 Hz, amplitude 1, written as even signal (only cosine components, see Fourier analysis) sampled with 32 Hz. Since the phases alternate between 0 ° and 180°, the phases are evaluated as a + and – sign and attributed to the amplitude. So, the subsequent harmonics alternate of sign. Folding frequencies are at 16, 32, etc. Hz. The 17th harmonic is folded to 15 Hz and its amplitude is added to the negative amplitude of the 15th harmonic, resulting in a much too small amplitude at 15 Hz. The component 47 Hz is first folded over 32

Hz to 17 Hz and then to 15 Hz. This final folding result is indicated by the lower curved arrow. Since the 47th harmonic has a 180^{o} phase shift, it is indicated negative. So its amplitude subtracts from the 15th harmonic. Bar length, dashed and solid are the theoretical amplitudes when a transform is made with a finite sample frequency. The horizontal stripe in the solid bars in the interval 0-16 Hz denote the final amplitude after folding.

Resuming, aliasing refers to an effect that causes different continuous signals to become indistinguishable (or aliases of one another) when sampled. if a signal comprises harmonics with a frequency >0.5f_m, the original signal cannot be completely reconstructed from the outcome of the DFT as an alias of the original signal recovered.

Aliasing can be prevented by filtering the signal before sampling, such that all harmonics >½f_m are deleted (see **More info**).
It can especially occur when the signal is periodic. Generally, biological signals are not periodic or at most pseudo-periodic (ECG), are noisy and have a spectrum that comprises most power in the low frequencies. Therefore, in practice filtering with a low pass Linear first order system and with f_m =1.5 the cutt off frequency of the filter is sufficient.

Application

Aliasing is an unwanted effect of sampling of any signal in the time or space domain (1D-3D).
The hardware of a signal processing apparatus generally prevents this drawback of sampling. When computational signal analysis software is used, combined with an analog to digital converter (ADC), generally the user himself has to provide against the effects. How signals can be processed effectively is discussed in Fourier transform and signal processing.

More Info

The description of **Principle** implies that everything goes well as long as there are no frequencies above ½f_m. If not, we get a spectrum between 0 and ½f_m. which is a mixture of the original spectrum (of the unsampled signal) above and below ½f_m. This is illustrated in Fig. 2. Since $f_{harmonic}$ can be >f_m, there are more folding frequencies; 0.5f_m + k·f_m, (0.5f_m modulo·f_m) with k is any non-zero positive integer. We have to fold a particular frequency so many times that it finally comes within the range 0 -½f_m. There we have to add it to the “genuine” (i.e. <0.5f_m) frequency component (harmonic). To add one should also *consider the phase* (vector addition is required).

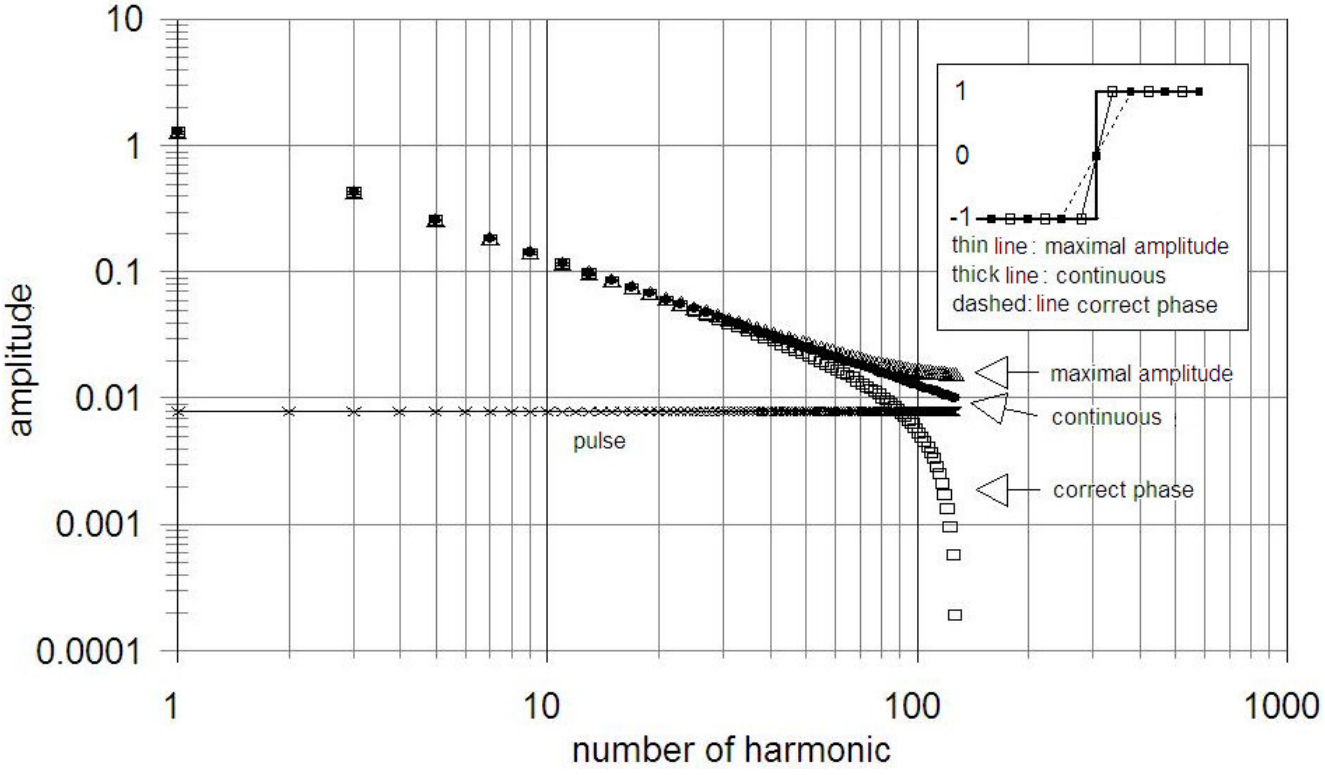


Fig. 3 Amplitude spectra of continuous block signal (A(f) = 4/(πkf); k is 1, 3, 5 etc.; dots), and after Fourier transform with sample frequency 256 Hz. The correct phase is obtained when at each zero crossing there is a sample (the filled squares of the inset). This yields the amplitude curve with open squares. The maximal amplitude (open triangles) is obtained when the samples are symmetrically just before and after the zero crossings (open rectangles in inset).

Choice of t=0

With a finite sample frequency it is important where in the signal t=0 is defined. This influences the outcome, and the more the closer the harmonic approaches $½f_m$. Fig. 3 gives the amplitude spectra of a continuous block, and a block sampled with 256 Hz, but with different positions of the first sample point (see inset). With the samples at the zero crossing, the phases are all correct: they alternate between 0 and 180°. Therefore, after folding the harmonics with $0.5f_m < f_{harmonic} < f_m$ are subtracted from those $<0.5f_m$. This results in too small amplitudes. The 127th component has obtained a final amplitude of $(4/\pi)\{1/(128-1) - 1/(128+1) + 1/(3X128-1) - 1/(3X128+1) \ldots.\}$. Calculating the power from the components, $1-127^{-2}$. This is directly found when the power is calculated in the time domain: we miss the power of the point in the zero-crossings. (Power in the time and frequency domain are the same, the Parseval theorema see Fourier analysis. With no samples at zero-crossings, the amplitudes are too large compared to the continuous block. When all sample points have the values +1 and −1, the power in both domains is exactly 1, but the phases are incorrect.

Digital Fourier transform of a pulse with a breadth of one sample point provide to a right amplitude spectrum to, what is also the case at the theoretical analysis of a pulse. Amplitude spectrum is constant until f_m, because all frequencies (under and above f_m) have the same amplitude. A phase shift off all sample points has no influence.

Preventing aliasing

From the above we learn that, irrespective how we sample, the outcome never has at the same time a correct amplitude and a correct phase. However, we can reduce the effect of folding by filtering as good as possible all frequencies $>½f_m$. In principle, $½f_m$ must be just larger than the highest harmonic with an amplitude > 0. In practise one uses the bandwidth of the signal. This is sufficient to obtain amplitudes and phases which enables reconstruction of the signal from the samples. However, this reconstruction requires an unrealizable filter that passes all frequencies $<½f_m$ unchanged while suppressing all others completely above $½f_m$. When realizable filters are used, some degree of oversampling is necessary. A low pass second order critical damped filter, which hardly shows attenuation or phase shift up to frequencies close to $½f_m$ (see Linear second order filter) is a very suitable anti-aliasing filter. In general a sampling frequency of 3 times $½f_m$ is sufficient to prevent aliasing.

When we view a digital photograph, the reconstruction (interpolation) is performed by a display or printer device, and by our eyes and our brain. If the reconstructed image differs from the original image, we are seeing an alias. An example of *spatial aliasing* is the Moiré pattern one can observe in a poorly pixelized image.

In video or cinematography, temporal aliasing results from the limited frame rate, and causes the wagon-wheel effect, whereby a spoked wheel appears to rotate too slowly or even backwards. Aliasing has changed its frequency of rotation. A reversal of direction can be described as a negative frequency.

Fourier transform and signal processing

Principle

Correct sampling, this means choosing a practical signal length T and the correct sample frequency is self evident. The latter is particular crucial for the results as shown in Fourier transform and Aliasing. In addition to low pass filtering, the hardware filtering should also be performed by a high pass filter before amplification to suppress too low frequencies, generated by drifting of the transducer, e.g. due to small temperature changes.

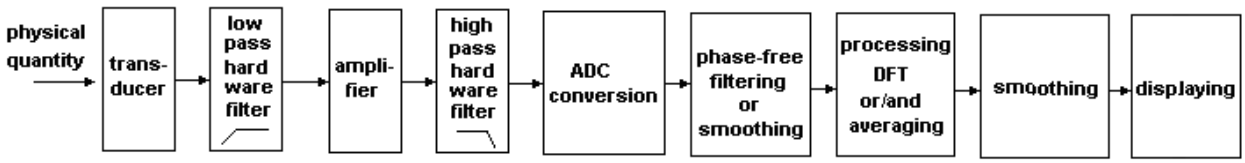


Fig. 1 Schematic presentation of signal processing from physical magnitude to presentation

Hardware filtering always causes signal distortion due to phase-shifts of the underlying harmonics. Therefore, hardware filtering should be reduced to what is minimal needed to prevent aliasing. Additional filtering can be done after sampling with the analog to digital converter (ADC). Then, the signal can be filtered phase-free by a software algorithm. Most convenient is to make a DFT of the signal, calculate the amplitude and phase spectra and correct the amplitude spectrum with the amplitude characteristic of the chosen filter. With a backward DFT the phase-free filtered signal is reconstructed.
Smoothing is another method of phase-free 'filtering' (it is a mathematical filter that has no hardware substitute). It removes fast signal variations by replacing the original sample amplitude by a new one which comprises the reduced amplitude of the sample and a fraction of the two neighboring sample amplitudes. The simplest procedure is taking ½ of the sample amplitude and add ¼ of each of the directly neighboring samples. The sum of the weighing factors is 1 and in this way, the mean of all samples is the same before and after smoothing. Remind that after filtering and smoothing, the power (proportional with squared amplitude) of the signal is diminished since the peaks in the signal are leveled off.
Averaging may respect spectra, made from subsequent (or partly overlapping) parts of equal duration of the signal or just subsequent parts of equal duration of the signal. These parts are 'synchronized' with a marker (trigger), for instance provided by a stimulus or by the signal itself, e.g. the maximum of the R-peak of the ECG.

More Info

Phase free filtering
An appropriate way to perform phase free filtering of a signal is to multiply the discrete Fourier transform (DFT) of the signal, $X_n=\mathcal{F}(x)$ (see Fourier transform) with the discrete amplitude characteristic A(n), the discrete equivalent of the continuous A(ω). With n going from 1 to N, A(n) should be made symmetric such that A(1) = A(N), A(2) = A(N-1), etc. The procedure is actually multiplication of the real and imaginary part of X_n with the amplitude attenuation of the filter at frequency n. When the phase should be preserved one also had to add the arguments of X_n and the complex transfer characteristic H_n (the discrete equivalent of H(jω), see Linear first order system). With the complex outcome the backward DFT is made, yielding the digitally filtered signal.
An alternative procedure is to filter the signal with preservation of phase (the regular way), invert the time of the outcome and filter it again (backward filtering). Now, two times the amplitudes are attenuated. Two times the phases are shifted by the same amount, but the first time with a negative sign and the second time with a positive sign and so resulting in no phase shift. In this way, one obtains always a symmetric result around t = 0. Fig. 2 gives the result of forward and backward filtering, so phase free, of an impulse with a first order low pass filter.
The phase free filtering applied two times in succession, so by multiplying $X_n=\mathcal{F}(x)$ with $(A(n))^2$ gives the same wave shape as forward plus backward (time reversal) filtering.
In addition, Fig. 2 also presents regular, this is hardware filtering of the impulse (1x filtered) and two times regular filtering (2x filtered). With hardware filtering one always obtains an output only existing for t ≥ 0. (The response to '2x filtered' can be seen as the integral of '1x filtered', but after some time the amplitudes leaks, or in hardware terms, the capacitor is discharged over the resistor (see Linear first order system). The first and last one show the dramatic difference in signal shape.

Smoothing
Smoothing is a discrete cross-correlation (see Stochastic signal analysis) performed with a symmetric weight function W(m). In continuous notation the cross-correlation function is:

$$\phi_{Wx}(\tau) = \lim(1/2T) \int_{T\to-\infty}^{T\to\infty} x(t-\tau)W(t)dt \qquad (1)$$

W(m) generally comprises M number of points and M is an odd number. The M numbers are mostly the values of the numbers of Pascal's triangle:

$$\binom{k}{m-1} = \frac{k!}{(m-1)!(k-(m-1))!}. \qquad (2)$$

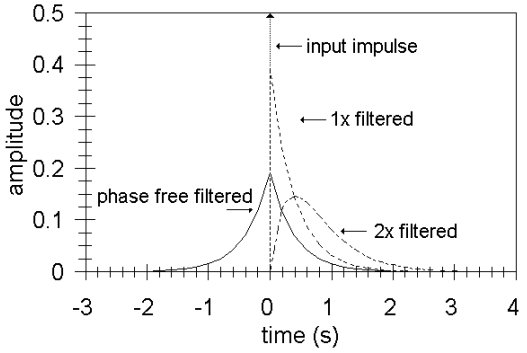


Fig. 2 Various ways of filtering an impulse by a first order low pass filter. The times of the sample points are the ticks along the time axis. See for more detail **More Info**.

W(m) is:

$$W(m) = \frac{\binom{k}{m-1}}{\sum_{m=1}^{m=M} \binom{k}{m-1}} \qquad (3)$$

where k is the number of the line in the triangle or the exponent of the binomial expression $(a+b)^k$. (k! = 1x2x3…k, the factorial of k). Notice that k=M-1 and that m goes from 1 to M.
With k=2, the weights become 0.25, 0.5 and 0.25. Making the autocorrelation (see Stochastic signal analysis) of this weight function with itself one obtains the weights 1/16, 4/16, 6/16, 4/16 and 1/16. This is the outcome of Eq. 3 with M=5.
With a signal length of N the number of operations is MxN. Phase-free filtering needs 2NlogN operations. Only for small values of N smoothing is about as fast as filtering.
Fig. 3 gives the smoothing of an impulse, with W(m) for M=3. This has been done 1 to 9 times repetitively. The more times, the more the result likes a Gaussian curve and the same holds for repetitive phase-free filtering.

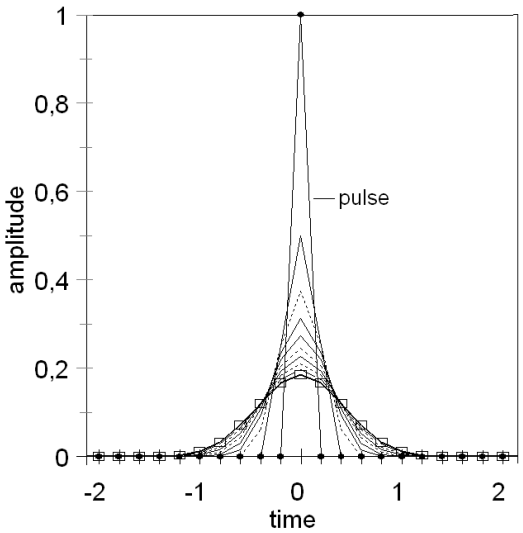


Fig. 3 Impulse, repetitively smoothed with weight function ¼, ½ and ¼ . Dots indicate samples of the impulse and open rectangles that of the 9 times smoothed pulse.

Averaging
This may concern averaging of subsequent pieces of equal length of signal or averaging of the spectra obtained with DFT applied to pieces of the signal. Dividing the signal in pieces can be given by a marker or trigger. This is the case when the signal is e.g. a physiological response to a repetitive physiological stimulus. Averaging gives an improvement of the signal to noise ratio of the square root of the number of pieces or spectra.

Halftime and time constant

Principle

Half-life or more practically halftime ($t_{½}$) is the time required for the decaying quantity N_0 to fall to one half of its initial value, $½N_0$. Knowing $t_{½}$, at any time the remaining quantity can be calculated:

$$N(t) = N_0 2^{-t/t_{1/2}}. \quad (1)$$

When t = $t_{1/2}$, then $N(t_{1/2}) = N_0 2^{-1} = ½N_0$, hence the name halftime. Thus, after 3 halftimes there will be $2^{-3} = 1/8$ of the original material left.
N(t) decays with a rate proportional to its own, instantaneous value. Formally, this can be expressed as the following differential equation, where *N* is the quantity and λ is a positive number called the decay constant.

$$\frac{dN}{dt} = -\lambda N_0. \quad (2)$$

The solution to this equation is:

$$N(t) = N_0 e^{-\lambda t}. \quad (3)$$

Generally, in (medical) physics and engineering, λ, a reciprocal time, is substituted by $1/\tau$. The Greek letter tau, *τ, is* the time constant of the decay process. The decay appears to be exponential (Fig. 1).

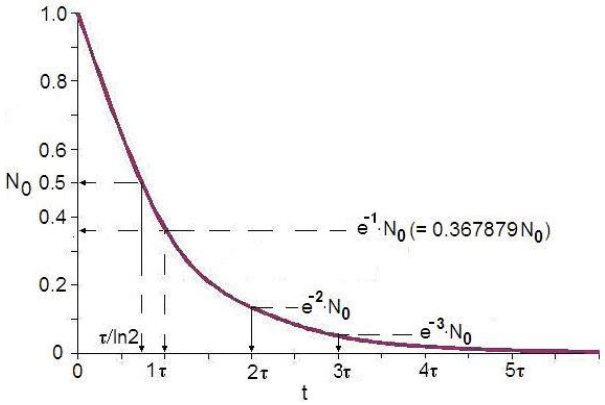


Fig. 1 Exponential decay with halftime $t_{1/2} = \tau/\ln 2$. Multiples of *τ* with their corresponding quantities are indicated.

Equation (1) and (3) look very similar and are strongly related. *τ* appears to be:

$$\tau = t_{1/2}\ln 2. \quad (4)$$

Substituting λ in (3) by $1/\tau$ and using (4) equation (1) is obtained.
In practice (medical, physical, control engineering systems), after 8 halftimes or 5 time constants N(t) is supposed to be zero ($< 0.01N_0$). So, after this time the process is assumed to have reach its asymptotic value (here zero).

Application

Applications are innumerous in science, technology and medicine. Trivial examples in medicine are the clearance of a substance in the blood, the radio active decay of a tracer molecule and the way in which a thermometer reaches for instance the body temperature of a patient. It is also basic in optics (Lambert Beer's law), spectroscopic applications, radio-diagnostics, radiotherapy, nuclear medicine, neurobiology etc.

More info

In (medical) physics and engineering τ characterizes the frequency response of a first-order, linear time-invariant (LTI) system (see Linear systems: general). A LTI system, simply said, has properties that do not vary. First-order means that the system can be described by a first order differential equation (2), an equation comprising the first derivative and no higher ones.
Examples include the electrical RC filter (comprising the resistor R and the capacitor C) and RL circuit (comprising the resistor R and the self-inductance L). (see for a description of the circuit elements R, C and L Electrical resistance, capacitance and self-inductance). It is also used to characterize the frequency response of various signal processing systems, such as the classical magnetic tapes, radio and TV transmitters and receivers and digital filters – which can be modeled or approximated by first-order LTI systems. Other examples include time constant used in control systems for integral and derivative action controllers.
The *frequency response* is given by the amplitude and phase characteristic. The first one gives the gain of the system as a function of the frequency of a sinusoidal input. The second one gives the phase shift of the output of the system relative to phase of the sinusoidal input. The time constant is related to the cut off frequency ω_0 of the first order LTI system:

$$\tau = 1/\omega_0,$$

where $\omega_0 = 2\pi f_0$ and f_0 the cut off frequency, i.e. the frequency of the sinusoidal input wave which output amplitude is $2^{-0.5}$ (= 0.71) times the input amplitude. .
The time constant also describes the output to a very elementary input signal, the impulse function. Its output has the same shape as the exponential decay of Fig. 1. After one τ the decay is 63.2% of its original value, and 36.8% (about $100/\tau$ %) remains. Another time of 4 τ is needed to reach the asymptote (zero).
See for more info about linear systems Linear systems: general and Linear first order system.

Linear first order system

Principle

A time invariant (see Linear systems: general) linear first order system is formally described by a linear first order differential equation with the time as independent variable. The output of the system is given by its input and the system characteristics. The latter form the time constant τ (see Halftime and time constant). The solution of the equation comprises an exponential function with time constant τ: $e^{-t/\tau}$.

The most used analogue of this mathematical system is the electric analogue, an electric closed circuit comprising a resistor R, a capacitor C (see Electrical resistance, capacitance and self-inductance) and a voltage V_s (or current) source (Fig. 1). They are called RC filters. There are two types: the low pass (or high frequency cut off) filter (Fig. 1a) and the high pass (or low frequency cut off) filter (Fig. 1b). They are used to diminish the high and the low frequencies in a signal, respectively.

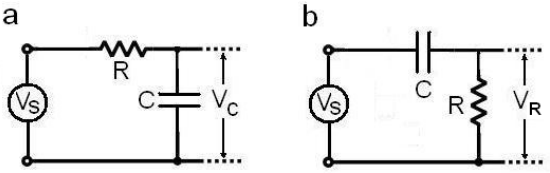


Fig. 1 Electrical first order systems with left the low pass and right the high pass system (or filter).

There are 3 types of special input signals. Normalized they are:
- the unit step function (with amplitude 1);
- the unit impulse function, infinite short and amplitude infinite, such that the time integral (area) is one;
- the sinusoidal signal (amplitude 1).

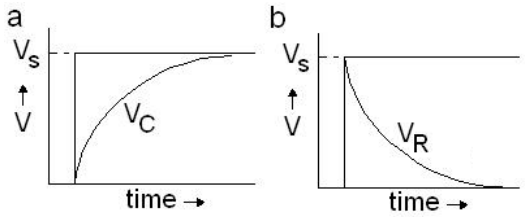


Fig. 2 Step responses measured over the capacitor and over the resistor of Fig.1a and b respectively.

Step response
The step response to the step function with amplitude V_s of the low (a) and high (b) pass filters of Fig. 1 are given in Fig. 2. They are:

- low pass filter (Fig. 1a, 2a): $V_{C,\,step} = V_s(1- e^{-t/\tau})$; (1a)
- high pass filter (Fig. 1b, 2b): $V_{R,\,step} = V_s e^{-t/\tau}$, (1b)

where $\tau = RC$.

Impulse response
Since the impulse is the time derivative of the step, the impulse response is the time derivative of the step response:

$V_{C,\,impulse} = V_s\tau^{-1}e^{-t/\tau}$; (2a)
$V_{R,\,impulse} = -V_s\tau^{-1}e^{-t/\tau}(t>0)$, with an impulse to $+\infty$ at $t = 0$. (2b)

Sinus response
The response to a continuing sinusoidal input is always a continuing sinusoidal output of the same frequency, but generally with a different amplitude and phase. The sinus response is dependent on the frequency, which is often expressed as angular frequency ω. Notice that $\omega = 2\pi f$ with f the frequency in Hz. The input/output ratio of the sinus amplitude as a function of frequency, A(ω), is visualized in the amplitude characteristic. The phase shift of the output sinus related to the phase of the input as a function of frequency, φ(ω), is visualized in the phase characteristic. Fig. 3 and 4 present both for the low and high pass filter. Expressed in equations it holds that:

$A_{low\ pass}(\omega) = 1/(1+ \omega/\omega_0)^{0.5}$ (3a)
$A_{high\ pass}(\omega) = 1/(1+ \omega_0/\omega)^{0.5}$ (3b)
$\varphi_{low\ pass}(\omega) = -\arctan\omega/\omega_0$ (4a)
$\varphi_{high\ pass}(\omega) = \arctan\omega_0/\omega$, (4b)

where $\omega_0=1/\tau$. Derivations (there are several) can be found in any basic textbook on physics. Notice that $A_{low\ pass}$ and $A_{high\ pass}$ are mutually mirrored along the vertical axis at $\omega/\omega_0 = 1$, and that $\varphi_{high\ pass}$ is shifted $+90^0$ with respect to $\varphi_{low\ pass}$.

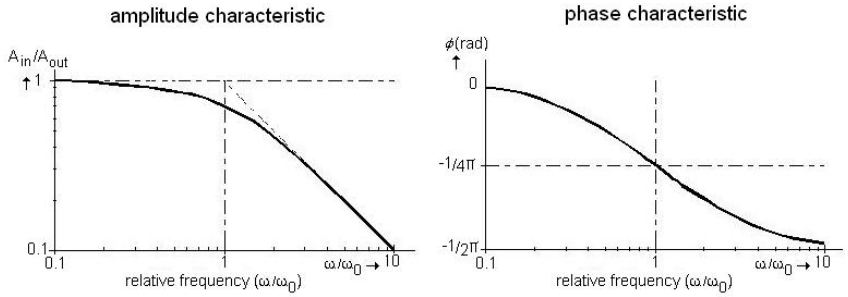


Fig. 3 Frequency characteristics of ***low*** pass first order system with a normalized frequency axis. The frequency characteristics presented as log-log and log-lin plots are called Bode plots.

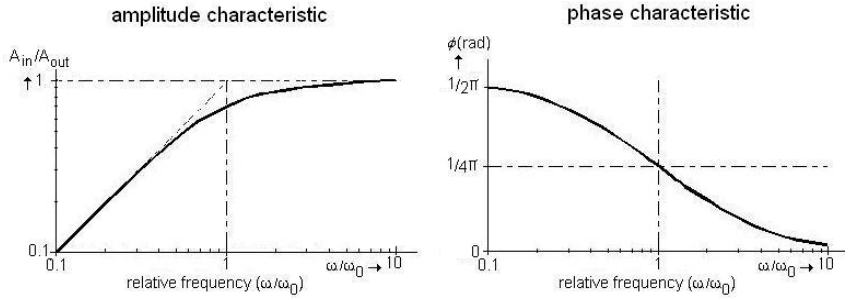


Fig. 4 As Fig. 3 for a *high* pass first order system.

A first order filter can also be made with a resistor and self-inductance. When the latter takes the place of the capacitance in Fig. 1a, a high pass filter is obtained, since high frequencies hardly pass the inductance, whereas for low frequencies the inductance is practically a short-circuit. Similarly, a low pass filter is obtained when the place of the capacitance in Fig. 1b is taken by the inductance. In practice this approach is a little more complicated because a coil also has a resistance (see Electrical resistance, capacitance and self-inductance). Therefore LC filters are seldom used.

Application

General
RC filters are applied in any electronic device. Their analogues exist in fluid dynamics and acoutics and phase-less 1D or 2D analogues exist in optics (lenses).
In numerical form they are used to filter digital signals in the time and spatial domain (2D, 3D or 3D-t).
Modeling
They are widely used in computational modeling of systems, also in medical and biological application. A classical example of a low pass filter in neurobiology is the decay of an action potential. Its characteristic frequency $\omega_0 = 1/\tau = 1/r_m c_m$, where r_m is the resistance across the membrane and c_m is the capacitance of the membrane. r_m is a function of the number of open ion channels and c_m is a function of the properties of the lipid bilayer.

More Info

In real notation, A and φ can be combined in a polar plot, in which A is the length of the vector and φ is the angle between the vector and the horizontal axis. For the low (high) pass filter, it is a semicircle in the right lower (upper) half-plane. ω follows this trajectory clockwise. Fig. 5 presents the polar plot for the low pass system.

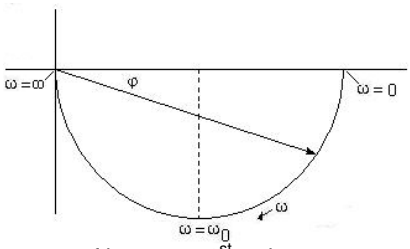


Fig. 5 Polar plot of frequency response of low pass 1st order system.

In jω notation, the transfer characteristic, comprising the real A(ω) and φ(ω), is denoted by H(jω). H(jω) can be found easily with the impedances of R and C (see Electrical resistance, capacitance and self-inductance). It always holds that H(jω) = output impedance/Σimpedances along the closed circuit, provided that the voltage source is ideal (impedance zero). Consequently, H(jω) of the low pass system is:

$$H(j\omega) = \frac{1/j\omega C}{R + 1/j\omega C} = \frac{1/RC}{j\omega + 1/RC} = \frac{1/\tau}{j\omega + 1/\tau} = \frac{\omega_0}{j\omega + \omega_0} \qquad (5a)$$

It holds that:

$A(\omega) = |H(j\omega)|$; (5b)

$\varphi(\omega) = \arg H(j\omega) = \mathrm{Im}\{(j\omega)\}/\mathrm{Re}\{H(j\omega)\}$. (5c)

The Laplace back transform yields the impulse response h(t):

$$\mathcal{L}^{-1}(H(j\omega) = h(t) \qquad (6)$$

The Laplace operator s is $s = \alpha + j\omega$, but in calculations of the transfer characteristics $\alpha = 0$. So, calculations based on (5) and (6) use the imaginary frequency jω. However the complex s-plane can be used, not only to calculate in an easy why A(ω) and φ(ω), but also to visualize the action of poles and zero's (see Linear system analysis). This is explained with the low pass 1st order system, which H(jω) is $H(j\omega) = \omega_0/(j\omega + \omega_0)$. ω_0 is a pole (root) of the denominator of H(jω).

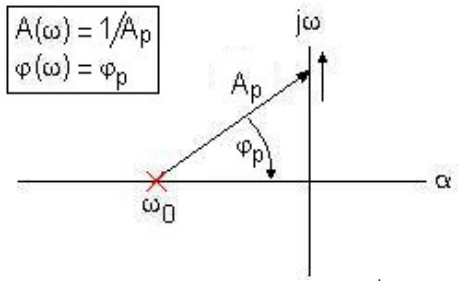


Fig. 6 Pole-zero plot of low pass 1st order system, in Laplace notation $H(s) = \omega_0/(s + \omega_0)$.

With poles and zero's (a zero is the root of the nominator) indicated in the complex s plane, this figure is called the pole-zero diagram. For the low pass 1st order system, the diagram, comprising only a pole at the left real axis, is presented in Fig. 6. A(ω) is found by moving from the origin along the positive jω axis (vertical arrow) and calculating the reciprocal of the length of the vector. This yields (3a). φ(ω), given by (4a), is found by turning the oblique side to the real axis (clockwise is a negative phase). For the high pass system Fig. 7 holds. It has also a zero in the origin. In Laplace notation it is:

$$H(s) = s/(s + \omega_0), \qquad (7)$$

with the nominator yielding the zero. The s in the nominator with its root in the origin has the action of pure differentiation. For sine waves this means adding $+90^o$ to the phase plot of the low pass system of Fig. 3b and adding a straight line with a slope of +1 through the point $(\omega_0,1)$ to the amplitude plot of Fig. 3a. Notice that adding this oblique line in this log/log plot means multiplication by ω. Now, the Bode plots of Fig. 4 are obtained. The inset of Fig. 7 gives the contribution of the pole and zero in A(ω) and φ(ω).

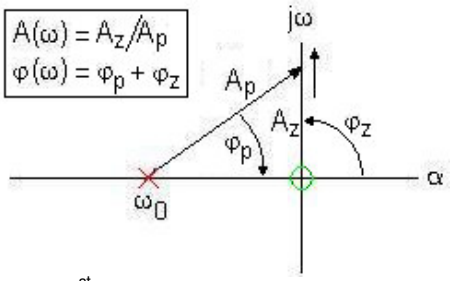


Fig. 7 Pole-zero plot of the high pass 1st order system $H(j\omega) = j\omega/(j\omega + \omega_0)$.

In general, multiplying some H(s) by the operator s yields differentiation of the output signal (any signal) and dividing H(s) by s means integration. This is related to the notion that the step response is the integral of the impulse response.
See for a more complete description of the Laplace approach in system theory Linear system analysis.

Linear second order system

Principle

From the point of view of basic theory of electricity a time invariant linear second order system comprises two frequency dependent linear components: i.e. two capacitors (separated by a resistance), two self-inductances (separated by a resistance) or one capacitor together with one self-inductance (see Electrical resistance, capacitance and self-inductance). All configuration have at least one resistor.
The point of view of the theory of electricity was chosen since that is most understandable. Moreover we assume that the signals are voltages.
Analogues of these systems can be found in other fields of physics e.g. mechanics, see Linear second order system, analogues.

Linear second order system comprising two first order system

A simple form, also from the point of practical realization is the serial configuration of two time invariant linear first order systems. Then we get a low pass together with a low pass system, two high pass systems or two high pass systems.
The amplitude attenuation of a serial system comprising two linear systems (see Linear first order system) can be found by multiplying the attenuations of both systems (if the second system does not load the first one. They must be separated by a 1x amplifier.) Consequently, a double low pass gives:

$$A_{\text{low pass1\&2}}(\omega) = A_{\text{low pas1}}(\omega) \times A_{\text{low pass2}}(\omega) = 1/(1+ \omega/\omega_{0,1})^{0.5} \times 1/(1+ \omega/\omega_{0,2})^{0.5}, \quad (1a).$$

where $\omega_0=1/\tau$ and τ is the time constant (see: Halftime and time constant).

The indices 1 and 2 indicate both low pass systems. When both are identical we get:

$$A_{\text{double low pass}}(\omega) = A_{\text{low pas}}{}^2(\omega) \times A_{\text{low pass}}(\omega) = 1/(1+ \omega/\omega_0) \quad (1b).$$

A high pass system and a low pass system (their order is irrelevant) give:

$$A_{\text{high low, low pass}}(\omega) = A_{\text{high pas}}(\omega) \times A_{\text{low pass}}(\omega) = 1/(1+ \omega_{0,\text{high}}/\omega)^{0.5} \times 1/(1+ \omega/\omega_{0,\text{low}})^{0.5}, \quad (2).$$

Linear second order system comprising one capacitor and one self-inductance

These systems show resonance for a particular frequency. The maximal resonance is close to the cut off frequency, which is generally indicated by ω_0 (or f_0). The extent of resonance is indicated by quality factor Q. The higher Q the more resonance near ω_0. Q and ω_0 are given by:

$$Q = R^{-1}\sqrt{(L/C)},\ \omega_0 = 1/\sqrt{(LC)}, \quad (3a)$$

with R the resistance, C the capacitance and L the self-inductance. By approximation the maximal amplitude is $\sqrt{Q}$ and found at ω_0. A practical way to estimate Q is with:

$$Q = \omega_0 \left| \omega_1 - \omega_2 \right|^{-1} \quad (3b)$$

For the low/high pass filter, the exact maximal amplitude is found at frequency:

$$\omega_{max}=(\omega_0{}^2 -/+ \tfrac{1}{2}\omega_0{}^2/Q^2)^{0.5}. \quad (3c)$$

Hence, ω_{max} is smaller/larger than ω_0 (for Q>4 less than 1%).

There are 3 configurations of this system, a low pass, band pass and high pass version, all with the same Q and ω_0, but with different A(ω) and Φ(ω).
The configuration of the low pass system is R, C and L in series with measuring the output over C.
Its amplitude A(ω), after some calculus, is:

$$A(\omega) = \frac{1}{(1+(1/Q^2-2)\omega^2/\omega_0{}^2+\omega^4/\omega_0{}^4)^{0.5}}, \quad (5)$$

and the phase:

$$\Phi(\omega) = \arctan - \frac{\omega/(Q\omega_0)}{(1-\omega^2/\omega_0{}^2)} \quad (6)$$

Fig. 1 represents A(ω) and Φ(ω) for a number of Q's. A(ω) and Φ(ω) together are called Bode plots.
For Q = 1, $A(\omega_0) = 1$ and the system shows a tiny maximum ($\sqrt{(4/3)}$ at $\omega = 0.5\omega_0\sqrt{2}$. This case is called a critical damped second order system for A(ω) and Φ(ω) most closely approaches the horizontal and oblique asymptotes.

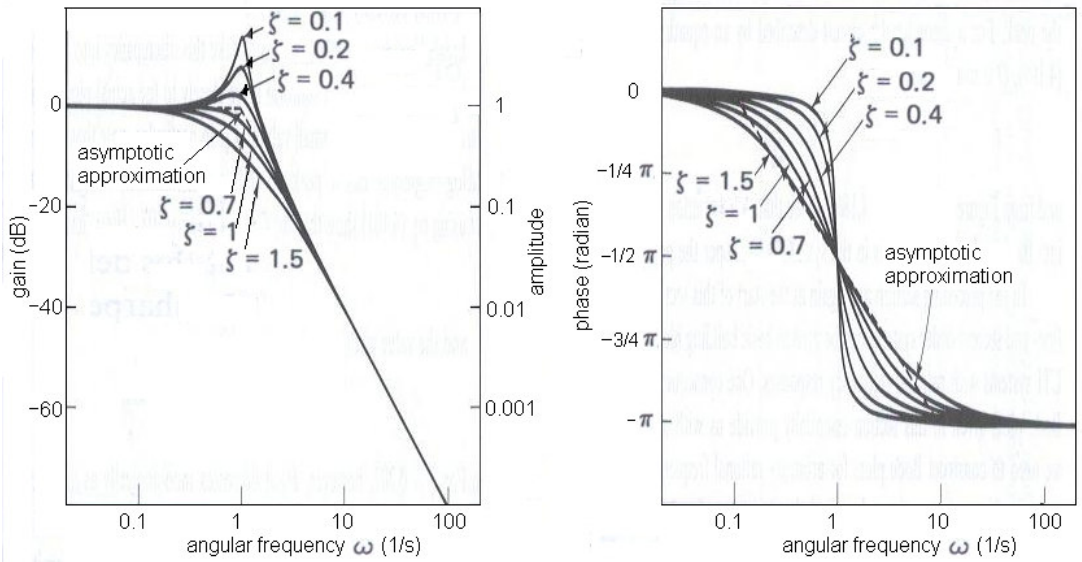


Fig. 1 Second order band-pass system with 6 values of Q ($\zeta = 0.5Q^{-1}$; ζ is called the damping ratio). At the left are the amplitude (or gain, in dB) characteristics, at the right the phase characteristics.

The system with L and C in series and measuring the output over R presents the band pass filter. Conceptually this is easy to understand. L has a high impedance for high frequencies and C for low frequencies (see Electrical resistance, capacitance and self-inductance). Therefore, only moderate frequencies pass: the band pass is realized.

Application

Numerous, in all kind of medical equipment which make use of transducers, signal measurement and retrieval and signal analysis. In addition, applications are found in the theory of biomechanics and in the medical applications of the theory of fluid dynamics. Examples are the rotation of the eye ball (see Torsion), the oscillation of the cupula (see Vestibular mechanics) in the semicircular canal and the vibration of the eardrum (see Physics of middle ear).

More Info

For calculation amplitude and phase it is more convenient to use the s-notation (from the Laplace transform with $s = j\omega$). Since $H(j\omega) = H(s)$ it applies that $A(\omega) = |H(j\omega)|$ and $Q(\omega) = \arctan \mathrm{Im}\{H(j\omega)\}/\mathrm{Re}\{H(j\omega)\}$.

When the output is measures over the capacitor, H(s) becomes:

$$a.\quad H(s) = (1/sC)/(sL+R+1/sC) = 1/(s^2LC+sRC+1) = \omega_o^2 / (s^2 + \omega_o s/Q + \omega_o^2) \qquad (7a)$$

When the output is measures over the resistance, we obtain:

$$b.\quad H(s) = R/(sL+R+1/sC) = sRC/(s^2LC+sRC+1) = (\omega_o s/Q) / (s^2 + \omega_o s/Q + \omega_o^2) \qquad (7b)$$

When the output is measures over the self-inductance, we obtain:

$$c.\quad H(s) = sL/(sL+R+1/sC) = s^2LC/(s^2LC+sRC+1) = s^2 / (s^2 + \omega_o s/Q + \omega_o^2) \qquad (7c)$$

Notice that the denominator of H(s) must be written as a polynomial of s to find its roots.

Frequency behavior of the denominator
In all three cases the denominator is the same with complex poles in:

$$s = -\tfrac{1}{2}R/L \pm j\,(1/LC - R^2/4L^2)^{1/2} = \alpha_k \pm j\omega_k. \qquad (8)$$

Poles can be found by calculating the roots of the denominator, here a quadratic equation. Consequently, a second order system has two poles. (The order of the polynomial of the denominator gives the number of poles, p.)

The parameter ω_k can have three values, depending on the value of the discriminator of the quadratic equation:

1. $\omega_k > 0$. Then Q<½ and both poles are located at the negative real axis (case 1. in Fig. 2). Hence, the system is a cascade of two first order systems, each with its own cut-off frequency. The frequency characteristic shows no resonance.
2. $\omega_k = 0$. Then Q = ½ and both poles coincide at the negative real axis (case 2. in Fig. 2). Hence, the system is a cascade of two identical first order systems with cut-off frequency ω_0. Again, the frequency characteristic shows no resonance.
3. $\omega_k < 0$. Then Q > ½ and both poles, located in the left complex half plane, have the real part α_k and the imaginary part $\pm\ \omega_k$ (case 3. in Fig. 2). Hence, they form a complex, conjugate pair and the frequency characteristic shows resonance.

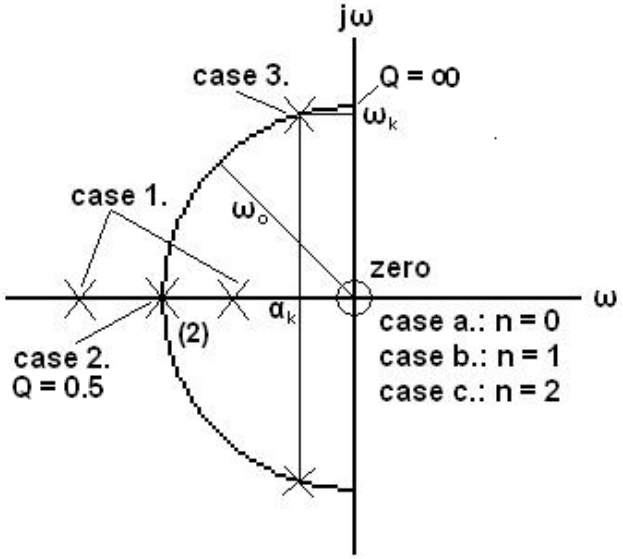


Fig. 2 p-n diagram. When case 1. evolves to 2. the two poles move to the point $(\omega_o,0)$, where they coincide (indicated by (2): two coinciding poles). From 2. to 3. the poles walk over the half circle, one up and one down, to form a conjugated pair. The circle has radius ω_o.

Frequency behavior of the nominator

The cases a., b. and c. only differ in their number of zero's, being none, 1 and 2 respectively. (The order of the polynomial of the nominator gives the number of zero's, n).

The steepness of negative slope is given by n-p. It is the power of asymptotic high frequency behavior. When n-p is negative there is attenuation: so $A(\omega) \sim \omega^{n-p}$. Since the system must be physical constructible, the condition is that $p \geq n$.

The zero's are located in the origin of the complex pole-zero plot (Fig. 2). Adding a zero in the origin is equivalent to a pure differentiation of the underlying differential equation describing the system. For the amplitude characteristic (notice double log) this means adding a line with slope +1 through the point $(\omega_0,1)$ and for the phase characteristic adding 90^o, (since d*sin*nx/dx = n*sin*(x+90)). This describes the transition from case a. to b. Going from b. to c. needs another differentiation. Hence, the number of zero's (n) in the origin gives the steepness of the positive slope of the asymptote: $A_{asymptote}(\omega) \sim \omega^n$.

From the p-n image, the amplitude and phase characteristic can be graphically constructed (see textbooks on system analysis).

Fig. 2 gives the p-n diagram for the three cases of the nominator and of the denominator.

Literature

Oppenheim A.V., Willsky A.S. and S. Hamid S. Signal and Systems, 2nd Edition, Prentice-Hall.
http://www.facstaff.bucknell.edu/mastascu/eControlHTML/SysDyn/SysDyn2.html#TheSystem

Linear second order system: analogues

Principle

In various sub-disciplines of classical physics there exists time invariant linear second order systems, which can be seen as analogues of each other. They all can be described by a 2nd order differential equation but preferably by their Laplace representation (s-notation).
Conceptually, they are most easily described by signals with voltage (or pressure or another equivalent dimension) as their dimension and their system constants (the elements, e.g. resistance), called impedance analogues. The generalized description focuses on the transfer characteristic, being the complex ratio of output-impedance/ system-impedance. Now, the voltage output signal of the system is by definition the voltage input signal times the system transfer. This approach avoids the problem of evolving the differential equations (see Linear first order system).
But the systems can also be described with the so-called mobility analogues (e.g. conductance). Now, the current (or its equivalent) is the basic variable of the equations.
Table 1 gives the electrical analogues of mechanical quantities.

Table 1. Electrical impedance and mobility analogues of mechanical quantities

Mechanical quantity		*Electrical analogues*			
		Impedance analogue		*Mobility analogue*	
Name	*Value*	*Name*	*Value*	*Name*	*Value*
mass	m	inductance	m	capacitance	m
compliance	C_m	capacitance	C_m	inductance	C_m
stiffness	S	capacitance	1/S	inductance	1/S
mechanical resistance	R_m	resistance	R_m	conductance	$1/R_m$
force	F	voltage	F	current	F
velocity	v	current	v	voltage	v
displacement	x	charge	x	flux	x

From ref. 1. See Electrical resistance, capacitance and self-inductance for the definitions of electrical these elements.

Table 2 gives the relations between electrical, hydrodynamical, mechanical and acoustical analogues.

Table 2 Relations between physical analogues.

Electrical			*Hydrodynamical*			*Mechanical*			*Acoustical*		
Name	*Symbol*	*Unity*	*Name*	*Symbol*	*Unity*	*Name*	*Symbol*	*Unity*	*Name*	*Symbol*	*Unity*
voltage	E	Volt	pressure	P	Pascal	force	F	Newton	pressure	P	Pascal
current	I	Ampère	volume-current	$\dot{Q}_H$	$m^3.s^{-1}$	velocity	v	$m.s^{-1}$	volume velocity	U	$m^3.s^{-1}$
charge	q	Coulomb	volume	Q_H	m^3	displacement	x	m	volume	Q_A	m^3
resistance	R	Ohm	hydrodynamic resistance	R_H	$N.s.m^{-5}$	mechanical resistance	R_M	$N.s.m^{-1}$	acoustic resistance	R_A	$N.s.m^{-5}$
capacitance	C	Farad	compliance	C_H	$m^5.N^{-1}$	mechanical compliance	C_M	$m.N^{-1}$	acoustic compliance	C_A	$m^5.N^{-1}$
inductance	L	Henry	inertance	M_H	$kg.m^{-4}$	mass	m	kg	acoustic inertance	M_A	$kg.m^{-4}$
impedance	Z	Ohm	impedance	Z_H	$N.s.m^{-5}$	mechanical impedance	Z_M	$N.s.m^{-1}$	acoustic impedance	Z_A	$N.s.m^{-5}$

From ref. 1.

Literature

L.H. van der Tweel and J Verburg. Physical concepts. In: Data in medicine: Collection, processing and presentation, R.S. Reneman and J. Strackee (eds.) Martinus Nijhoff, The Hague, p. 11-51.

Linear systems: general

Principle

System analysis is the branch of engineering, mostly electrical engineering that characterizes electrical systems and their properties. Many of the methods of system analysis can be applied to non-electrical systems, e.g. mechanical and acoustic systems and in the theory of flow of liquids.
The most simple and by far most applied linear system in calculations and in technology (electronic filters etc.) is the Linear first order system. It is very basic and its understanding is a necessity to understand the working of more complicated systems. More complicated systems (filters) are often a combination of several linear first order systems (see the mathematically more advanced contribution System analysis).
Linear systems can be distinguished in many ways.

Input and output
A system is characterized by how it responds to input signals. A system can have one or more input signals and one or more output signals. Therefore, a system can be characterized by its number of inputs and outputs. Single input with single output is most common. Another possibility is for instance multiple inputs (e.g. visual and auditory) and single output (eye blink). Here and in other contributions about system analysis, a system is supposed to be a *single-input-single-output system*. By far, the greatest amount of work in system analysis has been done with these systems, although several components (e.g. resistances or capacitors in an electric circuit) may obtain a different input signal.
It is often useful (or necessary) to break up a system into smaller pieces for analysis (see System analysis).
Since in physics and engineering systems process signals, properties of signals should first be defined before continuing the description of systems.

Analog and digital, continuous and discrete
Signals can be continuous (e.g. the EEG) or discrete in time (the time series of R peaks in the ECG or by approximation an action potential), as well as discrete in the values they take at any given time:
- Signals that are continuous in time and continuous in value are known as *analog signals*.
- Signals that are discrete in time and discrete in value are known as *digital signals*.

With this categorization of signals, a system can then be characterized as to which type of signals it deals with:
- A system that has analog input and analog output is known as an *analog system*.
- A system that has digital input and digital output is known as a *digital system*.

Systems with analog input and digital output or digital input and analog output are possible. However, it is usually easiest to break these systems up for analysis into their analog and digital parts, as well as the necessary analog to digital or digital to analog converter.
Another way to characterize systems is by whether their output at any given time depends only on the input at that time or perhaps on the input at some time in the past (or in the future!).

Memory and causality
Memory-less systems do not depend on any past input and systems *with memory* do depend on past input.
Causal systems do not depend on any future input. *Non-causal* or *anticipatory* systems do depend on future input. It is not possible to physically realize a non-causal system operating in "real time". However, they can be simulated "off-line" by a computer. They can give insight into the design of a causal system.
Analog systems with memory may be further classified as *lumped* or *distributed*. The difference can be explained by considering the meaning of memory in a system. Future output of a system with memory depends on future input and a number of coefficients (state variables), such as values of the input or output at various times in the past. If the number of state variables necessary to describe future output is finite, the system is lumped; if it is infinite, the system is distributed. In practice, most system in science as well as in medicine can be considered or approximated as lumped.

Linear and non-linear, superposition
A system is *linear* if the superposition principle holds. This means:

input	→ system →	output
a	→ system →	A
a	→ system →	B
a+b	→ system →	A+B

The principle also implies the scaling property: multiplying input *a* with k yields as output kA.

Further, it holds that when the input is a single sine wave with frequency f, the output solely comprises a sine wave with the same frequency f.
A linear system can be described by a linear differential equation. A system that behaves not linear is *non-linear*.

Time invariant and time variant
If the output of a system with equal input does not depend explicitly on time (hour of day, season etc.), the system is said to be time-invariant; otherwise it is time-variant. So, time invariant systems have time independent properties. But when, in terms of for instance an electric system one or more parts (e.g. a resistance) behaves noisy or changes otherwise in time (e.g. with ambient temperature, e.g. the thermistor, a temperature dependent resistor) then the system is time variant. Time-invariance is violated by aging effects that can change the outputs of analog systems over time (usually years or even decades).
In general, system theory considers linear time-invariant systems, the *LTI systems*.
Biological systems, described in terms of linear system analysis, are usually considered to be time-invariant when the time scale of investigation is restricted.

Deterministic and stochastic
An LTI system that will always exactly produce the same output for a given input is said to be deterministic. There are many methods of analysis developed specifically for *LTI* deterministic systems.
An LTI system that will produce slightly different outputs for a given input is said to be stochastic. Of coarse the average of many outputs should be the same as the output of its deterministic version. Its stochastic nature is caused by properties (e.g. a resistance or capacitor) that change in some unpredictable way in time. This means that the coefficients of the underlying differential equation are not constants. Thermal noise and other random phenomena ensure that the operation of any analog system, in fact every biological system, will have some degree of stochastic behavior. Despite these limitations, however, it is usually reasonable to assume that deviations from these ideals will be small.

The electric analogue
LTI systems are more easy to handle when they are studied as their electric analogues. To study them they are split, if possible, in a number of basic systems, be it first and second order systems. They can be arranged in parallel or serially.
First order systems always comprise an ideal resistance (the physical realization is an Ohmic resistor), and in addition one ideal capacitance (in practice a capacitor) or an ideal self-inductance (in practice a coil). Resistors are not ideal, they also have a small inductance and small capacity, and in the same way capacitors and coils are not ideal, but generally these imperfections can be ignored (not for the coil).
Second order systems comprise a resistance, capacitance and self-inductance. The electric behavior of these “building blocks” of linear systems is described in Electrical resistance, capacitance and self-inductance.

Application

Applications can be found in all branches of science and technology, and consequently also in (bio)medicine, but also in econometrics and social sciences.

More Info

As mentioned above, there are many methods of analysis developed specifically for LTI systems. This is due to their simplicity of specification. An LTI system is completely specified by its transfer function (which is a rational function for digital and lumped analog LTI systems). The transfer function (H) is actually the solution of the linear differential equation (for analog systems) or linear difference equation (for digital systems). However this solution is a function of frequency. The solution in the time domain (h(*τ*)) is the alternative description. This solution is very useful for two specific input signals, the unit impulse function (infinite at time zero and zero at any other time, with integral 1) and its integral, the unit step function.
The transfer function is not a function of the real frequency ω, but of the imaginary frequency jω. So, H(jω) is a function in the complex jω plane. To simplify notation, often Laplace notation, so H(s), is used.
From the transfer function the amplitude characteristic (eq. 5b) and phase characteristic (eq. 5c) can be derived. Together they give the frequency response. Which description, frequency or time domain, is most useful depends on the application.
The distinction between lumped and distributed LTI systems is important. A lumped LTI system is specified by a finite number of parameters, be it the zeros and poles of H(jω) (see System analysis) or the coefficients of its differential equation, whereas specification of a distributed LTI system requires more calculus (often a series expansion). Many bio-electric systems (brain, nerve axon) are actually distributed systems, but often they can be approximated by lumped systems.

Moiré pattern

Principle

A moiré pattern is an interference pattern created, for example, when two grids are overlaid at an angle, or when they have slightly different mesh sizes. Moiré patterns are often an undesired artifact of images produced by various digital imaging and computer graphics techniques, for example when scanning a halftone picture. This cause of moiré is a special case of aliasing (see Fourier transform and aliasing) due to undersampling a fine regular pattern.
Fig. 1a shows a moiré pattern. The interaction of the optical patterns of lines creates a real and visible pattern of thin and thick dark and light bands (see explanation below). More complicated line moiré patterns are created if the lines are curved or not exactly parallel. Moiré patterns revealing complicated shapes, or sequences of symbols embedded in one of the layers (in form of periodically repeated compressed shapes) are band and shape moiré patterns. One of the most important properties of a shape moiré pattern is its ability to magnify tiny shapes along either one or both axes, i.e. stretching. A common 2D example of moiré magnification occurs when viewing a chain-link fence through a second chain-link fence of identical design. The fine structure of the design is visible even at great distances.

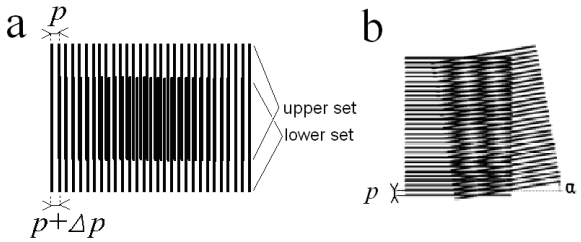


Fig. 1 Line moiré patterns. (a) The two sets of lines are superimposed in the mid-height of the figure. (b) Moiré pattern caused by rotation.

Explanation
Let us consider two patterns made of parallel and equidistant lines, e.g., vertical lines (Fig. 1a). The periodicity of the first pattern is p, the periodicity of the second is $p+\Delta p$, with $\Delta p > 0$. If the lines of the patterns are superimposed at the left of the figure, the shift between the lines increase when going to the right. After a given number of lines, the patterns are opposed: the lines of the second pattern are between the lines of the first pattern. Looking from a far distance gives pale zones when the lines are superimposed (there is white between the lines), and dark zones when the lines are "opposed". The middle of the first dark zone is obtained when the shift is equal to $p/2$. The n^{th} line of the second pattern is shifted by $n\Delta p$ compared to the n^{th} line of the first network. The middle of the first dark zone thus corresponds to $n\Delta p = p/2$. The distance d between the middle of a pale zone and a dark zone is $n = p/(2\Delta p)$. From this formula, we can see that:

- the larger p, the period width, the larger the distance between the pale and dark zones;
- the larger the discrepancy Δp the closer the dark and pale zones; a great spacing between dark and pale zones mean that the patterns have nearly the same periodicity.

Of course, when $\Delta p = p/2$, a uniformly grey figure results with no contrast.
The principle of the moiré pattern is similar to the Vernier scale.

Application

Applications are mainly technical. The moiré effect is used in strain measurement (see Elasticity 1: elasticity or Young's modulus): the pattern of the object as reference with superimposed the pattern of the deformed object by strain. A similar effect is obtained by the superposition of an holographic image (see Holography) of the object to the object itself: the hologram provides the reference and the difference with the object are the deformations, which appear as pale and dark lines.

More Info

In 2D there exist many types of moiré patterns. An example is that of a rotated pattern obtained by the superposition of two identical patterns rotated by an angle α (Fig. 1b). In its most simple form is made with two identical line patterns with the second pattern turned by an angle α. Seen from far, we can also see dark and pale lines: the dark lines correspond to the lines of nodes, that is, lines passing through the intersections of the two patterns. More complicated ones are obtained with curved lines. Taking a picture of a TV screen with a digital camera often produces severe moiré patterns.

Noise and thermal noise

Principle

In common use, the word noise means unwanted distribution of a signal or image, for instance in sound and the “snow” in an image (television, video, medical imaging techniques). It is generally generated in the electronics of equipment which produces sound or images.
In signal processing it can be considered data without meaning; that is, data that is not being used to transmit a signal, but is simply produced as an unwanted by-product. In statistics it is dispersion of data which makes it more difficult to draw conclusions. In information theory, however, noise is still considered to be information. In a broader sense, film grain or even advertisements in web pages can be considered noise.
Noise can block, distort, or change the meaning of a message in both human and electronic communication.
In many of these areas, the special case of thermal noise arises, which sets a fundamental lower limit to what can be measured or signaled and is related to basic physical processes at the molecular level.

In summary, there are many types of noise: thermal and electronic noise, acoustic and image noise, data noise, etc. In addition there is a distinction considering the frequency properties (i.e. the power spectral density, see Fourier transform, or amplitude spectrum, see Fourier analysis) as for instance white and pink noise. There is also a distinction considering the amplitude distribution of the noise, or better the noise signal, e.g. Gaussian noise, Poisson noise, binary noise (either 0 or 1), etc.

White noise
White noise is a random signal with a flat power spectral density. “White” is after the analogy with white light (all frequencies).
Pink noise or 1/f noise is a signal or process with a power spectral density that is proportional to the reciprocal of the frequency. Pink noise has an equal amount of energy per octave.

Acoustic noise
When speaking of noise in relation to sound, what is commonly meant is meaningless sound of greater than usual volume. Thus, a loud activity may be referred to as *noisy*. However, conversations of other people may be called noise for people not involved in any of them, and noise can be any unwanted sound such as the noise of e.g. dogs barking, traffic sounds spoiling the quiet. In audio, recording, and broadcast systems *audio noise* refers to the residual low level sound (usually hiss and hum) that is heard in quiet periods.

Application

Medical and human hearing White noise can be used to disorient individuals prior to interrogation and may be used as part of sensory deprivation techniques. White noise machines are used as sleep aids, as aid concentration by blocking out irritating or distracting noises in a person's environment and to mask tinnitus.

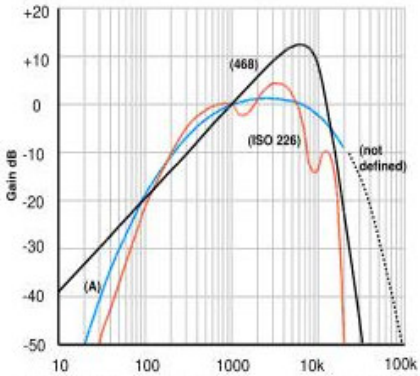


Fig. 1 Various grey weighing filters in use.

In audio engineering, noise can also refer to the unwanted residual electronic noise signal that gives rise to acoustic noise heard as 'hiss'. This noise is commonly measured using a specific u-shaped filter. This filter matches the spectrum of *Grey noise*, which is random noise with a psychoacoustic equal loudness. This means that, by approximation frequency bands with the width of the critical band at all frequencies

have equal loudness. (The critical bands are area's of about 1 mm at the basilar membrane which resonate at a pure tone.) An example of a filter is the inverted so-called A-weighing curve (curve A of Fig. 1.). In a concert hall or room, ideally, after filtering with grey noise filter a flat spectrum should remain. When the spectrum is not flat, by selective attenuation, it can be made flat (electronically, changing room acoustics, etc.). There are several grey noise filters to obtain equal loudness percepts. Filtering was formerly done by A-weighing, but nowadays in UK it is done with ITU-R 468 weighing. Elsewhere, ISO 226 is used, the most precise method (see Fig. 1), which very well matches the 40 phone curve.

General In the field of architectural acoustics, in order to dissemble distracting, undesirable noises in interior spaces, a low level of constant white noise is generated. White noise is commonly used in the production of electronic music. To set up the sound equalization (EQ) for a concert, a short burst of white or pink noise is released and monitored from various points in the venue so that the engineer can tell if the acoustics of the building naturally boost or cut any frequencies. Then the overall EQ is adjusted to ensure a balanced mix.
White noise can be used for frequency response testing of amplifiers and electronic filters. It will then automatically equalize each frequency band to get a flat response.
White noise is generally made with random number generators.

More Info

White, pink and red noise
The term white noise is also commonly applied to a noise signal in the spatial domain which has an autocorrelation (see Stochastic Signal Analysis) which can be represented by a delta function over the relevant space dimensions. The signal is then "white" in the spatial frequency domain.
Being uncorrelated in time does not, however, restrict the values a signal can take. Any distribution of values is possible (although it must have a zero mean). Even a binary signal which can only take on the values 1 or 0 will be white if the sequence of zeros and ones is statistically uncorrelated. Noise having a continuous distribution, such as a normal distribution, can of course be white. It is often incorrectly assumed that noise with a Gaussian amplitude distribution is necessarily white noise. However, neither property implies the other. Gaussianity refers to the way signal values are distributed, while the term 'white' refers only to the flat shape of the power spectral density.
An infinite-bandwidth, white noise signal is purely a theoretical construction. By having power at all frequencies, the total power of such a signal is infinite. In practice, a signal can be "white" with a flat spectrum over a defined frequency band.

Flicker noise, also known as 1/*f* noise, is a signal (or process) with a power spectral density that falls off reciprocally with f, yielding a pink spectrum. It occurs in almost all electronic devices, and results from a variety of effects, though always related to a direct current. The name arises from being intermediate between white noise ($1/f^0$) and red noise, $1/f^2$, more commonly known as Brownian noise (see Brownian motion).
Pink noise, the most common type of the more general 1/f noises with a power spectral density of the form: $S(f) \sim 1/f^{\alpha}$, where *f* is frequency and $0 < \alpha < 2$, with α usually close to 1. These "1/f-like" noises occur widely in nature (e.g. level of natural rivers and the sea, environmental temperature variations).

Electronic noise
Electronic noise exists in all circuits and devices as a result of thermal noise, also referred to as Johnson Noise. In any electronic circuit, there exist random variations in current or voltage caused by the random movement of the electrons carrying the current as they are jolted around by thermal energy. Lower temperature results in lower thermal noise. This same phenomenon limits the minimum signal level that any radio receiver can usefully respond to, because there will always be a small but significant amount of thermal noise arising in its input circuits. This is why sensitive medical apparatus use front-end low-noise amplifier circuits, and cooled with liquid nitrogen, such as for MRI, see MRI: general and Magnetoencephalography.

Thermal noise
Thermal noise is the electronic noise generated by the thermal agitation of the charge carriers (usually the
Electrons, but also ions) inside an electrical conductor at equilibrium, which happens regardless of any applied voltage.

Thermal noise is approximately white, meaning that the power spectral density is the same for all frequencies. Additionally, the amplitude of the signal has a Gaussian probability density function. Thermal noise is to be distinguished from shot noise which is caused by a finite number of energy carrying particles, e.g. electrons or photons). Shot noise is an additional current fluctuations that occurs at the very first instant when a macroscopic current starts to flow.

The power spectral density (proportional with voltage^2/Hz) of thermal noise is given by:

$$\overline{v}_n{}^2 = 4kTR, \quad (1)$$

where k is Boltzmann's constant (in J/K), T is the resistor's absolute temperature (K), and R is the resistor value (Ω). By approximation, at room temperature $\overline{v}_n$ is:

$$\overline{v}_n = 0.13\sqrt{R}\ (nV/\sqrt{Hz}) \quad (2)$$

For example, a resistor of 1 kΩ at an average temperature (300 K) has a resistor noise $\overline{v}_n$ of 4.07 nV/√Hz.
For noise at very high frequencies the above equation (1) does not apply. A more complicated formula, including Plank's constant is needed.

Radian and steradian

Radian

The circumference of a circle is $2\pi r$ with r the radius. The radian, an angle with symbol rad, is defined as the angle seen from the centre of the circle with an arc length of $1/(2\pi r)$. Any angle can be expressed in radians: $l_{arc}/(2\pi r)$, where l_{arc} is the arc length.

Steradian

The steradian (symbol: sr) is the derived SI unit of describing the 2D angular span in 3D-space, analogous to the way in which the radian describes angles in a plane. Just as the radian, the steradian is dimensionless. More precisely, the steradian is defined as "the solid angle subtended at the center of a sphere of radius *r* by a portion of the surface of the sphere having an area r^2." If this r^2 area is a circle, the solid angle is a simple cone subtending a half apex angle θ of $\cos^{-1}(1-1/2\pi) \approx 0.572$ rad or 32.77° (or with apex angle 65.54°).
Since the surface area of this sphere is $4\pi r^2$, then the definition implies that a sphere measures 4π steradians. By the same argument, the maximum solid angle that can be subtended at any point is 4π sr. A steradian can also be called a *squared radian.*

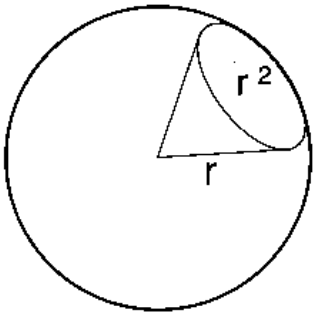


Fig. 1 A graphical representation of 1 steradian of a sphere with radius r.

A steradian is also equal to the spherical area with an angle of 1 radian. This area is equal to $1/4\pi$ of a complete sphere, or $(180/\pi)^2$ or 3282.8 square degrees.
As an example, radiant intensity can be measured in watts per steradian ($W \cdot sr^{-1}$).

Thermodynamics

Early in the industrial revolution, engineers tried to improve the efficiency of heat engines, such as steam engines, combustion engines but also refrigerators, which are all engines using some form of heat to do useful work. Fundamental research into general principles governing the efficiency of heat engines, such as diesel and gasoline engines, led to the branch of physics known as thermodynamics. It is the study of heat, thermal energy and work of a system. Thermodynamics deals only with the large-scale response of a system, which we can observe, and measure in experiments. Small-scale gas interactions are described by the kinetic theory of gases (see Compression and expansion).
There are three laws of thermodynamics and a more basic zero'th law. The concepts of entropy (see Thermodynamics: entropy) and enthalpy (see Thermodynamics: enthalpy) are also basic to thermodynamics.

Because knowledge of thermodynamics is now and then supposed in biomedical books and papers the first law (see Thermodynamics: first law), second law (see Thermodynamics: second law) and the concepts of entropy and enthalpy will be treated rather extensive. The zero'th law ((see Thermodynamics: zero'th law) and third law (see Thermodynamics: third law) are only discussed for completeness.

The chapter Thermodynamics for an ideal gas in the section **Gases and the Lung** gives in a table the various types of processes (isothermal, adiabatic, etc.), the relevant parameters (P, V, T, W, Q, U, S, *E*, n, R, c_p, c_v, γ) and their mathematical relationships.

Thermodynamics: enthalpy

Principle

In Thermodynamics and molecular chemistry, the enthalpy is a description of thermodynamic potential of a system, which can be used to calculate the heat transfer during a quasi-static process taking place in a closed thermodynamic system under constant pressure.
The enthalpy is defined by:

$$H = U + PV, \tag{1}$$

where (all units given in SI), *H* is the enthalpy (J), U is the internal energy, (J), P is the pressure of the system, (Pa) , V is the volume, (m^3). The expression PV is motivated by the following example of an isobaric process. Gas in a cylinder pushes a piston, maintaining constant pressure P. The work W done is W = Fx, where x is the displacement of the piston and the force on the piston is F = PA. Since Ax = V, W = PV, where P is a constant pressure and V the expansion of volume.

There exist a large number of numerical values of enthalpy changes under standard conditions (see Standard conditions for temperature and pressure).
The standard enthalpy change of vaporization, for example gives the enthalpy change when going from liquid to gas. These enthalpies are reversible; the enthalpy change of going from gas to liquid is the negative of the enthalpy change of vaporization. A common standard enthalpy change is the standard enthalpy change of formation, which has been determined for a large number of substances. The enthalpy change of any reaction under any conditions can be computed, given the standard enthalpy change of formation of all of the reactants and products.

Types of enthalpy in chemical processes under standard conditions

Standard enthalpy change of reaction is defined as the enthalpy change observed in a constituent of a thermodynamic system when, one mole of substance reacts completely.
Standard enthalpy change of formation is defined as the enthalpy change observed in a constituent of a thermodynamic system when, one mole of a compound is formed from its elementary antecedents under standard conditions**.**
Standard enthalpy change of combustion is defined as the enthalpy change observed in a constituent of a thermodynamic system, when one mole of a substance combusts completely with oxygen under standard conditions.

Standard enthalpy of atomization is defined as the enthalpy change required atomizing one mole of compound completely under standard conditions

Types of enthalpy in physical processes under standard conditions
Standard enthalpy of solution is defined as the enthalpy change observed in a constituent of a thermodynamic system, when one mole of a solute is dissolved completely in an excess of solvent under standard conditions
Standard enthalpy change of fusion (melting and freezing) is defined as the enthalpy change required to completely change the state of one mole of substance between solid and liquid states under standard conditions.
Standard enthalpy of vaporization is defined as the enthalpy change required to completely change the state of one mole of substance between liquid and gaseous states under standard conditions. *Example*: The energy needed to vaporize water, H_v at 100 °C is 40.66 kJ/mol, is more than five times the energy required to heat the same quantity of water from 0 °C.
Standard enthalpy of sublimation is defined as the enthalpy change required to completely change the state of one mole of substance between solid and gaseous states under standard conditions.
Standard enthalpy of denaturation is defined as the enthalpy change required denaturing one mole of compound under standard conditions.

More info

As an expansion of the first law of thermodynamics (see Thermodynamics: first law), enthalpy can be related to several other thermodynamic formulae. As with the original definition of the first law:

$$dU = \delta Q - \delta W, \quad (2)$$

where, δQ represents the infinitesimal amount of energy attributed or added to the system.
As a differential expression, with $\delta Q = T dS$ the value of H can be defined as:

$$dH = TdS + VdP, \quad (3)$$

where dS is the increase in entropy (J/K) (see Thermodynamics: entropy) with P constant.
In a more general form, the first law describes the internal energy with additional terms involving the chemical potential and the number of particles of various types. The differential statement for dH is then:

$$dH = TdS + VdP + \sum_i \mu_i dN_i. \quad (4)$$

where μ_i is the chemical potential for an i-type particle, and N_i is the number of such particles. It is seen that for an isobaric process, the VdP term is set to zero. When there is no change of the number of particles the $\mu_i dN_i$ terms are also zero, such that $dH = TdS$, the most simple interpretation of enthalpy.

The total enthalpy of a system cannot be measured directly; the *enthalpy change* of a system is measured instead. Enthalpy change (J) is defined by:

$$\Delta H = H_{final} - H_{initial}. \quad (5)$$

H_{final} is the final enthalpy of the system. In a chemical reaction, H_{final} is the enthalpy of the products.
$H_{initial}$ is the initial enthalpy of the system. In a chemical reaction, $H_{initial}$ is the enthalpy of the reactants.

An alternative description is to view the enthalpy in terms of entropy. With evaporation, the heat must be delivered by the surroundings to compensate for the increase in entropy when a liquid evaporated to gas. As the liquid and gas are in equilibrium at the boiling point (T_b), the Gibbs free energy $\Delta_v G = 0$, which leads to:

$$\Delta_v S = S_{gas} - S_{liquid} = \Delta_v H / T_b. \quad (6)$$

As neither entropy nor enthalpy vary greatly with temperature, it is normal to use the tabulated standard values without any correction for the difference in temperature from 298 K. (The enthalpy of water vaporization diminishes from 0 to 100 °C with 10%. The diminishing is caused by the increase of the kinetic energy.) A correction must be made if the pressure is different from 100 kPa, as the entropy of a gas is proportional to its pressure the entropies of liquids vary little with pressure, as the compressibility of a liquid is small.

For an exothermic reaction at constant pressure, the system's change in enthalpy is equal to the energy released in the reaction, including the energy retained in the system and lost through expansion against its surroundings. In a similar manner, for an endothermic reaction, the system's change in enthalpy is equal to the energy *absorbed* in the reaction, including the energy *lost by* the system and *gained* from

compression from its surroundings. A relatively easy way to determine whether or not a reaction is exothermic or endothermic is to determine the sign of ΔH. If ΔH is positive, the reaction is endothermic, that is heat is absorbed by the system due to the products of the reaction having a greater enthalpy than the reactants. On the other hand, if ΔH is negative, the reaction is exothermic, that is the overall decrease in enthalpy is achieved by the generation of heat.
Although enthalpy is commonly used in engineering and science, it is impossible to measure directly, as enthalpy has no reference point. Therefore, enthalpy can only accurately be used in a closed system. However, few real world applications exist in closed isolation, and it is for this reason that two or more closed systems cannot be compared using enthalpy as a basis, although sometimes this is done erroneously.

Thermodynamics: entropy

Principle

The thermodynamic entropy S is a measure of the disorder, randomness or chaos present in a system. An example is a broken cup that has less order and more chaos than an intact one. A highly random disordered system has more entropy. Heat or thermal energy causes individual atoms or molecules to increase their random motions. Increasing the internal heat energy of a system therefore increases its randomness, disorder, and thus the entropy. Gases, which are highly disorganized, have high entropy values. In contrast, solid crystals, the most organized form of matter, have very low entropy values.
Entropy can also be seen as a measure of the uniformity of the distribution of energy. Spontaneous changes tend to smooth out differences in temperature, pressure, density, and chemical potential that may exist in a system, and entropy is thus a measure of how far this smoothing-out process, the "dispersal" of energy, has progressed.
Entropy can also be seen as a variable (a so-called extensive state function) that accounts for the effects of irreversibility in thermodynamic systems.
The concept of entropy has evolved in the last 150 years and it has been interpreted from various disciplines, for instance astrophysics, chemistry, biochemistry and evolution biology. For these reasons, modern theory of entropy is certainly not simple and not easy to explain.
One of the notions of entropy is information entropy: it takes the mathematical concepts of statistical thermodynamics into areas of probability theory unconnected with heat and energy.

When a system's energy is defined as the sum of its "useful" energy, (e.g., that is used to push a piston), and its "useless energy", i.e., that energy which cannot be used for external work, then entropy may be visualized as the "useless" energy. The latter's energetic prevalence over the total energy of a system is directly proportional to the absolute temperature of the considered system.
Entropy, S, is not defined directly, but rather by an equation relating the change in entropy of the system to the change in heat of the system. For a constant temperature, the change in entropy, ΔS, is defined by the equation

$$\Delta S = \Delta Q/T, \qquad (1)$$

where $\Delta S = S_{final} - S_{initial}$, ΔQ is the net (supplied minus removed) motional energy ("heat") that is transferred reversibly to the system from the surroundings (or from another system in contact with the first system) divided by T, the absolute temperature at which the heat transfer occurs. During the isothermal and reversible transfer, the system goes from one state to another, and T is the absolute temperature at which the process is occurring. Eq. (1) directly shows that the increase in entropy is small when heat (this motional energy of the molecules) is added at high temperature and is greater when heat is added at lower temperature. Thus, for maximum entropy there is minimum availability for conversion into work and for minimum entropy there is maximum availability for conversion into work.

Processes can be reversible or irreversible (see also Thermodynamics: second law). Reversible implies that T has to stay the same while any energy is being transferred. An example is the melting of ice at 273.15 K. No matter what temperature the surroundings are, the temperature of the ice will stay at 273.15 K until the last molecules in the ice are changed to liquid water, i.e. until all the hydrogen bonds between the water molecules in ice are broken and new, less-exactly fixed hydrogen bonds between liquid water molecules are formed.

The example of melting ice

We consider a small 'universe', consisting of the 'surroundings' (a warm room) and 'system' (a glass with ice and cold water). In this universe, some heat energy δQ from the room (at 298 K or 25 °C) is transferred (dispersed) to the ice and water at its constant temperature T of 273 K (0 °C). The entropy of the system will change by the amount dS = δQ/T, in this example δQ/273 K. (The heat *δQ* for this process is the energy required to change water from the solid state to the liquid state, and is called the Enthalpy of fusion, i.e. the ΔH for ice fusion, being 6008 J/mol.) The entropy of the surroundings will decrease by an amount dS = −δQ/298 K. So in this example, the total entropy of the universe, S_{ice} + S_{room} increases. This is always true in spontaneous events in a thermodynamic system: the final net entropy after such an event is always greater than was the initial entropy.

An isentropic process occurs at constant entropy. For a reversible process, this is identical to an adiabatic process (see below). If a system has an entropy, which has not yet reached its maximum equilibrium value, a process of cooling may be required to maintain that value of entropy.

More Info

Internal energy is a property of the system whereas work done or heat supplied is not. In addition, for a reversible process, the total amount of heat added to a system can be expressed as δQ = TdS where T is temperature and S is entropy. Therefore, for a reversible process:

$$dU = TdS - PdV. \quad (2)$$

Since U, S and V are thermodynamic functions of state, the above relation holds also for non-reversible changes. The above equation is known as the fundamental thermodynamic relation.

In the case where the number of particles in the system is not necessarily constant and may be of different types, the first law (see Thermodynamics: first law) is extended to the *thermodynamic fundamental equation*:

$$dU = \delta Q - \delta W + \sum_i \mu_i dN_i = TdS - PdV + \sum_i \mu_i dN_i. \quad (3)$$

where $\sum_i$ is the sum over i types of particles, dN_i is the (small) number of type-i particles added to the system, and μ_i is the amount of energy added to the system when one type-i particle is added. The energy of that particle is such that the volume and entropy of the system remains unchanged. μ_i is known as the chemical potential of the type-i particles in the system.
When the process is not isothermally, then the mathematics becomes more complicated.

The concept of energy is central to the first law of thermodynamics (see Thermodynamics: first law), which deals with the conservation of energy and under which the loss in heat will result in a decrease in the internal energy of the thermodynamic system. Thermodynamic entropy provides a comparative measure of the amount of this decrease in internal energy of the system and the corresponding increase in internal energy of the surroundings at a given temperature.
Entropy provides a measure of the extent to which a heat engine can never completely recycle unused heat into work, but will always convert some of the heat into internal energy due to intermolecular interactions, which is not available to do work.
The concept of thermodynamic entropy is central to the second law of thermodynamics (see Thermodynamics: second law), which deals with the occurrence of spontaneously physical processes.

In a general sense the second law says that temperature differences between thermodynamic systems in contact with each other tend to even out and that work can be obtained from these non-equilibrium differences, but that loss of heat occurs, in the form of entropy, when work is done.
Entropy is also a measure of the instantaneous amount of energy in a physical system that cannot be used to do work without affecting the absolute temperature T of the system.
When a substance in a particular state of aggregation receives motional molecular energy (“heat”) from the surroundings, its temperature is raised, making its molecules move faster.
The energy being transferred from $T_{initial}$ to T_{final}, is directly given by:

$$\Delta S = C_p \cdot \ln(T_{initial}/T_{final}), \quad (4)$$

where C_p·is heat capacity with constant pressure.
Entropy is one of the three basic thermodynamic potentials: U (internal energy), S (entropy) and A (Helmholtz energy, physical useful energy). They are as follows related:

$$A(T,V) = U(S,V) - TS, \quad (5)$$

where V is volume.

The internal energy comprises kinetic energy due to the motion of molecules (translational, rotational, vibrational), the potential energy (vibrational and electric energy of atoms), the energy in all the chemical bonds and the energy of the free, conduction electrons in metals.

In chemistry the useful energy is the Gibbs free energy G:

$$G(T,p) = U + pV - TS, \quad (6)$$

where p is pressure.

U + pV is the enthalpy (see Thermodynamics: Enthalpy).

Statistical mechanics introduces calculation of entropy using probability theory to find the number of possible microstates at an instant. Now, entropy is defined as:

$$S = k \cdot \ln\Omega, \text{ or}$$
$$S = k \cdot \ln(1/p), \quad (7)$$

where k is Boltzmann's constant (1.38066×10^{-23} $J \cdot K^{-1}$) and Ω is the number of microstates for a given thermodynamic macrostate. A microstate is one of the many microscopic configurations of a system. At any given instant, the system is in one of these microstates, each of which is equally probable with probability *p* if the system is in equilibrium. Eq. (5) is purely theoretically since Ω can not be calculated. Moreover, it is ΔS which is of interest.
Statistical mechanical entropy is mathematically similar to Shannon entropy, which is part of information theory, where energy is not involved. This similarity means that some probabilistic aspects of thermodynamics are replicated in information theory.

Thermodynamics: zero'th law

Principle

If two thermodynamic systems are each in thermal equilibrium with a third, then they are in thermal equilibrium with each other, or in equations:

$$\text{if } T(A) = T(B),$$
$$\text{and } T(B) = T(C),$$
$$\text{then } T(A) = T(C), \quad (1)$$

where A, B and C are the systems and T denotes there absolute temperature.
When two systems are put in contact with each other, there will be a net exchange of energy between them unless or until they are in thermal equilibrium. This means that they contain the same amount of thermal energy for a given volume. In other words, they have the same temperature.
While this is a fundamental concept of thermodynamics, the need to state it explicitly as a law was not perceived until the first third of the 20th century, long after the first three laws were already widely in use, hence the zero numbering.

Thermodynamics: first law

Principle

The first law of thermodynamics is an expression of the more universal physical law of the conservation of energy: the increase in the internal energy ΔU of a system is equal to the amount of energy Q added by heating the system, minus the amount lost as a result of the work W done by the system on its surroundings. In formula:

$$\Delta U = Q - W. \quad (1a)$$

The internal energy is just a form of energy like the potential energy of an object at some height above the earth, or the kinetic energy of an object in motion. In the same way that potential energy can be converted to kinetic energy while conserving the total energy of the system, the internal energy of a thermodynamic system can be converted to either kinetic or potential energy. Like potential energy, the internal energy can be stored in the system. *Notice, however, that heat and work cannot be stored or conserved independently since they depend on the process.* The first law of thermodynamics allows for many possible states of a system to exist, but only certain states are found to exist in nature. The second of thermodynamics helps to explain this limitation. (States are the values of the state variables. In case of gases, the common ones are P, V and T.)

The amount of heat transferred into, or from a gas also depends on the initial and final states and the process, which produces the final state. The difference of the heat flow into the gas and the work done by the gas depends only on the initial and final states of the gas and does **not** depend on the process or path, which produces the final state. This indicates the existence of an additional variable, called the ***internal energy*** of the gas, which depends only on the state of the gas and not on any process. The internal energy is a state variable, just like the temperature or the pressure. The first law of thermodynamics defines the internal energy (U) as equal to the difference of the heat transfer (Q) ***into*** a system and the work (W) done ***by*** the system.

$$U_2 - U_1 = Q - W \qquad (1b)$$

The first law allows the possibility of a heat engine (for instance an ideal combustion engine) or other system in which the useful work output equals the energy input. In this case, the engine is 100% efficient. This is illustrated in Fig. 1.

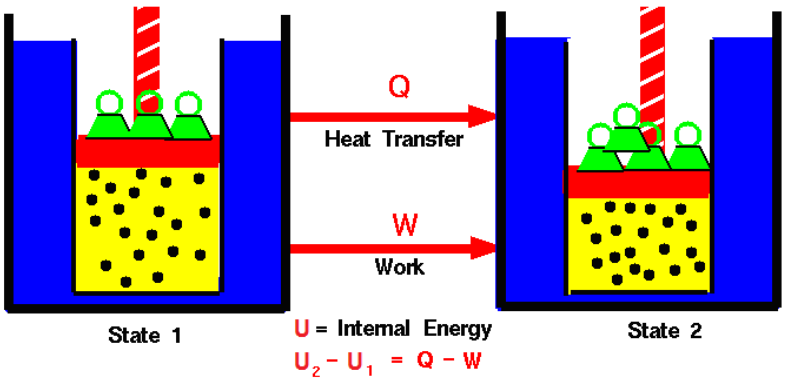


Fig. 1 Illustration of the first law.

The first law of thermodynamics basically states that a thermodynamic system can store or hold energy and that this internal energy is conserved. Heat transfer is a process by which energy from a high-temperature source is added to a low-temperature sink. In addition, energy may be lost by the system when it does mechanical work ***on*** its surroundings, or conversely, it may gain energy as a result of work done on it ***by*** its surroundings.

If a system is fully insulated from the outside environment, it is possible to have a change of state in which no heat is transferred into the system. A process, which does not involve heat transfer, is called an ***adiabatic*** process. The implementation of the first law of thermodynamics for gases introduces another useful state variable called the enthalpy H, the thermodynamic potential (see Thermodynamics: Enthalpy).

Types of basic thermodynamic processes

Relations between the classical thermodynamic variables of the gas laws, P, V, and T are often specified by holding one thermodynamic variable constant while the remaining two vary in a conjugated way.

An *isobaric process* occurs at constant pressure. An example would be to have a movable piston in a cylinder, so that the pressure inside the cylinder is always at atmospheric pressure, although it is isolated from the atmosphere. In other words, the system is dynamically connected, by a movable boundary, to a constant-pressure reservoir. Like a balloon the volume is contracting when the gas inside is heated. (For an ideal gas it holds that V/T = constant = PnR, where n is the number of moles and R the ideal gas constant). Since the volume changes work is done.

An *isochoric (or isovolumetric)* process is one in which the volume is held constant, meaning that the work done by the system will be zero. It follows that, for the simple system of two dimensions, any heat energy transferred to the system externally will be absorbed as internal energy. An isochoric process is also known as an isometric process. An example would be to place a closed tin can containing only air into a fire. To a first approximation, the can will not expand, and the only change will be that the gas gains internal energy, as evidenced by its increase in temperature and pressure (P/T = constant) or mathematically $\Delta Q = \Delta E$. We may say that the system is dynamically insulated, by a rigid boundary, from the environment
The temperature-entropy conjugate pair is concerned with the transfer of thermal energy as the result of heating.
An isothermal process occurs at a constant temperature. An example would be to have a system immersed in a large constant-temperature bath. Any work energy performed by the system will be lost to the bath, but its temperature will remain constant. In other words, the system is thermally connected, by a thermally conductive boundary to a constant-temperature reservoir (PV = constant).
An adiabatic process is a process in which there is no energy added or subtracted from the system by heating or cooling. For a reversible process, this is identical to an isentropic process. We may say that the system is thermally insulated from its environment and that its boundary is a thermal insulator. If a system has an entropy, which has not yet reached its maximum equilibrium value, the entropy will increase even though the system is thermally insulated. (See also Adiabatic compression and expansion.)

More Info

The above processes have all implicitly assumed that the boundaries are also impermeable to particles, such that the involved number of particles is constant. We may assume boundaries that are both rigid and thermally insulating, but are permeable to one or more types of particle. This will result in another number of processes.

The law of the conservation of energy in equation with infinitesimal terms is:

$$dU = \delta Q - \delta W. \quad (2)$$

dU is internal energy of a system equal to δQ the amount of energy added by heating the system, δW the amount lost as a result of the work done by the system on its surroundings. The infinitesimal heat and work are denoted by δ are not state functions but rather they are processes by which the internal energy is changed. In other words, they do not describe the state of any system. In mathematical terms, they are not exact differentials. The integral of an inexact differential depends upon the particular "path" taken through the space of thermodynamic parameters while the integral of an exact differential depends only upon the initial and final states.

When the system (.g. gas) expands the work done on the system is –PdV whereas in the previous formulation of the first law, the work done by the gas while expanding is PdV. In any case, both give the same result when written explicitly as:

$$dU = \delta Q - PdV. \quad (3)$$

A more general definition of the first law respects the internal energy: $\Delta U = Q + W + W'$, (4)
Where ΔU is the change in internal energy of a system during a process, *Q* is heat *added to* a system (J); that is, a positive value for *Q* represents heat flow *into* a system while a negative value denotes heat flow *out of* a system, *W* is the mechanical work done *on* (positive) or done *by* (negative) a system. W' is energy added by all other processes.
The first law may be stated equivalently in infinitesimal terms as:

$$dU = \delta Q + \delta W + \delta W', \quad (5)$$

where the terms now represent infinitesimal amounts of the respective quantities.

Literature
http://www.grc.nasa.gov/WWW/K-12/airplane/thermo2.html

Thermodynamics: second law

Principle

General

In simple terms the laws says that:

- Clausius (German physicist) statement: heat generally cannot spontaneously flow from a material at lower temperature to a material at higher temperature.
- In a system, a process that occurs will tend to increase the total entropy of the universe.
- Kelvin statement: it is impossible to convert heat completely into work in a cyclic process.

The Carnot cycle of a heat engine

A heat engine is any engine that uses heat energy to do some form of useful work. A heat engine absorbs thermal energy, or heat, from a heat source known as the hot reservoir and rejects waste heat to what is known as the cold reservoir. For example a steam engine operating by using a fire to boil water, which moves a piston, has the heated steam as the hot reservoir and earth's atmosphere, absorbing the waste heat, as the cold reservoir.
A heat engine absorbs energy from the hot reservoir, at a high constant temperature. At a different part of the cycle, called the Carnot cycle of the engine, it releases energy to the cold reservoir, at a lower constant temperature. It does not change its temperature while either absorbing or releasing thermal energy. This type of heat or energy transfer is isothermal or diabatic. The temperature changes occur when the engine is neither absorbing nor releasing thermal energy and these parts of the process are called adiabatic. During these parts expansion and compression occurs. The Carnot cycle is ideal when it allows 100% efficiency. This means that all supplied energy is transformed to work, but as the Kelvin statement says, this is impossible. It is close to ideal when the waste heat is dumped into a heat reservoir at a temperature of nearly absolute zero. In practice even that temperature is never reached.

A Carnot cycle engine absorbs and releases energy and has a compression and expansion phase. The four phases can be described as follows.
1. Reversible isothermal expansion of the gas at the "hot" temperature, T_1 (isothermal or diabatic heat addition). During this step (indicated by 1 in Fig. 1 and 2) the expanding gas causes the piston to do work on the surroundings. The gas expansion is propelled by absorption of quantity Q_1 of heat from the high temperature reservoir.
2. Reversible adiabatic expansion of the gas. For this step we assume the piston and cylinder are thermally insulated (closed system), so that no heat is gained or lost. Therefore, this step is adiabatic. During this step the temperature drops to T_2.
3. Reversible isothermal compression of the gas at the "cold" temperature, T_2. (isothermal heat rejection) Now the surroundings do work on the gas, causing quantity Q_2 of heat to flow out of the gas to the low temperature reservoir.
4. Adiabatic compression of the gas. Once again, we assume the piston and cylinder are thermally insulated. During this step, the surroundings do work on the gas, compressing it and causing the temperature to rise to T_1.
Fig. 1 gives the Carnot cycle with the four steps in a P-V diagram.

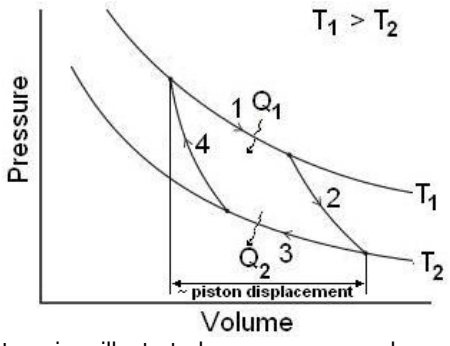


Fig. 1 A Carnot cycle of a heat engine, illustrated on a pressure-volume diagram to illustrate the work done. The piston displacement is proportional to the volume change. The vertical lines indicate the maximal volume change.

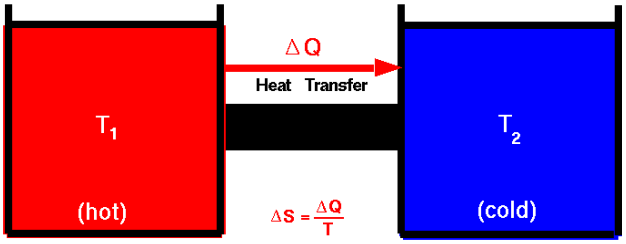


Fig. 2 Concept of the entropy change ΔS.

Entropy and the second law of thermodynamics

According to the first law (see Thermodynamics: first law) we can imagine a system, in which the heat is instead transferred from the cold object to the hot object. The cold object gets colder and the hot object gets hotter, but energy is conserved. Obviously such a system does not exist in nature and to explain this and similar observations, a second law of thermodynamics was proposed. For this law a new state variable is introduced called entropy. Entropy (S) has a variety of physical interpretations (see Thermodynamics: entropy), including the statistical disorder of the system. For our purposes, let us consider entropy to be just another property of the system, like pressure and temperature.
The change in entropy ΔS is equal to the heat transfer ΔQ divided by the temperature T.

$$\Delta S = \Delta Q/T \qquad (1)$$

For a given physical process, the combined entropy of the system and the environment remains constant if the process can be reversed. If we denote the initial and final states of the system by "i" and "f":

Reversible process: $S_f = S_i$
An example of a reversible process is ideally forcing a gas flow through a constricted pipe. Ideal means no boundary layer losses. As the flow moves through the constriction, the pressure, temperature and velocity change, but these variables return to their original values downstream of the constriction. The state of the gas returns to its original conditions and the change of entropy of the system is zero. Engineers call such a process an isentropic process (i.e. constant entropy).

Irreversible process: $S_f > S_i$
The second law states that if the physical process is irreversible, the combined entropy of the system and the environment must *increase*. The final entropy must be greater than the initial entropy for an irreversible process:
An example of an irreversible process is putting a hot object (1) in contact with an identical cold object (2). Eventually, they both achieve the same equilibrium temperature. If we then separate the objects, they remain at the equilibrium temperature and do not naturally return to their original temperatures. The process of bringing them to the same temperature is irreversible. In addition, the sum of heir entropies is increased. This can easily be understood when it is assumed that both reservoirs are identical except for their temperature (as illustrated in Fig. 2). Since $\Delta S_1 = -\Delta Q/T_1$) and $\Delta S_2 = \Delta Q/T_2$, it follows that $\Delta S_{total} = \Delta S_1 + \Delta S_2 > 0$. (Actually one has to calculate the integral over $\delta Q/T$.)

Application

Application is numerous, for example in the practice of combustion engines and refrigerators and for instance in medical research, which needs strong cooling.

More info

The first and second law can be combined to yield the Fundamental Thermodynamic Relation:

$$dU = TdS - PdV. \qquad (1)$$

Here, U is energy, T is temperature, S is entropy, P is pressure, and V is volume
The useful work that a heat engine can perform will therefore always be less than the energy put into the system.
There will always be some waste heat, for instance caused by some type of friction or other inefficiency. Engines must be cooled, as a radiator cools a car engine, because they generate waste heat. The maximum possible efficiency for an ideal heat engine is given by one Carnot cycle of compression and decompression. In the ideal case, according to the second law this is close to 100%. Thermodynamic engineers strive for this ideal but do not achieve it in real engines, like combustion or steam engines.

More insight in the second law is obtained by means of the T-S diagram of Fig. 3. In this diagram the heat energy is visualized by the area of a closed Carnot cycle or mathematically the circular integral of the loop:

$$Q = \oint TdS \,. \tag{2}$$

The second law of thermodynamics simply states that it is not possible to reach a temperature T_2 of absolute zero. In that case, there is no energy waste and thus all provided heat energy is transferred to mechanical energy. Now, the Carnot cycle engine reaches 100% efficiency. The actual efficiency of the engine is the area of the rectangle BCDE ($T_2 > 0$) divided by the area ACDF (ideal: $T_2 = 0$ K) of Fig. 3b, or expressed more simple:

$$\text{Efficiency (\%)} = 100(T_1 - T_2)/T_2 = 100(1 - T_2/T_1). \tag{3}$$

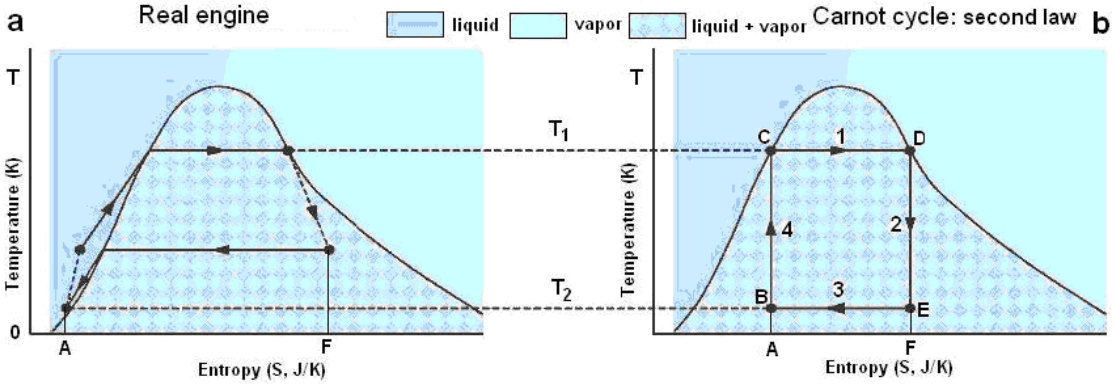


Fig 3 A Carnot cycle acting as a heat engine, illustrated on a temperature-entropy diagram. The cycle takes place between a hot reservoir at temperature T_1 and a cold reservoir at temperature T_2.

Fig. 3a shows the T-S diagram of a heat engine, which basically follows the second law. In addition to the imperfection of not reaching $T_2 = 0$ K, as the second law says, there are more imperfections. The area of the closed loop of fig. 1a has become smaller than area BCDE of Fig. 3b as is in accordance with the second law. The steps 2 and 4 of the cycle (see Fig. 1b) are not purely adiabatic since the lines indicating these steps are oblique instead of vertical. In addition, in practice the steps 1 and 3 are only nearly isothermal since the thermal isolation is imperfect.

Literature

http://www.grc.nasa.gov/WWW/K-12/airplane/thermo2.html

Thermodynamics: third law

Principle

The most common enunciation of third law of thermodynamics is:
As a system approaches absolute zero temperature, all processes cease and the entropy of the system approaches a minimum value. Ideally, this minimum can be zero in a perfect, pure crystal.

This law, the most theoretical of the four laws of thermodynamics, enters the field of atom physics and quantum numbers, which is outside the scope of this compendium.

Suspension

Principle

A suspension is a colloidal (see Colloid) dispersion (mixture) in which a finely-divided substance is combined with another substance, with the former being so finely divided and mixed that it doesn't rapidly settle out. In everyday life, the most common suspensions are those of solids in liquid water. Basically, all Emulsion (immiscible fluid in fluid mixtures) are also suspensions.
A suspension of liquid droplets or fine solid particles in a gas is called an *aerosol* (see Colloid). In the atmosphere these consist of fine dust, sea salt, cloud droplets etc.

Application

Suspensions are widely applied in medicine, daily life and industry. Common examples are ice cream, a suspension of microscopic ice crystals in cream. Mud or muddy water, is an emulsion where solid particles (sand, clay, etc.) are suspended in water. Paint is an emulsion where the solid pigment particles are mixed in a water-like or organic liquid.

More Info

The amount of energy determines the maximum size of particle that can be suspended. In the absence of additional energy (agitation), all particles down to colloidal size will eventually settle out into a distinct phase. Suspensions separate over some period of time, solutions never because the intermolecular forces between the different types of molecules are of similar strength to the attraction between molecules of the same type.

Wheatstone bridge

Principle

A Wheatstone bridge is used to measure an unknown electrical resistance by balancing two legs of a bridge circuit, one leg of which includes the unknown component.

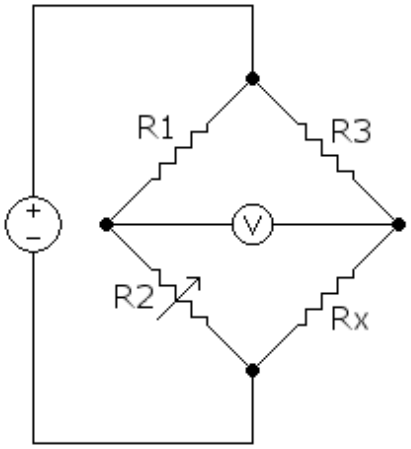


Fig. 1 Principle of the Wheatstone bridge

In Fig. 1, R_x is the unknown resistance to be measured; R_1, R_2 and R_3 are resistors of known resistance and the resistance of R_2 is adjustable (a potentiometer). If the ratio of the two resistances in the known leg (R_2 / R_1) is equal to the ratio of the two in the unknown leg (R_x / R_3), then the voltage between the two midpoints will be zero and no current will flow between the midpoints. The voltage is measured by a sensitive voltage meter, a sensitive galvanometer. R_2 is varied until this condition is reached. The current direction indicates if R_2 is too high or too low.
Detecting zero current can be done to extremely high accuracy. Therefore, if R_1, R_2 and R_3 are known to high precision, then R_x can be measured to high precision. Very small changes in R_x disrupt the balance and are readily detected.
At the balance point the value of R_x can be found from:
$R_x = R_2 \cdot R_3 / R_1$.
Variations on the Wheatstone bridge can be used to measure the capacitance of a condenser or the inductance of a coil.

Application

It is typically applied in biomedical apparatus, but mostly hidden for the user. In basic research, e.g. in the field of bio-electricity manually a minimum of some quantity can be adjusted.

Structure of matter and molecular phenomena

Adhesion

Principle

Molecules in liquid state experience strong intermolecular attractive forces. When those forces are between unlike molecules, they are said to be adhesive forces. The adhesive forces between water molecules and the walls of a glass tube are stronger than the cohesive forces (the attractive forces between like molecules). This leads to an upward turning meniscus at the walls of the vessel (see Capillary action).
The attractive forces between molecules in a liquid can be seen as residual electrostatic forces and are called van der Waals forces or van der Waals bonds (see Cohesion).
More generally, and from a macroscopically point of view, adhesion is the molecular attraction exerted between bodies in contact.
Notice that *in medicine*, an adhesion has a completely other meaning. It is a fibrous band of scar tissue that binds together normally separate anatomical structures. It usually occurs as a result of surgery, infection, trauma or radiation.

Application

Adhesion is of particular interest to (medical) biologists to understand the workings of cells and to engineers who wish to stick objects together.

More Info

Five mechanisms have been proposed to explain why one material sticks to another.

- *Mechanical adhesion*
  Two materials may be mechanically interlocked. Sewing forms a large-scale mechanical bond, Velcro forms one on a medium scale, and some textile adhesives form one at a small scale.
- *Chemical adhesion*
  Two materials may form a compound at the join. The strongest joins are where atoms of the two materials swap (ionic bonding) or share (covalent bonding) outer electrons. A weaker bond is formed if oxygen, nitrogen of fluorine atoms of the two materials shares a hydrogen nucleus (hydrogen bonding).
- *Dispersive adhesion*
  This is also known as adsorption. Two materials may be held together by van der Waals forces.
- *Electrostatic adhesion*
  Some conducting materials may pass electrons to form a difference in electrical charge at the join. This results in a structure similar to a capacitor and creates an attractive electrostatic force between the materials.
- *Diffusive adhesion*
  Some materials may merge at their interface by diffusion (see Diffusion: general). This may occur when the molecules of both materials are mobile and soluble in each other. This would be particularly effective with polymer chains where one end of the molecule diffuses into the other material. It is also the mechanism involved in sintering. When metal or ceramic powders are pressed together and heated, atoms diffuse from one particle to the next. This joins the particles into one.

What makes an adhesive bond strong?
The strength of the adhesion between two materials depends on which of the above mechanisms occur between the two materials, and the surface area over which the two materials contact. Materials that wet against each other tend to have a larger contact area than those that do not. Wetting depends on the surface energy (the disruption of chemical bonds that occurs when a surface is created) of the materials.

Brownian motion

Principle

Brownian motion is the random movement of particles suspended in a fluid. However, it is also a mathematical model, often called a Wiener process (continuous-time stochastic processes such as white noise generation) with slightly different properties than the physical process. The mathematical model has no direct relevance for medical physics and biophysics, and therefore will be discussed briefly.

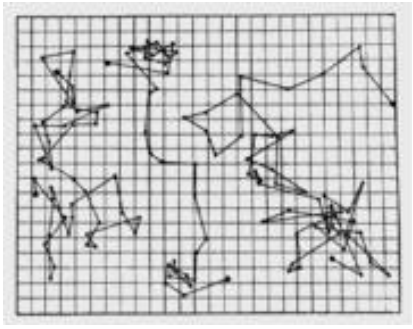


Fig. 1 Intuitive metaphor for Brownian motion

It is believed that Brown was studying pollen particles floating in water under the microscope. The pollen, as holds for all other kind of small particle are swimming randomly in water. One molecule of water is about 0.1 to 0.2 nm, (a hydrogen-bonded cluster of 300 atoms has a diameter of approximately 3 nm) where the pollen particle is roughly 1 µm in diameter, roughly 10,000 times larger in diameter than a water molecule. So, the pollen particle can be considered as a very large balloon constantly being pushed by water molecules. The Brownian motion of particles in a liquid is due to the instantaneous imbalance in the force exerted by the small liquid molecules on the particle.

Particles of a liquid (also a gas, so actually in a fluid) will show random movements due to the collisions with other liquid particles. Therefore, their displacement in a certain time t is less then linear with t. This also holds for the large molecules diluted or suspended in liquid. They make some 10^{13}-10^{15} collisions per second with liquid molecules. It appeared that the mean square displacement in the x-direction is proportional with t:

$$\overline{\Delta x^2} = Ct\,. \qquad (1)$$

C is a constant to be solved. A macromolecule with velocity v is subjected to a friction force F_f :

$$F_f = -fv, \qquad (2)$$

where f is the friction coefficient of the particle.
According to Stokes' law (see Stokes' law and hematocrit), applied to *spherical* particles f equals:

$$f = 6\pi\eta r, \qquad (3)$$

where η the dynamic viscosity of the medium and r the radius of the Brownian particle. This finally yields:

$$\overline{\Delta x^2} = \frac{kT}{3\pi\eta \cdot r}t\,, \qquad (4)$$

where k is Boltzmann's constant and T absolute temperature. t should be > 1 µs.
Example
Equation (4) applied to a blood platelet, supposing that it is by approximation spherical of shape (whereas in reality it is rather disk-like) with a radius of 0.001 mm, T = 310 K, and η of plasma is 2.0 mPa·s (20 ^{0}C) and t = 1, we get $\sqrt{\overline{\Delta x^2}} = 0.48$ µm per second, which should be visible under a light microscope. (Often reality is more complicated: plasma behaves as a so called non-Newtonian fluid. Applying a high stress means a high viscosity and a low stress a low viscosity. Consequently, the plasma under high pressure should show smaller Brownian motion.)

Application

Brownian motion occurs in cells, alive or dead. Experimental biomedical applications are found in nanomedicine and special microscopic techniques (e.g. size estimation, mobility of fluorescent macromolecules, estimating diffusion coefficients).

Mathematical models (Wiener models) have several real-world applications, for example stock market fluctuations.

More Info

The diffusion equation (see Diffusion: general) yields an approximation of the time evolution of the probability density function associated to the position of the particle undergoing a Brownian movement under the physical definition. The approximation is valid on short timescales and can be solved with the Langevin equation, a stochastic differential equation describing Brownian motion in a potential involving a random force field representing the effect of the thermal fluctuations of the solvent on the particle. The displacement of a particle undergoing Brownian motion is obtained by solving the diffusion equation under appropriate boundary conditions. This shows that the displacement indeed varies as the square root of the time, not linearly.
Brownian motion always goes on. This is not in conflict with the first law of thermodynamics (see Thermodynamics: first law), since the kinetic energy of the medium particles provide continuously energy. This will result in local cooling, but this is balanced by heat transport from areas with a higher temperature. In this way the 2^{nd} law of thermodynamics (see Thermodynamics: second law) is not violated. In other words, only locally and temporally the entropy (see: Thermodynamics: entropy) is diminished (but can also be enlarged), but the entropy of the total system will not change.

Mathematical models of Brownian motion
One-dimensional random walk takes place on a line. So, one starts at zero, and at each step one moves by a fixed amount along one of the two directions from the current point, with the direction being chosen randomly. The average straight-line distance between start and finish points of a one-dimensional random walk of n steps is $\pi^{-1}\sqrt{(2n)} \approx 0.8\sqrt{n}$. If "average" is understood in the sense of root-mean-square, then the average distance after n steps is $\sqrt{n}$ times the step length exactly, just as the time t of the physical Brownian movement.
Brownian motion is among the simplest continuous-time stochastic processes, and it is a limit of both simpler and more complicated stochastic processes as 3D random walk.

Capillary action

Principle

Capillary action described in the Cohesion-tension theory is considered a mix of cohesion and adhesion. Capillary action or capillarity (also known as capillary motion) is the ability of a substance to draw a liquid upwards against the force of gravity. The standard reference is a tube in plants but capillarity can be seen readily with porous paper. It occurs when the adhesive intermolecular forces between the liquid and a solid are stronger than the cohesive intermolecular forces within the liquid. The effect causes a concave meniscus to form where the liquid is in contact with a vertical surface. The same effect is what causes porous materials to soak up liquids.
A common object used to demonstrate capillary action is the *capillary tube*. When the lower end of a vertical glass tube is placed in a liquid, such as water, a concave meniscus forms. Surface tension pulls the liquid column up until there is a sufficient weight of liquid for gravitational forces to come in equilibrium with intermolecular adhesive forces. The weight of the liquid column is proportional to the square of the tube's diameter, but the contact area between the liquid and the tube is proportional only to the diameter of the tube, so a narrow tube will draw a liquid column higher than a wide tube. For example, a glass capillary tube 0.5 mm in diameter will lift a theoretical 28 mm column of water. (Actual observations show shorter total distances.) In addition, the angle of incidence (contact angle) can be calculated exactly (see the textbooks of physics).
With some materials, such as mercury in a glass capillary, the interatomic forces within the liquid exceed those between the solid and the liquid, so a convex meniscus forms and capillary action works in reverse. Now the liquid level in the tube is lower.

Application

In medicine Capillary action is also essential for the drainage of constantly produced tear fluid from the eye. Two cannaliculi of tiny diameter are present in the inner corner of the eyelid; their openings can be visualized with the naked eye when the eyelids are everted.
In botany A plant makes use of capillary force to draw water into its system (although larger plants also require transpiration to move a sufficient quantity of water to where it is required).
In hydrology capillary action describes the attraction of water molecules to soil particles.
In chemistry Chromatography utilizes capillary action to move a solvent vertically up in a plate or paper. Dissolved solutes travel with the solvent at various speeds depending on their polarity. Paper sticks for urine and pH tests are also applications.
In daily life Towels (fluid transfer from a surface) and sponges (the small pores) absorb liquid through capillary action. Some modern sport and exercise fabrics use capillary action to "wick" sweat away from the skin.

More info

The height h of a liquid column (m) is given by:

$$h = 2\gamma\cos\theta/\rho g r, \qquad (1)$$

where:
γ = surface tension (J/m² or N/m)
θ = contact angle, this is the angle between the meniscus (at the wall) and the wall
ρ = density of liquid (kg/m^3)
g = acceleration due to gravity (m/s²)
r = radius of tube (m)
For a water-filled glass tube in air at sea level,
γ = 0.0728 J/m² at 20 °C, θ = 20° (0.35 rad), ρ = 1000 kg/m^3 and g = 9.8 m/s². And so the height of the water column is given by:

$$h \approx 1.4\cdot10^{-5}/r \qquad (2)$$

Thus in a 2 m wide capillary tube, the water would rise an unnoticeable 0.014 mm. However, for a 1 mm wide tube, about the size of a hematocryte capillar, the water would rise 7 mm, and for a tube with radius 0.1 mm, the water would rise 14 cm.

Cohesion

Principle

Cohesion or cohesive attraction or cohesive force is the intermolecular attraction between (nearly) identical molecules. The cohesive forces between liquid molecules are responsible for phenomena such as Surface tension and capillary force.
Molecules in liquid state experience strong intermolecular attractive forces. When those forces are between like molecules, they are called cohesive forces. For example, cohesive forces hold the molecules of a water droplet together, and the strong cohesive forces constitute surface tension.
When the attractive forces are between unlike molecules, they are said to be adhesive forces (see Adhesion). When the adhesive forces between water molecules and the walls of a glass tube are stronger than the cohesive forces lead to an upward turning meniscus at the walls of the vessel. This is Capillary action. Mercury is an example of a liquid that has strong cohesive forces, as becomes clear from the very convex meniscus in the tube of a classical air pressure meter.

Application

There are many, in medicine, science and daily life. Often they are based on surface tension.
Clinical tests Normal urine has a surface tension of about 0.066 N/m, but if bile is present, (a test for jaundice) it drops to about 0.055. In the Hay test, powdered sulfur is sprinkled on the urine surface. It will float on normal urine, but sink if the bile lowers the surface tension.
Surface tension disinfectants Disinfectants are usually solutions of low surface tension. This allows them to spread out on the cell walls of bacteria and disrupt them.
Walking on water Small insects can walk on water. Their weight is not enough to penetrate the surface.
Floating a needle If carefully placed on the surface, a small needle can be made to float on the surface of water even though it is several times as dense as water.
Soaps and detergents They help the cleaning of clothes by lowering the surface tension of the water so that it more readily soaks into pores and soiled areas.
Washing The surface tension of hot water is lower and therefore it is a better "wetting agent" to get water into pores and fissures rather than bridging them with surface tension.

More Info

There are various phenomena, which are based on cohesion.

Surface Tension
Water is a polar molecule due to the high electronegativity of the oxygen atom, which is an uncommon molecular configuration whereby the oxygen atom has two lone pairs of electrons. When two water molecules approach one other they form a hydrogen bond. The negatively charged oxygen atom of one water molecule forms a hydrogen bond with a positively charged hydrogen atom in another water molecule. This attractive force has several manifestations. Firstly, it causes water to be liquid at room temperature, while other lightweight molecules would be in a gaseous phase. Secondly, it (along with other inter molecular forces) is one of the principal factors responsible for the occurrence of surface tension in liquid water.
Water at 20 °C has a surface tension of 0.073 N/m compared to 0.022 N/m for ethyl alcohol and 0.47 N/m for mercury. The latter high value is the reason why in a capillary filled with mercury the meniscus is very convex. The surface tension of water decreases significantly with temperature.
The Surface tension arises from the polar nature of the water molecule.
Cohesion in crystals
In crystals (of molecular-, ionic-, valence- and metal-type) many types of forces play a role such as van der Waals forces and forces of chemical bonds. A van der Waals force is the attraction between two molecules with positively and negatively charged ends. This polarity may be a permanent property of a molecule (Keesom forces) or a temporarily property, which occurs universally in molecules, as the random movement of electrons within the molecules may result in a temporary concentration of electrons at one end (London forces).

Evaporation and perspiration

Principle

Evaporation

With evaporation, the opposite process of condensation, the atoms or molecules of the liquid gain sufficient energy to enter the gaseous state. It is exclusively a surface phenomenon that occurs at all temperatures and should not be confused with boiling. Even at cool temperatures, a liquid can still evaporate, but only a few particles would escape over a long period of time. Boiling occurs throughout a liquid (Fig. 1) and is characterized by the boiling point (e.g. for H_20 at 1.0013 bar 100 °C). For a liquid to boil its vapor pressure must equal the ambient pressure and bubbles are generated in the liquid.
For particles of a liquid to evaporate, they must be located near the surface, and moving in the proper direction, and have sufficient kinetic energy to overcome the Surface tension. Only a very small proportion of the molecules meet these criteria, so the rate of evaporation is limited. Since the average kinetic energy of the molecules rises with temperature and a larger fraction of molecules reaches the requested velocity, evaporation proceeds more quickly at higher temperature. As the faster-moving molecules escape, the remaining molecules have lower average kinetic energy, and the temperature of the liquid thus decreases. This phenomenon is also called *evaporative cooling*.
For simplicity, from now the liquid is considered water and the gas air with some water vapor.

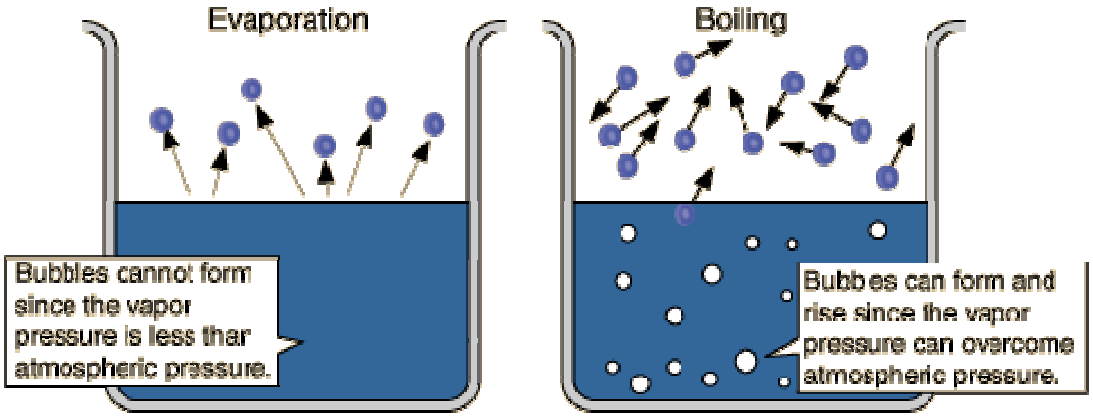


Fig. 1 Evaporation and boiling

Factors influencing rate of evaporation of water

- Evaporation is proportional with the difference in the partial water pressure (p_{H2O}) given by the temperature of the water and p_{H2O} in the air.
- Evaporation increases with the temperature of the water.
- The higher the speed of wind over the water surface, the faster the evaporation, which is due to constantly lowering the vapor pressure in the boundary layer. Wind speeds < 0.15 m/s are neglected.
- Other substances, especially polar ones (e.g. salts) decrease evaporation (Raoult's law). When a (monomolecular layer of) surfactant (a lipid-polar substance, see Surface tension) or a non-polar liquid with a smaller density (oil etc.) covers the water surface, evaporation is very strongly reduced.

The vapor pressure of saturation increases about exponential with the temperature, see Fig. 2.

Perspiration

Perspiration is the process of water evaporation from the human skin.
The process of evaporative cooling is the reason why evaporating sweat cools the human body. The cooling effect of flowing air along the (human) body is caused by convection but also by evaporation. Evaporation and so cooling happens also by expiration. At a body temperature of 37 °C, always, so irrespective the ambient pressure, the alveoli have a p_{H2O} of 47 mm Hg (= 62.6 mbar = 6260 Pa).
One square meter of a physiological salt solution evaporates about 6 dm^3 per day (see **More Info**). Assuming that the human skin is covered by a thin layer of physiological salt and that its surface is 1.9 m^2, then the perspiration is 1.9 m^2 x 6.0 $dm^3 \cdot m^{-2} \cdot day^{-1}$ = 11.4 dm^3/day. Actually, the loss is only 0.75 dm^3/day (see Body heat dissipation and related water loss). The reason is that the skin is generally not covered by a layer of physiological salt. Skin is not well permeable for water. Evaporation mainly occurs in the sweat glands, which cover only a small fraction of the skin. However, with extensive sweating (heavy endurance sport, sauna) the skin is covered with a thin layer of liquid and perspiration is some 0.5 dm^3/hour. Under these conditions, total sweat production is higher, but much drips off.

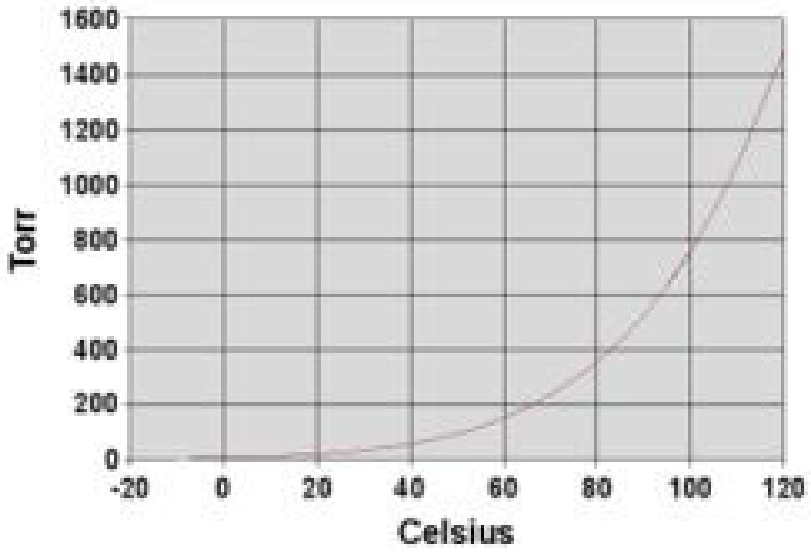


Fig. 2 Vapor pressure of equilibrium, p_{H2O}, as function of temperature. 760 Torr = 1atm = 1.013 bar. At 0 °C p_{H2O} is 6.2 mbar and at 100 °C 1001.3 mbar (= 1001.3 hPa).

More Info

Evaporation is a state change from liquid to gas, and as gas has less order than liquid matter, the entropy (see Thermodynamics: entropy) of the system is increased, which always requires energy input. This means that the enthalpy Thermodynamics: enthalpy)change for evaporation (ΔH_{liquid}) and the standard enthalpy change of vaporization or heat of evaporization (ΔH^0_{liquid}) is always positive, making it an endothermic process and subsequently, a cooling process.
If the evaporation takes place in a closed vessel, the escaping molecules accumulate as a vapor above the liquid. Many molecules return to the liquid and more molecules return as the density and pressure of the vapor increases. When the process of escape and return reaches equilibrium, the vapor is said to be "saturated," and no further change in either vapor pressure and density or liquid temperature will occur. For a system consisting of vapor and liquid of a pure substance, this equilibrium state is directly related to the vapor pressure of the substance, as given by the Clausius-Clapeyron relation:

$$\ln p^*_x = -\Delta\hat{H}_x/RT + B, \text{ where} \qquad (1)$$

p^*_x is the vapor pressure (bar)
$\Delta\hat{H}_x$ the heat of vaporization of liquid x (kJ/mole)
R is the gas constant (8.315 J/(mol·K)
T is the temperature (K)
B is a variable based on the substance and the system parameters.
$\Delta\hat{H}_{H2O}$ varies from 45.05 kJ/mol at 0 °C to 40.66 kJ/mol H_20 at 100 °C. This gives a strong dependency of p* in eq 1 from temperature.
Rewriting equation (1) for water, p^*_{H2O} is approximated by:

$$p^*_{H2O} = 1.27\cdot10^6\, e^{-5219/T} \text{ (bar)}. \qquad (2)$$

The rate of evaporation in an open system is related to the vapor pressure found in a closed system. If a liquid is heated with the vapor pressure reaching the ambient pressure, the liquid will boil.

The underlying physics of (1) is not too hard, but calculating the amount of liquid mass that is evaporated is another thing. The idea is that the fastest water molecules escape from a monomolecular layer of water, which accounts for the surface tension. The solution is an approximation, actually only for the state of equilibrium but for an open system, not in equilibrium, it works rather well.
The number of evaporating particles per second per unit area (ref. 1) is equal to:

$$N_e = (1/A)(v/d)e^{-W/kT}, \text{ where} \qquad (3)$$

- A is the cross sectional area of the particle (m^2, water molecule $0.057\cdot10^{-18}$ m^2),
- d the thickness of a monomolecular layer of the particles (m, water molecule ca. $0.27\cdot10^{-9}$ m),
- k Boltzmann's constant ($1.3805\cdot10^{-23}$ J/K).
- v the root mean square velocity of the liquid particle (m/s), a function of $T^{0.5}$: It can be calculated from Einsteins equation (ref. 2) of the Brownian motion (random movement of particles suspended in a fluid): $l^2 = (2kT/(3\pi\eta d)\cdot t$ where η is the dynamic viscosity coefficient, for water $\eta \approx 0.001$ Pa·s, at 37

^{o}C, *l* the free travel distance (estimated in water at 7 nm, nearly ten times less than the 66 nm in air) and t the time of the free path length. Rewriting gives $v = (2kT/(3\pi\eta d)l^{-1}$. This yields v =0.908 m/s).

- W is the energy needed to evaporate. $\Delta\hat{H}_{water}$ = 2428 kJ/kg at 37 ^{o}C, and a rough estimate of physiological salt is $\Delta\hat{H}_{physiological\ salt}$ = 2437 kJ/kg at 37 ^{o}C. Knowing Avogadro's number being $6.0225 \cdot 10^{23}$, $W_{water} = 7263 \cdot 10^{-23}$ J/molecule),
- The factor $e^{-W/kT}$ is the fraction of particles that have enough velocity to escape, so it presents also the probability (water at 310 K gives $42.6 \cdot 10^{-9}$, physiological salt $40.0 \cdot 10^{-9}$).

Finally, for water at 37 ^{o}C an evaporation of 7.48 mg·s^{-1}m^{-2} is found. This means lowering the surface with 6.4 mm per day or an evaporation of about 6.4 dm^3/m^2 per day (physiological salt 6.0 dm^3/m^2 per day). At 14 ^{o}C, experimentally (no air convections, humidity low) a surface lowering of 1.04 mm/day was found whereas 1.07 mm/day is predicted.

Literature

1. Feynman R.P., Leighton R.B. and Sands M. The Feynman lectures on Physics. Addison-Wesley, Reading, Mass., 1963.
2. Kronig R. (ed.). Leerboek der Natuurkunde, Scheltema & Holkema NV, Amsterdam 1966.

Mass spectrography

Principle

A mass spectrograph is used to separate electrically charged particles, for instance isotopes, according to their masses. In the mass spectrograph, beams of charged particles (ions) travel in a vacuum, pass through deflecting magnetic and electric fields (produced by carefully designed magnetic pole pieces and electrodes) and are detected by photographic plates. The degree of bending depends on the masses and electric charges of the ions. The detector measures exactly how far each ion has been deflected, and from this measurement, the ion's 'mass to charge ratio' can be worked out. From this information it is possible to determine with a high level of certainty what the chemical composition of the original sample was.

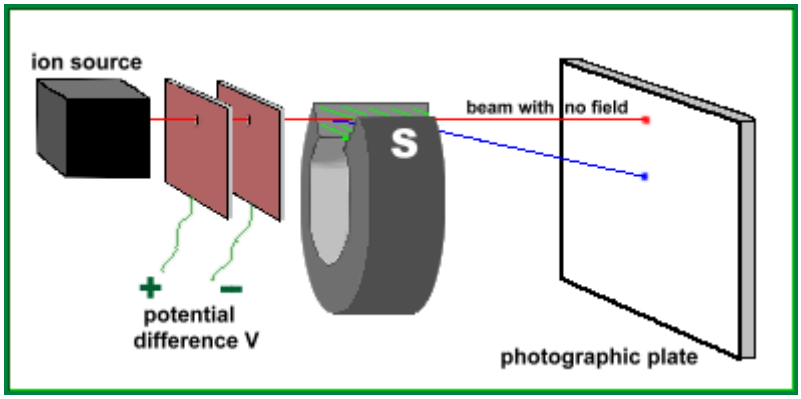


Fig. 1 A striped setup of a mass spectrograph for positive ions

The mass of an ion can be calculate directly:

$$m = \frac{B^2 q r^2}{2V}$$

where
B is the strength of the magnetic field provided by the magnet (in $N.Amp^{-1}.m^{-1}$),
r is the radius of curvature of the beam while in the field (m),
V is the voltage applied between the two plates (V),
q is the charge on an ion (1.60×10^{-19} coulomb if singly charged)
The equation shows that a smaller mass gives a smaller radius, due to the smaller momentum (mass·velocity) of the particle, provided that the ions enter the electric field with the same velocity.

Application

The mass spectrograph is very widely used in chemical analysis and in the detection of impurities in medicine, biology and (bio)chemistry. Applications include:

- identifying unknown compounds by their mass and/or fragments thereof;
- determining the isotopic composition of one or more elements in a compound;
- determining the structure of compounds by observing the fragmentation of the compound;
- quantifying the amount of a compound in a sample using carefully designed methods (mass spectrometry is not inherently quantitative);
- determining chemical or biological properties of compounds, proteins and peptides with a variety of other approaches, like age determination of graves (forensic medicine, medical archeology)

It is often combined with High performance liquid chromatography (HPLC, see Chromatography).

More info

Mass analyzers separate the ions according to their mass per charge. There are many types of mass analyzers. For these types see for instance Wikipedia. Usually they are categorized based on the principles of operation.

The common process comprises the following steps:
- The analyte is brought into some vacuum;
- it is ionized;
- the ions are accelerated in an electric field according to the Lorentz force acting on the charged particles.
Also in the two next steps the Lorentz force is the ruling physical law;
- by diaphragms and a velocity selector (a combination of electric and magnetic fields) the ions of interest are collected in a narrow beam;
- the beam enters a high vacuum chamber with two perpendicularly arranged magnetic fields. These bends the ions into semicircular paths ending at the photographic plate. The radius of this path depends upon the mass of the particles (all other factors, such as velocity and charge, being equal).
-, the position of the blackened spots on the plate makes possible a calculation of the isotope masses of the elements of the analyte. Nowadays an electron amplifier is used.
Depending on the selection process various versions of analyses exist.

A modification is a mass spectrometer, often used to measure the masses of isotopes. The quantity to measure is the mass-to-charge ratios of molecules and atoms. A mass spectrometer does not measure the kinetic energy of particles - all particles have the same known kinetic energy (or an integer multiple thereof, depending on the charge) - so it is disputable whether mass spectrometry strictly is a type of spectroscopy (see Spectroscopy).

Literature

http://www.worsleyschool.net/science/files/mass/spectrograph.html

Nuclear magnetic resonance (NMR)

Principle

Nuclear magnetic resonance (NMR) or zeugmatography is a non-invasive means of obtaining clinical images and of studying tissue metabolism in vivo.
Nuclei with an odd number of protons and/or neutrons possess a property called spin (rotating about its own axis). In quantum mechanics spin is represented by a magnetic spin quantum number. As atomic nuclei are charged, the spinning motion causes a magnetic moment in the direction of the spin axis (right-hand rule) according to the Lorentz force. This phenomenon is shown in Fig. 1. The strength of the magnetic moment is a property of the type of nucleus. Hydrogen nuclei (^{1}H), these are protons, which have the strongest magnetic moment, are in high abundance in biological material. Moreover, isotopes of H (deuterium and tritium) are not present in the body. Consequently hydrogen imaging is the most widely used MRI procedure.

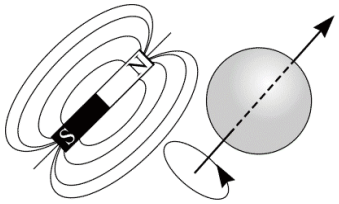
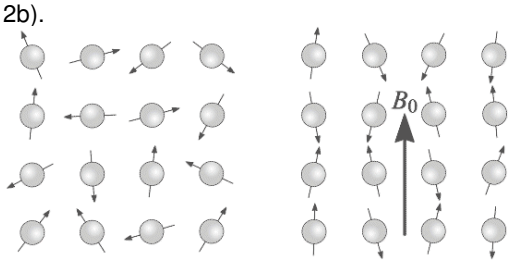


Fig. 1 A charged, spinning nucleus creates a magnetic moment which acts like a minuscule bar magnet (dipole).

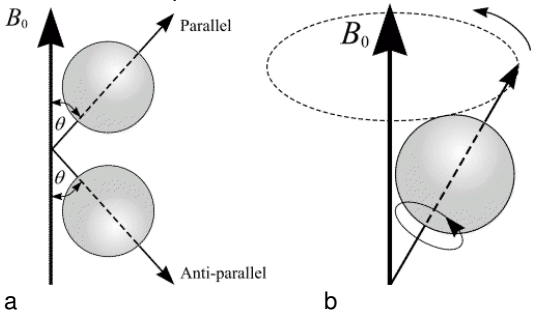
In the absence of an externally applied magnetic field, the magnetic moments of a collection of ^{1}H nuclei have random orientations (Fig. 2a). However, if an externally supplied magnetic field B_0 is imposed, the magnetic moments have a tendency to align parallel and anti-parallel with the external field (see Fig. 2b).



a b

Fig. 2 a. A collection of ^{1}H nuclei (spinning protons) in the absence of an externally applied magnetic field. b. Parallel and anti-parallel aligment of ^{1}H nuclei in an external magnetic field B_0.

There is a small mismatch, angle Θ, with the direction of B_0 (Fig. 3a). This is due to a precessing motion with a characteristic frequency as shown in Fig. 3b. This is analogous to the motion of a spinning top precessing in the earth's gravitational field. Atomic nuclei with the same magnetic spin quantum number as ^{1}H nuclei will exhibit the same effects - spins adopt one of two orientations in an externally applied magnetic field. The magnetic spin quantum number is denoted by m and has the values ±½. Elements whose nuclei have the same magnetic spin quantum number (-½ or +½) include ^{13}C, ^{19}F and ^{31}P. All three have one more neutron than the number of protons. ^{13}C is natural and stable, but its abundance is only 1.1%. ^{19}F and ^{31}P.are the only stable natural isotopes. Nuclei with higher magnetic spin quantum number will adopt more than two orientations.



a b

Fig. 3 (a) In the presence of an externally applied field, B_0, nuclei are constrained to adopt one of two orientations with respect to B_0. As the nuclei possess spin, these orientations are not exactly at 0 and 180^0 to B_0.. (b) A magnetic moment precessing around B_0. Its path describes the surface of a cone.

The Larmor equation expresses the relationship between the strength of a magnetic field, B_0, and the precessional frequency, F, of an individual spin.

$F = \gamma B_0$.

The proportionality constant γ is the gyromagnetic ratio of the nucleus. The precessional frequency, F, is also known as the Larmor frequency. For a hydrogen nucleus, the gyromagnetic ratio is 4257 Hz/Gauss. Thus at 1.5 Tesla (15,000 Gauss), F = 63.855 Megahertz.

Table 1

nucleus	γ (MHz/T)	abundance (%)
^{1}H	42.58	99.98
^{13}C	10.71	1.11
^{17}O	5.77	0.037
^{19}F	40.05	100
^{23}Na	11.26	100
^{31}P	17.23	100

Abundance ≡ 100[isotope]/[[sum of all isotopes]%

Radiofrequency field and MR signal

For a collection of ^{1}H nuclei, let the number of spins adopting the parallel and anti-parallel states be P1 and P2 respectively, with corresponding energy levels E1 and E2. E2 is greater than E1 causing P1 to be greater than P2. An obvious question is why do spins adopt the higher energy anti-parallel state? The answer is that spins of P2 may move to P1 if the exact amount of energy, $\Delta E = E2 - E1$ is supplied to the system. If the temperature of the system were absolute zero, all spins would adopt the parallel orientation P1. Thermal energy will cause P2 to be populated. At room temperature in a 1.5 Tesla magnetic field, there will typically be a population ratio P2:P1 equal to 100,000:100,006. This small difference gives a net polarization and consequently the opportunity to do measurements at al.
At any given instant, the magnetic moments of a collection of ^{1}H nuclei can be represented as vectors, as shown in Fig. 4. Every vector can be described by its components perpendicular to and parallel to B_0. For a large enough number of spins distributed on the surface of the cone, individual components perpendicular to B_0 cancel, leaving only components in the direction parallel to B_0. As most spins adopt the parallel rather than the antiparallel state, the net magnetisation M is in the direction of the B_0 field.

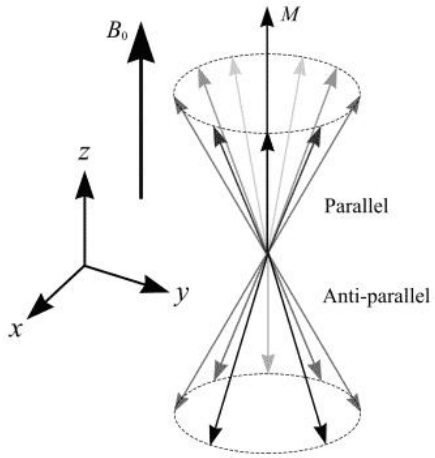


Fig. 4 A collection of spins at any given instant in an external magnetic field, B_0. A small net magnetisation, M, is detectable in the direction of B_0.

Suppose the direction of B_0 is aligned with the z-axis of Euclidean 3-space. The plane perpendicular to B_0 contains the x and y-axes. In order to detect a signal from ^{1}H nuclei, radio frequency (RF) energy must be applied. RF energy at the Larmor frequency causes nuclear spins to swap between parallel and anti-parallel states. This has an oscillatory effect on the component of the magnetisation M parallel to the z-axis. RF energy, like all electromagnetic radiation, has electric and magnetic field components. Suppose the magnetic field component is represented by B_1 and lies in the x-y plane. The x-y components of M will be made coherent by the B_1 field giving a net x-y component to M and hence

effectively cause M to tilt from the z direction into the x-y plane. This phenomenon is described further in Fig. 5.
The angle through which M has rotated away from the z-axis is known as the flip angle. The strength and duration of B_1 determine the amount of energy available to achieve spin transitions between parallel and anti-parallel states. Thus, the flip angle is proportional to the strength and duration of B_1. After pulses of 90^0 and 270^0. M has no z component and the population ratio P2:P1 is exactly one. A pulse of 180^0 rotates M into a position directly opposite to B_0, with greater numbers of spins adopting anti-parallel (rather than parallel) states. If the B_1 field is applied indefinitely, M tilts away from the z-axis, through the x-y plane towards the negative z direction, and finally back towards the x-y plane and z-axis (where the process begins again).

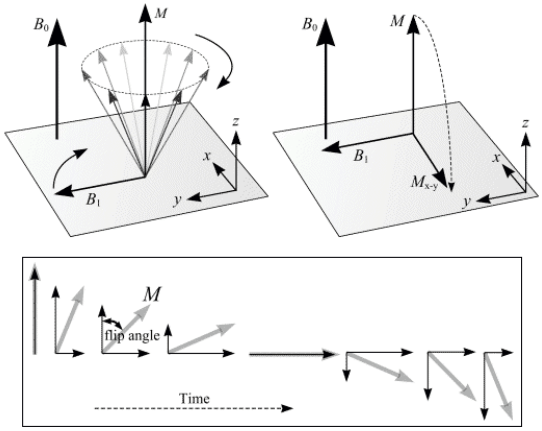
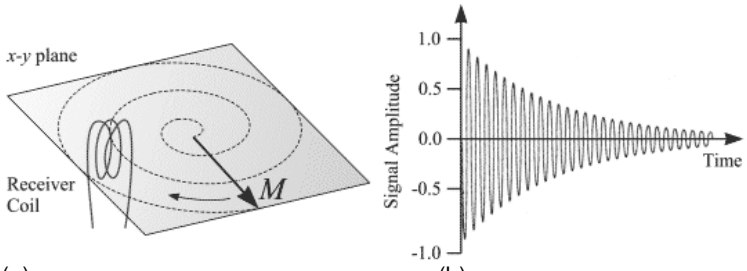


Fig. 5 (top) The effect of RF radiation on the net magnetisation M is to produce a second magnetic field Mx-y. M is tilted from its original longitudinal z-axis orientation, along the direction of the external magnetic field B_0, into the transverse x-y plane. (bottom) An illustration of flip angle, which is the angle through which M has rotated away from the z-axis.

Fig. 6a shows the situation after an RF pulse is applied that causes the net magnetisation vector M to flip by 90^0. M lies in the x-y plane and begins to precess about the B_0 axis. *M will induce an electromotive force in a receiver coil according to Faraday's law of magnetic induction. This is the principle of NMR signal detection.* It is from this received RF signal that an MR image can be constructed. Fig. 6b shows a graph of the voltage or signal induced in a receiver coil verses time. Such a graph, or waveform, is termed a free induction decay (FID). The magnitude of the generated signal depends on the number of nuclei contributing to produce the transverse magnetisation and on the relaxation times (see next section).



(a) (b)

Fig. 6 (a) After a 90^0 RF pulse, M lies in the x-y plane and rotates about the z-axis. The component of M in the x-y plane decays over time. An alternating current, shown in (b), is induced in the receiver coil.

Relaxation Processes
The return of M to its equilibrium state (the direction of the z-axis) is known as relaxation. There are three factors that influence the decay of M: magnetic field inhomogeneity, longitudinal T1 relaxation and transverse T2 relaxation. T1 relaxation (also known as spin-lattice relaxation) is the realignment of spins (and so of M) with the external magnetic field B_0 (z-axis). T2 relaxation (also known as T2 decay, transverse relaxation or spin-spin relaxation) is the decrease in the x-y component of magnetisation. For T1 relaxation and T2 relaxation see further MRI: T1 and T2 relaxation.

Magnet inhomogeneity
It is virtually impossible to construct an NMR magnet with perfectly uniform magnetic field strength, B_0. Much additional hardware is supplied with NMR machines to assist in normalising the B_0 field. However, it is inevitable that an NMR sample will experience different B_0's across its body so that nuclei comprising the sample (that exhibit spin) will have different precessional frequencies (according to the Larmor equation). Immediately following a 90^0 pulse, a sample will have Mx-y coherent. However, as time goes on, phase differences at various points across the sample will occur due to nuclei precessing

at different frequencies. These phase differences will increase with time and the vector addition of these phases will reduce Mx-y with time.

Adopted from Mike Puddephat dec 2001 http://www.easymeasure.co.uk/principlesmri.aspx

Paramagnetism, diamagnetism and magnetophoresis

Paramagnetism

Paramagnetism is the tendency of atomic magnetic dipoles to align within an external magnetic field. Paramagnetic materials attract and repel like normal magnets when subjected to a magnetic field. This effect occurs due to quantum-mechanical spin as well as electron orbital angular moment (the vector cross product of radius and mass times velocity). This alignment of the atomic dipoles with the magnetic field tends to strengthen it and is described by a relative magnetic permeability μ slightly greater than unity (or, equivalently, a small magnetization or positive magnetic susceptibility X_v).

Application

In medical Magnetic resonance imaging (see MRI: general)the paramagnetism of contrast substances as barium sulfate is applied to enhance contrast.
Hb is diamagnetic when oxygenated but paramagnetic when deoxygenated. Therefore, an important application is the paramagnetism of deoxy-Hb. The magnetic resonance (MR) signal of blood is therefore variable depending on the level of oxygenation. These differential signals can be detected using an appropriate MR pulse sequence as Blood Oxygenation Level Dependent (BOLD) contrast (see MRI; functional MRI)).
The existence of four unpaired electrons in deoxy-Hb can give rise to a process called magnetophoresis (see below).

More info

Paramagnetism requires that the atoms individually have permanent dipole moments even without an applied field, which typically implies partially filled electron shells. In pure paramagnetism, these atomic dipoles do not interact with one another and are randomly oriented in the absence of an external field, resulting in zero net magnetic moment. If they *do* interact, they can spontaneously align or anti-align, resulting in ferromagnetism (permanent magnets) or antiferromagnetism respectively. (With the latter, occurring below a certain temperature, the magnetic moments of the particles, align in a regular pattern with neighboring spins (on different sublattices) in opposite directions.. Only in a external field there is magnetization since the opposite sublattice magnetizations do not cancel. Above the aforementioned temperature, the material is typically paramagnetic.)
In atoms with no permanent dipole moment, e.g. for filled electron shells, a weak dipole moment can be *induced* in a direction *anti*-parallel to an applied field, an effect called diamagnetism. Paramagnetic materials also exhibit diamagnetism, but the latter effect is typically orders of magnitude weaker. Paramagnetic materials in magnetic fields will act like magnets but when the field is removed, the magnetic alignment is quickly disrupted. In general, paramagnetic effects are small. They are, as ferrimagnetism, temperature dependent. (Ferrimagnetism, occurring below the Curie temperature, is similar as antiferromagnetism, but the sublatices comprise either Fe^{2+} of Fe^{3+}. This results in a small net magnetization.)

In order to be paramagnetic, i.e. to be attracted, the electrons align themselves in one direction in a magnetic field. The following example may clarify this. Carbon has an electron configuration with the six orbitals: 1s^2, 2s^2, 2p^2. The way that this would look is 1s^2: in this orbital we have one arrow (the momentum) going up and one arrow going down. 2s^2: in this orbital there is one arrow going up and one arrow going down. 2p^2: this has three orbitals with one arrow going up in the first orbital, one arrow going up in the second orbital and nothing (no electron) in the third orbital. Since in 2p^2 both electrons have the same spin, a permanent dipole moment results and there is a strong paramagnetic behavior in an magnetic field in which all the dipoles are lined up. In a similar way it can be understood that the alkaline-earth metals, and e.g. Pt are paramagnetic. For ions, basically the same concept holds to explain paramagnetism.

Although most components of the human body are weakly diamagnetic, many organisms have been shown to contain small amounts of strongly magnetic materials, usually magnetite (Fe_3O_4; ferrimagnetism). The most extreme case is that of magnetotactic bacteria. With their magnetite particles they orient themselves in the earth's magnetic field. Magnetite crystals have also been found in pigeons, honeybees, many mammals, and even in the human brain, but in proportionately much smaller amounts than in the bacteria.
It seems very unlikely that there is enough magnetite within the human body to provide a possible mechanism to explain the highly speculative effects of magnetic therapy.

Diamagnetism

Diamagnetism is the exact opposite as paramagnetism. A diamagnetic substance is repelled, very weakly, by a magnet and is only exhibited in the presence of an externally applied magnetic field as the result of changes in the orbital motion of electrons. Consequently, diamagnetic substances have a relative magnetic permeability smaller than unity (or negative magnetic susceptibility).
The magnetic field creates a magnetic force on a moving electron, being F=qv x **B**. (see Lorentz force). This force changes the centripetal force on the electron, causing it to either speed up or slow down in its orbital motion. This changed electron speed modifies the magnetic moment of the orbital in a direction against the external field.
It should be noted that dia- and paramagnetism can change when the aggregation state changes and also when an element becomes an ion.

Application

Medical MRI generally utilizes the diamagnetism of water, which relies on the spin numbers of excited hydrogen nuclei in water of the tissue under study. Some nuclei with non-zero spin numbers are (to some extent) parallel or anti-parallel aligned in the strong magnetic field of the MR apparatus (some 2-7 Tesla). The vast quantity of aligned nuclei in a small volume sum to produce a small detectable change in field.
Similar as paramagnetism of deoxy-Hb, the diamagnetism of oxy-Hb is applied in functional MRI (see MRI: functional MRI). This technique is often applied to establish the change (relative) of oxygen consumption in particular brain structures when e.g. performing some intellectual or sensory-motor task.
A particularly fascinating phenomenon involving diamagnets is that they may be levitated in stable equilibrium in a very strong magnetic field, with no power consumption. The magnetic moment is proportional to the applied field **B**. This means that the magnetic energy of diamagnets is proportional to $\mathbf{B}^2$, the intensity of the magnetic field. In very strong fields, a thin slice of pyrolytic graphite (strongly diamagnetically), a live frog, and water amongst other things have been levitated with success.

More Info

Consider two electron orbitals; one rotating clockwise and the other counterclockwise, say in the plane of this page. An external magnetic field into the page will make the centripetal force on an electron rotating clockwise increase, which increases its moment out of the page. That field would make the centripetal force on an electron rotating counterclockwise decrease, decreasing its moment into the page. Both changes oppose a magnetic field into the page. The induced net magnetic moment is very small in most everyday materials. An example is He. As helium has two orbitals, one with spin-arrow facing up and one with a spin-arrow facing down. Together with the angular moments of both orbitals, hardly any magnetic momentum remains in the atom. Consequently, as other noble gases, Ne is diamagnetic.
All materials show a diamagnetic response in an applied magnetic field; however for materials which show some other form of magnetism, the diamagnetism is completely overwhelmed. Substances which only, or mostly, display diamagnetic behavior are termed diamagnetic materials, or diamagnets.
Materials that are said to be diamagnetic are those which are usually considered by non-physicists as "non magnetic", and include water and DNA, most organic compounds such as oil and plastic, and many metals. Superconductors may be considered to be perfect diamagnets.

Magnetophoresis

The existence of unpaired electrons in the four heme groups of deoxy-Hb gives this molecule paramagnetic properties as contrasted to the diamagnetic character of oxy-Hb. Based on the measured magnetic moments of Hb and its compounds, and on the relatively high Hb concentration of human erythrocytes, differential migration of these cells is basically possible if exposed to a high magnetic field. In this way it is possible to measure the migration velocity of deoxy-Hb-containing erythrocytes, exposed to a mean magnetic field of 1.40 T (Tesla) and a mean gradient of 0.131 T/mm, a process called magnetophoresis (e.g. *Biophys J.* 2003;84:2638-45).

Particle-X-ray interaction

Principle

According to all kind of physical principles and laws atomic and molecular particles can interact with radiation. X-ray photons impinging upon matter generally lose energy. Fig. 1 summarizes the most important interactions.

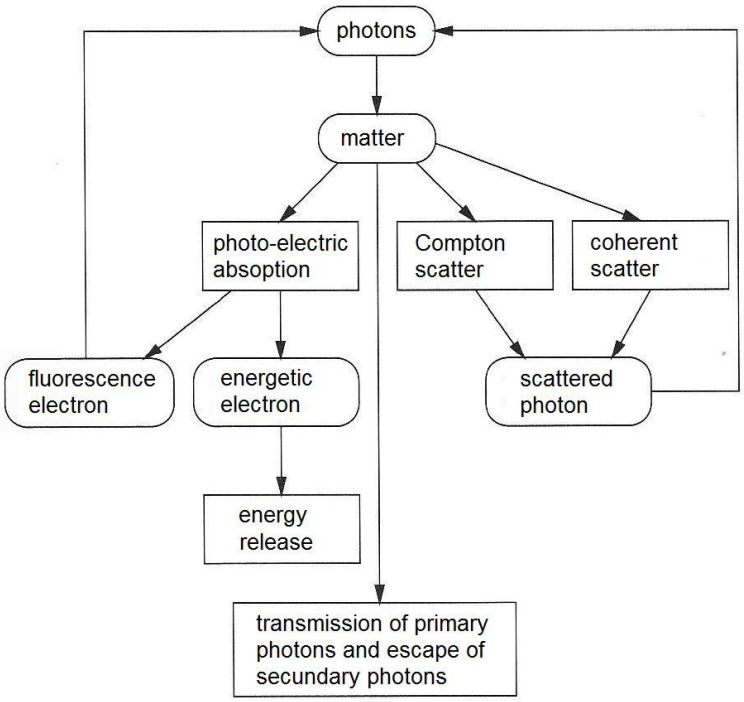


Fig. 1 Diagram of photon-matter interactions.

The three most important interaction processes, which are mostly accompanied with energy loss, are coherent scatter, photo-electric absorption and the Compton effect.

Coherent (classical) scatter
When the energy of an X-ray is small (low-energy photons) compared to the binding energy of the electrons, then there is scatter from the whole atom without release of energy. Hence, the scatter is elastic.

Photo-electric absorption
In this case the photon, in fact a mid-energy photon, looses all its energy. The released energy is used to compensate the binding energy and the remaining energy is used by the escaped electrons to gain kinetic energy. A free electron can occupy the free place in the electron orbit at a lower energy level by emitting a fluorescent photon.
The X-ray machine, X-ray microscopy and Fluorescopy make use of this process.

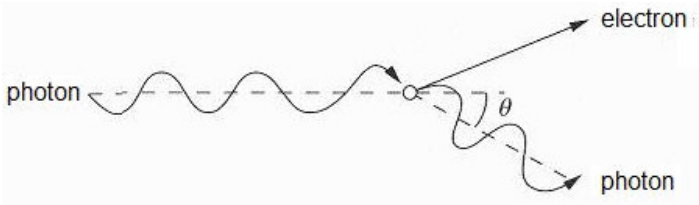


Fig. 2 Compton effect.

Compton effect.
The Compton effect is incoherent scatter, hence emittance of photons with various wavelengths. The primary high photon energy (X-ray or gamma ray: 50-1000 KeV) is a multiple of the electron binding

energy, such that the binding energy can be neglected. Hence, the interaction of the photon can be considered as being with a free electron. The energy of the scattered photon is always diminished, hence the scatter is inelastic and such that the wavelength difference is:

$\Delta\lambda = \lambda_0 (1-\cos\theta)$.

where θ is the angle of scatter, see Fig. 2. λ_0 is the Compton wavelength, being 2.43 pm. With backscatter, the difference is 2x2.43 pm and with scatter exactly perpendicular it is 2.43 pm (see Fig. 3). A high-energy photon, so with a short wavelength, transfers much more energy than a low-energy photon.

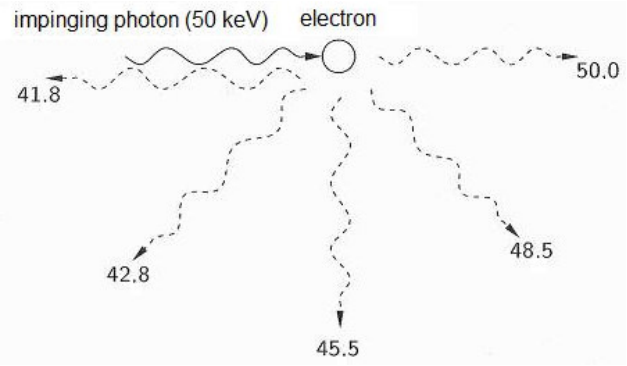


Fig. 3 Compton effect with various angles of scatter.

The Compton effect is basic in radiotherapy and gamma spectroscopy.

Pair production
Very high energetic photons (>1.022 MeV) impinging upon a nucleus can produce the emittance of an electron-positron pair, called pair production. This process has no applications in (bio)medicine. The electron-positron generation with PET is based on radioactive decay of a radionuclide.

Secondary effects
Fast electrons (photo-electric or Compton electrons) lose their energy mainly by ionization and excitation. This respects generally electrons in outer orbits with low binding energy. Per ionization typically 34 eV is transferred, and consequently one fast electron can produce a whole trace of positive and negative ions, some 1500 pairs when starting with 50 keV. The remaining released energy is used for excitation of atoms. When the ions return to their neutral state by absorbing or releasing an electron, a fluorescent photon can be emitted. Due to the massive generation of ions, X-radiation is also called indirect ionizating radiation.
The secondary photons or fluorescent photons can show in their turn interaction.

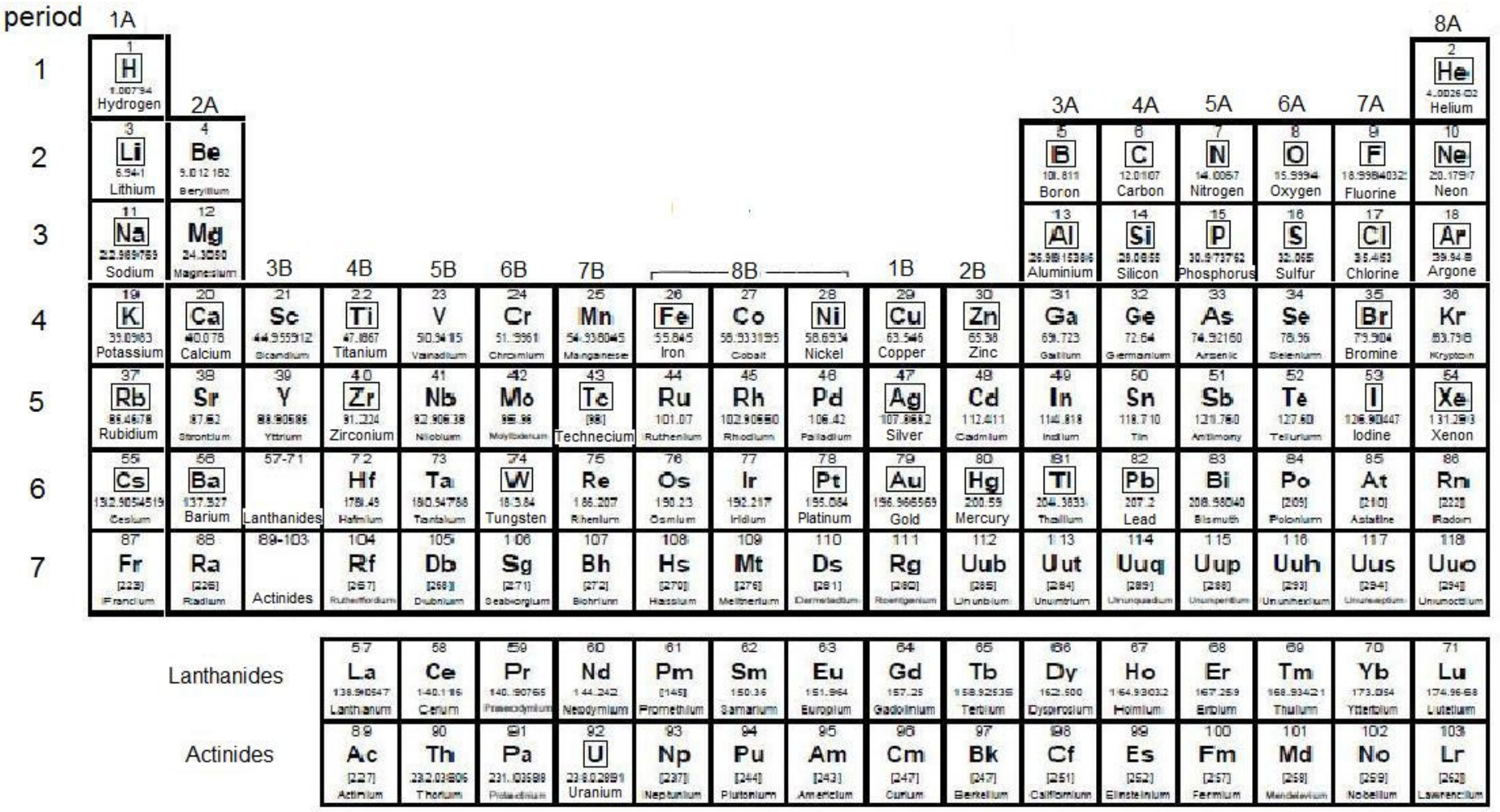
Summarizing, when X-radiation hits matter, several things happen:
- a part of the photons are undisturbed transmitted;
- a great number of ion-pairs arises;
- in all directions photons, especially from mater with a high atomic number, are emitted and moreover fluorescent photons, which are typically for the types of atoms in the matter.

Periodic system of elements

Periodic system of elements or with its full name the Periodic table of the chemical elements has 7 periods and 18 groups, arranged from 1A, 2A, 1B to 5B, 8B (3 groups) , 6B , 7B and 3A to 8A.

The periods give the electron shells, which are present, and the groups the filling of the shells.
On the left are the metals and on the right the nonmetals. These groups are further subdivided. Group 1A contains the alkali metals, group 2A the alkaline earth metals, group 7A the (natural) halogens and group 8A the (natural) noble gases.

In each box, an element is presented, except in the box of the Lanthanides (rare earth metals, nowadays also called Lanthanoids) and the box of the Actinides (also a kind of rare earth metals, which are all radioactive, nowadays also called Actinoides).From top to bottom, the box gives the atomic number, the symbol, the atomic weight and the international unless an English name is known. When the symbol is written within a small box, then the respecting element is mentioned in this compendium.


period
1A
2A
3B
4B
5B
6B
7B
8B
1B
2B
3A
4A
5A
6A
7A
8A
1 H Hydrogen
2 He Helium
3 Li Lithium
4 Be Beryllium
5 B Boron
6 C Carbon
7 N Nitrogen
8 O Oxygen
9 F Fluorine
10 Ne Neon
11 Na Sodium
12 Mg Magnesium
13 Al Aluminium
14 Si Silicon
15 P Phosphorus
16 S Sulfur
17 Cl Chlorine
18 Ar Argon
19 K Potassium
20 Ca Calcium
21 Sc Scandium
22 Ti Titanium
23 V Vanadium
24 Cr Chromium
25 Mn Manganese
26 Fe Iron
27 Co Cobalt
28 Ni Nickel
29 Cu Copper
30 Zn Zinc
31 Ga Gallium
32 Ge Germanium
33 As Arsenic
34 Se Selenium
35 Br Bromine
36 Kr Krypton
37 Rb Rubidium
38 Sr Strontium
39 Y Yttrium
40 Zr Zirconium
41 Nb Niobium
42 Mo Molybdenum
43 Tc Technecium
44 Ru Ruthenium
45 Rh Rhodium
46 Pd Palladium
47 Ag Silver
48 Cd Cadmium
49 In Indium
50 Sn Tin
51 Sb Antimony
52 Te Tellurium
53 I Iodine
54 Xe Xenon
55 Cs Cesium
56 Ba Barium
57-71 Lanthanides
72 Hf Hafnium
73 Ta Tantalum
74 W Tungsten
75 Re Rhenium
76 Os Osmium
77 Ir Iridium
78 Pt Platinum
79 Au Gold
80 Hg Mercury
81 Tl Thallium
82 Pb Lead
83 Bi Bismuth
84 Po Polonium
85 At Astatine
86 Rn Radon
87 Fr Francium
88 Ra Radium
89-103 Actinides
104 Rf Rutherfordium
105 Db Dubnium
106 Sg Seaborgium
107 Bh Bohrium
108 Hs Hassium
109 Mt Meitnerium
110 Ds Darmstadtium
111 Rg Roentgenium
112 Uub Ununbium
113 Uut Ununtrium
114 Uuq Ununquadium
115 Uup Ununpentium
116 Uuh Ununhexium
117 Uus Ununseptium
118 Uuo Ununoctium
Lanthanides
57 La Lanthanum
58 Ce Cerium
59 Pr Praseodymium
60 Nd Neodymium
61 Pm Promethium
62 Sm Samarium
63 Eu Europium
64 Gd Gadolinium
65 Tb Terbium
66 Dy Dysprosium
67 Ho Holmium
68 Er Erbium
69 Tm Thulium
70 Yb Ytterbium
71 Lu Lutetium
Actinides
89 Ac Actinium
90 Th Thorium
91 Pa Protactinium
92 U Uranium
93 Np Neptunium
94 Pu Plutonium
95 Am Americium
96 Cm Curium
97 Bk Berkelium
98 Cf Californium
99 Es Einsteinium
100 Fm Fermium
101 Md Mendelevium
102 No Nobelium
103 Lr Lawrencium

Surface tension

Principle

In physics, surface tension is an effect within the surface layer of a liquid (gas-liquid interface) that causes the layer to behave as an elastic sheet. It is the effect that allows insects to walk on water, and that causes Capillary force.

Fig. 1 Surface tension prevents this flower from submerging.

Flat surfaces
Surface tension is caused by the attraction between the liquid molecules, due to various intermolecular forces. In the bulk of the liquid, each molecule is pulled equally in all directions by neighboring liquid molecules, resulting in a net force of zero. At the surface of the liquid, the molecules are pulled inwards by other molecules deeper into the liquid, but there are no liquid molecules on the outside to balance these forces. So, the surface molecules are subject to an inward force of molecular attraction which is balanced by the resistance of the liquid to compression. There may also be a small outward attraction caused by air molecules, but as air is much less dense than the liquid, this force is negligible.
Surface tension is measured in Newton/meter ($N \cdot m^{-1}$) and represented by the symbol γ, σ, or *T* and is defined as the force along a line of unit length parallel to the surface or work done per unit area. In other words, the surface tension ($N \cdot m^{-1}$) is equivalent to energy per square meter ($J \cdot m^{-2}$). This means that surface tension can also be considered as free surface energy. If a surface with surface tension γ is expanded by a unit area, then the increase in the surface's stored energy is also equal to γ.

Drops and gas bubbles
With a completely flat water surface there is no force, which tries to pull a liquid molecule outside the liquid, provided that the gas above the surface is saturated with the molecules of the liquid. However, with a gas bubble in a liquid, the interface attempts to have an as small as possible surface, since doubling the bubble surface doubles the increase in the surface's stored energy. Consequently the surface tension is directed toward the centre of the bubble. Without a counteracting force the bubble will vanish. This can only be prevented when the bubble pressure P_{bubble} balances the surface pressure P_{ST} due to the surface tension plus the ambient pressure P_{amb} in the liquid being the sum of hydrostatic and atmospheric pressure:

$$P_{bubble} = P_{ST} + P_{amb}, \text{ and} \qquad (1)$$

$$P_{ST} = 2\gamma/r$$

where r the radius. P_{bubble} is the sum of the partial pressures of the composing types of gas in the bubble ($\Sigma P_{bubble\ gases}$).

The above equation holds for a pure liquid like water. However, in biological systems there are generally molecules with a hydrophobic and hydrophilic part, like fatty acids. These molecules form a monomolecular membrane at the interface with the hydrophobic end inside the gas and the hydrophilic end in the water.

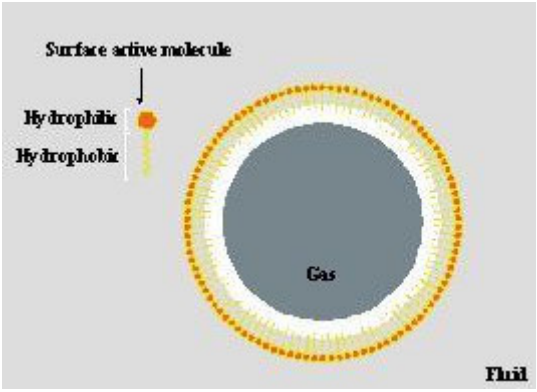


Fig. 2 Skin of surface active molecules surrounding a gas bubble.

Just as the water molecules at the interface “pull” towards each other in surface tension, the surface active molecules, the surfactant, “push” against each other. So, this force is directed outward. This counteracts the effect of surface tension, and therefore eliminates the loss of gas in the bubble by outward diffusion. This membrane reduces the motion of gas molecules from the bubble to the liquid. $P_{surfactant}$ is:

$$P_{surfactant} = 2\gamma_c/r$$

γ_c accounts for the springy “push back” effect of the surfactants. Equation (1) becomes:

$$\Sigma P_{bubble\ gases} + 2\gamma_c/r = P_{amb} + 2\gamma/r. \qquad (2)$$

Due to the surface tension liquid drops, e.g. the small water droplets in fog or floating oil drops in liquid are spherical.

Application

The surfactant is of crucial importance to maintain the shape of the alveoli. Without surfactant they will shrink immediately. The surfactant decreases the effect of γ by a factor of ten. Lung surfactant and surfactant disorders are subject of extensive research.
Venous gas embolism (VGE) and decompression sickness (DCS) of divers is due to the occurrence of pathological nitrogen bubbles in tissues and blood. Generally, VGE occurs during the ascent, but especially after the ascent and DCS occurs mostly after the ascent. The occurrence of these bubbles relies on surface tension of the bubble skin and pressure differences.

More info

A soap bubble comprises a double (or more) molecular membrane of fatty acids (the ‘skin’) covered by a thin water film at both sides. The inside directed force yielded by the surface tension of both water films is counteracted by the outward directed force provided by the skin. The net effect of both surface tension evokes a pressure within the bubble. Since there is an outer and an inner interface the constant 2 in eq. (2) should be doubled to 4.
Eq. (2) suggests equilibrium, but this cannot be reached. The outer water film will vaporize due to the low p_{H2O} of the outside air. This makes the bubble slightly growing. Also N_2 and O_2 enter the bubble due to partial pressure differences whereas H_2O cannot leave easily the bubble since the skin is a barrier for this polar molecule. This also makes the bubble growing. Also soap molecules and water collect at the bottom due to gravity. These processes destabilize the bubble leading to rupture from top to bottom.

A quantity related to surface tension is the energy of Cohesion, which is the energy released when two bodies of the same liquid become joined after removal of the boundary of separation. Since this process involves the removal of surface from each of the two bodies of liquid, the energy of cohesion is equal to twice the surface energy. A similar concept, the energy of Adhesion applies to two bodies of different liquids.

Scanning tunneling microscopy

Principle

With scanning tunneling microscopy (STM) one can visualize surfaces at the atomic level. STM probes the density of states of a material using tunneling current. The density of states (DOS), a concept of quantum mechanics, of a system describes the number of states at each energy level that is available to be occupied. A high DOS means that there are many states available for occupation. The STM is based on the concept of quantum tunneling, an effect of quantum mechanics.
When a conducting tip is brought very near to a metallic or semiconducting surface, a specific voltage difference between the two can allow electrons to tunnel through the vacuum between them. For low voltages, this tunneling current is locally dependent on DOS. Variations in current as the probe passes over the surface are translated into an image. Electrons behave as waves of energy, and in the presence of a potential, assuming the 1D case, the energy levels of the electrons are given by solutions of the Schrödinger's equation.

For STM, good resolution is considered to be 0.1 nm lateral resolution and 0.01 nm depth resolution. The STM can be used not only in ultra high vacuum but also in air and various other liquid or gas, and at temperatures ranging from near 0 Kelvin to a few hundred degrees Celsius.

Procedure
First the tip is brought into close proximity, some 0.4 - 0.7 nm, of the sample by some coarse sample-to-tip control. This distance d is the equilibrium position between attractive ($0.3 < d < 1$ nm) and repulsive ($d < 0.3$ nm) interactions. Once tunneling is established, piezoelectric (see Piezoelectricity) transducers are implemented to move the tip in three directions. As the tip is scanned across the sample in the x-y plane, the density of states and therefore the tunnel current changes. This change in current with respect to position can be measured itself, or the height, z, of the tip corresponding to a constant current can be measured. These two modes are called constant height mode and constant current mode, respectively.
In constant current mode, feedback electronics adjust the height by a voltage to the piezoelectric height control mechanism. This leads to a height variation and thus the image comes from the tip topography across the sample and gives a constant charge density surface; this means contrast on the image is due to variations in charge density.
In constant height, the voltage and height are both held constant while the current changes to keep the voltage from changing; this leads to an image made of current changes over the surface, which can be related to charge density. The benefit to using a constant height mode is that it is faster, as the piezoelectric movements require more time to register the change in constant current mode than the voltage response in constant height mode.

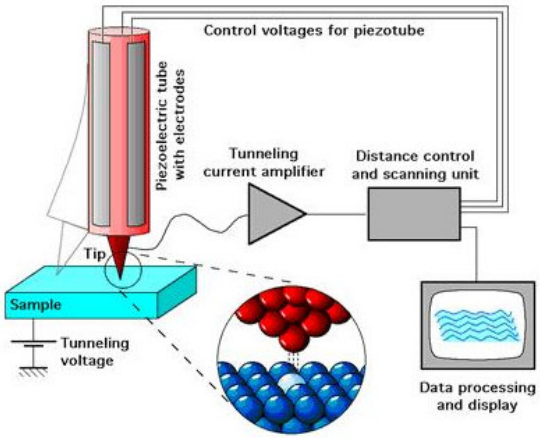


Fig. 1 Schematic view of an STM

Application

Material science at nanometer resolution and indirectly the applications of Atomic force microscopy.

More Info

The resolution of an image is limited by the radius of curvature of the scanning tip of the STM. Additionally, image artifacts can occur if the tip has two tips at the end rather than a single atom; this

leads to “double-tip imaging,” a situation in which both tips contribute to the tunneling. Therefore it has been essential to develop processes for consistently obtaining sharp, usable tips. Recently, carbon nanotubes have been used in this instance.

Due to the extreme sensitivity of tunnel current to height, proper vibration isolation is imperative for obtaining usable results.
Maintaining the tip position with respect to the sample, scanning the sample in raster fashion and acquiring the data is computer controlled. The computer is also used for enhancing the image with the help of image processing as well as performing quantitative morphological measurements.

In addition to scanning across the sample, information on the electronic structure of the sample can be obtained by sweeping voltage and measuring current at a specific location. This type of measurement is called scanning tunneling spectroscopy.

Other STM Related Studies
Many other microscopy techniques have been developed based upon STM. These include Photon scanning tunneling microscopy, which uses an optical tip to tunnel photons; Scanning tunneling potentiometry, which measures electric potential across a surface; and Spin polarized scanning tunneling microscopy, which uses a ferromagnetic tip to tunnel spin-polarized electrons (see Electron Spin Resonance) into a magnetic sample.
Other STM methods involve manipulating the tip in order to change the topography of the sample. This is attractive for several reasons. Firstly the STM has an atomically precise positioning system which allows very accurate atomic scale manipulation. Furthermore, after the surface is modified by the tip, it is a simple matter to then image with the same tip, without changing the instrument.

Waves and wave phenomena

Doppler principle

Principle

The Doppler effect, discovered by Christian Doppler in 1842, is the apparent *change in frequency* (and wavelength) of a wave that is perceived by an observer moving relative to the source of the waves. The waves can be electromagnetic (visible light, X-ray, radio-waves, gamma radiation, etc.), sound waves, gravity waves, surface waves at a liquid (water) etc. For sound waves, the velocity of the observer and the source are reckoned relative to the transmitting medium.

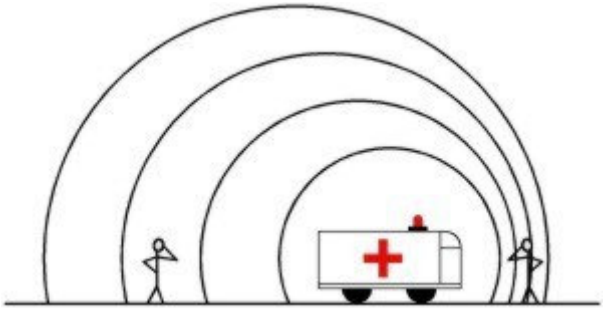


Fig. 1 Sound waves emanating from an ambulance moving to the right.

The total Doppler effect results from either motion of the source or motion of the observer. Each of these effects is analyzed separately. For waves which do not require a medium (such as light) only the relative difference in velocity between the observer and the source needs to be considered. Fig. 1 visualizes how the sound of an ambulance are compressed (perceived frequency increase) in front of the ambulance and 'diluted' (frequency decrease) behind it.
Remind that the frequency of the sounds that the source *emits* does not actually change.

Conceptual explanation
The following analogy helps to understand the Doppler principle. Someone throws one ball every second in your direction. Assume that balls travel with constant velocity. If the thrower is stationary, you will receive one ball every second. However, if he is moving towards you, you will receive balls more frequently than that because there will be less spacing between the balls. The converse is true if the person is moving away from you. So it is actually the *wavelength*, which is affected; as a consequence, the perceived frequency is also affected.
If the moving source is emitting waves through a medium with an actual frequency f_0, then an observer stationary relative to the medium detects waves with a frequency f given by:

$$f = f_0\left(\frac{c}{c + v_s}\right) \quad (1)$$

where c is the speed of the waves in the medium and v_s is the speed of the *s*ource with respect to the medium (negative if moving towards the observer, positive if moving away), with the observer on the pathway of the source (radial to the observer). With v_s is<<c and Δf, the frequency shift, being $f - f_0$, and applying (1) Δf is:

$$\Delta f = f_0 \frac{v_s}{c} \quad (2)$$

A similar analysis for a moving observer and a stationary source yields the observed frequency (the observer's velocity being represented as v_o):

$$f = f_0\left(1 + \frac{v_0}{c}\right) \quad (3)$$

A stationary observer perceives the moving ambulance siren at different pitches depending on its relative direction. The siren will start out higher than its stationary pitch, slide down as it passes, and continue lower than its stationary pitch as it recedes from the observer. The reason the siren slides is because at the moment of passing there is some distance between the ambulance and you. If the siren approached you directly, the pitch would remain constant (as v_s is only the radial component) until the vehicle hit you, and then immediately jump to a new lower pitch. The difference between the higher pitch and rest pitch (when $v_0 = 0$) would be the same as the lower pitch and rest pitch. Because the vehicle

passes by you, the radial velocity does not remain constant, but instead varies as a function of the angle between your line of sight and the siren's velocity:

$$f = f_0\left(\frac{c}{c + v_s \cos\theta}\right) \quad (4)$$

where θ is the angle between the object's forward velocity and the line of sight from the object to the observer.

Applications

The Doppler effect is broadly applied to measure the flow velocity of blood in vessels and the heart with ultrasound. A limitation is that the Ultrasound beam should be as parallel to the blood flow as possible. Other limitations are absorption at interfaces and scatter (e.g. on blood cells). Δf, the Doppler shift is 2 times that of 2) since the emitted ultrasound beams impinges *and* reflects on the blood cells. Δf is generally some hundreds of Hz and can directly made audible by a microphone. The ultrasound Doppler flowmetry method has its equivalent by applying a laser: Laser Doppler flowmetry, which is especially used for hypodermal vessels.

Contrast enhanced ultrasound (CEU) using gas-filled microbubble contrast media can be used to improve velocity or other flow-related medical measurements.
Velocity measurements of blood flow are also used in other fields of Echography (obstetric, neurological).
Instruments such as the *Laser Doppler velocimeter* (LDV), and *Acoustic velocimeter* (ADV) have been developed to measure velocities in a fluid flow.

Measurement of the amount of gas bubbles in the venous system is performed in diving medicine. They are generally measured in the pulmonary artery by a Doppler probe at the 3rd intercostal space. Gas bubbles in the circulation may result in decompression sickness.

"Doppler" has become synonymous with "velocity measurement" in medical imaging. But in many cases it is not the frequency shift (Doppler shift) of the received signal that is measured, but the phase shift (*when* the received signal arrives).

The Doppler effect is also a basic 'tool' in astronomy (red shift, temperature measurements by line broadening) and daily life radar (navigation, speed control).

More info

The LDV (also known as laser Doppler anemometry, or LDA) is a technique for measuring the direction and speed of fluids like air and water. In its simplest form, LDV crosses two beams of collimated, monochromatic light in the flow of the fluid being measured. A microscopic pattern of bright and dark stripes forms in the intersection volume. Small particles in the flow pass through this pattern and reflect light towards a detector, with a characteristic frequency indicating the velocity of the particle passing through the probe volume. LDV may be unreliable near solid surfaces, where stray reflections corrupt the signal. The ADV emits an acoustic beam, and measure the Doppler shift in wavelengths of reflections from particles moving with the flow. This technique allows non-intrusive flow measurements, at high precision and high frequency.

The echo-Doppler technique The description of velocity measurement holds for a continuous emitted ultrasound. With pulsed Doppler, short periods of emitting and receiving are alternated. By adjusting the period length the depth of reflection can be selected. In this way, motion of deeper layers is not disturbing the analysis.
The Doppler method and echography are combined in one apparatus, the *echo-Doppler* and yields a duplex image: in the black and white echo-image, direction and velocity are depicted in red for approaching the probe and blue for leaving. The more blue or red the higher the speed. The most common application is Doppler echocardiography.

Lissajous figure

Principle

In mathematics, a Lissajous figure is the graph of a 2D-motion which occurs when two sinusoidal oscillations with the same frequency or different frequencies (a and b), and each with its own amplitude (A and B) and with a phase difference (φ), are added as vectors with one sinus along the horizontal and one along the vertical axis.
In equations:

$$x = A\sin(at+\varphi) \quad \text{and} \quad y = B\sin(bt), \tag{1}$$

where t is time.
Depending on the ratio a/b, the values of A, B and φ, the figure is simple or complicated. The conditions A=B, a/b=1 and φ=0 yields a line, changing φ to φ=π/2 (radians) a circle is obtained and making A≠B an ellipse. Other ratios of a/b produce more complicated curves (Fig. 1), which are closed only if a/b is a rational fraction (a and b integers). The perception of these curves is often a seemingly 3D-trajectory of motion, an example of a visual illusion.

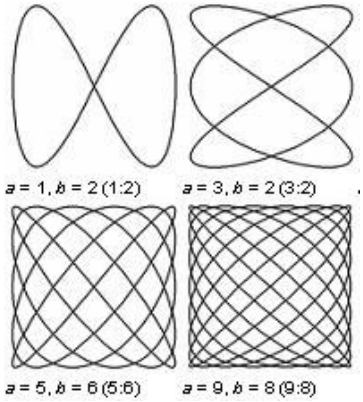


Fig. 1 Examples of Lissajous figures with φ = π/2, a odd, b even, |a – *b*| = 1.

More info

Also parabolas and more complicated figures can be obtained, as depicted in Fig. 2, which shows the relationship of frequency ratio and phase shift φ.

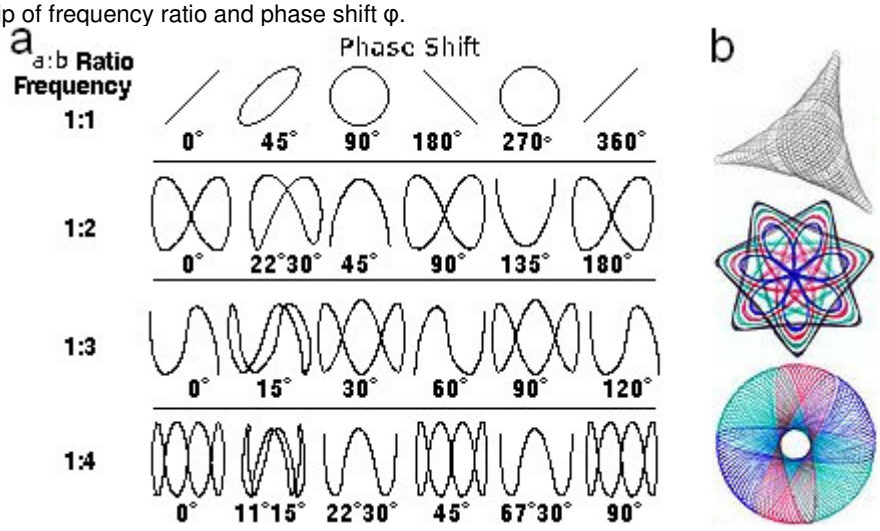


Fig. 2. a. The influence of the phase difference b. Example of spirographs.

Even though they look similar, *spirographs* are different as they are generally enclosed by a circular boundary where a Lissajous curve is bounded by a square (±1, ±1).

Biomechanics and tissue elasticity

Atomic force microscopy

Principle

Atomic force microscopy (AFM) or scanning force microscopy is a very high-resolution type of Scanning probe microscopy, with a resolution of a fraction of a nanometer, more than 1000 times better than the optical diffraction limit (see Light: diffraction).
The precursor to AFM is the Scanning tunneling microscopy. AFM is one of the foremost tools for imaging, measuring and manipulating matter at nanoscale. The term 'microscopy' in the name is actually a misnomer because it implies looking, while in fact the information is gathered by "feeling" the surface with a mechanical probe.
To achieve the resolution, the AFM depends in part on the sharpness of the probe tip and the precision and quietness of its control system. The main parts of the control system are 1) a high-precision positioning actuator in the (x-y) plane of the specimen, usually piezoelectric system; 2) for the tip displacement in the vertical direction (z) a highly sensitive position detection system with a cantilever and optical detection (Fig. 1a) or with a fixed probe and displacement system with a piezo-electric element mounted along the vertical axis of the sample stage; 3) sophisticated control electronics for all axis.

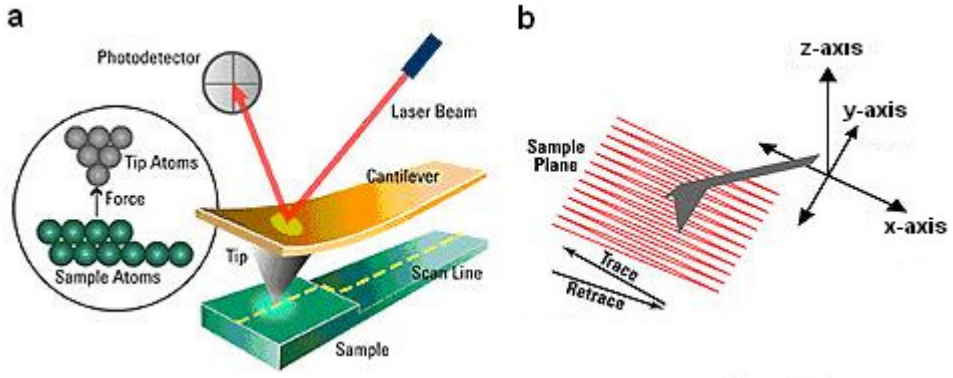


Fig.1 Basic principle of AFM. a. Optical detection system withy cantilever deflection. b. Scanning principle.

The AFM consists of a microscale cantilever with a sharp tip (the probe) at its end that is used to scan the specimen surface. The tip, made of Si, SiO_2 or Si_3N_4 with radius curvature of about 10 nm.
The cantilever is typically Si or silicon nitride with a tip radius of curvature on the order of nanometers. When the tip is brought into proximity of a sample surface, various types of forces between the tip and the sample lead to a deflection of the cantilever according to Hooke's law (see Elasticity and Hooke's law and Stiffness). Typically, the deflection is measured using a Laser spot reflected from the top of the cantilever into an array of photodiodes. Other methods that are used include optical Interferometry, capacitive sensing or piezoresistive (see Piezoelectricity) AFM cantilevers. These cantilevers are fabricated with piezoresistive elements that act as a strain gauge. Using a Wheatstone bridge, strain in the AFM cantilever due to deflection can be measured, but this method is not as sensitive as laser deflection or interferometry.
A feedback mechanism is employed to adjust the tip-to-sample distance to maintain a constant force between the tip and the sample. Traditionally, the sample is mounted on a piezoelectric tube, that can move the sample in the *z* direction for maintaining a constant force, and the *x* and *y* directions for scanning the sample. Alternatively a 'tripod' configuration of three piezo crystals may be employed, with each responsible for scanning in the x,y and z directions. This eliminates some of the distortion effects seen with a tube scanner.
The AFM can be operated in a number of modes, depending on the application. In general, possible imaging modes are divided into static (also called contact) modes and a variety of dynamic (or non-contact) modes.

The AFM has several advantages over the scanning Electron microscopy (SEM). Unlike EM which provides a 2D image, AFM provides a true 3D surface profile. Additionally, samples do not require any special treatments and most AFM modes can work perfectly well in ambient air (no high vacuum) or even a liquid environment, enabling the study of biological macromolecules and even living organisms.

In principle, AFM can provide higher resolution than SEM. It has been shown to give true atomic resolution in ultra-high vacuum. This application is comparable in resolution to Scanning tunneling microscopy and transmission EM (see Electron microscopy: Transmission EM).

A disadvantage of AFM compared with the Scanning electron microscope (SEM) is the image size. The SEM can image an area on the order of mm x mm with a depth of field on the order of millimetres. The AFM can only image a maximum height on the order of a µm and a maximum scanning area of around 150 by 150 µm. Further, scanning is relatively slow.

Application

In science and technology and especially in biomedical sciences. In addition to resolve #D structure, also for instance resistivity, temperature and elasticity can be measured.

More Info

The tip, made of Si, SiO_2 or Si_3N_4, have a radius curvature of about 10 nm. The forces are of the order of 1 nN which can result in cantilever deflections of some 0.1 nm. The resolution has an ultimate limit. Apart from technical limitations this is determined by the thermal vibrations of the involved particles, the atoms or molecules. Reference 1. gives an extensive description of the underlying physics and technology.

Imaging modes

The primary modes of operation are the static (contact) mode and dynamic mode.
Static or contact mode In the static mode operation, the attractive forces can be quite strong, causing the tip to 'snap-in' to the surface. Thus static mode AFM is almost always done in contact where the overall force is repulsive. Consequently, this technique is typically called 'contact mode'. The force between the tip and the surface is kept constant during scanning by maintaining a constant deflection and changes in the oscillation amplitude or phase provide the feedback signal for imaging. In AM, changes in the phase of oscillation can be used to discriminate between different types of materials on the surface. AM can also be operated either in the non-contact or in the intermittent contact regime. In ambient conditions, most samples develop a liquid meniscus layer. Because of this, keeping the probe tip close enough to the sample for short-range forces to become detectable while preventing the tip from sticking to the surface presents a major hurdle for the non-contact dynamic mode in ambient conditions.
Dynamic mode In the dynamic mode with frequency modulation (FM) or amplitude modulation (AM), the cantilever is externally oscillated at or close to its resonance frequency (see Linear second order system). In FM, the oscillation amplitude, phase and resonance frequency are modified by tip-sample interaction forces; these changes in oscillation with respect to the external reference oscillation provide information about the sample's characteristics. In FM, stiff cantilevers provide stability very close to the surface and as a result, this technique was the first AFM technique to provide true atomic resolution in ultra-high vacuum conditions changes in the oscillation frequency provide information about tip-sample interactions. Frequency can be measured with very high sensitivity and thus the frequency modulation mode allows for the use of very stiff cantilevers. FM has also been used in the non-contact regime to image with atomic resolution by using very stiff cantilevers and small amplitudes in an ultra-high vacuum environment.

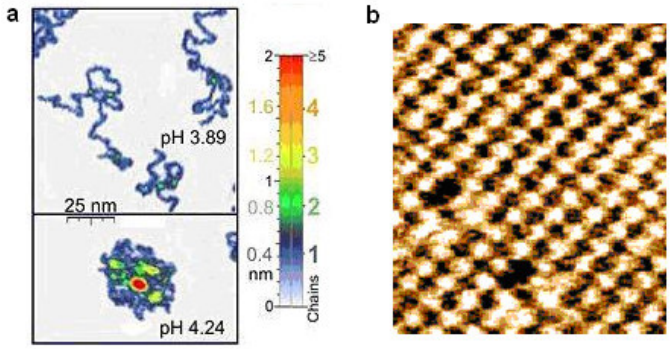


Fig. 2 Examples of AFM. a. Single polymer chains (0.4 nm thick) recorded in a tapping mode with different pH. Green locations of the two-chains-superposition correspond to 0.8 nm thickness. b. The atoms of a NaCl crystal.

Tapping mode Dynamic contact mode (also called intermittent contact or tapping mode) was developed to bypass this problem. In dynamic contact mode, the cantilever is oscillated up and down at near its resonance frequency by a small piezoelectric element mounted in the AFM tip holder, such that it comes in contact with the sample with each cycle, and then enough restoring force is provided by the

cantilever spring to detach the tip from the sample. Tapping mode is gentle enough even for the visualization of adsorbed single polymer molecules (for instance, 0.4 nm thick chains of synthetic polyelectrolytes) under liquid medium.

Force-distance measurements
Another application is the measurement of force-distance curves (a few pN versus tens of nm). Here, the AFM tip is approached towards and retracted from the surface and the static deflection of the cantilever is monitored as a function of piezo displacement.

Identification of individual surface atoms
The AFM can be used to image and manipulate atoms and structures on a variety of surfaces. The atom at the apex of the tip "senses" individual atoms on the underlying surface when it forms incipient chemical bonds with each atom. Because these chemical interactions subtly alter the tip's vibration frequency, they can be detected and mapped.
The trick is to first measure these forces precisely for each type of atom expected in the sample. Each different type of atom can be identified in the matrix as the tip is moved across the surface. Such a technique has been used now in cell biology.

Literature
http://www.bphys.uni-linz.ac.at/bioph/download/Principles%20of%20Atomic%20Force%20Microscopy.pdf
http://www.afmuniversity.org/index.cgi?CONTENT_ID=33
F. Giessibl, Advances in Atomic Force Microscopy, Reviews of Modern Physics 75 (3), 949-983 (2003).

Ballistocardiography

Principle

Ballistocardiography is a noninvasive technique for the assessment of cardiac function by detecting and measure the recoil (the reaction) of the human body due to the blood that the heart is currently pumping (the action). It is the derivative of the momentum (mass x velocity) and consequently has the dimension mass x length/time^2.
To be more precise a ballistocardiogram (BCG) measures the impact of blood colliding with the aortic arch, which causes the body to have an upward thrust (reaction force) and then the downward thrust of the blood descending. The ballistocardiogram is in the 1-20 Hz frequency range.
One example of the use of a BCG is a ballistocardiographic scale, which measures the recoil of the person's body that is on the scale. A BCG scale is able to show a persons heart rate as well as his weight.
Sensors are often hidden in the upholstery of a chair mounted on the scale and the electronics is also hidden. In this way the subject is not aware of recording.

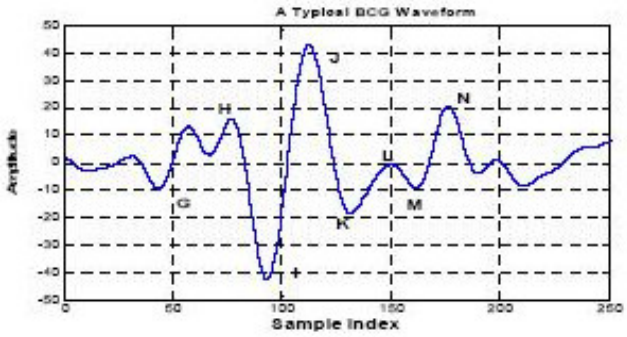


Fig. 1 A BCG signal with spikes and wave complexes (from ref. 1).

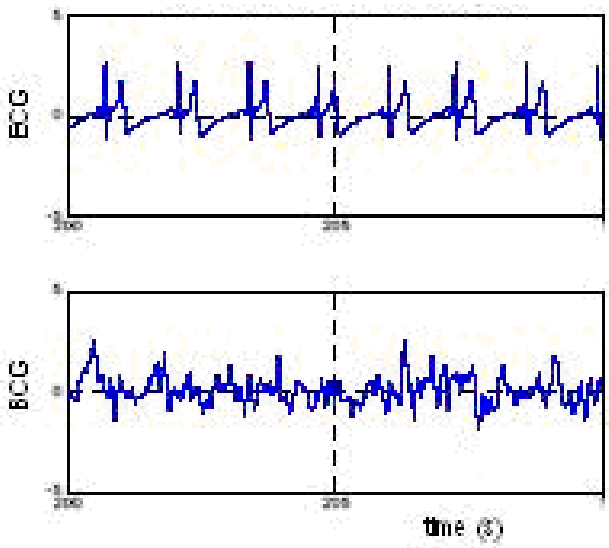


Fig. 2 ECG and BCG (lower panel) records of a normal subject using 1-45 Hz band pass filter for ECG and 1-10 HZ for BCG. (Motion artifacts in BCG signal are not removed).

Application

The charm of the method is that no electrodes are needed to be attached to the body during measurements. This provides a potential application to assess a patient's heart condition at home. Recordings can be transmitted real time to the clinic.
The BCG can show main heart malfunctions by observing and analyzing the BCG signal.

BCG is used in heamodynamic modeling for calculating pulsatile fluid flows with large Womersley numbers in large arteries around the heart and valves.
A 3D version of BCG has been used in spaceflight during free-floating microgravity.

More info

A BCG can be recorded from the surface of body with accelerometers, specific piezoelectric foil sensors or charge-capacitive sensors (e.g. EMFi sensor). The measured signal is amplified by a charge amplifier.
Often the BCG is recorded together with the ECG (one lead by strips of copper on the arm rests to measure R-peaks for heart rate), respiratory rate, respiratory amplitude, and body movements, all wireless. This integrated technique is also called the static charge-sensitive-bed (SCSB) method. It is for instance used for recording body movements after exercise and during sleep.
As holds for many biological signals (EEG, ECG etc.) analysis is nowadays often performed by wavelet analysis or an or another component analysis. The analysis may be proceeded by applying an artificial neural networks (ANNs).

Literature

1. www.arehna.di.uoa.gr/Eusipco2005/defevent/papers/cr1069.pdf -
2. Akhbardeh A., Junnila S., Koivistoinen T., Värri A. Applying Novel Supervised Fuzzy Adaptive Resonance Theory (SFART) Neural Network and Biorthogonal Wavelets for Ballistocardiogram Diagnosis. Proceedings of the 2006 IEEE, International Symposium on Intelligent Control, Munich, Germany, October 4-6, 2006.
3. Xinsheng Yu; Dent, D.; Osborn, C. Classification of ballistocardiography using wavelet transform and neural networks. Engineering in Medicine and Biology Society, 1996. Bridging Disciplines for Biomedicine. Proceedings of the 18th Annual International Conference of the IEEE. 1996, 3,:937 - 938

Biomechanics

Principle

Biomechanics is the research and analysis of the mechanics of living organisms on multiple levels, from the level of molecules to tissue, organ and organism.
Some simple examples of biomechanics research include the investigation of the forces that act on limbs, static as well as dynamic, the aerodynamics of bird and insect flight, hydrodynamics of swimming in fish, and locomotion in general across all forms of life, from individual cells to whole organisms. The biomechanics of human motion is called kinesiology.

Applied mechanics, and mechanical engineering disciplines such as fluid mechanics and solid mechanics, play prominent roles in the study of biomechanics. By applying the laws and concepts of physics, biomechanical mechanisms and structures can be simulated and studied.
Loads and deformations can affect the properties of living tissue. For example, the effects of elevated Blood pressure on the mechanics of the arterial wall and bone growth in response to exercise can remodel the vessel or bone anatomy.
Biomechanics in sports science attempts to improve performance in athletic events through modeling, simulation, and measurement.

Application

The following fields of research and application can be considered as Biomechanics.

Fluid dynamics
applies to gases and liquids (see Poiseuille's Law, Reynolds Number, Flow: entrance effect and entrance length, Flow in a bended tube, Flow in bifurcations, Flow through a stenosis, Navier-Stokes equations) of circulation (see Blood flow, Bernoulli's and Pascal's Law, Womersley number) and the respiratory system (Lung gas transport 1, basic principles, Lung gas transport 2, pressure, volume and flow, Lung gas transport 3, resistance and compliance).

Biomechanics of bones and soft tissues
Bones are anisotropic (different properties in the different directions) and so bones are strongest along one particular axis and less strong along the two other axes, with the latter two mostly of the same strength.
Soft tissues such as skin, tendon, ligament and cartilage are combinations of matrix proteins and fluid. The function of *tendons* is to connect muscle with bone and is subjected to tensile loads. Tendons must be strong to facilitate movement of the body while at the same time remaining compliant to prevent damage to the muscle tissues. *Ligaments* connect bone to bone and therefore are stiffer than tendons but are relatively close in their Tensile strength. Cartilage, on the other hand, is primarily loaded in compression and acts as a cushion in the joints to distribute loads between bones.
The stress-strain relations can be modeled using Hooke's Law, in which they are related by linear constants known as the Young's modulus or the elastic modulus, the shear modulus and Poisson's ratio. See for detailed descriptions Elasticity and Hooke's law, Elasticity 1: elastic or Young's modulus, Elasticity 2: shear modulus, Elasticity 3: compressibility and bulk modulus, Stiffness and Tensile strength.

Biomechanics of muscle
The biomechanics of muscle, being skeletal muscle (striated), cardiac muscle (striated) or smooth muscle (not striated) is highly dynamic and often not linear.

More Info

Biomechanics of soft tissue
The theory rather well applies to bone but much les to soft tissues due to non-linear behavior and evident visco-elasticity properties. With visco-elasticity there is energy dissipation, or Hysteresis, between the loading and unloading of the tissue during mechanical tests. Some soft tissues can be preconditioned by repetitive cyclic loading to the extent where the stress-strain curves (see Tensile strength) for the loading and unloading portions of the tests nearly overlap.

Elasticity and Hooke's law

Principle

Hooke's law of elasticity is an approximation which states that the amount by which a material body is deformed (the *strain*) is *linearly related to* the force (the *stress*) causing the deformation. Materials for which Hooke's law is a useful approximation are known as linear-elastic or "Hookean" materials.
As a simple example, if a spring is elongated by some distance ΔL, the restoring force exerted by the spring F, is proportional to ΔL by a constant factor k, the spring constant. Basically, the extension produced is proportional to the load. That is,

$$F = - k \Delta L. \quad (1a)$$

The negative sign indicates that the force exerted by the spring is in direct opposition to the direction of displacement. It is called a "restoring force", as it tends to restore the system to equilibrium. The potential energy stored in a spring is given by:

$$U = 0.5\, k\, \Delta L^2. \quad (1b)$$

This comes from adding up the energy it takes to incrementally compress the spring. That is, the integral of force over distance. This potential can be visualized as a parabola on the U-ΔL plane. As the spring is stretched in the positive L-direction, the potential energy increases (the same thing happens as the spring is compressed). The corresponding point on the potential energy curve is higher than that corresponding to the equilibrium position (ΔL = 0). Therefore, the tendency for the spring is to decrease its potential energy by returning to its equilibrium (un-stretched) position, just as a ball rolls downhill to decrease its gravitational potential energy. If a mass is attached to the end of such a spring and the system is bumped, it will oscillate with a natural frequency (or, in other words, resonate with a particular frequency, see Linear second order system).

For a Hookean material it also holds that ΔL is reciprocally proportional to its cross sectional area A, so $\Delta L \sim A^{-1}$ and ΔL ~ L. When this all holds, we say that the spring is a linear spring. So, Hooke's law, i.e. equation (1a) holds. Generally ΔL is small compared to L.
For many applications, a rod with length *L* and cross sectional area A, can be treated as a linear spring. Its relative extension, the *strain* is denoted by ε and the tension in the material per unit area, the *tensile stress*, by σ. Tensile stress or tension is the stress state leading to expansion; that is, the length of a material tends to increase in the tensile direction.
In formula:

$$\varepsilon = \Delta L/L, \quad \text{(dimensionless)} \quad (2a)$$
$$\sigma = E\varepsilon = E\Delta L/L = F/A, \quad (\mathrm{N/m^2} \equiv \mathrm{Pa}) \quad (2b)$$

where ΔL is the extension and *E* the modulus of elasticity, also called Young's modulus. Notice that a large *E* yields a small ΔL. *E* is a measure of the stiffness (not the stiffness to prevent bending; see Stiffness) and reciprocal to the mechanical compliance. As (2b) says, *E* is the ratio σ/ε and so the slope of the stress/strain (σ/ε) curve, see Fig. 1.

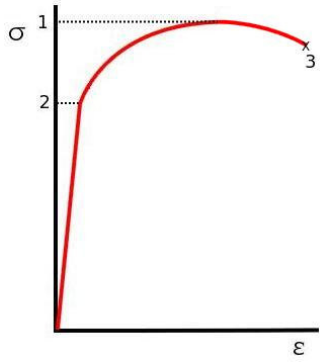


Fig. 1 Stress-strain curve. The slope of the linear part is by definition *E*. 1. ultimate strength, 2 limit of proportional stress. 3 rupture. Notice that ε is the *fraction* of extension (dimensionless) and σ the tensile *force/area*.

The modulus of elasticity *E* is a measure of the stiffness of a given material to prevent linear extension. Hooke's law is only valid for the linear part, the elastic range, of the stress-strain curve (see Fig. 1). When the deforming force increases more and more, the behavior becomes non-linear, i.e. the stress-strain curve deviates from a straight line. In the non-linear part *E* is defined as the rate of change of stress and strain, given by the slope of the curve of Fig. 1, obtained by a tensile test. Outside the linear range for many materials like metals stress generally gives permanent deformation.
With further increasing stress the slope diminishes more and more with increasing strain and materials like metals liquidize. Finally it ruptures. See Tensile strength for more info how materials with different mechanical properties, such as bone and tendon, behave before rupturing.

Materials such as rubber, for which Hooke's law is never valid, are known as "non-hookean". The stiffness of rubber is not only stress dependent, but is also very sensitive to temperature and loading rate.

Applications

Strength calculations of skeleton and cartilage structures, of tendons (especially the Achilles and patella tendon, also aging studies) and the elastic behavior of muscles, blood vessels and alveoli. Further aging of bone, thrombolysis (compression modulus, see **More Info**).There are also applications in the hearing system: calculations of the mechanical behavior of the tympanum, middle ear ossicles, auditory windows and the cochlear membranes.
Failure strains (points of rupture) are important to know, especially for biomaterials. For tendons and muscle (unloaded) they go up to 10%, sometimes until 30%. Ligaments have occasionally a failure strain up to 50%. Spider fiber reaches 30% and rubber strands more than 100%. However, these high values all include plastic deformation.
Rubber is a (bio)material with one of the most exotic properties. Its stress-strain behavior exhibits the so called Mullins effect (a kind of memory effect) and Payne effect (a kind of strain dependency) and is often modeled as hyperelastic (strain energy density dependency). Its applications, also in medicine, are numerous.

More Info

Application of the law as described above can become more complicated.

Linear vs non-linear This has already been discussed. There exists all kind of non-linearities, depending on the material (steel-like, aluminum-like, brittle-like as bone, elastic-like as tendons).
Isotropic and anisotropic. Most metals and ceramics, along with many other materials, are isotropic: their mechanical properties are the same in all directions. A small group of materials, as carbon fiber and composites have a different *E* in different directions. They are anisotropic. Many biomaterials, like muscle, bone, tendon, wood, are also anisotropic. Therefore, *E* is not the same in all three directions. Generally in the length of the structure it is different than in both par-axial directions. Now, σ and ε comprise each 3x3 terms. This gives the tensor expression of Hooke's Law and complicates calculations for biomaterials considerably. Often there is circular symmetry (muscle, tendon) which brings the dimensionality down to 2D (2D elasticity)
Inhomogenity Sometimes, a (biological) material is not homogeneous in some direction, so *E* changes along some axis. This happens in trabecular bone when it is denser at the surface.

Literature
http://www.vki.ac.be/research/themes/annualsurvey/2002/biological_fluid_ea1603v1.pdf

Elasticity 1: elastic or Young's modulus

Principle

There are three primary elastic moduli, each describing a different kind of deformation. These are the elasticity or Young's modulus *E*, shear modulus G (see Elasticity 2: Shear Strength) and the bulk modulus K (see Elasticity 3: compressibility and bulk modulus):
Because all elastic moduli can be derived from Young's modulus, the latter is often referred to simply as the *elastic modulus*.
The modulus of elasticity *E* is a measure of the stiffness of a given material or its resistance to be stretched when a force is applied to it. The elastic modulus is defined as the slope of the curve of tensile stress and tensile strain:

$$E = \Delta\sigma / \Delta\varepsilon, \qquad (1)$$

where relative extension (strain) is denoted by ε and the tension in the material per unit area, the tensile stress, by σ. Because strain ε is a unit-less ratio, σ and *E* are measured in Pa. For small values of ε and σ, *E* is constant for some materials (the stress strain curve is linear). But often, especially with biomaterials, generally the curve is nowhere a straight line.
For the linear behavior (straight line) we have:

$$E = \frac{\sigma}{\varepsilon} = \frac{F / A}{\Delta L / L} \qquad (2)$$

Table 1 gives some values of *E* for some materials applied in prostheses and biomaterials (room temperature):
Table 1 Young's modulus for some (bio)materials

material	*E* (GPa)	G (GPa)
rubber	0.001	
ZYLON PBO	138-167	
steel	200	75.8
titanium		41.4
vessel wall	0.0001-0.001	
muscle	0.00001	0.00006
tendon	0.1	
bone	0.1-10	4.5 – 8.0*

* human pipe bones

For solids G (Elasticity 2: shear modulus) is about halve K (Elasticity 3: compressibility and bulk modulus). See for a further description Elasticity and Hooke's law.
Aging makes many biomaterials more brittle with lowering ultimate strength. For bone this is 2%/decade with a 2% increase of *E*.

Application

They are found in the field of the biomechanics of bone and tendon structures in basic medical science. The modulus of elasticity (and other material properties) is of importance for all kind of implants, such as stents, tubes, wires, cardial valves, artificial hearts, etc. and for dental and orthopedic prostheses etc.

More Info

When an object has an isotropic molecular crystalline structure, *E* as well as K are larger than G. In other words, the crystal resist stronger against strain and compression than against shear. Strain gives a distance increase of some particle with the neighboring particles in the direction of the applied force and a decrease in the two perpendicular directions, but to a smaller amount. The latter depends on the Poisson's ratio (see Elasticity 3: compressibility and bulk modulus). In the unstressed state this potential energy, the result of solely internal forces, is minimal. Whit strain, external force is added and the potential energy rises.
Compression gives a distance decrease of some particle with all neighboring particles that gives the largest change in potential energy. With shear there is only a distance change in one dimension. Shear and strain makes the crystal anisotropic. This can change various material properties.
Young's modulus is a physical consequence of the Pauli exclusion principle.

Elasticity 2: shear modulus

Principle

The shear modulus or modulus of rigidity (G or sometimes S) is a type of elastic modulus which describes an object's tendency to shear (the deformation of shape at constant volume) when acted upon by opposing forces; it is defined as shear stress over shear strain.

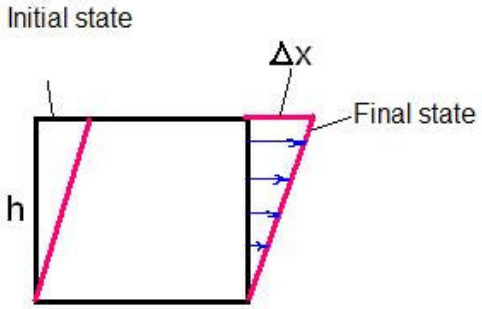


Fig. 1 Deformation by shear stress.

$$G = \frac{F/A}{\Delta x/h} = \frac{Fh}{\Delta xA}, \qquad (1)$$

where F/A is shear stress (force F over cross area A) and $\Delta x/h$ is shear strain (distance of deformation Δx over height h).
The shear modulus is one of several quantities, such as compressibility, stiffness, hardness, and toughness (see Tensile strength) for measuring the strength of materials. All of them arise in the generalized Hooke's law (see Elasticity and Hooke's law).
Anisotropic materials such as wood and bone are poorly described by elastic moduli.
G and *E* are related by the Poisson's ratio μ (see Elasticity 3: compressibility and bulk modulus):

$$G = E/(2+\mu) \qquad (2)$$

Shear modulus is usually measured in GPa. See Elasticity 1: elastic or Young's modulus for the values of G of some materials.

Application

G is smaller than *E* as (2) shows. Therefore, bone fracture mostly occurs when bended or twisted which gives a high shear. In contrast, tendons rupture by tensile stress.

More Info

In solids, there are two kinds of sound waves, pressure waves and shear waves. The speed of sound for shear waves is controlled by the shear modulus.
The shear modulus is part of the derivation of viscosity.

Elasticity 3: compressibility and bulk modulus

Principle

The bulk modulus *(K)* describes volumetric elasticity, or the tendency of an object's volume to deform when under pressure. It is defined as volumetric compressive stress over volumetric compression, and is the inverse of compressibility.

Compressive strength is the capacity of a material to withstand axially directed pushing forces. When the limit of compressive strength is reached, materials are crushed (bone, concrete). An object under high pressure may undergo processes that will affect the stress-strain curve (see Tensile Strength), due to chemical reactions.

Application

Material choice for the design of prostheses and strength calculations of biomaterials like bone. When a mammalian or avian tubular bone with thin walls is bended, in the outer curvature there is tensile stress and in the inner curvature compressive stress. The shear stress (see Elasticity 2: Shear Modulus) is small but becomes more important the thicker the wall.
Bone can withstand greater 1D compressive stress than tensile stress. K of bone, ca. 50 GPa, is some 2-4 times that of concrete.

More info

The Poisson's ratio μ
Here, for simplicity it is assumed that in the three dimensions material properties are the same (a so called isotropic material). When a beam is stretched by tensile force (see Tensile Strength), its diameter (d) changes. The change in diameter Δd, generally a decrease, can be calculated:

$$\Delta d/d = -\mu\, \Delta L/L, \qquad (1a)$$

where L is the length of the beam, and ΔL the length increase by the tensil force F (see also Tensil Strength). The constant of proportionality μ is the *Poisson's ratio*. Since ΔL/L = F/(*E*A) with A is the cross sectional area it holds that:

$$\Delta d/d = -\mu\, F/(EA). \qquad (1b)$$

The range of μ is from 0 to 0.5, and mostly 0.2 < μ < 0.4. For metals μ ≈ 1/3 and for rubber μ ≈ 0.5, which means that rubber hardly changes volume. Some materials behave atypically: increase of d when stretched. This happens with polymer foam.
Since generally ΔL << L, it can easily been proven that the change in volume ΔV (generally increase) is:

$$\Delta V = d^2\Delta L + 2dL\,\Delta d. \qquad (2a)$$

Substitution in (1a) and dividing by $V = d^2L$ gives the relative volume change of the beam:

$$\Delta V/V = (1-2\,\mu)\Delta L/L = (1-2\,\mu)F/EA \qquad (2b)$$

With the material constants *E* and μ the changes in shape of an isotropic elastic body can be described when an object is stretched or compressed along one dimension. Objects can be subjected to a compressive force in all directions. For this case a new material constant, the compressive strength or bulk modulus K, comprising *E* and μ, is introduced. The bulk modulus can be seen as an extension of Young's modulus *E* (see Elasticity 1: elastic or Young's modulus) to three dimensions.

Derivation of K
When an elastic body, here as an example a cube (edge length b), is compressed from all sides by a pressure increase of ΔP, then the change of the edge length Δb is:

$$\Delta b = -\frac{1-2\mu}{E} b \cdot \Delta P \qquad (3a)$$

The volume change is by approximation:

$$\Delta V \approx 3b^2\,\Delta b. \qquad (3b)$$

After substitution of Δb of (3b) in (3a) it follows that:

$$\frac{\Delta V}{V} = -\frac{3(1-2\mu)}{E}\Delta P \qquad (4a)$$

The bulk modulus K is defined as:

$$K = \frac{E}{3(1-2\mu)} \qquad (4b)$$

The value of μ determines whether K is smaller (brittle materials), about equal (many metals) or larger (highly elastic materials like rubber) than *E*. Finally, It follows that

$$\frac{\Delta V}{V} = -\frac{1}{K}\Delta P \qquad (4c)$$

This equation says that the higher the compressive strength (K), the smaller the volume changes. The material constant K (in Pa) is known for many materials.

Elasticity of the aorta

Nico A.M. Schellart, Dept. of Med. Physics, AMC

Principle

When the heart contracts, blood is pumped into the aorta. Then, pressure P_a, flow and volume V of the aorta increase. We consider now firstly how under assumed validity of the law at of Hooke and Laplace the P-V relation will be. Firstly, we calculate using the law of Laplace how the flow of the aorta relates to P_a.

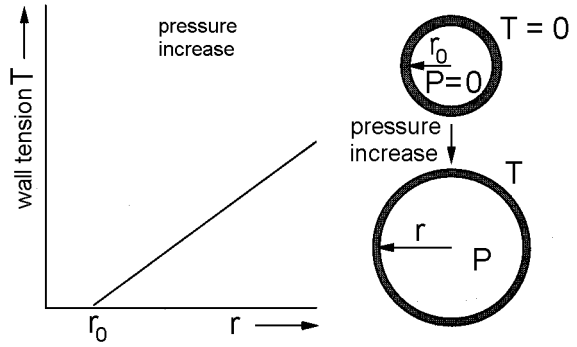


Fig. 1 Pressure increase causes dilatation of the cylinder

Since the force in the aorta wall is $F_w = 2P_a r_i l$, where r_i the aorta inner radius and *l* its length. Without overpressure, the half inner circumference of the halve of the aorta cylinder is $w = \pi r_{i,0}$ with $r_{i,0}$, the inner radius before the blood ejection of the left ventricle (Fig. 1, right upper part). Then the material cross section is $A = l \cdot d_0$. After ejection the inner radius is increased to $r_{i,t}$. and the change of w, the strain of the elastic aorta wall, is $\Delta w = \pi(r_{i,t} - r_{i,0})$. The wall stress goes from (assumed) zero (equilibrium before ejection) to a value given by Hooke's law:

$$\sigma = E\varepsilon = E\Delta L/L = F/A, \qquad (1)$$

(see Elasticity and Hooke's law), and after ejection, we obtain:

$$\sigma = F_w/A = E\Delta w/w = E(r_{i,t} - r_{i,0})/r_{i,0}. \qquad (2)$$

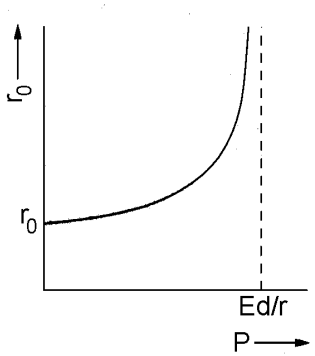


Fig. 2 The relation between the radius r of a cylinder and the pressure in the cylinder

This gives to how the wall tension increases at a cylinder, if by pressure increase the radius of the cylinder increases. In literature one finds except wall stress σ (in N/m^2) also wall tension T= Pr, generally given in N/m. Substitution of $T = P_a r_i$ in (2) provides (after conversion):

$$r_{i,t} = r_{i,0}(1- P_a r_{i,0}/Ed)^{-1}. \quad (3)$$

This function is illustrated in Fig. 2.
When the pressure in the aorta increases, also the volume V increases. Using $V = \pi r_{i,0}^2 l$ it follows that:

$$V(P_a)= V_0(1- P_a r_{i,0}/Ed)^{-2}. \quad (4)$$

All values must be expressed in preferably the SI system (otherwise convert, e.g. 100 mmHg ~ 13600 N/m^2 = 13.6 kPa). At the age of 50 years, realistic values for the aorta are $R_{i,0}$ = 5.6 mm, E = 5.105 N/m^2, d = 2 mm, V_o=40 mL and *l* = 40 cm.
Fig. 3 depicts equation (4). Above 70 mmHg, this P-V relation ceases to be quadratic. V increases much less and finally the relation becomes sigmoid. The model, derived from the linear approach of elasticity (law of Hooke), only holds well at low values of P_a. At higher values of P_a, collagen in the vessel wall will dominate the properties of the elastine.

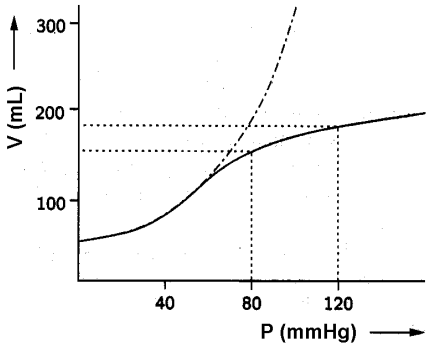


Fig. 3 Aorta volume V as a function of the pressure of the aorta

More Info

As a measure for the ductility of the aorta (compliance) C is used, either in its dynamic (C_{dyn}) or static (C_{stat}) form. C_{dyn} is the pressure derivative of the volume/pressure curve (drawn line in Fig. 3): $C_{dyn} = dV/dP_a$. Fig. 4 shows the impact of the compliance of the aorta. The drawn line represents the pressure in the left ventricle and the dotted line that in the aorta. An approach of the dynamic compliance between 80 and 120 mmHg (the regular values) is obtained by taking dV/dP_a at 100 mmHg, which can be approximated by $C_{dyn,100}$ = (185-160)/(120- 80) = 0.625 mL/mmHg). The static compliance C_{stat} is V/P_a. For P_a = 100 mmHg it follows that $C_{stat,100} \approx 170/100 = 1.7$ mL/(mmHg). When the volume-pressure relation of the aorta is approached by a straight line between the origin and the point with pressure P_a (the ' work point '), the slope of this line represents static compliance. This definition of C_{stat} holds for the aorta as a whole, here with an assumed length of 40 cm. In the literature, C_{stat} is sometimes expressed per unit length of the vessel. The compliance of the aorta results in a rather regular flow, in spite of the pulse-wise time course due to the contractions of the heart.

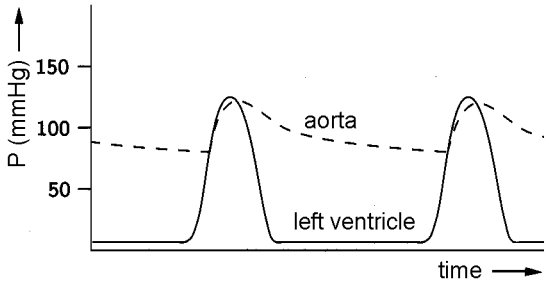


Fig. 4 Effect of the compliance on the aorta pressure

Fig. 4 shows that after closing the aorta valve P_a (and volume) decreases about exponentially. The total pulsation of P_a is much smaller than that of the ventricle due to the high aorta compliance. If the aorta would be a rigid tube, the blood flow would be more strongly pulsating. This will result in a larger mechanical load of the rest of the circulatory system as follows from the Windkessel model (see Blood pressure: Windkessel model) of the circulatory system. Aging diminish the 'empty' volume V_0 of the aorta. This hardly affects C_{stat} but it seriously decreases C_{dyn}: from approximately 0.6 mL/mmHg at the age of 50 to approximately 0.2 mL/mmHg at 80 years.

Literature

Van Oosterom, A and Oostendorp, T.F. Medische Fysica, 2nd edition, Elsevier gezondheidszorg, Maarssen, 2001.
N. Westerhof, Noble M.I.M and Stergiopulos N. Snapshots of hemodynamics: an aid for clinical research and graduate education, 2004, Springer Verlag.

Laplace's law

Principle

The law of Laplace holds for fluid or gas filled hollow objects (e.g. blood vessels, heart, alveoli). It gives the relation between transmural pressure P, and wall tension, T. With a soap bubble as an example, the classical approach, this tension is directly related to the Surface tension and consequently has the dimension N/m.
For a circular cylinder with length *l* and an inner radius r_i., as model of a blood vessel, P acts to push the two halves apart with a force equal to P times the area of the horizontal projection of the cylinder, being $2l \cdot r_i$ (Fig. 1). It acts on both halves, so the force F_P is:

$$F_P = 2P l r_i. \quad (1a)$$

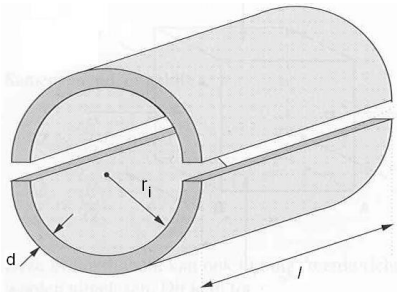


Fig. 1

The two halves are kept together by wall stress, σ_w, acting in the wall only. The related force F_w, visualized in Fig. 2, is thus:

$$F_W = 2\sigma_w l \cdot d, \quad (1b)$$

where d the wall thickness. 2d*l* is two times the longitudinal cross area of the wall itself, at which this force is acting (Fig. 2). This force is in equilibrium with F_P and thus: $2Pl r_i = 2\sigma_w l \cdot d$. This results in the form of the law of Laplace as mostly used in hemodynamics. It presents the relation between P within the lumen and σ:

$$\sigma_w = P r_i d^{-1}. \quad (2)$$

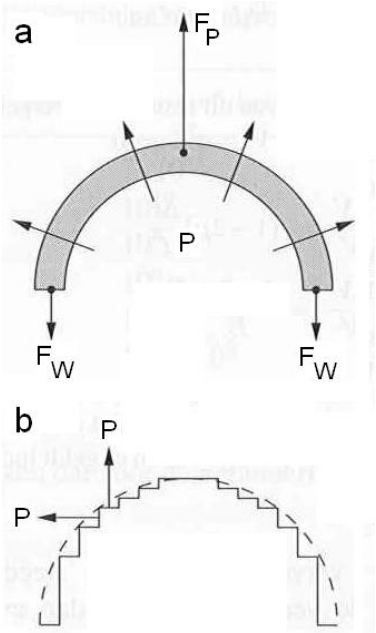


Fig. 2 (a) Visualization of the forces acting on the wall and in the wall. (b) Cancellation of the sideward directed forces due to P. The horizontal projection of the inside of the cylinder and P directly gives F_P (eq. 1a).

Stress has the dimension N/m^2, the same as pressure. In other words it says that pressure and wall stress are related by the ratio of radius over wall thickness. With constant d and P, but increasing r_i, σ_w increases proportionally with r_i, At a certain r_i, ultimate strength is reached (see Tensile strength) and rupture occurs (along the length axis of the cylinder).

Another measure is the wall tension T, the force per unit length of the cylinder:

$$T_W = F_W (2l)^{-1} = P r_i. \quad (3a)$$

For a sphere, a similar derivation holds and the result is $\sigma_w = \frac{1}{2} P r_i / d$ and T_W is:

$$T_W = \frac{1}{2} P r_i. \quad (3b)$$

Application

The law is of importance in (experimental) vascular medicine and cardiology, and experimental pulmology.
Assuming a simple shape such as a sphere, or circular cylinder, the law may be applied to the ventricular wall in diastole and systole, as well as to vessel walls. The law of Laplace can also be used in the contracting heart, where the force is generated in the wall and the rise in pressure is the result of the contracting muscle.
Since the wall cross sectional area of a blood vessel is constant, d decreases inversely with r_i, and consequently $\sigma_w \sim r_i^2$. Therefore, with hypertension, rupture of a vein or aneurism may occur with a small increase of r_i.

More Info

A more general form of the Laplace equation holding for a hollow organ of any shape is $T_W = P r_c$, where r_c, is the (local) curvature radius. The “mean” curvature radius of any 3D shape is defined as:

$$1/r_c = 1/r_1 + 1/r_2, \quad (4)$$

where r_1 is the radius of the largest tangent circle and r_2, that of the largest one perpendicular on the plane of the former. For a sphere $r_1 = r_2$, and for a cylinder $r_1 = \infty$ and $r_2, = r_i$. After calculating r_c for the cylinder and sphere respectively (3a) and (3b) directly follow.

The law of Laplace applies to any geometry, irrespective of whether the material is linear or nonlinear elastic or if the wall is thin or thick. The only limitation of Laplace's law is that it yields the average wall stress and thus it cannot give any information on the stress distribution across the wall. For cylindrical geometries, and assuming linearly elastic (Hookean, see Elasticity and Hooke's law) material the distribution of circumferential stress or hoop stress across the wall thickness can be approximated by:

$$\sigma_{w,r} = Pr_i^2 (1+ r_o^2/r^2)/ (r_o^2- r_i^2), \quad (5)$$

where r_o and r_i are the external and internal radius, respectively, and r is the position within the wall for which local stress is calculated.
The relevant force related to (local) wall stress of the heart muscle of the left ventricle is often calculated for the ventricular “equatorial” plane. It is $F = P \cdot A_e$, where A_e is equatorial cavity cross-sectional area and P luminal pressure. The wall stress σ_w is given by F/A_w, with A_w the equatorial cross-sectional area of the muscle ring. Thus $\sigma_w = P \cdot A_e/A_w$.
Many other relations between wall force or stress and ventricular pressure have been reported, but since measurement of wall force is still not possible, it is difficult to decide which relation is best.

Relation to the Young’s modulus
For a relatively thin arterial wall ($d \ll r_i$ and incompressible) one can use Laplace's Law to derive the following expression for the incremental elastic modulus:

$$E = (r_i^2/d)\Delta P/\Delta r_i. \quad (6)$$

For thick walls, as is often the case in arteries, the Young’s modulus (see Elasticity 1: elastic or Young’s modulus) is best derived from the measurement of pressure and radius:

$$E = 3\, r_i^2\, r_o(\Delta P/2\Delta r_o)/(r_o^2- r_i^2) \quad (7)$$

Literature
Van Oosterom, A and Oostendorp, T.F. Medische Fysica, 2nd edition, Elsevier gezondheidszorg, Maarssen, 2001.
N. Westerhof, Noble M.I.M and Stergiopulos N. Snapshots of hemodynamics: an aid for clinical research and graduate education, 2004, Springer Verlag.

Scanning Probe Microscopy

Principle

Scanning Probe Microscopy (SPM) forms images of surfaces using a mechanical probe that scans the specimen. An image of the surface is obtained by mechanically moving the probe in a raster scan of the specimen, line by line, and recording the probe-surface interaction as a function of position. The 2D grid of data points provides the information to construct a 3D- image. SPM was founded with the invention of Scanning tunneling microscopy.
Many scanning probe microscopes can image several interactions simultaneously. The manner of using these interactions to obtain an image is generally called a mode.
The resolution varies somewhat from technique to technique. This depends largely on the ability of piezoelectric actuators (see Piezoelectricity) to execute motions with a precision and accuracy at the atomic level or better on electronic command. One could rightly call this family of technique 'piezoelectric techniques'.
Probes are cut or etched from Pt/Ir or Au wires.
Well-known types of SPM are Atomic force microscopy and Scanning tunneling microscopy followed by Magnetic force microscopy.

Application

In science and technology and especially in biomedical sciences.

More info

Advantages of scanning probe microscopy

- The resolution of the microscopes is not limited by diffraction (see Light: diffraction), but only by the size of the probe-sample interaction volume, which can be as small as a few picometres. Hence, the ability to measure small local differences in object height is a novelty. Laterally, the probe-sample interaction extends only across the tip atom or atoms involved in the interaction.
- The interaction can be used to modify the sample to create small structures (nanolithography).
- Specimens do not require a partial vacuum but can be observed in air at standard temperature and pressure or while submerged in a liquid reaction vessel.

Disadvantages of scanning probe microscopy

- The detailed shape of the scanning tip is sometimes difficult to determine. Its effect on the resulting data is particularly noticeable if the specimen varies greatly in height over lateral distances of 10 nm or less.
- The scanning techniques are generally slower in acquiring images, due to the scanning process. Like all scanning techniques, the embedding of spatial information into a time sequence opens the door to uncertainties in lateral spacing and angle. This can arise from specimen drift, feedback loop oscillation, and mechanical vibration.
- The maximum image size is generally smaller.

Stiffness

Principle

Stiffness is the resistance of an elastic body to deflection or bending by an applied force.
In solid mechanics, a material behaves elastically if it changes shape due to an applied load, and that when the load is removed, recovers its original shape. According to solid mechanics theory, *every* material will change shape when loads are applied to it (even very small loads). Furthermore, every material will undergo elastic deformation as long as the loads are kept under a certain limit. This limit is known as the yield strength (see Tensile strength and other subjects about elasticity) of the material, and is one way of defining its strength.
The elasticity of a solid is inversely proportional to its stiffness. Stiffness, when corrected for the dimensions of the solid, becomes modulus of elasticity *E*, an intrinsic material property (see Elasticity 1: elastic or Young's modulus). Considering *E*, stiffness should be interpreted as the resistance against linear strain, not as stiffness against bending. However, the latter is directly related to *E* and of principle importance in daily engineering, but also in bio-engineering and bio-mechanics.

The bending-stiffness *k* (N/m) of a body that deflects a distance δ under an applied force F (Fig. 1) is:

$$k = F/\delta \qquad (1)$$

The elastic deflection δ (m) and angle of deflection φ (radians) in the example of Fig. 1, a cantilever (single sided clamped) beam with width B and height H is:

$$\delta = F \cdot L^3 / (3 \cdot E \cdot I) \qquad (2a)$$

$$\varphi = F \cdot L^2 / (2 \cdot E \cdot I) \qquad (2b)$$

where
F = force acting on the tip of the beam
L = length of the beam (span)
E = modulus of elasticity

$$I = \text{area of momentum} = B \cdot H^3/12 \qquad (3)$$

where H is height and B width.

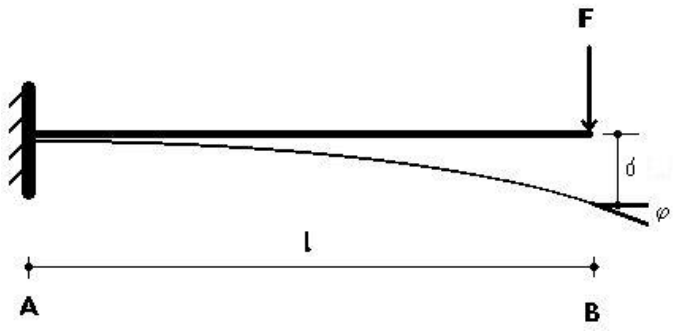


Fig. 1 Cantilever beam with load at the tip.

From (2a) it follows that the ratio L/H is the most determining factor. Doubling L with the same H and B increases deflection 8 fold. For given L, F and *E*, the extend of bending is only dependent on *I*, that is given by the cross sectional size and shape. With the momentums of a circular rod versus a square beam and a pipe versus a square box girder, with the conditions that the outer cross sectional area's are equal (on the basis of anatomical space limitations) and the cross sectional material area's are also the same (equal ''investment" of grow and "maintenance"), it can be proved that irrespective wall thickness the ratio $\delta_{pipe}/\ \delta_{box\ girder} = 12/4\pi = 0.955$. This small draw back of the tubular shape of bones is probably compensated by a more efficient development of round structures versus structures with sharp edges.

Application

In the field of mechanics of biomaterials, stiffness is mainly of importance for bones.
Equation (1) implies that a high *E* is sought when deflections are undesirable (orthopedic prostheses, like bars, plates, joints), while a relatively low *E* is required when (some) flexibility is needed (flexible orthopedic prostheses, stents).
Tubular bones can generally be considered as round tubes when the bone tissue in the medulla is ignored. (Actually, the mesh of small bone-beams of the medulla considerably increases stiffness of the tubular bone). The middle section (tubular structure) of the human femur and tibia have a medulla/total diameter ratio of ca. 0.55. That of the humerus is ca. 0.6.
Using the same amount of bone material, tubular bending stiffness increases with a factor $(2k^2-1)$ where k is the factor of increase of the outer diameter with respect to the diameter of the massive rod. For the femur and tibia this means an increase in stiffness of some 67% compared to the rod design (k=1.15). Wall thickness can also be expressed in k: $0.5k-0.5(k^2-1)^{0.5}$. However, the increase in outer diameter is not unlimited. Other mechanical (and anatomical) properties and constraints determine an optimal diameter.
The diaphysis of the human femur has a practically circular outer diameter, more than any other human tubular bone. However, the medulla is less cylindrical than the outer diameter with a larger width in the antero-posterior direction (59% of diameter) than in the medio-lateral direction (46%). The reversed may be expected since in the anterior-posterior direction a femur should be stiffer than in lateral direction. Forces act mainly in the median plane, not in the medio-lateral (coronal) plane. However, in the medial plane limb stiffness (with the knee as a hinge-joint) is also provided by tendons and muscles, whereas in the coronal plane this contribution is much less. The cross section of the humerus is more concentric, yielding more equal stiffness in both directions. The forces acting on the humerus are probably also more equal in the different planes than for the femur.

More Info

The above example of the femur can be approximated by a square box girder with a rectangular inner shape as depicted in Fig. 2. Its area moment is $(BH^3 - bh^3)/12$. Starting from a square inner shape material is taking away from the anterior and posterior walls and used for an increase of wall thickness of the lateral sides, such that the total material spent is the same. Normalizing the moment of the square profile at 1, and calculating the moment for the dashed inner shape, the area moment for the anterior/posterior and lateral direction is (1-0.81 x 0.36) and (1-0.16 x 0.36) respectively. The latter is clearly larger (33%), consequently bending in the medial plane is larger (see (1)).
In addition to stiffness, bending rupture and yield strength are of importance. Taking also the bone density of the medulla into account, making calculations more realistic, the femur seems to be more designed for high yield strength (see Tensile strength) and to withstand bending fracture more than stiffness.
The above makes clear that even for parts of extremities with static conditions biomechanics is already complicated. To understand static mechanics better we need the *concept of the neutral line*, which is beyond the scope of this compendium.

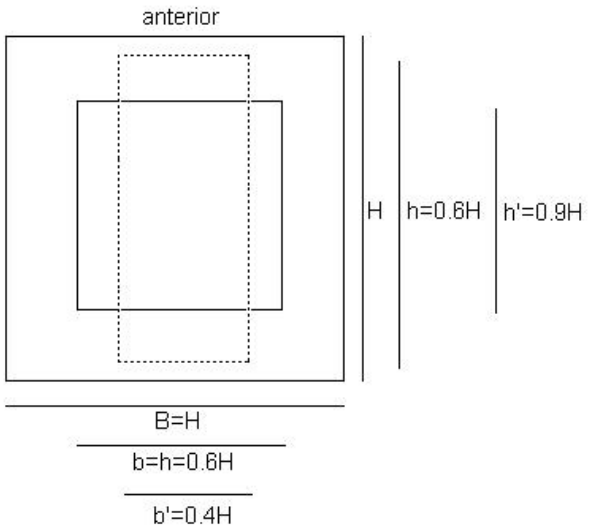


Fig. 2 Basic approximation of a non-concentric tubular bone cross section. The box girder with dashed inner circumference has the same material cross section as the square one.

As both F and δ are vectors, in general their relationship is characterized by a stiffness matrix for the 3 dimensions. The displacement can, in general, refer to a point distinct from that where the force is applied and a complicated structure will not deflect purely in the same direction as an applied force. The stiffness matrix enables such systems to be characterized in straightforward terms. For a mathematical treatise see http://www.efunda.com/formulae/solid_mechanics/mat_mechanics/hooke_orthotropic.cfm. Modern engineering design software can calculate stiffness, bending strength etc. for complicated 3-D (biological) structures.

Literature

Évinger S, Suhai B, Bernáth B, Gerics B, Ildikó Pap I, and Gábor Horváth G. How does the relative wall thickness of human femora follow the biomechanical optima? An experimental study on mummies. J Exp Biol 208, 899-905 (2005)
Drapeau MSM and Streete MA. Modeling and remodeling responses to normal loading in the human lower limb, Am J Physic Anthropol 129:403–409 (2006)
Elizabeth Weiss, Humeral Cross-Sectional Morphology From 18th Century Quebec Prisoners of War: Limits to Activity Reconstruction, Am J Physic Anthropol, 126:311–317 (2005).

Tensile strength

Principle

Tensile stress or tension is the stress state leading to expansion; that is, the length of a material tends to increase in the tensile direction. The volume of the material stays more or less constant. Therefore in a uni-axial material the length increases in the tensile stress direction and the other two directions will (with most materials) decrease in size. The par-axial decrease is given by the Poisson's ratio μ. Tensile stress is the opposite of compressive stress. Objects designed to resist tensile stress are ropes, nails, bolts etc. Beams or rod subjected to bending have tensile stress as well as compressive stress (see Elasticity 3: compressibility and bulk modulus) and/or shear stress (Elasticity 2: shear modulus). Tensile stress may be increased until the reach of tensile strength, namely the *limit state* of stress. This is the maximum amount of tensile stress that it can be subjected to before failure. The definition of failure can vary according to material type and design methodology. There are three typical definitions of tensile strength:

- Yield Strength - The stress a material can withstand without permanent deformation. This is the linear part of the stress-strain curve.
- Ultimate Strength - The maximum stress a material can withstand. This is generally higher than the breaking strength.
- Breaking Strength - The stress value on the stress-strain curve at the point of rupture.

The various definitions of tensile strength are shown in Fig. 1 and Fig. 1 of Elasticity and Hooke's law.

The table gives the some yield strengths and ultimate strength of some materials used in prostheses and some biomaterials:

Material	Yield strength (MPa)	Ultimate strength (Mpa)	Density (g/cm^3)
Polypropylene	12-43	19.7-80	0.89-0.93
Stainless steel AISI 302	520	860	7.9
Titanium Alloy (6% Al, 4% V)	830	900	4.51
Nylon, type 6/6	45	75	1.14
Kevlar		3600	1.4
ZYLON (PBO)		5800	1.56
Rubber		15	1.1-1.2
Spider silk	1150 (?)	1200	1.3
Pine Wood (parallel to grain)	11	40	0.61
Bone (limb)	17	130	1.8
Patella tendon		60	0.8 – 2.7
Dentine	17		2.1-3
Enamal	67		

Many metals have a linear stress-strain relationship up to a sharply defined yield point. For stresses below this yield strength all deformation is recoverable, and the material will relax into its initial shape

when the load is removed. For stresses above the yield point, a portion of the deformation is not recoverable, and the material will not relax into its initial shape. This unrecoverable deformation is known as plastic deformation. Sometimes, these plastic deformations can be large without fracture (in metals such as being drawn into a wire): the material is ductile. For many technical applications plastic deformation is unacceptable, and the yield strength is used as the design limitation. Brittle materials such as concrete, carbon fiber and bone (more or less) do not have a yield point at all. They behave very different from ductile materials. A stress-strain curve for a typical brittle material is shown in Fig. 1.

Applications

Material properties of artificial materials of human prostheses (orthopedic, vascular, artificial heart valves, wires etc.) are of ultimate importance since they determine the risk of a defect and the interval of displacement. For cardiovascular applications they are crucial. Much research is done to improve their mechanics.

The stiffness of compact bone tissue to prevent extension depends on the bone from which it is taken. Fibular bone has a Young's modulus *E* (see Elasticity 1: elastic or Young's modulus) about 18% greater, and tibial bone about 7% greater than that of femoral bone. The differences are associated with differences in the histology of the bone tissue. In biomaterials, tensile strength is different for the different directions. Bone is elastically anisotropic, i.e. its properties depend on direction. Such behavior is unlike that of steel, aluminum and most plastics, but is similar to that of wood. For example for the human femur, in longitudinal direction *E* is 135 MPa, but transversally only 53 MPa.
Severe tendon injuries nearly always result in chronic lengthening. The failure strain of tendons is about 10% of their rest length.

More info

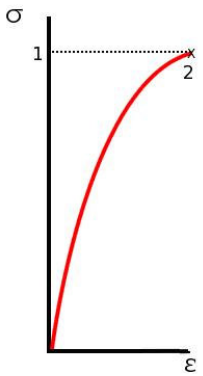


Fig. 1 Stress-strain curve typical of a brittle material (e.g. bone). 1. ultimate strength, 2. rupture.

Tensile strength is measured in Pa (≡ N/m^2), The breaking strength of a rope or tendon is specified in Newton without specifying the cross-sectional area of the rope (e.g. supraspinal tendon 900 N). This is often loosely called tensile strength, but this not a strictly correct use of the term.
In brittle materials such as rock, concrete, cast iron, tensile strength is negligible compared to the compressive strength and it is assumed zero for most engineering applications.
Tensile strength can be measured for liquids as well as solids. For example, when a tree draws water from its roots to its upper leaves by transpiration the column of water is pulled upwards from the top by capillary action and this force is transmitted down the column by its tensile strength.

Single-walled carbon nanotubes made in academic labs have the highest tensile strength of any material yet measured, with labs producing nanotubes with a tensile strength of 63 GPa (63,000 MPa) well below its theoretical tensile strength of 300 GPa (300,000 MPa). As of 2004, however, no macroscopic object constructed using a nanotube-based material has had a tensile strength remotely approaching this figure, or substantially exceeding that of high-strength materials like Kevlar.

A parameter derived from the stress-strain curve is toughness, the resistance to fracture of a material when stressed. It is defined as the amount of energy that a material can absorb before rupturing, and can be found by finding the area (i.e., by taking the integral) underneath the stress-strain curve. It is the same as $-0.5k\Delta L^2$ where k is the spring constant and ΔL the strain (see Elasticity and Hooke's law). Toughness, often expressed as the *Modulus of Toughness*, is measured in J/m^3.

Literature
http://silver.neep.wisc.edu/~lakes/BoneAniso.html

Torsion

Principle

Torsion occurs when an object is twisted or screwed around its axis. Torsion can be the result of an applied torque. It is a kind of shear stress (see Elasticity 2: shear modulus). For circular sections, the shearing stress at a point on a transverse plane is always perpendicular to the radius at that point.
The torsion coefficient is a property of torsion springs. It is the torque required to twist a body through an angle of one radian (1 rad = 360/π) and is usually denoted by *K*. Therefore it is given as

$$K = \tau/\theta, \quad (\text{Nm/rad}) \tag{1}$$

where τ is the torque and θ is the angle in radians. This equation is analogous to the spring constant given by Hooke's law (see Elasticity and Hooke's law) only used for linear strain along an axis.
Torque can be thought of informally as "rotational force". The force applied to a lever, multiplied by its distance from the lever's fulcrum (pivot point), is the torque. Torque, τ, also called moment or couple, is commonly measured in Nm. More generally, one may define torque as the *vector product*:

$$\tau = r \times F, \tag{2}$$

where F is the force vector and r is the vector from the axis of rotation to the point on which the force is acting. This assumes the force is in a direction at right angles to the straight lever.
Since τ is a vector, it cannot be expressed in Joule, a scalar unit. The rotational analogues of force, mass, and acceleration are torque, moment of inertia and angular acceleration respectively. For example, a force of 3 N applied 2 m from the fulcrum exerts the same torque as 1 N applied 6 m from the fulcrum.
The joule is also defined as 1 Nm, but this unit is not used for torque. Since energy can be thought of as the result of "force times distance", energy is always a scalar whereas torque is the vector product "force cross distance". A torque τ of 1 Nm applied through a full revolution will require an energy E of exactly 2πJ (from $E = \tau \cdot \theta = 1 \times 2\pi$ J).

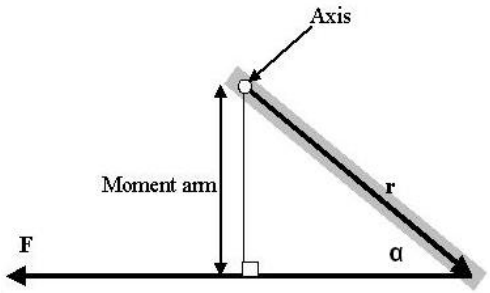


Fig. 1 Moment-arm diagram

If a force of magnitude F is at an angle α from the displacement arm of length *r* (and within the plane perpendicular to the rotation axis, see Fig. 1), then from the definition of cross product, the magnitude of the torque arising is:

$$\tau = r\sin(\alpha) \times F. \tag{3}$$

Application

Torque is applied in the models about axial rotation in joint and about the saccadic eye movements.
A simplified model of eye rotation is that of Westheimer, which comprises inertia J due to the mass of the eye ball, friction B due to the ball/orbita friction, and stiffness K due to the muscles and tendons. Together they yield a second order linear system (see More Info and Linear systems). The system is visualized in Fig. 2.
Other important applications are the torque calculations in a hip and shoulder joint, specifically in the field orthopedic revalidation, orthopedics of elderly, and sports medicine.

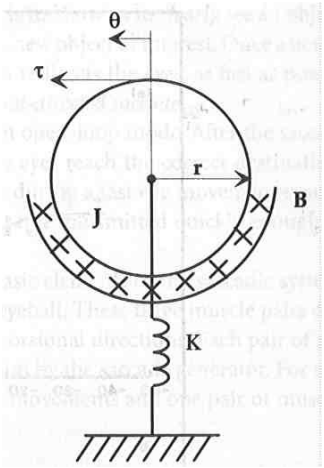


Fig. 2 The eyeball position is given by angle Θ. The stiffness is visualized by the symbol of a spring.

More info

Static equilibrium
For an object to be in static equilibrium, not only must the sum of the forces be zero, but also the sum of the torques (moments) about any point. So, for the forces F, the sum of the forces requirement is $\Sigma F = 0$ and, and for the torque $\Sigma\tau = 0$. That is, to solve statically determinate equilibrium problems in 3-D we need 2 times 3 (dimensions) is 6 equations.

Torque as a function of time
Torque is the time derivative of angular momentum (L), just as force is the time derivative of mass times velocity. For multiple torques acting simultaneously:

$$\Sigma\tau = dL/dt \quad (1a)$$

$$L = \omega \cdot I = 2\pi f m r^2, \quad (1b)$$

where ω is angular velocity, *I* moment of inertia, f rotation frequency (rotations/s) and m mass. So, if *I* is constant,

$$T = I \cdot d\omega/dt = I\alpha \quad (2)$$

where α is angular acceleration, a quantity usually measured in rad/s^2.

Relationship between torque and power
If a force is allowed to act through a distance, it is doing mechanical work. Similarly, if torque is allowed to act through a rotational distance, it is doing work. Power is the work per unit time. However, time and rotational distance are related by the angular speed where each revolution results in the circumference of the circle being traveled by the force that is generating the torque. This means that torque that is causing the angular speed v_{ang} is doing work and the generated power may be calculated as:

$$P = \tau \cdot v_{ang}. \quad (3)$$

Mathematically, the equation may be rearranged to compute torque for a given power output. However, in practice there is no direct way to measure power whereas torque and angular speed can be measured directly.
Consistent units must be used. For metric SI units torque is in Nm and v_{ang} in rad/s (not revolutions per second). For different units of power, torque, or angular speed, a conversion factor must be inserted into the equation. For example, if v_{ang} in is measured in revolutions instead of radians, a conversion factor of 2π must be added because there are 2π radians in a revolution:

$$P = \tau \cdot 2\pi \cdot n_{rev}, \quad (4)$$

where n_{rev}, the rotational speed, is in revolutions per unit time.
For a rotating object, the *linear distance* covered at the circumference in a radian of rotation is the product of the radius with the angular speed. That is: linear speed = radius x angular speed. By definition, linear distance = $v \cdot t = r \cdot v_{ang} \cdot t$.

By substituting F = r/τ (from the definition of torque) into the definition of power, being P = F times linear distance/t , the power can be calculated:

$P = (\tau/r)\,(r \cdot n_{ang}) \cdot t)/t = \tau \cdot n_{ang}$. (5)

If the rotational speed is measured in revolutions per unit of time, the linear speed and distance are increased proportionately by 2π in the above derivation to give (4).

The model of rotatiuon of the eye ball

This model can be written as:

$$\tau(t) = Jd\theta/dt^2 + Bd\theta/dt^2 + kd\theta/dt^2 \quad (6).$$

The solution in Laplace notation is:

$$H(s) = \frac{\omega_n^2 / K}{s^2 + 2\zeta\omega_n s + \omega_n^2}, \quad (7)$$

where ω_n = 120 rad/s and ζ = 0.7. Consequently, the system is well damped.

Literature

Enderle J.D. The fast eye movement control system. In: The biomedical engineering handbook, Bronzino (ed), CRC Press, Boca Raton, 1995, pp 2473-2493.

Transport

Bernoulli's and Pascal's Law

Principle

The hydrostatic pressure in a tube is visualized in Fig. 1 and is given by Pascal's Law:

$$P_1 + \rho g h_1 = P_2 + \rho g h_2 = \text{constant}, \quad (1)$$

where
ρ = fluid density (kg/m^3);
g = acceleration due to gravity on Earth (m/s^2);
h = height from an arbitrary point in the direction of gravity (m).
P = pressure somewhere imposed in the tube (N/m^2, Pa).
The second term is due to gravity. Fig. 1 illustrates the hydrostatic condition in the tube.

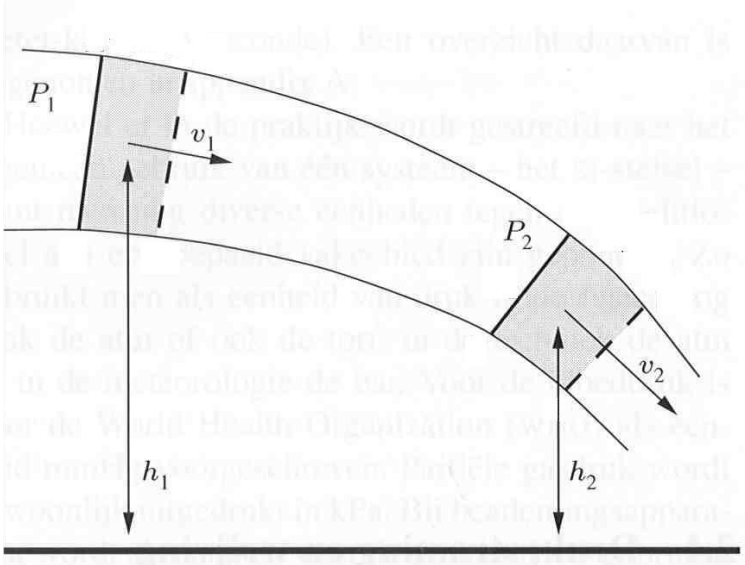


Fig. 1 Visualizing of Pascal's law and Bernoulli's equation in a tube.

By adding an inertia term, ½ρv^2, the kinetic energy per unit mass of the moving liquid with velocity v, Pascal's Law evolves to Bernoulli's equation:

$$\tfrac{1}{2}\rho v^2 + \rho g h + P = \text{constant} \quad (2)$$

Resuming, in fluid dynamics, Bernoulli's equation, describes the behavior of a fluid (gases included) moving in the tube. More precisely, the flow is along a streamline (the line indicating the trajectory of a fluid particle). There are typically two different formulations of the equations; one applies to incompressible fluids and (2) holds. The other applies to compressible fluids (see **More Info**).

These assumptions must be met for the equation to apply:
- inviscid flow viscosity (internal friction) = 0;
- steady flow;
- incompressible flow, *ρ* is constant along a streamline. However, *ρ* may vary from streamline to the streamline;

For constant-density steady flow, it applies throughout the entire flow field. Otherwise, the equation applies along a streamline.
A decrease in pressure occurs simultaneously with an increase in velocity, as predicted by the equation, is often called Bernoulli's principle.
Bernoulli's equation (2) under the above conditions and considering one dimension is a special, simple case of the Navier-Stokes equations.

Applications

In devices and instruments to calculate pressure in gas and liquid flows. For instance, in science, the Bunsen burner, the water-jet pump (venturi-principle), the Pitot tube. In the (path)physiology of the vascular and airways system to calculate and measure pressure, for instance before and after a stenosis (see Flow through a stenosis). For flow in the vascular and airways system many refinements are needed as those of **More Info**, and even more than that. Two approaches are feasible, a semi-analytic approach with applying practical 'rules' of technical liquid dynamics or a numerical approach with mathematical tools like the finite element approach.

More info

A second, more general form of Bernoulli's equation may be written for compressible fluids, in which case, following a streamline, we have:

$$v^2/2 + \varphi + w = \text{constant} \qquad (3)$$

Here, φ is the gravitational potential energy per unit mass, which is just φ = gh in the case of a uniform gravitational field, and w is the fluid enthalpy per unit mass, which is also often written as h. Further it holds that:

$$w = \varepsilon + P/\rho. \qquad (4)$$

ε is the fluid thermodynamic energy per unit mass, also known as the specific internal energy or "sie". With constant P, the term on the right hand side is often called the Bernoulli constant and denoted *b*. For steady inviscid flow (see Gas Laws) with no additional sources or sinks of energy, *b* is constant along any given streamline. Even more generally when *b* may vary along streamlines, it still proves a useful parameter, related to the "head" of the fluid.
Until now internal friction is considered to be zero. However, this is generally not true. Therefore, classically, the pressure drop over a tube (volume flow Q, length L and diameter d) with incompressible and with compressible flow is described with:

$$\Delta P = k_1 Q + k_2 Q^2 \qquad (5)$$

The first term applies for the laminar, the viscous component of the resistance and the 2nd for the turbulent part. Both are not implied in (2) and (3). The laminar part comprises the friction within the tube, and extra friction introduced by irregularities like non-smoothness of the tube wall, constrictions (see Flow through a stenosis), bends (see Flow in a bended tube) and bifurcations (see Flow in bifurcations). The same irregularities can also contribute to the second term when the flow through the irregularities is turbulent on basis of the Reynolds number . For transitional flow both terms contribute.
The influence of non-smoothness of the tube wall is expressed in a resistance factor λ. For Re<2300 it is irrelevant (always laminar). For 4000<Re<10^5 the pressure drop ΔP is given by the Darcy-Weisbach equation:

$$\Delta P = 0.5\lambda \cdot \rho \cdot L \cdot v^2/d = 8 \cdot \pi^{-2} \cdot \lambda \cdot \rho \cdot L \cdot d^{-5} \cdot Q^2, \qquad (6)$$

where
$\lambda = 0.316/Re^{0.25}$ (according to the Blasius boundary layer).
For gas flow in a smooth tube there is an expression covering the various types of flow (ref. 1):

$$\Delta P_{tube,gas} = 8 \cdot \pi^{-7/4} \cdot \eta^{1/4} \cdot \rho^{3/4} \cdot L \cdot d^{-19/4} \cdot Q^{7/4}, \qquad (7)$$

where η is the dynamic gas viscosity (Pa·s).

Literature

1. Clarke J.R. and Flook V. Respiratory function at depth, in: The Lung at Depth, Lundgren C.E.G. and Miller J.N. (Ed), Marcel Dekker Inc., New York, Basel, pp 1-71,1999.
2. Van Oosterom, A and Oostendorp, T.F. Medische Fysica, 2nd edition, Elsevier gezondheidszorg, Maarssen, 2001.
3. Wikipedia

Blood flow

Principle

Blood flow in arteries can be characterized by the combination of three key phenomena:
internal laminar (or sometimes turbulent) flow, with and without flow irregularities due to a stenosis, a curvature or a bifurcation;
pulsatile flow, diminishing from aorta to capillaries.
compliant arterial wall boundaries;
Each of them has a dramatic effect on the flow pattern.

More Info

For aorta flow and geometry the entrance effect due to the aortic valve (see Flow: entrance effect and entrance Length), characterized by the entry length, is at least 150 cm, which is far greater than the length of the aorta. The flow in the aorta thus cannot be characterized as fully developed, i.e. laminar Poiseuille flow (see Poiseuille's Law). In addition to the entrance effect the proximal aorta is bended. The bend causes extra resistance and therefore pressure drop (see Flow in a bended tube).
In the bend, there is strong asymmetry due to the centripetal forces at the outer curvature. Branching causes also entry phenomena such as asymmetries in the velocity patterns (see Flow in bifurcations). Also complicated secondary flows perpendicular at the axial flow direction may occur, and even flow separation, all of which are far more difficult to analyze than simple steady-state fully developed Poiseuille flow. Secondary flows in curvatures and bifurcations are characterized by a swirling, helical component superimposed on the main streamwise velocity along the tube axis. In fact, all the larger arteries of the circulatory system, including the epicardial coronary vessels are subject to entrance effects.
In large vessels the inertia character overwhelms the viscid character. The inertance with the dimension of mass can be expressed as $\mathbf{L} = \rho \cdot l/A$, where ρ is the density, l the tube length and A the wetted tube area. Then the pressure drop ΔP over distance l is equal to $\Delta P = \mathbf{L} \cdot \dot{V}$ with $\dot{V}$ the volume flow (L/s).
Furthermore, flow in the larger arteries is, in general, not Poiseuille flow due to the pulsatile character, especially in the aorta. In the ascending aorta of large mammals viscous effects of the entrance region are confined to a thin-walled boundary layer. The core is characterized as *largely inviscid*, caused by the heavy pulsatile character. These factors, together with specific density and viscosity are comprised in the Womersley number, which can be considered as the pulsatile version of the Reynolds number. With high numbers inertia dominates, yielding a rather well flat flow front. With low numbers viscosity dominates, yielding parabolic-like flows, however skewed towards the outer wall. An example for this is the flow in the left common coronary artery. In other coronary arteries the Reynolds numbers are much lower, the viscous effects are more dominant and flow is laminar. The velocity profile in many regions will be more like a parabolic Poiseuille flow, except that there will be skewing of this profile due to vessel curvature and branching. Also, significant entrance effects may result in the blunting of the velocity profiles.

Although there have been numerous fluid-dynamic studies of secondary flow phenomena, instrumentation limitations have prevented in vivo observations.
An additional complication introduced by the geometry of the arterial system is flow separation from and reattachment to the wall, causing recirculation zones. This phenomena in pulsatile flows is extremely complicated.

Literature

Author N. Westerhof, Noble MIM and Stergiopulos N. Snapshots of hemodynamics: an aid for clinical research and graduate education, 2004, Springer Verlag.
http://mss02.isunet.edu/Students/Balint/bloodflow.html.

Blood pressure

See for a description, the methods of measurement, models and special issues of blood pressure the chapters:

Blood pressure: (Central) venous
Blood pressure: description and measurement
Blood pressure: models
Blood pressure: pulse pressure
Blood pressure: Windkessel model

Blood pressure: models

Principle

Blood pressure (BP) is the pressure exerted by the blood at right angles to the walls of the blood vessels P_i minus the environmental or ambient pressure P_a, so:

$$BP = P_i - P_a. \quad (1)$$

Unless indicated otherwise, BP refers to systemic arterial BP, i.e. the pressure in the large arteries delivering blood to body parts other than the lungs, such as the brachial artery (in the arm). The pressure of the blood in other vessels is lower than the arterial pressure. BP values are generally stated in mmHg, but can be converted to an SI-unit, i.e. in Pascals. The conversion is:

$P = \rho_{Hg} g h/1000$, (Pa)

where ρ_{Hg} is the specific density of Hg (kg/m^3), g the gravitational acceleration (m/s^2) and h the value of BP in mmHg. Hence, P = 0.1328·BP (kPa).

Mean arterial pressure (MAP)
The mean arterial pressure (MAP) is defined as the average arterial pressure during a single cardiac cycle. It is a result of the heart pumping blood from the veins back into the arteries. The up and down fluctuation of the arterial BP results from the pulsatile nature of the cardiac output (see Blood pressure: pulse pressure). The pulse pressure is determined by the interaction of the stroke volume versus the volume and elasticity of the major arteries.
MAP can be calculated by:

MAP = (stroke volume x systemic resistance) + CVP,

where CVP is central venous pressure (see Blood pressure: (Central) venous). Mostly, CVP can be neglected. The first term at the right is the heamodynamic analog of Ohm's law for an electric circuit (V=iR). Cardiac output represents the efficiency with which the heart circulates blood throughout the body. Since MAP stroke volume and systemic resistance are not easy to measure, a simple equation of approximation has been developed:

$$MAP = P_{dias} + (P_{sys} - P_{dias})/3. \quad (2)$$

It shows that MAP is nearer the level of diastolic than systolic pressure.

Factors influencing BP
The physics of the circulatory system, as of any fluid system, are very complicated. Many physical and physiological factors influence BP. Some physical factors are:
Heart rate The higher the heart rate, the higher BP (assuming no change in stroke volume).
Blood volume The higher the blood volume, the higher the cardiac output.
Resistance The higher the resistance, the higher the BP. Resistance is related to size (the larger the blood vessel, the lower the resistance), as well as the smoothness of the blood vessel walls. Smoothness is reduced by the buildup of fatty deposits on the arterial walls (arteriosclerosis). Deposits can affect the laminar character of the flow (see Poiseuille's law and Reynolds number). Various substances (vasoconstrictors and vasodilators) change vessel diameter, thereby changing BP.

Viscosity or thickness of the fluid Increase of viscosity results in increase of resistance and so of BP. Anemia reduces viscosity, whereas hyperemia increases viscosity. Viscosity also increases with blood sugar concentration.

More Info

Usually, the systolic pressure P_{sys} amounts to 120 mmHg, or about 16 kPa. At this P_{sys} and an air pressure P_a of 1 atm, about 100 kPa, the total, absolute pressure in the blood vessel P_i is:

$$P_i = P_{sys} + P_a = 116 \text{ kPa}. \quad (3)$$

16 kPa is equivalent to the pressure of a column of water with a height of 1.63 m.

The flow speed of blood in the body is 10 up to 100 cm/s (respectively during diastole and systole) in the aorta, approximately 10 cm/s in the arteries and approximately 0.1 cm/s in capillares. According to the law of Bernoulli (see Bernoulli's and Pascal's Law) it holds that:

$$P_i + \rho gh + \tfrac{1}{2}\rho v^2 = c, \quad (4)$$

where c is a constant. This means that: $P_{sys} + \rho gh + \tfrac{1}{2}\rho v^2 = c - P_a$. Since the ambient pressure is the same in all tissues, it holds that:

$$P_{sys} + \rho gh + \tfrac{1}{2}\rho v^2 = \text{constant}. \quad (5)$$

At a flow speed v = 50 cm/s (the largest flow speeds in the blood vessels are of this order of size) it holds that $\tfrac{1}{2}\rho v^2 = 0.13$ kPa. Since P_{sys} is 16 kPa, the term $\tfrac{1}{2}\rho v^2$ in the law of Bernoulli, which we may call the flow-pressure, is negligible with respect to the blood pressure P_{sys}. However, the term ρgh is not negligible as is explained in Blood pressure: body posture.

See also:
Blood pressure: (Central) venous
Blood pressure: description and measurement
Blood pressure: body posture
Blood pressure: pulse pressure
Blood pressure: Windkessel model

Blood pressure: (central) venous

Venous pressure
Venous pressure is the blood pressure (BP) in a vein or in the atria of the heart. It is much less than arterial BP, with common values of 5 mmHg in the right atrium and 8 mmHg in the left atrium. Measurement of pressures in the venous system and the pulmonary vessels plays an important role in intensive care medicine but requires invasive techniques.

Central venous pressure
Central venous pressure (CVP) describes the pressure of blood in the thoracic vena cava, near the right atrium. CVP reflects the amount of blood returning to the heart and the ability of the heart to pump the blood into the arterial system. It is a good approximation of right atrial pressure, which is a major determinant of right ventricular end diastolic volume (right ventricular preload).
CVP can be measured by connecting the patient's central venous catheter to a special infusion set which is connected to a small diameter water column. If the water column is calibrated properly the height of the column indicates the CVP.

Blood pressure: description and measurement

Principle

Blood pressure (BP) is the pressure exerted by the blood at right angles to the walls of the blood vessels minus the environmental pressure, see Blood pressure: models.
Systolic and diastolic adult BP of the brachial artery are typically 120 (16 kPa) and 80 mmHg (10 kPa) respectively. In addition to systolic and diastolic BP, the *mean arterial pressure* (MAP), see Blood pressure: models and *pulse pressure* (see Blood pressure: pulse pressure) are other important quantities.
BP varies from one heartbeat to another and throughout the day (in a circadian rhythm); they also change in response to stress (exercise etc.), nutritional factors, drugs, or disease.

Measurement
Invasive measurement
Arterial BP is most accurately measured *invasively* by placing a cannula into a blood vessel and connecting it to an electronic pressure transducer. This is done in human and veterinary intensive care medicine, anesthesiology, and for research purposes.
Non-invasive measurement
The non-invasive auscultatory and oscillometric measurements are simpler and quicker, have no complications, and are less unpleasant and painful, at the cost of somewhat lower accuracy and small systematic differences in numerical results. These methods actually measure the pressure of an inflated cuff at the points where it just occludes blood flow (systolic BP) or just permits unrestricted flow (diastolic BP).
The classical auscultatory method uses a stethoscope (for listening to the so-called Korotkoff sounds), a sphygmomanometer (upper arm cuff attached to a mercury or aneroid manometer).
Basic digital BP monitors are relatively inexpensive, making it easy for patients to monitor their own BP. Their accuracy can vary greatly; most have not been certified for accuracy by an approved authority. Upper arm, rather than wrist, monitors usually give readings closer to auscultatory. Some meters are automatic, with pumps to inflate the cuff without squeezing a bulb.

Fig. 1 Auscultatory method An aneroid sphygmomanometer with stethoscope

Oscillometric methods are used in the long-term measurement. The equipment is the same as for the auscultatory method, but with an electronic pressure sensor (transducer) fitted in the electronically operating cuff. The manometer is an electronic device with a numerical readout and checked periodically.
The cuff is inflated to a pressure initially in excess of the systolic BP ($BP_{systolic}$), and than reduced below $BP_{diastolic}$ over a period of about 30 s. When blood flow is nil (pressure > $BP_{systolic}$) or unimpeded (cuff pressure <$PB_{diastolic}$), cuff pressure will be essentially constant. When blood flow is present, but restricted, the cuff pressure will vary periodically in synchrony with the cyclic expansion and contraction of the brachial artery, i.e., it will oscillate. The values of $PB_{systolic}$ and $PB_{diastolic}$ are computed from the raw measurements and displayed.
Oscillometric monitors do not give entirely meaningful readings in certain “special conditions” such as arterial sclerosis, arrhythmia, preeclampsia, pulsus alternans, and pulsus paradoxus.
In practice the different methods do not give identical results; an algorithm and experimentally obtained coefficients are used to adjust the oscillometric results to give readings which match the auscultatory as well as possible. Some equipment uses computer-aided analysis of the instantaneous BP waveform to determine the systolic, mean, and diastolic points.
The term NIBP, for Non-Invasive BP, is often used to describe oscillometric monitoring equipment.

More Info

Regulation of BP
The endogenous regulation comprises the baroreceptor reflex, the renin-angiotensin system (RAS) and aldosterone release. This steroid hormone stimulates Na-retention and K-excretion by the kidneys. Since Na^+ is the main ion that determines the amount of fluid in the blood vessels by osmosis, aldosterone will increase fluid retention, and indirectly, BP.

The physics of the circulatory system, as of any fluid system, are very complicated (see e.g. Elasticity of the aorta, Navier-Stokes equations, Windkessel model and Blood pressure: models).
Many physical and physiological factors influence BP. Cardiac output is heart rate times stroke volume. It represents the efficiency with which the heart circulates blood throughout the body.

For *factors influencing BP* see Blood pressure: models.

For *Mean arterial pressure (MAP)* see Blood pressure: models.

For *Orthostatic hypotension* see Blood pressure: influence of posture.

Blood pressure: influence of posture

Principle

The larger arteries are low resistance have high flow rates that generate only small drops in pressure. For instance, with a subject in the supine position, blood traveling from the heart to the toes typically only experiences a 5-mmHg drop in mean pressure.
Sometimes BP (blood pressure) drops significantly when a patient stands up from sitting and even stronger for a reclining position. This is orthostatic hypotension; gravity reduces the rate of blood return from the body veins below the heart back to the heart, thus reducing stroke volume and cardiac output. A few seconds are needed for recovery and if too slow or inadequate, the individual will suffer reduced blood flow to the brain, resulting in dizziness and potential blackout. Increases in G-loading, such as routinely experienced “pulling Gs” by supersonic jet pilots, greatly increase this effect. Repositioning the body perpendicular to gravity largely eliminates the problem.

Application

Clinical applications are e.g. the treatment of a reduced orthostatic reflex and edema in legs and feet. In practice, BP is measured in the upper arm, hence at the elevation of the heart, to obtain a measurement not biased by a difference in elevation.

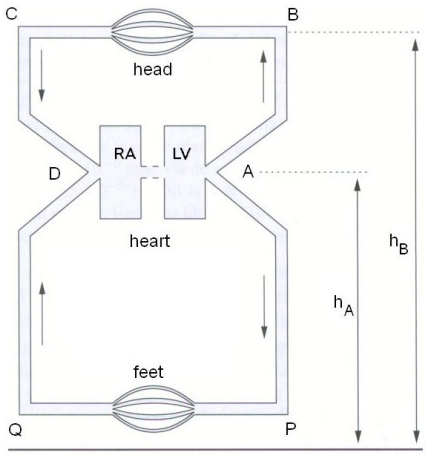


Fig. 1 Schematic represent the problem of the influence of the body posture on BP. Ra right atrium, LV left ventricle.

More Info

Fig. 1 conceptually summarizes the effect of body posture on BP. Here the lung circulation is omitted. Since the arterioles present by far the largest resistance in the circulation, the resistance (i.e. friction) in arteries can be neglected. Therefore, in the arteries with their hydrostatic differences in height, solely the law of Bernoulli (see Bernoulli's and Pascal's Law) suffices. The arterioles are short and hydrostatic pressure differences over their length can be neglected. Hence, in contrast to the arteries, for the arterioles Bernoulli's law is irrelevant whereas Poiseuille's Law is dominant.
At position D (the right atrium), blood pressure P_D is approximately 0 kPa, hence equal to ambient pressure. P_A (left ventricle) is ca. 16 kPa (the value during systole). BP at any other point is found with the laws of Bernoulli and Poiseuille. Application of the law of Bernoulli gives for blood pressures at the points A and B with up right position (consider that the term ρv^2 is negligible):

$$P_A + \rho g h_A = P_B + \rho g h_B, \quad (1)$$

with ρ is the density of blood and g the constant of gravity. Hence:

$$P_B = P_A + \rho g(h_B - h_A) = 16 - 5 = 11 \text{ kPa}, \quad (2)$$

assuming that in supine position (h_B - h_A) is approx. 50 cm. Hence, in the head with supine position, BP is 5 kPa lower, for sure a relevant amount. Since $P_C = P_D - 5$ kPa = – 5 kPa, the arterio-venous difference in the head is the same as the difference between right atrium and left ventricle.
Similar calculations, assuming that h_A = 120 cm, yield that P_P and P_Q + are 28 and 12 kPa respectively, so 17 kPa higher than in the head. This can cause several clinical disorders, e.g. edema in the feet and lower legs.

In reclining position, hydrostatic differences are nil and the arterio-venous BP-difference in the head is practically the same as in up-right position meaning that the flow is the same. Going to the up-right position, head BP lowers and as a reaction, the diameter of the arteries (with their elastic walls) diminishes, causing a decrease in flow. This effect is still reinforced because the stroke volume of the heart becomes smaller by blood pooling in the abdomen and legs. Now, the orthostatic control system regulates the diameter of the arteries, such that the flow is maintained irrespective body posture. When the posture change is to fast, e.g. from a sleeping (extra low BP and cardiac output) reclining position to a standing position (e.g. for nightly urinating) even a healthy male can show a transient dizziness.

Literature

Convertino VA. High sustained +Gz acceleration: physiological adaptation to high-G tolerance. J Gravit Physiol. 1998;5:51-4. Review.
Convertino VA. G-factor as a tool in basic research: mechanisms of orthostatic tolerance. J Gravit Physiol. 1999;73-6.

Blood pressure: pulse pressure

Principle

Formally, pulse pressure is systolic minus diastolic blood pressure. It can be calculated by:

Pulse pressure = stroke volume/compliance (Pa or mmHg).

Compliance is 1/elastance for a hollow organ, see Compliance (hollow organs).
Usually, the resting pulse pressure in healthy adults, sitting position, is about 40 mmHg. The pulse pressure increases with exercise due to increased stroke volume and reduced total peripheral resistance, up to pulse pressures of about 100 mmHg while diastolic pressure remains about the same or even drops (very aerobically athletic individuals). The latter effect further increases stroke volume and cardiac output at a lower mean arterial pressure. The diastolic drop reflects a much greater fall in total peripheral resistance of the muscle arterioles in response to the exercise (recruitment of a greater proportion of red versus white muscle tissue).

Application

Low values

If resting pulse pressure < 40 mmHg, the most common reason is an error of measurement. If it is genuinely low, e.g. 25 mmHg or less, the cause may be low stroke volume, as in congestive heart failure and/or shock. This interpretation is reinforced if the resting heart rate is relatively rapid, e.g. 100-120 mmHg, reflecting increased sympathetic nervous system activity.

High values
If the usual resting pulse pressure is consistently greater than 40 mmHg, e.g. 60 or 80 mmHg, the most likely basis is stiffness of the major arteries, aortic regurgitation (a leak in the aortic valve), an extra path for the blood to travel from the arteries to the veins, hyperthyroidism or some combination. (A chronically increased stroke volume is also a technical possibility, but very rare in practice.) Some drugs for hypertension have the side effect of increasing resting pulse pressure irreversibly. A high resting pulse pressure is harmful and tends to accelerate the normal ageing of body organs, particularly the heart, the brain and kidneys.
A high pulse pressure is an important risk factor (20% increase) for heart disease. A 10 mm Hg increase in pulse pressure increases the risk of major cardiovascular complications and mortality.

Blood Pressure: Windkessel model

Principle

The Windkessel model consists of four anatomical components: left ventricle, aortic valve, arterial vascular compartment, and peripheral flow pathway (see Fig. 1).

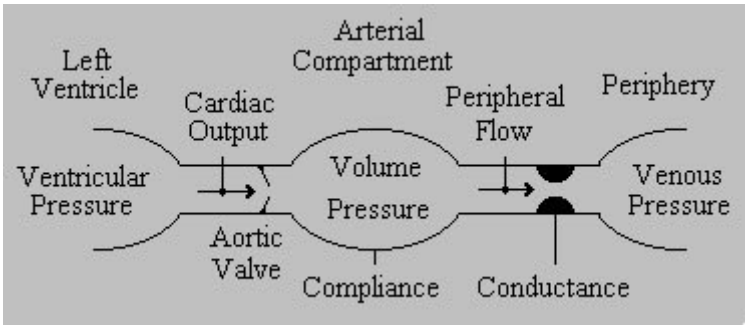


Fig. 1 Compartments of the Windkessel model.

The basic model is a closed hydraulic circuit comprised of a water pump connected to a chamber. The circuit is filled with water except for a pocket of air in the chamber. (*Windkessel* is the German word for air-chamber.) As water is pumped into the chamber, the water both compresses the air in the pocket and pushes water out of the chamber, back to the pump. The compressibility of the air in the pocket simulates the elasticity and extensibility of the major artery, as blood is pumped into it by the heart ventricle. This effect is commonly referred to as *arterial compliance*. The resistance, which the water encounters while leaving the Windkessel and flowing back to the pump, simulates the resistance to flow encountered by the blood as it flows through the arterial tree from the major arteries, to minor arteries, to arterioles, and to capillaries, due to decreasing vessel diameter. This resistance to flow is commonly referred to as *peripheral resistance*.
In terms of system analysis, the variable $I(t)$ ($mL \cdot s^{-1}$) is the flow of blood from the heart to the aorta (or pulmonary artery). It is assumed that the stroke volume of the heart is independent of the resistance and so $I(t)$ is the independent variable. The dependent variable $P(t)$ (mmHg) is the blood pressure in the aorta (or pulmonary artery). Further, the system parameter C ($mL \cdot mmHg^{-1}$) is the compliance, this is the constant ratio of air volume to air pressure (see Compliance (hollow organs))
R ($mmHg \cdot min \cdot mL^{-1}$) is the peripheral resistance or flow-pressure proportionality constant ($I/P=R$ as in Ohms law) in the systemic (or pulmonary) arterial system.

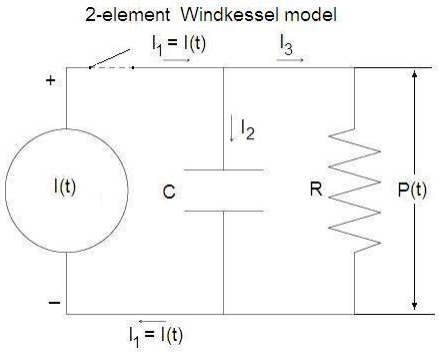


Fig. 2 Windkessel model with 2 elements.

Now, I(t) can be expressed as a function of P(t) with R and C as parameters. Assuming that C is constant and flow I through the pipes connecting the air chamber to the pump follows Poiseuille's law and is proportional to the fluid pressure, the system behaves as a first order low pass system (or filter, see Linear systems: first order). It can be described with a simple differential equation. This equation describes the relationship between the imposed current, I(t), and the resulting time-varying electrical potential, P(t), in the electrical analog depicted in Fig. 2. At the start of the diastole, at t = 0, C is loaded and has the starting pressure P_0. (In the aorta this is the systolic pressure.) During the diastole there is no blood flow from the heart, so I(t) = 0. Therefore, the source of the pressure, the heart, is disconnected (closed aorta valves), which action is presented by opening the switch in the upper left branch. The equation can be solved exactly for P(t):

$$P(t) = P_0 e^{(t\text{-}tdias)/(RC)}), \qquad (1)$$

where t_{dias} is the time at the start of diastole (for simplicity, one can take (t_{dias} = 0), and P_0 is the blood pressure in the aorta (or pulmonary artery) at the start of diastole. RC is the time constant (see Half-time and time constant), often denoted by τ (tau), which characterizes the speed of decay of P(t). After the time of 1τ, P has decreases to a fraction 1/e=0.368 of its original value (here P_0).
Equation (1) holds for a single beat. After some 6τ $P(t) \approx 0$. Since the interval between te beats is much smaller than 6τ, some pressure at the end of cardiac cycles remains. This is the diastolic pressure.

A more advanced model of is the 3-element Windkessel model, which adds a resistive element between the pump and the air-chamber to simulate resistance to blood flow due to the aortic or pulmonary valve. It has been used to model blood pressure and flow in mammalian and avian animal models. This resistance depends of the instant in the cardiac cycle. It is very large (infinite) during the diastole and has a varying magnitude during the systole. It can be modeled by a fixed resistor combined with a switch.
Here is a schematic illustration of the electrical circuit corresponding to the 3-element Windkessel model:

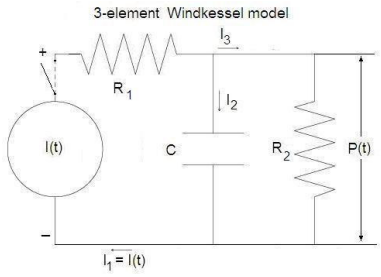


Fig. 3 Windkessel model with 3 elements.

Using the same circuit analysis technique as for the 2-element Windkessel model circuit, the differential equation for the 3-element Windkessel model becomes a little more complicated, but it still comprises only a first order time derivative, belonging to a first order differential equation. However, when we measure the aortic pressure, that is behind R_1, then we obtain the same solution is in the 2-element model, with the switch opened and R substituted by R_1.
Until now, the hydrodynamic inertia of the blood flow has been ignored. Its electric analog is an inductance L, represented by a coil. The drop in electrical potential across an inductor with self-inductance, L, is L(d I(t)/dtL.

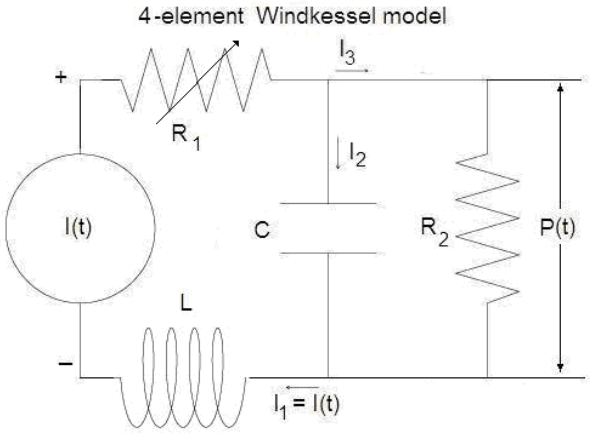


Fig. 4 Windkessel model with 4 elements.

Following the theory of circuit analysis, one finds a 2nd order differential equation (since the 2nd derivative also occurs) for this 4-element Windkessel model. R_1 is presented as a variable resistance (indicated by the arrow) to model the whole cardiac cycle.

L has units of mass per length to the fourth power. Note that for L = 0, this 4-element Windkessel equation reduces to the 3-element Windkessel equation. Also, during diastole when I(t) and its derivatives vanish, we can solve P(t) and get the same exponentially decreasing pressure function with decay time constant R_2C as in the 3-element Windkessel model.
Until know, the period of the systole was ignored. When we introduce the blood ejection during the systole, L also plays some role. For the systolic ejection one should choice a realistic time function as done in the upper panel of Fig. 5. After numerical parameterization Fig. 5 gives a result of a simulation.

Application

Windkessel models are frequently used to describe the load faced by the heart in pumping blood through the pulmonary or systemic arterial system, and the relation between blood pressure and blood flow in the aorta or the pulmonary artery. Characterizing the pulmonary or systemic arterial load on the heart in terms of the parameters that arise in Windkessel models, such as arterial compliance and peripheral resistance, is important, for example, in quantifying the effects of vasodilator or vasoconstrictor drugs. Also, a mathematical model of the relationship between blood pressure and blood flow in the aorta and pulmonary artery is useful, for example, in the development and operation of mechanical heart and heart-lung machines. If blood is not supplied by these devices within acceptable ranges of pressure and flow, the patient will not survive. The model is also applied in research with isolated hearts.

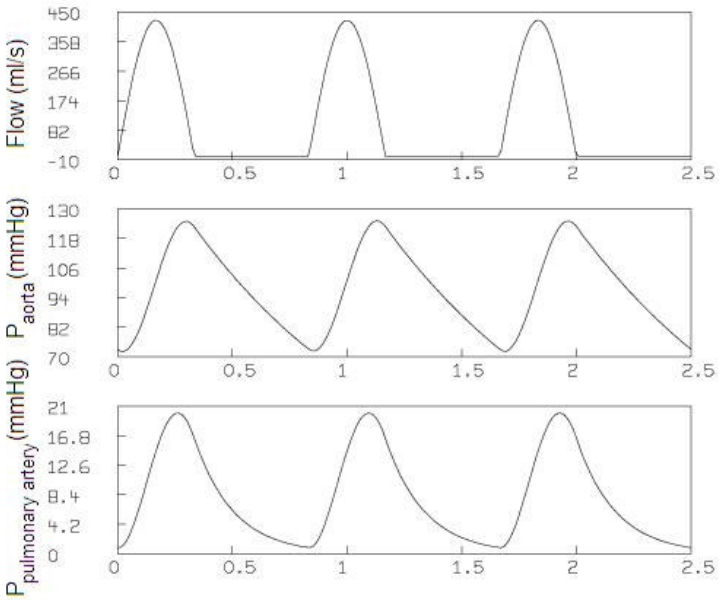


Fig. 5. Results of a simulation with the 4-element Windkessel model.

More Info

Often more complicated modifications of the Windkessel models have been applied in experimental research. For instance the resistance R_1 is often replaced by an impedance (so a compliance is added), which is time dependent.

Literature

N. Westerhof, Noble M.I.M and Stergiopulos N. Snapshots of hemodynamics: an aid for clinical research and graduate education, 2004, Springer Verlag.

Body heat conduction and Newton's Law of cooling

Principles

Heat conduction is the transmission of heat across matter. Heat transfer is always directed from a higher to a lower temperature, in accordance with he second law of thermodynamics (see Thermodynamics: second law). The donor is refrigerating and the acceptor is warming. Denser substances are usually better conductors; metals are excellent conductors.

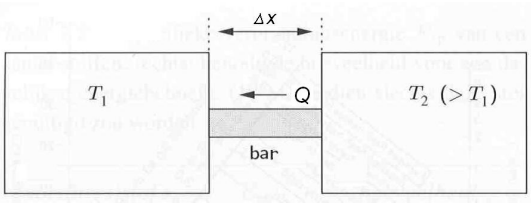


Fig. 1 Heat transfer in a bar, clamped between two solid objects with a homogeneous, constant temperature.

The law of heat conduction, also know as *Fourier's law*, states that the time rate of heat flow Q through a slab of a perfectly insulated bar, as shown in Fig.1, is proportional to the gradient of temperature difference:

$Q = k \cdot A \cdot \Delta T / \Delta x$, or more formally: (1a)
$dQ/dt = k \cdot A \cdot dT/dx$ (1b)

A is the transversal surface area, Δx is the distance of the body of matter through which the heat is passing, *k* is a conductivity constant (W/(K.m)) and dependent on the nature of the material and its temperature, and Δ*T* is the temperature difference through which the heat is being transferred. See for a simple example of a calculation **More Info**. This law forms the basis for the derivation of the heat equation. The R-value is the unit for heat resistance, the reciprocal of the conductance. Ohm's law (voltage = current·resistance) is the electrical analogue of Fourier's law.
Conductance of the object per unit area U (=1/R) is:

$U = k/\Delta x$ (2a)

and so Fourier's law can also be stated as:

$Q = U \cdot A \cdot \Delta T.$ (2b)

Heat resistances, like electrical resistance, are additive when several conducting layers lie between the hot and cool regions, because heat flow *Q* is the same for all layers, supposing that A also remains the same. Consequently, in a multilayer partition, the total resistance is:

$1/U = 1/U_1 + 1/U_2 + 1/U_2 + \ldots..$ (3)

So, when dealing with a multilayer partition, the following formula is usually used:

$$Q = \frac{A \cdot \Delta T}{\Delta x_1 / K_1 + \Delta x_2 / K_2 + \Delta x_3 / K_3 + \ldots} \quad (4)$$

where *K* should be read as *k* and $\Delta x_1 + \Delta x_2 + \ldots = 1$.
When heat is being conducted from one fluid to another through a barrier, it is sometimes important to consider the conductance of the thin film of fluid, which remains stationary next to the barrier. This thin film of fluid is difficult to quantify, its characteristics depending upon complicated conditions of turbulence and viscosity, but when dealing with thin high-conductance barriers it can sometimes be quite significant.

Application

In problems of heat conduction, often with heat generated by electromagnetic radiation and Ultrasound. Generation by radiation can be performed for instance by lasers (e.g. dermatological and eye chirurgy),

by microwave radiation, IR and UV (the latter two also cosmetic). Specific applications are thermographic imaging (see Thermography), low level laser therapy (LLLT, Photobiomodulation), thermo radiotherapy and thermo chemotherapy of cancers. Then, laws of radiation also play a role (see Wien's displacement lawand Stefan-Boltzmann law. Other fields of application are space and environmental medicine (insulating clothing), cardio-surgery with a heart-lung machine.
Heat transfer calculations for the human body have several components: internal heat transfer by conduction and by perfusion (blood circulation) and external by conduction and convection. The transport by perfusion complicates calculations substantially.

More info

Newton's law of cooling
Newton's law of cooling states that the rate of heat loss of a body is proportional to the difference in temperatures between the body and its surroundings. This form of heat loss principle, however, is not very precise; a more accurate formulation requires an analysis of heat flow based on the heat equation in an inhomogeneous medium. Nevertheless, it is easy to derive from this principle the exponential decay (or increase) of temperature of a body. If $T(t)$ is the temperature of the body as a function of time, and T_{env} the temperature of the environment then its derivative is:

$$dT(t)/dt = -\tau\,(T - T_{env}) \qquad (5)$$

where $1/\tau$ is some positive constant. Solving (5) gives:

$$T(t) = T_{env} + (T_{t=0} - T_{env})e^{-t/\tau}. \qquad (6)$$

The constant τ appears to be the time constant and the solution of (6) is the same as (de)charging a condenser in a resistance-capacitance circuit from its initial voltage $V_{t=0}$ to the final obligatory voltage V_{obl} (see Linear first order system).

In, fluid mechanics (liquids and gases) the Rayleigh number (see Rayleigh, Grashof and Prandtl Number) for a fluid is a dimensionless number associated with the heat transfer within the fluid. When the Rayleigh number is below the critical value for that fluid, heat transfer is primary in the form of conduction; when it exceeds the critical value, heat transfer is primarily in the form of convection (see Body heat dissipation and related water loss).

An example of heat loss when submerged
The problem is: how many energy passes the skin of a human jumping into a swimming pool, supposing that:
heat transfer in the body is solely determined by conduction;
the heat gradient in the body is time invariant;
the water temperature surrounding the body remains at the initial temperature;
the mean heat gradient of the body is $\Delta T/\Delta x = 1.5\ ^{o}C/cm = 150\ ^{o}C/m$
body surface 2.0 m^2 (an athletic body of 75 kg and 188 cm)
$T_{skin} = 25\ ^{o}C$;
$T_{water} = 25\ ^{o}C$;
$k_{water} = 0.6$ W/m.s
Since dQ/dt and T/dx are time and distance invariant, (1a) can directly applied:
$Q = -0.6 \cdot 2.0 \cdot 150 = 180$ W.
For a subject of 25 year basal metabolism is 43.7 W/m^2, implying that Q is about twice basal metabolism. Even this very basic approach illustrates the very high energy expenditure of for instance a swimmer, even in warm water. With $T_{water} = 13\ ^{o}C$, the skin and deeper tissues will cool down fast and soon the temperature gradient in the body is doubled, resulting in a 360 W loss. This makes clear that with normal sized people in such cool water, exhaustion, next hypothermia and finally consciousness and drowning is a process of about an hour.
Several physical multi-compartment models have been developed to calculate heat transfer in the human body (see **Literature**).

Literature
ASHRAE, Fundamental Handbook, Ch. 8 Physiological Principles, Comfort and health. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, 1989.
http://www.ibpsa.org/%5Cproceedings%5CBS1999%5CBS99_C-11.pdf

Body heat dissipation and related water loss

Principles

The type of energy, e.g. heat, motion and radiation, produced by the body (nearly) all depends on chemical processes. With a constant body mass (no grow or weight loss) and in rest nearly all energy is transformed to heat, also the small amount of mechanical energy of heart and ventilation. All heat is transferred to the environment and there are various components of this heat transfer.
Calculation of the components can be performed by applying (semi-)empirical equations (curve fitting of experimental data) or by applying physical laws, as will be done here.
Table 1 summarizes all components of heat release with their numerical values of a normal shaped man in rest and when performing heavy endurance sport (running). Some of the components are bidirectional, in other words heat can also be collected. This holds for radiation (sun bathing) and convection (as extreme example the hair dryer). The five components are dependent on biometric factors and to calculate numerically their contribution, this factors should by defined. This is done in **More Info**.

Table 1 Components of heat release

Heat release of male, 40 year, 75 kg, 175 height, body area 1.9 m^2		
Components of release (v_{air} < 0.15 m/s)	seated in rest	running 15 km/h
C_{res}, expiration of gas	1.2	41.8
E_{res}, evaporated water in expired air	9.3	332
R_{skin}, radiation	30.6	116.3
C_{skin}, convection along the skin	14.6	50.3#
E_{skin}, perspiration and evaporation via skin	21.1*	685**
Total heat release	77	1226

lower limit (see **More Info**), *sweating 16 mL/h, ** sweating 2 L/h.

Two very small components, heat conduction (air-body) and sound production (mainly heart) and the small heat conduction (via the buttocks or feet) are neglected.
The dissipated heat of the runner is 16 times that of the seated person in rest. The runner has to dissipate this heat otherwise he will suffer from hyperthermia. Without dissipation, a rough calculation yields a body temperature increase of 0.22 K/min. In rest, E_{res} and E_{skin} (for wind speed v < 0.15 m/s) together is 23.4 W, about ⅓ of the total, but during running ca. ¾. R_{skin} is 45 W, about ⅔ of the total loss of 79 Watt in rest. During running loss by radiation and convection become less important. Total water losses by evaporation and perspiration is 46 mL/h for the seated person in rest and 1342 mL/h for the runner (see More Info), 29 times more. The calculations show the enormous difference between both conditions.

Total loss in rest should be nearly the same as the basal metabolism or basal energy expenditure (BEE), which is equal to:

$$\text{male:} \quad BEE = 66.67 + 13.75W + 5H - 6.76A \ (\text{kCal/day}), \tag{1a}$$

$$\text{female:} \quad BEE = 665.1 + 9.56W + 1.85H - 4.68A, \tag{1b}$$

where H is height (cm), W is weight (kg), A is age (year). With $BEE_{Watt} = 0.0484\ BEE_{kCal/h}$, our model has an BEE of 82.4 Watt. The energy expenditure in rest and reclining is about 2% lower, but seated it is some 8% higher. All together, the calculations of heat losses and heat production are well in accordance.

The components can be calculated as follows. **More Info** gives details and the calculations of the values of Table 1.

Heat release by expiration

$$C_{res} = m \cdot c_p \cdot \Delta T, \tag{2}$$

with m the mass of gas expired per second, ΔT the temperature difference between inspired and expired gas and c_p the specific heat coefficient (the amount of energy to increase the temperature of one unit of mass of gas with 1° C under constant pressure) of the expired gas.

Heat release by expired water vapor

$$E_{res} = m_{H2O} \cdot \Delta H_{H2O}, \quad (3)$$

with m_{H2O} the mass of water vapor and ΔH_{H2O} the specific evaporation heat of water:

Heat release by radiation
The peak wavelength of the infrared radiated light by the human body is about 10 µm (calculated from the skin temperature with Wien's dispacement law) The law of Stefan-Boltzmann says that a black emitting body radiating in an infinite space with 0 K as background temperature emits $\sigma \cdot A_{body} T^4_{body}$ Watt. The constant σ is the constant of Stefan-Boltzmann, being 5.7 x 10^{-8} W/(K^4.m^2). Extending the law for grey bodies (law of Kirchhoff) and a temperature > 0 K the equation becomes:

$$R_{body} = \varepsilon_{body} \cdot \sigma \cdot A_{body} (T^4_{body} - T^4_{background}) \approx 0.5 \cdot \varepsilon_{body} \cdot \sigma \cdot A_{body} \Delta T(\Delta T + 2T_{background})^3, \quad (4)$$

where $\Delta T = T_{body} - T_{wall}$ and ε_{body} the emittance coefficient ($0.8 < \varepsilon_{body} < 1.0$, and thin-clothed 0.95).

Convection along the skin
Under many conditions, the human body releases heat to the surrounding air. (But the process can reverse, for instance by entering a warm room).The underlying mechanism is that the air particles colliding with the skin (the "wall") obtain a larger momentum (mass times velocity) at the cost of the velocity of the Brownian motion of the skin particles. And so, the air particles increase their velocity, and consequently the boundary layer of air covering the skin obtains a higher temperature. This heated layer has a lower specific density than the cooler air at a larger distance. In rest, this difference causes a laminar ascend of the boundary layer. This is the process of heat release by convection. Heat release by convection is hard to calculate and various approaches can be found in literature (see **Literature**). With laminar convection the problem is easier than with turbulence, although still complicated.
A laminar gas flow has a Rayleigh number (Ra) between 10^4 and 10^8. (see Rayleigh, Grashof and Prandtl Number). Convection currents with a velocity $v_{air} < 0.15$ m/s (generally indoor) appear to be laminar since Ra is about 1.0×10^6. Under the above conditions the refrigeration law of Newton (see Body heat conduction and Newton's Law of cooling) applies:

$$C_{skin} = \alpha \cdot A \cdot \Delta T, \quad (5)$$

where α the heat convection coefficient, A the area of the body and ΔT the difference in temperature between skin and ambient air. The parameter α is $1.35 \cdot (\Delta T/H)^{1/4}$ at 1 bar and 20 °C, with H the effective height of the subject.

Perspiration and evaporation via the skin
Perspiration is the process of water evaporating though the skin driven by the vapor pressure difference in the outer skin and the lower vapor pressure of the surrounding air. Evaporation is the process of sweat evaporation from the skin surface. The release is calculated by:

$$E_{skin} = m \cdot \Delta H_{skin\ water}, \quad (6)$$

where m is the loss of mass of liquid and ΔH the specific evaporation heat. In rest with a low skin temperature there is no sweat production, but with a high skin temperature there is some sweat production.

Application

These are in aerospace, altitude, and sports medicine, in occupational medicine for heavy exertion and extreme environmental conditions. Further air-conditioning (general, hospitals and commercial aviation) and clothing industry.
A typical application Dehydration during long flights is not caused by the low humidity in the cabin. The extra loss of water due to the humidity is only about 350 mL/day (see **More Info**). The actual reasons are a too low liquid intake by food, at all drinking to few, and acute altitude diuresis.

More info

To clarify the equations the components of heat loss will be calculated in examples. First a human heat-model (standard subject) should be specified:
- 40 years, male.
- Body weight W = 75 kg, length H = 175 cm, body area $A = 0.007184 \cdot W^{0.425} \cdot H^{0.725} = 1.90\ m^2$.

- In rest, seated (on a poorly heat-conduction seat, indoor) and running during heavy endurance sport (running, 15 km/hour).
- Sweat production 0.26 in rest and 33 mL/min when running. Thin clothing.
- RMV (minute respiratory volume of inspiration) is 5.6 L/min in rest, in sitting position. During running 200 L/min.
- $FI_{N2}/FE_{N2} = 1.06$ (N_2 fraction in inspired air/N_2 fraction in expired gas. From this ratio $RMV_{expiration}$ (= $RMV \cdot FI_{N2}/FE_{N2}$) is calculated.
- Temperature of expired air is 310 K, independent of the ambient air temperature (actually there is a small dependency).
- Temperature of the skin T_{skin} is 303 K (30 ^oC) at rest and 310 K when running.
- $\Phi_m = 88.2$ W (seated some 10% more than lying). Φ_m is total metabolic power. Φ_m is age and sex dependent (see above).
- Air velocity: indoor 0.15 m/s: outdoor, produced by the runner, 4.17 m/s.

C_{res}, *expiration of gas*

$$C_{res} = \{RMV_{exp} \cdot \rho_0 \cdot (273/T) \cdot p/60\} \cdot c_{p,air} \cdot \Delta T, \text{ with:} \quad (2a)$$

$RMV_{exp} = (FI_{N2}/FE_{N2})RMV = 1.06 \cdot RVM$ L/min. RVM is 5.6 and 200 L/min;
$\rho_0 = 1.29$ kg/m^3, the specific density of air at 273.15 K and 1 bar;
T = 310 K, the temperature T of the expired gas is;
p = 1 bar, the ambient pressure;
60, the conversion factor from minute to second;
$c_{p,air} = 1.00$ kJ·kg^{-1}K^{-1}, the specific heat capacity (at 0 ^oC);
$\Delta T = 12$ K (ambient temperature is 298 K).
After completing all values, 1.17 and 66.2 W is found for *rest* and *running* respectively.

A completely different, empirical approach (ref. 1) is :

$$C_{res} = 0.0014\ \Phi_m\ (307 - T_{ambient}), \quad (2b)$$

where Φ_m the total metabolic power.
In a commercial aircraft, C_{res} is about 0.8^2 smaller since both ρ_0 and $c_{p,air}$ are reduced with 20%.

E_{res}, *evaporated water in expired air*
Starting from $RMV_{exp} = RMV \cdot FI_{N2}/FE_{N2}$, considering the fraction of evaporated water vapor, correcting for temperature, considering seconds, the volume per second (m^3/s) is found. Via the molecular volume of 22.4 m^3 and the molecular weight of water (m_{H2O}), and ΔH_{H2O} (2260 kJ/kg), the evaporation in kg/s is found.

$$E_{res} = \{[RMV \cdot (FI_{N2}/FE_{N2}) \cdot (FE_{H2O} - FI_{H2O}) \cdot (273/310)/60]/22.4\} m_{H2O} \cdot \Delta H_{H2O} \quad (3a)$$

Between brackets the volume of water vapor in m^3/min at 273 K is found ($FE_{H2O} = 0.0618$ and $FI_{H2O} = 0.003$, only about 9% humidity, 310 K is body temperature). For the standard subject in rest E_{res} becomes 9.3 W and 332 W for the runner.
The water loss is 14.8 and 529 mL/h respectively.
There is no pressure dependency (mountaineering, diving) since the alveolar p_{H2O} is always 6.3 kPa.
Consequently, in a commercial aircraft, E_{res} is about the same supposing $FI_{H2O} = 0.003$, which means very dry air. With a normal humidity (60%), water loss is about 30% less. The example shows that water loss by expiration can be neglected in rest.

An empirical approach, modified after ref. 1, is:

$$E_{res} = 0.0173\ \Phi_m (5.87 - p_{H2O,ambient}/1000). \quad (3b)$$

R_{body}, *Radiation*
Subject in rest within a room Often, the body is surrounded at a finite distance by a wall. Whereas the body radiates in all directions to the walls, each small part of wall basically radiates to the rest of the wall and to the subject. A new parameter C is introduced to control these effects. It comprises the surface of the body and the wall, and the emittance factor of the body and the wall. Using the approximation of ref. (3) the result is:

$$R_{body} \approx C \cdot \sigma \cdot A_{body}\, \Delta T (\Delta T + 2T_{background})^3/2, \quad (4a)$$

The parameter C is defined as $1/C = 1/\varepsilon_{body} + (A_{body}/A_{wall})(1/T_{wall} - 1)$. The effective area of the sitting body $A_{s\text{-}skin}$ is 1.33 m^2, (70% of A_{skin},) and $\epsilon_{wall} = 0.9$. With a room of 70 m^3, C appears to be 5.4×10^{-8} Wm^{-2}K^{-1}. With $T_{skin} = 303$ K and $T_{wall} = 298$ K the radiant power is 29.4 W/ m^2. In a similar way the

radiant power of the chamber wall in the direction of the subject can be calculated. It amounts to 10.1 W/m^2. Since the absorption coefficient of the subject is about 0.65, the net dissipation is 29.4 – 6.6= 23.0 W/m^2. For the sitting standard subject this finally yields R_{skin} = 30.6 W. When $A_{wall}>>A_{skin}$, in the definition of C, the second term at the right can be ignored.
Subject running outdoor Equation (4) with T_{skin} = 310 K yields 116.3 W.

A simple approach for most typical indoor conditions (ref. 1) is:

$$R'_{skin} = 4.7 \cdot A \cdot \Delta T \text{ (W)}. \quad (4b)$$

This results in 31.3 W (sitting subject). For a large ΔT with a high body temperature this linear approximation is less accurate.

C_{skin} Convection along the skin
Subject in rest within a room The dependency of the heat convection coefficient α on the effective height H_e implies a dependency on body posture. For lying, sitting and standing the effective surface is some 65%, 70% ad 73% of total body area respectively, and H_e is some 17, 80 and 91% of actual height. This yields values for α of 2.7, 2.2 and 1.8. With $\Delta T = T_{skin} - T_{air} = 303 - 298 = 5$ K and completing the equation $C_{skin} = \alpha \cdot A \cdot \Delta T$ for reclining, seated and standing posture C_{skin} is 16.8, 14.6 and 12.5 W respectively, values closely together.
Subject running outdoor The runner has a ΔT of 12 K. His effective height is supposed to be 85% of the actual height and his effective surface 100%. This three changed values give an increase of a factor of 4.02 compared to standing in rest. Supposing that the factor α still applies for an air velocity of 15 km/h, than the convection loss of the runner is 50.3 W.
Air velocities v > 1 m/s give a substantial increase in C_{skin}: the factor of proportionality is $(v \cdot p)^{0.6}$, where p is the ambient pressure. From this factor the wind-chill temperature factor can be calculated. The ratio of the air velocities yields a chill ratio of 7.35. This yields 92 W. The original value of 50.3 W is too low and the latter an upper limit. The factor α needs correction, since the thermal diffusivity (see Rayleigh, Grashof and Prandtl Number) is much higher in turbulent convection (the runner) than with laminar convection. The deviation from the upper limit depends on the clothing of the runner (extend of turbulence).
Foggy air augments C_{skin}.
Heat release by laminar convection is pressure dependent: $C_{p\ bar} = p^{1/4} C_{1\ bar}$. Consequently in aircrafts and at altitude it is less and in hyperbaric chambers (for hyperbaric oxygen treatment) it increases. The dependency on pressure can be clarified conceptually and only qualitatively as follows. With a higher pressure, there is a higher density and so more collisions with the wall. This effect augments the heat release. But a higher density means that in the gas there are also more collisions, reducing the "diffusion" of the heat. Also the flow of the convection behaves different. This all results in the exponent of ¼.

E_{skin} *Perspiration and evaporation via the skin*

$$E_{skin} = m \cdot \Delta H_{skin\ water}. \quad (5)$$

The mass (m) of sweat and perspirated skin water is 8.7×10^{-6} kg/s (0.75 kg/day, supposing both contributions are the same at a skin temperature of 30 °C), and $\Delta H_{skin\ water}$ is 2428 kJ/kg (slightly higher than for pure water due to the dissolved minerals). The calculation yields E_{skin} = 21.1 W.
Supposing that 50% of the sweat production of the runner evaporates, perspiration and evaporation is $282\ 10^{-6}$ kg/s ml is consequently E_{skin} = 685 W. With a large sweat production the calculation is actually more complicated since part of the sweat evaporates and the remaining part is cooled and drips off.
E_{skin} reduces with pressure (altitude) since the water particles make less collisions with the air particles which hampers their diffusion in the surrounding air. In an aircraft cabin, the dry air increases the water loss due to perspirated by some 30%, as holds for the water loss of E_{res}. Evaporation also increases some 30%, due to a higher convection.

Literature

1. ASHRAE, Fundamental Handbook, Ch. 8 Physiological Principles, Comfort and health. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, 1989.
2. Bernards J.A. and Bouwman L.N. Fysiologie van de mens. 5ft edition. Bohn, Scheltema & Holkema, 1988.
3. Polytechnisch Zakboekje, 1993, PBNA, Arnhem

Cauterization

Principle

Cauterization by heat
This is burning of tissue with a hot cauter for removing or burning arteries to stop them from bleeding. Cautery can also mean the branding of e.g. a livestock. Formaly, cauterization was used to stop heavy bleeding (amputations) by a piece of heated metal placed onto the wound. Later special medical instruments called cauters were used to cauterize e.g. arteries. The present form is electrocautery. The basics of electrocautory equipment is similar as that of Electrosurgery.
Cauterization by refrigeration
Removal of tissue (warts) and stopping of bleedings can also be performed by a cold probe, cooled by e.g. liquid nitrogen. This technique has been evolved to cryosurgery.
Chemical cautery
Many chemical reactions can destroy tissue and some are used routinely in medicine, most commonly for the removal of small skin lesions (i.e. warts or necrotized tissue) or hemostasis.

Application

Electrocauterization is the process of destroying tissue with electricity and is a widely used technique in modern surgery, as for instance cutting through soft tissue i.e. abdominal fat in a laparotomy or breast tissue in a mastectomy, and especially small vessels bleedings (larger vessels being ligated). Applications of *cryosurgery* respect general surgery, gynecology, otorhinolaryngology and skin oncology.
Chemical cauterizing is performed by e.g. silver nitrate and cantharidin. The former is bounded in a small stick that is dipped into water and pressed onto the lesion to be cauterized for a few moments. Cantharidin, an extract of the blister beetle, causes epidermal necrosis and blistering (warts).

More Info

Nasal Cauterization
Recurrent nose bleeds are most likely caused by an exposed blood vessel. In a bleeding-free period, it can be cauterized. The different methods of cauterization to stop the nose bleeding include burning the affected area with acid, hot metal, Lasers, or silver nitrate. Sometimes liquid nitrogen is used as a less painful alternative, though less effective. Topically applied cocaine make this procedure less uncomfortable and cocaine is the only local anesthetic which also produces vasoconstriction.

Chromatography

Principle

Chromatography is a family of chemical techniques, analytically and preparatively, for the separation of mixtures. It involves passing the mixture containing the analyte, in the "mobile phase", often in a stream of solvent, through the "stationary phase." The stationary phase retards the passage of the sample components. When components pass through the system at different rates they become more and more separated in time. Each component has a characteristic time of passage through the system, called the "retention time." Chromatographic separation is achieved when the retention time of the analyte differs from that of other components in the sample.
The mixture is carried by liquid or gas and is separated into its component parts as a result of differential distributions of the solutes as they flow over a stationary liquid or solid phase. Various techniques rely on the differential affinities of substances for a gas or liquid mobile medium and for a stationary absorbing medium through which they pass, such as paper, gelatin, alumina or silica.
The (chemical) physics underlying all types of chromatography concern various cohesive and adhesive forces (see Cohesion, Adhesion, Capillary forces and Surface tension) and diffusion (see Diffusion: general) with the analyte, mobile and stationary phase playing the key roles.

Application

In many clinical disciplines but especially in internal medicine. Also in many biomedical, biochemical and chemical research and industrial.

More Info

Retention
The retention is a measure of the speed at which a substance moves in a chromatographic system. In continuous development systems where the compounds are eluted with the eluent, the retention is usually measured as the *retention time* R_t or t_R, the time between injection and detection. In interrupted development systems like thin layer chromatography, the retention is measured as the *retention factor* R_f, defined by:

R_f = distance moved by compound / distance moved by eluent.

Since it is hard to standardize retention, a comparison is made with a standard compounds under absolutely identical conditions.
A chromatographic system can be described as the mobile and stationary phases being in equilibrium. The partition coefficient K is based on this equilibrium, defined as K = [solute in stationary phase] / [solute in mobile phase]. K is assumed to be independent of the concentration of the analyte, and can change if experimental conditions are changed, for example temperature. As K increases, it takes longer for solutes to separate. For a column of fixed length and flow, the retention time (t_R) and retention volume (V_r) can be measured and used to calculate K.

Chromatographic techniques

Paper chromatography

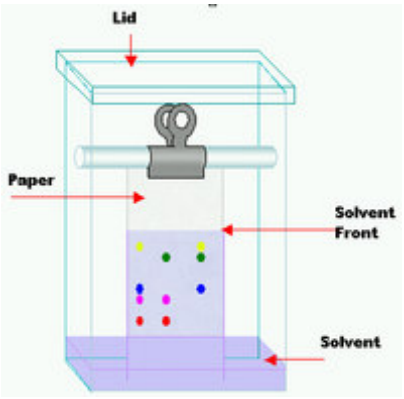


Fig. 1 Principle of a paper chromatograph

In paper chromatography a small spot of solution containing the sample is applied to a strip of chromatography paper (Fig. 1). This sample is adsorbed onto the paper. The paper is then dipped into a

suitable solvent (such as ethanol or water) and placed in a sealed container. As the solvent rises through the paper by capillary forces it meets the sample mixture. This starts to travel up the paper with the solvent. Cohesive and adhesive interactions with the paper make different compounds travel at different rates. The process takes some hours. The final chromatogram can be compared with other known mixture chromatograms for identification. This technique demonstrates very well the principle, but at the moment it has only educational relevance.

Thin layer chromatography (TLC)

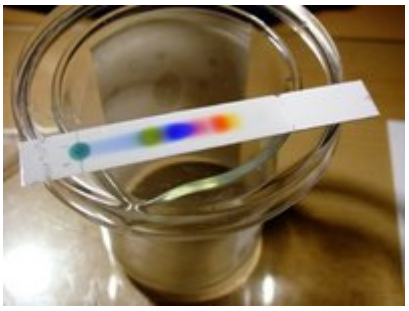


Fig. 2 Separation of black ink on a TLC plate

In TLC the stationary phase is a thin adsorbent layer like silica gel, alumina, etc. on a flat carrier like a glass plate, a thick aluminum foil, etc. (Fig. 2). The process is similar to paper chromatography but runs faster and separates better. It is used for monitoring chemical reactions and analysis of reaction products. Colorless spots of the compounds are made visible by a fluorescent dye (see Fluorescence) in the adsorbent under UV light. R_f values should be the same regardless of the extent of travel of the solvent, and in theory they are independent of a single experimental run. They do depend on the solvent used, and the type of TLC plate. Nowadays relevance is mainly educational.

Column chromatography

Column chromatography utilizes a vertical glass column filled with some form of solid support with the sample to be separated placed on top of this support. The rest of the column is filled with a solvent which, under the influence of gravity, moves the sample through the column. Similarly to other forms of chromatography, differences in rates of movement through the solid medium are translated to different exit times from the bottom of the column for the various elements of the original sample.

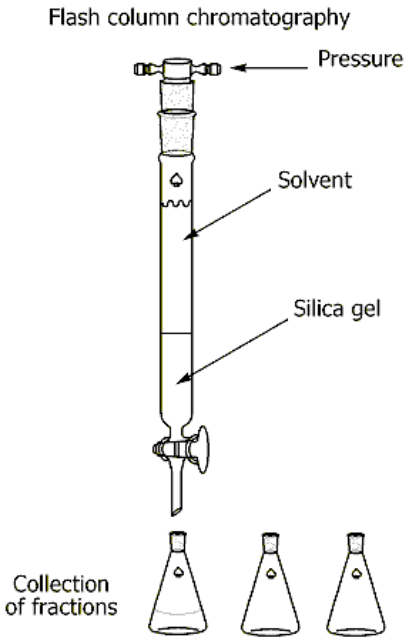


Fig. 3 Principle of a flash column chromatography setup

The flash column chromatography (Fig. 3) is very similar to the traditional column chromatography, except for that the solvent is driven through the column by applying positive pressure. It is faster and gives better separation. Miniaturized (disposable columns), this technique is widely applied.

Gas(-liquid) chromatography G(L)C
Gas-liquid chromatography is based on a partition equilibrium of analyte between a liquid stationary phase and a mobile gas. The mobile phase is a carrier gas, usually an inert gas such as He or N_2, and the stationary phase is a microscopic layer of liquid on an inert solid support inside of a very long and very thin column. It is useful for a wide range of non-polar analytes, but poor for thermally labile molecules. It is often combined with Mass spectrography.

Ion exchange chromatography
Ion exchange chromatography is a column chromatography that uses a charged stationary phase. It is used to separate charged compounds including amino acids, peptides and proteins. The stationary phase is usually an ion exchange resin that carries charged functional groups which interact with oppositely charged groups of the compound to be retained. Bound compounds can be eluted from the column by gradient elution (solvent composition with a time-gradient, e.g. in salt concentration or pH) or isocratic elution (solvent composition constant). Ion exchange chromatography is commonly used to purify proteins using Fast Protein Liquid Chromatography (FPLC).

Immobilized metal ion affinity chromatography, IMAC
IMAC is a popular and powerful way to purify proteins. It is based on the specific covalent binding between histidine or other unique amino acids (either naturally or grafted with recombinant DNA techniques) and various immobilized metal ions, such as copper, nickel, zinc, or iron.

High performance liquid chromatography, HPLC
HPLC is a form of column chromatography used frequently in biochemistry and analytical chemistry. The analyte is forced through a column (stationary phase) by a liquid (mobile phase) at high pressure, which decreases the time the analytes have to diffuse within the column. Diffusion within the column leads to broad peaks and loss of resolution. Less time on the column then translates to narrower peaks in the resulting chromatogram and thence to better resolution and sensitivity (discrimination from ‘’background noise’). Another way to decrease time the analyte stays on the column is to change the composition of the mobile phase over a period of time (a solvent time-gradient). HPLC is often combined within one apparatus with a mass spectrograph or gas(-liquid) chromatograph.

Reversed phase (RP) liquid chromatography
RP-HPLC was developed for large polar biomolecules. Like the name implies the nature of the stationary phase is reversed. RP-HPLC consists of a nonpolar stationary phase and a polar mobile phase. One common stationary phase is special treated silica. The retention time is longer when the mobile phase is more polar. This is the reverse of the situation which exists when normal silica is used as the stationary phase.

Gel permeation chromatography GPC
GPC also known as size exclusion chromatography or Sephadex gel chromatography, separates molecules on basis of size. Smaller molecules enter a porous media and take longer to exit the column, hence larger particles leave the column first. GPC is good for determining polymer molecular weight distribution, but has a low resolution.

Affinity chromatography
Affinity chromatography is based on selective non-covalent interaction between an analyte and specific molecules. It is often used in the purification of proteins (or better protein constructs).

There are many other versions of chromatography, see Wikipedia or textbooks on analytic chemistry.

Diffusion: Fick's laws

Principle

Fick's first law
Fick's first law is used in steady-state diffusion (see Diffusion: general), i.e., when the concentration within the diffusion volume does not change with respect to time ($J_{in} = J_{out}$).

$$J = -D\frac{\partial \phi}{\partial x}, \quad (1)$$

where
J is the diffusion flux in dimensions of (mol m^{-2} s^{-1});
D is the diffusion coefficient or diffusivity, (m^2 s^{-1});
φ is the concentration (mol m^{-3});
x is the position (m).
D is proportional to the velocity of the diffusing particles, which depends on the temperature, viscosity of the fluid and the size of the particles according to the Stokes-Einstein relation. For biological molecules the diffusion coefficient normally ranges from 10^{-11} to 10^{-10} m^2 s^{-1}.

Fick's second law
Fick's second law is used in non-steady or continually changing state diffusion, i.e., when the concentration within the diffusion volume changes with respect to time.

$$\frac{\partial \phi}{\partial t} = -D\frac{\partial^2 \phi}{\partial x^2}, \quad (2)$$

Where:
ϕ is the concentration (mol m^{-3});
t is time (s);
D is the constant diffusion coefficient (m^2 s^{-1});
x is the position (m).
With some calculation, it can be derived from the First Fick's law and the mass balance.

Application

Equations based on Fick's law have been commonly used to model transport processes in foods, porous soils, semiconductor doping process, etc.. In biomedicine, transport of biopolymers, pharmaceuticals, in neurons, etc. are modeled with the Fick equations.
Typically, a compound's D is ~10,000x greater in air than in water. CO_2 in air has a diffusion coefficient of 16 mm^2/s, and in water, its coefficient is 0.0016 mm^2/s.

Physiological examples
For example for the steady states, when concentration does not change by time, the left part of the above equation will be zero and therefore in one dimension and when *D* is constant, the equation (2) becomes $0 = D\, d\varphi/dx^2$. The solution for the concentration φ will be the linear change of concentrations along *x*. This is what by approximation happens in the alveolar-capillary membrane and with diffusion of particles from the liquid to liquid phase in the wall of capillaries in the tissues.

Biological perspective
The first law gives rise to the following formula:

$$\text{Rate of diffusion} = KA(P_2 - P_1) \quad (3)$$

It states that the rate of diffusion of a gas across a membrane is:
- Constant for a given gas at a given temperature by an experimentally determined factor, K
- Proportional to the surface area over which diffusion is taking place, A
- Proportional to the difference in partial pressures of the gas across the membrane, $P_2 - P_1$

It is also inversely proportional to the distance over which diffusion must take place, or in other words the thickness of the membrane. The factor K comprises this thickness.
The exchange rate of a gas across a fluid membrane can be determined by using this law together with Graham's law (see Diffusion: Graham's law.

More Info

Temperature dependence of the diffusion coefficient
The diffusion coefficient at different temperatures is often found to be well predicted by

$$D = D_0 e^{\frac{-E_A}{RT}}, \quad (4)$$

where:
D_0 is the maximum diffusion coefficient at infinite temperature,
E_A is the activation energy for diffusion (J/mol),
T is absolute temperature (K),
R is the gas constant in dimensions of (J/(K·mol)).

3D diffusion
For the case of 3-dimensional diffusion the Second Fick's Law looks like:

$$\frac{\partial \phi}{\partial t} = D\nabla^2 \phi \quad (5)$$

where ∇ is the del operator. In a 3D system with perpendicular coordinates (*x*, *y*, *z*), this is a Cartesian coordinate system $\mathbf{R}^3$, del is defined as:

$$\nabla = i\frac{\partial}{\partial x} + j\frac{\partial}{\partial y} + k\frac{\partial}{\partial z}, \quad (6)$$

where i, j and k are the unit-vectors in the direction of the respective coordinate (the standard basis in $\mathbf{R}^3$).

Finally, if the diffusion coefficient is not a constant, but depends upon the coordinate and/or concentration, the Second Fick's Law looks like:

$$\frac{\partial \phi}{\partial t} = \nabla \cdot (D\nabla \phi). \quad (7)$$

Diffusion: general

Principle

Diffusion is the net action of particles (molecules, atoms, electrons, etc.), heat, momentum, or light whose aim is to minimize a concentration gradient. A concentration gradient is the difference between the high concentration and the low concentration. It also includes the speed of the process. Diffusion can be quantified by measuring the concentrations gradient.
More formally, diffusion is defined as the process through which speed a thermodynamic system at local thermodynamic equilibrium returns to global thermodynamic equilibriums, through the homogenization of the values of its intensive parameters (or bulk parameters, such as viscosity, density, melting point etc.).

In all cases of diffusion, the net flux of the transported quantity (atoms, energy, or electrons) is equal to a physical property (diffusivity, thermal conductivity, electrical conductivity) multiplied by a gradient (a concentration, thermal, electric field gradient). So:
Diffusion ≡ some conductivity times a gradient.
Noticeable transport occurs only if there is a gradient. For example, if the temperature is constant, heat will move as quickly in one direction as in the other, producing no net heat transport or change in temperature.

The process of diffusion minimizes free energy and is thus a *spontaneous* process. An example of diffusion is the swelling of pasta, where water diffuses into the sponge-like structure of the dry and stiff pasta.
The different forms of diffusion can be modeled quantitatively using the diffusion equation, (see **More Info**) which goes by different names depending on the physical situation. For instance - steady-state bi-molecular diffusion is governed by Fick's laws (see Diffusion: Fick's law, steady-state thermal diffusion is governed by Fourier's law. The generic diffusion equation is time dependent, and as such applies to non-steady-state situations as well.

The second law of thermodynamics states that in a spontaneous process, the entropy of the universe increases. Entropy is a measure of how far a spontaneous physical process of smoothing-out differences has progressed, for instance a difference in temperature or in concentration (see further Thermodynamics: entropy). Change in entropy of the universe is equal to the sum of the change in entropy of a system and the change in entropy of the surroundings. A system refers to the part of the universe being studied; the surrounding is everything else in the universe. Spontaneous change results in dispersal of energy. Spontaneous processes are not reversible and only occur in one direction. No work is required for diffusion in a closed system. Reversibility is associated with equilibrium. Work can be done on the system to change equilibrium. Energy from the surroundings decrease by the amount of work expended from surroundings. Ultimately, there will be a greater increase in entropy in the surroundings than the decrease of entropy in the system working accordingly with the second law of thermodynamics (see Thermodynamics: second law).

Types of diffusion
Diffusion includes all transport phenomena occurring within thermodynamic systems under the influence of thermal fluctuations (i.e. under the influence of disorder; this excludes transport through a hydrodynamic flow, which is a macroscopic, ordered phenomenon). Well known types are the diffusion of atoms and molecules, of electrons, (resulting in electric current), Brownian motion (e.g. of a single particle in a solvent), collective diffusion (the diffusion of a large number of (possibly interacting) particles), effusion of a gas through small holes, heat flow (thermal diffusion), osmosis, isotope separation with gaseous diffusion.

Application

Application, i.e. occurrence, in nature, dead and alive, can be found nearly everywhere. In living bodies it is found interstitionally as well as in living cells. Many man made apparatus, also for medical purposes, make use of the principle of diffusion. Some very obvious examples of applications are in the techniques of Chromatography, Electrophoresis, Oxygen analysis, Spectroscopy, Thermography, the calculation of body heat conduction and dissipation (see Body heat conduction and Newton's Law of cooling) etc.

Diffusion in biological systems
Specific examples in biological systems are diffusion across biological membranes, of ions through ion channels, in the alveoli of mammalian lungs across the alveolar-capillary membrane. In the latter

process the diffusion is Knudson-like diffusion, in other words it is also dependent on space restrictions (the small size of the providing (alveoli) and recipient (capillaries) volumes. Another type is facilitated diffusion (passive transport across a membrane, with the assistance of transport proteins)

Numeric example

By knowing the diffusion coefficient of oxygen in watery tissue, the distance of diffusion and the diffusion gradient, then the time required for the diffusion of oxygen from the alveoli to the alveolar capillaries can be calculated. It appears to be some 0.6 ms (see **More Info**).

More Info

Diffusion equation

The diffusion equation is a partial differential equation, which describes the density fluctuations in a material undergoing diffusion. It is also used in population genetics to describe the 'diffusion' of alleles in a population.

The equation is usually written as:

$$\frac{\partial \phi}{\partial t} = \nabla \cdot D(\phi) \nabla \phi(\vec{r}, t) \qquad (1)$$

where ϕ is the density of the diffusing material, *t* is time, D is the collective diffusion coefficient, $\vec{r}$ is the spatial coordinate and the nabla symbol ∇ represents the vector differential operator del (see also Fick's Laws). If the diffusion coefficient depends on the density then the equation is nonlinear. If *D* is a constant, however, then the equation reduces to the following linear equation:

$$\frac{\partial \phi}{\partial t} = D \nabla^2 \phi(\vec{r}, t)\,, \qquad (2)$$

also called the heat equation.

Diffusion displacement

The diffusion displacement can be described by the following formula:

$$\langle r_k^2 \rangle = 2kD't, \qquad (3)$$

where *k* is the dimensions of the system and can be one, two or three. D' is the diffusion coefficient, but now also per unit of concentration difference of the particles and *t* is time. For the three-dimensional systems the above equation will be:

$$\langle x^2 \rangle + \langle y^2 \rangle + \langle z^2 \rangle = \langle r_3^2 \rangle = 6D't \quad , \qquad (4)$$

where < > indicates the mean of the squared displacement of the individual particles.

Physiological numeric example

Calculate the time required for O_2 diffusion across the alveolar-capillar membrane. By knowing D' of oxygen in watery tissue (suppose 15×10^{-10} $m^2 s^{-1} bar^{-1}$), the 1D-distance of diffusion x (suppose the distance between alveolar and capillary volume is 0.4 µm) and the O_2 diffusion gradient Δp is 0.09 bar (suppose partial pressures of 0.14 bar in alveoli and 0.05 bar in the pulmonary artery) and supposing that the system is ideal (half infinite at both sides and a constant gradient), then the time required for the diffusion of O_2 from the alveoli to the alveolar capillaries can be calculated from (3) but now defined for a gas:

$$t = x^2/(2D\Delta p) = 0.16 \times 10^{-12}/(2 \times 15 \times 10^{-10} \times 0.09) = 0.6 \times 10^{-3} \text{ s.} \qquad (5)$$

Actually, this time should be a little longer due to the circular shape of the capillaries. See further Diffusion: Fick's laws, for the diffusion equations and Diffusion: Grahams Law for particle velocity of diffusing particles.

Diffusion: Graham's law

Principle

Graham's law is also known as Graham's law of effusion. Effusion is the process where individual molecules of a gas flow through a hole without collisions. This will occur if the diameter of the hole is considerably smaller than the mean (collision-)free path of the molecules. At 1 bar, the free path is of the oder of 70 nm. The law states that the rate of effusion of a gas is inversely proportional to the square root of the molecular mass of its particles. This formula can be written as:

$$v_1/v_2 = (m_2/m_1)^{0.5}, \quad (1)$$

where:
v_1 is the rate of effusion of the first gas;
v_2 is the rate of effusion for the second gas;
m_1 is the molar mass of gas 1;
m_2 is the molar mass of gas 2.

Graham's law is most accurate for molecular effusion which involves the movement of one gas at a time through a hole. It is only approximate for diffusion of one gas in another or in air, as these processes involve the movement of more than one gas.
According to the kinetic theory of gases, the absolute temperature T (Kelvin) is directly proportional to the average kinetic energy of the gas molecules ($0.5mv^2$). This can be derived from:
$\langle v^2 \rangle = 3RT/(N_A m)$,
where $\langle v^2 \rangle$ the mean of the squared velocity of the particles, R the molar gas constant (= 8315 J/kmol·K), $N_A m$ the gas mass with N_A the Avogadro's number (see Gas laws).
Thus, to have equal kinetic energies and so temperature, the velocities of two different molecules would have to be in inverse proportion to the square roots of their masses. Since the rate of diffusion is determined by the average molecular velocity, Graham's law for diffusion could be understood in terms of the molecular kinetic energies being equal at the same temperature.

Application

The law is of important for processes of diffusion of a gas into another gas or gas mixture (see Diffusion: general). For diffusion across a liquid membrane, gas concentration should be low and the membrane thin.

Graham's Law can also be used to find the approximate molecular weight of a gas if the rates are measured and the molecular weight of one of the gases is a known.

Graham's law was the basis for separating $^{235}UF_6$ from $^{238}UF_6$. Both isotopes of uranium, as element, are found in natural uranium ore, but the 235-isotope about 100 times less. By repeated diffusion through porous barriers the slightly lighter ^{235}U isotope is enriched.

Electrophoresis

Principle

Electrophoresis is the movement of an electrically charged substance under the influence of an electric field. This movement is due to the Lorentz force, which acts on the charge of the particle under study and which is dependent on the ambient electrical conditions. This force is given by:

$$F = qE \quad (1)$$

F (a vector) is the Lorentz force, q is the charge (a scalar) of the particle, E is the electric field, a vector. The resulting electrophoretic migration is countered by forces of friction, such that the rate of migration is constant in a constant and homogeneous electric field:

$$F_f = vf\,, \quad (2)$$

where v is the velocity and f is the frictional coefficient. Since in the stationary condition forces of friction and Lorentz force become the same, it holds that:

$$qE = vf \quad (3)$$

The electrophoretic mobility μ is defined as:

$$\mu = v/E = q/f\,. \quad (4)$$

The expression (4) above applied only to charged molecules (ions) at a low concentration and in a non-conductive solvent. Poly-ionic molecules are surrounded by a cloud of counter-ions which alter the effective electric field applied on the molecule. This render the previous expression a poor approximation of what really happens in an electrophoretic apparatus.

Application

Electrophoresis is used as an preparative and analytical tool in molecular biology.
Gel-electrophoresis is an application of electrophoresis in molecular biology, especially in DNA techniques. The gel-electrophoresis apparatus uses a positive and a negative charged pole. The (macro)molecule, e.g. DNA is loaded on the negatively charged pole and pulled through the gel toward the positive pole.
The charge of a DNA molecule is provided by negative phosphate groups. The content of the buffers (solutions) and gels used to enhance viscosity greatly affects the mobility of macromolecules. The gel used in the procedure is typically an agarose or a polyacrylamide gel, depending on the type of molecule studied. The thickness of the gel is typically ca. 8 mm. Within the gel is a tightly woven matrix that the molecules must pass through as they are moving from one pole to the other. The smaller molecules can weave in and out of the matrix of the gel with more ease, compared with larger molecules. Wells, or rectangular openings, are formed along one edge of the gel. These wells mark the different lanes in which a sample may be loaded. Agarose is applied to separate large DNA fragments (100-50,000 base pairs) with modest resolution (some 50 base pairs). Polyacrylamide is useful for small fragments (10-1000 pairs) with single base-pair resolution.
Modifications are e.g. gradient (detergent) gel-electrophoresis and (water cooled) rapid agarose gel electrophoresis.

More Info

The mobility depends on both the particle properties (e.g., surface charge density and size) and solution properties, e.g., ionic strength, electric permittivity, and pH. (Permittivity describes how an electric field affects and is affected by a medium, e.g. air or water). For high ionic strengths, an approximate expression for the electrophoretic mobility μ_e is given by the equation:

$$\mu_e = \varepsilon \cdot \varepsilon_0 \cdot \zeta / \eta, \quad (5)$$

where ε is the dielectric constant (relative permittivity) of the liquid, ε_0 is the permittivity of vacuum, η is the viscosity of the liquid, and ζ is the zeta potential (i.e.electrokinetic potential of the plane between the attached fluid and the mobile fluid (called slipping plane) in the electric double layer) of the particle (see *Double layer (interfacial)* in Wikipedia).

Electrosurgery

Principle

Electrosurgery is the application of a high-frequency electric current to tissue as a means to remove lesions, staunch bleeding, or cut tissue. It is based on the generation of local heat dissipated by a piece of tissue when electric current is flowing through it. The tissue can be considered as an electric resistor. To perform electrosurgery, a voltage source is applied to the fine surgical electrode or probe (electric knife) and a 2nd electrode with the tissue in between. The current can be adjusted by changing the voltage of the source. The dissipated power is $P = V^2/R$ (in Watts) which directly follows from Ohms law ($V = iR$) and the relation $P = iV$).
The electrode does not heat up since the resistance of the metal electrode and metal wire is so much smaller than that of the tissue that very little power is expended inside the metal conductors. Electrosurgery is performed using a device called an Electrosurgical Generator, sometimes referred to as an RF Knife.

The change in temperature of an object is inversely proportional to its (specific) heat capacity (in $J \cdot kg^{-1} \cdot K^{-1}$). For water (near 20 °C) this is 4184 $J \cdot kg^{-1} \cdot K^{-1}$. This value can also be used for watery tissues. Since the heat needed is proportional to the mass of the object, the heated mass is limited by using small probes. By applying a high current density (current/area), which is achieved by a small electrode tip, the resistance of the small volume of tissue adjacent to the tip is subjected to a large current density. The generated heat will now easily burn the tissue at the electrode.

The human nervous system is very sensitive to low-frequency (0 Hz to ca. 1000 Hz) electricity, which stimulates the nervous system. At even low currents low-frequency electricity causes electric shock, which may involve acute pain, muscle spasms, and/or cardiac arrest. The sensitivity decreases with increasing frequency and at frequencies > 100 kHz, electricity does not stimulate the nervous system. To avoid electric shock, electrosurgical equipment operates in the frequency range of 200 kHz to 5 MHz.

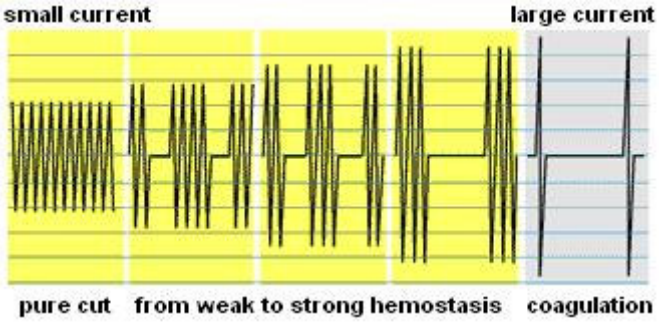


Fig. 1 Modes of operation.

Fig. 1 illustrates the various operation modes for the various aims.

Application

Electrosurgery can be used to cut, coagulate, desiccate, or fulgurate tissue. Its benefits include the ability to make precise cuts with limited blood loss.
Although electrosurgical devices may be used for the cauterization of tissue in some applications (e.g. hemorrhoid surgery), electrosurgery is usually used to refer to a quite different method than that used by many dedicated electrocautery devices. The latter uses heat conduction (see Body heat conduction and Newton's Law of cooling) from a hot probe heated by a direct current (much in the manner of a soldering iron), whereas electrosurgery uses alternating current to directly heat the tissue itself (electric diathermy), while the probe tip remains relatively cool.
Different waveforms of the electric current can be used for different electrosurgical procedures. For cutting, a continuous single frequency sine wave is generated. This produces rapid heating. At the cellular level, rapid heating causes tissue cells to boil and burst. At a larger scale, the ruptured cells create a fine tear in the tissue, creating a clean incision.
For coagulation, the sine wave is modulated by turned on and off in rapid succession. The overall effect is a slower heating process, which causes cells to coagulate. The proportion of on time to on+off time, the duty cycle, can be varied to allow control of the heating rate.

Dermatological applications are removal of skin tags, removal/destruction of benign skin tumors and warts. For several aims it is now often preferred by dermatologists over laser surgery and cryosurgery.

Savety
High power *monopolar* surgery requires a good electrical contact between a large area of the body and the return electrode to prevent severe burns (3rd degree) in unintended areas on the skin and beneath the skin of (anesthetized) patients. To prevent unintended burns, the skin should be clean and dry and a conductive jelly should be used. Proper electrical grounding practices must be followed in the electrical wiring of the building. It is also recommended to use a newer electrosurgical unit that includes alarms for ground circuit interruption.
Safety is essentially improved when the electric circuit is inductively uncoupled. This is performed by a primary coil in the generator and a secondary coil in the circuit of the probe.

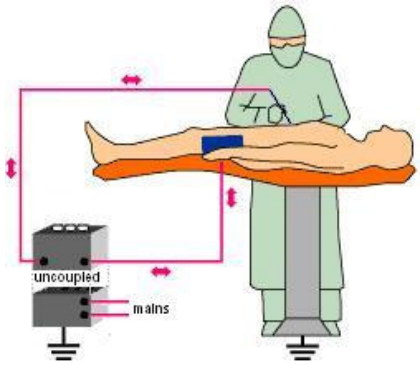


Fig. 2 Bipolar setup with the circuit uncoupled from the mains by coils. The double headed arrow indicates the currents direction which alternates every cycle of the current (A.C. or alternating current).

Medical numerical example
A strongly simplified calculation yields the order of magnitude of the strength of currents and voltages applied. Suppose that a small blood vessel (say 2 mm in diameter) should be coagulated. This is performed by applying a current, which coagulates at 237 ^{o}C a sphere with a diameter of 2.8 mm (volume 0.23 ml). This requires 0.23 g x 4.2 J/(g x K) x 200 K = 10 J (tissue density is 1 g/ml). This is applied in 5 s, so 2 J/s, is 2 Watt. Supposing an electrode resistance of 10 KOhm (bipolar set up), then 140 V (= $P^{0.5}V^{0.5}$ = (10000 x 2)$^{0.5}$) is needed and the current is 140 V/10 KOhm = 14 mA (calculated by using $P = V^2/R$). This holds for continues current (DC or direct current), whereas the currents is applied intermittent in high frequency bursts. Supposing a duty cycle of 10% and a sinusoidal current, then the amplitude of the voltage is $2^{0.5}$ X 10 X 140/2 is about 1000 V. The 1000 V (generating 71 mA effectively) again underlines the absolute necessity of save grounding and shows that uncoupling is the ultimate save approach.

More Info

Electrosurgical modalities
Monopolar and Bipolar There are two circuit topologies: *monopolar* and *bipolar*. The *bipolar modality* is used less often. Voltage is applied to the patient using a special forceps, with one tine connected to one pole of the A.C. (alternating current) voltage source and the other tine connected to the other pole of the voltage source. When a piece of tissue is held by the forceps, a high frequency electrical current flows from one to the other forceps tine, through the intervening tissue. In the *monopolar modality* the patient lies on top of the *return electrode*, a relatively large metal plate or a relatively large flexible metalized plastic pad (somewhere attached to the body), which is connected to the other wire of the current source. The surgeon uses a single, pointed, probe to make contact with the tissue. The electrical current flows from the probe tip through the body and then to the return electrode. In the monopolar modality the heating is also very precisely confined to the tissue that is near the probe tip since the current rapidly spreads out laterally in the body, causing a quadratic decrease in the current density with the distance.
Spark gap or *fulguration modality* It is a low-powered monopolar electrosurgery performed on conscious outpatients (dermatology). At low power, this technique requires no return electrode or patient-contact-plate since at the very high frequencies and low currents, the parasitic capacitance between the patient's body and the machine's ground potential is large enough to close the electric circuit. If such a spark is small, it can cause relatively minor shocks or burns, but these are not a problem in a low-powered setting with conscious patients because they are immediately noticed. For high-power or surgical anesthesia settings, however, a ground pad is always necessary to insure that all such stray ground currents enter the machine safely through a large-skin-surface contact, and dedicated wire.

Flow cytometry

Principle

Flow cytometry is a technique for counting, examining, and sorting microscopic particles suspended in a stream of fluid. It allows simultaneous multiparametric analysis of the physical and/or chemical characteristics of single cells flowing through an optical and/or electronic detection apparatus.

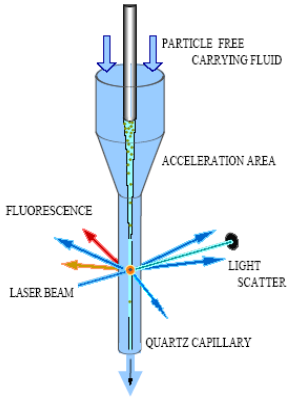


Fig. 1 Principle of fluorescence cytoflow meter with 3 emitted wavelengths, forward and side scatter.

A beam of light (usually laser light) of a single wavelength is directed onto a hydro-dynamic focussing stream of fluid. With hydrodynamic focusing a "wall" fluid called the sheath fluid is being pumped through. The sample is injected into the middle of the sheath flow. If the two fluids differ enough in their velocity or density, they do not mix: they form a two-layer stable flow, with the sheath enveloping the sample in a linear, steady flow. The diameter of the sample flow is of µm magnitude.
Each suspended particle passing through the beam scatters the light (see Light: scatter) in some way (Fig. 1). The scatter is in line with the light beam (Forward Scatter or FSC) and perpendicularly to it (Side Scatter (SSC). Fluorescent chemicals found in the particle or attached to the particle may be excited to emit light (see Fluorescence). This combination of scattered and fluorescent light is picked up by the detectors and the evoked signals are analyzed.
FSC correlates with the cell volume and SSC depends on the inner complexity of the particle (i.e. shape of the nucleus, the amount and type of cytoplasmatic granules or membrane roughness). Some flow cytometers have eliminated the need for fluorescence and use only light scatter for measurement. Other flow cytometers form images of each cell's fluorescence, scattered light and transmitted light.

Application

Applications include research in molecular biology (fluorescence tagged antibodies in transplantation, hematology, tumor immunology and chemotherapy, genetics and sperm sorting in IVF), pathology, immunology, plant biology and marine biology (auto-fluorescent properties of photosynthetic plankton, Fig. 2). In protein engineering, flow cytometry is used in conjunction with yeast display and bacterial display to identify cell surface-displayed protein variants with desired properties.

A flow cytometer has 6 main components:
- a flow cell: a liquid stream carries and aligns the cells so that they pass one by one through the light beam for sensing;
- a light source: various types of high pressure lamps and at present mostly lasers are used (HeNe for 632 nm, red; Ar for488 nm, blue-green; Kr for 341 nm, blue light);
- dichroic filters (see Light: beam splitter) for the various emitted wavelengths;
- detectors (photomultiplier tubes);
- an electronic amplifier;
- ADC (analogue to digital converter) generating FSC and SSC as well as digital fluorescence signals.
- a computer for analysis of the signals.

Modern flow cytometers are able to analyze several thousand particles/s, in "real time", and can actively separate and isolate particles having specified properties. A flow cytometer is similar to a microscope (see Optical microscopy), except that instead of producing an image of the cell, flow cytometry offers

"high-throughput" (for a large number of cells) automated quantification of set parameters. To analyze solid tissues a single-cell suspension must first be prepared.

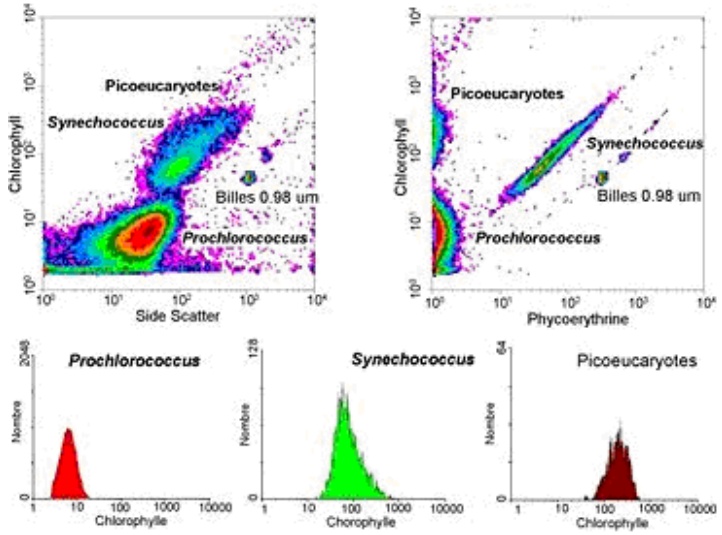


Fig. 2 Analysis of a marine sample of photosynthetic picoplankton by flow cytometry showing three different populations

Commercial flow cytometers can be used with a large number of reagents, such as fluorescently-labeled antibodies and analysis software.
Modern instruments usually have multiple lasers and many fluorescence detectors. Increasing the number of lasers and detectors allows for multiple antibody labeling, and can more precisely identify a target population by their phenotype. Certain instruments can even take digital images of individual cells, allowing for the analysis of fluorescent signal location within or on the surface of cells.
The data generated by flow-cytometers can be plotted 1D (lower row of Fig. 2 with number versus, here, fluorescence), 2D (upper left panel Fig. 2 with fluorescence versus side scatter) or even 3D. The plots are often made on logarithmic scales. Because different fluorescent dyes' emission spectra overlap, signals at the detectors have to be compensated electronically as well as computationally.
The fluorescence labels that can be used will depend on the lamp or laser used to excite the fluorochromes and on the detectors available.

More Info

Fluorescence-activated cell sorting (FACS)
Fluorescence-activated cell sorting is a specialized type of flow cytometry with sorting a heterogeneous mixture of biological cells into various containers, one cell at a time, based upon the specific scattering and fluorescent characteristics of each cell.
The cell suspension is entrained in the center of a narrow, rapidly flowing stream of liquid. The flow is arranged so that there is a large separation between cells relative to their diameter. A vibrating mechanism causes the stream of cells to break into individual droplets. The system is adjusted so that there is a low probability of more than one cell being in a droplet. Just before the stream breaks into droplets the flow passes through a fluorescence measuring station where the fluorescent character of interest of each cell is measured. An electrical charging ring is placed just at the point where the stream breaks into droplets. A charge is placed on the ring based on the immediately prior fluorescence intensity measurement and the opposite charge is trapped on the droplet as it breaks from the stream. The charged droplets then fall through an electrostatic deflection system that diverts droplets into containers based upon their charge. In some systems the charge is applied directly to the stream and the droplet breaking off retains charge of the same sign as the stream. The stream is then returned to neutral after the droplet breaks off.
Magnetic bead sorting
An alternative sorting method is magnetic bead sorting. Instead of fluorescent molecules, antibodies are conjugated to iron particles. After staining, cells can be purified using a strong magnet. It is less expensive than FACS and large numbers of cells can be used but it is less sensitive than FACS to rare populations of cells

Literature
http://biomicro.sdstate.edu/younga/VET791/VET791.pdf

Flow: entrance effect and entrance length

Principle

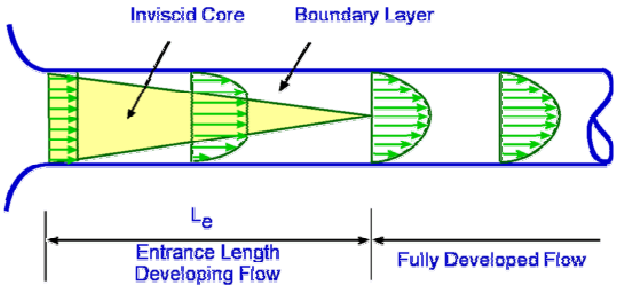


Fig. 1 Flow at the entrance to a tube

Consider a flow entering a tube and suppose that the entering flow is uniform, so inviscid. As soon as the flow 'hits' the tube many changes take place. The most important of these is that viscosity imposes itself on the flow and the friction at the wall of the tube comes into effect. Consequently the velocity along the wall becomes zero (also in tangential direction). The flow in the layer adjacent to the wall decelerates continuously. This layer is what is called the boundary Layer. Viscous effects are dominant within the boundary layer. Outside of this layer is the inviscid core with uniform flow where viscous effects are negligible or absent.
The boundary layer is not a static phenomenon; it is dynamic. It grows meaning that its thickness increases as we move downstream. From Fig. 1 it is seen that the boundary layer from the walls grows to such an extent that they all merge on the centre line of the tube. Once this takes place, inviscid core terminates and the flow is all viscous. The flow is now called a *fully developed flow*; the velocity profile is parabolic or Poiseuille flow (see Poiseuille's Law). Once the flow is fully developed the velocity profile does not vary in the flow direction. In fact in this region the pressure gradient and the shear stress in the flow are in balance. The length of the tube between the start and the point where the fully developed flow begins is called the Entrance Length, denoted by L_e. The entrance length is a function of the Reynolds number Re of the flow.
The distance needed to restore a laminar inviscid flow to a parabolic laminar flow is called the entrance length, being:

$$L_{e,laminar} = C \cdot D \cdot Re, \qquad (1)$$

where D is the tube diameter and C a constant, dependent on the diameter change and abruptness of the geometrical change of the entrance. It ranges from about 0.06 (fast conical change) to 0.03 (smooth, slender conical change). The latter value holds rather well for most anatomical geometries, except for the heart-vessel and mouth-trachea transitions.

Application

Hemodynamics and flow in the airways system (see Flow through a stenosis, Flow in curvatures and Flow in bifurcations).
Numerical example Supposing that the aorta diameter is 3 cm and Re is 1400, which holds in rest, then $L_{e,turbulent}$ is about 250 cm. This is much more than the distance of the first large bifurcations after the aortic arch. Also when we take (conceptually) into account the aorta bend and the bifurcations in the aorta bend (which will disturb the process of restoration), the pulsatile character of the aorta flow and the compliance of the wall (which support restoration) Poiseuille flow is not restored.

More info

For turbulent flow holds:

$$L_{e,turbulent} = 4.4D \cdot Re^{1/6}. \qquad (2)$$

At the critical condition, i.e., Re_d =2300, the L_e/D ratio, the entrance length number, for a laminar flow is 138. From this condition, with transitional flow to full turbulent flow at Re = 10000 the ratio *diminishes* to 18.6.

Literature

http://www.aeromech.usyd.edu.au/aero/ pipeflow/node7.html#node9

Flow in a bended tube

Principle

When steady laminar fluid in a long straight tube enters a bend, every fluid element of it must change its direction of motion. It must acquire a component of acceleration at right angles to its original direction, and must therefore experience a force in that direction. This most fast moving elements, those at the axis, will change their original path less rapidly than slower moving ones, on account of their greater inertia ($\frac{1}{2}\rho v^2$). Their original velocity vector obtains an axial (the axis of the bend) component and a component perpendicular to the axis. At the axis within the plane of the bend, this force, perpendicular to the axial force, is directed to the outside wall of the bend. The same holds for neighboring positions, close the central plane of the bend. This results in an outward rather uniform flow. In consequence, the fluid originally at the outside of the bend has a tangential flow component along the wall in the direction of the inside of the bend. And so, a transverse circulation or secondary motion is set up, as shown in Fig. 1a. The further moved into the bend, the original parabolic velocity profile at the entrance becomes more and more skewed in the plane of the bend (Fig. 1b).
Faster moving fluid near the outside of the bend means that the axial velocity profile in the plane of the bend is distorted from its symmetric shape to the M-shape shown in Fig. 1c. The reason is that the fluid being swept back around the side walls is still traveling faster than that pulled towards the center of the tube from the inside wall.
A similar pattern of flow is to be expected far from the entrance in a long, continuously curved tube, like the aorta bend, whatever the initial velocity profile.

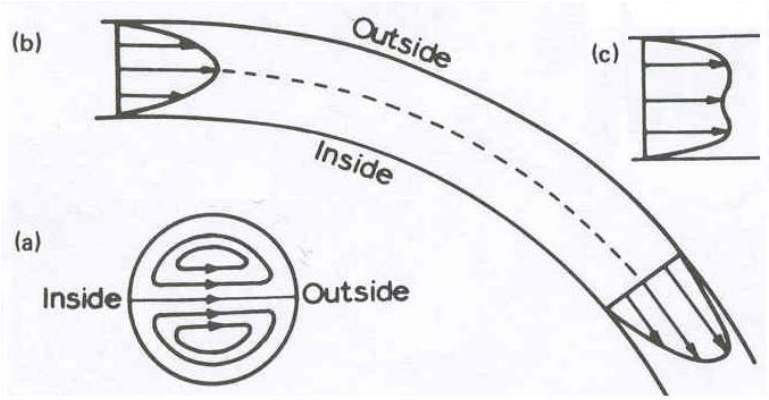


Fig. 1 Laminar flow profile in a curvature. c is perpendicular to b.

Application

In fluid dynamics of blood flow through vessels and flow in the airways. Curvatures in the airways system are generally combined with bifurcation (see Flow in bifurcations). They may increase resistance and pressure drop compared to a bifurcation with straight tubes. In the vascular system bends are very common, for instance in the brain, and can increase resistance considerably. Vessels in joints can be bend easily 90^0 and even more than 135^0. In the bend, the vessel, in particular veins, will be flattened, another effect of increasing resistance pressure drop. Supposing a vessel is compressed to half its diameter, whereas the perimeter remains the same and the new shape is elliptical, than the area is decreased by 16%. When the volume flow remains the same, then the Reynolds number increases by 16%.

With a sharp bend over 90^0 with a curvature ratio of 1.5, as occurs with the *aorta bend*, a curvature coefficient (resistance increment; see **More Info**) ζ_c of 0.17 is reached. In physical rest, with a cardiac output of 6 L/min and v is 0.15 m/s, Reynolds number Re of the aorta is about 1350, supposed the aorta is straight, there is no entrance effect (see Flow: entrance effect and entrance length) of the aortic valve and the flow is steady. Taking the pulsatile blood flow into account the maximal velocity is about 0.6 m/s. Taking the bending and the entrance effect (entrance coefficient ζ_e at least 0.2) also into account, Re will reach a temporarily maximum of Re much higher than 1350. Although with pulsatile flow the critical Re is higher than 2100, the flow in the aorta is supposed to have transitional flow, also at rest. With very high cardiac outputs (30 L/min), as occur during heavy endurance sport, the aortic flow is turbulent and this will proceed in the largest arteries.
In an 180^0 strong bend, as in the aorta, and high flows, which occur during heavy exercise, there arise curling, spiralising, asymmetrically distributed vortices (Dean vortices).

More info

As long as the ratio of the curvature radius (r_c) and tube diameter (D), being r_c/D, is sufficiently large, the flow depends on the value of the parameter De (the Dean number):

$$De = Re\ (D/r_c)^{1/2}. \quad (1)$$

Dean numbers up to 150 hold for the above described flow behavior.
If the Reynolds number is increased sufficiently, the flow will become turbulent. However, the critical Reynolds number for transition is much greater in a curved tube than in a straight one. The transition occurs at critical De numbers, which are r_c/D dependent. Once the flow is turbulent, the dissipation, and hence pressure drop, rises dramatically. However, the fully developed turbulent mean velocity profile is similar to that in laminar flow.

There is no good theory available for large Dean numbers with turbulent flow. With such flow, the velocity profile is blunt and the flow inviscid. At the entrance of the bend, the flow profile, in contrast to that of Poiseuille flow (see Poiseuille's Law), is skewed to the inside. However, the maximum value of the shear rate at the wall, initially on the inside wall, switches over to the outside wall at a distance of only about one diameter from the entrance of the bend. Consequently, the skewed profile flips over to the outside.
At the outside wall of the bend, the rate of energy dissipation by viscosity is high, due to the high velocity gradient. Therefore the pressure gradient required to maintain a given flow rate is also higher than in Poiseuille flow. Since the distortion of the velocity profile increases for increasing Dean number (i.e. for increasing Reynolds number in a given bend), the ratio of actual pressure drop to Poiseuille pressure drop in a given length of tube will also increase.

Analytical calculation of the flow in a bend is very complicated. Here follows a practical technical approach.
The resistance increase due to the curvature can be expressed as the curvature resistance coefficient or loss-factor ζ_c. It is dependent on curvature angle β and the ratio of curvature radius (r_c) and tube diameter (D) r_c/D. Up to β is about 22.5 ° there is no r_c/D dependency. At 45°, normal values in vascular anatomy, a ratio of 10 yields a factor 0.07. In technical handbooks for fluid transport, ζ_c is tabulated for β and curvature ratio.

Literature

Pedley TJ et. al, Gas flow and mixing in the airways, In: West, J.B. (ed.) *Bioengineering Aspects of the Lung*. New York: Marcel Dekker, 1977.

Flow in bifurcations

Principle

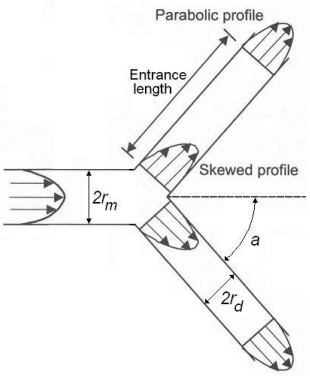


Fig. 1 Bifurcation

As Fig. 1 shows, a bifurcation changes the circular symmetric parabolic fluid flow (of a liquid or gas) in a circular mother tube to a skewed profile in the daughter branches. The radius ratio of mother and

daughter branches, r_d/r_m, is energetically optimal for a ratio of $2^{-1/3} = 0.79$. This gives a daughters/mother area ratio of 1.25. For a symmetric bifurcation a semi-bifurcation angle *a* of 37.5^0 (=arctan($2^{-1/3}$)) is optimal. For this angle the resistance increase is only some 10%.

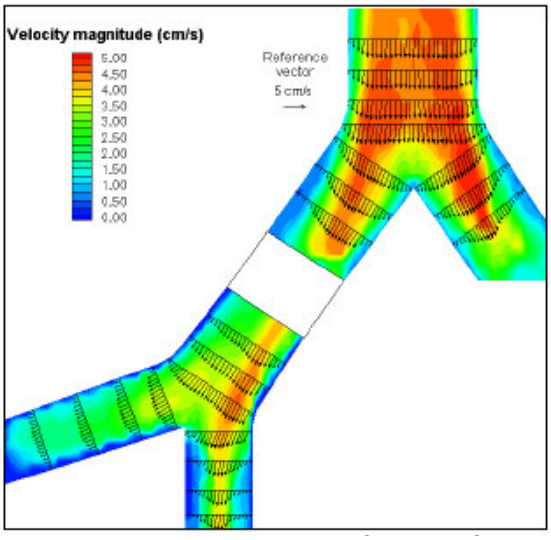


Fig. 2 Model of the flow profile of air in the trachea bifurcation

Application

Vascular and airways flow, and flow in the kidney. They play an important role in the study of the origination and grow of arteriosclerosis and aneurisms.

In the airways tree of the lungs some 26 generations (orders) of bifurcations, from trachea to the most remote alveoli can be distinguished (see Lung gas transport 1: basic principles). In mammals, the bifurcation ratio r_d/r_m, is on average 0.78 and the bifurcation angle 67^o, 8^o less than the optimum. Human midrange generations show higher angles: 79^o. The two semi-bifurcation angles are more similar for higher generations. From the 2^{nd} to the 10^{th} generation the human bifurcation ratio r_d/r_m is very close to 0.79, but for higher generations the ratio slowly increases and the highest generations hardly show a diameter decrease from mother to daughter. The first generation bifurcation shows daughter diameters which are much smaller than the optimum. However, above optimizations holds for Poisseuille flow (see Poisseuille's Law), and this does not hold for the human trachea.
In the vascular system of mammals an averaged value of 0.68 is found for r_d/r_m. In the myocardium, r_d/r_m goes from 0.79 (the optimum) for the capillaries to about 0.72 for arteries of some mm diameter. Nevertheless, the optimization is poor since the variation is very large with many much too small or large daughters. Also the bifurcations are rather asymmetric, going from 0.8 (= $r_{small\ daughter}/r_{large\ daughter}$) of capillaries to about 0.2 of mm-sized arteries. For energetic reasons, asymmetry in diameter is accompanied by asymmetry in the semi-bifurcation angle. The human brain the mean bifurcation angle is 74^o with tens of degrees of variation. Branching angles in muscles can deviate substantial from their fluid dynamic minimum cost optimum. Classically, minimum cost according to Murray's Law is obtained when:

$(r_{d1}/r_m)^3 + (r_{d2}/r_m)^3 = 1$.

With $r_{d1} = r_{d2}$ the smallest minimum (with the ratio 0.79) is obtained. This solution is about the same as that on the basis of calculations with the Womersley number for laminar flow with minimal local wave reflection as criterion. For a given radius ratio, the optimal angles can be calculated.
The experimentally obtained exponents of Murray's law range from 2.5 to 2.8.

More info

Fig. 3 illustrates in more detail the flow patters in a bifurcation.
When the direction of flow is inverted, in the mother tube four secondary loops can arise (Fig. 4). Secondary flows are characterized by a swirling, helical component superimposed on the main streamwise velocity along the tube axis, see also Flow in a bended tube. A complicating factor is wall compliance, especially of importance with pulsatile flow, another factor which makes the pattern more complicated.

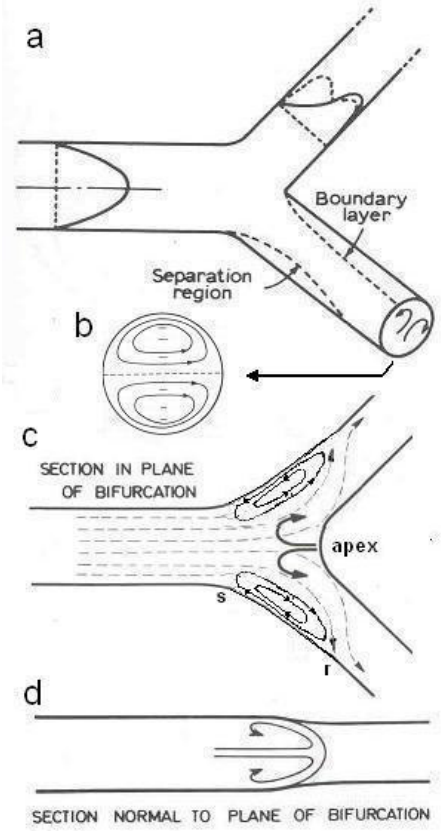


Fig. 3 a. The dashed profile in the upper daughter branch is the velocity profile perpendicular on the plane of the bifurcation. b. the dashed line is in the plane of the bifurcation (the symmetry plane). c. Stream lines with points of separation and reattachment indicated. d Streamline of flow impinging onto the apex.

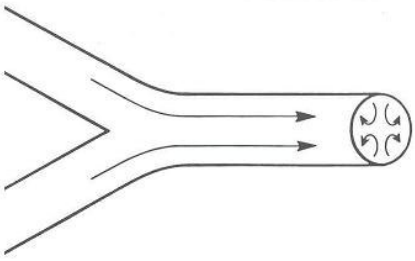


Fig. 4 Inversed flow in a bifurcation.

Literature

Canals M et. al. A simple geometrical pattern for the branching distribution of the bronchial tree, useful to estimate optimality departures. Acta Biotheor. 2004;52:1-16.
Frame MD, Sarelius IH. Energy optimization and bifurcation angles in the microcirculation. Microvasc Res. 1995 Nov;50(3):301-10.
Murray CD. The physiological principle of minimum work. I. The vascular system and the cost of blood volume. Proc Nat Acad Sci 12: 207-214, 1926.
Pedley TJ et. al, Gas flow and mixing in the airways, In: West, J.B. (ed.) *Bioengineering Aspects of the Lung*. New York: Marcel Dekker, 1977.
VanBavel E, Spaan JA. Branching patterns in the porcine coronary arterial tree. Estimation of flow heterogeneity. Circ Res. 1992;71:1200-12.
http://www.vki.ac.be/research/themes/annualsurvey/2002/biological_fluid_ea1603v1.pdf

Flow through a stenosis

Principle

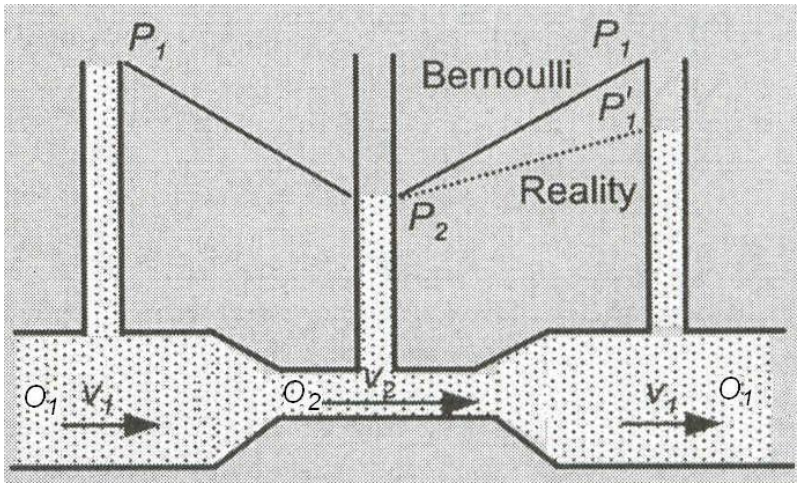


Fig. 1 The pressure after the stenosis P'_1 is smaller than P_1. due to losses at the entrance and exit of the stenosis (see **More Info**). P is pressure, O area and v mean flow velocity.

With incompressible flow of liquid through a stenosis (Fig. 1) in a horizontal tube, after the constriction, according to Bernoulli's law (see Bernoulli's and Pascal's Law) the pressure drops. Since h is constant (horizontal), the gravity term can be neglected and it holds that $\frac{1}{2}\rho v^2 + P = \text{constant}$. Since in the tube the volume flow $\dot{V} = O_1 v_1 = O_2 v_2 = \text{constant}$ and with $O_2 = \beta O_1$, it holds that $v_2 = \beta^{-1} v_1$, and $P_2 = \beta^{-2} P_1$. The constriction can result in a change of the type of flow. For instance, before the constriction it can be laminar, for circular diameters in accordance with Poiseuille's law, and after the constriction slightly or actually turbulent, depending on the Reynolds number . However, since $\dot{V}$ is constant and v applies to the mean flow velocity in the tube, pressure changes are not dependent on the type of flow, in accordance with Bernoulli's law.

Application

In hemodynamics. To maintain the same $\dot{V}$, a stenosis demands a higher driving force. Many stenosis increases substantially energy expenditure, be it cordially or pulmonary and will finally give rise to disorders. The severity of a stenosis is expressed as % area occlusion: $100(1-\beta)\%$.

More info

A constriction increases the tube resistance due to the decreased diameter over the length of the constriction, due to the (sudden) narrowing. For laminar flow this increase is given by Poiseuille's law. To calculate the pressure at the exit of the constriction Bernoulli's law defined as $\frac{1}{2}\rho v^2 + \rho g h + P = \text{constant}$ (ρ is specific density, g gravity constant), is inadequate since it does not account for the entrance effect, and the widening of the constriction. The equation is extended to:

$$\frac{1}{2}\rho v^2 + \rho g h + \Sigma w_i + P = \text{constant}, \qquad (1)$$

where Σw_i is the summed losses in the irregularities in the tube, such as sudden changes in diameter and curvatures. With this extension and elaborating Σw_i it is more common to define Bernouilli's law in terms of equivalent height:

$$\frac{1}{2}v^2/g + h + \Sigma 8\pi l_i v_i \eta / g\rho O_i + \Sigma \frac{1}{2}\zeta_i v_i^2/g + P/\rho g = \text{constant}, \qquad (2)$$

where η is dynamic viscosity. The effects are dependent on the type of flow. The third term accounts for the constriction itself and the tube-approximation of the smooth irregularities of entrance and exit, all with Poiseuille-like flow. l_i is the length of the irregularity and v_i the mean flow velocity. The factor ζ in the fourth term is the resistance coefficient of an irregularity. The effects of narrowing and widening are very strongly dependent on the smoothness of the transition. The range of change of the narrowing is a factor of 10, with a rounded entrance (say angle of narrowing up to 20^o, $\zeta=0.05$ and abrupt, $\zeta=0.45(1-\beta)$, as extremes. Here, β is the contraction coefficient: β = (smaller cross section area)/(larger cross section

area). For a widening $\zeta=(1/\beta - 1)^2$. Then, for a surface ratio of 2 ζ is 1, yielding a pressure drop $\Delta P = -½v^2/g$. This shows that exit effects can be very large. Moreover, it works over a long distance before parabolic laminar behavior is restored (see Flow: entrance effect and entrance length).
A practical equation to calculate the pressure drop over a vascular stenosis with turbulent flow is the approximation:

$$\Delta P = 8\pi l v \eta / O_1 + \zeta v^2 \rho (O_1/O_2 - 1)^2, \qquad (3)$$

where v is the velocity before the stenosis, l the total stenosis length and ζ is about 1.5.

HR_{max}

Principle

HR_{max} is the maximum heart rate (HR) that a person should achieve during maximal physical exertion. Research indicates that it is most closely linked to a person's age; a person's HR_{max} will decline during live. The speed at which it decline over time is related to fitness: the more fit a person is, the more slowly it declines.
HR_{max} is measured during periodically increase of the intensity of exercise (for instance, when a treadmill is being used by increase in speed or slope of the treadmill, or by interval training) until the subject can no longer continue, or until certain changes in heart function are detected in the ECG (at which point the subject is directed to stop). Typical durations of such a test range from 10 to 20 minutes. There are many prediction equations for HR_{max}. One of the most precise prediction equations, published in a meta-analysis (ref. 1), is:

$$HR_{max} = 208 - 0.7 \times \text{age (beats/min)}, \quad (1)$$

where age is in years.
It is independent of sex.
It is slightly too high for people performing frequently endurance sport (a decrease of about ¾ point per hour endurance sport/week with a maximum of 10 points). The equation yields probably too high values for people older than 65 years. (A reduction of 1 point for every year above 65 seems likely). People who have participated in sports and athletic activities in early years will have a higher HR_{max} than those less active as children.

Application

HR_{max} is utilized frequently in the fitness industry, specifically during the calculation of target heart rate (ref. 2) when prescribing a fitness regime.

Literature

1 Tanaka H, Monahan Kevin D and. Seals DR Age-predicted maximal heart rate revisited2001J Am Coll Cardiol 37, 153-156.
2 Wikipedia
3 McArdle, WD., Katch FI, Katch VL, , Exercise Physiology: Energy, Nutrition & Human Performance, Lippincott Williams and Wilkins, 2001.

Navier-Stokes equations

Principle

The Navier-Stokes equations are a set of equations that describe the motion of fluids (liquids and gases, and even solids of geological sizes and time-scales). These equations establish that changes in momentum (mass x speed) of the particles of a fluid are simply the product of changes in pressure and dissipative viscous forces (friction) acting *inside* the fluid. These viscous forces originate in molecular interactions and dictate how *sticky* (viscous) a fluid is. Thus, the Navier-Stokes equations are a dynamical statement of the balance of forces acting at any given region of the fluid.
They are one of the most useful sets of equations because they describe the physics of a large number of phenomena of academic and economic interest. They are useful to model weather, ocean currents (climate), water flow in a pipe, motion of stars inside a galaxy, flow around a wing of an aircraft. They are also used in the design of aircraft and cars, the *study of blood flow*, the design of power stations, the analysis of the effects of pollution, etc.
The Navier-Stokes equations are partial differential equations which describe the motion of a fluid, so they focus on the *rates of change* or *fluxes* of these quantities. In mathematical terms these rates correspond to their derivatives. Thus, the Navier-Stokes for the simplest case of an ideal fluid (i.e. incompressible) with zero viscosity states that acceleration (the rate of change of velocity) is proportional to the derivative of internal pressure. Poiseuille's Law and Bernoulli's equation are special cases of 1D Navier-Stokes.
The fluid motion is described in 3D space, and densities and viscosities may be different for the 3 dimensions, may vary in space and time. Since the flow can be laminar as well as turbulent, the mathematics to describe the system is highly complicated. In practice only the simplest cases can be solved and their exact solution is known. These cases often involve non turbulent flow in steady state (flow does not change with time) in which the viscosity of the fluid is large or its velocity is small (small Reynolds number).
For more complicated situations, solution of the Navier-Stokes equations must be found with the help of numeric computer models, here called computational fluid dynamics.
Even though turbulence is an everyday experience it is extremely hard to find solutions for this class of problems. Often analytic solutions cannot be found.
Reducing the models to 1D, as is often done in fluid dynamics of blood vessels, makes the problem handsome, see e.g. http://www.math.ist.utl.pt/~mpf2005/abstracts/contributed_moura_vergara.pdf and ftp://ftp.inria.fr/INRIA/publication/publi-pdf/RR/RR-5052.pdf

Application

Theoretical medical applications in the vascular and airways system, the latter with compressible fluid dynamics.

More Info

Those who like to avoid e.g. the mathematical comprehensive Wikipedia paper on the Navier-Stokes equations can read the more friendly paper on ''Incompressible Navier-Stokes equations reduce to Bernoulli's Law'', http://home.usit.net/~cmdaven/navier.htm.
In this paper the incompressible Navier-Stokes vector-form equation, a nonlinear partial differential equation of second order (in dimensionless variables):

$$v_t + (v\cdot\nabla)v = -\nabla p + \mu\nabla^2 v + \rho g. \qquad (1)$$

(where v is a vector representing the velocity of an infinitesimal element of mass at a point in 3-D space, p is the scalar pressure at the same point, ρ is the mass density at the point and is assumed constant throughout the medium, μ is the viscosity of the medium, and g is a constant vector acceleration due to some constant external force on the infinitesimal element, usually taken to be gravity) can finally be reduced to Bernoulli's Law in a 4D (3D and time) vector form:

$$\tfrac{1}{2}\rho V^2 = P + \rho g Z + \text{Const} \qquad (2)$$

(where V is the analytic 4-D velocity, P is the 4D analytic vector pressure field, g is a constant acceleration which can be imposed in an arbitrary direction, and Z is a vector representing arbitrary displacement in 4-D space). By reducing this equation to one spatial dimension and taking time-invariant flow the traditional scalar Bernoulli's equation is obtained.

Modified after Wikipedia and http://home.usit.net/~cmdaven/navier.htm.

Osmosis

Principle

Osmosis is the net movement of solvent (mostly water) across a membrane permeable to the solvent, but not the solute from a region of high solvent potential (concentration) to an area of low solvent potential. Hence, the transport is from the less-concentrated (hypotonic), to the more-concentrated (hypertonic) solution, which tends to reduce the difference in concentrations. It is a physical process without input of energy. Osmosis releases energy, and can be made to do work, as when a growing tree-root or swelling piece of wet wood splits a stone.

The process effect can be countered by increasing the pressure of the hypertonic solution, with respect to the hypotonic. The osmotic pressure is defined to be the pressure required to maintain equilibrium, with no net movement of solvent. Osmotic pressure is a colligative property, meaning that the property depends on the molar concentration of the solute but not on its identity.
Osmosis, being the result of diffusion across a semi-permeable membrane is important in biological systems as many biological membranes are semipermeable. In general, these membranes are impermeable to organic solutes with large molecules, such as polysaccharides, while permeable to water and small, uncharged solutes. Permeability may depend on solubility properties, charge, or chemistry as well as solute size. Water molecules travel through the plasma cell membrane, tonoplast (vacuole) or protoplast in two ways. The one is by diffusing across the phospholipid bilayer directly and the other is via aquaporins (small transmembrane proteins similar to those in facilitated diffusion and in creating ion channels). Osmosis provides the primary means by which water is transported into and out of cells. The turgor pressure of a plant cell is largely maintained by osmosis, across the cell membrane, between the cell interior and its relatively hypotonic environment.

Consider a permeable membrane with apertures small enough to allow water (solvent) molecules, but not larger solute molecules, to pass through. When this membrane is immersed in liquid it is constantly hit by molecules of the liquid, in motion due to their thermal kinetic energy. In this respect, solute and solvent molecules are indistinguishable. At a molecular scale, every time a molecule hits the membrane it has a defined likelihood of passing through. Here, there is a difference: for water molecules this probability is non-zero; for solute molecules it is zero.
Suppose the membrane is in a volume of pure water. In this case, since the circumstances on both sides of the membrane are equivalent, water molecules pass in each direction at the same rate; there is no net flow of water through the membrane.
If there is a solution on one side, and pure water on the other, the membrane is still hit by molecules from both sides at the same rate. However, some of the molecules hitting the membrane from the solution side will be solute molecules, and these will not pass through the membrane. So water molecules pass through the membrane from this side at a slower rate. This will result in a net flow of water to the side with the solution. Assuming the membrane does not break, this net flow will slow and finally stop as the pressure on the solution side becomes such that the movement in each direction is equal: dynamic equilibrium. This could either be due to the water potential on both sides of the membrane being the same, or due to osmosis being inhibited by pressure.
The osmotic pressure p (often written by π or Π) is given by van 't Hoff's law:

$$p = cRT = \rho h, \qquad (1)$$

where c is the molar concentration (= n/V with V the volume and n number of moles) of the solute and h the equivalent hydrostatic height difference (see Fig. 1). c is supposed to be much smaller than the solvent molar concentration), R the gas constant, T absolute temperature. Equation (1) is identical to the pressure formula of an ideal gas. To make it more precise, it should however be corrected for the volume occupied by the solute (see **More Info**).

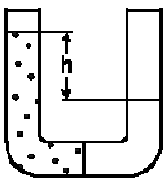


Fig. 2 Visualization of the osmotic pressure by the hydrostatic pressure difference: $p = \rho h$. ρ the specific density of the solvent (generally water) and h the difference in height of the menisci.

Reverse osmosis

The osmosis process can be driven in reverse against the concentration gradient of the solute by applying a pressure in excess of the osmotic pressure.

Forward osmosis
Osmosis may be used directly to achieve separation of water from a "feed" solution containing unwanted solutes. A "draw" solution of higher osmotic pressure than the feed solution is used to induce a net flow of water through a semi-permeable membrane, such that the feed solution becomes concentrated as the draw solution becomes dilute. The diluted draw solution may then be used directly (as with an ingestible solute like glucose), or sent to a secondary separation process for the removal of the draw solute. This secondary separation can be more efficient than a reverse osmosis process would be alone, depending on the draw solute used and the feed-water treated.

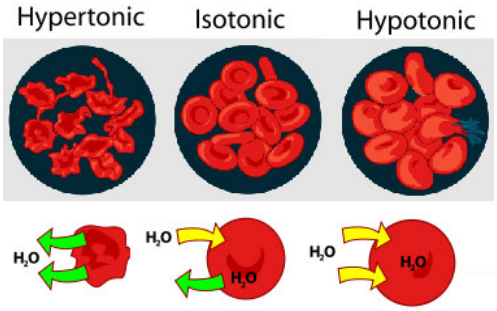


Fig. 2 Effect of different solutions on blood cells

Example of cost to desalinate sea water
With R = 8.315 (J/(K·mol), and T = 300 K. The amount of salt in seawater is about 33 g/L. NaCl (molecular weight 23+35.5= 58.5) is the dominant solute, completely ionized, so the particle concentration is double the molar concentration. This yields c = 1128 mol/m^3, and an osmotic pressure:
$p = 1128 \cdot 8.315 \cdot 300 = 28.1 \cdot 10^5$ Pa = 28.1 bar (28.1 kg/cm^2).
In order to push v liters of solvent through a membrane area A of 1 cm^2, the equivalent distance d of travel is 10 m. The work E needed is:
$E = Fd = pAd = pv.$ (2)
And consequently for v is 1 liter (= 10^{-3} m^3) E becomes:
$E = 28.1 \cdot 10^5 \cdot 10^{-3} = 2810$ Nm/L = 278 kg·m/L.
or, 2780 J/L = 0.66 kcal/L. This is the *theoretical minimum* cost to desalinate sea water. Boiling 1 L of water (20 ^{0}C) requires about 620 kcal, about 1000 times more.

Application

In unusual environments, osmosis can be very harmful to organisms. For example, freshwater and saltwater aquarium fish placed in water of a different salinity than that they are adapted to will die quickly, and in the case of saltwater fish, rather dramatically. Another example of a harmful osmotic effect, but often practiced, is the use of table salt to kill leeches and slugs.

- If the medium is hypotonic the cell will gain water through osmosis and will swell.
- If the medium is isotonic there will be no net movement of water across the cell membrane.
- If the medium is hypertonic the cell will lose water by osmosis and so shrinks.

The reverse osmosis technique is commonly applied in desalination, water purification, water treatment, and food processing. Recent advances in pressure exchange and the ongoing development of low pressure membranes have significantly reduced the costs of water purified by reverse osmosis. Forward osmosis is an area of ongoing research, focusing on applications in desalination, water purification, water treatment, food processing, etc.

More Info

Osmosis can also be explained via the notion of entropy (see: Thermodynamics: entropy), from statistical mechanics. As above, suppose a permeable membrane separates equal amounts of pure solvent and a solution. Since a solution possesses more entropy than pure solvent, the second law of thermodynamics states that solvent molecules will flow into the solution until the entropy of the combined system is maximized. Notice that, as this happens, the solvent loses entropy while the solution gains entropy. Equilibrium, hence maximum entropy, is achieved when the entropy gradient becomes zero.

Osmotic pressure

As mentioned before, osmosis may be opposed by increasing the pressure in the hypertonic region. A more precise equation of osmotic pressure is:

$$pV = -RT\cdot\ln(1-n) \quad (3)$$

Equation (1) is an approximation under the condition that $c = n/V$ is very small. Then it holds that $-\ln(1-n) \approx n$ and equation (1) follows.
When the solute concentration c_2 is not very small compared to the solvent concentration c_1, equation (1) does not well hold. A good approximation, taking gas vapor pressures into account is:

$$p = (RT/V_1)\cdot\ln(p^0_1/p_1), \quad (4)$$

where V_1 is the pure molar volume of the solvent, p^0_1 and p_1, the vapor pressures of pure solvent and the solvent in the solution, respectively, all at the same P and T. Unfortunately, p_1 is not easy to obtain. Another approximation, assuming the mixture is still ideal, the solute is a non-electrolyte and the so called activity coefficient of the solvent γ_1 is 1, the concentration of pure solvent c_1 and of the solute c_2 should be taken into account (see ref. 1):

$$p = -(RT/V_1)\cdot\ln(c_1,/(c_1+ c_2)). \quad (5)$$

This prevent to high, non empirical pressures. See for other approximations ref. 2.

Literature

1. Biomedical Microdevices, 2002, 4:4, 309-321
2. Am J Physiol Regulatory Integrative Comp Physiol 237:95-107, 1979.

Pitot tube

Basic principle

A Pitot tube is an instrument to measure flow.

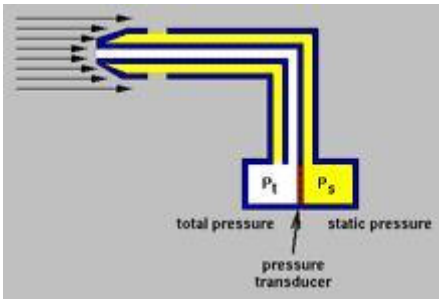


Fig. 1 Pitot tube

The basic instrument, see Fig. 1, consists of two coaxial tubes: the interior tube is open to the flow (i.e. opening perpendicular), while the exterior tube is open at ninety degrees to the flow (i.e. opening parallel). Manometers are used to measure the difference between these two pressures and using Bernoulli's equation the flow rate of the fluid or gas can be calculated.

Application

Pitot tubes are applied in the pulmonary pneumotachograph. Also it is applied in ultrasound flow measurement and fluid flow. Further it is widely applied in technology, like measuring relative velocity (not ground velocity) of a plane (ρ can be calculated by measuring P_{static} and outside temperature)

More Info

The flow field outside the Pitot tube is everywhere homogeneous except around the entrance of the interior tube. In accordance with the law of Bernoulli (see Bernoulli's and Pascal's Law), in the flow field, so at al all stream lines at all positions, it holds that $½\rho v^2 + P = \text{constant}$. The term ρgh is omitted since a difference in height can be neglected. At the entrance the law also holds but there is no flow within the internal tube. The velocity pressure or propelling pressure (caused by the force of the fluid which tries to flow into the closed tube interior), being $½\rho v^2$, is added to P_{static}, and so:

$$P_{entrance} = ½\rho v^2 + P_{static}. \quad (1)$$

$P_{entrance}$, the total pressure is also called the stagnation pressure. In the exterior tube, with an opening parallel to the flow, there is also no flow, so no velocity pressure. Here, only the P_{static} will be registered. $P_{entrance} - P_{static}$ is measured by the pressure transducer. When ρ is known v can be calculated.
.

Poiseuille's Law

Basic Principle

A laminar flow within a pipe or tube, in medicine for instance a blood vessel or a tube of the airways system, has a velocity which increases with the distance from the wall of the tube, provided that the viscosity is everywhere the same. This increase is parabolic with the maximum of the parabola in the centre of the tube and consequently the velocity is there maximal (Fig. 1).

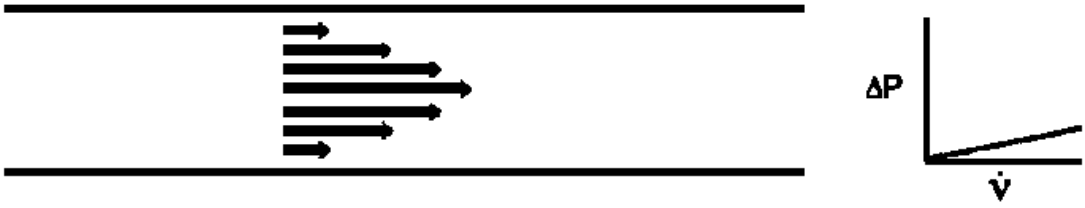


Fig. 1 Laminar flow in accordance with Poiseuille's Law.

Laminar flow is dependent on the viscosity of the gas, but not on its density (ρ). The quantity relating the pressure drop Δp over a certain length of the tube and the stationary volume flow $\dot{V}$ (in m^3/s) is the resistance R of the tube, analogue to the electric law of Ohm (V= i·R). It is given by:

$$\Delta p = \dot{V} \cdot R_{tube}, \qquad (1)$$

with Δp (Pa) the pressure difference over the length (m) of the tube. Δp is the driving 'force' to create flow. R_{tube} of an ideal tube with radius r (m) and length L is:

$$R_{tube} = 8 \cdot \eta \cdot L/(\pi \cdot r^4) \qquad \text{(in Pa/ } m^3\text{)}. \qquad (2)$$

where η the dynamic viscosity (for air 17.1 x 10^{-6} and 19.0 x 10^{-6} Pa·s at 0 and 37 °C respectively, blood ca. 1 x 10^{-3} Pa·s). For a formal derivation of the law see the textbooks of physics.

Substituting 2) in 1) gives the law of Poiseuille:

$$\dot{V} = \pi \cdot \Delta p_{tube} \cdot r^4/(8 \cdot \eta \cdot L), \qquad (3)$$

The flow, as a function of the distance from the centre, the axis of the tube, is parabolic as illustrated in Fig. 1.

Application

Applied in the calculation of flow of blood through the vessels or heart (fluid mechanics of cardiovascular system) and the flow of air and expiratory gas through the airways (see Lung gas transport 1-3).

Some numeric examples
With extreme exercise (heart rate 200 beats/min, stroke volume 100 mL) the mean velocity in the aorta (diameter 3 cm), supposing laminar flow would be 0.47 m/s (or 1.7 km/h). At the axis, the velocity with laminar parabolic flow is exactly twice as much. With 0.47 m/s the current is in between laminar and full grown turbulent since the Reynolds number is about 4200. In the main arterial branches (0.1 cm diameter, with in total 3000 within the human body), Reynolds number is about 5, and consequently the flow is laminar.
A similar calculation holds for the trachea. With extreme exercise, the inspiratory volume of a top-endurance athlete is 200 L/min. With a diameter of the trachea of 2 cm, the peak maximal velocity at the axis supposing laminar flow would be 42.4 m/s (supposing that the duration of inspiration and expiration are equal with a square wave profile in time). However, with a mean velocity of 21.2 m/s this means that the Reynolds number is about 16500, and so turbulent and 8 times larger than laminar flows allows.

More Info

Bifurcations increase the resistance (see Flow in bifurcations). With an optimal angle and laminar flow, this is about 10%. It is caused by the change from an axial symmetric parabolic current profile to an

asymmetric laminar profile. The resistance increase can change the character of the flow: from laminar before the bifurcation to transitional after the bifurcation, and from transitional to turbulent. These so called entrance effects fade away after some distance (see Flow: entrance effect and entrance length and also Lung gas transport 1, basic principles). So, with laminar, symmetric flow before the bifurcation, the profile is again axis symmetric after a distance of some five tube diameters
After bifurcations of arteries or arterioles, the susceptibility for plaque formation is substantially increased when the flow is not anymore laminar.
Constrictions, like that of the glottis or a stenosis (airways and blood vessels) also can change the character of the flow (see Flow through a stenosis).

Literature

Van Oosterom, A and Oostendorp, T.F. Medische Fysica, 2nd edition, Elsevier gezondheidszorg, Maarssen, 2001.

Rayleigh, Grashof and Prandtl Numbers

Principle

These three numbers have to do with heat transport by natural *convection*, laminar or turbulent, and not with heat conduction. They do not comprise current velocity. For forced convection (e.g. in case of wind), laminar or turbulent, these numbers are irrelevant. Then, Reynolds number and the Nusselt number are applied. For the latter, the reader is referred to technical textbooks.

Rayleigh number

In fluid mechanics (gases and liquids), the Rayleigh number is a dimensionless number associated with the heat transfer within the fluid. When the Rayleigh number is below the critical value for that fluid, heat transfer is primary in the form of conduction when it exceeds the critical value; heat transfer is primarily in the form of convection. The latter can be laminar or turbulent.

Grashof number

In natural convection the Grashof number plays the same role that is played by the Reynolds number in forced convection. The buoyant forces are competing with viscous forces and at some point they overcome the viscous forces and the flow is no longer nice and laminar.

Prandtl Number

The Prandtl number is from a conceptual point of view the ratio of the thickness of the velocity boundary layer to thermal boundary layer. The former is the layer where the (macroscopic) particle velocity in the moving medium (e.g. air) varies from zero (at the surface of the e.g. the body) to that of the moving medium (the speed of air surrounding the body). The thermal boundary layer is the layer which comprises the full difference in temperature. When Pr=1, the boundary layers coincide.
Typical values of the Prandtl number are:

Material	Pr
Gases	0.7-1.0
Air 20 ^{o}C	0.71
Water	1.7-13.7
Oils	50-100,000

When Pr is small, it means that heat diffuses very quickly compared to the velocity (momentum). This means the thickness of the thermal boundary layer is much bigger than the velocity boundary layer.

Application

Medical applications are rare since heat transport by natural convection is seldom of interest. It can play a role in aerospace, environmental and sports medicine.

More info

The *Rayleigh number* is the product of Grashof which describes the relationship between buoyancy and viscosity within a fluid and Prandtl number, which describes the relationship between momentum diffusivity and thermal diffusivity.
For free convection near a vertical smooth surface, the Rayleigh number is:

$$Ra = Gr \cdot \mathrm{Pr} = \frac{g \cdot \beta}{\nu \cdot \alpha} \Delta T \cdot L^3 \qquad (1)$$

where
Ra = Rayleigh number
Gr = Grashof number
Pr = Prandtl number
g = gravity constant (N/kg)
β = thermal expansion coefficient (1/K)
ΔT = temperature difference between surface and quiescent temperature (K)
L = characteristic length (mostly effective height, for a plate it is the actual height) (m).
v = kinematic viscosity = η/ρ (= dynamic viscosity/density) (m^2/s)
α = thermal diffusivity = λ/(ρ·c_p) where λ is the heat conduction coefficient (W/(m·K), ρ the specific density (kg/m^3) and c_p the specific heat coefficient at constant pressure (J/(kg·K).

Free convection over a horizontal surface (plate) also the Nusselt number is relevant (the reader is referred to technical textbooks).

The *Grashof number* is defined as:

$$Gr = \frac{BuoyancyForces}{ViscousForces} = \frac{g \cdot \beta \cdot \Delta T \cdot L^3}{\nu^2} \qquad (2)$$

For air of 20 °C, Gr = $78 \cdot 10^9 \cdot \Delta T \cdot L^3$. For a vertical plate, the flow transitions to turbulent around a Grashof number of 10^9.

The *Prandtl number*, which is specially applied for heat transfer and which comprises some fluid properties, is defined as:

$$\mathrm{Pr} = \frac{\nu}{\alpha} = \frac{\eta \cdot c_p}{\lambda} \qquad (3)$$

The Prandtl number is the ratio of momentum diffusivity (kinematic viscosity) to thermal diffusivity.

Literature
http://www.coolingzone.com/Content/Library/Tutorials/Tutorial%201/DNHT.html

Reynolds Number

Principle

The Reynolds number indicates whether the mass transport trough a tube (blood vessel, airways of ventilatory system, etc.) is laminar, for instance parabolic according to Poiseuille's law (Fig. 1a; see *Poiseuille's Law*) or turbulent (Fig. 1b).

The Reynolds number is defined as:

$$Re = 2v \cdot r \cdot \rho/\eta = 2\,\dot{V} \cdot \rho/(\pi \cdot r \cdot \eta) \qquad (1)$$

with v the mean velocity (m/s) in the tube, r the characteristic (see **More Info**, for circular diameters the radius) tube radius (m), ρ density (kg/m^3), η dynamic viscosity (Pa·s) and $\dot{V}$ the volume flow (m^3/s).

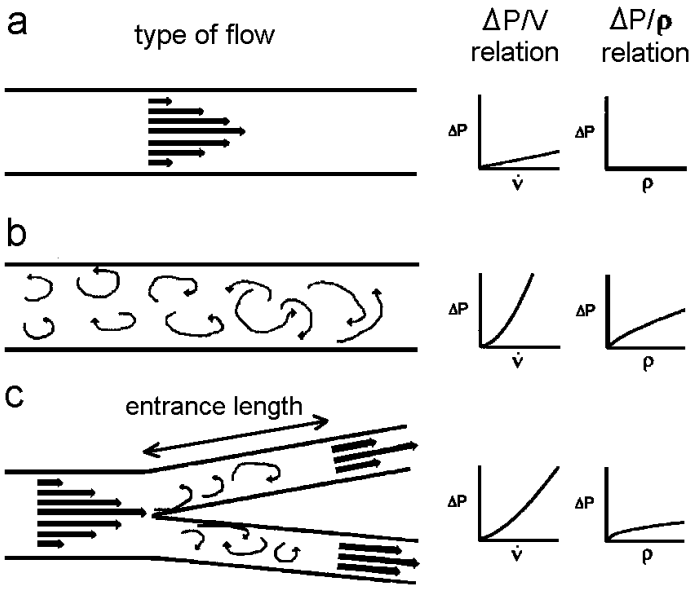


Fig. 1 Types of flow. See **Principle** and **More Info**.

Depending on the Reynolds number, three states of flow can be distinguished. When Re<2000, than the flow is laminar (Fig. 1a) and when Re>10000 the flow is turbulent (Fig. 1b). In between, there is a transitional state. There is a 4th state that occurs at the orifice of a tube in some cavity or after a bifurcation (Fig. 1c).
Fig. 1 visualizes the various types of flow.

a) represents laminar flow, here parabolic according to Poiseuille's law. The $\Delta p/\dot{V}$ relation is a strait line with the slope equal to the resistance of the tube, R_{tube} (see Poiseuille's Law).

b) presents turbulent flow. The $\Delta p/\dot{V}$ relation is a curved line and the resistance of the tube increases (nearly) quadratic with the flow. An increase in the specific density ρ of the medium (as can hold for an inspiring gas) asks for a higher Δp, the driving force of the flow, to obtain the same flow as with a lower density. This can easily be seen from the $\Delta p/\dot{V}$ relation and Δp/ρ relation. The velocity in the tube is nearly independent of the distance to the wall, so by approximation rectangular (at the wall itself, the velocity is zero). With transitional flow, the velocity is in between a parabola and a rectangle.

c) represents the flow in case of a bifurcation. The $\Delta p/\dot{V}$ and Δp/ρ relation is in between those of laminar and turbulent flow. When the turbulent flow after the bifurcation changes to a laminar flow, this laminar flow starts with an axis-asymmetry before becoming axis-symmetric. This asymmetry increases the resistance only some 10% (see Flow in bifurcations). When the flow is very low, the flow remains laminar along the total entrance length (see Flow: entrance effect and entrance length).
Many papers etc. address and model mathematically the flow in the vascular and airways system, also for bifurcations.

Application

Reynolds numbers are used to characterize the flow of blood through the vessels (fluid mechanics of cardiovascular system) and the flow of air and expiratory gas through the airways (see the various chapters about Lung Mechanics). Laminar flow can become turbulent after a stenosis. This substantially increases the resistance, for instance of the vascular system and so increases the load of the heart muscle.
When the diameter of a vessel with Poiseuille flow increases 19% with the same driving pressure, flow resistance R halves ($R \sim r^{-4}$; see Poiseuille's Law), $\dot{V}$ doubles and so does the Reynolds number. With this increase in diameter, the transition from laminar to turbulent occurs at twice the original volume flow. Such a diameter increase can easily be established in arterioles. This point is further addressed in Blood flow .

More info

Reynolds number may be interpreted as the ratio of two forces that influence the behavior of fluid flow in the boundary layer. Reynolds number also play a rol when a fluid moves over or around an object. For simplicity we will restrict to a plate with characteristic length L. The two forces are the inertia forces and viscous forces:

$$\mathrm{Re} = \frac{InertialForces}{ViscousForces} \approx \frac{\rho \cdot v^2 / L}{\eta \cdot v / L^2} = \frac{\rho \cdot v \cdot L}{\eta} \qquad (2)$$

When the Reynolds number is large, the inertia forces are in command. Viscous forces dominate the boundary layer when the Reynolds number is small. When the number is low enough the flow over the whole plate is laminar, even when there are tiny local disturbances. As the inertia forces get bigger, the viscosity can no longer maintain order and these tiny disturbances grow into trouble makers and there is a transition to turbulent flow.
Another important quantity of the boundary layer that is influenced by the Reynolds number is its thickness. As the Reynolds number increases, the viscous layer gets squeezed into a smaller distance from the surface.
The value of Reynolds number beyond which the flow is no longer considered laminar is called the *critical* Reynolds number.
Calculation of the Reynolds number is easy as long as you identify the characteristic length and use the right velocity and a consistent set of units. For flow over a flat plate, the characteristic length is the length of the plate and the characteristic velocity is the free stream velocity. For *tubes* the *characteristic length is the tube diameter* and the *characteristic velocity is the average velocity* through the tube obtained by dividing the volumetric flow rate by the cross-sectional area (we are assuming that the tube is full, of course). For tubes with a non-circular cross-section, the characteristic length is the hydraulic or wetted diameter defined as:

$$L=4A/C, \qquad (3)$$

where A is the cross-sectional area of the tube and *C* is the wetted perimeter.
For an ellipse with a and b the half long and short axes, a good approximation (that of Ramanujan) of the circumference or perimeter *C* is:

$$C = \pi[3(a+b) - ((3a+b)(a+3b))^{0.5}] \qquad (4)$$

It can easily be verified that for a circular tube the hydraulic diameter equals the tube diameter. For non-circular tubes the average velocity is used as the characteristic velocity. Suppose that a vessel or airway tube is compressed to half its diameter, whereas the perimeter remains the same and the new shape is elliptical, than the area is decreased by 16%. When the volume flow remains the same, then according to (1) the Reynolds number increases by a 16%.
With turbulent (gas)flow, tube resistance $R_{tube,turbulent}$ is proportional with:

$$R_{tube,turbulent} \sim L^{-4.25} \qquad (5)$$

Consequently, $R_{tube,turbulent}$ is increased by a factor 2.086.

The situation gets complicated when a tube system has many velocity and length scales such as the flow inside the vascular system or the airways system (see Lung gas transport, airways resistance and compliance).

Some typical values of Reynolds number:

Spermatozoa ~ 1×10^{-4} (until Re = 0.1 flow is laminar around smooth particles)
Blood flow in brain ~ 1×10^{2}
Blood flow in aorta ~ 1×10^{3}
Onset of turbulent flow ~ 2.3×10^{3} (pipe flows) - 10^{6} (boundary layers)
Person swimming ~ 4×10^{6}
Aircraft ~ 1×10^{7}
Blue Whale ~ 3×10^{8}

Here are some numerical values of dynamic viscosities of liquids and air:

η_{air} $17.1\cdot10^{-6}$ at 0^{o} C (Pa·s)
η_{water} $1002\cdot10^{-6}$ ·at 20^{o} C (Pa·s)

η_{blood}	$2700 \cdot 10^{-6}$·at 37° C, hematocrit ca. 40% (Pa·s), strongly dependent on hematocrit (with pathological values such as 60-70% ca. 10 mPa·s. Notice that blood is non-Newtonian fluid resulting in a stress dependent η)
$\eta_{blood\ plasma}$	ca. $1500 \cdot 10^{-6}$·at 37 °C (Notice that blood is non-Newtonian fluid resulting in a stress dependent η)
ρ_{air}	1.29 kg/m^3 times 273/T (according to the Gas laws) with the temperature in Kelvin
ρ_{water}	998 kg/m^3 at 20 °C (for physiological temperatures independent of absolute temperature T).

Stokes' law and hematocrit

Principle

Stokes' law describes the motion of a small spherical object in a viscous fluid. The particle is subjected to the downward directed force F_g of gravity and the buoyant force F_A, also known as the force exerted by Archimedes' law. The both forces are:

$$F_g = (4/3)\pi r^3 \rho_p g \quad (1a)$$
$$F_A = (4/3)\pi r^3 \rho_f g, \quad (1b)$$

where r is the radius (the Stokes radius, see **More Info**) and *g* is the acceleration due to gravity (9.8 m/s^2). The resulting force is:

$$F_r = F_g - F_A = (4/3)\pi r^3 g(\rho_p - \rho_f.) \quad (2)$$

The resulting force is directed downward when the specific density of the particle ρ_p is higher than that of the fluid ρ_f (e.g. a glass bead) and upward when it is smaller (a polystyrene sphere). Here it is supposed that the former case holds, and so the sphere will move downward.

As soon as the sphere starts moving there is a third force, the frictional force F_f of the fluid. Its direction is opposite to the direction of motion. The total resulting force is:

$$F_{tot} = F_r - F_f. \quad (3)$$

As long as F_{tot} is positive, the velocity increases. However, F_f is dependent on the velocity. Over a large range of velocities the frictional force is proportional to the velocity (v):

$$F_f = 6\pi r \eta v \quad (4).$$

where η is the dynamic fluid viscosity. Expression (4) is Stokes' law.
After some time the velocity does not increase anymore but becomes constant. Then equilibrium is reached. In other words, F_r is canceled by F_f and so $F_{tot} = 0$. From now on the particle has a constant velocity. The equilibrium or setting velocity v_s can be calculated from (2), (3) and (4) with $F_{tot} = F_r, - F_f = 0$. The result is:

$$v_s = (2/9)r^2 g(\rho_p - \rho_f)/\eta, \quad (5).$$

The proportionality of v_s with r^2 means that doubling the radius gives a reduction of a factor of four for the setting time.
Equation (5) only holds under ideal conditions, such as a very large fluid medium, a very smooth surface of the sphere and a small radius.

Application

The law has many applications in science, e.g. in earth science where measurement of the setting time gives the radius of soil particles.

Blood cells
In medicine, a well known application is the precipitation of blood cells. After setting, on the bottom of a hematocrit tube are the red cells, the erythrocytes, since they are large and have the highest density. In the middle are the white cells, the leucocytes despite their often larger volume. However, they are less dense and especially less smooth, which slows their speed of setting. On top, and hardly visible, is a thin band of the much smaller platelets, the thrombocytes. The relative height of the band (cylinder) with the red cells is the hematocrit. Although the red cells are not spherical and the medium is not large at a

all (a narrow hematocrit microtube) the process still behaves rather well according to the law of Stokes. Red cells can clotter to money rolls, which sett faster (of importance for the hematologist).
The equivalent radius of a red blood cell can be calculated from the various parameters: heamatocrit (assume 0.45 L/L), setting velocity (0.003 m/hour = $0.83 \cdot 10^{-6}$ m/s), density of red cells (1120 kg/m^3) and plasma (1000 (kg/m^3). Everything filled out in eq. (5) yields r = 3.5 µm. Actually the red cell is disk-shaped with a radius of about 3.75 µm and a thickness of 2 µm. When the volume is by approximated by a cylinder (the inner thickness is less, but the edge is rounded), the equivalent sphere diameter is 2.76 µm. The too large estimate from Stoke's law is due to the strongly deviating shape of the red cell from the sphere, causing a lower setting velocity.

Centrifugation
An important application is the process of centrifugation of a biochemical sample. The centrifuge is used to shorten substantially the setting time. In this way proteins and even smaller particles can be harvested, such as radio nucleotides (enrichment of certain isotopes of uranium in an ultracentrifuge). With centrifugation, the same equations hold, but the force of gravity g should be replaced by the centrifugal acceleration *a*: $a = 4\pi^2 f^2 R$, where f is the number of rotations/s and R the radius of the centrifuge (the distance of the bottom of the tube to the center). In biochemistry, *a* can easily reach 10^4 g and in physics even 10^6 g.

More Info

More formally, Stokes' law is an expression for the frictional force exerted on spherical objects with very small Reynolds numbers (e.g., very small particles) in a continuous fluid with some viscous by solving the small fluid-mass limit of the generally unsolvable Navier-Stokes equations.
The Reynolds number holds for a fluid (liquid or gas) flowing in a pipe, channel etc., but also for an object flowing (blood cell, air craft) in a fluid. This directly follows from the definition of the Reynolds number:

$$\text{Re} = \text{inertial force/frictional force} = \rho v_s L/\eta = v_s L/\nu, \qquad (6)$$

where v_s is velocity, L the characteristic length and ν the kinematical fluid viscosity ($\nu = \eta/\rho$).
With volume flow of a fluid, v_s is the mean flow velocity and for motion of an object in a fluid it is the velocity with respect to the fluid.
For flow in a circular tube, L is the diameter, and for a spherical object it is $2r_h$. The radius r_h is the Stokes radius or hydrodynamic radius. When a hard object is spherical it is (practically) its actual radius. More precisely, the hydrodynamic radius comprises solvent (hydro) and shape (dynamic) effects. It is in between half the largest diameter (the effective or rotational radius) and the equivalent radius, the radius of a sphere with the same volume and weight. Knowing r_h, the diffusion coefficient (D) in the fluid can be calculated. However, calculating r_h itself is not possible.
Going back to the Reynolds number, it appears that Stokes' law holds rather well when Re of the object (particle) is less than or of order 1.

When an object falls from rest, its velocity v(t) is:

$$v(t) = \frac{mg}{b}(1 - e^{-bt/m}) \qquad (7)$$

where *b* the drag coefficient. v(t) asymptotically approaches the terminal velocity $v_t = mg/b$. For a certain *b*, heavier objects fall faster.
Stokes law assumes a low velocity of a small falling particle. Stokes drag has a coefficient b equal to $b = 6\pi r\eta$. This is the coefficient used in equation (4). The derivation of b is easy for the parameters r and η, but the problem is the factor 6π. A factor 2 π is caused by a pressure effect and a factor 4π by friction, (For a derivation see ref. 1 or 2).
For example, consider a small sphere with radius *r* = 1 µm moving through water at a velocity v of 10 µm/s. Using 10^{-3} as the dynamic viscosity of water in SI units, we find a drag force of 0.2 pN. This is about the drag force that a bacterium experiences as it swims through water.

Literature
G. K. Batchelor, An Introduction to Fluid Dynamics, Cambridge, 1967, 2000.
www.aemes.mae.ufl.edu/~uhk/STOKESDRAG.pdf

Womersley number

Principle

The largely inviscid and heavy pulsatile character of the flow in large blood vessels and airways is expressed in the Womersley number α, defined as:
$\alpha = r^2 \cdot \omega \cdot \rho / \eta$
where r is the tube radius, ω·the frequency of the oscillations of the flow, ρ the density and η the dynamic viscosity. The physics of pulsatile flow can be derived from the Navier-Stokes equations.
The Womersley number can be considered as the pulsatile version of the Reynolds number. With high numbers (> 7) inertia dominates, yielding a rather well blunted or flat flow front. With low numbers viscosity dominates yielding parabolic-like flows, however skewed towards the outer wall. With α = 2 the flow profile is practically parabolic-like. Fig. 1 illustrates the various types.

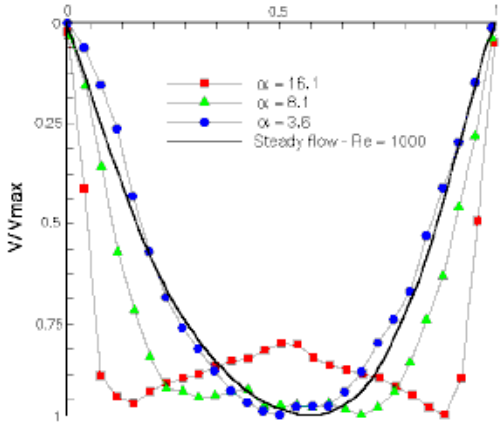


Fig. 1 Data of a model of the pulsatile flow profile of air in the trachea with various Womersley numbers compared to steady Poiseuille flow with Reynolds number Re = 1000. Relative scales. From http://www.vki.ac.be/research/themes/annualsurvey/2002/biological_fluid_ea1603v1.pdf

Application

Vascular and airways flow, described by the oscillatory flow theory. This theory is only relevant in the conduit tubes with there large diameters and turbulent flow. For the small tubes laminar flow holds.

More info

Since the heart does not generate a single sine wave but a series of harmonics (see Fourier analysis) the flow profile is found by adding the various harmonics. The relation between pressure drop and flow as given above is the so-called longitudinal impedance of a vessel segment.
In the periphery with small blood vessels or airways (small r) and little oscillation (α ca. 1), there is no need for the oscillatory flow theory and we can describe the pressure-flow relation with Poiseuille's law.
For the very large conduit arteries and airways, where α > 10, friction does not play a significant role and the pressure-flow relation can be described with inertance alone. For α values in between, the combination of the resistance plus the inductance approximates the oscillatory pressure-flow. Models of the entire arterial and airways system have indicated that, even in intermediate size arteries, the oscillatory flow effects on velocity profiles are mediocre. The main contributions of the tree to pressure and flow wave shapes are due to branching, non-uniformity and bending of the tubes. Thus for global transport phenomena, a segment of a tube can be described, in a sufficiently accurate way, by an inertance in conduit tubes, and a resistance in peripheral arteries and bronchioles and alveolar tubes. The oscillatory flow theory is, however, of importance when local phenomena are studied. For instance, detailed flow profiles and calculation of shear stress at the tube wall require the use of the oscillatory flow theory.

Gases and the Lungs

Adiabatic compression and expansion

Principle

To understand the concept of adiabatic phenomena, first Boyle's law is applied to the simple set-up of Fig. 1. It is assumed that both compartments have constant volume and temperature.

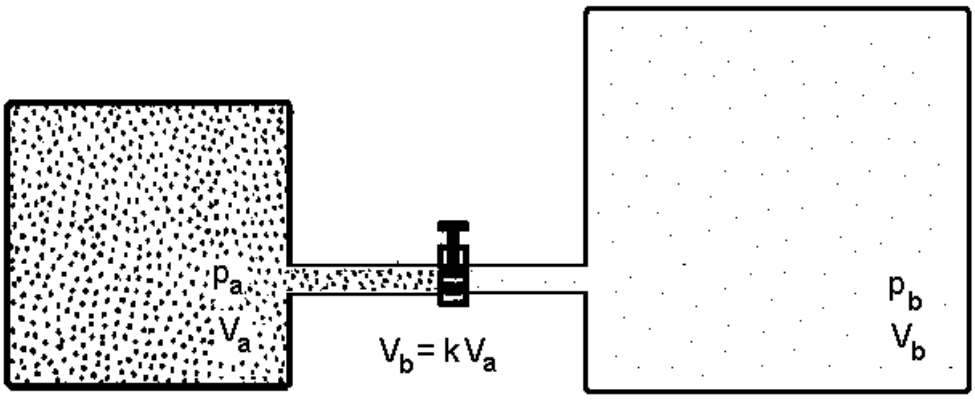


Fig. 1
The amount of gas in both compartments together is proportional $p_a \cdot V_a + p_b \cdot V_b$ which holds with the closed valve (at time zero) as well as after equilibration of the pressures (at time t) when the valve has been opened. So, it holds that :

$$p_{a,0}V_a + p_{b,0}V_b = p_t(V_a + V_b). \quad (0)$$

With given $p_{a,0}$, $p_{b,0}$, V_a and V_b = p_t can be calculated. However, this is only true when the temperature T in the whole system remains constant, i.e. isothermic compression and expansion holds. In daily life, a bicycle pump becomes warm when a tire is inflated. This is due to heat transfer from the heated compressed air in the pump to the tube. When the heat produced by compression is not given off to the environment but "absorbed" by the gas it self, or when the energy needed for expansion is not provided by the environment but provided by the expanding gas itself, than Boyle's law does not hold. Hence, in a heat-isolated system a compressing gas heats and an expanding gas cools down. These processes are called adiabatic compression and expansion. The deviation from Boyle's law can be very substantial (Fig. 2), as will be explained conceptually with the example of Fig. 1.

When the valve has a very narrow opening, the process of equilibration takes so much time (tens of minutes) that the whole process is isothermic. There is enough time for heat release and uptake between the system and the environment. When the equilibration is very fast (about a second; pipe and valve diameter large) the process is practically adiabatic and the changes of temperature can rather well be calculated.

For ideal gases under pure adiabatic conditions the p-V relation is:

$$p \cdot V^{\gamma} = \text{constant}, \quad (1)$$

with γ a numerical value greater than 1 and depending on the type of gas. γ is the so-called c_p/c_V ratio, the ratio of heat capacity of the gas (c_p) with constant p and the specific heat capacity of the gas (c_V) with constant V. It depends on the number of atoms in the gas molecule. For mono-atomic gases, e.g. the noble gases, γ is 5/3. For diatomic gases, for instance air, γ is 7/5 (= 1.4). More precisely, c_p/c_V is :

$$c_p/c_V = c_p/(c_p - R), \quad (2)$$

where R the ideal gas constant (see Gas laws).

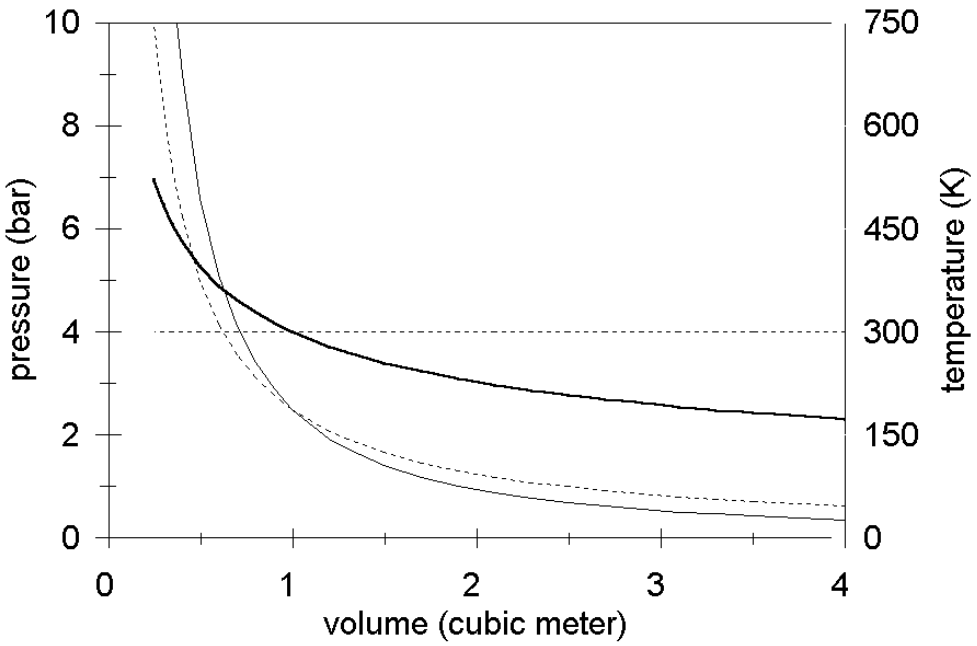


Fig.2 Isothermic and adiabatic p-V relation. The dashed curve gives the p-V relation according to Boyle's law and the dashed straight horizontal line the temperature belonging to it. The solid curve represents the adiabatic p-V relation and the thick solid curve gives the adiabatic temperature. n = 0.1 kmol.

As follows from Eq (1), reducing a volume by a factor k^1 gives a pressure increase much higher than k (being $k^{1.4}$). Conceptually, one can say that a factor k is due to the volume decrease according to Boyle and that the remaining factor, $k^{0.4}$, is caused by the temperature increase.
An isothermal p-V curve according to Boyle and an adiabatic p-V curve are depicted in Fig. 2. Numerically, the resulting temperature effect can be very substantial. A further description of adiabatic processes can be found in More info. Since textbooks seldom give examples how to calculate the adiabatic effect, an example is given in More Info.

Application

Adiabatic effects are seldom applied and in contrary, are generally unwanted. So, prevention is required. (Semi)-artificial breathing apparatus (home care pulmonary diseases, pulmonology, IC, surgery theatre) should have valves, tubes etc. designed in such a way that valves are not blocked by freezing and the breathing gas is not cooled down by adiabatic effects. Adiabatic effects also play a role in the technology of (clinical) hyperbaric chambers, high pressure gas tanks and diving regulators.
In the human (and mammalian) body), compression and expansion of gas filed cavities, adiabatic effects play now role. Even when one dives deep into the water from some tens of meters altitude, the temperature rise in the cavities is physiologically irrelevant since the process is sufficiently isotherm.
A medical imaging method, Optoacoustic Imaging is based on adiabatic processes.

More Info

The exponent γ is the c_p/c_V ratio, the ratio of heat capacity of the gas (c_p) with constant p and the specific heat capacity of the gas (c_V) with constant V. The complicated derivation can be found in physical textbooks. Here, a conceptual explanation is given. Related theory can be found in Thermodynamics: entropy.
The process of compression always has the same energetic cost, so independent of the velocity of the compression. This external energy is added to the intrinsic energy of the gas, which is actually the kinetic energy of the particles. The kinetic energy of the particles is proportional with the temperature. When the external energy is supplied instantaneously, it is completely added to the kinetic energy of the particles. When the compression is slow, then during the time of compression, the particles have the time to transfer a part of their kinetic energy to the vessel wall. The result is that the temperature of the gas raises less and the temperature of the wall increases. When the compression is extremely slow, the gas remains its temperature since all external energy is (indirectly) transferred to the environment. Now, the compression is isothermic and Boyle's law holds.
For instantaneous compression, the increase can be calculated as follows. Suppose that volume V with pressure p is compressed to V' yielding a new pressure p'. Applying Eq. (1) means that:

$$p \cdot V^{\gamma} = p'(V')^{\gamma}. \quad (3)$$

Before compression and after compression the ideal gas law (see Gas laws) holds, so that pV = nRT and p'V' = nRT' (T' is the new temperature), and some more calculation yields:

$$\Delta T = T' - T = T((V/V')^{(\gamma - 1)} - 1) \quad (4)$$

This relation also holds for volume expansion.
In a similar way, for a given change of pressure it can be derived that:

$$\Delta T = T((p/p')^{(\gamma-1)/\gamma} - 1). \quad (5)$$

Example of calculation
Suppose, everything ideal, that a gas volume at 300 K (27 ^{o}C) is instantaneously reduced in volume by a factor 2.5 and that there is no heat transport. According to Eq. (4) this yields $\Delta T = 300 \times (2.5^{0.4} - 1) = 133$ K. However, due to the temperature increase, a much higher factor of pressure increase is obtained (being $2.5^{1.4} = 3.6$) rather than the factor of 2.5 according to Boyle's law. Actually, a compression-factor of 1.92 ($= 2.5^{1/1.4}$) is required to obtain p = 2.5 bar. Then, the temperature increase becomes $300 \times (1.92^{0.4} - 1) = 90$ K.
The equation $p \cdot V^{\gamma}$ = constant works also the other way around. Suppose that the pressure p is increased instantaneously to k·p by a factor of k. Then, according to Boyle, the volume V reduces with a factor k, but now, the reduction is only a factor $k^{1/1.4}$. Again, in the pressure increase, a "Boyle factor" $k^{1/1.4}$ and a "temperature factor" $k^{0.4/1.4}$ are comprised. Of course, these two factors together have to yield k ($= k^{1/1.4} \cdot k^{0.4/1.4}$). Now, with k = 2.5 and T = 300 K the temperature increase is again 90 K ($2.5^{0.4/1.4} \times 300$ K).

In practice, it appears that the processes of compression and expansion are seldom purely adiabatic or purely isothermic. When the process is in between the calculation is complicated and, moreover in practice valves may block by freezing, for instance the valve of Fig.1. This may result in oscillatory behavior between blocking and unblocking by freezing and unfreezing.

Capnography

Principle

A capnograph is an instrument used to measure the CO_2 concentration of the expired air. Generally it is based on infrared (IR) analysis with a single beam emitter and measurements of non-dispersive IR absorption with a solid state mainstream sensor and ratiometric measurements of red/IR absorption, similar as with oximetry (see Pulse oximetry).

Molecular Correlation Spectrography (MCS with IR)
Laser-based technology is used to generate an IR emission that precisely matches the absorption spectrum of CO_2 molecules. The high emission efficiency and extreme CO_2 specificity and sensitivity of the emitter-detector combination allows for an extremely short light path which allows the use of very small sample cell (15 mm^3). This in turn permits the use of a very low flow rate (50 mL/min) without compromising accuracy or responsive time. This is in contrast to conventional CO_2 IR method, where the sampling flow rate is 150 mL/min.

Raman Spectrography
Raman Spectrography uses the principle of Raman scattering for $C0_2$ measurement. The gas sample is aspirated into an analyzing chamber, where the sample is illuminated by a high intensity monochromatic argon laser beam. The light is absorbed by molecules which are then excited to unstable vibrational or rotational energy states (Raman scattering). The Raman scattering signals (Raman light) are of low intensity and are measured at right angles to the laser beam. The spectrum of Raman scattering lines can be used to identify and quantify all types of molecules in the gas phase.

Mass Spectrography
The mass spectrograph separates molecules on the basis of mass to charge ratios (see Mass spectrography).

Applications

Capnography is widely used in clinical practice, such as pulmonology, anesthesiology and in the IC-unit. Mass spectrometers are quite expensive and bulky to use at the bedside. They are either "stand alone," to monitor a single patient continuously, or "shared," to monitor gas samples sequentially from several patients in different locations (multiplexed). Tens of patients may be connected to a multiplexed system (with a rotary valve) and monitored every 2 or 3 min.

More Info

In addition to this ‘‘classical’’ methods new techniques are photo-acoustic and magneto-acoustic technique for CO_2 monitoring in the open unintubated airway.

Acoustic impedance measurement (PAS) PAS is a still experimental approach (see ref. 1), based on an acoustic impedance measurement of the air with an electro-acoustic sensor coupled to an acoustic resonator. When IR energy is applied to a gas, the gas will try to expand which leads to an increase in pressure. The applied energy is delivered in pulses causing pulsation of the pressure. With a pulsation frequency lying within the audible range, an acoustic signal is produced. This is detected by a microphone. The impedance characteristic is depending on the sound velocity within the gas mixture contained in the acoustic resonator. The relation between the acoustic impedance and the CO_2 concentration is approximately linear. Since the sound velocity (and so the acoustic impedance) is also dependent on temperature and humidity, the outcome should be corrected for these parameters.
Potential advantages of PAS over IR spectrometry are:

- higher accuracy;
- better reliability;
- less need of preventive maintenance;
- less frequent need for calibration.

Further, as PAS directly measures the amount of IR light absorbed, no reference cell is needed and zero drift is nonexistent in PAS. The zero signal is reached when there is no gas present in the chamber. Despite being a superior method of measurement of CO_2, photoacoustic spectrography did not gain as much popularity as IR spectrography.

Literature

1. http://www.mrtc.mdh.se/publications/1006.pdf
2. http://www.capnography.com/Physics/Physicsphysical.htm

Compression and expansion

Principle

To understand the concept of diabatic (i.e. isothermic) phenomena, first Boyle's law is applied to the simple set-up of Fig. 1.

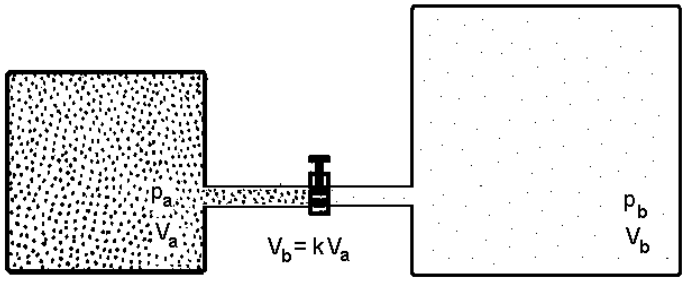


Fig. 1

The amount of gas in both compartments together is $p \cdot V_a + p \cdot V_b$ which holds with the closed valve (at time zero, t_0) as well as after equilibration of the pressures (at time t) when the valve has been opened. So, it holds that:

$$p_{a,0} \cdot V_a + p_{b,0} \cdot V_b = p_t \cdot (V_a + V_b). \quad (1)$$

With given $p_{a,0}$, $p_{b,0}$, V_a and V_b at t_0, p_t can be calculated. When $V_b = jV_a$ and $p_{b,0} = kp_{a,0}$ then p_t becomes:

$$p_t = p_a(1+jk)/(1+j). \quad (2)$$

However, this is only true when the temperature T in the whole system remains constant during the whole process of pressure equilibration, i.e. the process is isotherm or diabatic. When this is not the case, then the left compartment cools down and the right one heats up. For more info about this adiabatic process, see Adiabatic Compression and expansion.
Another condition is that the gas is ideal, i.e. the particles do not interact and they have no size. When pressures (so densities) are low, both conditions apply well and (1), which is actually based on Boyle's law (see Gas Laws), can be used. When they do not hold, the Van der Waals corrections are necessary.

Application

Innumerous in science, technology and so indirectly in medicine, e.g. in breathing apparatus and especially in pulmonology.
The Van der Waals corrections are used for mass calculations of commercials gases in high-pressure tanks e.g. applied in medicine, especially with the expensive helium.

More Info

The Van der Waals correction has two constants *a* and *b* (Table 1). The correction factor "*a*" is needed for the interaction between the gas particles (attraction coefficient), and a correction factor "*b*" for the volume occupied by the gas particles.

Table 1.

Molecule, atom or mixture	Van der Waals constants *a* 10^3J.m^3/kmol^2	*b* 10^{-3}m^3/kmol
He	3,5	22
H_2	25	26
O_2	140	31
N_2	140	39
CO_2	360	44
H_2O	550	30.5
air	140	37.4

b of air interpolated from weighted *b*'s of O_2 and N_2

According to Van der Waals:

$$(p + an^2/V_m^2)\,(V_m - nb) = nRT, \qquad (3a)$$

where V_m the total volume and n the number of kmoles in the volume V. Constant b increases and constant a reduces pressure. Table 1 gives for some gases the numerical values of the two constants. When a and b are zero, then the van der Waals equation degenerates to Boyle's law. To calculate p, (3) can be rewritten:

$$p = nRT/(V_m - nb) - an^2/V_m^2, \qquad (3b)$$

With normal temperatures and pressures of about 100 bar or more, the Van der Waals correction makes sense. For instance, a tank filled with air at 300 bar and 290 K comprises 7.8% less (assumed that $a_{air} = 37.4 \times 10^{-3}$). At 200 bar the interaction effect dominates the particle-volume-effect, but at 300 bar the situation is reversed. The p-V diagram of Fig. 2 also illustrates the rather complicated behavior of the correction. Since Boyle's law is independent of the type of particles, the straight line of Boyle (log-log diagram) holds for both air and helium.

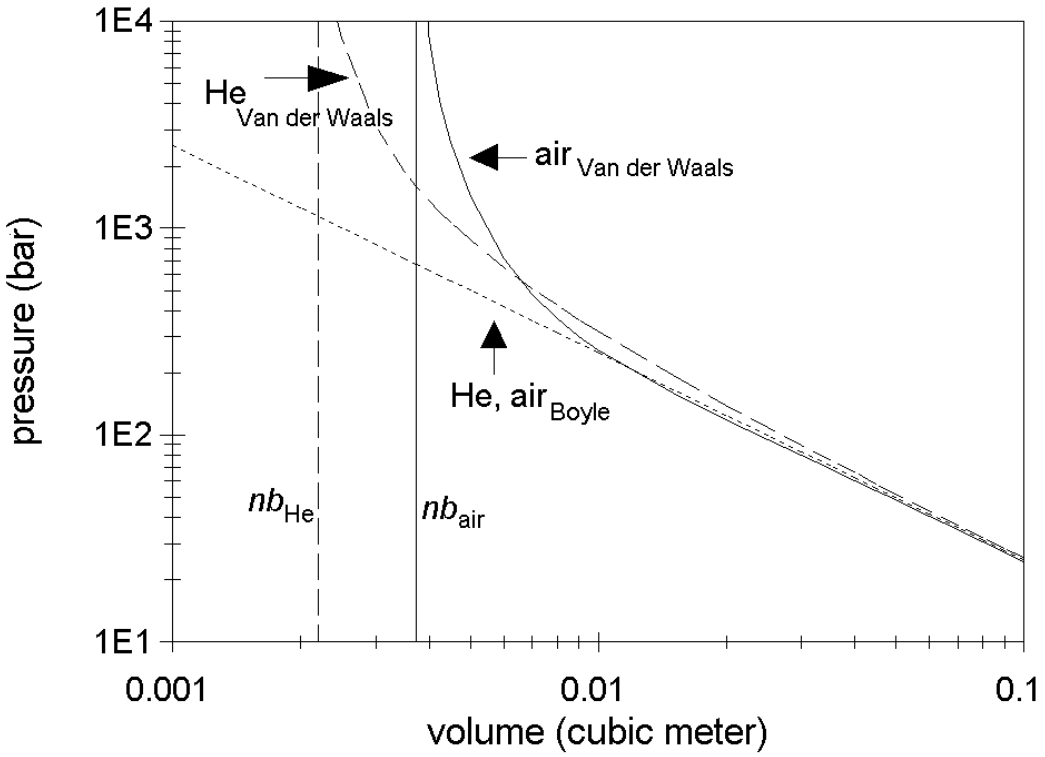


Fig. 2 Comparison of p/V relation according to the Law of Boyle and according to the Van der Waals equation for air and He. The curves are calculated for n = 0.1 kmol and T = 300 K (1E4 = 10^{-4}).

The Van der Waals curve of He shows already strong deviations (about 10%) from Boyle's law at 100 bar, since the interaction effect is weak compared to the particle-volume-effect. The p-V curve approaches the straight Boyle line from above, but the air-curve first crosses the Boyle line and then approaches the line from below. For low pressures the Boyle line is the asymptote for the Van der Waals curves. The rather surprising behavior of the Van der Waals equation is due to the fact that it is a cubic equation in n and in V.
For very precise calculations in gas mixtures at pressures beyond 50 bar, this correction is even not precise enough. Then, the theory of virial coefficients, taking also into account the (first and higher order) interactions between the types of particles in a mixture, is applied (see textbooks of physics).

Gas laws

Laws of Boyle, Gay-Lussac, Avogadro and Dalton, and the ideal gas law

Basic principles

A gas is one of the three aggregation states of a substance, solid, fluid and gaseous. A gas is compressible and often a mixture. It occupies all available space uniformly and completely, has a small specific mass, diffuses and mixes rapidly, and is mono- (He), di- (N_2, O_2, CO), tri- (CO_2, O_3), or poly-atomic (NH_3, methane etc). In a so-called ideal gas, the particles (being molecules or atoms) have no size and do not influence each other since they show no mutual attraction (no Cohesion).
Gas particles move at random with velocities of many hundreds m/s (Table 1). The mean particle-particle distance is in the nm-range. Gas volume is empty for ca. 99.9% and therefore the gas particles compared to liquids infrequently collide with each other. At 300 K and 1 bar the mean free path of H_2 is ca. 66 nm, more than 10 times than in the liquid phase. Collisions with constant temperature (isotherm) are pure elastic, also with a wall at the same temperature.
The gaslaws are those of *Boyle* (and Mariotte), *Gay-Lussac* (and that of Charles), *Avogadro* and *Dalton*, and the *Ideal gas law*,

Table 1

Particle	diameter of particle nm	velocity v (273 K) m/s	km/h	molecular mass m	c_p/c_V ratio or γ *
He	0.027	1304	4690	4.003	1.66
H_2		1838	6620	2.016	1.41
O_2	0.034	461	1740	32.00	1.4
N_2	0.037	493	1860	44.011	1.29
H_2O(vapor)	0.027			18.016	1.33

* $c_p/c_V = c_p/(c_p-R)$ (see Boyle's law, the ideal gas law and Adiabatic compression and expansion).

Applications

These laws are fundamental for anesthesiology, the (patho)physiology of pulmonology, the medicine of diving, especially with respect of the occurrence of decompression sickness, hyperbaric medicine (HBO, the application of pure O_2 as breathing gas under *h*yper*b*aric conditions), aviation (also recreational types), aerospace medicine and mountaineering. The law of Dalton is basic for the physiology of respiration.

More info

These laws can only be applied when there is integral heat exchange with the environment. Then the process is called isothermic or diabatic. In practice this seldom holds. For instance when a high-pressure helium tank is filled in the factory it raises in temperature (adiabatic compression). This compression is so fast that the produced heat has not the time to be transferred to the environment. For adiabatic (non-isothermic compressions or expansions see Adiabatic compression and expansion). The ideal gas laws assume an isothermic process and, self-evident with an invariable amount of gas mass.

The law of Boyle (and Mariotte)

In the derivation of Boyle's law it is assumed that the gas particles make elastic collisions with the wall surrounding the gas. The collisions exert a force on the wall, which is proportional to the number of particles per unit of volume, their velocity and their mass. Pressure (p) is defined as force (F) per area (A). By doubling the density of the gas (this is doubling the number of particles in a given volume), the number of collisions doubles, and hence the exerted force and so the pressure. Hence, the equation:

$$p_1 \cdot V_1 = p_2 \cdot V_2 = \text{constant} \quad \text{or}$$
$$p_1/p_2 = V_2/V_1, \text{ or } p = \text{constant} \cdot V^{-1}, \qquad (1)$$

is obtained. This law holds well for moderate gas densities (pressures < 300 bar, with regular temperatures) and under isotherm conditions, i.e. the process is diabatic. For higher pressures the condition that the total volume of the particles can be neglected and that the particles do not influence

each other does not hold any longer. The Law of Boyle is refined with the van der Waals corrections, see Compression and expansion.

The law of Gay-Lussac

This law states that the ratio of pressure and absolute temperature is constant provided that volume of the gas is constant:

$$p_1/ p_2 = T_1/T_2 = \text{constant}. \quad (2a)$$

It has been proved that the squared velocity $\langle v^2 \rangle$ is proportional with T and reciprocal to the gas mass:

$$\langle v^2 \rangle = 3RT/(N_A m), \quad (2b)$$

where $\langle v^2 \rangle$ the mean of the squared velocity of the particles, R the molar gas constant (= 8315 J/kmol·K), $N_A m$ the gas mass with N_A the Avogadro's number and m the particle mass.
Conceptually, the correctness of the law can be understood by realizing that $\frac{1}{2} \cdot m \cdot v^2$ is kinetic energy of a particle. So, for a certain type of particle an increase in T gives an increase of v^2 and consequently of p. When p and *n* are constant it holds that:

$$V_1/T_1 = V_2/T_2 = \text{constant}, \textit{ the law of Charles}. \quad (2c)$$

The ideal gaslaw (of Boyle and Gay-Lussac)

This is a combination of the law of Boyle and the law of Gay-Lussac:

$$pV = nRT, \quad (3)$$

with *n* the number of kmoles of the gas and with R the molar gas constant (= 8315 J/kmol·K).
It holds that $pV = \frac{1}{3} \cdot N_A m \langle v^2 \rangle$ = constant and that $\langle v^2 \rangle = 3RT/(N_A m)$. After substitution of $\langle v^2 \rangle$ and applying this for *n* moles the law follows.

The laws of Avogadro

$$V_1/n_1 = V_2/n_2 = \text{constant} \quad (4)$$

Since equal volumes of ideal gasses at equal pressure and equal temperature comprise an equal number of particles, the law follows directly.

The law of Dalton

The pressure of a mixtures of gasses is the sum of the pressure of the individual gasses (defined as the partial pressures) since the kinetic energy ($\frac{1}{2} \cdot m \cdot \langle v^2 \rangle$) of all types of particles, irrespective their type, is the same: $m\langle v^2 \rangle = 3RT/N_A$ = constant (see law of Gay-Lussac). So:

$$p_{total} = p_1 + p_2 + p_3 .... \quad (5)$$

Gas volume units, STPD, BTPS and ATPS

Since the mass of gas in a unit volume is dependent on pressure and temperature (see Gas laws) they have to be specified with their pressure and temperature.
Three quantities are used.
STPD: Standard Temperature and Pressure Dry, so at 1 atm (760 mm Hg), 0^o C and $p_{HO2} = 0$;
BTPS: Body Temperature Pressure Saturated, defined at 37^o C, ambient pressure and saturated with water vapor (p_{H2O} = 6.3 kPa or 47 mm Hg);
ATPS: Ambient Temperature Pressure Saturation, so at ambient temperature and pressure, with p_{HO2} saturated.

Respiratory volumes are usually reported at BTPS. On the other hand, moles of gas (i.e. O_2, CO_2 production) are commonly reported at STPD.
Sometimes volume measurements are performed neither at STPD nor at BTPS, but at ATPS.
Conversion among these conditions can be easily performed by applying the Gas laws for ideal gases, with volume proportional to absolute temperature T, and inversely proportional to pressure P.
As an example, if a spirometer is used to measure the tidal volume (V_{tidel}) of a subject in a mountain region where barometric pressure P=450 mm Hg, and the spirometer is at 23^o C (knowing that the pressure of water vapor at saturation at 23^o C is 21 mm Hg, and that at a body temperature of 37^o C is 47 mm Hg) ATPS can be converted to BTPS. For convenience pressures are all in mm Hg, so recalculation in bar or Pascal is not required.

$$V_{BTPS}/V_{ATPS} = ((T_{BTPS} / T_{ATPS})) \cdot ((P_{ATPS}) / P_{BTPS})) \quad (1a)$$
$$= ((273+37)/ (273+23)) \cdot ((450-21)/(450-47))$$

hence,

$$V_{tidel,BTPS} = 1.1149\ V_{tidel,ATPS}\ . \quad (1b)$$

Similarly, a BTPS volume can be converted to a STPD volume:

$$V_{STPD} = ((P_{BTPS} -47)/760) \cdot (273/310) \cdot V_{BTPS} = 0.00116 \cdot V_{BTPS} \quad (2)$$

When P_{BTPS} = 760 mm Hg, then:

$$V_{STPD} = ((760-47)/760) \cdot (273/310) \cdot V_{BTPS} = 0.826\ V_{BTPS} \quad (3)$$

Hot wire anemometry

Principle

In the hot wire anemometry an electrically heated wire (or screen or film) is placed in the gas pathway, which is cooled by the gas flow (Fig. 1). The degree of cooling depends on the gas flow rate, which can thus be calculated.

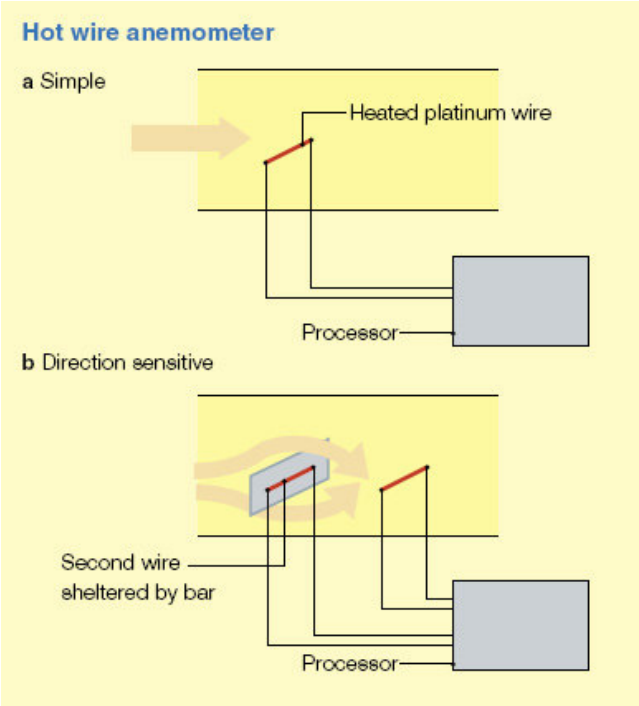


Fig. 1 (from http://www.frca.co.uk/article.aspx?articleid=100390)

Application

Especially in pulmonology, also neonatal, and IC unit.

More Info

The hot wire (usually Pt) is incorporated into a balanced Wheatstone bridge circuit. Cooling the wire changes its resistance and unbalances the bridge. Most designs work on the constant temperature principle, whereby a correcting current is applied through the hot wire to compensate for the cooling effect of the gas, maintaining a constant wire temperature and thus restoring the balance in the Wheatstone bridge. This current is measured and from it the gas flow rate is determined. To compensate for changes in the gas temperature, a second wire is usually incorporated, which is maintained at ambient temperature. Minor corrections are also made according to the gas composition to accommodate the variation in specific heat capacity. Hot wire anemometry is generally extremely accurate.

The cooling effect occurs with flow in either direction, and so to measure exhaled tidal volume the hot wire anemometer is placed in the expiratory branch of the circuit. It can be modified to provide information about the direction of flow by using an additional heated wire placed just downstream from a small bar, as shown in Fig. 1b. This bar shelters the wire from the full cooling effects of flow in one direction but not the other, and thus inspiratory and expiratory flows can be calculated separately. For this purpose the sensor must be placed in the Y-mouth-piece before the bifurcation. This technique is particularly useful to control neonatal ventilation.

Lung gas transport 1, basic principles

Theorie of gas flows in the airways system
Pulmonary gas transport is afflicted with a huge number of physiological and clinical variables. Possibly this is a reason for its step motherly position in the medical curriculum.
Streams of gas and liquid in pipes behave on a gliding scale from orderly, i.e. laminar, to chaotic, i.e. turbulent, with a transitional stage in between; see Fig. 1 of Poiseuille's Law. Types of flow are classified according to Reynolds number, Re. For Re < 2200 the flow is laminar and for Re > 10000 the flow is turbulent. For laminar flow in circular tubes Poiseuille's Law holds (parabolic velocity profile).
The airways can be described as a system of many-fold bifurcating tubes as illustrated in Fig. 1.

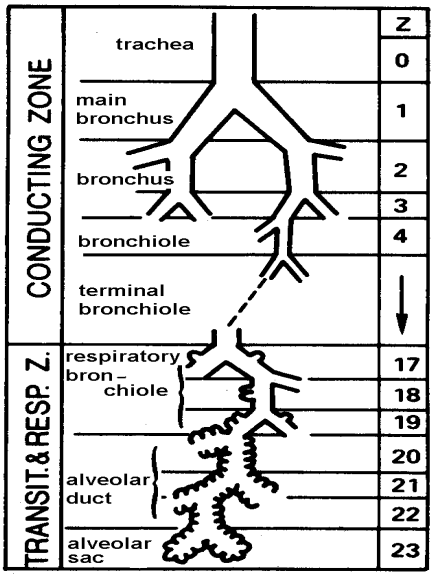


Fig. 1 Weibel model of airways with its generation numbers indicated (see text). Conduction zone: transport; transition zone: transport and gas exchange; the respiratory zone (alveoli) gas exchange. (Modified from Weibel and Gil, 1977).

The trachea is called the 0^{th} generation, the bifurcation of the trachea is of the 1^{st} order and the two main bronchi are the 1^{st} generation. The trachea always has the highest Re and the glottis still more. Even for highest inspiratory volumes (RMV, Respiratory Minute Volume, > 120 L/min), the high generations have laminar flow.

Basic lung gas transport
The driving force of volume flow is the transpulmonal pressure difference Δp_{tp}, between the pressure in the tissue directly surrounding the lung and the pressure in the mouth cavity. For simplicity flow is considered as stationary and not pulsatile (by approximation sinusoidally) as holds for normal breathing. Generally it is measured in the esophagus proximal of an intra-esophageal inflated balloon. When the pressure needed to inflate the alveoli is subtracted, the resulting difference Δp varies from 100 Pa (= 1 cm H_2O; RMV about 6 L/min) to 800 Pa (high flows) and under special conditions (e.g. airways disorders) it can go up to about 3000 Pa. Δp is distributed over the whole bifurcating network from mouth up to the last generation.

Application

The Weibel model is widely applied for flow calculations. It is useful to obtain insight in flow behaviour with disorders in the upper and lower airways of patients of all ages.

More Info

The driving force of volume flow (i.e. volume/time) $\dot{V}$ in a tube of length L is the pressure difference Δp_{tube} over L. ($\dot{V}$ = 2·RMV/60 in L/s, the factor 2 is often, but not always, already implied in the equations.) When Poiseuille's law holds, then the flow-velocity v increases linear with Δp_{tube}. Laminar flow decreases linear with the gas viscosity (but is independent of its density ρ). Therefore, the

resistance of laminar flow is called the viscous resistance. Note that laminar flow is not always parabolic. Poiseuille's Law only holds for parabolic flow.

With $\Delta p_{trachea}$ = 1 Pa, Poiseuille's law gives a $\dot{V}_{trachea}$ of about 70 L/min. Is this flow still laminar? $Re_{trachea}$ appears to be 5560 (at 35 °C), so the flow is in between laminar and turbulent. Therefore the actual $\dot{V}_{trachea}$ is considerably lower. Imperfections (non-smooth wall, slightly varying diameter, curvature etc.) further diminish $\dot{V}_{trachea}$. However, most effective is the constriction of the glottis causing substantial pressure loss and non-laminar behavior due to the constriction effect (see Flow through a stenosis) and the entrance effect (see Flow: entrance effect and entrance Length). In conclusion, even for low flows ($\dot{V}$ = 0.3 L/s; nearly rest) Poiseuille's law is not valid in the trachea. A higher Δp than given by Poiseuille's law is needed to obtain this flow.

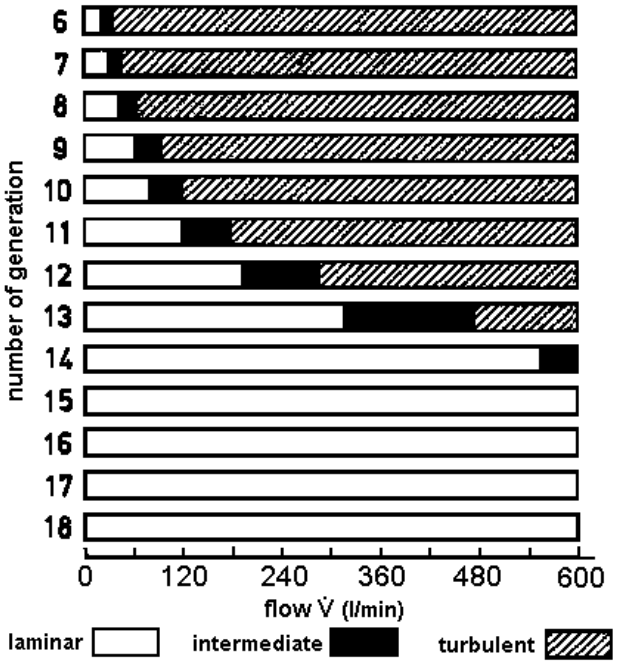


Fig. 2 Types of flow in the airway generations (after Hrnčíř, 1996, with increased generation numbers by 5-9 depending on the flow, to account for the entrance effects). A flow at the maximum of ventilation of 600 L/min can be reached by an endurance top-athlete during short lasting extreme exertion.

Also bifurcations disturb Poiseuille flow (see Flow in bifurcations). Heavy disturbance can give intermediate or turbulent flow.

After the entrance length, dependent on the diameter of the daughter branch and its Re, in the daughter branch the flow regains a laminar profile. Subsequent orders of bifurcations are so close together that for the upper generations the current fails to change to laminar before the next bifurcation is reached. However, from a certain critical generation laminar flow is always realized.

For ½ L/s (as with walking) this is at about generation 10. The two effects, the entrance phenomenon and the flow-profile dependency on Re, results in a critical generation with a lower generation number the higher the flow. For the lowest flow (0.2L/s) this is about generation 5 and for the highest flow (10 L/s) generation 15.

All together, the P-V relation of the airways system is non-linear. The non-laminar behavior implies that Δp is proportional to V^e with 1<e<7/4. This is discussed in Lung gas transport 2, pressure-flow relation.

Literature

Hrnčíř, E. Flow resistance of airways under hyperbaric conditions. Physiol. Res., 45, 153-158, 1996.
1962.
Weibel E.R. and Gil J. Structure-function relationships at the alveolar level. In: Bio-engineering aspects of the lung, West, J.B. (ed), Lung Biology in health and Disease, Vol. 3, Marcel Dekker Inc., New York and Basel, 1977

Lung gas transport 2, pressure, volume and flow

Principle

The volume-pressure relationships

$\dot{V}$ (L/s) is the time derivative of volume, V. V and $\dot{V}$ are measured with a spirometer, generally as function of time (see Spirometry). Pressures are measured with a barogram, as function of time (Fig. 1). The peak-peak value of the barogram and spirogram, dynamically obtained (quite breathing), are Δp_{tp} and V_{tidal} (=0.5 $\dot{V}$ /respiratory frequency) respectively. Combining both, the V-P relation is obtained (Fig. 1, right), that is not linear but elliptic caused by non-equal airways resistances of in- and expiration. The expiratory airflow is limited by dynamic airway compression, the downstream distally narrowing of the airways up to a *choke point* or equal pressure point. But the main cause of the hysteresis is pneumatic behaviour of the alveoli. During one breathing cycle the elliptic V-P loop is completed. With a high breathing frequency the ellipse becomes rounder and with a high V_{tidal} larger.
The V-P relation can also be statically measured with forced ventilation, generally with anaesthetised subjects (Fig. 2). With a high V_{tidal} the resulting 'loop' becomes flatter and strongly sigmoid when V_{tidal} = TLC-RV (total lung capacity minus residual volume).
Another diagram obtained with spirometry is the V- $\dot{V}$ diagram obtained during maximal forced expiration. Its maximum provides insight in the mechanical properties of the airways and its shape about flow limitations.
In the P-V diagram, the dashed axis of the ellipse gives the dynamic lung compliance ($C_{L,dyn}$) and the surface of the ellipse the work done to overcome the viscous airways resistance R_{aw}.

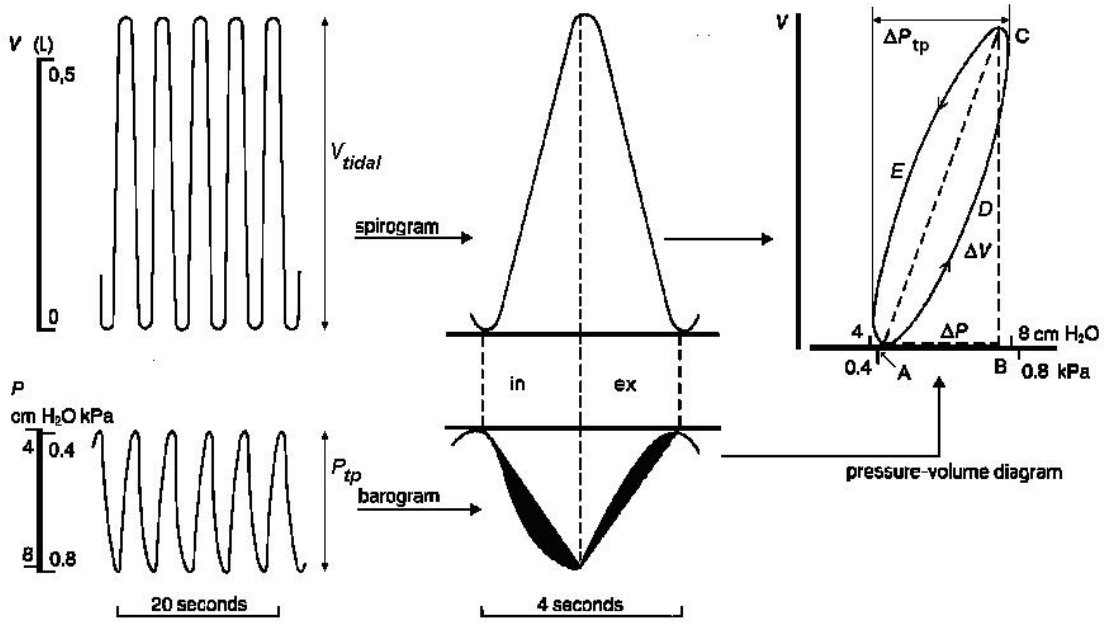


Fig. 1. The black area in the middle lower diagram accounts for the elliptic shape of the P-V loop (right diagram). It gives the dynamic lung compliance $C_{L,dyn} = \Delta V/p\Delta$. ($P_{tp}$ is Δp_{tp}).

Application

Performing heavy work requires an RMV of about 37.5 L/min. This is nearly 7 times the value in rest (RMV = 5.6 L/min, or $\dot{V}_{O2}$ = 0.32 O_2/min). The supposed Δp is 600 Pa, and consequently Δp_{tp} is about 1000 Pa when the effect of the compliance (mainly alveoli) is added. This value can be maintained for a long time.
What does this means for (mild) allergic asthma? Suppose, that (as a model) due to (very) cold inspired air there is an acute constriction of 15% of the bronchioles and smaller airways. Then $\mathbf{R}_{aw}$ (the airways resistance) increases by a factor of about 1.6 ($\approx 0.5 + 0.5 \times 0.85^{-19/4}$, see Fig.1 and **More Info** of Lung gas transport 3, resistance and compliance), yielding a Δp of 1060 Pa. Adding the requested pressure to overcome the elastance **E** (the resistance of the alveoli to be inflated), Δp_{tp} becomes about 1500 Pa. This is much too fatiguing to maintain e.g. running and it has a high risk of escalation (more constriction and hypercapnia).

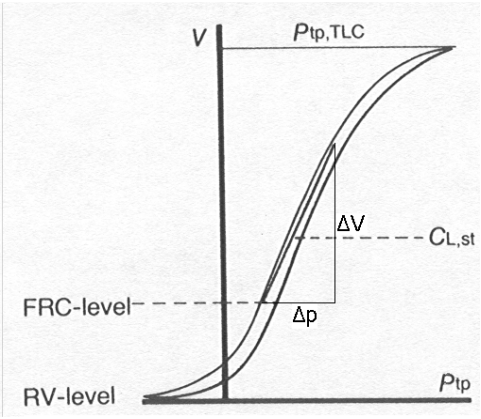


Fig. 2 P_{tp}-V diagram recorded under static conditions (passive ventilation). $\Delta p_{tp} = p_{mouth} - p_{oesophagus}$. $C_{L,stat}$ (*static* lung compliance) = $\Delta V/\Delta p$. FRC functional residual capacity, TLC total lung capacity, RV the residual volume, all in L.

Chronic smokers (>10 cigarettes/day, > 15 years) have poor gas transport since their $\mathbf{R}_{aw}$ is at least 25% higher (resulting in $\Delta p \approx 1100$ Pa). They have many bullae, which, according to the changed surface tension, decreases the elastance **E**. However, it is mostly found that **E** is increased since the alveolar tissue stiffness is substantially increased. They have a reduced alveolar volume. Conclusion: chronic smoking and endurance sport are pulmonary incompatible.

The shape of the hysteresis loop of the static and dynamic P-V diagram is diagnostically of great importance. In healthy lungs, $C_{L,dyn}$ halfs when the breathing frequency (f) changes from 10/min to maximal (>40/min). The static narrow sigmoid P-V diagram changes to pear-shaped with f = 30/min and V_{tidal} maximal.
With disorders the shape of the loop is more complicated, especially with bronchoconstriction and dilatation induced by drugs.
With emphysema and normal breathing, the static Δp-V loop is shifted to left and for maximal ventilation the loop is wider. TLC and static C_L are increased and Δp_{tp} decreased. With bronchodilatation all these effects are reversed. Pulmonary fibrosis causes a wider and more horizontal loop and so an increased p_{tp} and decreased TLC and $C_{L,stat}$.

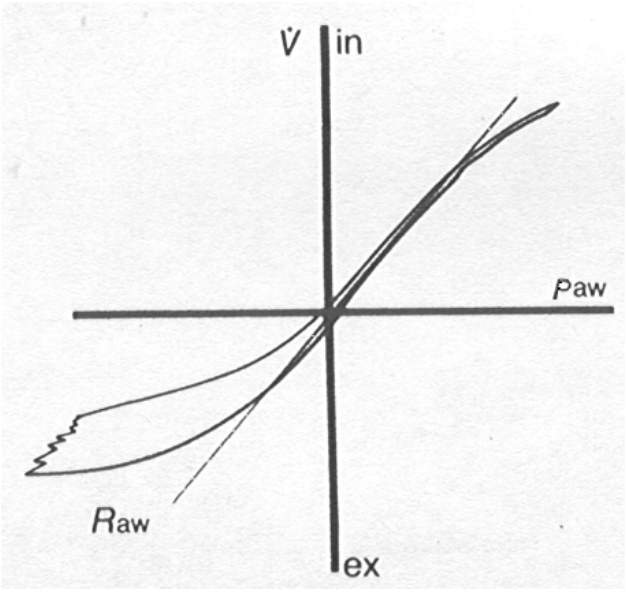


Fig. 3 p_{aw}-$\dot{V}$ diagram with $p_{aw} = p_{mouth} - p_{alveolar}$. The latter is measured with body Plethysmography.

Another diagnostically important diagram obtained with Spirometry is p_{aw}-$\dot{V}$ diagram (Fig. 3), measured during quiet breathing. It is useful to find abnormal changes of R_{aw} (= $d(\dot{V})/d(p_{aw})$.

In general, obstructive and restrictive disorders have different outcomes for the various volume measures, as Table 1 shows.

Table 1

	Obstructive	restrictive
FEV1/FVC ratio	↓	↑
FEV1	↓	↓
FVC	↓	↓
TLC	↑	↓
RV	↑	-
FRC	-	↓

More Info

The airways have a mixed laminar and turbulent behaviour, yielding the classical equation:

$$\Delta p = \text{resistance times current} = (K_1 + K_2 \cdot \dot{V}\,)\,\dot{V} = K_1 \cdot \dot{V} + K_2 \cdot \dot{V}^2, \quad (1)$$

with the laminar term linear and turbulent term quadratic with $\dot{V}$. K_1 represents the laminar flow-resistance. Simplified, it comprises the viscosity of the gas, and the dimensions of the airways. K_2 accounts for the extra resistance due to irregularities of the walls of bronchi etc., to constrictions and bifurcations, resulting together in turbulent flow. The behaviour of the alveoli with their surface tension and the counteracting effect of the alveolar surfactant (see Surface Tension) are not yet incorporated. In addition to this, there is a small force caused by the elasticity (the stiffness to prevent strain, for hollow organs the elastance **E** (see Compliance (hollow organs)) of the tissue of alveoli and bronchioli. Surface tension and elasticity together constitute the static elasticity **E** of the system. For a further treatment of **E** see Lung gas transport 3, airways resistance and compliance. The semi-theoretical approximation with **E** implied is given by the equation:

$$\Delta p_{tp} = \mathbf{E} \cdot V_{tidal} + K_1 \cdot \dot{V} + K_2 \cdot \dot{V}^2. \quad (2)$$

(The product $\mathbf{E} \cdot V_{tidal}$ has the well-known electric analogue $Q/C = V$, with Q the charge of a capacitor and analogous to V, *C* the capacitance (compliance) of a capacity and analogous to $1/\mathbf{E}$, and V the electric potential, analogous to pressure.) Since Eq. (2) is very sensitive for the value of **E** it is inappropriate to calculate the constants or Δp_{tp}. Ignoring **E** of airways and alveoli, and with low, constant flows (or slowly changing), such that the effects of inertia can also be ignored, analogue to the electric law of Ohm ($V = i \cdot \mathbf{R}$), it holds that:

$$\Delta p = \mathbf{R}_{aw} \cdot \dot{V}, \quad (3a)$$

where $\mathbf{R}_{aw}$·the airways resistance. With turbulent flow, tube resistance is proportional to $\dot{V}^{3/4}$ (see Lung gas transport 3, airways resistance and compliance). This reduces (3a) to:

$$\Delta p = k \cdot \dot{V}^{7/4}. \quad (3b)$$

Including mouth and glottis, the constant k is nearly 120, and replacing by RMV, the equation becomes:

$$\Delta p' = 0.3 \cdot RMV^{7/4}. \quad (3c)$$

For intermediate flow the exponent is in between 1 and 7/4. Eq. (3b) is theoretically a better alternative for (1).

Literature

Nunn, J.F. Applied Respiratory Physiology. 2nd edition. Butterworths, London-Boston, 1977.
Tammeling, G.J. and Quanjer, Ph.H. Contours of Breathing, Boehringer Ingelheim BV, 1978, Haarlem.

Lung gas transport 3, resistance and compliance

Principle

The ventilatory system is characterized by a resistance, an inert (mass) component and a compliance. The latter is 1/stiffness = 1/elastance=1/**E**. **E** is measured in kPa/L. A high elastance counteracts inflation of the lungs. The airways, the alveoli, the lung tissue and chest wall all contribute. The inert airways and alveolar component is always neglected. The same holds for the lung tissue and wall inertia up to moderate ventilation.
Resistances and compliances can be measured and modelled, but both approaches are not trivial.

Resistance of the airways tree
Empirically, $\mathbf{R}_{aw}$ is related to the total lung capacity TLC (in litre) and age (year) according to Tammeling and Quanjer (1978):

$$\mathbf{R}_{aw} = 410((0.0021A+0.4)TLC)^{-1}, \qquad (Pa{\cdot}L^{-1}{\cdot}s) \qquad (1)$$

where TLC is 10.67L – 12.16 (male) with L the length (m). $\mathbf{R}_{aw}$ can also be measured from the dynamically recorded P_{aw} - $\dot{V}$ curve (see Lung gas transport 2, pressure, volume and flow). For a subject of 40 years and L=1.75 m this yields a $\mathbf{R}_{aw}$ of 130 $Pa{\cdot}l^{-1}{\cdot}s$.
Finally, $\mathbf{R}_{aw}$ can be estimated from the Weibel model of the airways system (see Fig. 1 of Lung gas transport 1, basic principles) by adding $\mathbf{R}_{tube}$ of all tubes together (see **More Info**). The airways tree appears to have a resistance **R** of 29 $Pa{\cdot}L^{-1}{\cdot}s$ Since $\mathbf{R}_{aw}$ is proportional to $\dot{V}^{3/4}$, $\mathbf{R}_{aw}$ becomes:

$$\mathbf{R}_{aw} = k{\cdot}\mathbf{R}{\cdot}\dot{V}^{3/4} = k{\cdot}29\,\dot{V}^{3/4}. \qquad (2a)$$

With k is 4 (see More Info) and rounded $\mathbf{R}_{aw}$ is:

$$\mathbf{R}_{aw} = 120\,\dot{V}^{3/4}. \qquad (2b)$$

For $\dot{V}$ =1 L, (2b) is close to that found with the empirically equation (1). Chest wall resistance is about twice $\mathbf{R}_{aw}$ (lung tissue resistance can be neglected), which brings total pulmonary resistance at about 400 $Pa{\cdot}L^{-1}{\cdot}s$, about the mean of literature data (Nunn 1977).

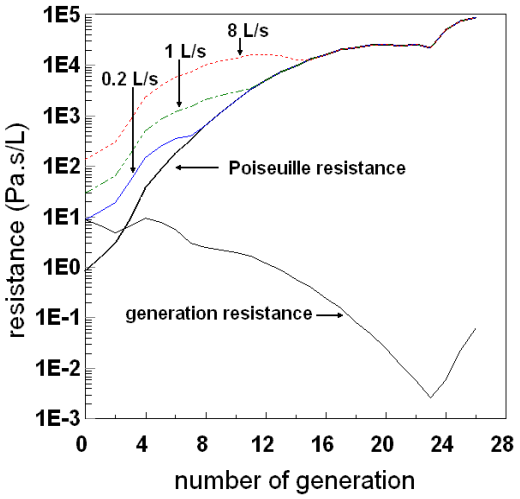


Fig. 1 The curves indicated by flows are the resistances of individual branches according to (3a) and the Poisieulle resistance according to (3b). The generation resistance, obtained by dividing the individual tube resistance by the number of parallel tubes in a single generation. The curve was calculated for $\dot{V}$ =0.2 L/s (awake, in rest). $\dot{V}$ =1 L/s and $\dot{V}$ =8 L/s are equivalent to normal and extreme endurance sport respectively. Generation 0 is the trachea.

Compliance of the airways system

The determining factors for the alveolar compliance are the surface tension of the alveoli (and alveolar ducts) and the counteracting effect of the alveolar surfactant (see SurfaceTension).
The healthy $C_{L,dyn}$ is about 1.8 and 1.0 L/kPa of lung and chest wall respectively. $C_{L,dyn}$ can be measured from the dynamic p_{tp}-V diagram. These values are about 20% larger than the static compliances.
The measurement of $C_{L,stat}$ is done passively (anaesthetized subjects or cadavers) with aid of the static P_{tp}-V diagram (see Lung gas transport 2, pressure, volume and flow). At an age of 25 year, C_{stat} (lung and chest wall together) is 2.7 L/kPa. It decreases logarithmically to 1.8 L/kPa at 75 year (*Respiration* 1989;55:176-80). This is in line with the age dependent increase of the bulk modulus K (see Elasticity 3: compressibility and bulk modulus) and shear modulus G (Elasticity 2: shear modulus) (see ref. 4). The age effect increases with p_{tp}. The change of K and G is mainly an alveolar wall property. Mostly it is found that smoking decreases $C_{L,stat}$.

Application

$\mathbf{R}_{aw}$ can be measured with several methods, such as body Plethysmography. Constrictions in the alveolar ducts (emphysema) considerably increase $\mathbf{R}_{aw}$. Many other disorders affect the lower airways, as chronic bronchitis, bronchiectasis, cystic fybrosis, asthma and bronchiolitis. This seems to be conflicting withy Fig. 1, but **E** is not taken into account, which virtually enhances $\mathbf{R}_{aw}$ substantially. but Generally they produce such severe constrictions that the $\mathbf{R}_{aw}$ increases substantially. $\mathbf{R}_{aw}$ of smokers seems about 20% enlarged.
The age effect of K and especially G considerably attributes to the increase of stiffness of lung parenchyma. Lungs with chronic obstructive pulmonary disease and lung emphysema with a1-antitrypsin deficiency have an increased K. The effect is consistent with the behavior found in other organs, such as systemic arteries.

More Info

Modelling $\mathbf{R}_{aw}$
With turbulent flow, i.e. a high $\dot{V}$, the tube resistance $\mathbf{R}_{tube}$ is (modified after Clarke and Flook, 1999, for 1 bar air pressure):

$$\mathbf{R}_{tube} = 3.2 \cdot 9.9 \cdot L \cdot D^{-19/4} \cdot \dot{V}^{3/4} \qquad (Pa \cdot L^{-1} \cdot s) \qquad (3a)$$

where L and D (diameter) in cm and $\dot{V}$ the stationary flow in L/s. The constant 9.9 comprises the dynamic viscosity η (18.8 µPa.s at 310 K) and ρ (1.14 g/L). The constant 3.2 comprises the effect of entrance length (=0.06Re·D) and tube imperfections (not perfectly cylindrical, round and smooth, and with curvatures). For Poiseuille flow, i.e. a low $\dot{V}$, $\mathbf{R}_{tube}$ is:

$$\mathbf{R}_{tube} = 1.2 \cdot 0.76 \cdot L \cdot D^{-4} \qquad (3b)$$

The transition from weak turbulent to laminar flow occurs at the generation number where the coloured curves of Fig. 1 merge the Poiseuille resistance curve. The lower curve indicates that only the first, say 10, generations (the upper airways) contribute substantially to $\mathbf{R}_{aw}$.

Due to the complicated geometry of the system, numeric results of modelling of $\mathbf{R}_{aw}$ only give rough approximations. $\mathbf{R}_{aw}$ can be found by summation of all tube resistances $\mathbf{R}_{tube}$ over all generations for a given $\dot{V}$. In all generations there is pressure drop. The resistance of the airways tree is:

$$R = k \cdot 2^{5/2} \cdot \pi^{-7/4} \cdot \eta^{1/4} \cdot \rho^{3/4} \cdot \sum_{i=0}^{i=26} (L_i \cdot D_i^{-9/4} \cdot (v / 2^i)^{3/4}) / 2^i , \qquad (4)$$

where k a constant, η the dynamic viscosity, ρ the gas density and i the generation number. v (= $\dot{V}/2^i$) is the flow per tube and 2^i the number of tubes in the i-th generation. The nominator 2^i is implied since in the i-th generation 2^i tubes are parallel. L_i and D_i can be obtained from the Hansen Ampaya non-scaled model with downscaled tube numbers for generation 24-26 (since paths did already end in sacs). The first five generations contribute most to the resistance. For $\dot{V}$ = 1 L/s, and all constants, lengths and diameters completed, **R** is about 29 $Pa \cdot l^{-1} \cdot s$ with k=1.
Under the assumption that some bifurcations are missing and due to tube imperfections **R** is larger. Also mouth and glottis should be included. Especially constriction and bifurcations (entrance effects) substantially increase **R**. Finally the flow is not stationary but dynamically (sinusoidal breathing). All together, this result in a correction factor k estimated at 4.

Literature
1. Clarke JR and Flook V. Respiratory function at depth, in: The Lung at Depth, Lundgren C.E.G. and Miller J.N. (Ed), Marcel Dekker Inc., New York, Basel, pp 1-71,1999.

2. Nunn JF, Applied Respiratory Physiology, 2nd ed. Butterworths, London-Boston, 1977.
3. Tammeling, GJ and Quanjer, PhH. Contours of Breahting, Boehringer Ingelheim BV, 1978, Haarlem.
4. Lai-Fook SJ and Hyatt RE. Effects of age on elastic moduli of human lungs. *J Appl Physiol* 89: 163-168, 2000.

Lung gas transport 4, cost of breathing

Principle

Theoretically a well as experimentally, the cost of respiration P_{res} is hard to estimate due to the non-linear behaviour of the resistance of the airways-alveolar-long tissue-chest wall system. Empirically it can be performed indirectly from O_2 consumption and from the area of the hysteresis loop of the V-p_{tp} diagram (see Fig. 1, 2 of Lung gas transport 2, pressure, volume and flow). As a rule of thumb it holds that $P_{res,rest}$ =0.1 Watt and at MVV, the Maximal voluntary ventilation (maintained for ca. 15 s) 0.5-1.0 J/breath. With a maximal breathing frequency (ca. 60/min), the maximal P_{res} is 30-60 Watt (e.g. Hesser et al. 1981). These estimates hold for adult, healthy subject up to an age of about 60 year.
The direct application of the general law 'power ≡ resistance times current^2 ' is the basis of two models. There are various models to estimate the resistance. Model [1], the classical one, divides the airways resistance $\mathbf{R_{aw}}$ in a constant and a term linear dependent on $\dot{V}$: $\mathbf{R_{aw}} = K_1 + K_2 \cdot \dot{V}$ (see Lung gas transport 2, pressure, volume and flow). In model [2] airways resistance is $\mathbf{R_{aw}} = k \cdot \dot{V}^{3/4}$. Model [3] utilizes the before mentioned histeresis loop of the V-p_{tp} diagram

When the visco-elastic and plasto-elastic losses in the lung tissue and chest wall tissue are taken into account, although hard to estimate, the total cost P_{res} is assumed to be a factor k' higher (twice for low flow to about triple for high flow) than $P_{vis,aw}$. The three models yield the following.

Model [1]

$$P_{res} = 0.06 + 0.0011 \cdot RMV^2 + 8.8 \cdot 0^{-6}\ RMV^3 \quad (k'=3) \qquad (1a)$$

This equation is recalculated from empirical measurements (Milic-Emili and D'Angelo, 1997) and RMV calculated from $\dot{V}$. The constant 0.06 was added to obtain 0.1 W at rest.

Model [2]

$$P_{res} = 0.1 + 31 \cdot 10^{-6} \cdot RMV^{11/4} \qquad (k'=3) \qquad (1b)$$

The 0.1 was added to obtain 0.1 W at rest. To obtain this equation the viscous cost of the airways was multiplied by 3 to account for other resistive losses.

Model [3]:

This model is imprecise since the shape of the loop changes considerably with V_{tidal} and frequency. Actually, viscous wall cost should be implemented too. Total cost is the area ADCEA of the above mentioned Fig. 1, multiplied by ca. 3.0 to account for viscous wall losses. Approximations systematically yield too low values (about factor of 5). Experimentally, the reported values are mostly determined by the integral of $\Delta p \cdot \dot{V}$, which can give useful results.
Fig. 1 gives the results of models [1] and [2].

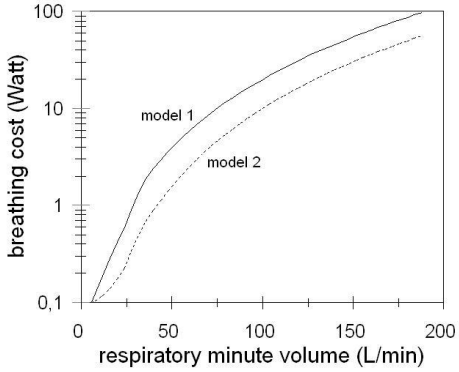


Fig. 1 Breathing cost of subject with normal height (175 cm) and weight (75 kg).

Application

Fig. 1 shows that for healthy lungs the cost is irrelevant except for the artificial condition of maximal breathing without other exertion and for the highest RMVs (nearly 10% of total energy expenditure). For most patients in (nearly) rest, the cost of breathing is such low that it can nearly always be afforded. However, moderate and high RMV can be impossible.
A low temperature of the air results in a higher airway resistance (especially for asthma patients). The reduced MVV reduces the maximal exertion. Only in extreme cases (e.g. nearly blocking of the airways) it becomes a limiting factor. The same holds for smokers, but here the main cause is their reduced diffusing capacity of the lung epithelium. Acute respiratory distress syndrome (ARDS; inflammation of the of the lung parenchyma) of adults and Infant respiratory distress syndrome (IRDS; lack of surfactant) gives rise to a severe increase of breathing cost. These patients may die due to exhaustion.

More Info

Model [1] The power of breathing with strict laminar conditions is $P_{res} = \Delta p \cdot \dot{V}$. For the airways system with its non-laminar flow, the classical approach is to substitute Δp by $\mathbf{E} \cdot V_{tidal} + K_1 \cdot \dot{V}_{in} + K_2 \cdot \dot{V}_{in}^2$ (eq. (2) in Lung gas transport 2, pressure, volume and flow), and doubling the cost due to the expiration. With a low V_{tidal}, the "elastic energy" stored during inspiration in the alveoli is used during the expiration (elastic recoil), although there is always some loss. This can be minimized by chosen an optimal breathing frequency and V_{tidal} for a given RMV. With RMV is 14 L/min, the optimal V_{tidal} is close to 1.2 litre. Completing the constants, P_{res} becomes (see Milic-Emili and D'Angelo, 1997):

$$P_{aw} = 3.95 \cdot \dot{V}_{in}^2 + 1.95 \cdot \dot{V}_{in}^3 \text{ (W)} \qquad (2a)$$

Replacing $\dot{V}_{in}$ (in L/s) by RMV and accounting for the other resistances, (2a) becomes (1a).
Model [2] An approach via the general law 'power ≡ pressure times current' is

$$P_{aw} = 120 \cdot 1000^{1.75} \cdot \dot{V}^{1.75}, \qquad (3a)$$

obtained by substitution of Δp by $\Delta p = 120 \dot{V}^{3/4}$ (Δp in Pa s/L and $\dot{V}$ in L/s) and converting L in m^3. Accounting for the other losses (k'=3), adding the constant 0.1 W and expressing $\dot{V}$ in RMV (L/min) P_{res} is found with (1b).
Model [3] Breathing cost can also be calculated with a more theoretically based model that also implies the tissues surrounding the lung (Tammeling and Quanjer, 1978). When a volume, e.g. a spirometer or a plastic back is filled with air, the work E is proportional with p·V (Joule). E is comprised of two parts, a viscosity (or resistance) part E_{vis}, given by the viscosity of the gas, and an elastic part (reciprocal to the compliance *C*) E_{el}. Both comprise a component of the thoracal wall E_W and of the airways system, E_L. Total work E for one inspiration then becomes:

$$E_{ins} = E_{W,vis} + E_{W,el} + E_{L,vis} + E_{L,el} \qquad (4a)$$

For $E_{L,el}$ it holds that:

$$E_{L,el} = \tfrac{1}{2} V_{tidal} \cdot \Delta p_{L,el}\ , \qquad (4b)$$

with $\Delta p_{L,el}$ the pressure to overcome the elastance of the alveoli and lung tissue during quit breathing. Eq. (4b) is actually the area of the triangle ABCA of Fig. 1 of Lung gas transport 2, pressure, volume and flow. However, $E_{L,el}$ is regained when the expiration takes place. The same recoil holds for $E_{W,el}$, which mainly comprises the compliance of the chest wall. At maximal airflow or during the early part of the expiratory maneuver, the main driving pressure is the intrinsic elastic recoil of the lungs. What remains for the cost are both viscous parts. $E_{L,vis}$ is the area of the hysteresis loop of Fig. 1. For very large flows the cost of the $P_{L,el}$ of inspiration and that of expiration do not longer cancel; there is a net loss which increases progressively with V and the same holds for $P_{W,el}$.

Literature

Hesser CM, Linnarsson D, Fagraeus L. Pulmonary mechanisms and work of breathing at maximal ventilation and raised air pressure. J Appl Physiol. 1981, 50:747-53.
Milic-Emili j and D'Angelo E, Work of breathing. In: The Lung, Crystal RG, West JB, Barnes PJ and Weibel ER, Lipencott-Raven, Philadelphia, New York, 1437-1446, 1997.
Tammeling, GJ and Quanjer, PhH. Contours of Breathing, Boehringer Ingelheim BV, 1978, Haarlem.

Oxygen analysis

Principle

Oxygen is the most vital gas for almost all biological species and a key-element in many chemical reactions. There are many physical and chemical principles to measure O_2 in a gas mixture and dissolved in fluid.

In gas mixtures the following techniques are most common:
Paramagnetic principle The analyzer measures the paramagnetic susceptibility of O_2.
High temperature zirconia sensor A type of electrochemical analyzer.
O_2 analysis using electrochemical cells The electrochemical O_2 sensor operates more or less like a battery.
Coulometric Based on the same principle as conventional electrochemical analyzers.

In liquids the following techniques are most applied:
Hersch cell It is based on electrolytic chemical reactions.
Polarographic A galvanic process with O_2 permeating through a membrane.
Optical fluorescence In this fluorescence application the presence of O_2 slows or reduces the amount of fluorescence.
Absorption spectroscopy The sensor uses laser diode absorption spectroscopy in the visible spectrum. See **More info** for a detailed description.

Application

Numerous in experimental medicine, pulmonology, clinical physiology, cardiology etc., space, environmental and sport medicine and in biochemistry, industrial microbiology etc.

More Info

In gas mixtures

Paramagnetic principle
The analyzer measures the paramagnetic susceptibility of O_2 (see Paramagnetism, diamagnetism and magnetophoresis) in the sample gas mixture by means of a magneto-dynamic measuring cell. The physical property which distinguishes O_2 from other gases is its much higher paramagnetism compared to other common gases. Advantages are the fast response time, high flow rates (up to 1 L/min), small sample volumes (2 mL), the extremely low drift, the absolute linearity and the negligible cross sensitivity against other sample gas components.
The measuring cell with a small mirror at its center is mounted in a strong inhomogeneous magnetic field. The paramagnetic O_2 strengthens the forces on the diamagnetic measuring cell and causes a shifting which is detected by a system consisting of light beam, mirror and a photo cell.
A compensation current is induced via the feedback coil on the measuring cell that leads to a zero voltage. The required current is linearly proportional to the O_2 concentration.

High temperature zirconia sensor
A zirconia O_2 sensor uses a stabilized zirconia ceramic. Platinum electrodes are painted on the outside and inside of the O_2 sensor. The sensor is heated above 600^{o} C. At this temperature, the crystal lattice structure expands allowing O_2 ions to migrate through the O_2 sensor. Oxygen breaks down into O_2-ions at the electrodes of the sensor and travels through the sensor between the electrodes. The flow of O_2-ions is either outward or inward of the O_2 sensor depending on the O_2 concentration in the sample gas compared to the O_2 concentration in a reference gas. Advantages of this type of O_2 sensor are a wide measuring range and fast response. The sensor gives mV output, which is converted to $\%O_2$.

O_2 analysis using electrochemical cells
The electrochemical O_2 sensor operates much like a battery. Oxygen gas flows to an electrode and becomes a negatively charged hydroxyl OH-ion. This ion moves through an electrolyte in the O_2 sensor to a positively charged electrode typically made of lead. The OH-ion reacts with Pb and releases electrons. The electron flow is measured and be converted to an O_2 concentration. Advantages of this type of O_2 sensor include ability to measure O_2 in hydrocarbon or solvent streams, accurate, and inexpensive. Low O_2 measurement down to 0.01ppm is possible.

Coulometric The non-depleting coulometric cell (referring to Coulomb) operates on essentially the same principle as conventional electrochemical analyzers. However, the non-depleting electrodes provide the capability for continuous O_2 measurement in the sample gas with no degradation of the electrodes (no consumable lead anode), and, thus, false low O_2 readings due to electrode degradation have been eliminated. Sample gas diffuses through a simple gas diffusion barrier to the cathode in an electrolyte solution. Oxygen is reduced at this electrode to hydroxyl ions. Assisted by a KOH electrolyte, the ions migrate to the anode, where they are oxidized back to O_2. Unlike replaceable galvanic electrodes used as the driving mechanism for the reaction, an external electromagnetic force of 1.3 V DC drives the reaction. The resulting cell current is directly proportional to the O_2 concentration in the gas stream. Because of the inherent stability of the electrodes, sensitivity levels to less than 5ppb (parts per billion) of O_2 can be achieved.

Liquid dissolved O_2

Hersch cell It is based on electrolytic chemical reactions and detects O_2 at ppb level. The galvanic cell (or Hersch cell) is composed of essentially two electrodes immersed in an electrolytic fluid, usually concentrated KOH. No external potential is applied.
In the membrane version, O_2 diffuses into the sensor, usually through a Teflon membrane, where it is reduced electrochemically to hydroxide at the silver cathode. The HO-ions migrate to the Pb anode, where Pb is reacted to form lead oxide. The current generated from the reduction/oxidation reaction is proportional to the O_2 concentration of the sample gas. Therefore, the sensor has an absolute zero.
In the version without a membrane the electrodes are exposed to the wastewater, which is used as the electrolyte. The sensor is calibrated in O_2 saturated water. As water has different pH and conductivity values to that of wastewater, there can be no certainty that the measurement, especially below 2 ppm (parts per million) is accurate.

Polarometric (membrane) (See Voltammetry) This is also called, and more correctly amperometric since the voltage is held constant. The anode and cathode are immersed in an electrolyte, into which O_2 permeates through a membrane. It differs from a galvanic sensor in that the anode has to be polarized at the specific galvanic voltage (of the electro potential series) of O_2, being –650 mV. Without O_2 in the solution, also a small current is flowing. The polarographic sensor needs calibration with liquids with known O_2 concentrations. The voltage-concentration relation is linear. As the sensor ages, its zero offset changes.

Optical fluorescence This is sometimes also called luminescence. In this Fluorescence application the presence of O_2 slows or reduces the amount of fluorescence. As light is constantly produced in a bandwidth rather than an absolute wavelength, there is no absolute zero. However, as the measurement is frequency (rate of change) based, there is no drift as long as the signal strength is reasonable. Measurement in absolute dark (contamination, sensor damage) is a prerequisite. A major problem is that the fluorophore is soluble in water. So the material needs to be bonded to another material, e.g. a ruthenium (fluorophore)/silicon (bond) matrix, which impedes or stops the degradation of the fluorescing material.

Absorption spectroscopy The sensor uses Laser diode absorption Spectroscopy in the visible spectrum, similar to the absorption method used to measure CO_2, N_2O, and anesthetic agents in the IR (infra red) spectrum.
However, O_2 absorption is just visible (760 nm) without interference or absorption by other ventilation or anesthetic gases. The emission and the absorption line width of O_2 are very narrow, < 0.01 nm, compared to some 100 nm for the CO_2 absorption band at atmospheric pressure.
As the O_2 concentration increases, the light intensity is attenuated, and the photodetector response (thermally adjusted emitted frequency) varies linearly with the O_2 concentration.

Almost all dissolved O_2 analyzers are calibrated using air as the reference. At 25°C, the saturation value of O_2 in water is 8.4 ppm.

In addition to this ''classical'' methods new techniques are for instance photo-acoustic and magneto-acoustic techniques for O_2 monitoring in the open un-intubated airway.

Plethysmography

Principle

Plethysmograpy provides the most accurate measure of volumes, such as that of the whole body (submerged in a water filled chamber), the lungs and extremities.

Use for lungs

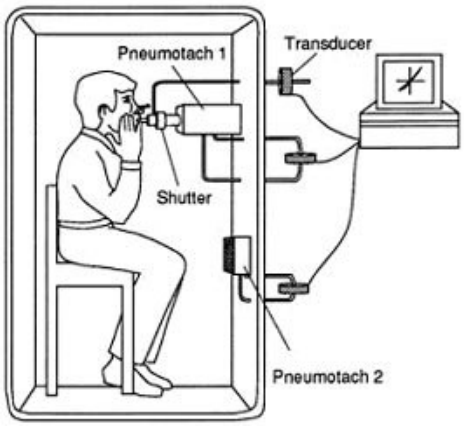


Fig 1. The lung-plethysmograph and the principle of measurement

It measures the functional residual capacity (FRC) as well as the total lung capacity (TLC).

Use for limbs

Some plethysmograph devices are attached to arms, legs or other extremities and used to determine circulatory capacity.

Application

Plethysmography is the most accurate measure of lung volumes. The difference in full versus empty lungs can be used to assess diseases.
Since with a plethysmograph also pressures and flows can be measured, it can be used for precise measurement of resistance and compliances. Therefore it is ideal for assessing airway passage restrictions and airways obstruction reversibility. An obstructive disease will show increased FRC because some airways do not empty normally, while a restrictive disease will show decreased FRC. Body plethysmography is particularly appropriate for patients who have air spaces which do not communicate with the bronchial tree; in such patients the gas dilution method (see Pulmonary tests and instruments) would give an incorrectly low reading.

More info

At the end of normal expiration (FRC remains) and, during a second measurement, at the end of maximal inspiration (lung content is V_{lung} or TLC), the mouthpiece is closed. The patient is then asked to make an inspiratory effort. As the patient tries to inhale the glottis is closed and the lungs expand due to decreasing pressure. This expands the lung volume (Fig. 2). This, in turn, increases the pressure with ΔP_{box} within the box since it is a closed system and the volume of the body compartment has increased by $\Delta V_{box} = -\Delta V_{lung}$ (provided that temperatures are constant). This is measured with pneumotach2 (see Pulmonary tests and equipment) after opening it. Since the initial pressure P_{box} and the ΔP_{box} (manometer measurements) are also known ΔV_{lung} can be found with Boyle's Law:

$$\Delta V_{lung} = -V_{box}\Delta P_{box}/(\Delta P_{box}+P_{box}), \quad (1)$$

where V_{box} the empty box volume minus the patient volume (estimated from weight and body fat measurement).

By applying the law again, now for the lung, FRC and TLC can be is found:

$$V_{lung} = -\Delta V_{lung}(\Delta P_{lung}+P_{lung})/\Delta P_{lung}, \quad (2)$$

with both pressures known from the manometer in the mouthpiece.

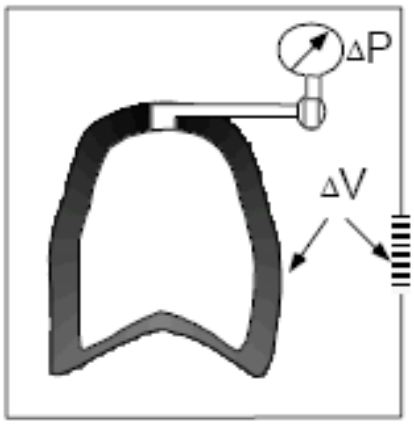


Fig. 2 Principle of determination

Pneumotachography

Principle

The pneumotachograph or the pneumotach provides accurate flows over the entire physiological range in the most demanding applications and exercise, pulmonary function testing, and nutritional assessment. The measurement is often based on a bi-directional differential pressure Pitot tube. The output is unaffected by condensation, turbulence, gas viscosity, temperature or position.

More Info

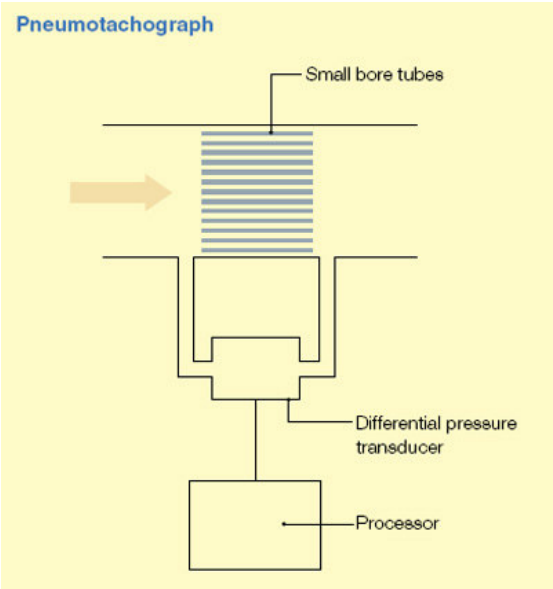


Fig. 1 (from http://www.frca.co.uk/article.aspx?articleid=100390)

Another approach is to measure a pressure drop across a resistance in the gas pathway. The measurement is rapidly and accurately using a differential pressure transducer, from which flow rate and

volume are calculated. The resistance is usually designed to produce laminar flow, so that the flow rate is directly proportional to the measured pressure drop. This is achieved using a series of small-bore tubes arranged in parallel (Fig. 1), through which the gas flow must pass. A heating element is sometimes incorporated to prevent the build-up of condensation that could compromise accuracy. The total resistance added by the pneumotachograph should be small so that it can be used in a spontaneously breathing patient.
Measurement can be made at various points in the breathing system (or ventilator), and a pair of sensors is often used so that inspired and expired tidal volumes can be measured independently. In addition to the differential pressure across the chamber, the absolute pressure in the airway can be easily measured. When linked to the recorded tidal volume, compliance can be calculated (see Lung gas transport 1, basic principles) and displayed in real time.

Pulmonary tests and instruments

Working principles of volumetric and flow measurements

Direct measurement
Gas volumes (and associated flows) can be measured directly using bulk filling of an enclosed space of known volume. Instruments using direct measurement include the industrial gas meter, vitalograph and water-displacement spirometer. Because of their intractability, their use in clinical practice is limited.

Indirect measurement
For clinical use, measurement is usually made indirectly, using a property of the gas that changes in parallel to flow or volume and which can be more easily determined.
Pressure drop across a resistance
– as flow occurs through a resistance, a pressure drop occurs. This effect can be used to calculate flow by keeping the resistance constant and measuring the pressure change as the flow varies, as in a pneumotachograph.
Mechanical movement
– flowing gas has kinetic energy related to its velocity, which can be converted into a measurable value by rotation of a vane (e.g. a spirometer) or bending a flexible obstruction, transducing this to produce an electrical signal.
Heat transfer
– gas flowing past a heated element acts to cool it, as in a hot-wire anemometer.
Ultrasound interference
– the velocity of an ultrasound signal is increased by a gas flowing alongside it in the same direction, and decreased if the gas is flowing against it.

Volume Units
Since the mass of gas in a unit volume is dependent on pressure and temperature (see Gas laws) they have to be specified with their pressure and temperature. Common are STPD, BTPS and ATPS. See for their definitions Gas volume units, STPD, BTPS and ATPS.

Volumetric and flow measurements

Vitalograph
The vitalograph is used specifically to record a single vital capacity breath. Its design uses an expanding bellows. Volume is displayed on the vertical axis and time on the horizontal axis, so that the pattern of expiration is shown as well as the volume.

Forced **E**xpiratory **V**olume in 1 second
The FEV1 is the most widely used parameter to measure the mechanical properties of the lungs and is measured with e.g. a spirometer (see Spirometry). FEV1accounts for the greatest part of the exhaled volume from a spirometric maneuver and reflects mechanical properties of both the large airways and medium-sized airways. In a normal flow-volume loop, the FEV1occurs at about 75% of the forced vital capacity (FVC). This parameter is reduced in both obstructive and restrictive disorders. In obstructive diseases, FEV1 is reduced disproportionately to the FVC and is an indicator of flow limitation. In restrictive disorders FEV1, FVC and total lung volume are all reduced, and in this setting FEV1 is a measure of volume rather than flow.

Gas dilution method
The method measures lung volumes and is based on:

– starting at the functional reserve capacity (FRC),
– known volume of tracer gas at known concentration,
– measurement of final concentration after rebreathing.
It is easily performed, but less accurate when an obstruction is present.

Hot wire anemometry see Hot wire anemometry

Mechanical flow transducer
A device using mechanical movement is the flow transducer, also used in IC-ventilators. The gas flow is split so that measurement is made in a small side channel. This comprises a thin metal disc supported on a flexible pin, which is mounted perpendicular on the flow direction (Fig. 1). This results in bending backwards by the flow. A strain gauge (comprising piezo crystals, see Piezoelectricity) situated immediately behind the pin is compressed as it is bent, with a force dependent on the flow. The resulting electrical signal is processed to calculate the flow rate with a high degree of accuracy.

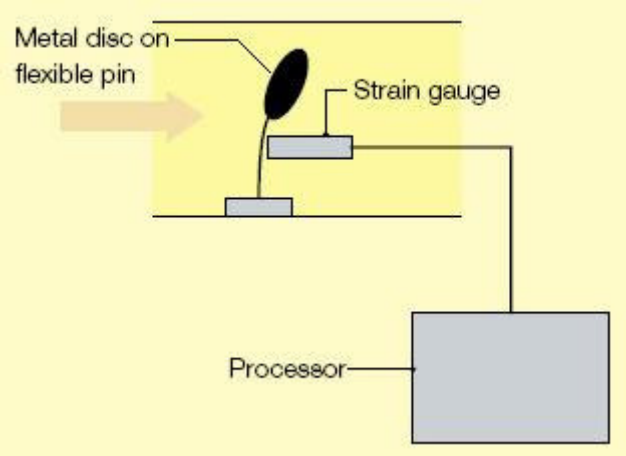


Fig. 1 Mechanical flow transducer

Peak Flow Meter
A peak flow meter is a small, hand-held device used to measure the degree of restriction in the airways. It measures the Peak Expiratory Flow, PEF.
The meter is a specialized vane meter (see below) that measures the maximum flow rate only, without calculation of the associated volume. It is modified so that movement of the vane by expiratory gas results in a steadily enlarging pathway for gas escape. The final position to which the bar has moved corresponds to the peak expiratory flow rate.
Its use (in the clinic and at home) is to assess conditions such as asthma, where the problem is largely confined to airway resistance, which limits the expiratory flow rate. Although its use is limited in this respect, it is a very simple and reliable bedside test.

Plethysmography, see Plethysmography

Pneumotachography, see Pneumotachography

Spirometer see Spirometry

Vane meter
The most common vane meter is the Wright's spirometer, in which the gas flow is directed tangentially to strike a rotating vane in the gas pathway (Fig. 2). Originally the design was strictly mechanical, but modern versions use a light source and photodetector positioned across the vane to count its rotation. It is lower at low flows (because of friction) and higher at high flows than a linear relationship would predict because of the momentum (mass times velocity).

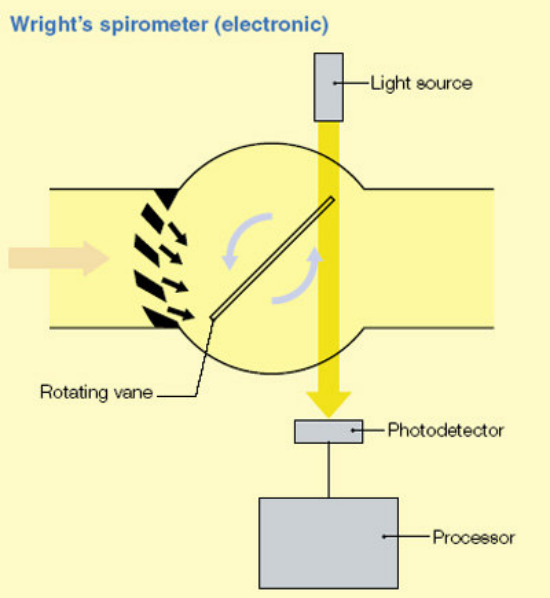


Fig. 2 Vane flow meter

Ultrasonic flow meters work on the principle that when an ultrasound signal is being transmitted within a flowing gas, the medium in which the sound pulse propagates, its velocity changes in proportion to that of the gas flow. When the gas flow and ultrasound signal are in the same direction, an increase in signal velocity occurs, causing a decrease of the time of arrival. Conversely, when the signal is against the direction of gas flow, arrival time is lengthened. The change it time is:

$$\Delta t = V_{flow}\cos\alpha/L, \quad (1)$$

where α is the angle between sensor and flow direction and L the length between transmitter and receiver. (1) directly gives the flow speed and volume flow ($\pi r^2 V_{flow}$), where r is tube radius. Notice that sound speed is dependent on temperature. In dry air at 0 °C it is 331.5 m/s and at 37 °C 353.2 m/s. Raising the humidity from 0 to 100% (expired air) gives an increase of about 3%.

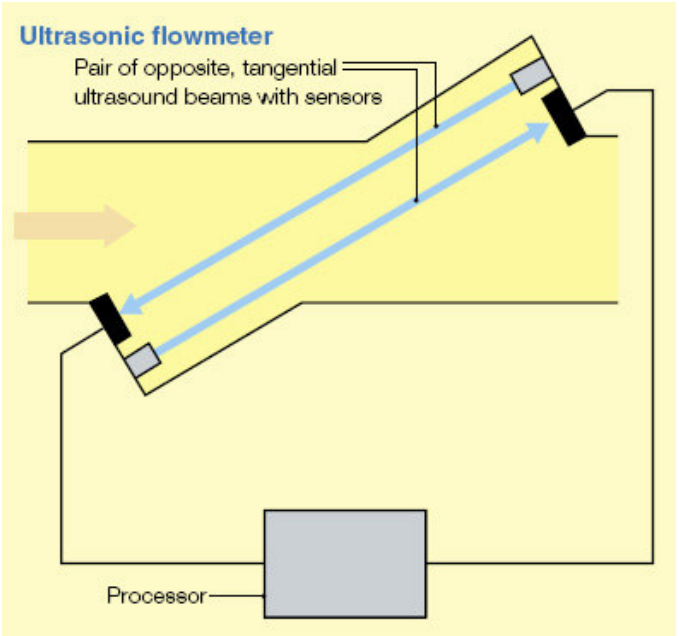


Fig. 3 Ultrasonic flow meter

The usual design is to incorporate a pair of ultrasound beams aimed in opposite directions, each with a sensor.

Some other instruments

Capnography See Capnography

Diffusing Capacity for CO, DLCO
DLCO measures transfer of a soluble gas, mostly CO, from the airspace to the blood. It is presented as mL CO·min^{-1}·mm Hg^{-1} (STPD, see Gas volume units, STPD, BTPS and ATPS). The alveolar volume (VA) at which the DLCO is measured is also commonly reported (BTPS). The ratio of DLCO to VA is also commonly reported as the DL/VA or simply D/VA.
The method is based on a single deep breath with a known concentration of CO for a defined amount of time and measurement of exhaled CO. The CO taken up is calculation. The uptake is proportional to surface area for gas exchange. 'Confounders' are Hb available to bind CO, CO present in blood and thickness of alveolar membrane. It is possible to correct for them. Noble gases, often used in experimental research, don't have these drawbacks. The outcome is evaluated in relation to the measured vital capacity.
The outcome is reduced in case of destruction of alveoli, emphysema, pulmonary fibrosis and lung infiltration, loss of pulmonary vascular bed, pulmonary hypertension, and dysfunction of pulmonary embolism.
It is normal with asthma and elevated in increased pulmonary blood flow or volume (exercise, mild congestive heart failure). See for applications, (contra)indications etc.
http://www.rcjournal.com/online_resources/cpgs/sbcmdc99cpg.html

Measurement of resistance and compliance with the interrupter technique
Measurements of airway resistance using the interrupter technique (Rint) is useful for assessing lung function, especially in children of preschool age since cooperation of the subject can be minimal. The principle is that, following an instantaneous interruption of airflow at the airway opening (by closing a valve), there is an instantaneous equilibration of pressure between the alveoli and the airway opening, the mouth (behind the valve). The technique further assumes that there is a single value of alveolar pressure. Following the occlusion, a rapid change in pressure is seen which is equal to the airways resistive fall in pressure between the alveoli and the airway opening. Dividing this pressure change by the flow occurring immediately before the occlusion gives Rint (analog to Ohm's law). The initial rapid pressure change is followed by a second slower change in pressure which is related to the visco-elastic properties of the respiratory tissues, together with any gas redistribution following the occlusion. This effects are dependent on the timing of the occlusion in the phase of respiration.
The equipment includes a rapidly closing valve that occludes the airway for 100 ms before allowing normal respiration to resume. In practice, Rint is rather constant during the tidal cycle. Generally, Rint is measuring during the expiratory phase of respiration at peak tidal flow.
The high frequency (100-2000 Hz) version of this technique is suitable to estimate airway wall resistance and compliance, especially in infants with wheezing. The impedance data (comprising resistance and compliance, see Impedance and Lung gas transport 3: resistance and compliance), spectrally analysis yield also information about airway geometry (diameters), e.g. to quantify bronchoconstriction, or bronchodilation. The measure is only slightly influenced by lung and chest wall tissues.
The forced oscillation technique (FOT) is a method to assess resistances and compliances. A typical application is assessment of bronchial hyperresponsiveness. FOT employs small-amplitude pressure oscillations superimposed on normal breathing.
See for more info http://thorax.bmjjournals.com/cgi/content/full/58/9/742 and
http://erj.ersjournals.com/cgi/content/full/22/6/1026

Pulse Oximetry see Pulse oximetry

Oxygen analyzers see Oxygen analysis

Literature

http://www.frca.co.uk/article.aspx?articleid=100390 (Most figures are adopted from this source)
http://thorax.bmjjournals.com/cgi/content/full/58/9/742.)
http://erj.ersjournals.com/cgi/content/full/22/6/1026

Pulse oximetry

Principle

Pulse oximetry is a non-invasive method to monitor the blood oxygenation in the capillary bed, expressed as arterial oxy-Hb/total Hb in % and called SaO_2. (SpO_2, mostly used instead of SaO_2, is an undefined, sloppy acronym). Modern meters also measure pulse rate and pulse strength.

The principle of pulse oximetry is based on the red and infrared (IR) light absorption characteristics of oxygenated and deoxygenated Hb. Oxy-Hb absorbs more IR light and allows more red light to pass through (Fig. 1). Deoxy-Hb absorbs more red light and allows more infrared light to pass through.

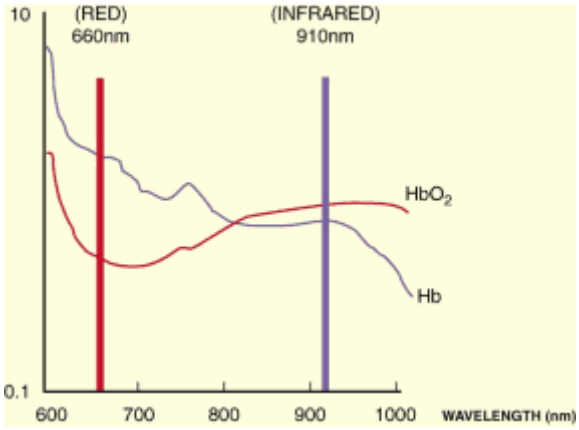


Fig. 1 (from http://www.frca.co.uk/article.aspx?articleid=100390)

Pulse oximetry uses a light emitter with red and infrared light (IR) emitting diodes (LEDs, at 660 nm and 910-940 nm, respectively) that shines through a reasonably translucent site with good blood flow. A red and IR light emitter is (usually) placed on top of a fingertip, toe, pinna or lobe of the ear and a photodetector (red and IR) just opposite at the other side. The diodes flash at a rate of approximately 30/s. The diodes are switched on in sequence, with a pause with both diodes off. This allows compensation for ambient light. The microprocessor analyses the changes in light absorption during the arterial pulsatile flow and ignores the non-pulsatile component of the signal (which results from the tissues and venous blood).

Application

It is especially useful in an intensive care setting, for assessment of emergency patients, determining the effectiveness of or need for supplemental O_2, and monitoring during anesthesia. In addition, it is increasingly used in space, aviation, altitude, sports and environmental medicine.
Recent pulse oximetry technology shows significant improvement in the ability to read through motion (including shivering) and low perfusion (cold or vasoconstriction, e.g. due to vasopressor agents); thus making pulse oximetry very useful. Falsely high or falsely low readings will occur when Hb is bound to CO or cyanide. Therefore, under specific conditions, arterial blood gas tests are still used to supplement or validate pulse oximeter readings.

It should be noted that this is a measure solely of oxygenation and is not a substitute for blood gas analyses since it gives no indication of CO_2 levels, blood pH, or $NaHCO_2$.

More Info

There are two methods of sending light through the measuring site: transmission and reflectance. In the transmission method, the emitter and photodetector are opposite of each other with the measuring site in-between. The light can then pass through the site. In the reflectance method, the emitter and photodetector are next to each other on top of the measuring site. The light bounces from the emitter to the detector across the site. The transmission method is the most common type used and for the discussion below the transmission method will be implied.
The oxygen saturation is estimated by measuring the transmission of light through the pulsatile tissue bed. This is based on the Beer-Lambert law: the intensity of transmitted light through a material (here the tissue) decreases exponentially as a function of the concentration of the absorbing substance *and* decreases exponentially as a function of the path length through the material. The law is applied for both

types of molecules *and* for both wavelengths, yielding two equations with two unknown concentrations and the unknown path length. Since not concentrations themselves are relevant but the concentration ratio, this ratio can basically be solved.
The R/IR is compared to a “look-up” table (made up of empirical formulas) that convert the ratio to a SaO_2 (SpO_2) value. Typically an R/IR ratio of 0.5 equates to approximately 100% SaO_2, a ratio of 1.0 to approximately 82% SaO_2, while a ratio of 2.0 equates to 0% SaO_2.
For a reliable calculation some complications should be eliminated. The major problem is to differentiate between the arterial pulsation, the Hb light absorption and the absorption in intermediate tissues (e.g. skin) and venous blood. However, with each heart beat there is a surge of arterial blood, which momentarily increases arterial blood volume across the measuring site. This results in more light absorption during the surge.
The light absorbed by non-pulsatile tissues is constant (DC). The alternating absorption component (AC) is the result of pulsatile blood pulsations. The photodetector generates a voltage proportional to the transmitted light. The AC component of the wave accounts for only 1-5% of the total signal. The high frequency of the diodes allows the absorption to be calculated many times per second. This reduces movement effects on the signal.
Since peaks occur with each heartbeat or pulse, the term “pulse oximetry” was coined.
The microprocessor analyses both the DC and AC components at 660 nm and 940 nm. Modern pulse oximeters may use more than two wavelengths.

Spirometry

Basic Principles

Spirometry, the most common of the Pulmonary Function Tests (PFTs), is the measurement of the amount (volume) and speed of air that can be inhaled and exhaled. Results are usually given in both raw data (e.g. L/s) and percent predicted, i.e. the test result as a percent of the "predicted values" for the patients of similar characteristics (height, weight, age, sex, and sometimes race).
Spirometer comes in many different varieties. Many produce a Flow-Volume diagram, see Fig. 1. Most spirometers also display a Volume-Time curve, with volume (L) along the vertical axis.

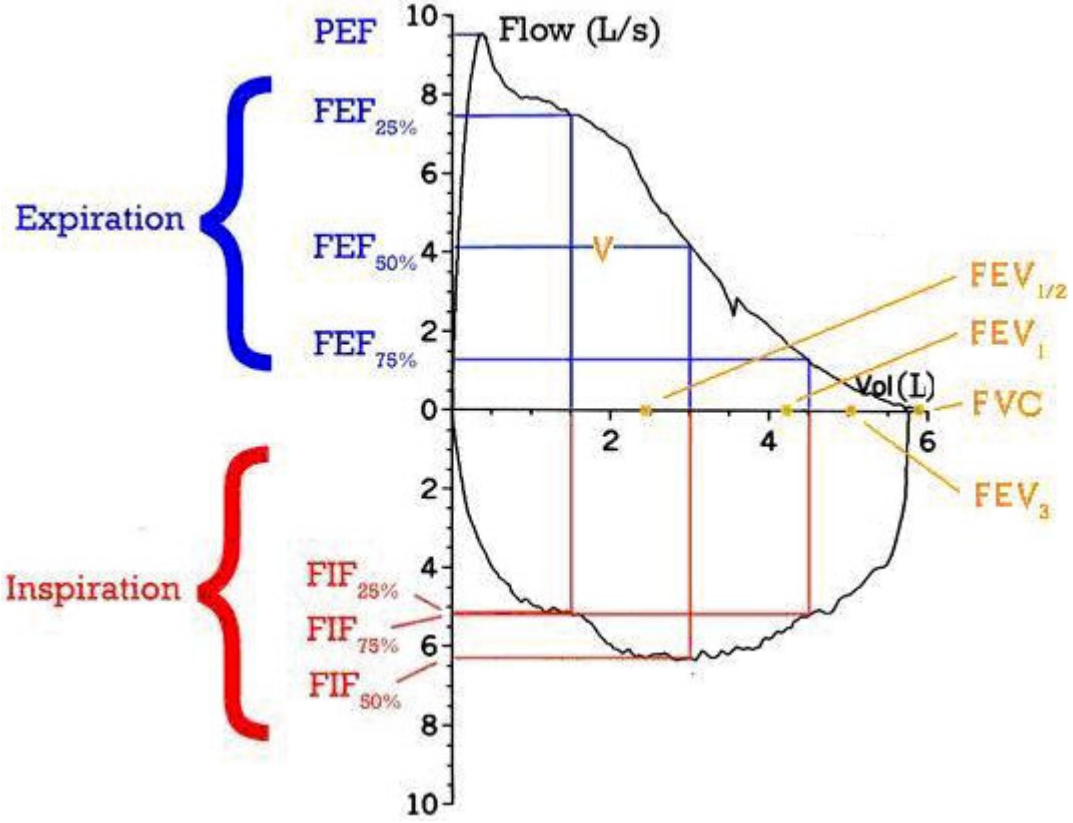


Fig. 1 Flow-Volume diagram. Positive values represent expiration, negative values represent inspiration. The trace moves clockwise for expiration followed by inspiration. (Note the FEV_1, $FEVA_{1/2}$ and FEV_3 values are arbitrary in this graph and just shown for illustrative purposes, they must be recorded as part of the examination). For the definitions of the various parameters see the list below.

The basic spirometry test itself is simple, and varies slightly depending on the equipment used. Generally, the patient is asked to take the deepest breath he can, and then exhale into a machine as hard as possible, for as long as possible, followed by a rapid deep breath in. Sometimes, the test will be preceded by a period of quiet breathing in and out from the machine, or the rapid breath in will come before the hard exhalation. During the test, soft clips are used to prevent air escaping through the nose. Sometimes, filter mouthpieces are used to prevent the spread of germs, or in other cases a length of tubing is placed between the patient and the machine. The test is repeated, usually at least three times and often up to as many as eight times, to insure that the results are accurate and repeatable. With provocation tests some drug is administered to study a lung disorder or to study hypersensitivity to some allergic agent.
Due to the patient cooperation required, spirometry can only be used on children old enough to comprehend and follow the instructions given (typically about 4-5 years old), and only on patients that are able to understand and follow instructions - thus, this test is not suitable for patients that are unconscious, heavily sedated, or have limitations that would interfere with vigorous respiratory efforts. Other types of PFTs are available for infants (see e.g. Rint in Pulmonary tests and instruments) and unconscious persons.

Explanation of Common Test Values

FVC: Forced Vital Capacity - This is the total amount of air that you can forcibly blow out after full inspiration, measured in liters.
FEV 1: Forced Expiratory Volume in 1 second - This is the amount of air that can be forcibly blow out in one second, measured in liters. Along with FVC it is considered one of the primary indicators of lung function.
FER: This is the ratio of FEV1/FVC, which showing the amount of the FVC that can be expelled in one second. In healthy adults this should be approximately 0.8.
PEF: Peak Expiratory Flow - This is the speed of the air moving out of your lungs at the beginning of the expiration, measured in L/s.
FEF 25-75% or 25-50%: Forced Expiratory Flow 25-75% or 25-50% - This is the average flow (or speed) of air coming out of the lung during the middle portion of the expiration (also sometimes referred to as the MMEF, for maximal mid-expiratory flow).
FIF 25-75% or 25-50%: Forced Inspirtory Flow 25%-75% or 25%-50% - This is similar to FEF 25%-75% or 25%-50% except the measurement is taken during inspiration.

Application

Spirometry is an important tool used for assessing lung diseases such as asthma, cystic fibrosis, and COPD. Flow-volume diagrams are of great help to diagnose low or high located constrictions in the airways.

More Info

Over the years volumetric spirometers with a water bell were the most common ones. For flow-spirometry for instance the Fleisch-pneumotach (differential pressure measurement (on the basis of Bernoulli's principle). The turbine (rotation velocity of a hand-held small turbine screw), the Pitot tube (principle of the water jet steam) or the Hot wire anemometery can also be the instrumental basis of a flow-spirometer.

Standard conditions for temperature and pressure

In physical sciences, standard sets of conditions for experimental measurements allow comparisons to be made between sets of data obtained under different conditions.
The most used standards are those of the International Union of Pure and Applied Chemistry and the National Institute of Standards and Technology (IUPAC) but are far from being universal standards. Other organizations have established a variety of alternative definitions for their standard reference conditions. The current version of IUPAC's standard is a temperature of 0 ^{o}C (273.15K, 32 ^{o}F) and an absolute pressure of 100 kPa (14.504 psi), while NIST's version is a temperature of 20 ^{o}C (293.15K, 68 ^{o}F) and an absolute pressure of 101.325 kPa (= 1 atm = 14.696 psi).

Thermodynamic equations for an ideal gas

$PV^n = Constant$

Process	Isobaric	Isochoric	Isothermal	Adiabatic
Variable =>	Pressure	Volume	Tempemature.	No Heat Flow
Quantity Constant =>	$\Delta P = 0$	$\Delta V = 0$	$\Delta T = 0$	$Q = 0$
n	0	∞	1	$\gamma = C_p/C_v$
First Law	$\Delta U = Q - W$	$\Delta U = Q$ $W = 0$	$\Delta U = 0$ $Q = W$	$\Delta U = -W$ $Q = 0$
Work $W = \int P dV$	$P(V_2 - V_1)$	0	$P_1 V_1 \ln\left(\frac{V_2}{V_1}\right)$	$\frac{P_1V_1 - P_2V_2}{\gamma - 1}$
Heat Flow Q	$m C_p (T_2 - T_1)$	$m C_v (T_2 - T_1)$	$P_1 V_1 \ln\left(\frac{V_2}{V_1}\right)$	0
Heat Capacity	C_p	C_v	∞	0
Internal Energy $\Delta U = U_2 - U_1$	$m C_v (T_2 - T_1)$	$m C_v (T_2 - T_1)$	0	$m C_v (T_2 - T_1)$
Enthalpy $\Delta H = H_2 - H_1$ $H = U + PV$	$m C_p (T_2 - T_1)$	$m C_p (T_2 - T_1)$	0	$m C_p (T_2 - T_1)$
Entropy $\Delta S = S_2 - S_1$ $= \int dQ/T$	$mC_p \ln \frac{T_2}{T_1}$	$mC_v \ln \frac{T_2}{T_1}$	$nR \ln \frac{V_2}{V_1}$	0*
Ideal Gas Relations $\frac{P_1V_1}{T_1} = \frac{P_2V_2}{T_2}$ $PV = NkT$	$P_1 = P_2$ $\frac{V_1}{T_1} = \frac{V_2}{T_2}$ $\frac{T_1}{T_2} = \frac{V_1}{V_2}$	$V_1 = V_2$ $\frac{P_1}{T_1} = \frac{P_2}{T_2}$ $\frac{T_1}{T_2} = \frac{P_1}{P_2}$	$T_1 = T_2$ $P_1V_1 = P_2V_2$ $\frac{P_1}{P_2} = \frac{V_2}{V_1}$	$Q = 0$ $(S_1 = S_2)$ * $P_1V_1^\gamma = P_2V_2^\gamma$ $\frac{T_1}{T_2} = \left(\frac{V_2}{V_1}\right)^{\gamma-1}$

* For Reversible Processes

n c_v = m C_v **c_p - c_v = R** **n R = N k** **n c_p = m C_p**
γ = C_p/C_v = c_p/c_v = Ratio of Specific Heats
C_p = Constant Pressure Specific Heat Capacity (J/kg/ °C)
C_v = Constant Volume Specific Heat Capacity (J/kg/ °C)
c_p = Molar Constant Pressure Heat Capacity (J/mole/oC)
c_v = Molar Constant Volume Heat Capacity (J/mole/oC)

VO_{2max}

Principle

VO_{2max} is the maximum amount of oxygen, in mL, one can use in 1 min per kg of bodyweight, hence in mL/kg·min. It is also called the (mass) specific VO_{2max}. VO_{2max} is also known as maximal O_2 consumption or maximal O_2 uptake of the whole body, the absolute VO_{2max}. Then, it is mostly expressed in L/min for the whole body. For experimental purposes and for comparing the aerobic performance of for instance endurance sports athletes, the former expression is highly preferred.

Measuring VO_{2max}

Accurately measuring VO_2 max involves a ramp test (treadmill or cyclo-ergometer) in which exercise intensity is progressively increased until exhaustion while measuring the rate and O_2 concentration of the inhaled and exhaled air. (The name ergometer comes from the erg, an old energy measure. It is defined as defined as 1 cm.g/s).

A simple but adequate test for a rough estimate is the well-known Cooper test in which the distance (in km) covered by running or swimming 12 minutes is measured. For running the estimate of VO_{2max} is:

$$VO_{2max} = (\text{distance} - 0.505)45 \text{ (mL/min·kg)}. \quad (1)$$

There exist many other more or less reliable tests and VO_{2max} models (calculators) to estimate VO_{2max} (see ref. 1 and 2). The parameters of the calculators are always age and BMI (body mass index, mass/height^2 in kg/m^2), and often also sex and a measure of endurance sport activity, generally hours/week. Sometimes HR_{max} is used (ref. 2) by applying the Fick principle (see **More Info**). VO_{2max} declines with age, as illustrated in Table 1, which presents norm values as function of age (calculated from ref. 1).

Table 1 VO_{2max} (mL/min·kg)

Age (year)	<30	30-40	40-50	50-60	60-70
men	39	36½	33	31½	29½
women	35	32½	30	28	27

or in formula:

$$VO_{2max,man} = 48.4 - 0.43A + 0.0020A^2$$
$$VO_{2max,woman} = 44.1 - 0.42A + 0.0024A^2, \quad (2)$$

where A is age (year, > 20).

Application

VO_{2max} is utilized in sports medicine, and in daily life in the fitness industry and sports, especially with top athletes.

More info

.

Fick Equation

VO_2 (L/min) is properly determined by the Fick Equation (see Diffusion: Fick's laws):

$$VO_2 = Q(C_aO_2 - C_vO_2), \quad (3)$$

where Q is the cardiac output (beats/time x stroke volume, i.e. HR·SV) of the heart, C_aO_2 is the arterial oxygen content, and C_vO_2 is the venous oxygen content. By applying the principle two times, for HR_{max} and HR_{rest}, finally the following expression for VO_{2max} is obtained:

$$VO_{2max} = (HR_{max}/HR_{rest})\cdot(SV_{max}/SV_{rest})\cdot((C_aO_2 - C_vO_2)_{max}/(C_aO_2 - C_vO_2)_{rest})\cdot VO_{2rest} \text{ (L/min)}. \quad (4)$$

Since the ratio's (SV_{max}/SV_{rest}) and $(C_aO_2 - C_vO_2)_{max}/(C_aO_2 - C_vO_2)_{rest}$, and the specific VO_{2rest} are rather constant over subjects, they are about 3.4, 1.3 and 3.4 (mL/kg/min) respectively, one obtaines:

$$VO_{2max} = (HR_{max}/HR_{rest})\cdot 3.4\cdot 1.3\cdot 3.4\cdot W = 15\cdot W(HR_{max}/HR_{rest}), \text{ (mL/min)} \quad (5)$$

where W is body weight (kg).
For the specific VO_{2max} one obtaines:

$$VO_{2max} = 15(HR_{max}/HR_{rest}), \quad (mL/min \cdot kg) \tag{6}$$

The equations hold in between 20 and 50 years and are sex invariant.

VO_{2max} levels
VO_{2max} varies considerably in the population. The scores improve with training. In endurance sports, such as cycling, rowing, cross-country skiing and running, VO_{2max} values above 75 mL/kg/min are rather common. World class cyclists and cross-country skiers typically exceed 80 mL/kg/min and a rare few may exceed 90 for men. Women top athletes generally exceed 70 mL/kg/min. A competitive club athlete might achieve a VO_{2max} of ca. 70 mL/kg/min.

Literature

MArdle, WD., Katch FI, Katch VL, , Exercise Physiology: Energy, Nutrition & Human Performance, Lippincott Williams and Wilkins, Section 3, Chap 7, 11 and 17, 2001.

Uth N,Sorensen H, Overgaard K, Pedersen PK.. Estimation of VO2max from the ratio between HRmax and HRrest--the Heart Rate Ratio Method. 2004 Eur J Appl Physiol 91 111-115.

Light and Optics

CCD camera

Principle

A charge-coupled device (CCD) is an image sensor, consisting of an integrated circuit containing an array of linked, or coupled, light-sensitive capacitors. This device is also known as a Color-Capture Device.
The capacitors are the classical components of the CCD camera, but more and more photodiodes are the fundamental collecting units of the CCD.
Physics of operation An image is projected by a lens on the capacitor array (actually a 2D array or matrix), causing each capacitor to accumulate an electric charge proportional to the light intensity at that location. A one-dimensional array, used in line-scan cameras, captures a single slice of the image, while a two-dimensional array, used in video and still cameras, captures the whole image or a rectangular portion of it. Once the array has been exposed to the image, a control circuit causes each capacitor to transfer its content to its neighbor. The last capacitor in the array dumps its charge into an amplifier that converts the charge into a voltage. By repeating this process, the control circuit converts the entire content of the array to a varying voltage, which it samples, digitizes and stores in memory.

Application

CCDs are used in medical Fluoroscopy, optical and UV Spectroscopy and in all kind of basic cellular research. Frequent science applications are in astrophysics. Daily life applications are digital photography and "1D" CCD grids are applied in fax machines.
CCDs containing grids of pixels are used in digital cameras, optical scanners and video cameras as light-sensing devices. They commonly respond to 70% of the incident light (meaning a quantum efficiency of about 70%) making them more efficient than photographic film, which captures only about 2% of the incident light.

More info

CCDs are typically sensitive to infrared light, which allows infrared photography, night-vision devices, and zero lux (or near zero lux) video-recording/photography. Because of their sensitivity to infrared, CCDs used in astronomy are usually cooled to liquid nitrogen temperatures, because infrared black body radiation (see Body heat dissipation and related water loss and Wien's displacement law) is emitted from room-temperature sources. One other consequence of their sensitivity to IR is that infrared from remote controls will often appear on CCD-based digital cameras or camcorders if they don't have infrared blockers. Cooling also reduces the array's dark current, improving the sensitivity of the CCD to low light intensities, even for ultraviolet and visible wavelengths.
Thermal noise, dark current, and cosmic rays may alter the pixels in the CCD array. To counter such effects, an average of several exposures is made. The average of images taken with the shutter closed is necessary to lower the random noise. Once developed, the "dark frame" average image is then subtracted from the open-shutter image to remove the dark current and other systematic defects in the CCD (dead pixels, hot pixels, etc.).

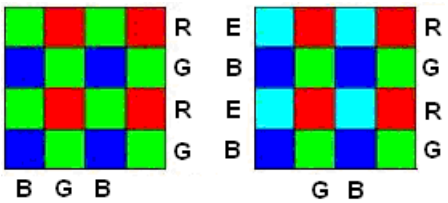


Fig. 1 Left the Bayer filter with twice the green cells and right the RGBE filter with as fourth color cyan. All four colors have the same frequency of occurrence.

Digital color cameras generally use a Bayer or RGBE filter. These filters are color filters in a matrix arrangement. After passing the filer matrix, the light is detected by a matrix of photosensors.
Better color separation can be reached by three-CCD devices (3CCD) and a dichroic beam splitter prism, (see Dichroism) and Light: beam splitter) that splits the image into red, green and blue components. Each of the three CCDs is arranged to respond to a particular band of wavelengths. Some semi-professional digital video camcorders (and all professionals) use this technique.

Chemoluminescence and Bioluminescence

Principle

Chemoluminescence

Chemoluminescence is the emission of light as the result of a chemical reaction. The reaction may occur in the liquid phase or in the gas phase. Most simply, given reactants **A** and **B**, with an excited intermediate ◊, the reaction is:

$$[\mathbf{A}] + [\mathbf{B}] \rightarrow [\Diamond] \rightarrow [\text{Products}] + \text{light}. \quad (1)$$

The decay of the excited state [◊] to a lower energy level is responsible for the emission of light. In theory, one photon of light should be given off for each molecule of reactant, so Avogadro's number of photons per mole. In actual practice, non-enzymatic reactions seldom exceed 1% quantum efficiency. For example, in the liquid phase, if [A] is luminol and [B] is hydrogen peroxide in the presence of a suitable catalyst the reaction is:

$$\text{luminol} + H_2O_2 \rightarrow \text{3-aminophthalate}[\Diamond] \rightarrow \text{3-aminophthalate} + \text{light}. \quad (2)$$

A standard example of chemoluminescence in the laboratory setting is found in the luminol test, where evidence of blood is taken when the sample glows upon contact with iron. A daily live example is a lightstick.

Fig. 1 Lightsticks

Applications

Analysis of organic species: useful with enzymes, where the substrate is not directly involved in chemoluminescence reaction, but the product is a reactant of the chemoluminescence reaction. Further environmental gas and liquid analysis for determining small amounts of impurities or poisons in air. Typical example is NO determination with detection limits down to 1 ppb (parts per billion).

More info

Enzymatic chemoluminescence (ECL) is a common technique for a variety of detection assays in biology. An horseradish peroxidase molecule (HRP) is tethered to the molecule of interest (usually by immunoglobulin staining). This then locally catalyzes the conversion of the ECL reagent into a sensitized reagent, which on further oxidation by hydrogen peroxide, produces an excited triplet (a set of three quantum states of a system, each with total spin $S = 1$), e.g. of carbonyl which emits light when it decays to the singlet (S = 0). The result is amplification of antibody detectability.
When chemoluminescence takes place in living organisms, the phenomenon is called bioluminescence.

Bioluminescence

Bioluminescence is the production and emission of light by a living organism as the result of a chemoluminescence reaction during which chemical energy is converted to light energy.
Bioluminescence is really a form of "cold light" emission; less than 20% of the light is generated by thermal radiation. It should not be confused with fluorescence, phosphorescence or refraction of light.
An example is bioluminescence by dinoflagellates at the surface of seawater when the surface is agitation (e.g. by a swimmer or a copepode). The λ_{max} is at ca. 472 nm and the emittance has the remarkable efficiency of more than 50%,

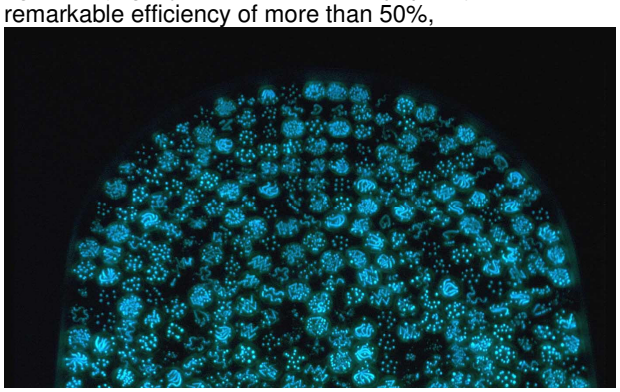


Fig. 2 Image of hundreds of agar plates cultured with a species of bioluminescent marine bacteria.

Bioluminescence may be generated by symbiosis of organisms carried within a larger organism. It is generated by an enzyme-catalyzed chemoluminescence reaction, wherein a luceferin (a kind of pigment) is oxidized by a luceferase (a kind of enzyme). ATP is involved in most instances. The chemical reaction can be either extra- or intracellular process. The expression of genes related to bioluminescence in bacteria is controlled by the lux operon.

Application

Luciferase systems are widely used in the field of genetic engineering as reporter genes (green fluorescent protein, see Fluorescence).
The structure of photophores, the light producing organs in bioluminescent organisms, are being investigated by industry (glowing trees, organisms needing watering), food quality control, detecting bacterial species and studies into potential applications as for tagging domestic animals.

More Info

All cells produce some form of bioluminescence within the electromagnetic spectrum, but most is neither visible nor noticeable to the naked eye. Every organism's bioluminescence is unique in wavelength, duration, timing and regularity of flashes.
90% Of deep-sea marine life is estimated to produce bioluminescence in one form or another. Many marine invertebrates have bioluminescence, like planktons, microbes, corals, clams, jelly fish, nudibranchs, crustaceans (lobsters, squids etc.), echinoderms (sea stars, sea urchins etc.). Most marine light-emission belongs in the blue and green part of spectrum, the wavelengths that are less absorbed than long wavelengths. However, certain jawless fish emit red and IR light.
Non-marine bioluminescence is less widely distributed, but with more color variety. Well-known forms of land-bioluminescence are fireflies and New Zealand glow worms. Other insects (and larvae), worms (segmented), arachnoids, fish and even species of fungi have bioluminescent abilities.
Most forms are brighter (or only exist) at night, following a circadian rhythm.
It is thought to play a direct role in camouflage, attraction, repulsion and communication. It promotes the symbiotic induction of bacteria into host species, and may play a role in colony aggregation.

Dichroism

Principle

Dichroism has two related but distinct meanings in optics. With the first one, a dichroic material causes visible light to be split up into distinct beams of different wavelengths (colors), not to be confused with dispersion as happens in a prism (see Light and Light: refraction). With the 2nd one, light rays having different polarizations (see Light: polarization), are absorbed by different amounts.
Which meaning of *dichroic* is intended can usually be inferred from the context. A mirror, a filter or beam splitter (see Light: beam splitter) is referred to as *dichroic* in the wavelength-separating first sense; a dichroic crystal or material refers to the polarization-absorbing second sense.

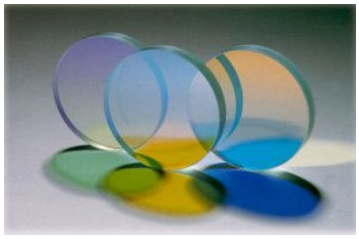


Fig 1 Dichroic filters

Application

General
The most common example is the dichroic filter. Dichroic filters operate using the principle of interference (see Light: diffraction and Huygens' principle). Alternating layers of an optical coating are built up upon a glass substrate, selectively reinforcing certain wavelengths of light and interfering with other wavelengths. By controlling the thickness and number of the layers, the wavelength of the bandpass can be tuned and made as wide or narrow as desired. Because unwanted wavelengths are reflected rather than absorbed, dichroic filters do not absorb much energy during operation and so become much less warm as absorbance filters (see for absorbance Lambert-Beer law).
Other examples of the wavelength type of dichroism are the dichroic mirror and the dichroic prism. The latter is used in some camcorders (miniaturized video cameras), which uses several coatings to split light into red, green and blue components. This is also applied in the CCD camera.
Medicine and food industry
The dichroism of optically active molecules (see **More Info**) is used in the food industry to measure syrup concentration, and in medicine as a method to measure blood sugar (glucose) in diabetic people.

More Info

The original meaning of *dichroic* refers to any optical device, which can split a beam of light into two beams with different wavelengths.
Basically, a dichroic filter has more than one transmission peak, with transmission frequencies harmonically related. However, in practice one needs nearly always a filter with one transmission peak. These are the filters, which are called interference filters. The higher harmonics, with have much lower transmission, are mostly attenuated by an absorbance filter. Side bands (similar of those of Fig. 1 in Light: diffraction) have very small transmission. They can strongly be diminished by stacking two identical filters, but on the cost of transmission of the principal peak.

The second meaning of dichroic refers to a material in which light in different polarization states, traveling through it, experience a varying absorption. The term comes from observations of the effect in crystals such as tourmaline. In these crystals, the strength of the dichroic effect varies strongly with the wavelength of the light, making them appear to have different colors when viewed with light having differing polarizations. This is more generally referred to as pleochroism, and the technique can be used to identify minerals.
Dichroism also occurs in optically active molecules, which rotate linearly polarized light (see Light: polarization). Depending on the 3-D molecular structure the rotation is left (levorotatory) or right (dextrorotatory). This is known as circular dichroism. Glucose is dextrorotatory and fructose strongly levorotatory. However, basically an optic active substance has a dextrorotatory and levorotatory version.
Dichroism occurs in liquid crystals (substances with properties between those of a conventional liquid, and those of a solid crystal) due to either the optical anisotropy of the molecular structure (resulting in more than one refractive index, see Light: Snell's law) or the presence of impurities or the presence of dichroic dyes.

Endoscopy

Principle

Endoscopy refers to looking inside the human body using an *endoscope*. Endoscopy is a minimally invasive diagnostic technique used to assess the interior surfaces of an organ by inserting a tube into the body. The instrument may have a rigid (borescope) or flexible tube (fiberscope) and not only provide an image for visual inspection, photography and video-scopy, but also enable taking biopsies and retrieval of foreign objects. Endoscopy is the vehicle for minimally invasive surgery.

An endoscope comprises an eyepiece, the ocular, producing a parallel exit bundle or virtual image and light source to illuminate the object on one end. At the other end is an objective lens producing a real image (see Light: the ideal and non-ideal lens). Both are linked by a tube mounting an optical fiber system (see Fiber optics). So, the fundamental principle of operation is transmitting optical information through a bundle of optical fibers such that an image can be observed. However, a classical *boroscope* may comprise instead of the fiber system a whole series of lenses as transmission system. Basically, an endoscope is a kind of microscope. The light source may provide wide band visible light, whether or not spectrally scanned, and for specific applications narrow band, (near) IR light. For Fluorescence (natural or artificial), e.g. applied in examination the esophagus, UV light can be used. Often, the light is provided by a Laser. An additional channel allows entry of air, fluid, as well as remote control of medical instruments such as biopsy forceps or cytology brushes.

Fig. 1 A flexible endoscope.

Application

Endoscopy is applied in nearly every medical discipline, also for outdoor patients. Endoscopes can be divided into two distinct categories according to their medical application. These are the regular, macroscopic endoscopes such as the gastroscope, colonscope, and bronchoscope to inspect epithelial surfaces. The second category comprises the miniaturized types. They includes ultrathin endoscopes for use as ophthalmic endoscopes, angioscopes and needlescopes. Ultra-thin endoscopes are used for robotic surgery. The latter, with a diameter less than one mm, have been developed to examine very small parts of internal organs. The images of *ultrathin needlescopes* contain 2,000 to 6,000 pixels with a smallest resolution of about 0.2 mm. They can be inserted into for instance mammary glands to detect breast cancer at early stages.
A borescope is used in arthroscopy (and also in engineering).

Non-medical uses are in architectural design (pre-visualization of scale models) and internal inspection of complicated technical systems and examination of improvised explosive devices by bomb robots.

More Info

The type of fibers used is dependent on the type of the illuminating light and the image specifications. Often, the fibers to deliver the light, the light guide, (mostly with coherent light, so a light beam with a single frequency and phase) and those to transmit the image information, the image guide, are of different types of fibers (see Fiber optics).
Recent developments are fiber-optic fluorescence imaging systems. Until recently, fiber-based fluorescence imaging was mainly limited to epifluorescence and scanning confocal modalities (confocal micro-endoscopy) (see Light microscopy: confocal). New classes of photonic crystal fibers (see Fiber optics) facilitate ultra-short pulse delivery for fiber-optic two-photon fluorescence imaging. This can be combined with two-photon fluorescence and second harmonic generation microscopy, miniaturized in a nonlinear optical endoscope based on a double-clad photonic crystal fiber to improve the detection

efficiency and a MEMS (MicroElectroMechanical System) mirror to steer the light at the fiber tip (see Fiber optics, Light microscopy: two-photon fluorescence, and the chapters about light microscopy). Another new application is combining laser holographic interferometry with an endoscope (see Holography, Huygens'principle and Interferometry). Another combination is laser Doppler imaging (see Doppler principle) of blood flow with endoscopy.
With the application of robotic systems, telesurgery was introduced as the surgeon could operate from a site physically removed from the patient.
Wireless capsule endoscopy is another emerging technology. This technique uses an ingestible capsule comprising a miniature camera with a MEMS mirror for scanning and a transmitter. In this way some part of the gastrointestinal tract can be visualized. Nowadays, this application in the esophagus is more or less standard, but other parts of the tract are still experimentally examined due to peristaltic movements. MEMS technology may provide a solution for this.

Fiber optics

Principle

An optical fiber is a cylindrical light-isolated waveguide that transmits light along its axis by the process of total internal reflection (see Light: Fresnel equations). The fiber consists of a *core* surrounded by a *cladding* layer. To confine the optical signal in the core, the refractive index of the core must be greater than that of the cladding.
Optical fibers may be connected to each other by connectors or by *splicing* that is joining two fibers together to form a continuous optical waveguide. The complexity of this process is more difficult than splicing copper wire.

Multimode fiber (MMF)

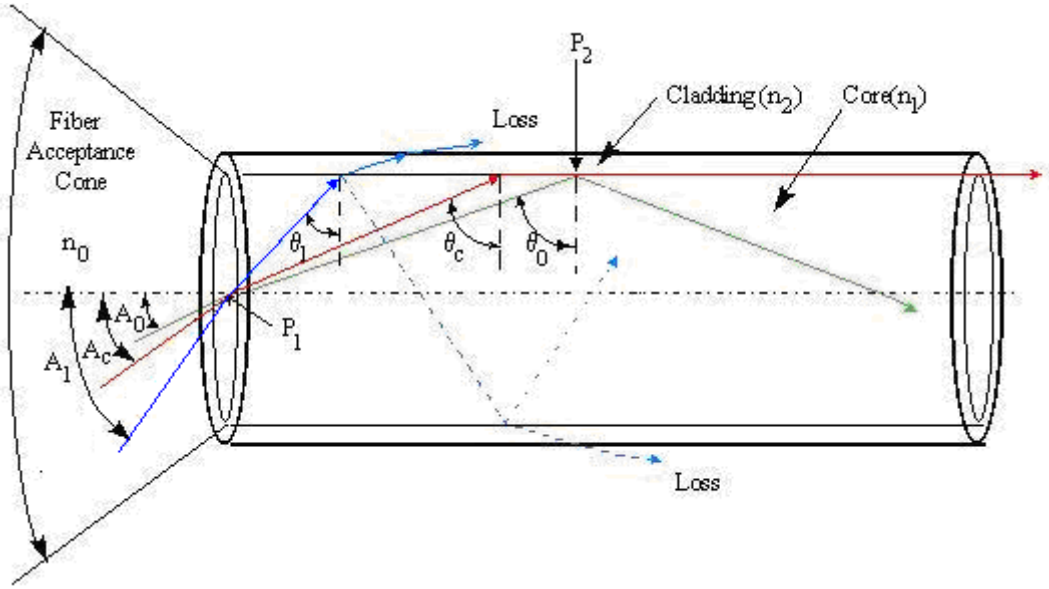


Fig. 1 The propagation of light through a SI-MMF.

A fiber with large core diameter (some tens of µm up to hundreds of µm) behaves in accordance with geometric optics. It is called a *multimode fiber* (MMF) since there are various modes of vibration given by the wave equations (see textbooks of physics).
When the boundary between the core and cladding is abrupt, the MMF is called a *step-index* (SI) fiber. When it is gradual it is a *graded-index* (GRIN) MMF.
In a SI MMF, rays of light are guided along the fiber core by total internal reflection. Rays (for instance the green one in Fig. 1) are completely reflected when they meet the core-cladding boundary at a higher angle (measured relative to a line normal to the boundary) than the critical angle Θ_c, the minimum angle for total internal reflection. The red ray in Fig. 1 impinges with the angle Θ_c. Θ_c is $n_{gladding}/n_{core}$.
Rays that meet the boundary at a lower angle (the blue one in Fig. 2) are lost after many repeated reflections/refractions from the core into the cladding, and so do not convey light and hence information along the fiber.
Θ_c determines the acceptance angle of the fiber, also expressed in the numerical aperture NA ($\equiv n_0 \cdot \Theta_c$). A high NA allows light to propagate down the fiber in rays both close to the axis and at various angles, allowing efficient sending of light into the fiber. However, this high NA increases the amount of dispersion as rays at different angles have different path lengths and therefore take different times to traverse the fiber. This argues for a low NA.

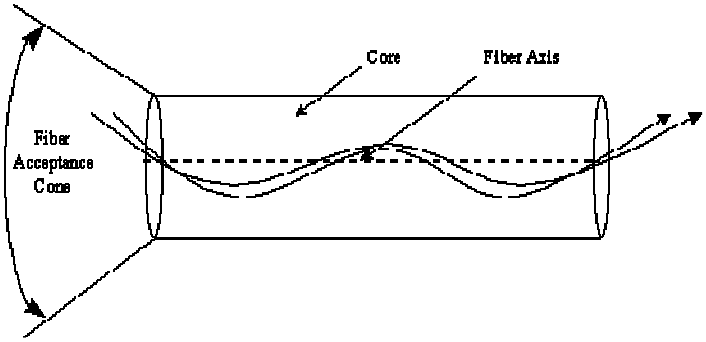


Fig. 2. Paths of light rays in a GRIN fiber. The refractive index changes gradually from the center to the outer edge of the core.

In a GRIN fiber, n_{core} decreases continuously from the axis to the cladding. This causes light rays to bend smoothly as they approach the cladding, rather than reflecting abruptly from the core-cladding boundary. The resulting curved paths reduce multi-path dispersion because the undulations (see Fig.2) diminish the differences in path lengths. The difference in axial propagation speeds are minimized with an index profile which is very close to a parabolic relationship between the index and the distance from the axis.
SI and GRIN fibers suffer from Rayleigh scattering (see Light: scattering), which means that only wavelengths between 650 and 750 nm can be carried over significant distances.

Singlemode fiber (*SMF*)
This single glass fiber (core diameter generally 8 - 10 µm) has the axial pathway as solely mode of transmission, typically at near IR (1300 or 1550 nm). It carries higher bandwidth than multimode fiber, but requires a light source with a narrow spectral width. Single-mode fiber have a higher transmission pulse rate and cover up to 50 times more distance than multimode, but it also costs more. The small core and single light-wave virtually eliminate any distortion that could result from overlapping light pulses (little pulse dispersion), providing the least signal attenuation and the highest transmission speeds of any fiber type.

Image transmission
Since an individual fiber can transmit only a spot of a certain wavelength and intensity. A large number of single fibers must be aligned and fused together to transmit an image. This means assembly of optical fibers in which the fibers are ordered in exactly the same way at both ends of the bundle to create an image. This type of fiber bundle is called a coherent bundle or image guide bundle. On the other hand, the assembly of optical fibers that are bundled but not ordered is called an incoherent bundle. An optical fiber which is incapable of producing an image is used in medical endoscopes, boroscopes, and fiberscopes as a light guide. The light guide, as well as the image guide, is essential to construct an image in any optical instrument. Light guides are much less expensive and easy to produce compared to image guides and are designed to maximize light carrying ability. In an image guide, the amount of image detail (resolving power) depends on the diameter of each fiber core. Generally, the individual fibers of a light guide are much thicker than fibers (MMFs) in image guides because resolution is not a factor.

Application

Medical
Optical fibers are used in transducers and bio-sensors for the measurement and monitoring of for instance body temperature, blood pressure, blood flow and oxygen saturation levels. In medical applications, the fiber length is so short (less than a few meters) that light loss and fiber dispersion are not of concern. Glass optical fibers are used in most endoscopes and are made with SI fibers.
Optical fibers are also used as transmission lines in equipment that is very sensitive to disturbance by electric fields, such as EEG amplifiers. At the other hand, they are applied to prevent the generation of a magnetic field due to current flowing in an electric cable. Even very small current produce magnetic fields strong enough to disturb Magnetoencephalographic recordings (see Magnetoencephalography). All these applications are based on SI MMFs.
The principle of fiber optics is also found in nature. Long rods, as found in for instance frogs, and the visual sensory cells of arthropods, the ommatidia, act as wave guides.
General
Optical fiber cables are frequently used in ICT applications (such as cable television and all kind of data transport). For far distance transmission SMFs are used. MMFs can only be used for relative short distances, e.g. for ICT applications in a building.

More info

Consider Fig. 1 again. The ray incident on the face of the SI fiber at angle A_1 will be refracted weakly inside the core and refracted substantially into the cladding. The angle A_c is referred to as the maximum acceptance angle since θ_c is the critical angle for total internal reflection and all smaller angles are totally reflected. The angles A_c and θ_c are determined by the refractive indices of core and cladding. Therefore, a ray incident on the core-cladding boundary at an angle less than θ_c will not undergo total internal reflection and finally will be lost.

In Fig. 2 at the point P_1 it holds that:

$$n_o \sin A_c = n_1 \sin (90-\theta_c) \quad (1)$$

Also at the point P_2:

$$n_1 \sin \theta_c = n_2 \sin (90) = n_2 \text{ or } \Theta_c = \arcsin(n_2/n_1) \quad (2)$$

Together they give:

$$n_0 \sin A_c = n_1 \cos \theta_c = (n^2_1 - n^2_2)^{1/2} = NA, \text{ or}$$

NA is the numerical aperture of the SI fiber and is defined as the light-gathering power of an optical fiber. When the face of the fiber is in contact with air ($n_0 = 1$ for air), $NA = \sin\theta_c$. When $n_2/n_1 = 0.99$, then θ_c is 8.1^o and A_c is 12.2^o.

When total internal reflection occurs, there is also light transmission in the gladding, the evanescent wave (a very nearby standing wave). This can cause light leakage between two adjacent fibers even when the diameter of a fiber is many times greater than the wavelength. In SMFs, the energy transmitted via the evanescent wave is a significant fraction.
In SI fibers, the light rays zigzag in straight lines between the core/cladding on each side of the fiber axis. In GRIN fibers, the light travels in a curved trajectory, always being refracted back to the axis of the fiber (Fig. 2). At angles > Θ_c, light never reaches the outer edge of the fiber. At angles< Θ_c, the light enters the adjacent fiber, traverses the guide and is absorbed on the periphery of the guide as in the case of the SI guide.

Glass optical fibers are mostly made from silica (SiO_2) with a refractive index of about 1.5. Typically the difference between core and cladding is less than one percent. For medical applications, due to the required properties (optical quality, mechanical strength, and flexibility) also plastic optical fibers are used. Plastic fibers have the advantages of much simpler and less demanding after-processing and plastic fibers are lighter and of lower cost than glass fibers.
Plastic is common in step-index multimode fiber with a core diameter of 1 mm and have more propagation losses than glass fiber (1 dB/m or higher).

A fiber with a core diameter < $10 \cdot \lambda$ cannot be modeled using geometric optics, but must be analyzed as an electromagnetic structure, by solution of the electromagnetic wave equation, which describes the propagation of electromagnetic waves (see textbooks of physics).
The number of vibration modes in a SI MMF can be found from the V number:

$$V = (2\pi r/\lambda) (n^2_1 - n^2_2)^{1/2}, \quad (3)$$

where r is the core radius and λ wavelength. When $n_0 = 1$, than V becomes $(2\pi r/\lambda)NA$. When $V < 2.405$ only the fundamental mode remains and so the fiber behaves as a SMF.
The electromagnetic analysis may also be required to understand behaviors such as speckles that occur when coherent light (same frequency and phase) propagates in a MMF. (A speckle pattern is a random intensity pattern produced by the mutual interference of coherent wave fronts that are subject to phase differences and/or intensity fluctuations. See also Huygen's Principle.) Speckles occur with Interferometry and its applications.

A new type of crystals, photonic crystals led to the development of the photonic crystal fiber (PCF). (Photonic crystals are periodic optical (nano)structures that are designed to affect the motion of photons in a similar way as periodicity of a semiconductor crystal affects the motion of electrons.) These fibers consist of a hexagonal bundle of hollow microtubes embedded in silica with in the center the fiber of photonic crystal. A PCF guides light by means of diffraction from a periodic structure, rather than total internal reflection. They can carry higher power than conventional fibers.

Fluorescence

Principle

Fluorescence, as other types of luminescence, is mostly found as an optical phenomenon in cold bodies (in contrast to incandescence, a process with a flame), in which a molecule absorbs a high-energy photon, and re-emits it as a lower-energy photon with a longer wavelength (λ). The energy difference between the absorbed and emitted photons ends up as molecular vibrations, finally in the form of heat. Usually the absorbed photon is in the UV, and the emitted light is in the visible range, but this depends on the absorbance curve and the shift to the higher emitted λ (Stokes shift: Δλ) of the particular fluorophore (the molecule with the fluorescent structure).
The process can be described by:

$$S_1 \rightarrow S_2 + h/\lambda_{emitted}. \quad (1)$$

The system starts in state S_1, and after the fluorescent emission of a photon with energy hv, it is in state S_2 where h is Planck's quantum mechanical constant, being $6.626 \cdot 10^{-34}$ Js.

Applications

There are many natural and synthetic compounds that exhibit fluorescence, and they have a number of medical, biochemical and industrial applications (fluorescent lighting tubes).
The fluorophore attached by a chemical reaction to bio-molecules enables very sensitive detection of these molecules.

Examples are:

Automated sequencing of DNA by the chain termination method
Each of four different chain terminating bases has its own specific fluorescent tag. As the labeled DNA molecules are separated, the fluorescent label is excited by a UV source, and the identity of the base terminating the molecule is given by the wavelength of the emitted light.

DNA detection
The compound ethidium bromide, when free to change its conformation in solution, has very little fluorescence. Ethidium bromide's fluorescence is greatly enhanced when it binds to DNA, so this compound is very useful in visualizing the location of DNA fragments in agarose gel electrophoresis (see Electrophoresis).

The DNA microarray
Immunology and immonohistochemistry An antibody has a fluorescent chemical group attached, and the sites (e.g., on a microscopic specimen) where the antibody has bound can be seen, and even quantified, by fluorescence.

FACS, fluorescent-activated cell sorting
Fluorescence resonance energy transfer and similar techniques has been used to study the structure and conformations of DNA and proteins. This is especially important in complexes of multiple biomolecules.

Calcium imaging Aequorin, from the jellyfish Aequorea victoria, produces a blue glow in the presence of Ca^{2+} ions (by a chemical reaction). Other fluorescent dyes are calcium orange and the intracellular indicator Indo-1. It has been used to image calcium in cells in real time, especially in neurobiological applications. This technique has a long history in research of hippocampus slices. Imaging at the light microscopic and confocal level (see Confocal microscopy) is also used to explore the contribution of inward calcium currents and calcium release in relation to synaptic transmission in neurons. Specific applications are analyses of neuronal networks and synaptic plasticity, often studied with the Patch-clamp technique and Voltage clamp technique. This techniques may use the voltage sensitive Ca^{2+} dyes Fluo, Ca-green en Fura.

More Info

The success with aequorin has led to the discovery of Green Fluorescent Protein (GFP), an important research tool. GFP and related proteins are used as reporters for any number of biological events including sub-cellular localization. Levels of gene expression are sometimes measured by linking a gene

for GFP production to another gene. Fluorescent calcium indicator proteins (FCIPs) are Ca^{2+}-sensitive GFP variants. Fig. 1 shows an example of light-evoked Ca^{2+} responses in retinal ganglion cells**.**

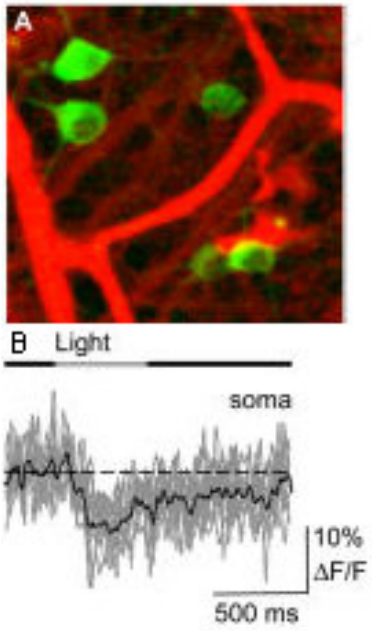


Fig. 1. Intact, light-sensitive retinal whole mount. A Blood vessels are red and active retinal ganglion cells are green (FCIP-positive). B Light-stimulus-evoked Ca^{2+} response (black trace; gray traces are single trials) measured in the soma with ΔF/F the relative fluorescence changes. After PLoS Biol. 2004; 2(6): e163.

Also, many biological molecules have an intrinsic fluorescence that can sometimes be used without the need to attach a chemical tag. Sometimes this intrinsic fluorescence changes when the molecule is in a specific environment, so the distribution or binding of the molecule can be measured. Biliburin, for instance, is highly fluorescent when bound to a specific site on serum albumin. Zinc protoporphyrin, formed in developing red blood cells instead of hemoglobin when iron is unavailable or lead is present, has a bright fluorescence and can be used to detect these abnormality.

Fluorescence resonance energy transfer (FRET)

Principle

Fluorescence resonance energy transfer (FRET) is a distance-dependent interaction between the electronic excited states of two dye molecules in which excitation is transferred from a donor molecule to an acceptor molecule *without emission of a photon*. This energy transfer mechanism is termed "Förster resonance energy transfer" (FRET; called after the German scientist Förster) or dipole-dipole resonance energy transfer. When both molecules are fluorescent, the term "fluorescence resonance energy transfer" is often used, although the energy is not actually transferred by fluorescence, as illustrated in Fig. 1.

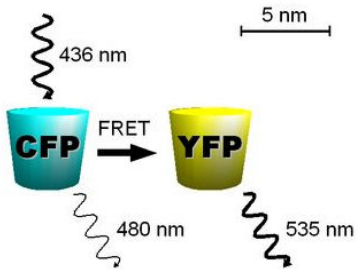


Fig. 1 Principle of FRET. The emitted 480 nm signal is much weaker than the 535 signal emitted by the acceptor YFP. When the distance between the two chromophores is too large, YFP is not excited and CFP emits a strong 480 nm signal.

Primary Conditions for FRET

- Donor and acceptor molecules must be in close proximity (typically 1–10 nm).
- The absorption spectrum of the acceptor must overlap the fluorescence emission spectrum of the donor (see Fig. 1).
- Donor and acceptor transition dipole orientations must be approximately parallel.

The efficiency of FRET is dependent on the inverse sixth power of the intermolecular separation, making it useful over distances comparable with the dimensions of biological macromolecules. Thus, FRET is an important technique for investigating a variety of biological phenomena that produce changes in molecular proximity (1-10 nm). When FRET is used as a contrast mechanism, co-localization of proteins and other molecules can be imaged with spatial resolution beyond the limits of conventional optical microscopy.

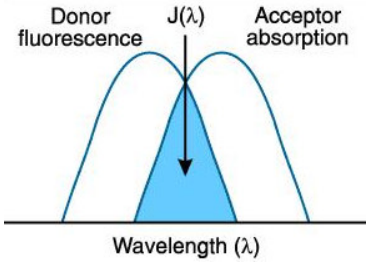


Fig. 2 Overlap of the energy spectra.

Donor/acceptor pairs

In most applications, the donor and acceptor dyes are different, in which case FRET can be detected by the appearance of sensitized fluorescence of the acceptor or by quenching of donor fluorescence. When the donor and acceptor are the same, FRET can be detected by the resulting fluorescence depolarization.

Non-fluorescent acceptors such as dabcyl have the particular advantage of eliminating the potential problem of background fluorescence resulting from direct (i.e., nonsensitized) acceptor excitation.

Typical values of the distance between donor and acceptor are 3 to 6 nm, e.g. fluorescein-tetramethylrhodamine 5.5 nm and fluorescein-fluorescein 4.4 nm.

The most used FRET pair for biological use is a cyan fluorescent protein (CFP)-yellow fluorescent protein (YFP) pair. Both are color variants of green fluorescent protein (GFP). While labeling with organic fluorescent dyes is laborious, GFP variants can be easily attached to a host protein by genetic engineering.

Application

In fluorescence microscopy (see Optical microscopy, fluorescence), fluorescence confocal laser scanning microscopy (see Optical microscopy: confocal laser scanning), as well as in molecular biology, FRET is a useful tool to quantify molecular dynamics, such as protein-protein interactions, protein-DNA interactions, and protein and nucleic acid structure conformational changes. For monitoring the complicated formation between two molecules, one of them is labeled with a donor and the other with an acceptor, and these fluorophore-labeled molecules are mixed. When they are dissociated, the donor emission is detected upon the donor excitation. When the donor and acceptor are in proximity (1-10 nm) due to the interaction of the two molecules, the acceptor emission is predominantly observed because of the intermolecular FRET from the donor to the acceptor.
FRET can further be used for receptor/ligand interactions, immunoassays, probing interactions of single molecules, automated DNA sequencing, and nucleic acid hybridization.

More Info

Förster Radius
The distance at which energy transfer is 50% efficient (i.e., 50% of excited donors are deactivated by FRET) is defined by the Förster radius (R_o). The magnitude of R_o is dependent on the spectral properties of the donor and acceptor dyes. It is equal to:

$$R_0^6 = 8.8 \times 10^{-28} \kappa^2 n^4 Q_0 J, \quad (1)$$

where
κ^2 is the dipole orientation factor (range 0-4);
n is the refractive index quantum yield of the medium;
Q_0 is the fluorescence of the donor in the absence of the acceptor;
J is the spectral overlap integral calculated as:

$$J = \int f_D(\lambda)\varepsilon(\lambda)\lambda^4 d\lambda, \quad (2)$$

where
f_D is the normalized donor emission spectrum, and
ε_A is the acceptor molar extinction coefficient.

$\kappa^2 = 2/3$ is often assumed. This value is obtained when both dyes are freely rotating and can be considered to be isotropically oriented during the excited state lifetime. If either dye is fixed or not free to rotate, then $\kappa^2 = 2/3$ is not a valid, but the deviation is small.
The efficiency of FRET can be expressed as:
- the intensity difference of the two emitted signals (535 nm minus 480 nm of the example), divided by the emitted donor intensity;
- in terms of the effect of distance, $R_0^6/(r^6 + R_0^6)$;
- in terms of the effect of half time (life time), the difference between life time without acceptor and with acceptor divided by that without acceptor.

Related methods
A limitation of FRET is the requirement for external illumination to initiate the fluorescence transfer, which can lead to background noise in the results from direct excitation of the acceptor, or photo-bleaching. To overcome this difficulty, Bioluminescence Resonance Energy Transfer (or BRET) has been developed. This technique uses a bioluminescent luciferase rather than CFP to produce an initial photon emission compatible with YFP.
A different, but related, mechanism is Dexter Electron Transfer (where the energy transfer is triplet-triplet).
An alternative method to detecting protein-protein proximity is Bimolecular fluorescence complementation (BiFC), where two halves of a YFP are fused to a protein. When these two halves meet they form a fluorophore after about 60 s - 1 hr.

Fluoroscopy

Principle

Fluoroscopy is an imaging technique commonly used to obtain real-time images of the internal structures. In its simplest form, a fluoroscope consists of an X-ray source and fluorescent screen between which the object, the patient, is placed. Modern fluoroscopes couple the screen to a CCD video camera allowing still images or images to be played on a monitor or an X-ray image intensifier for digital imaging. A CCD camera (charge-coupled device) is an image sensor, consisting of an integrated circuit containing a matrix of coupled, light-sensitive capacitors).

Fig. 1 A modern fluoroscope

The x rays are attenuated in dependence on the type of structure of the body. They cast a shadow of the structures on the fluorescent screen. Images on the screen are produced as the unattenuated x rays are absorbed by atoms, which process gives rise to the emission of free electrons with a high kinetic energy (the photoelectric effect). While much of the energy given to the electrons is dissipated as heat, a fraction of it is given off as visible light by exiting atoms in the fluorescent molecules. Then, by ‘’de-excitation’’ light is emitted (see Fluorescence), the fluorescent process and this produce the image.

Application

Common fields of application are the gastrointestinal tract (including administration of barium, and enteroclysis), orthopedic surgery (operation guidance), angiography of the leg, heart and cerebral vessels, urological surgery (e.g. retrograde pyelography, i.e. injection of contrast liquid into the ureter (resulting in retrograde flow) in order to visualize the ureter and kidney). implantation of cardiac rhythm devices (pacemakers, implantable cardioverter defibrillators and cardiac resynchronization devices).

Risks
The risk on radiation damage by ionizing should be balanced with the benefits of the procedure to the patient. Although the length of a typical procedure often results in a relatively high absorbed dose, digitization of the images captured and flat-panel detector systems has reduced the radiation dose. Radiation doses to the patient depends especially on length of the procedure, with typical skin dose rates quoted as 20-50 mGy/min (Gy is Gray, the applied dose. 1 Gy is 1 J/kg tissue). Because of the long length of some procedures, in addition to standard cancer-inducing stochastic radiation effects, deterministic radiation effects have also been observed ranging from mild erythema, equivalent of a sun burn, to more serious burns. While deterministic radiation effects are a possibility, they are not typical of standard fluoroscopic procedures. Most procedures, sufficiently long in length to produce radiation burns, are part of necessary life-saving operations.

More Info

X-ray Image Intensifiers

At present, the original X-ray image intensifiers are replaced by CCD cameras or modern image intensifiers, which no longer use a separate fluorescent screen. Instead, a cesium iodide phosphor is deposited directly on the photocathode of the intensifier tube. The output image is approximately 10^5 times brighter than the input image. This *brightness gain* is comprised of a *flux gain* (amplification of photon number) and *minification gain* (concentration of photons from a large input screen onto a small output screen). Each of them approximates a gain of a factor of 100. This gain is such that quantum noise, due to the limited number of X-ray photons, is now a significant factor limiting image quality.

Flat-panel detectors
Also flat-panel detectors replace the image intensifier in fluoroscope design. They have increased sensitivity to X-rays, and therefore reduce patient radiation dose. They have also a better temporal resolution, reducing motion blurring. Contrast ratio is also improved: flat-panel detectors are linear over a very wide latitude, whereas image intensifiers have a maximum contrast ratio of about 35:1. Spatial resolution is approximately equal.
Since flat panel detectors are considerably more expensive they are mainly used in specialties that require high-speed imaging, e.g., vascular imaging and cardiac catheterization.

Imaging concerns
In addition to spatial blurring factors, caused by such things as the Lubberts effect, (non-uniform response of an imaging system at different depths), K-fluorescence reabsorption (reabsorption in the K-orbit of the atom) and electron range, fluoroscopic systems also experience temporal blurring due to system lag. This temporal blurring has the effect of averaging frames together. While this helps reduce noise in images with stationary objects, it creates motion blurring for moving objects. Temporal blurring also complicates measurements of system performance for fluoroscopic systems.

Holography

Principle

Holography is an advanced form of photography that allows an image to be recorded in 3-D. This technique can also be used to optically store, retrieve, and process information.

Fig. 1 Identigram as a security element in an identity card

Several types of holograms can be made. The first holograms were "transmission holograms", which were viewed by shining laser light through them and looking at the reconstructed image at the other side. A later refinement, the "rainbow transmission" hologram allowed viewing by white light and is commonly seen today on *credit cards* as a security feature and on product packaging. These versions of the rainbow transmission holograms are now commonly formed as surface relief patterns in a plastic film, and they incorporate a reflective aluminum coating which provides the light from "behind" to reconstruct their imagery. Another kind of common hologram is the true "white-light reflection hologram" which is made in such a way that the image is reconstructed naturally using light on the same side of the hologram as the viewer.

Technical description
The difference between holography and photography is best understood by considering what a black and white photograph actually is: it is a point-to-point recording of the intensity of light rays that make up an image. Each point on the photograph records just one thing, the intensity (i.e. the square of the amplitude of the electric field) of the light wave that illuminates that particular point. In the case of a color photograph, slightly more information is recorded (in effect the image is recorded three times viewed through three different color filters), which allows a limited reconstruction of the wavelength of the light, and thus its color. Recent low-cost solid-state lasers are performed to make holograms

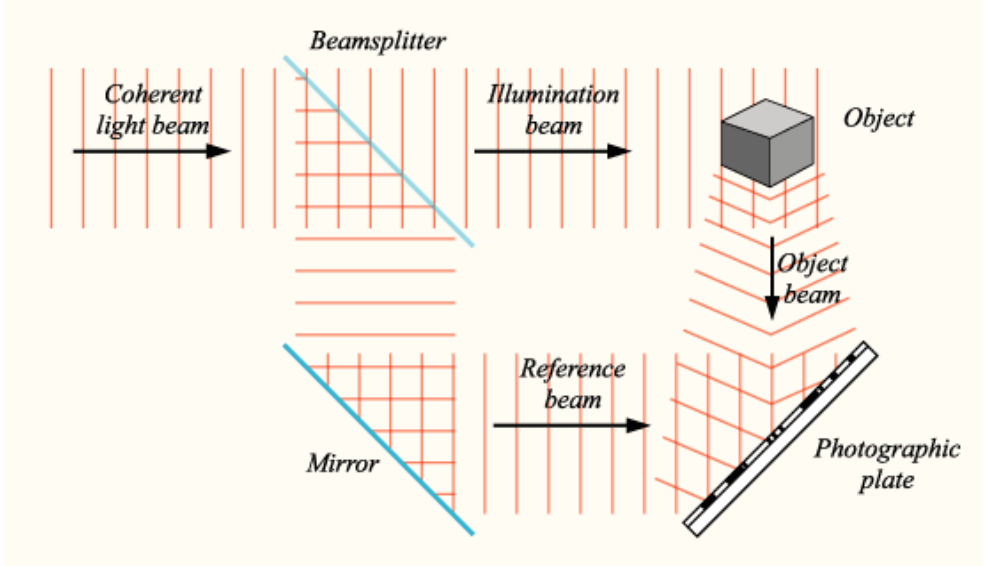


Fig. 1 Principle of making a hologram. (See Light: beam splitter for its working principle.)

Light, being a wave phenomenon, is characterized also by its phase. In a photograph, the phase of the light from the original scene is lost, and with it the three-dimensional effect. In a hologram, information from both the intensity and the phase is recorded. When illuminating the hologram with the appropriate light, it diffracts part of it into exactly the same wave (up to a constant phase shift invisible to our eyes) which emanated from the original scene, thus retaining the three-dimensional appearance. Also color holograms are possible.

Holographic recording process
To produce a recording of the phase of the light wave at each point in an image, holography uses a *reference beam* (see Fig. 1) which is combined with the light from the object (the *object beam*). If these two beams are coherent, optical interference (see Huygens'principle and Light: diffraction) between the reference beam and the object beam, due to the superposition of the light waves, produces a series of intensity fringes that can be recorded on standard photographic film. These fringes form a type of diffraction grating on the film, which is called the hologram. The central miracle of holography is that when the recorded grating is later illuminated by a substitute reference beam, the original object beam is reconstructed, producing a 3D image.

These recorded fringes do not directly represent their respective corresponding points in the space of a scene (the way each point on a photograph represents a single point in the scene being photographed). Rather, a small portion of a hologram's surface contains enough information to reconstruct the entire original scene, but only what can be seen from that small portion as viewed from that point's perspective. This is possible because during holographic recording, each point on the hologram's surface is affected by light waves reflected from all points in the scene, rather than from just one point. It is as if, during recording, each point on the hologram's surface were an eye that could record everything it sees in any direction. After the hologram has been recorded, looking at a point in that hologram is like looking "through" one of those eyes.
To demonstrate this concept, you could cut out and look at a small section of a recorded hologram; from the same distance you see less than before, but you can still see the entire scene by shifting your viewpoint laterally or by going very near to the hologram, the same way you could look outside in any direction from a small window. What you lose is the ability to see the objects from many directions, as you are forced to stay behind the small window.

Holographic reconstruction process
When the processed holographic film is illuminated once again with the reference beam, diffraction from the fringe pattern on the film reconstructs the original object beam in both intensity and phase (except for rainbow holograms). Because both the phase and intensity are reproduced, the image appears three-dimensional; the viewer can move his or her viewpoint and see the image rotate exactly as the original object would.

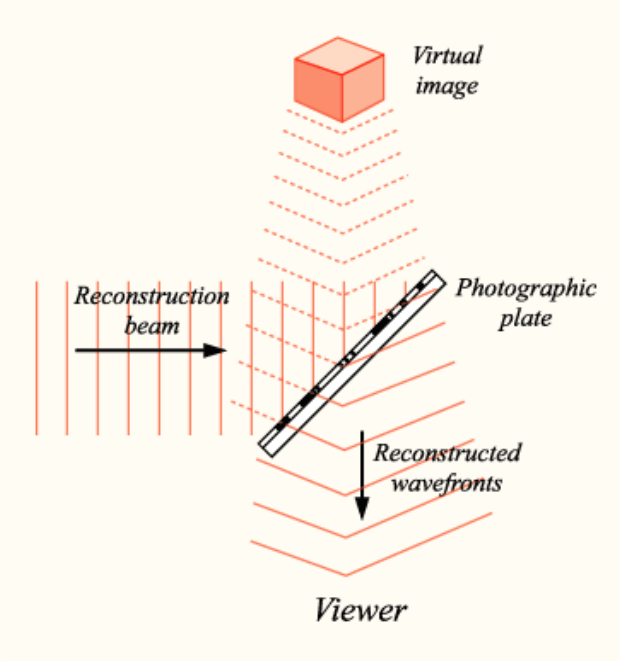


Fig. 2 Principle of reconstruction of the image.

It is possible to store the diffraction gratings that make up a hologram as phase gratings or amplitude gratings of various specific materials.

Application

High tech applications are numerous in (astro)physics, but large scale application in mass storing medical images etc. is coming soon. On bank notes, credit cards etc. it are normal safely features. Because of the need for coherent interference between the reference and object beams, Laser light is used to record holograms. But formerly other coherent light sources such as Hg-arc lamps (see Light:

sources) have also been applied. In simple holograms, the coherence length of the beam determines the maximum depth the image can have. The coherence length L is:

$$L = \lambda^2/(n\,\Delta\lambda), \quad (1)$$

where λ is the central wavelength of the source, *n* is the refractive index of the medium, and Δλ is the spectral width of the source. In sunlight and incandescent light holograms on credit cards have depths of a few mm. A good holography laser will typically have a coherence length of several meters.

Holography can be applied to a variety of uses other than recording images.
Holographic data storage stores information at high density inside crystals or photopolymers and has the potential to become the next generation of popular storage media with possibly 1 gigabit/s writing speed and 1 terabit/s readout speed since terabyte disks are nearly commercially available.
An alternate method to record holograms is to use a digital device like a CCD camera instead of a conventional photographic film. This approach is often called *digital* holography.

More Info

Dynamic holography
The discussion above describes static holography, with sequentially recording, developing and reconstructing. A permanent hologram is produced.
There exist also holographic materials which don't need the developing process and can record a hologram in a very short time (optical parallel processing of the whole image). Examples of applications of such real-time holograms include phase-conjugate mirrors ("time-reversal" of light), optical cache memories, image processing (pattern recognition of time-varying images) and optical computing.
The fast processing compensates the fact that the recording time. The optical processing performed by a dynamic hologram is much less flexible than electronic processing. On one side one has to perform the operation always on the whole image, and on the other side the operation a hologram can perform is basically either a multiplication or a phase conjugation. But remember that in optics, addition and Fourier transform (see Fourier analysis) are already easily performed in linear materials, the second simply by a lens. This enables some applications like a device that compares images in an optical way.

Holonomic brain theory
The fact that information about an image point is distributed throughout the hologram, such that each piece of the hologram contains some information about the entire image, seemed suggestive about how the brain could encode memories. For instance spatial frequency encoding by cells of the visual cortex was best described as a Fourier transform of the input pattern. Also the cochlea makes a Fourier transform. This holographic idea leaded to the term "holonomic".

Huygens' principle

Principle

The Huygens principle is a method of analysis applied to problems of wave propagation. This holds for macroscopic phenomena (optical devices such as lenses, prisms etc. very much larger than the wavelength). It recognizes that each point of an advancing wave front is in fact the center of a source of a new train of waves and that the advancing wave as a whole may be regarded as the sum of all the secondary waves arising from points in the medium already traversed.

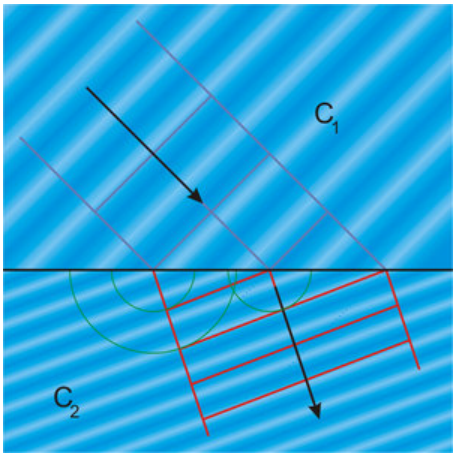


Fig. 1 Huygens' principle applied to refraction when a light beam goes from medium with a high speed of propagation C1 and consequently low refractive index to medium C2 with a low speed and high refractive index.

The Huygens' principle also holds for (near-)microscopic phenomena (optical devices such as apertures and slits of the order of a wavelength). It simply states that a large hole can be approximated by a collection of many small holes so each is practically a point source (whose contribution is easy to calculate). A point source generates waves that travel spherically in all directions). Similarly a relatively wide slit is composed of many narrow ones (subslits), and adding the waves produced by each produces the diffraction pattern (see Light: diffraction). For example, if two rooms are connected by an open doorway and a sound is produced in a remote corner of one of them, a person in the other room will hear the sound as if it originated at the doorway. As far as the second room is concerned, the vibrating air in the doorway is the source of the sound. The same is true of light passing the edge of an obstacle, but this is not as easily observed because of the short wavelength of visible light.

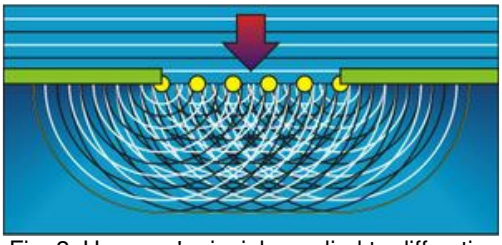


Fig. 2 Huygens' principle applied to diffraction.

The above views of wave propagation helps better understand a variety of wave phenomena, such as refraction and diffraction. The former is visualized in Fig. 1 and the latter in Fig. 2.

Interferometry

Principle

Interferometry is the technique of superimposing (interfering) two or more waves, to detect differences between them. Interferometry works because two waves with the same frequency that have the same phase will add to each other while two waves that have opposite phase will subtract. Typically, in an interferometer, a wave is split into two (or more) coherent parts, which travel different paths, and the parts are then combined to create interference. When the paths differ by an even number of half-wavelengths, the superposed waves are in phase and interfere constructively, increasing the amplitude of the output wave. When they differ by an odd number of half-wavelengths, the combined waves are 180° out of phase and interfere destructively, decreasing the amplitude of the output. Thus anything that changes the phase of one of the beams by only 180°, shifts the interference from a maximum to a minimum. This makes instruments with interferometers sensitive for changes of the phase of a wave, such as path length or refractive index.
Early interferometers principally used white light sources. Modern researchers often use monochromatic light sources like lasers, and even the wave character of matter can be exploited to build interferometers (e.g. with electrons, neutrons or even molecules).

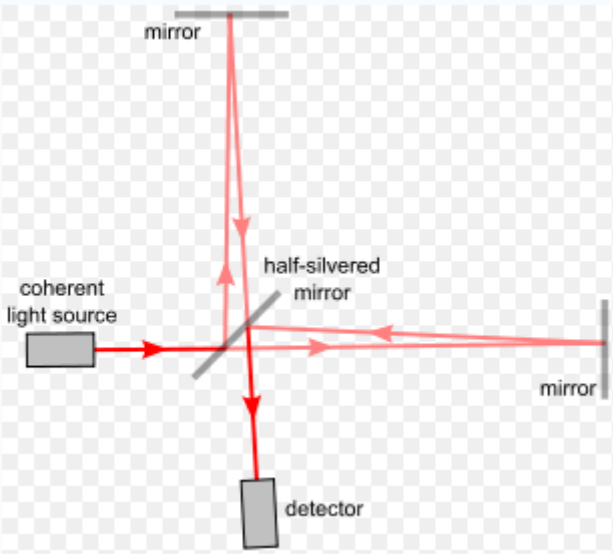


Fig. 1 A Michelson interferometer. In the control condition, both mirrors are equidistant to the semi-transparent mirror.

The classical interferometer is the Michelson(-Morley) interferometer. In a Michelson interferometer, the basic building blocks are a monochromatic source (emitting one wavelength), a detector, two mirrors and one semitransparent mirror (often called beam splitter; see Light: beamsplitter). These are put together as shown in the Fig. 1.
There are two paths from the (light) source to the detector. One reflects off the semi-transparent mirror, goes to the top mirror and then reflects back, goes through the semi-transparent mirror and to the detector. The other one goes through the semi-transparent mirror, to the mirror on the right, reflects back to the semi-transparent mirror, then reflects from the semi-transparent mirror into the detector.
If these two paths differ by a whole number (including 0) of wavelengths, there is constructive interference and a strong signal at the detector. If they differ by a whole number plus a half wavelength (so 0.5, 1.5, 2.5 ...) there is destructive interference and a weak signal.
Interferometry can also be done with white light, but then the path length of coherence, the coherence length L (see **More Info** for equation) is much shorter:

Application

Interferometry is applied in a wide variety of fields in science. It is also the basic technique of Optical coherence tomography. There exists also a Doppler modification (ref. 1).

More Info

The coherence length is the propagation distance from a coherent source to a point where an electromagnetic wave maintains a specified degree of coherence (the normalized correlation of electric fields, in its simplest form the average of the normalized crosscorrelations (see Stochastic signal analysis) within an ensemble of waves). The significance is that interference will be strong within a coherence length of the source, but not beyond it. In long-distance transmission systems, the coherence length may be reduced by propagation factors such as dispersion, scattering, and diffraction.
The coherence length *L* is given approximately by:

$$L = \lambda^2/(n\Delta\lambda) = c/n\Delta f, \quad (1)$$

where λ is the central wavelength of the source, n is the refractive index of the medium, Δλ is the spectral width of the source, c is the speed of light in a vacuum and Δf is the bandwidth of the source. A more practical definition of the coherence length is the optical path length difference of a self-interfering laser beam which corresponds to a 50% fringe visibility, where the fringe visibility is defined as:
$V = (I_{max} - I_{min})/ (I_{max} + I_{min})$ where I is the fringe intensity.
Helium-neon lasers have a typical coherence length of 20 cm, while semiconductor lasers reach some 100 m.

Special types of interferometry
Coherent interferometry
Coherent interferometry uses a coherent light source (e.g. a He-Ne neon laser), and can make interference with large difference between the interferometer path length delays. The interference is capable of very accurate (nm) measurement by recovering the phase.
One of the most popular methods of interferometric phase recovery is phase-shifting by piezoelectric transducer phase-stepping. By stepping the path length by a number of known phases (minimum of three) it is possible to recover the phase of the interference signal.
The applications are nm-surface profiling, microfluidics (DNA chips, lab-on-a-chip technology), mechanical stress/strain, velocimetry.
Speckle Interferometry
In optical systems, a speckle pattern is a field-intensity pattern produced by the mutual interference of partially coherent beams that are subject to minute temporal and spatial fluctuations. This speckling effect is most commonly observed in fiber optics.
Holography
A special application of optical interferometry using coherent light is Holography, a technique for photographically recording and re-displaying 3D scenes.
Low-coherence interferometry
This type utilizes a light source with low temporal coherence such as white light or high specification femtosecond lasers. Interference will only be achieved when the path length delays of the interferometer are matched within the coherence time of the light source. It is suited to profiling steps and rough surfaces. The axial resolution of the system is determined by the coherence length of the light source and is typically in the µm-range.
Optical coherence tomography is a medical imaging technique based in low-coherence interferometry, where subsurface light reflections are resolved to give tomographic visualization. Recent advances have striven to combine the nm-phase retrieval with the ranging capability of low-coherence interferometry.

References
1. Talley DG et al. , Sankar SV, Bachalo WD.Phase-Doppler Interferometry with Probe-to-Droplet Size Ratios Less Than Unity. II. Application of the Technique. Appl Opt. 2000 Aug 1;39(22):3887-93.

Lambert-Beer law

Principle

The Lambert-Beer law, also known as Beer's law or the Beer-Lambert-Bouguer law relates the absorption of light to the properties of the material through which the light is traveling.

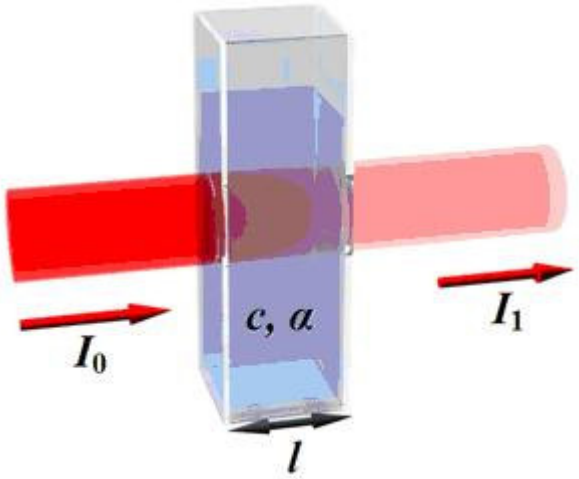


Fig. 1 Beer-Lambert absorption of a beam of light as it travels through a cuvette of size l.

There are several ways in which the law can be expressed:

$A = \alpha l c$, (1)
$I_1/I_0 = 10^{-\alpha l c}$, and (2)
$A = -\log (I_1/I_0)$ (3)

with $\alpha = 4\pi k/\lambda$

where:
A is absorbance or extinction
I_0 is the intensity of the incident light
I_1 is the intensity after passing through the material
l is the distance that the light travels through the material (the path length)
c is the concentration of absorbing species in the material
α is the absorption coefficient or the molar absorptivity of the absorber, the logarithmic decrement per unit path length per mol
λ is the wavelength of the light
k is the extinction coefficient.

In essence, the law states that there is an exponential dependence between the transmission of light through a substance and the concentration of the substance, and also between the transmission and the length of material that the light travels through. Thus if l and α are known, the concentration of a substance can be deduced from the amount of light transmitted by it.
The units of c and α depend on the way that the concentration of the absorber is being expressed (see More info).

Application

The law's link between concentration and light absorption is the basis behind the use of Spectroscopy to identify substances.
Further the law is applied in many medical, scientific and technical instruments.

More info

If the material is a liquid, it is usual to express the absorber concentration c as a mole fraction (the number of moles divided by the total number of moles in the liquid, so the mole fraction is dimensionless. The units of α are thus reciprocal length (e.g. cm^{-1}). In the case of a gas, c may be expressed as a density (units of reciprocal length cubed, e.g. cm^{-3}), in which case α is an absorption cross-section and has units of length squared (e.g. cm^2). If concentration c is expressed in moles per unit volume, α is a molar absorptivity usually given in units of mol cm^{-2}. In spectroscopy, often extinction (E = clk) is used.
The value of the absorption coefficient α varies between different absorbing materials and also with wavelength for a particular material. It is usually determined by experiment.
The law tends to break down at very high concentrations, especially if the material is highly scattering (see Light: scattering). If the light is especially intense or c is high (effect of self- screening), nonlinear optical processes can cause deviations.

Laser

Principle

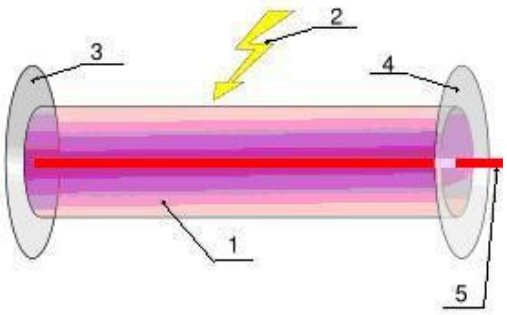
A laser (*L*ight *A*mplification by *S*timulated *E*mission of *R*adiation) is an optical source that emits photon in a coherent beam (i.e. when the beam is split in many tiny waves, actually consisting of a single photon, they are all in phase, have the same polarization and form together a single well-defined wave front). Laser light is typically near-monochromatic, i.e. consisting of a single wavelength or color, and emitted in a narrow beam. This is in contrast to common light sources, such as the incandescent light bulb, which emit incoherent photons in almost all directions, usually over a wide spectrum of wavelengths. Laser action is based on quantum mechanics and thermodynamics theory.

Application

In industry, technology, science and consumer electronics etc. the number of applications is enormous. In medicine, the laser scalpel is used for surgical techniques. Lasers are also used for dermatological procedures including removal of birthmarks, tattoos etc. Laser types used in dermatology include ruby (694 nm), alexandrite (755 nm), pulsed diode array (810 nm), and the various YAG laser. The YAG lasers emit infrared light. The YAG lasers are also applied in ophthalmology.
LASIK surgery is *la*ser-assisted in *si*tu *k*eratomileusis. First a low-power laser creates a topographic map of the cornea. Then an excimer laser (193 nm) is used to remodel the corneal stroma. The laser vaporizes tissue without causing damage.
Lasers are also used as the light delivery system in a fibro scope (see Fiber optics) for Endoscopy.
safety The coherence and low divergence of laser light means that it can be focused by the eye into an extremely small spot on the retina, resulting in localized burning and permanent damage in seconds or even faster. Lasers are classified into safety classes numbered I (inherently safe) to IV. Even scattered light can cause eye and/or skin damage. Laser products available for consumers, such as CD players and laser pointers are usually in class I, II, or III.

More info

A laser is composed of an active laser medium and a resonant optical cavity. Fig. 1 gives the principle of producing a laser beam.



1. Active laser medium, 2. Laser pumping energy, 3. Mirror, 4. Partial mirror, 5. Laser beam.

The gain medium is a material of controlled purity, size, and shape. The material itself is for instance a helium-neon gas or rubidium gas. The gain medium uses a quantum mechanical effect called stimulated emission to amplify the beam. This is the process by which, when perturbed by a photon, matter may lose energy resulting in the creation of another photon. The perturbing photon is not destroyed in the process (as with absorption), and the second photon is created with the same phase and frequency (or wavelength) as the original. These are the photons of the laser beam.
For a laser to operate, the gain medium must be "pumped" by an external energy source, such as electricity or light (from a classical source such as a flash lamp, or another laser). The pump energy is absorbed by the laser medium to produce exited states. (An exited state exists of an electron in a higher orbit than the lowest possible one, the ground state, so with higher energy than the ground state that is more energy than the absolute minimum). The mirrors enable multiple reflection so that the photons remain for a longer time in medium. In this way more easily excited particles are created and the number of particles in one excited state considerably exceeds the number of particles in some lower state. In this condition, an energy providing optical beam passing through the medium produces more stimulated emission than stimulated absorption. The process can be thought of as optical amplification, and it forms the basis of the laser (or for radio waves the maser).

Light

Principle

Light is the part to the electromagnetic spectrum that is visible to the animal eye. The study of light and the interaction of light and matter is termed optics.
The elementary particle that defines light is the photon. The three basic dimensions of light (or better electromagnetic radiation) are:

- intensity, which is related to the human perception of brightness of the light;
- frequency (or wavelength), perceived by humans as the color of the light, and
- polarization (or angle of vibration), which is only weakly perceptible by humans under ordinary circumstances.

Due to the wave-particle duality of matter light simultaneously exhibits properties of both waves and particles.
Here follows a description of the most important features of light.

Speed of light The speed of light in a vacuum is exactly 299,792,458 m/s (fixed by definition).
Refraction When light goes from the one to another medium, it is refracted (see Light: refraction) and reflected (see Fresnel equations).
Dispersion Since refraction is frequency dependent, the refracted beam is decomposed in its various frequencies (or wavelengths) which all have their own angle of refraction. The classical way to achieve this is with a prism, see Fig. 1.

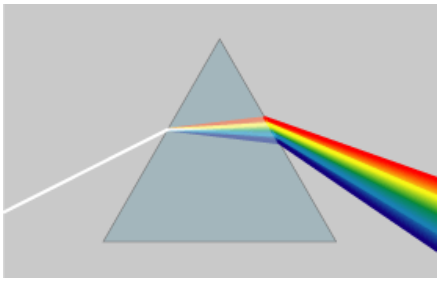


Fig. 1 Dispersion of a light beam in a prism.

The visible spectrum (see Fig. 1).
Electromagnetic radiation from 400 to 700 nm is called visible light or simply light. However, some people may be able to perceive wavelengths from 380 to 780 nm. A light-adapted eye typically has its maximum sensitivity at around 555 nm, which is in the green region (see: Luminosity function). In day light, so with photopic vision, the different wavelengths are detected by the human eye and then interpreted by the brain as colors. The spectrum does not, however, contain all the colors that the human eyes and brain can distinguish. For instance, brown and pink are absent. See Vision of color to understand why.
The *optical spectrum* includes not only visible light, but also ultraviolet (UV) at the short wavelength (high frequency) end and infrared (IR) at the long wavelength end. Some animals, such as bees, can see UV radiation while others, such as pit viper snakes, can see IR light.

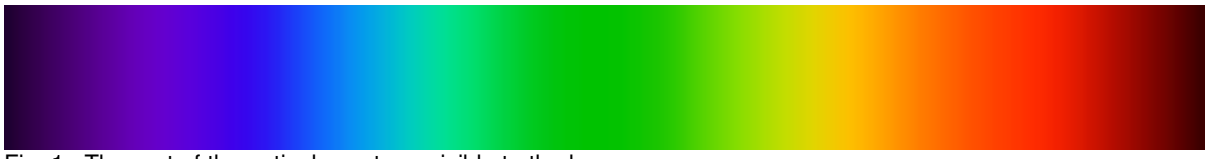


Fig. 1 The part of the optical spectrum visible to the human eye.

Polarization With reflection and refraction light is also polarized to some extent (see Light: polarization). Polarization describes the direction of the electric oscillation in the plane perpendicular to the direction of propagation.
Diffraction This refers to phenomena associated with wave propagation, such as the bending, spreading and interference of waves emerging from an aperture on the order of the wavelength (pinhole and narrow-split experiments). For more explanation see Light: diffraction.
Absorption When light propagates trough a medium some of its energy is absorbed by the medium (see Lambert-Beer law). In general, all or most of the absorbed energy is transformed to heat. The part that is not transformed to heat can be emitted as radiation (see Chemoluminescence and

Bioluminescence, Fluorescence, Phosphorescence) or transformed to electric current (the photoelectric effect, see **More Info**).
Scattering Scattering of is a process whereby light (and sound or moving particles), are forced to deviate from a straight trajectory by one or more localized non-uniformities in the medium through which it passes. This also includes deviation of reflected radiation from the angle predicted by the law of reflection (called *diffuse* reflections). An example is scattering of light in the eye lens and intraretinal scatter. See further Light: Scattering.

Theories about light

Classical particle theory (Newton)
Light was assumed to be composed of corpuscles (particles of matter) which were emitted in all directions from a source. This theory cannot explain many of the properties of light. It wrongly assumed a higher speed in a denser medium. The classical particle theory was finally abandoned around 1850.

Classical wave theory (Huygens)
Light was (and is) assumed to be emitted in all directions as a series of waves in a medium (Fig. 3). As waves are not affected by gravity, it was assumed that they slowed down upon entering a denser medium. It can explain phenomena such as refraction, polarization, dispersion and diffraction. It was wrongly assumed that light waves would need a medium for transmission (like sound waves indeed need).

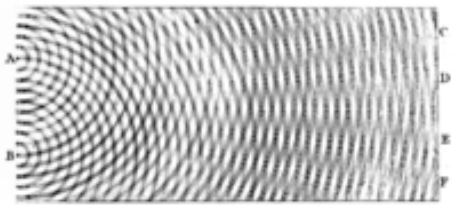


Fig. 3 Interference of the waves emitted by two sources.

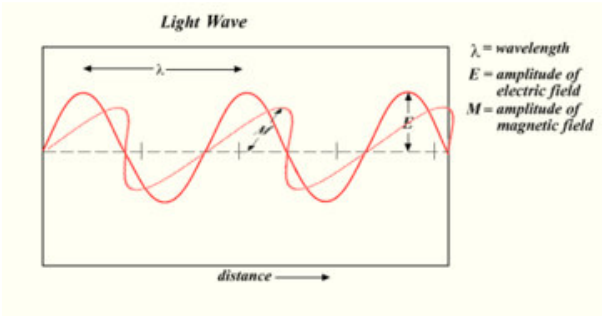


Fig. 4 A linearly-polarized light wave frozen in time and showing the two oscillating components of light that generate an electric and a magnetic field perpendicular to each other and to the direction of motion.

Electromagnetic theory
The angle of polarization of a beam of light as it passed through a polarizing material could be altered by a magnetic field, an effect now known as Faraday rotation. It is one of the arguments that light is a high-frequency electromagnetic vibration, which could propagate even in the absence of a medium such as the “ether”. The frequency range of light is only a very small part of the whole electromagnetic range. Other parts of the electromagnetic spectrum are applied in e.g. radio, radar, television and electromagnetic imaging (see Spectroscopy).

Application

Medical Many medical instruments and apparatus are based on light for imaging, as do also prostheses like spectacles etc. All this things will not be discussed. Here some applications based on UV and IR light are mentioned.
UV radiation is not normally directly perceived by humans except in a much delayed fashion, as overexposure of the skin. UV light can cause sunburn, or skin cancer. Underexposure can cause vitamin D deficiency. However, because UV is a higher frequency radiation than visible light, it very easily can cause materials to fluorescence visible light.

Thermography is performed with a camera using IR light. In general heating of the skin or the whole body by radiation is caused by IR light. However, any intense radiation can have the same effect. Other examples are UV and IR spectroscopy (see Spectroscopy).
Technical IR cameras convert IR light to visible light. Depending on their application we distinguish night-vision binoculars, cameras. These are different from image intensifier cameras, which only amplify available visible light.

More info

The Special Theory of Relativity
The wave theory explains nearly all optical and electromagnetic phenomena, but some anomalic phenomena remained that could not be explained:
- the constant speed of light,
- the photoelectric effect,
- black body radiation.
The constant speed of light contradicted the mechanical laws of motion, which stated that all speeds were relative to the speed of the observer. This paradox was resolved by revising Newton's laws of motion into Einstein's special theory of relativity.
The photoelectric effect, being the ejection of electrons when light strikes a metal surface, causing an electric current to flow out. The explanation is given by the wave-particle duality and quantum mechanics.
A third anomaly involved measurements of the electromagnetic spectrum emitted by thermal radiators, or so-called black bodies (see Wien's displacement law and Body heat dissipation and related water loss). The explanation is given by the *Quantum theory*. The theory of black body radiation says that the emitted light (and other electromagnetic radiation) is in the form of discrete bundles or packets of energy. These packets were called quanta, and the particle of light was given the name photon, just as other particles, such as an electron and proton. A photon has an energy, E, proportional to its frequency, f:

$$E = hf = hc/\lambda, \quad (1)$$

where h is Planck's constant (= $6{,}623 \cdot 10^{-34}$ Js), λ is the wavelength and *c* is the speed of light. Likewise, the momentum (mass times speed) p of a photon is also proportional to its frequency and inversely proportional to its wavelength:

$$p = E/c = hf/c = h/\lambda. \quad (2)$$

Wave-particle duality and of *quantum electrodynamics*
The modern theory that explains the nature of light is the wave-particle duality, founded by quantum theory. More generally, the theory states that everything has both a particle nature and a wave nature, and various experiments can be done to bring out one or the other. The particle nature is more easily discerned if an object has a large mass, but also particles, such as electrons and protons exhibited wave-particle duality. The quantum mechanical theory of light and electromagnetic radiation culminated with the theory of quantum electrodynamics, or QED.

Light: beam splitter

Principle

A beam splitter is an optical device that splits a beam of Light in two. In its most common form, it is a cube, made from two triangular glass prisms, which are glued together by resin (Fig. 1). The thickness of the resin layer is adjusted such that (for a certain wavelength) half of the light incident through one "port" (i.e. face of the cube) is reflected and the other half is transmitted. Polarizing (see Light: polarization) beam splitters, use birefringent materials (two instead of one refractive index), splitting light into beams of differing polarization.

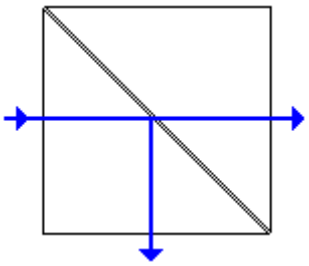


Fig. 1 Schematic representation of a beam splitter cube

Another design is a *half-silvered mirror*. This is a plate of glass with a thin coating of (nowadays) aluminum. The coating is such thick that light incident at 45° is half transmitted and half reflected.

Application

Beam splitters are found in all kind of (medical) equipment, especially in microscopes, spectroscopes, instruments with laser and ophthalmic instruments.

More Info

A third version of the beam splitter is a trichroic mirrored prism assembly which uses trichroic optical coatings to split the incoming light into three beams, one each of red, green, and blue (see Dichroism). Such a device was used in multi-tube color television cameras and in 3-color film movie cameras. Nowadays it is often applied in CCD cameras. Other applications are LCD screens and LCD projectors. (An LCD (liquid crystal display) is a low powered electronically-modulated optical thin, flat panel consisting of color or monochrome pixels filled with liquid crystals and arrayed in front of a light source (backlight) or reflector).

Light: diffraction

Principle

Diffraction refers to various phenomena associated with wave propagation, such as the bending, spreading and interference of waves emerging from an aperture. It occurs with any type of wave, including sound waves, water waves, electromagnetic waves such as light and radio waves, and matter displaying wave-like properties according to the wave–particle duality (see Light). While diffraction always occurs, its effects are generally only noticeable for waves where the wavelength is on the order of the size of the diffracting object (e.g. slit) or aperture.
Much theory has been developed to describe diffraction patterns of a pinhole and of one, tow and more slits. They all rely on the Huygens' principle. Fig. 1 gives the pattern of a single slit.
In slit experiments, narrow enough slits can be analyzed as simple wave sources. A slit reduces a wave 3-D problem into a simpler 2-D problem.

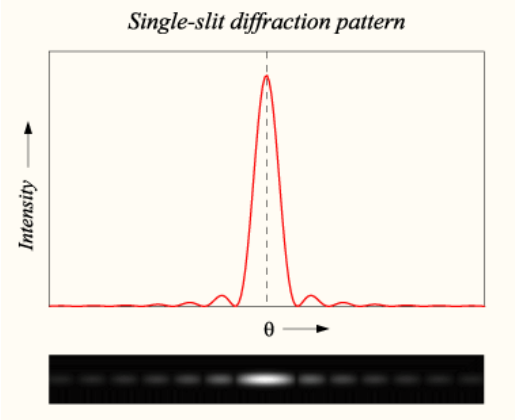


Fig. 1 Graph and image of single-slit diffraction

It is possible to calculate exactly the intensity of the diffraction pattern as a function of angle in the case of single-slit diffraction. Since the angles of the maxima and minima in the pattern are dependent on the

wavelength λ, diffraction gratings impart angular dispersion (decomposed in wavelengths or “colors”, see Light) on a beam of light.

Application

There are many applications in specific fields of physics, and technology, and so indirectly in medicine. Diffraction plays a major role in light- and electron microscopy.
Diffraction is used in X-ray crystallography, called *Bragg diffraction*, (see **More Info**) to deduce the structure of a crystal from the angles at which X-rays are diffracted from it. The most common demonstration of Bragg diffraction is the spectrum of colors seen reflected from a compact disc: the closely-spaced tracks on the surface of the disc form a diffraction grating, and the individual wavelengths of white light are diffracted at different angles from it.

More info

The angels of the maxima and minima are inversely proportional to the slit or aperture diameter. So, the smaller the diffracting object the 'wider' the resulting diffraction pattern. When the diffracting object is repeated, for example in a diffraction grating the effect is to create narrower maxima on the interference fringes.
It is mathematically easier to consider the case of so called far-field or *Fraunhofer diffraction*, where the slits or pinholes are far from the point where the pattern is observed, so in the far-field, i.e. distance >> λ. The more general case is known as near-field or Fresnel diffraction, and involves more complicated mathematics. Here, far-field diffraction is considered, which is commonly observed in nature.
Diffraction through a circular aperture
For diffraction through a circular aperture, there is a series of concentric rings surrounding a central disc (together called the Airy pattern illustrated in Fig. 2)).
A wave does not have to pass through an aperture to diffract. Any beam of light undergoes diffraction and spreads in diameter. This effect limits the minimum size d of a spot of light formed at the focus of a lens, known as the diffraction limit:

$$d = 1.22\,\lambda f/a, \qquad (1)$$

where λ is the wavelength of the light, f is the focal length of the lens, and a is the diameter of the beam of light, or (if the beam is filling the lens) the diameter of the lens. The spot contains about 70% of the light energy.
By use of Huygens' principle, it is possible to compute the diffraction pattern of a wave from any shaped aperture. If the pattern is observed at a sufficient distance from the aperture, in the far-field, it will appear as the two-dimensional Fourier transform of the function representing the aperture.

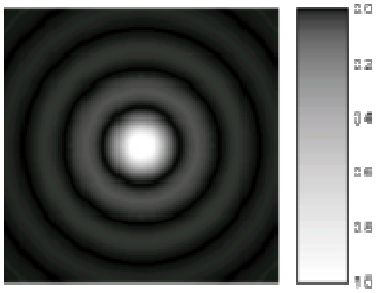


Fig. 2 The image of an Airy pattern. (The grey scale intensities have been adjusted to enhance the brightness of the outer rings of the pattern.)

Diffraction through slits
Fig. 3 illustrates the principle of interference with two slits in the far-field description. The two “rays” passing the two slits of Fig. 3 are supposed to be parallel (the prerequisite of the Fraunhofer diffraction). In this case they have an angle such that there is a path length difference of 0.5 λ. This creates the first minimum in the interference pattern, since the both waves extinct each other (destructive interference). Fig. 4 gives all maxima (with crossings of both spherical wave fronts) and minima (black crossings). The angular positions of the multiple-slit minima correspond to path length differences of an odd number of half wavelengths:

$$a\sin\theta = \tfrac{1}{2}\lambda(2m+1), \qquad (2)$$

where m is an integer that labels the order of each minimum, a is the distance between the slits

and θ is the angle for destructive interference. The maxima are at path differences of an integer number of wavelengths:

$$a\sin\theta = m\lambda. \quad (3)$$

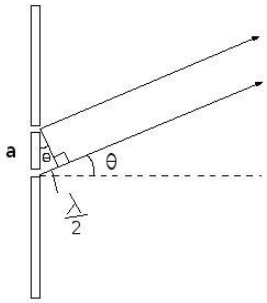


Fig. 3 Principle of interference. The slit are perpendicular on the plane of the screen.

We can calculate that for three waves from three slits to cancel each other the phases of slits must differ 120^0, thus path difference from the screen point to slits must be λ/3, and so forth. In the limit of approximating the single wide slit with an infinite number of subslits the path length difference between the edges of the slit must be exactly λ to get complete destructive interference for indidence angle θ (and so a dark stripe on the screen).

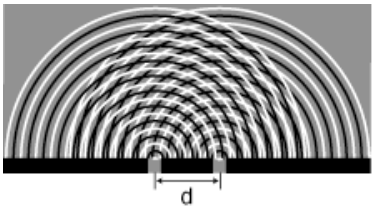


Fig. 4 Double-slit diffraction and interference pattern. The slits are perpendicular on the plane of the paper.

Bragg diffraction
Diffraction from multiple slits, as described above, is similar to what occurs when waves are scattered from a periodic structure, such as atoms in a crystal or rulings on a diffraction grating. Each scattering center (e.g., each atom) acts as a point source of spherical wave fronts; these wave fronts undergo constructive interference to form a number of diffracted beams. The direction of these beams is described by *Bragg's law*:

$$m\lambda = 2L\sin\theta, \quad (4)$$

where L is the distance between scattering centers and m is an integer known as the *order* of the diffracted beam.

Diffraction of particles
The experimental prove of the diffraction of "particles," such as electrons, was one of the powerful arguments in favor of quantum mechanics and the existence of the wave-particle duality. The wavelengths of these particle-waves are small enough that they are used as probes of the atomic structure of crystals.

Light: Fresnel equations

Principle

The Fresnel equations, describe the behavior of light when moving between media of differing refractive indices (see Light: refraction). The *reflection* of light that the equations predict is known as Fresnel reflection and the refraction is described by Snell's law.

When light moves from a medium of a given refractive index n_1 into a second medium with refractive index n_2, both reflection and refraction of the light may occur.

The fraction of the intensity of incident light that is reflected from the interface is given by the *reflection coefficient R*, and the fraction refracted by the *transmission coefficient T*. The Fresnel equations assume that the two media are both *non-magnetic*.

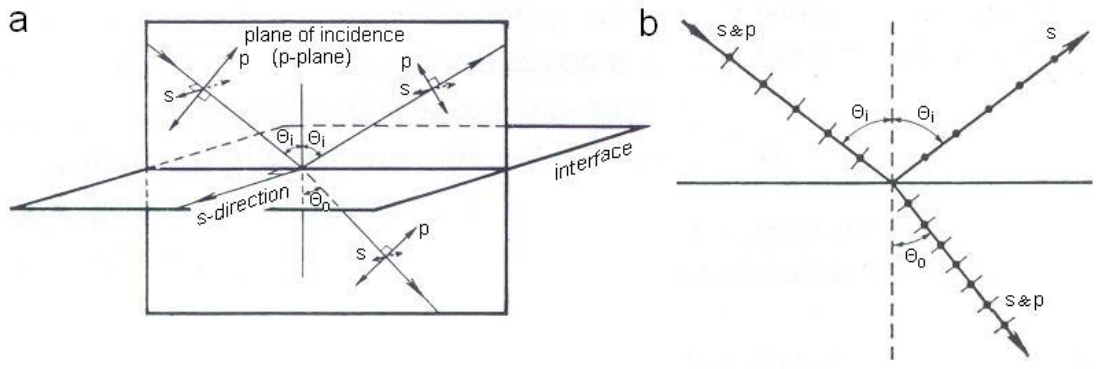


Fig. 1. a. Visualization of the decomposition of the unpolarized beam is the p and s-polarized parts. b. The angle of incidence is Brewster's angle. The reflected beam is purely s-polarized.

The calculations of R and T depend on the polarization of the incident ray. If the light is polarized with the electric field of the light perpendicular to the plane of Fig. 1a (s-polarized), the reflection coefficient is given by R_s. If the incident light is polarized in the plane of the diagram (p-polarized), the R is given by R_p.
A graphical representation of the calculation of R_s and R_p as function of θ_i is given in Fig. 2.

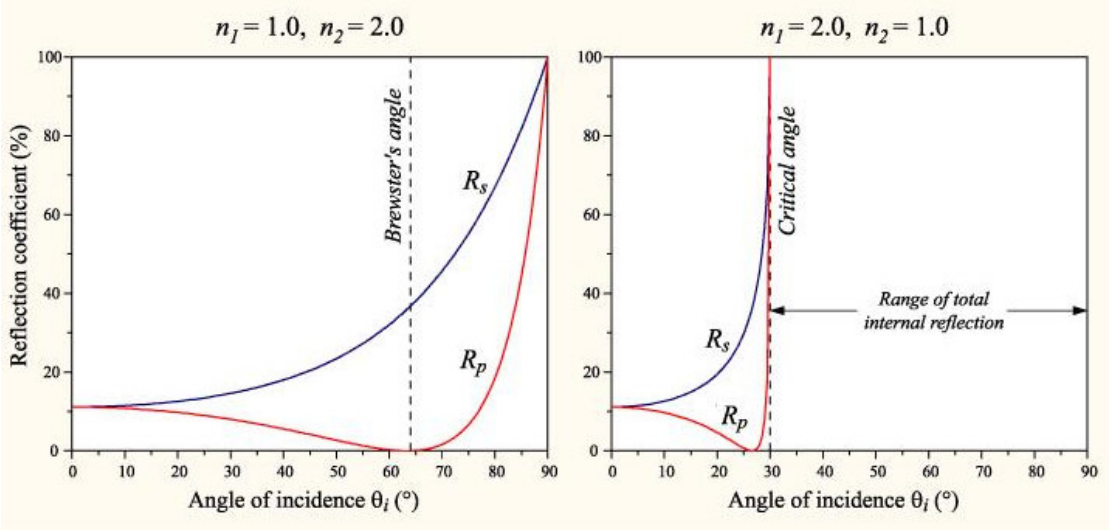


Fig. 2 Left, R_s and R_p for a light beam going from a medium with a low refractive index to one with a high index. Right, R_s and R_p for a light beam going from high to low refractive index.

At one particular angle for a given n_1 and n_2, the value of R_p goes to zero and a s-polarized incident ray is purely refracted. This angle is known as Brewster's angle, and is around 56° for a glass medium in air or vacuum.
When moving from a more dense medium into a less dense one (i.e. $n_1 > n_2$), exceeding an incidence angle known as the critical angle, all light is reflected and $R_s = R_p = 1$. This phenomenon is known as total internal reflection (see Snell's law). The critical angle is approximately 41° for glass in air.
If the incident light is unpolarized (containing an equal mix of s- and p-polarizations), the reflection coefficient is $R = (R_s + R_p)/2$.

When the light is about perpendicular to the interface ($\theta_i \approx \theta_t \approx 0$), R and T are given by:

$$R = \{(n_1 - n_2)/(n_1 + n_2)\}^2, \text{ and} \quad (1)$$
$$T = 1 - R. \quad (2)$$

For common (clean!) glass, R is about 4%. Note that reflection by a window is from the front side as well as the back side, and that some of the light bounces back and forth a number of times between the two sides. The combined reflection coefficient for this case is 2R/(1 + R).

Application

Applying Brewster's angle is a way to produce pure polarized light. Repeated reflection and refraction on thin, parallel layers is responsible for the colors seen in oil films on water, used in optics to make reflection free lenses and perfect mirrors, etc.

More info

The general equations for R_s and R_p are:

$$R_s = \left(\frac{n_1 \cos\Theta_i - n_2 \cos\Theta_t}{n_1 \cos\Theta_i + n_2 \cos\Theta_t} \right)^2 \quad (3a)$$

and:

$$R_p = \left(\frac{n_1 \cos\Theta_t - n_2 \cos\Theta_i}{n_1 \cos\Theta_t + n_2 \cos\Theta_i} \right)^2 \quad (3b)$$

where θ_p can be derived from θ_i by Snell's law.

In each case T is given by $T_s = 1 - R_s$ and $T_p = 1 - R_p$.
R and S correspond to the ratio of the *intensity* of the incident ray to that of the reflected and transmitted rays. Equations for coefficients corresponding to ratios of the electric field amplitudes of the waves can also be derived, and these are also called "Fresnel equations". Here, we use squared amplitude (just as in the above equations n^2). Therefore, the intensity is expressed in Watts, more precisely in W/steradian (see Light: photometric and radiometric units of measure and Radian and steradian). The above equations are approximations. The completely general Fresnel equations are more complicated (see e.g. Wikipedia).

Light: the ideal and non-ideal lens

Principle

The ideal lens has the following properties:

- All light rays emanating from a point of light, an object point, at the object distance from the lens, are refracted by the lens through one point, the image point.
- Rays hitting the lens parallel to its optic axis are refracted through its focus.
- Rays hitting the lens at its centre are not refracted.

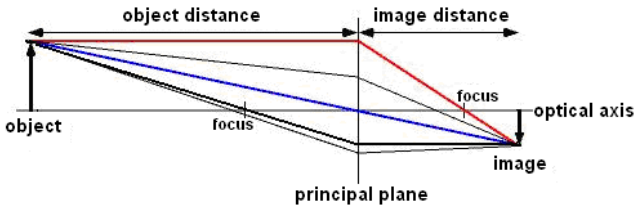


Fig. 1 Simple formal description of a positive lens, presented by a single principal plane (the vertical line). The image is real, i.e. can be made visible at a piece of paper.

The lens equation relates object distance, image distance and lens power:

$$1/d_{object} + 1/d_{image} = 1/f = \varphi, \quad (1)$$

where d is distance and φ lens power. It holds for *thin* lenses.

Note that if $d_{object} < f$, d_{image} becomes negative, the image is positioned on the same side of the lens as the object. This imaginary image cannot be projected on a screen, but it can be observed when looking from the right through the lens. A magnifying-glass creates this kind of image. The magnification is given by:

$$M = -d_{image}/d_{object} = f/(f - fd_{object},) \quad (2)$$

where M is the magnification factor. If $|M|>1$, the image is larger than the object. Notice the sign convention here shows that, if M is negative, as it is for real images, the image is upside-down with respect to the object (Fig. 1) and smaller. For virtual images, M is positive, the image is upright and larger than the object.
In (1), negative lenses 1/f obtains a minus sign. A negative lens cannot make an image at all. It is always practiced together with a positive lens, generally to compensate for lens errors (see **More Info**).
When two ideal lenses are put next to each other, their powers add. In formula: $\varphi = \varphi_1 + \varphi_2$. This does not apply if the lenses are not close together. Then, $\varphi = \varphi_1 + \varphi_2 - d_{12}\varphi_1\varphi_2$ where d_{12} is the distance between both lenses.
Real lenses do not behave exactly like ideal lenses but thin lenses are fairly ideal. Therefore, the term thin lens is often used as a synonym for ideal lens. However, the most ideal lenses available in practice (such as camera and microscope objectives) are often very thick, being composed of several lenses with different refractive indices.

Application

A lens shows its most ideal behavior when only its central part is used. An example is the photopic adapted eye with its very large φ but small pupil, since good imaging is needed at the fovea centralis. Therefore, ideal lens optics can be useful in elementary eye optics (see Optics of the eye).

Lenses are applied in innumerous medical apparatus and research equipment.
Lenses and the theory of lenses play a fundamental role in all versions of light microscopy and in ophthalmology.
The human cornea-lens system has comparatively small lens errors (see below), seen its enormous refractive power (about 58 dptr). Only astigmatism can be corrected well.

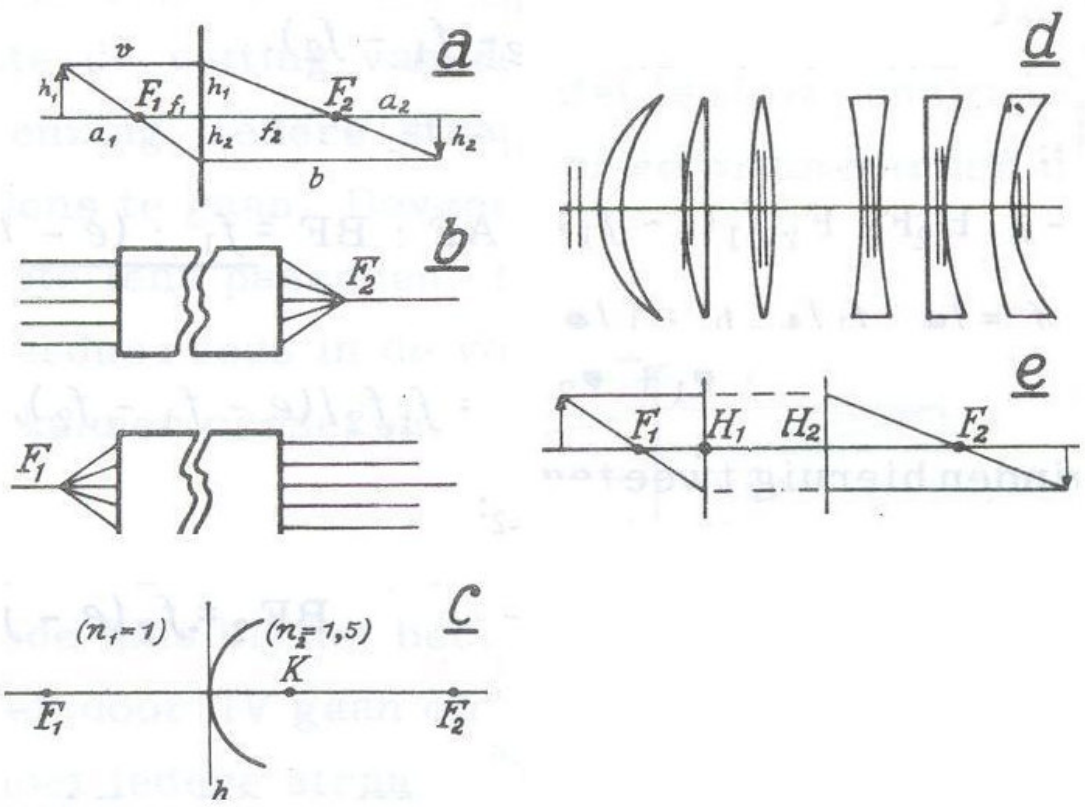


Fig. 2 Principle planes a the lens concept with one main plane, b and e with two main planes, c two media separated by a spherical interface, for instance the cornea (with K the center of the sphere. K is

also called the principal point). An object in the medium with n=1.5 produces a diverging beam at the left. This is related to the fact that $F_2 > F_1$. d position of the two main planes for various types of lenses.

More info

Strenght
If the thickness of the lens is small compared to the both radii of curvature, r_1 and r_2, then the thin lens approximation can be made, and *f* is then estimated as:

$$1/f = (n_{lens}/n_{medium} - 1)(1/r_1 - 1/r_2) \quad (3)$$

Notice that for e.g. a double convex lens r_1 and r_2 have opposite signs since both sphere centers are located opposite.

Principal planes
A more precise description is applying two principal planes. Fig. 2 illustrates how to handle this concept. Spectacle lenses are generally convex-concave (positive, left one Fig. 2e) or convex-convex (negative, right one Fig. 2e). As Fig. 2d indicates, their principal planes lay outside the lens itself. Therefore, for accurate calculations, the single-principal-plane concept is generally not adequate. The theory of the two principal planes, illustrated in Fig. 2e, holds for any system of lenses.

Monochromatic lens errors
These aberrations occur when light comprising even a single wavelength is used. Most monochromatic aberrations are 3rd order errors, since their size is dependent on the summed powers of lens diameter d and object size from the axis h (e.g. d^2h), yielding a power of 3.

Spherical aberrations Even a thin lens does not produce an image *point* as suggested by Fig. 1, but a small blurred spot, which intensity falls off with eccentricity. The diameter of the circle increases with d^3.
Coma This is an imperfection, which remains when spherical aberration is solved in a lens system. It is the comet-like blurring spot (with the expanding tail pointing outwards) of the image of a parallel beam which hits the lens not-parallel to its optical axis. It is proportional with d^2h.
Astigmatism This happens when for two (perpendicular) planes through the optical axis the power φ is different. It produces an ellipse with blurred and sharp sectors when a circle is imaged. It is proportional with dh^2.
I*mage curvature* Another error, occurring with large objects and large image angle is the curvature of the plane if the image. Depending on the lens system, the image is cushion shaped or ton shaped when displayed on a straight plane. This error is also proportional with dh^2.
Image distortion The distortion is pincushion shaped or barrel shaped. The error is proportional with h^4.

Chromatic aberration
A non-monochromatic error is chromatic aberration. This is caused by the wavelength dependency of the refractive index, called dispersion (see Light: refraction), resulting in a φ dependent on wavelength. Finally this results in an image for each wavelength with a shorter d_{image} the shorter the wavelength. It can be minimized in an achromatic doublet lens, consisting of two lenses of different types of glass.

Diffraction "eror
"When a small aperture is hold in front of a positive lens, then a parallel axial beam also produces a circular spot. This is due to the wave character of light. The diameter of the spot is distance between the two first order minima of the single-slit diffraction pattern:

$$d = 2.44(\lambda/d_{lens})(f/n_0), \quad (4)$$

where λ is the wavelength and n_0 the refractive index of the medium between lens and spot.

Since various errors occur simultaneously the image of a single point produces a complicated figure as image.

Literature
For more info see student textbooks of physics, Wikipedia chapters and books of optics.

Light: polarization

Principle

In electrodynamics, polarization is the property of electromagnetic waves, such as light, that describes the direction of their transverse electric field. More generally, the polarization of a transverse wave describes the direction of oscillation in the plane perpendicular to the direction of propagation. Longitudinal waves, such as sound waves, (see Sound and Acoustics) do not exhibit polarization, because for these waves the direction of oscillation is along the direction of propagation.

The simplest manifestation of polarization to visualize is that of a plane wave, which is a good approximation to most light waves (a plane wave is a wave with infinitely long and wide wave fronts). All electromagnetic waves propagating in free space or in a uniform material of infinite extent have electric and magnetic fields perpendicular to the direction of propagation. Conventionally, when considering polarization, the electric field vector is described and the magnetic field is ignored since it is perpendicular to the electric field and proportional to it. The electric field vector may be arbitrarily divided into two perpendicular components labeled *x* and *y* with z indicating the direction of propagation, see Fig. 1. The x vector is at the left (red) and the y-vector at the right (green). For a simple harmonic wave (sinus), the two components have the same frequency, but the amplitude and phase may differ. In the example of Fig. 1, the projection of the movement of the electric vector (blue in Fig. 1) on the horizontal x-y plane creates the purple line (the most simple Lissajous figure).

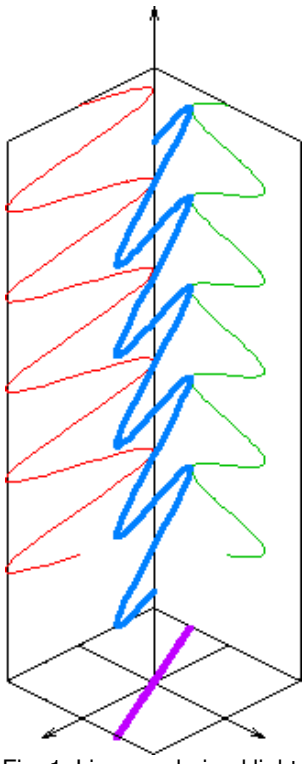


Fig. 1 Linear polarized light.

In Fig. 1, the two orthogonal (perpendicular) components are in phase. Since the ratio of the strengths of the two components is constant, the direction of the electric vector (the vector sum of these two components) is constant. And since the tip of the vector traces out a single line on the plane perpendicular on the direction of propagation, this special case is called *linear polarization*. The direction of this line depends on the relative amplitudes of the two components.

Application

Light reflected by shiny transparent materials is partly or fully polarized, except when the light is normal (perpendicular) to the surface. A polarizing filter, such as a pair of polarizing sunglasses, can be used to observe this by rotating the filter while looking through. At certain angles, the reflected light will be

reduced or eliminated. Polarizing filters remove light polarized at 90° to the filter's polarization axis. If two polarizers are placed atop one another at 90 ° angles to one another, no light passes through. Linear polarization is used in polarization microscopy.
Polarization by scattering is observed as light passes through the atmosphere. The scattered light produces the brightness and blue color in clear skies. This partial polarization of scattered light can be used to darken the sky in photographs, increasing the contrast (Fig. 2). This effect is easiest to observe at sunset, on the horizon at a 90 ° angle from the setting sun.

Fig. 2 Left unpolarized, right partly polarized. The effects of a polarizer on the sky in a color photograph. The right picture has the polarizer, the left does not.

Brightness of images of the sky and clouds reflected from horizontal surfaces are drastic reduced, which is the main reason polarizing filters are often used in sunglasses.
Rainbow-like patterns, visible through polarizing sunglasses, are caused by color-dependent *birefringent* (double refraction, see Snell's Law) effects, for example in toughened glass (e.g. car windows) or items made from transparent plastics. The role played by polarization in the operation of liquid crystal displays (LCDs) is also frequently apparent to the wearer of polarizing sunglasses, which may reduce the contrast or even make the display unreadable.
[
Biology
The naked human eye is weakly sensitive to polarization. Polarized light creates a very faint pattern near the center of the visual field, called Haidinger's brush. With sun light this pattern, a yellow horizontal figure 8 shaped spot with a bleu spot above and below the figure 8, is very difficult to see, but with practice one can learn to detect polarized light with the naked eye.
Many animals are apparently capable of perceiving the polarization of light, which is generally used for navigational purposes, since the linear polarization of sky light is always perpendicular to the direction of the sun. This ability is very common among the insects, including bees, which use this information to orient their communicative dances. Polarization sensitivity has also been observed in species of octopus, squid, cuttlefish, and mantis shrimp. This ability is based on the polarizing effect of the anisotropic microstructure of the photoreceptor cells of these animals.

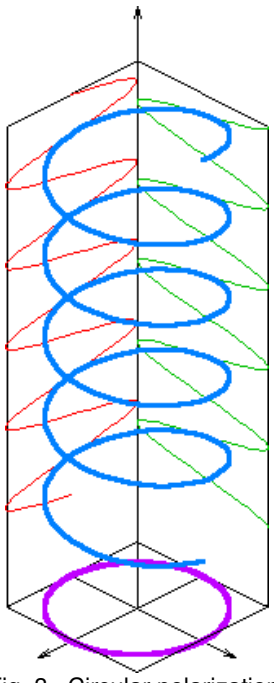


Fig. 3 Circular polarization.

More info

In Fig. 3 the two orthogonal components have exactly the same amplitude and are exactly (plus or minus) 90^o out of phase. In this special case the electric vector traces out a circle in the plane of projection, so this special case is called circular polarization. The direction of rotation depends on the sign of the phase difference.
All other cases, that is where the two components are not in phase and either do not have the same amplitude and/or are not 90^o out of phase are called elliptical polarization because the electric vector traces out an ellipse in the horizontal plane.

In nature, electromagnetic radiation is often produced by a large number of individual radiators, producing waves independently of each other. In general there is no single frequency but rather a spectrum of different frequencies present, and even if filtered to an arbitrarily narrow frequency range, there phases and planes of polarization are different. This type of light is described as *incoherent*. However, this does not mean that polarization is only a feature of coherent radiation. Incoherent radiation may show statistical correlation between the components of the electric field, which can be interpreted as *partial polarization*. In general it is possible to describe a beam of light as the sum of a completely incoherent part (no correlations) and a completely polarized part. One may then describe the light in terms of the *degree of polarization*, and the parameters of the polarization ellipse.

A more complete description, fundamental and mathematically can be found in Wikipedia.

Light: refraction

Principle

When light passes through a transparent substance, such as air, water or glass, its speed is reduced, and it undergoes refraction. The reduction of the speed of light in a denser material can be indicated by the refractive index, n, which is defined as:

$$n = c/v,$$

where c is the speed in vacuum and v the speed in the medium.
Thus, n = 1 in a vacuum and n > 1 in (any) matter.
When a beam of light enters a medium from vacuum or another medium, it keeps the same frequency. So, its “color” remains the same. However, its wavelength changes since speed changes. If the incident beam is not orthogonal to the edge between the media, the direction of the beam will change.
If the refractive indices of two materials are known for a given frequency, then one can compute the angle by which radiation of that frequency will be refracted as it moves from the first into the second material from Snell's law.

Application

The refractive index of a material is the most important property of any optical system that uses refraction. It is used to calculate the focusing power of lenses, and the dispersive power of prisms. Since the refractive index is a fundamental physical property of a substance, it is often used to identify a particular substance, confirm its purity, or measure its concentration. The refractive index is used to measure solids (glasses and gemstones), liquids, and gasses. Most commonly it is used to measure the concentration of a solute in an aqueous solution. A refractometer is the instrument used to measure the refractive index. For a solution of sugar, the refractive index can be used to determine the sugar content. Refraction of light by lenses is used to focus light in magnifying glasses, spectacles and contact lenses (see The ideal and non-ideal lens) and microscopes (see Light microscopy).

More info

The refractive index n of a material is defined by:

$$n = (\varepsilon_r \mu_r)^{0.5}, \qquad (1)$$

where ε_r is the medium's dielectric constant (relative permittivity), and μ_r is its relative magnetic permeability. For a non-magnetic material, μ_r is very close to 1, therefore n is approximately $\varepsilon_r^{0.5}$.
Remind that ε_r is strongly frequency dependent. So for water, one should not take its static ε_r, (with frequency zero) which is about 81!
The effect of the frequency dependency of n (except in vacuum, where all frequencies travel at the same speed, c) is known as dispersion, and it is what causes a prism to divide white light into its constituent spectral colors. It explains rainbows, and is the cause of chromatic aberration in lenses (see The ideal and non-ideal lens).
Anisotropy
The refractive index of certain media may be different depending on the polarization and direction of propagation of the light through the medium. This is known as birefringence or anisotropy (see Snell's Law).
There are some other phenomena related to refraction like non-linear behavior of n (e.g. Kerr effect), inhomogeneity etc. For this, the reader consults physical textbooks.

Light: scattering

Principle

Scattering is a process whereby some forms of radiation, such as light or moving particles, are forced to deviate from a straight trajectory by non-uniformities in the medium. It occurs also with sound. In conventional use, this also includes deviation of reflected radiation from the angle predicted by the law of reflection (see Light: refraction). Reflections that undergo scattering are often called *diffuse* reflections and unscattered reflections are called *specular* (mirror-like) reflections.
The types of non-uniformities that can cause scattering, the *scatterers* or *scattering centers* are for instance particles, bubbles, droplets, cells in organisms, density fluctuations in fluids and surface roughness.
Scattering can be distinguished between two broad types, *elastic* and *inelastic*. Elastic scattering involves no change of radiation energy, but inelastic scattering does. If the radiation loses a significant proportion of its energy, the process is known as *absorption*. This is governed by the Lambert-Beer law.
Major forms of elastic light scattering are *Rayleigh scattering* and so called *Mie scattering*. Inelastic EM scattering effects include Brillouin scattering (see **More Info**).
With Rayleigh scattering of light, or EM radiation, it holds that particle diameter $d < 0.1\lambda$ (λ is wavelength) of the light. It occurs when light travels in transparent solids and liquids, but especially in gases. Rayleigh scattering is proportional to λ^{-4}.
If $d > \lambda$, light is not separated in all its wavelengths and the scattered light appears white, as do salt and sugar.

Light scattering is one of the two major physical processes that contribute to the visible appearance of most objects. The other is absorption. Surfaces described as *white* owe their appearance almost completely to the scattering of light by the surface. The absence of surface scattering leads to a shiny or glossy appearance. Light scattering can also give color to some objects, usually shades of blue (as with the sky, the human iris, and the feathers of some birds), but resonant light scattering in nanoparticles can produce different highly saturated and vibrant hues.

Application

In medicine
Scattered light is the image forming light in dark field microscopy. Scattered sound plays the same role in Echography. However, in many applications of computer-generated imagery one tries to minimize the disturbance influence of scattering.
Some specific way of scattering is of importance for detecting DNA, proteins and for Fluorescence.
In ophthalmology it is of importance with respect of the quality of the eye media. Scattering in the eye media, especially in the eye lens disturb clear vision, especially under scotopic conditions. In vision research diffusers (see More Info) are often applied.
In daily life
Why is the sky blue? This effect occurs because blue photons hit the air molecules in the earth's atmosphere and are scattered down to the earth's surface. Red photons are not affected by the particles and pass on through the earth's atmosphere. This causes blue light to be scattered down to the earth's surface which makes the sky appear blue.

During sunrise and sunset the sun's light must pass through a much greater thickness of the atmosphere to reach an observer on the ground. This extra distance causes multiple scatterings of blue light, but relatively little scattering of red light. This is seen as a pronounced red-hued sky in the direction towards the sun: an orange-red sun, which is yellow during daytime.
For the sun high overhead, sunlight goes through a much smaller atmospheric layer, so little scattering takes place. This is why the sky close to the overhead sun in midday appears mostly white, the sun's color.

More info

When radiation is only scattered by one localized scattering center, this is called *single scattering*, but mostly scattering centers are grouped together, and *multiple scattering* occurs. The main difference between the effects of single and multiple scattering is that single scattering can usually be treated as a random phenomenon and multiple scattering is usually more deterministic. Single scattering is often described by probability distributions.
With multiple scattering, the final path of the radiation appears to be a deterministic distribution of intensity as the radiation is spread out. This is exemplified by a light beam passing through thick fog. Multiple scattering is highly analogous to diffusion and the terms *multiple scattering* and *diffusion* are interchangeable in many contexts. Optical elements designed to produce multiple scattering are thus known as *diffusers* or Lambertian radiators or reflectors.

Not all single scattering is random. A well-controlled laser beam can be exactly positioned to scatter off a microscopic particle with a deterministic outcome. Such situations are encountered in radar scattering from e.g. a car or aircraft.

Rayleigh scattering
The inherent scattering that radiation undergoes passing through a pure gas or liquid is due to microscopic density fluctuations as the gas molecules move around.
The degree of Rayleigh scattering varies as a function of particle diameter d, λ, angle, polarization (see Light: polarization), and coherence (see Light). The intensity *I* of light scattered by a single small particle from a beam of unpolarized light of intensity I_0 is given by:

$$I = I_0 \frac{(1+\cos^2\theta)}{2l^2}\left(\frac{2\pi^2}{\lambda}\right)^4\left(\frac{n^2-1}{n^2+2}\right)^2\left(\frac{d}{2}\right)^6 \qquad (1)$$

where I is the distance to the particle, θ is the scattering angle, n is the refractive index (see Light: refraction) of the particle.
The angular distribution of Rayleigh scattering, governed by the $(1+\cos^2\theta)$ term, is symmetric in the plane normal to the incident direction of the light, and so the forward scatter equals the backwards scatter.

Mie scattering
For larger diameters the shape of the scattering center becomes much more significant and the theory only applies well to spheres, spheroids (2 equal axes) and ellipsoids (3 unequal axes).

Both Mie and Rayleigh scattering of EM radiation can undergo a Doppler shift (see Doppler principle) by moving of scattering centers.
At values $d/\lambda > 10$ the laws of geometric optics are mostly sufficient to describe the interaction of light with the particle, and at this point the interaction is not usually described as scattering.
Another special type of EM scattering is coherent backscattering. A description of this phenomenon is beyond the scope of this compendium.

Tyndall effect
This is the effect of light scattering on particles in colloid systems, such as emulsions (see Colloid and Emulsion). The effect distinguishes between these types of colloids. It is proportional to d^6 and hardly on λ.

Brillouin scattering
This occurs when light in a medium (such as water or a crystal) interacts with density variations and changes its path. When a medium is compressed n changes and the light's path necessarily bends. The density variations may be due to acoustic modes, vibration phenomena in crystals (phonons) or temperature gradients.

Light: sources

Principle

The most common processes of emission of light are based on:
- burning;
- heating (thermal);
- line-emission;
- radio activity.

Heating The most common light sources are thermal: a body at a given temperature emits a characteristic spectrum of black body radiation. Examples include sunlight, glowing solid particles in flames and incandescent light bulbs. These bulbs emit only 10% of their energy as visible light, 20% is lost by heat conduction and convection and the remainder as infrared. The peak of the blackbody spectrum is in the infra red (IR) for relatively cool objects like human beings (see Wiens displacement law, Thermography). As the temperature increases, the peak shifts to shorter wavelengths, producing first a red glow, then a white one (metal heating from "red hot" or "white hot), and finally a blue color as the peak moves out of the visible part of the spectrum and into the ultraviolet ".
Line emission Atoms emit and absorb light at characteristic energies. This produces emission lines (a bright line in a continuous spectrum) in the atomic spectrum. Emission can be spontaneous, as in light-emitting diodes, gas discharge lamps (such as Ne-lamps, Hg-vapor lamps, etc.), and flames (light from the hot gas itself. For example, Na in a gas flame emits characteristic yellow light). Emission can also be stimulated, as in a Laser or a maser (applied in the kitchen microwave).
Certain substances produce light by Fluorescence (fluorescent lights). Some substances emit light slowly after excitation by more energetic radiation, known as Phosphorescence. Phosphorescent materials can also be excited by bombarding them with subatomic particles, a process known as cathodoluminescence (used in cathode ray tube televisions). Certain chemicals produce visible radiation by Chemoluminescence and bioluminescence.
Certain other mechanisms can produce light: electroluminescence (a light emitting material in response to an electric current passed through it, or to a strong electric field), sonoluminescence (emission of short bursts of light from imploding bubbles in a liquid when excited by sound. An example is the snapping of a specialized shrimp claw), visible cyclotron radiation (acceleration of a free charged particle, such as an electron), scintillation (a flash of light produced in certain materials when they absorb ionizing radiation), radioactive decay, etc.

Table 1 lists overall luminous efficacy and efficiency for various light sources:

Category	Type	Overall luminous efficacy (lm/W)	Overall luminous efficiency
Combustion	candle	0.3	0.04%
Incandescent	40 W tungsten	12.6	1.9%
	quartz halogen	24	3.5%
	high-temperature	35	5.1%
Fluorescent	5-24 W compact	45-60	6.6%-8.8%
	28 W tube (T5)	104	15.2%
Light-emitting diode	white LED	26-70	3.8%-10.2%
Arc lamp	Xe-arc lamp	30-150	4.4%-22%
	Hg-Xe arc lamp	50-55	7.3%-8.0%
Gas discharge	high pressure Na lamp	150	22%
	low pressure Na lamp	183 - 200	27%

Radio active processes Particles produced by radioactive decay, moving through a medium faster than the speed of light in that medium can produce visible *Cherenkov radiation* (characteristic "blue glow" of nuclear reactors). Other examples are radioluminescent paint and self-powered tritium lighting, formerly and nowadays used on watch and clock dials.

Application

Medical applications are multiple in apparatus, instruments, etc. In Thermography the human body can be considered as an IR light source.
In daily live and industry, applications are innumerous.
For two reasons Na-lamps for road lighting are optimal. Their efficacy is high and since they emit monochromatic light, the eye does not suffer from chromatic aberration (see Light: the ideal and non-ideal lens).

More info

See Wikipedia and the internet.

Light: units of measure

Principle

Photometry is measuring the *perceived* brightness to the human eye. It is distinct from radiometry, see below. The human eye is not equally sensitive to all wavelengths. Photometry attempts to account for this by weighting the measured power at each wavelength with a factor that represents how sensitive the eye is at that wavelength. The standardized model of the eye's response to light as a function of wavelength is given by the photopic luminosity function (see Vision of color). Photometry is based on the eye's photopic response, and so photometric measurements will not accurately indicate the perceived brightness of sources in scotopic conditions.

Many different units are used. The adjective "bright" can refer to a lamp which delivers a high luminous flux (measured in lumens, lm), or to a lamp which concentrates the luminous flux it has into a very narrow beam (in candelas, cd). Because of the ways in which light can propagate through three-dimensional space, spread out, become concentrated, reflects off shiny or matte surfaces, and because light consists of many different wavelengths, the number of fundamentally different kinds of light measurement is large, and so are the numbers of quantities and units that represent them.

Table 1 gives the quantities and units are used to measure the quantity or “brightness" of light.

Table 1 SI photometry units

Quantity	Symbol	Abbr SI unit.	Notes
Luminous energy	Q_v	lm·s	units are sometimes called Talbots
Luminous flux	F	lm	also called *luminous power*
Luminous intensity	I_v	cd	an SI base unit, 1/683 W at 555 nm
Luminance	L_v	cd/m^2	units are sometimes called nits
Illuminance	E_v	lx	Used for light incident on a surface
Luminous emittance	M_v	lx	Used for light emitted from a surface
Luminous efficacy		lm/W	ratio of luminous flux to radiant flux; theoretical maximum is 683

lx is lux. Cd is candela, 1 lx = 1 cd.sr/m^2. 1 lm = 1 cd.sr (sr is steradian, a solid angle of about 66^o, see Radian and steradian).

Radiometry is measuring the power emitted by a source of electromagnetic radiation (see Light). Radiometric quantities use unweighted *absolute power*, so radiant flux is in watts (versus luminous flux in lumens).
Photometric versus radiometric quantities Every quantity in the photometric system has an analogous quantity in the radiometric system. Some examples of parallel quantities include:

- Luminance (photometric) and radiance (radiometric)
- Luminous flux (photometric) and radiant flux (radiometric)
- Luminous intensity (photometric) and radiant intensity (radiometric)

A comparison of the watt and the lumen illustrates the distinction between radiometric and photometric units.

Table 2 SI radiometry units

Quantity	Symbol	Abbr. unit	Notes
Radiant energy	Q	J	energy
Radiant flux	Φ	W	radiant energy per unit time, also called *radiant power*
Radiant intensity	I	$W \cdot sr^{-1}$	power per unit solid angle
Radiance	L	$W \cdot sr^{-1} \cdot m^{-2}$	power per unit solid angle per unit *projected* source area. Sometimes confusingly called "intensity".
Irradiance	E	$W \cdot m^{-2}$	power incident on a surface. Sometimes confusingly called "intensity".
Radiant emittance or exitance	M	$W \cdot m^{-2}$	power emitted from a surface. Sometimes confusingly called "intensity".
Spectral radiance	L_λ or L_v	$W \cdot sr^{-1} \cdot m^{-3}$ *or* $W \cdot sr^{-1} \cdot m^{-2} \cdot Hz^{-1}$	commonly measured in $W \cdot sr^{-1} \cdot m^{-2} \cdot nm^{-1}$
Spectral irradiance	E_λ or E_v	$W \cdot m^{-3}$ *or* $W \cdot m^{-2} \cdot Hz^{-1}$	commonly measured in $W \cdot m^{-2} \cdot nm^{-1}$

Light bulbs are distinguished in terms of power (in W) but this power is not a measure of the amount of light output, but the amount of electricity consumed by the bulb. Because incandescent bulbs all have fairly similar characteristics, power is a guide to light output. However, there are many lighting devices, such as fluorescent lamps, LEDs, etc., each with its own efficacy. Therefore, power is also used for the *amount of emitted light*. See for more details Light: sources.
The lumen is the photometric unit of light output and defined as amount of light given into 1 sr (steradian, see Radian and steradian) by a point source of 1 Cd (the candela). The candela, a base SI unit, is defined as the I_v of a monochromatic source with frequency 540 10^{12}Hz, and I = 1/683 W/sr. (540 THz corresponds to about 555 nm, the wavelength to which the human eye is most sensitive. The number 1/683 was chosen to make the candela about equal to the standard candle, the unit which it superseded).
Combining these definitions, we see that 1/683 W of 555 nm green light provides 1 lm. The relation between watts and lumens is not just a simple scaling factor. The definition tells us that 1 W of pure green 555 nm light is "worth" 683 lm. It does not say anything about other wavelengths. Because lumens are photometric units, their relationship to watts depends on the wavelength according to how visible the wavelength is. Infrared (IR) and ultraviolet (UV) radiation, for example, are invisible and do not count. According to photopic spectral luminous function, 700 nm red light is only about 4% as efficient as 555 nm green light. Thus, 1 W of 700 nm red light is "worth" only 27 lm.
Table 2 gives the quantities and units are used to measure the output of electromagnetic sources.

Application

Photometric measurement is based on photodetectors. The lighting industry uses more complicated forms of measurement, e.g. obtained with spherical photometers and rotating mirror photometers.

More Info

Non-SI photometry units for luminance are the Footlambert, Millilambert and Stilb and for illuminance the foot-candle and the Phot.

Microscopy

Principle

Microscopy is any technique for producing visible images of structures too small to be seen by the human eye, using a microscopy or other magnification tool. There are 4 main types of microscopy;

- optical,
- X-ray,
- Electron,
- scanning probe microscopy.

Optical and electron microscopy involves the diffraction, reflection, or refraction of radiation incident upon the subject of study, and the subsequent collection of this scattered radiation in order to build up an image. This process may be carried out by wide field irradiation of the sample (for example standard light microscopy and transmission electron microscopy) or by scanning of a fine beam over the sample (for example confocal microscopy and scanning electron microscopy). Scanning probe microscopy involves the interaction of a scanning probe with the surface or object of interest.

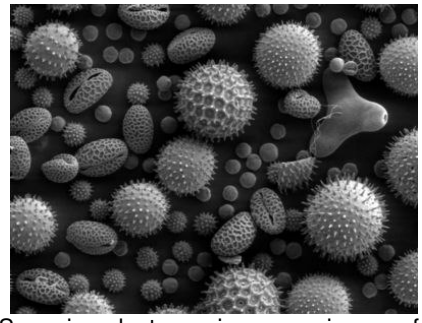


Fig. 1 Scanning electron microscope image of pollen.

Application

The development of microscopy revolutionized biology and medicine for basic research and applications and remains an essential tool in sciences as (bio)chemistry and as with many others.

More Info

Optical microscopy
The principal compounds to make images with light are lenses (see Light: the ideal and non-ideal lens). See for a further description Optical microscopy.

X-ray microscopy
The principal compounds to make images with reflected radiation are mirrors and interference plates. See further X-ray microscopy.

Electron Microscopy
For light microscopy the wavelength of the light limits the resolution to around 200 nm. In order to gain higher resolution, the use of an electron beam with a far smaller wavelength is used in electron microscopes. The main compounds to make images are coils to deflect the X-rays.
Transmission electron microscopy (TEM) is similar to the compound optical microscopy, by sending an electron beam through a very thin slice of the specimen. The resolution is 0.05 nanometer (2005).
Scanning electron microscopy (SEM) visualizes details on the surfaces of cells and particles and gives a very nice 3D view. It gives results much like the stereo light microscope with a resolution in the lower range of that of TEM. See further Electron microscopy.

Scanning probe microscopy
The basic method of the various types (atomic and ultrasonic force, and scanning tunneling) is the use of a solid-state probe tip in vicinity (near field) of an almost flat object. The probe scans in some way the surface of the object. See further Scanning probe microscopy.

Optical coherence tomography (OCT)

Principle

Optical Coherence Tomography, or OCT, is a technique for obtaining sub-surface images of translucent or opaque materials at a resolution equivalent to a low-power microscope. It is effectively 'optical ultrasound', imaging reflections from within tissue to provide cross-sectional images.
It provides images of tissue morphology at a far higher resolution (better than 10 µm) than is possible with other imaging modalities such as MRI or ultrasound.

The key benefits of OCT to the user are:

- live sub-surface images at near-microscopic resolution;
- enables instant, direct imaging of tissue morphology;
- no preparation of the sample or subject is required;
- no ionizing radiation – allowing safe use in office, laboratory or clinic.

OCT can deliver much higher resolution because it is based on light and optics, rather than sound or radio frequency radiation. It works as follows: an optical beam is projected into the subject, and light is reflected from the layers and sub-surface artefacts as the beam penetrates. Most of this light is scattered on its way back to the surface. Scattered light has lost its original direction and therefore cannot be used for imaging: this is why scattering material such as tissue appears opaque to the human eye. However, a very small proportion of the light is not scattered but reflected (as with a mirror). This light is detected and used in an OCT microscope for optical imaging.
The reflected light has the unique property that it is coherent, so that it can be detected in an OCT instrument using a device called an optical interferometer (see Interferometry). Essentially, the interferometer is used to separate the useless, incoherent, scattered light from the valuable, coherent, reflected light. It also provides the depth and intensity information from the light reflected from a sub-surface feature, enabling an image of it to be built up, rather like an echo-sounder. In this respect, OCT is more similar to ultrasound than to MRI or to PET.
The technique is limited to imaging some mm below the surface in tissue, because at greater depths the proportion of light that escapes without scattering is vanishingly small. No special preparation of the specimen is required, and images can be obtained 'non-contact' or through a transparent window or membrane. It is also important to note that the laser output from the instruments is low – eye-safe near-infra-red light is used – and no damage to the sample is therefore likely.

Basic working principle
Optical coherence tomography (OCT) is an interferometric, (the technique of superimposing or interfering two or more waves, to detect differences between them, see also Huygens' principle) non-invasive optical tomographic imaging technique offering mm-penetration (mostly 2-3 mm, sometimes more) with µm-scale axial and lateral resolution and cross-sectional imaging capabilities.
OCT works through low-coherence Interferometry. In conventional interferometry with long coherence length (laser interferometry), interference of light occurs over a distance of meters. In OCT, this interference is shortened to a distance of µms, thanks to the use of broadband light sources (broad range of wavelengths, e.g. white light). Light with broad bandwidths can be generated by using super luminescent diodes (super bright LEDs) or lasers with extremely short pulses (femtosecond lasers, since this pulse comprises many frequencies as follows from a Fourier transform).
Light in an OCT system is broken into two pathways (arms), a sample arm (containing the item of interest) and a reference arm (usually a mirror). The combination of reflected light from the sample arm and reference light from the reference arm gives rise to an interference pattern, but only if light from both arms have travelled the "same" optical distance ("same" meaning a difference of less than a coherence length). By a scanning motion of the mirror in the reference arm, a reflectivity profile of the sample can be obtained (this is time domain OCT). Areas of the sample that reflect back a lot of light will create greater interference than areas that don't. Any light that is outside the short coherence length will not interfere. This reflectivity profile, called an A-scan contains information about the spatial distances and location of structures within the item of interest (e.g. determination of the eye length for calculation of intraocular lens power). A cross-sectional tomograph (B-scan)) may be achieved by laterally combining a series of these axial depth scans (A-scan), e.g. to produce a 2-D, cross-sectional view of the vitreous and retina. En face imaging (C-scan) at an acquired depth is possible depending on the imaging engine used.

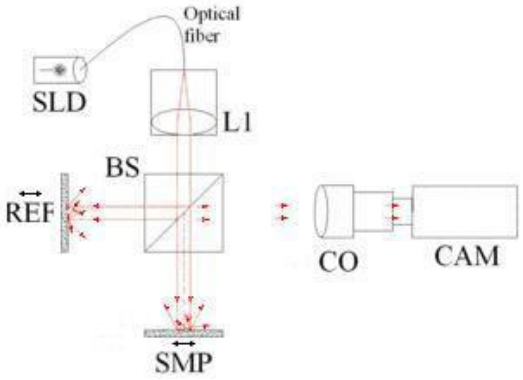


Fig. 1 Full-field OCT optical design. Components include: super-luminescent diode (SLD), convex lens (L1), 50/50 Beam splitter (BS), camera objective (CO), CMOS-DSP camera (CAM), reference (REF) and sample (SMP).

The basic optical setup typically consists of an interferometer (Fig. 1, typically Michelson type) with a low coherence, broad bandwidth (white) light source. The horizontal black arrow in Fig. 1 off the sample represents the scanning movements in the sample plane (x,y). The scattered beam of the sample comes from a range of depths (z axis). The black arrow off the reference indicates the scanning in the z direction. (The scattered beam from the reference has no depth info; it comes from the surface). The scattered beam of each sample depth interferes with the scattered beam of the reference when the both path lengths are identical. The camera functions as a 2D detector array, and by performing a sequence of depth (z) scans a non-invasive 3D imaging device is achieved.
The classical type of OCT is Time Domain OCT (see **More Info**).

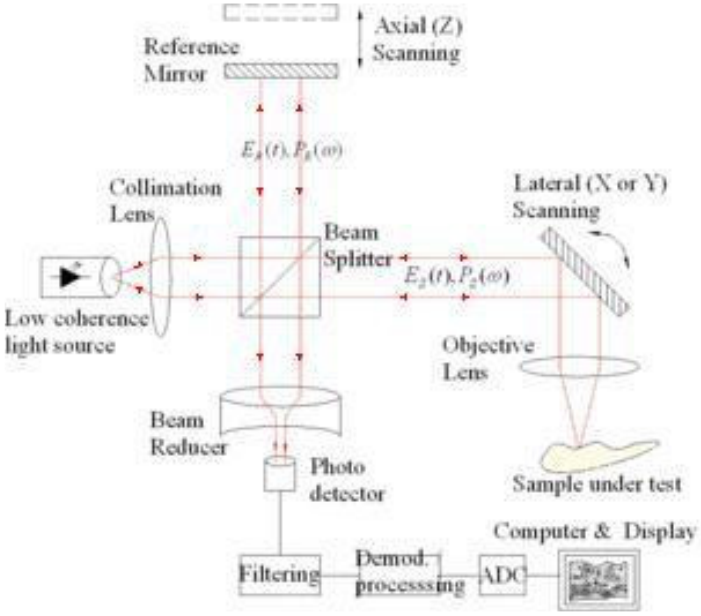


Fig. 2 Typical optical design of single point OCT. Scanning the light beam on the sample enables non-invasive cross-sectional imaging up to 3 mm in depth with micrometer resolution.

Application

OCT is widely applied, especially in *ophthalmology* and other tissue imaging requiring µm resolution and mm penetration depth.
OCT has important advantages over e.g. medical ultrasonography (see Echography), MRI, (see MRI: general)) and confocal microscopy (see Optical microscopy: confocal laser scanning). These techniques are not suited to morphological tissue imaging at µm-scale. The former two having poor resolution and the latter are lacking mm-penetration depth.
A new development is high-definition OCT (HD-OCT) with a better resolution along the z-axis as the important improvement. Notice that OCT resolution is dependent on coherence length of the beam and on NA (numerical aperture), whereas in light microscopy only NA determines the resolution. Another

development is functional OCT: subtracting to OCT images, with the second image obtained by strongly stimulating the retina with light. At present OCT can also be used (via fiber optics) to visualize the coronary vessel wall.

More info

Time domain OCT
The path length of the reference arm is translated longitudinally in time. A property of low coherence interferometry is that interference, i.e. the series of dark and bright fringes, is only achieved when the path difference lies within the coherence length (see Interferometry) of the light source. This interference is called autocorrelation in a symmetric interferometer (both arms have the same reflectivity), or crosscorrelation in all other cases. The envelope of this modulation changes as path length difference is varied, where the peak of the envelope corresponds to path length matching. The interference of two partially coherent light beams can be expressed in terms of the source intensity, I_s, as:

$$I = k_1 I_s + k_2 I_s + 2(k_1 I_s \cdot k_2 I_s)^{0.5} \cdot \mathrm{Re}\,[\gamma(\tau)], \quad (1)$$

where $k_1 + k_2 < 1$ represents the interferometer beam splitting ratio and $\mathrm{Re}[\gamma(\tau)]$ the real part of the complex degree of coherence, i.e. the interference envelope and carrier dependent on reference arm scan or time delay τ. Due to the coherence gating effect of OCT the complex degree of coherence is represented as a Gaussian function expressed as:

$$\gamma(\tau) = e^{-\left(\frac{\pi \Delta v \tau}{2\sqrt{\ln 2}}\right)^2} e^{-j 2\pi v_0 \tau}, \quad (2)$$

where Δv represents the spectral width of the source in the optical frequency domain, and v_0 is the centre optical frequency of the source. In equation (2), the Gaussian envelope is amplitude (1st term) modulated by an optical carrier (2nd term). The peak of this envelope represents the location of sample under test microstructure, with an amplitude dependent on the reflectivity of the surface.
In OCT, the Doppler-shifted optical carrier has a frequency expressed as:

$$f_{Doppler} = 2 f_0 v_s / c, \quad (3)$$

where f_0 is the central optical frequency of the source, v_s is the scanning velocity of the path length variation, and c is the speed of light. The axial and lateral resolutions of OCT are decoupled from one another; the former being an equivalent to the coherence length of the light source and the latter being a function of the optics. The coherence length Lof a source and hence the axial resolution of OCT is defined as:

$$L_c = (2\ln 2/\pi)(\lambda_0^2/\Delta\lambda). \quad (4)$$

Frequency Domain OCT (FD-OCT)
In frequency domain OCT the broadband interference is acquired with spectrally separated detectors (either by encoding the optical frequency in time with a spectrally scanning source or with a dispersive detector, like a grating and a linear detector array). Since the Fourier transform of the autocovariance function is the power spectral density the depth scan can be immediately calculated by a Fourier-transform from the acquired spectra, without movement of the reference arm. This feature improves imaging speed dramatically, while the reduced losses during a single scan improve the signal to noise proportional to the number of detection elements. The parallel detection at multiple wavelength ranges limits the scanning range, while the full spectral bandwidth sets the axial resolution.

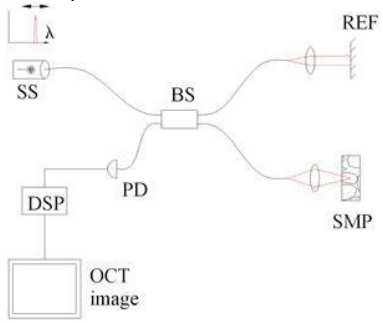


Fig. 3 Spectral discrimination by Time Encoded Frequency Domain OCT. Components include: swept source or tunable laser (SS), beam splitter (BS), reference mirror (REF), sample (SMP), photodetector (PD), digital signal processing (DSP).

Spatially Encoded Frequency Domain OCT (also Fourier Domain OCT)
SEFD-OCT extracts spectral information by distributing different optical frequencies onto a detector stripe (line-array CCD or CMOS) via a dispersive element (see Fig. 4). Thereby the information of the full depth scan can be acquired within a single exposure. However, the large signal to noise advantage of FD-OCT is reduced due the lower dynamic range of stripe detectors in respect to single photosensitive diodes, resulting in an SNR (signal to noise ratio) advantage of ~10 dB at much higher speeds. The drawbacks of this technology are found in a strong fall-off of the SNR, which is proportional to the distance from the zero delay and a sin(z)/z type reduction of the depth dependent sensitivity because of limited detection line width. (One pixel detects a quasi-rectangular portion of an optical frequency range instead of a single frequency, the Fourier-transform leads to the sin(z)/z behavior). Additionally the dispersive elements in the spectroscopic detector usually do not distribute the light equally spaced in frequency on the detector, but mostly have an inverse dependence. Therefore the signal has to be re-sampled before processing, which can not take care of the difference in local (pixel wise) bandwidth, which results in further reduction of the signal quality.

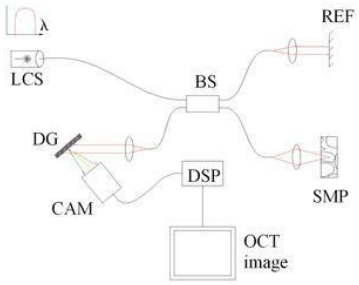


Fig. 4 Spectral discrimination by SEFD-OCT. Components include: low coherence source (LCS), beam splitter (BS) (Light: beam splitter), reference mirror (REF), sample (SMP), diffraction grating (DG) and full-field detector (CAM) act as a spectrometer, and digital signal processing (DSP).

Time Encoded Frequency Domain OCT (also called swept source OCT)
TEFD-OCT tries to combine some of the advantages of standard TD and SEFD-OCT. Here the spectral components are not encoded by spatial separation, but they are encoded in time. The spectrum either filtered or generated in single successive frequency steps and reconstructed before Fourier-transformation. By accommodation of a frequency scanning light source (i.e. frequency scanning laser) the optical setup (see Fig. 5) becomes simpler than SEFD, but the problem of scanning is essentially translated from the TD-OCT reference-arm into the TEFD-OCT light source. Here the advantage lies in the proven high SNR detection technology, while swept laser sources achieve very small instantaneous bandwidths (is line width) at very high frequencies (20-200 kHz). Drawbacks are the nonlinearities in the wavelength, especially at high scanning frequencies, the broadening of the line width at high frequencies and a high sensitivity to movements of the scanning geometry or the sample (below the range of nanometers within successive frequency steps).

Full-field OCT
Focusing the light beam to a point on the surface of the sample under test, and recombining the reflected light with the reference will yield an interferogram with sample information corresponding to a single A-scan (Z axis only). Scanning of the sample can be accomplished by either scanning the light on the sample, or by moving the sample under test. A linear scan will yield a 2D data set corresponding to a cross-sectional image (X-Z axes scan), whereas an area scan achieves a 3D data set corresponding to a volumetric image (X-Y-Z axes scan).

Single point (confocal) OCT
Systems based on single point, or flying-spot time domain OCT, must scan the sample in two lateral dimensions and reconstruct a three-dimensional image using depth information obtained by coherence-gating through an axially scanning reference arm (Fig. 2). Two-dimensional lateral scanning has been electromechanically implemented by moving the sample using a translation stage, and using a microelectro-mechanical system scanner.

Parallel OCT
Parallel OCT using a CCD camera has been used in which the sample is full-field illuminated and en face imaged with the CCD, hence eliminating the electromechanical lateral scan. By stepping the reference mirror and recording successive *en face* images, a 3D representation can be reconstructed.

Optical microscopy

Principle

The optical microscope, often referred to as a "light microscope", is a type of microscope which uses visible, UV (for fluorescence) or sometimes IR light and a system of lenses (see Light: the ideal and non-ideal lens) to magnify images of small samples. Optical microscopes are the oldest and simplest of the microscopes.
There are two basic configurations of the optical microscope in use, the simple (one lens) and compound (many lenses) one.

Simple optical microscope
The simplest and oldest microscope is a microscope that uses only one convex lens for magnification,. Today only the hand-lens and the loupe belong to this type.
Magnification
Suppose the minimal viewing distance of a subject is D (by convention it is standardized at 250 mm), then for the angle of seeing (α) with the unaided eye of a subject with height h it holds that tanα = h/D. With a focus length of the loupe f and viewing with an unaccommodated eye, for the angle of vision α' holds that tanα' = h/f. Then the magnification M is:

$$M = \tan\alpha'/\tan\alpha = D/f. \qquad (1)$$

Notice that by convention this magnification is called the angular magnification, whereas it is calculated as the *linear* magnification. Formally, the angular magnification is α'/α'. For small angles both definitions give the same number. With D=250 mm and taking a botanical loupe with f = 25 mm then M =10. Stamp loupes reach only about M is 2 to 5.

Compound optical microscope
Fig. 1 shows a classical compound *microscope* and Fig. 2 its most basic optics. It comprises a single glass lens of short focal length for the objective and another single glass lens for the eyepiece or ocular lens. Modern microscopes of this kind are more complicated, with multiple lens components in both objective and eyepiece assemblies. These multi-component lenses are designed to reduce aberrations, particularly chromatic aberration (by using a doublet lens, i.e. 2 lenses kitted together) and spherical aberration (see Light: the ideal and non-ideal lens).
In modern microscopes the mirror is replaced by a lamp unit providing stable, controllable illumination.

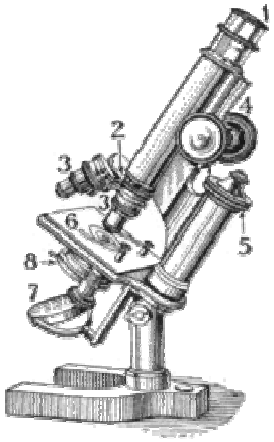


Fig. 1. Basic microscope elements. 1. ocular lens, 2. objective turret, 3. objective lenses, 4. coarse adjustment knob, 5. fine adjustment knob to image a specific plane of the object, 6. object holder or stage, 7. mirror to reflect the light of some light source, 8. diaphragm to select the center of the light beam and a condenser.

Working principle
The optical components of a modern microscope are very complicated and for a microscope to work well, the whole optical path has to be very accurately set up and controlled. Despite this, the basic optics is simple.
The objective lens is, at its simplest, a very high powered magnifying glass i.e. a lens with a very short focal length (f_{obj}). This is brought very close to the specimen (object) being examined so that the light from the object comes to a focus about 160 mm inside the microscope tube. This creates an enlarged image of the subject. This image is inverted and can be seen by removing the eyepiece and placing a

piece of tracing paper over the end of the tube. By careful focusing a rather dim image of the object, much enlarged can be seen. It is this *real image* that is viewed by the eyepiece lens that provides further enlargement.

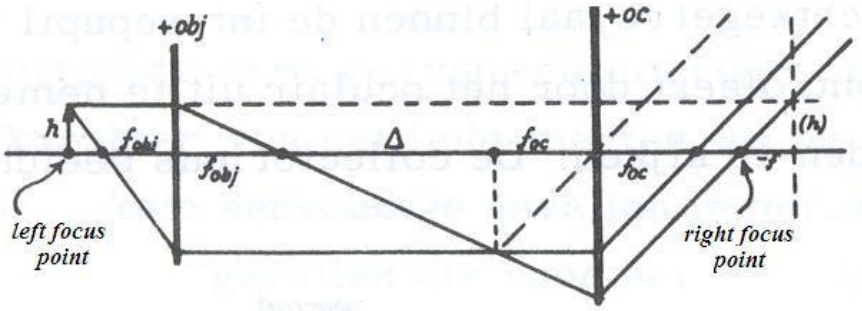


Fig. 2 Basic design of a compound microscope.

The magnification of the objective is $M_{obj} = \Delta/f_{obj}$. The focal length f of the total system is:

$$f = -f_{obj}f_{oc}/\Delta. \quad (2a)$$

And consequently, the magnification with respect to the unaided eye (not-accommodated) is:

$$M = 250/f, \quad (2b)$$

with all sizes in mm. M can reach theoretically 1000x, but with very limited depth of field. In practice M is less, see **More Info**.

Basic components of the optical microscope

Ocular

Generally, the ocular is a compound lens in a cylinder, which is made of two (or more) lenses, one near the front and one near the back of the eyepiece tube. In many designs, the virtual image comes to a focus between the two lenses of the eyepiece, the first lens bringing the real image to a focus and the second lens enabling the eye to focus on its virtual image. Headaches and tired eyes after using a microscope are signs that the eye is being forced to focus at a close distance rather than at infinity. Typical values for oculars include X5, X10 and X20. With apochromatic objectives (triplet objectives; three or more lenses kitted together) chromatic aberration is practically perfectly corrected. With these objectives certain combinations with the ocular give optimal performance.

Objective

The objective lens is a compound lens (x4, x5, x10, x20, x40, x80 and x100) in a cylinder mounted with some others in a circular, rotatable nose piece. The weaker objectives have two doublets (2 lenses kitted together). The 4th objective in the nose piece is mostly an oil immersion lens (100×). It often comprises two single lenses and 2 or more triplet lenses. When used, then oil is filing the space between cover slip and the lens. The refractive indexes of the oil and of the cover slip are closely matched, minimizing refraction.

Object holder

The object holder or stage is a platform below the objective which supports the object being viewed. In the centre is a circular hole through which light shines to illuminate the object. The stage usually has arms to hold slides (mostly 25x75 mm).

Light source

The controllable source illuminates the object, generally from below (from above it is an inverted microscope). The emitted beam of the source is restricted by a diaphragm. The simplest, daylight is directed via a mirror.

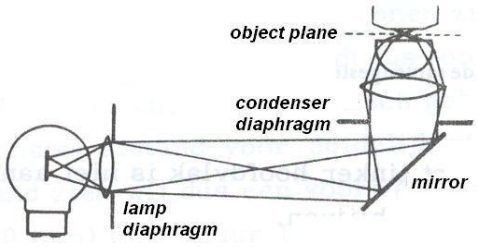


Fig. 3 Abbe condenser.

Condenser

Generally the diaphragm of the light source is focused through an optical device, the condenser with filters available to manage the quality and intensity of the light. The condenser diaphragm enables that the beam is well within the aperture of the first condenser lens (a double-convex lens). The 2nd lens is a convex-planar lens (very strong) with the planar side just below the stage. Such a condenser is called an Abbe condenser (Fig. 3).

Modern microscopes may have many more features, including transmission/reflection illumination, filters, apparatus for phase and differential interference contrast microscopy, digital cameras, etc.

Bright field microscopy is the basic technique and uses transmitted white light, with illuminated from below and observed from above. This generally implies low contrast, low apparent resolution due to the blur of out of focus material. Advantages are simplicity and the minimal sample preparation.
An *inverted microscope* is a microscope with its light source and condenser on the top above the stage pointing down, and the objective and turret are below the stage pointing up. Inverted microscopes are useful for observing living cells or organism at the bottom of e.g. a tissue culture flask.

More Info

Limitations of light microscopes
Ocular There are two major ocular types, the Huygens and the Ramsden ocular. The former has two lenses of the same material at a distance of half the sum of both focal lengths. These results in a chromatically corrected ocular, but an ocular micrometer cannot be applied. This can be done with the later one, which has lenses of about the same power. The object focal plane is slightly before the first lens. Since the intermediate image is real, in contrasts to the Huygens' ocular, at that image position the micrometer is inserted.
Resolution Microscopes are limited in their ability to resolve fine details by the properties of light and the refractive materials of the lenses. Diffraction of light is the principle problem. Its depends on A_N, the numerical aperture of the optical system and λ and it sets a definite limit (*d*) to the optical resolution. Assuming that optical aberrations optical aberrations (see Light: the ideal and non-ideal lens) are negligible, the resolution or diffraction limit d for two non-self-luminous particle in the object illuminated by an extraneous, non coherent source is:

$$d = 1.22\lambda/(2A_N) = 1.22\lambda/(2n\sin\tfrac{1}{2}u), \quad (3)$$

where n is the refractive index of the medium (air or immersion oil) *u* is the aperture of the objective, the angle seen from the object. u is defined by $\tan\tfrac{1}{2}u$ = a/object distance, where a the diameter of the beam of light measured at the objective, or (if the beam is filling the lens) the diameter of the lens.

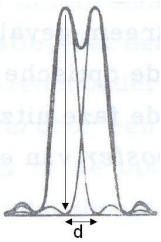


Fig. 4 Point resolution d with the overlapping images (point spread functions) of two points .

The equation directly follows from the single narrow split optics with the factor 1.22 accounting for a pinhole instead of a split. (see also *Light: diffraction).*
The resolution for a grid with *coherent light* is:

$$d = \lambda/n\sin\tfrac{1}{2}u. \quad (4)$$

This is (ignoring the factor 1.22) two times less than the point-resolution. This is caused by the fact that for resolving the grid, besides the principle maximum, also the first order maximum is needed. For the point resolution only the principle diffraction maximum is needed. Graphically, the resolution is found when the maximum of the one point coincides with the minimum between the principal and the 1st order maximum of the image of the other point (Fig. 4). This yields the factor 2.
The lower resolution of the grid can be compensated by using an oblique entering light beam (Fig. 5).

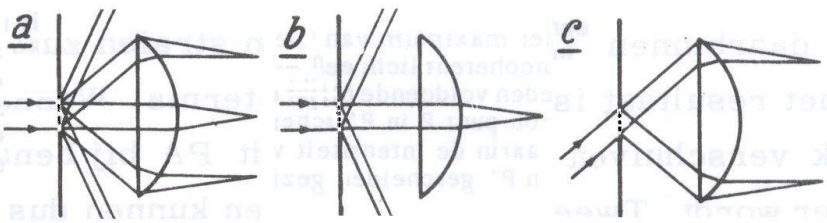


Fig. 5 a. Coarse grid which is visible since the 2 beams of 1st order maxima are captured by the lens. b. Invisible fine grid with the 1st order beams outside the lens. C. Just visible grid with 'one' (half circle) 1st order beam inside the lens.

The highest practical A_N in air is about 0.8 (theoretically 0.95), and with oil up to 1.3. With λ = 440 nm, deep violet that is just visible, air as medium, the highest point resolution d is 330 nm and with oil immersion 210 nm. For green light (550 nm) these numbers are about 25% higher. Other designs can offer an improved resolution when observing self-luminous particles (coherent), which is not covered by Abbe's diffraction limit. See http://micro.magnet.fsu.edu/optics/timeline/people/abbe.html for Abbe's work in light microscopy

Magnification As long as humans use their eyes for direct microscopical observation, it makes no sense to design better microscopes than the resolution of the human eye, which is considered to be an angle of 1'. This yields a maximal magnification that is sufficient. Supposing that we like to see 2 points separated at angle α = 4' (= 1/860 rad) then the so called useful magnification M_{useful} , being the resolution of the naked eye divided by that of the microscope is:

$$M_{useful} = 250\alpha/d = 250\alpha/(1.22\lambda/2A_N) = (250/860)x2A_N/1.22\lambda = (1000/860)A_N/2.44\lambda, \quad (5)$$

where A_N and λ the wavelength in mm. With λ = 600 nm (yellow light, the mean of white light):

$$M_{useful} \approx 820A_N. \quad (6)$$

Going to the limit of 1', $200A_N$ is found.

Types of optical microscopes and techniques

There exist many specialized types of optical microscopy.

- Stereo microscopy (2 objectives and 2 oculars, low M).
- Oblique illumination microscopy (highlights otherwise invisible features).
- Dark field microscopy (contrast improving of unstained, transparent objects).
- Differential interference contrast microscopy (differences in optical density will show up as differences in relief).

These four are described in Optical microscopy: specific techniques.

- Deconvolution microscopy (contrast improvements by mathematical calculations).
- Sub-diffraction microscopy techniques (by using near-field optics instead of the common far field approach), such as NSOM (Near-field scanning optical microscopy with a light source and/or a detector at nm-scale), ANSOM (apertureless NSOM with fluorescence), fitting the PSF (point spread function), PALM & STORM (Photo-activated localization microscopy and Stochastic optical reconstruction microscopy), structured illumination (breaking the diffraction limit and post-exposure analysis).

These types are described in Optical microscopy: super-resolution techniques.

- Other important types are phase contrast optical microscopy, see Optical microscopy: phase contrast and Fluorescence microscopy (see Optical microscopy: fluorescence), Confocal laser scanning microscopy (improved image by less disturbance of non-focused light, see Optical microscopy: confocal laser scanning).

Drawing with the optical microscope

Modern microscopes provide photo-micrography and image recording electronically. However, then only one thin plane is ever in good focus. Therefore, the experienced microscopist will, in many cases, still prefer a hand drawn image rather than a photograph. With knowledge of the subject one can accurately convert a 3D image into a precise 2D drawing. This holds for instance for dendritic and axonal trees. The creation of careful and accurate micrographs requires a microscopical technique using a monocular eyepiece. It is essential that both eyes are open and that the eye that is not observing down the microscope is instead concentrated on a sheet of paper on the bench besides the microscope. With practice, and without moving the head or eyes, it is possible to accurately record the observed details by tracing round the observed shapes by simultaneously "seeing" the pencil point in the microscopical image. Practicing this technique also establishes good general microscopical technique. It is always less tiring to observe with the microscope focused so that the image is seen at infinity and with both eyes open at all times.

Photomicrography

For simple purposes a digital camera can be placed directly to the ocular, and the image can captured with a steady hand. However, the edges of an image are darker than the centre (vignettation). This image may be blurred due to camera shake.

A more professional, but also more costly solution is to use a tube adapter. With this method, the ocular is removed and an adapter is fitted into the microscope tube. The adapter acts as a mechanical and

optical interface between microscope and a digital camera (CCD camera). This makes it possible to avoid motion blurs and vignettation effects, leading to a higher quality of the image.
The best solution is to attach the digital camera with an optical-mechanical adapter via a 2nd tube, the phototube. A firm mechanical connection is particularly important, because even the small movements (vibrations) of the camera strongly reduce the image quality. The two oculars continue to be used for the visual observation of the object.

Optical microscopy: confocal microscopy

Principle

Confocal microscopy is an optical imaging technique used to increase micrograph contrast and/or to reconstruct 3D images by using a spatial pinhole to eliminate out-of-focus light or flare in specimens that are thicker than the focal plane. A recent modification of confocal microscopy is confocal laser scanning microscopy (see: Optical microscopy: confocal laser scanning microscopy).

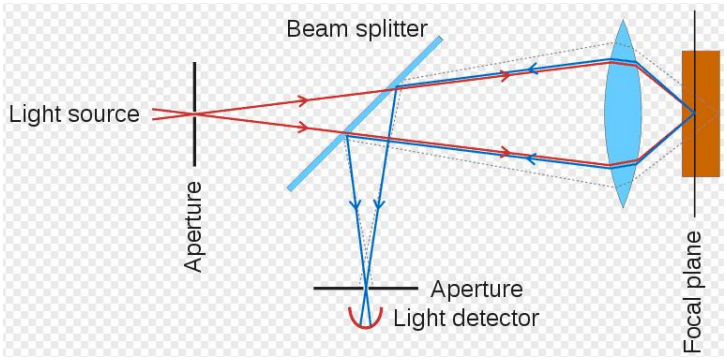


Fig. 1 Basic design of confocal microscope. For beam splitter see Light: beam splitter.

In a conventional (i.e., wide-field) fluorescence microscope (see Optical microscopy: fluorescence), the entire specimen is flooded in light from a light source. Due to the conservation of light intensity transportation, all parts of the specimen throughout the optical path will be excited and the fluorescence detected by a photodetector or a camera. In contrast, a confocal microscope uses point illumination and a pinhole in an optically conjugate plane in front of the detector to eliminate out-of-focus information. Only the light within the focal plane can be detected, so the image quality is much better than that of wide-field images. As only one point is illuminated at a time in confocal microscopy, 2D or 3D imaging requires scanning the specimen with a rectangular pattern of parallel lines. The thickness of the focal plane is defined mostly by the square of the numerical aperture NA of the objective lens ($NA = n \sin\theta$, n is refractive index, θ is the half-angle of the maximum cone of light entering or exiting the lens), and also by the optical properties of the specimen, especially n. These microscopes also are able to see into the image by taking images at different depths.

Application

This technique has gained popularity in the scientific and industrial communities. Typical applications include life sciences and semiconductor inspection.
Confocal microscopy provides the capacity for direct, noninvasive, serial optical sectioning of intact, thick, living specimens with a minimum of sample preparation as well as a marginal improvement in lateral resolution.

More Info

Three types of confocal microscopes are commercially available:
- Confocal laser scanning microscopes,
- Spinning-disk (Nipkow disk) confocal microscopes and
- Programmable Array Microscopes (PAM).

Confocal laser scanning microscopy yields better image quality than Nipkow and PAM, but the imaging frame rate was very slow (less than 3 frames/second) until recently; spinning-disk confocal microscopes

can achieve video rate imaging—a desirable feature for dynamic observations such as live cell imaging. Confocal laser scanning microscopy has now been improved to provide better than video rate (60 frames/second) imaging by using MEMS (Microelectromechanical systems) based scanning mirrors.

Optical microscopy: confocal laser scanning microscopy

Principle
Confocal Laser scanning Microscopy (CLSM or LSCM) generates an image by a completely different way than the normal visual bright field microscope. It gives slightly higher resolution, but more importantly it provides optical sectioning without disturbing out-of-focus light degrading the image. Therefore it provides sharper images of 3D objects. This is often used together with fluorescence microscopy (see Optical microscopy: fluorescence).
Computer reconstructed 3D-images of spatially complicated objects are obtained by in-focus images of thick specimens, a process known as *optical sectioning*.

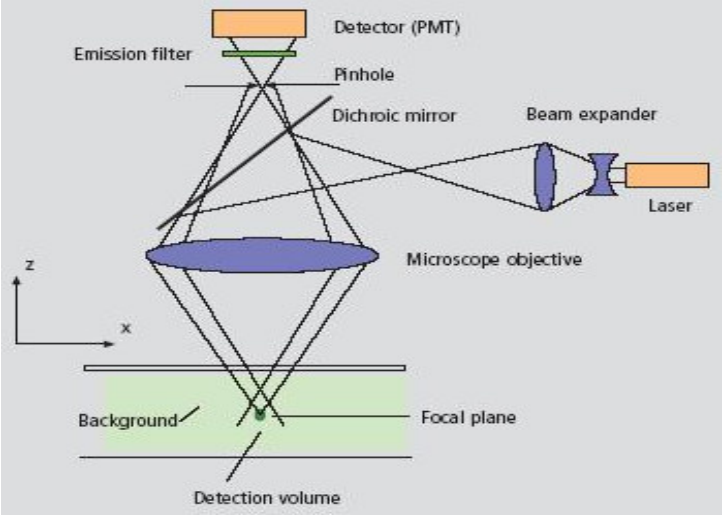


Fig. 1 Beam path in a confocal LSM.

A microscope objective (see Optical microscopy) is used to focus a laser beam onto the specimen, where it excites Fluorescence, for example. The fluorescent radiation is collected by the objective lens (see Light: the ideal and non-ideal lens) and efficiently directed onto the detector via a dichroic (see Dichroism) beam splitter (see Light: beam splitter). The interesting wavelength range of the fluorescence spectrum is selected by an emission filter, which also acts as a barrier blocking the excitation laser line. The pinhole is arranged in front of the detector (see Fig. 1), on a plane conjugate to the focal plane of the objective. Light coming from planes above or below the focal plane is out of focus when it hits the pinhole, so most of it cannot pass the pinhole and therefore does not contribute to form the image. After passing the pinhole, the fluorescent light is detected by a photodetection device, a photomultiplier tube (PMT) or avalanche photodiode (a specialized photodiode based on ionization), transforming the light signal into an electrical one that is recorded by a computer.

The pinhole is the extra and very relevant optical part compared to conventional fluorescence microscopy (see Optical microscopy: fluorescence) resulting in sharper images and permits one to obtain images of various vertical axis planes of the sample. The pinhole not only blocks the out-of-focus light but also permits the optical sectioning along the z-axis planes (yielding the *z stacks* of the sample).

The detected light originating from an illuminated volume element within the specimen represents one pixel in the resulting image. As the laser scans over the plane of interest, a whole image is obtained pixel-by-pixel and line-by-line, and finally plane by plane results in the 3D image. The brightness of an image pixel corresponds to the relative intensity of detected fluorescent light. The beam is scanned by using one or more oscillating mirrors. Information can be collected from different focal planes by raising or lowering the microscope stage.

Confocal microscopy also provides a substantial improvement in lateral resolution and the capacity for direct, noninvasive, serial optical sectioning of intact, thick, living specimens with a minimum of sample

preparation. Because CLSM depends on fluorescence, a sample usually needs to be treated with fluorescent dyes to make objects visible. However, the actual dye concentration can be low to minimize the disturbance of biological systems: some instruments can track single fluorescent molecules. Also, transgenic techniques can create organisms that produce their own fluorescent chimeric molecules.
CLSM is a scanning imaging technique in which the resolution obtained is best explained by comparing it with another scanning technique like that of the scanning electron microscope (SEM). Do not confuse CLSM with phonograph-like imaging, AFM (Atomic force microscopy (AFM)) or STM (Scanning tunneling microscopy (STM)), for example, where the image is obtained by scanning with an atomic tip over a conducting surface.
The resolution of the 3D diffraction pattern and the focal volume is not given by the numerical aperture (see Fiber optics) of the system's objective lens and the wavelength of the laser used since only light generated in a small volume element is detected at a given time during the scanning process. The optimal size of the pinhole is the size of the point spread function (the central peak) of the objective lens. This strongly increases the resolution at detector level. Without pinhole, a much larger volume element of the sample contributes to the detector signal.

Application

CLSM is widely-used in biological and biomedical disciplines, from cell biology and genetics to microbiology and developmental biology.
Clinically, CLSM is used in the evaluation of various eye diseases, and is particularly useful for imaging, qualitative analysis, and quantification of endothelial cells of the cornea. It is used for localizing and identifying the presence of filamentary fungal elements in the corneal stroma in cases of keratomycosis, enabling rapid diagnosis and thereby early institution of definitive therapy. Research into CLSM techniques for endoscopic procedures is also showing promise.
Because CLSM depends on fluorescence, a sample usually needs to be treated with fluorescent dyes to make objects visible. However, the actual dye concentration can be low to minimize the disturbance of biological systems: some instruments can track single fluorescent molecules. Also, transgenic techniques can create organisms that produce their own fluorescent chimeric molecules (such as a fusion of GFP, green fluorescent protein with the protein of interest).

CLSM is also used for archeological determination of age (e.g. the Magdalen papyrus).

More Info

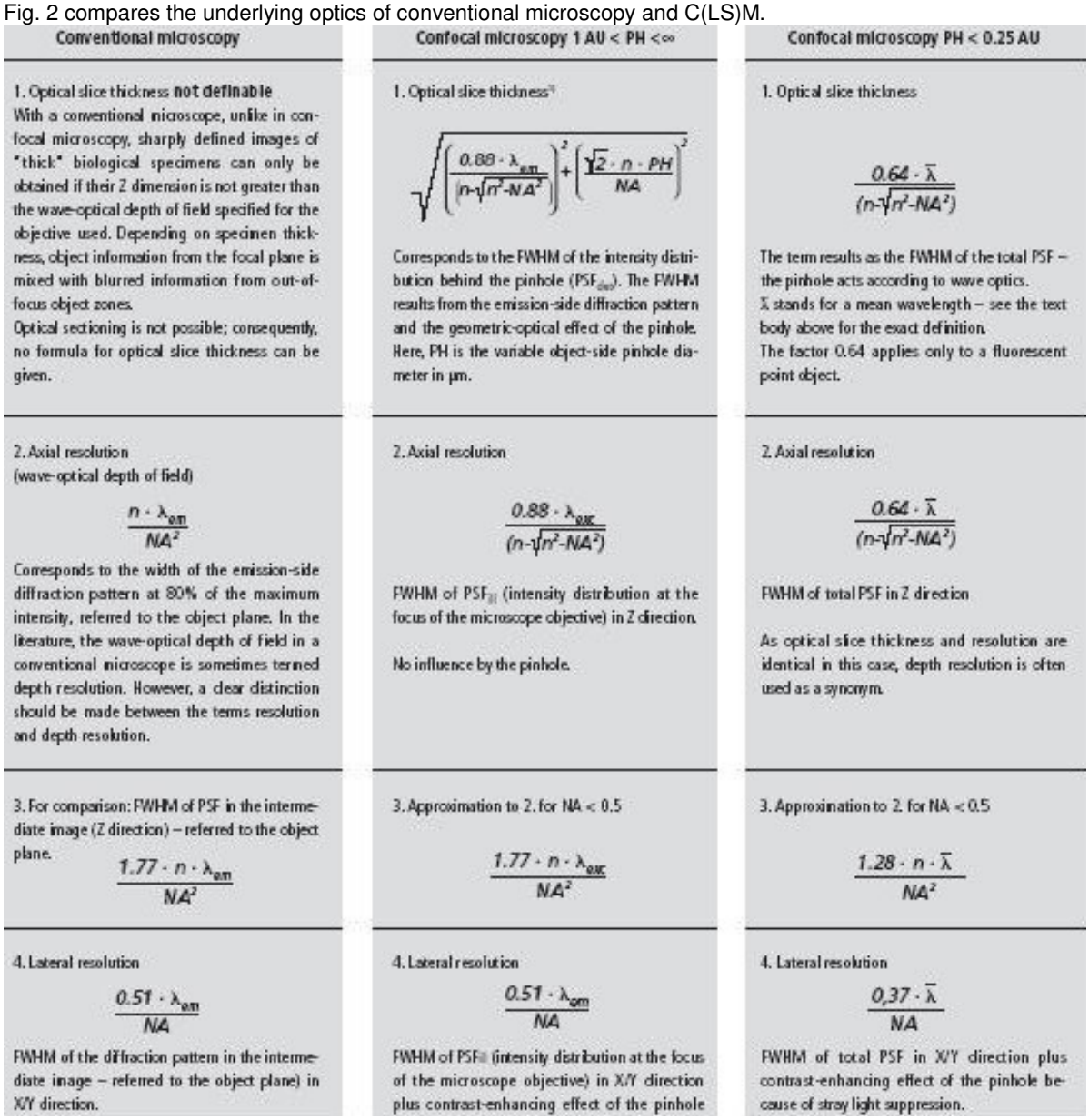
The servo controlled oscillating mirrors performing the scanning across the sample in the horizontal plane have a low reaction latency and the scan speed determines the signal-to-noise ratio. This is due to the small number of photons typically available in fluorescence microscopy. One can compensate for this effect by using more sensitive photodetectors or by increasing the intensity of the illuminating laser point source. Increasing the intensity of illumination later risks excessive bleaching or other damage to the specimen of interest, especially for experiments in which comparison of fluorescence brightness is required.

Resolution enhancement
The obtained CLSM resolution is best explained by comparing it with another scanning technique like that of scanning Electron microscopy (SEM).
The size of the pixel based scanning volume is determined by the spot size (close to diffraction limit) of the optical system because the image of the scanning laser is not an infinitely small point but a 3D diffraction pattern. The size of this diffraction pattern can be reduced by blocking the higher orders of the diffraction pattern. For example, if the pinhole diameter is set to 1 Airy unit (see Light: diffraction, More info) then only the first order of the diffraction pattern makes it through the aperture to the detector, thus improving resolution at the cost of a slight decrease in brightness.

The resolution limit in confocal microscopy depends not only on the probability of illumination but also on the probability of creating enough detectable photons, so that the actual addressable volume being associated with a generated light intensity is smaller than the illuminated volume.
Depending on the fluorescence properties of the used dyes, there is a more or less subtle improvement in lateral resolution compared to conventional microscopes. However, with light creation processes with much lower probabilities of occurrence such as frequency doubling or a second harmonic generation (SHG, see also Fourier analysis), the volume of addressing is reduced to a small region of highest laser illumination intensity, substantially improving lateral resolution. SHG is a nonlinear process, in which photons interacting with a nonlinear material are effectively "combined" to form new photons with twice the energy, and therefore half the wavelength of the initial photons. Unfortunately, the probability decrease in creation of detectable photons negatively affects the signal-to-noise ratio. One can compensate for this effect by using more sensitive photo detectors or by increasing the intensity of the

illuminating laser point source. Increasing the intensity of illumination later risks excessive bleaching or other damage to the specimen of interest, especially for experiments in which comparison of fluorescence brightness is required.

Fig. 2 compares the underlying optics of conventional microscopy and C(LS)M.



Conventional microscopy	Confocal microscopy 1 AU < PH <∞	Confocal microscopy PH < 0.25 AU
1. Optical slice thickness **not definable** With a conventional microscope, unlike in confocal microscopy, sharply defined images of "thick" biological specimens can only be obtained if their Z dimension is not greater than the wave-optical depth of field specified for the objective used. Depending on specimen thickness, object information from the focal plane is mixed with blurred information from out-of-focus object zones. Optical sectioning is not possible; consequently, no formula for optical slice thickness can be given.	1. Optical slice thickness[1] $\sqrt{\left(\frac{0.88 \cdot \lambda_{em}}{n-\sqrt{n^2-NA^2}}\right)^2+\left(\frac{\sqrt{2} \cdot n \cdot PH}{NA}\right)^2}$ Corresponds to the FWHM of the intensity distribution behind the pinhole (PSF_{det}). The FWHM results from the emission-side diffraction pattern and the geometric-optical effect of the pinhole. Here, PH is the variable object-side pinhole diameter in µm.	1. Optical slice thickness $\frac{0.64 \cdot \bar{\lambda}}{(n-\sqrt{n^2-NA^2})}$ The term results as the FWHM of the total PSF – the pinhole acts according to wave optics. $\bar{\lambda}$ stands for a mean wavelength – see the text body above for the exact definition. The factor 0.64 applies only to a fluorescent point object.
2. Axial resolution (wave-optical depth of field) $\frac{n \cdot \lambda_{em}}{NA^2}$ Corresponds to the width of the emission-side diffraction pattern at 80% of the maximum intensity, referred to the object plane. In the literature, the wave-optical depth of field in a conventional microscope is sometimes termed depth resolution. However, a clear distinction should be made between the terms resolution and depth resolution.	2. Axial resolution $\frac{0.88 \cdot \lambda_{exc}}{(n-\sqrt{n^2-NA^2})}$ FWHM of PSF_{II} (intensity distribution at the focus of the microscope objective) in Z direction. No influence by the pinhole.	2. Axial resolution $\frac{0.64 \cdot \bar{\lambda}}{(n-\sqrt{n^2-NA^2})}$ FWHM of total PSF in Z direction As optical slice thickness and resolution are identical in this case, depth resolution is often used as a synonym.
3. For comparison: FWHM of PSF in the intermediate image (Z direction) – referred to the object plane. $\frac{1.77 \cdot n \cdot \lambda_{em}}{NA^2}$	3. Approximation to 2. for NA < 0.5 $\frac{1.77 \cdot n \cdot \lambda_{exc}}{NA^2}$	3. Approximation to 2. for NA < 0.5 $\frac{1.28 \cdot n \cdot \bar{\lambda}}{NA^2}$
4. Lateral resolution $\frac{0.51 \cdot \lambda_{em}}{NA}$ FWHM of the diffraction pattern in the intermediate image – referred to the object plane) in X/Y direction.	4. Lateral resolution $\frac{0.51 \cdot \lambda_{em}}{NA}$ FWHM of PSF_{II} (intensity distribution at the focus of the microscope objective) in X/Y direction plus contrast-enhancing effect of the pinhole	4. Lateral resolution $\frac{0,37 \cdot \bar{\lambda}}{NA}$ FWHM of total PSF in X/Y direction plus contrast-enhancing effect of the pinhole because of stray light suppression.

Fig. 2 Comparison of conventional microscopy and C(LS)M . Abbreviations and symbols: α Aperture angle of a microscope objective, AU Airy unit (diameter of Airy disc, the 3dB width of the central spot in the point spread function, see Light and Optical microscopy), dpix Pixel size in the object plane, FWHM Full width at half maximum of an intensity distribution (e.g. optical slice), n Refractive index of an immersion liquid, NA Numerical aperture (see Fiber optics) of a microscope objective, PH Pinhole; diaphragm of variable size arranged in the beam path to achieve optical sections, PSF Point spread function, RU Rayleigh unit (10^{10} photons m^{-2} s^{-1}), SNR Signal-to-noise ratio. From http://www.zeiss.com/c1256d18002cc306/0/f99a7f3e8944eee3c1256e5c0045f68b/$file/45-0029_e.pd, an excellent explanation, basically as well as advanced.

Optical microscopy: fluorescence

Principle

This method can be extremely sensitive, allowing the detection of single molecules. Many different fluorescent dyes can be used to stain different structures or chemical compounds.

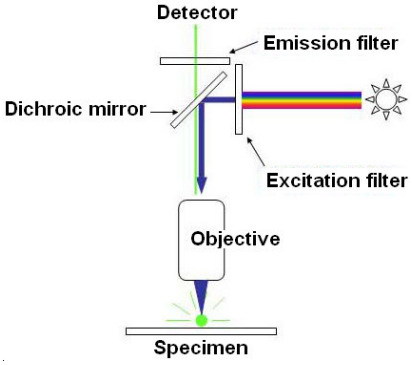


Fig. 1 Principle of fluorescence microscopy. The light source is mostly a xenon or mercury arc-discharge lamp. Only blue light reaches the specimen and by fluorescence green light is emitted.

To block the excitation light from reaching the observer or the detector, filter sets of high quality are needed. These typically consist of an excitation filter selecting the range of excitation wavelengths, a dichroic mirror (see Dichroism), and an emission filter blocking the excitation light. Most Fluorescence microscopes are operated in the epi-illumination mode (illumination and detection from one side of the sample, see below) to further decrease the amount of excitation light entering the detector.
See also total internal reflection fluorescence light microscopy (see **More Info**).

Application

One particularly powerful method is the combination of antibodies coupled to a fluorochrome, e.g. fluorescein or rhodamine as in immunostaining. The antibodies can be made tailored specifically for a chemical compound. For example, one strategy often in use is the artificial production of proteins, based on the genetic code (DNA). These proteins can then be used to immunize rabbits, which then form antibodies which bind to the protein. The antibodies are then coupled chemically to a fluorochrome and then used to trace the proteins in the cells under study. Many uses of immunofluorescence have been outmoded by the development of highly-efficient fluorescent recombinant proteins containing fluorescent protein domains, e.g. green fluorescent protein (GFP). Use of such "tagged" proteins allows much better localization and less disruption of protein function. Genetically modified cells or organisms directly express the fluorescently-tagged proteins, which enables the study of the function of the original protein in vivo.
Since fluorescence emission differs in wavelength from the excitation light, a fluorescent image ideally only shows the structure of interest that was labeled with the fluorescent dye. This high specificity led to the widespread use of fluorescence light microscopy in biomedical research. Different fluorescent dyes can be used to stain different biological structures, which can then be detected simultaneously, while still being specific due to the individual color of the dye.

More Info

In cell and molecular biology, a large number of processes in or near cellular membranes such as cell adhesion, secretion of neurotransmitters and membrane dynamics have been studied with conventional fluorescence microscopy. However, fluorophores that are bound to the specimen surface and those in the surrounding medium exist in an equilibrium state. When these molecules are excited and detected with a conventional fluorescence microscope, the resulting fluorescence from those fluorophores bound to the surface is often overwhelmed by the background fluorescence due to the much larger population of non-bound molecules.
Fluorophores lose their ability to fluoresce as they are to strongly illuminated (photobleaching). The use of more robust fluorophores is another option.

Epifluorescence microscopy
This method, instead of transmitting the excitatory light through the specimen it is passed through the objective onto the specimen. Since only reflected excitatory light filters through, the transmitted light is filtered out, giving a much higher intensity. Epifluorescence microscopy is widely used in life sciences.

Total internal reflection fluorescence microscope

Total internal reflection fluorescence microscope (TIRFM) is a modification with which a thin region of a specimen, usually less than 200 nm can be observed.

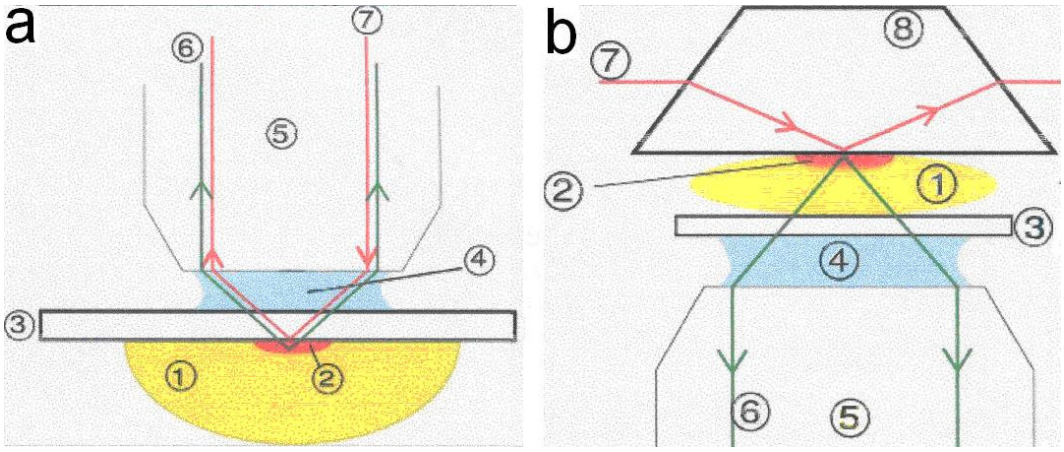


Fig. 2 a. Principle of *epi*-TIRFM. 1. Specimen, 2. Evanescent wave range, 3. Cover slip, 4. Immersion oil, 5. Objective, 6. Emission beam (signal), 7. Excitation beam. b. Principle of *trans*-TIRFM. 1. to 7. as figure a., 8. quartz prism.

A TIRFM uses evanescent waves to selectively illuminate and excite fluorophores in a restricted region of the specimen immediately adjacent to the glass-water interface. An evanescent (means "tends to vanish") wave is a near field (until λ/π from the source) standing wave exhibiting exponential decay with distance (see further Light: Fresnel diffraction). Evanescent waves are generated only when the incident light is totally reflected at the glass-water interface. Due to exponential decay from the interface it penetrates to a depth of only approximately 100 nm into the sample medium. Thus the TIRFM enables a selective visualization of surface regions such as the basal plasma membrane (which are about 7.5 nm thick) of cells as shown in the figure above. The selective visualization of the plasma membrane renders the features and events on the plasma membrane in living cells with high axial resolution.
TIRFM can also be used to observe the fluorescence of a single molecule, making it an important tool of biophysics and quantitative biology.

Another new fluorescence microscopical technique is 2-photon fluorescence microscopy. (see Optical microscopy: 2-photon fluorescence).

Optical microscopy: 2-photon fluorescence

Principle

Multi-Photon Fluorescence Microscopy is a relatively novel imaging technique in cell biology. It relies on the quasi-simultaneous absorption of two or more photons (of either the same or different energy) by a molecule. During the absorption process, an electron of the molecule is transferred to an excited-state electron orbital in the molecule. The molecule (i.e. the fluorophore) in the excited state has a high probability (> 10 %) to emit a photon during relaxation to the ground state. Due to radiationless relaxation in vibrational levels, the energy of the emitted photon is lower compared to the sum of the energy of the absorbed photons.

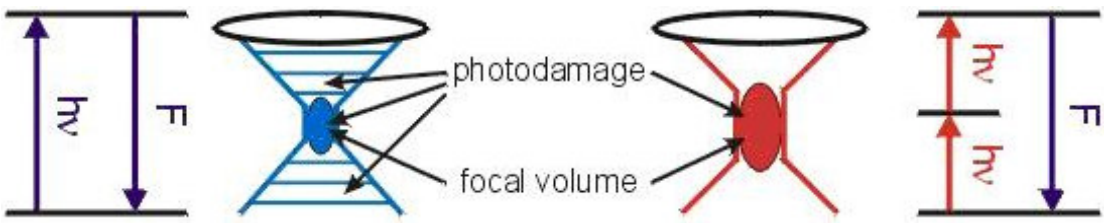


Fig. 1 Principle of fluorescence induced by one-photon absorption (left) and two-photon absorption (right).

While the resolution in two-photon fluorescence microscopy (2PFM) is less good, photo-damage is lower and penetration depth is higher compared to single-photon (confocal) fluorescence microscopy (1PFM) due to the large wavelength (often near IR) of the excitation light. The multi-photon absorption process needs a very high density of photons (0.1 - 10 MW/cm^2) from a ps-to-fs-pulsed light source. This is because the virtual absorption of a photon of non-resonant energy lasts only for a very short period (10^{-15} - 10^{-18}s). During this time a second photon must be absorbed to reach an excited state. Multi-photon fluorescence microscopy in cell biology uses pulsed lasers (pulse width in the range of 20 fs to 5 ps, typically ~ 100 fs) with high repetition rates (10 -100 MHz).

2PFM is the most common multi-photon fluorescence application in cell biology for two reasons. First, the best performing Laser (Ti:Sapphire laser) amply covers the spectral region of the often used near IR with pulse widths of 100 - 150 fs and repetition frequencies of about 75 MHz. Second, the performance of fluorescence microscopy relies very much on the use of highly specific dye molecules (e.g. pH- or other ion indicators), originally applied with classical microscopy using near UV and visible light sources. Therefore, most of the dyes absorb in the near-UV and the visible spectral region. These well-characterized dyes are also used in 2PFM. For multi-photon fluorescence microscopy, using the Ti:Sapphire laser for excitation, well-characterized dyes, which fluorophores should absorb in the far UV, have not yet been developed.

Application

Applications are in cell biology, fundamental as well as experimental clinical.

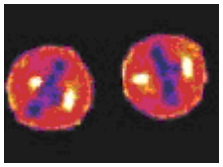


Fig. 2 2PFM images of two dividing HEK293 cells (blue = low intensity; red = medium intensity; yellow = high intensity). The fluorescence originates from a Calmodulin-EGFP fusion protein, which binds to Myosin Light Chain Kinase (at the cell membrane) and to CaM-Kinase II (in the meiotic spindle poles) during cell division.

More Info

The most important differences of 2PFM compared to 1PFM are the quadratic dependence on the average power (linear in 1PFM) and the power-of-four dependence on the NA of the microscope objective (quadratic in 1PFM).
2PFM has several advantages over commonly used fluorescence microscopy techniques. If the average excitation power is moderate, the two-photon absorption process takes place only in the focus of the laser beam and thereby provides a three-dimensional resolution. Radial and axial resolutions of well below 0.5 µm and 1 µm, respectively, are achieved for a typical excitation wavelength (800 nm) by using microscope objectives with a high numerical aperture. The axial resolution of 2PFM is a great advantage compared to the classical 1PFM. A similar axial resolution can be achieved in 1PFM using a confocal set-up (see Optical microscopy: confocal microscopy). The three-dimensional spatial resolution in 1PFM is compromised by lower signal intensity compared to 2PFM. Therefore, confocal 1PFM requires higher excitation intensities. This reduces the observation time of a photo-vulnerable object like cells or biological tissues.

The use of near-IR radiation in 2PFM enhances the penetration depth and reduces image deterioration due to scattering when passing through biological tissue. Since the elastic scattering of light (see Light: scattering) is proportional λ^{-n} (with an exponent between 4 as holds for Rayleigh scattering and 2.2 with scattering in between the former and Tyndall scattering, depending on the material), the scattering - a major contributor to image deterioration - is an order of magnitude less pronounced in 2PFM. It allows imaging in 10 times deeper regions compared to 1PFM.

Two of the major limiting factors in the use of fluorescence microscopy are photo-bleaching and photo-damage. Both are limited to the focal region in 2PFM, whereas in 1PFM the upper and lower regions of the excitation light cone are affected. Although 2PFM has a slightly lower resolution and requires complicated excitation sources, it can be still advantageous due to the much longer observation time. This allows following physiological events longer in time.

An additional advantage of the fs-to-ps pulsed excitation with MHz repetition rates is the capability of fluorescence lifetime imaging (FLIM). The fluorescence lifetime is obtained by repetitive measurement of the time laps between the excitation pulse and the fluorescence photon (TCSPC, Time-Correlated Single Photon Counting). Each pixel contains information on both the fluorescence lifetime and the "classical" fluorescence. Fluorescence lifetime imaging is valuable especially for intracellular measurements, where the absolute number of fluorophores, either membrane permeable organic fluorophores or autofluorescent proteins (e.g. GFPs), can not be determined in a quantitative way.

Modified after: http://www.fz-juelich.de/inb/inb-1//Two-Photon_Microscopy/

Optical microscopy: phase contrast

Principle

Phase contrast microscopy improves visibility of transparent and colorless objects by making use of the differences in phase. These differences result from differences in refractive index (the optical density) of the various parts of the object. However, phase differences are not visible to the human eye. The design is such that the phase differences are transformed to amplitude differences, which are in the plane of the image seen as contrast differences. For example, the nucleus in a cell will show up darkly against the surrounding cytoplasm. Contrast is good; however, it is not for use with thick objects. Frequently, a halo is formed even around small objects, which obscures detail.

Application

The phase contrast microscope is widely used for examining such specimens as biological tissues. The phase contrast microscope is a vital instrument in biological and medical research. When dealing with transparent and colorless components in a cell, dyeing is an alternative but at the same time stops all processes in it. The phase contrast microscope has made it possible to study living cells, and cell division is an example of a process that has been examined in detail with it.

More info

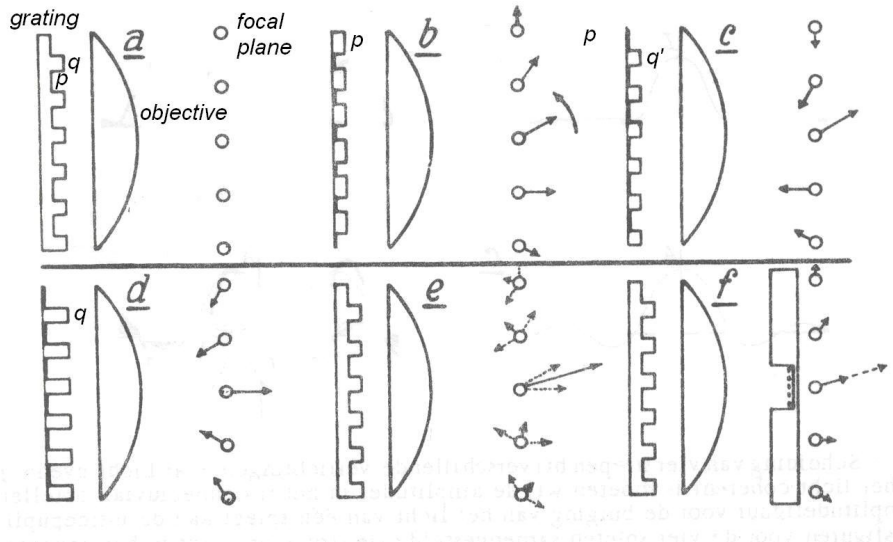


Fig. 1 Principe of phase delays of principle maximum and submaxima of an amplitude grating.

The working principle can be understood by using as an object a phase grating (see for the optics of gratings Light: diffraction). This is a grating with alternating thickness (*p* and *q*) of a transparent material (glass etc.) as depicted in Fig. 1 a. Fig. 1b gives the phase of the principle maximum in the middle and those of the first and second order maxima of the thin parts (p). Fig. 1c gives those of the thick parts, but for the moment, the thickness is considered the same as in Fig. 1b (*q'*). The central principal maximum has the same phase, but the others are opposite. The gratings of b. and c. added together do not form a grating at all. This will only yield a principal maximum, no higher order maxima.

Therefore, in c. the phases of the higher order maxima are opposite. In Fig. 1 d, the actual thickness *q* is taken into account. This yields an additional phase shift of all maxima. Fig. 1e gives the vectorial addition of b. and d. The deviation of e. with respect to b. is an extra shift of the submaxima of 90^o. By changing artificially the phase of the principal maximum with 90^o and providing extra absorption by a phase plate, the relative phases are the same is in b. Grating e. that is actually an amplitude grating can be seen as a grating with different zones of refractive index (*p* and *q*). In conclusion, phase differences are transformed to contrast differences.

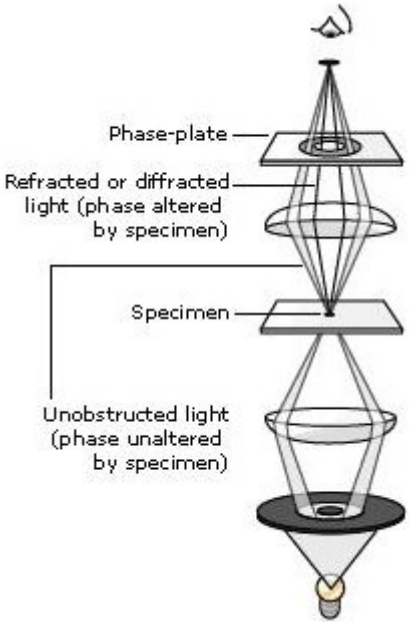


Fig. 2 Basic design of phase contrast microscope

Fig. 2 gives the basic light pathway. The system consists of a circular annulus in the condenser, which produces a cone of light with no light in the middle of the cone. This annular shaped cone is superimposed on a similar sized ring within a special objective, the phase-objective. Every objective has a different sized ring, so for every objective another condenser setting has to be chosen.

Optical microscopy: specific techniques

Stereo microscopy

Fig. 1 Stereo microscope

Principle The stereo or dissecting microscope uses two separate optical paths with two objectives and two eyepieces to provide slightly different viewing angles to the both eyes. In this way it produces a 3D visualization of the sample being examined.

Application The stereo microscope is often used to study the surfaces of solid specimens or to carry out close work such as sorting, dissection, microsurgery, watch-making, small circuit board manufacture or inspection.
More info Great working distance and depth of field are important qualities which are inversely correlated with resolution. The resolution is maximally that of an average 10× objective in a compound microscope, and often much lower. The useful magnification factor is up to 100×.
Video dual CCD camera pickups can been fitted to stereo microscopes, allowing the images to be displayed on a high resolution LCD monitor. Software converts the two images to an integrated anachrome (made of hues in between two complementary colors (see Vision of color)) 3D image, for viewing with plastic red and cyan glasses. These colors are complementary (see Vision of color). Also other modern stereo-vision viewing can be used to help vision of depth. The results are viewable by a group wearing the glasses.

Oblique illumination microscopy
The use of oblique (from the side) illumination gives the image a 3D appearance and can highlight otherwise invisible features. A more recent technique based on this method is *Hoffmann's modulation contrast*, a system found on inverted microscopes for use in cell cultures.

Dark field microscopy
Principle Dark field microscopy improves the contrast of unstained, transparent specimens. Dark field illumination uses an annular carefully aligned light beam to minimize the quantity of directly-transmitted (not-scattered) light entering the objective. Only the light scattered by the sample enters the objective. Therefore it produces a dark, almost black, background with bright objects on it. The annular beam is produced by a central stop before the condenser. Although image contrast is strongly improved, the technique does suffer from low light intensity in the final image and continues to be affected by low apparent resolution.
Application Detecting and counting small objects in suspension (e.g. microorganisms) or for instance tissue slides with developed radioactive grains.
More info *Rheinberg illumination* is a special variant of dark field illumination in which transparent, colored filters are inserted just before the condenser so that light rays at high aperture are differently colored than those at low aperture (i.e. the background to the specimen may be blue while the object appears self-luminous yellow). Other color combinations are possible but their effectiveness is quite variable.

Phase contrast microscopy, see Optical microscopy: phase contrast

Differential interference contrast (DIC) or Nomarski microscopy

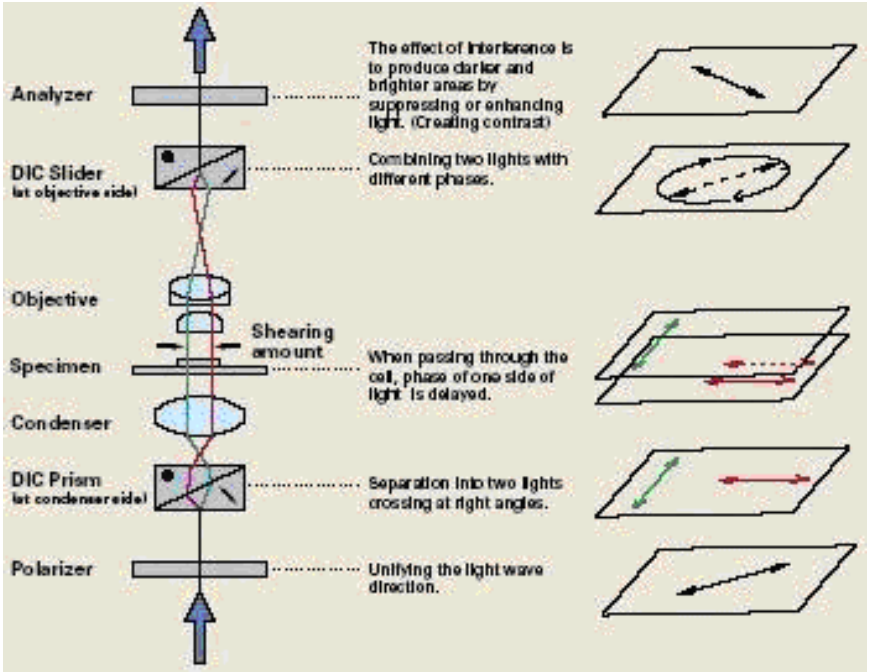


Fig. 2 Principle of the light pathways in Nomarski microscope. The dashed red arrow in the specimen plane present the beam without phase delay.

Principle Superior and much more expensive than phase contrast microscopy is the use of interference contrast. Differences in optical density will show up as differences in relief. A nucleus within a cell will actually show up as a globule in the most often used version, the Nomarski method. However, this is an *optical effect*, and the relief does not necessarily resemble the true shape! Contrast is very good and the condenser aperture can be used fully open, thereby reducing the depth of field and maximizing resolution.
Application Microscopy of living tissue cultures, microorganism, small invertebrates (nematodes) etc.
More info This method consists of a special prism (a Nomarski-modified Wollaston prism) in the condenser that splits light in two beams with their direction of polarization perpendicular to each other (see Light: polarization). The difference in spatial position between the two beams, the shearing amount, is less than the resolution of the objective.
After passage through the specimen and objective lenses, the beams are reunited by a 2nd Nomarski prism (DIC Slider). In a homogeneous specimen, all rays recombine to light with the same and unchanged polarization (linear polarized beam at 45^0), indicated by the black dashed arrow in DIC slider plane at the right of Fig. 2. Now, no contrast is being generated. However, the delayed beam recombines with the reference beam to a circular polarized beam. After the analyzer the 45^0, polarized light is blocked, yielding black and the circular polarized is passed yielding a bright spot in the image. In this way, the spatial variety in refractive index position is transformed to a brightness image of the specimen. Such varieties can be strong, for instance at the boundary a nucleus within the cytoplasm. For the various applications there are different DIC sliders.

Fluorescence microscopy (see Optical microscopy: fluorescence)

Confocal laser scanning microscopy (see Optical microscopy: confocal laser scanning)

In addition to these techniques optical microscopy has the versions of ultraviolet (UV) and infrared (IR) microscopy, the later including photothermal microscopy. and reflection contrast microscopy (RCM). More complicated are holographic interference microscopy, holographic phase contrast microscopy (see Holography), laser projection microscopy with brightness amplification, scanning optical microscopy, nonlinear microscopy with 2nd harmonic generation, Raman (see Raman spectroscopy) and optoacoustic microscopy (see Optoacoustic imaging).

Optical microscopy: super-resolution techniques

Mathematical technique

Deconvolution microscopy
If one considers a small light source (essentially a bright infinitesimal small spot), light coming from this spot and imaged by an ideal lens (no aberrations, see Light: the ideal and non-ideal lens), the image spreads out from the center. The image is the point spread function (PSF). It is proportional to the function $(\sin x/x)^2$ (x is the distance from the optical axis (z-axis)) and circle-symmetric around the z-axis (see Light: diffraction, Fig. 1 and 2).
Since any image is made up of a large number of such small (e.g. fluorescent) light sources the image is convolved by the PSF (see for convolution Stochastic signal analysis).
Knowing this PSF means, than it is possible to reverse this process to a certain extent by computer based methods commonly known as deconvolution. It is applied in deconvolution microscopy to remove the PSF-blur. The various algorithms can be roughly classified in *non restorative* and *restorative* methods. While the non restorative methods can improve contrast by removing out of focus light from focal planes, only the restorative methods can actually reassign light to it proper place of origin. This can be an advantage over other types of 3D microscopy such as confocal microscopy, because light is not thrown away but "reused". For (pseudo-)3D deconvolution one typically provides a series of images derived from different focal planes plus the knowledge of the PSF which can be either derived experimentally or theoretically from knowing all contributing optical components of the microscope.

Physical techniques: sub-diffraction techniques

The spatial limit of focusing is about half λ, but this is not a true barrier, because this diffraction limit is only true in the far-field (>λ/π). Localization precision can be increased with many photons and careful analysis. The following explores some approaches to image objects smaller than 250 nm.

Near-field scanning optical microscopy (NSOM)
The principle is using a light source and/or a (fluorescence) detector that is itself nm-scale. Diffraction is actually a far-field effect: the light from an aperture is the Fourier transform (see Fourier analysis) of the

aperture in the far-field. But in the near-field (<λ/π), all of this is not necessarily the case. NSOM forces light through the tiny tip of a pulled glass fiber with an aperture on the order of tens of nm. When the tip is brought to nm's away from a molecule, the resolution is not limited by diffraction but by the size of the tip aperture (because only that one macromolecule will see the light coming out of the tip). An image can be built by a raster scan (parallel lines comprised of pixels) of the tip over the surface to create an image.
The main down-side to NSOM is the limited number of photons you can force out a tiny tip, and the minuscule collection efficiency (if you are trying to collect fluorescence in the near-field). Other techniques such as ANSOM (see below) try to avoid this drawback.

ANSOM
ANSOM is aperture-less NSOM: it uses a tip very close to a fluorophore to enhance the local electric field the fluorophore sees. Instead of forcing photons down a tiny tip, this technique creates a local bright spot in an otherwise diffraction-limited spot. Basically, the ANSOM tip is like a lightning rod which creates a hot spot of light.

Stimulated emission depletion (STED)
This method uses two laser pulses. The first pulse is a diffraction-limited spot that is tuned to the absorption wavelength, so excites any fluorophores in that region; an immediate second pulse is red-shifted to the emission wavelength and stimulates emission back to the ground state, thus depleting the excited state of any fluorophores in this depletion pulse. The trick is that the depletion pulse goes through a phase modulator that makes the pulse illuminate the sample in the shape of a donut, *so the outer part of the diffraction limited spot is depleted and the small center can still fluoresce*. By saturating the depletion pulse, the center of the donut gets smaller and smaller until they can get resolution of tens of nanometers. This technique also requires a raster scan like NSOM and standard confocal laser scanning microscopy (see Optical microscopy: confocal laser scanning).

Fitting the PSF
The methods above (and below) use experimental techniques to circumvent the diffraction barrier, but one can also use analysis to increase the ability to know where a nanoscale object is located. The PSF, limited by diffraction, can be used to locate the center of the PSF and thus the location of the fluorophore. The precision depends on the number of photons collected. This analysis localizes single fluorophores to a few nm. This, of course, requires collecting *many* photons.

Photo-activated localization microscopy (PALM) & Stochastic optical reconstruction microscopy (STORM)
What fitting a PSF is to localization, PALM is to "resolution". This term is here used loosely to mean measuring the distance between objects, not true optical resolution. The basic premise of both techniques is to fill the imaging area with many dark fluorophores that can be photoactivated. Because photoactivation is stochastic, only a few, well separated molecules "turn on". Then a Gaussian function is fitted to the PSF for high localization. After the few bright dots photobleach, another flash of the photoactivating light activates random fluorophores again and the PSFs are fit of these different well spaced objects. This process is repeated many times, building up an image molecule-by-molecule; and because the molecules were localized at different times, the "resolution" of the final image can be much higher than that limited by diffraction.
The major problem with these techniques is that to get these beautiful pictures, it takes on the order of hours to collect the data. This is certainly not the technique to study dynamics (fitting the PSF is better for that).

Structured illumination
This technique applies wide-field structured-illumination (SI) with patterned light and relies on both specific microscopy protocols and extensive software analysis post-exposure. Because SI is a wide-field technique, it is usually able to capture images at a higher rate than confocal-based schemes like STED. The patterned light increases the resolution by measuring the fringes in the Moiré pattern (from the interference of the illumination pattern and the sample) which are used computationally to restore the image.
SI enhances spatial resolution by collecting information from frequency space outside the observable region. This process is done in reciprocal space: the Fourier transform of 2D space (FT, see Fourier transform) of an SI image contains superimposed additional information from different areas of reciprocal space; with several frames with the illumination shifted by some phase, it is possible to computationally separate and reconstruct the FT image, which has much more resolution information. The reverse FT returns the reconstructed image with a gain of resolution of a factor 2 (2 because the SI pattern cannot be focused to anything smaller than half the wavelength of the excitation light). To further increase the resolution, you can introduce *nonlinearities*, which show up as higher-order harmonics in the FT. A sinusoidal saturating excitation beam produces the distorted fluorescence intensity pattern in the emission. This nonlinearity yields a series of higher-order harmonics in the FT.

Each higher-order harmonic in the FT allows another set of images that can be used to reconstruct a larger area in reciprocal space, and thus a higher resolution down to 50-nm.
The main problems with SI are that, in this incarnation, saturating excitation powers cause more photo damage and lower fluorophore photostability, and sample drift must be kept to below the resolving distance. The former limitation might be solved by using a different nonlinearity (such as stimulated emission depletion or reversible photoactivation, both of which are used in other sub-diffraction imaging schemes); the latter limits live-cell imaging and may require faster frame rates or the use of some usual markers for drift subtraction. Nevertheless, SI is certainly a strong contender for further application in the field of super-resolution microscopy.

Literature
For STED see http://www.mpibpc.gwdg.de/groups/hell/STED.htm.

Phosphorescence

Principle

Phosphorescence is a specific type of photoluminescence, related to fluorescence, but distinguished by slower time scales associated with quantum mechanically forbidden energy state transitions of electron orbits.
Phosphorescence is a process in which energy stored in a substance is released very slowly and continuously in the form of glowing light. The reason of the characteristically slow rate of energy release is that on a microscopic level, the probability for the process of emitting light to occur is very low, almost being forbidden by quantum mechanics. A Phosphorescent substance is at the same time also fluorescent as is illustrated in Fig. 1.

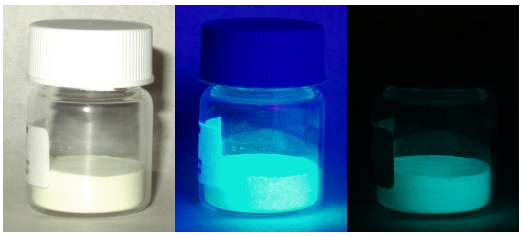


Fig. 1 Phosphorescent powder under visible light (left), ultraviolet light, and total darkness (right).

Applications

Industrial (cathode ray tubes, analog TV sets, watch dials and dials of other consumer equipment), tryptophan phosphorescence spectroscopy and phosphorescence in (pathogenic) bio-aerosols are bio-medical applications. Further labeling of biomolecules by a phosphorophore, a carrier of phosphoresce. The delayed emission and the temperature dependency promises future biomaterial identification.

More Info

Most photoluminescent events, in which a chemical substrate absorbs and then re-emits a photon of light, are fast, on the order of 10 ns! However, for light to be absorbed and emitted at these fast time scales, the energy of the photons involved (i.e. the wavelength of the light) must be carefully tuned according to the rules of quantum mechanics to match the available energy states and allowed transitions of the molecule. In the special case of phosphorescence, the absorbed photon energy undergoes an unusual transition into a higher energy state (usually a triplet state, a set of three quantum states of a system, each with total spin $S = 1$). This state is metastable and transition to the initial state is forbidden. As a result, the energy can become trapped in a state with only quantum mechanically "forbidden" transitions available to return to the lower energy state. Emission occurs when thermal energy raises the electron to a higher state from which it can de-excite. Therefore, phosphorescence is temperature dependent.
Most phosphorescent compounds are still relatively fast emitters, with triplet lifetimes on the order of milliseconds. However, some compounds have triplet lifetimes up to minutes or even hours, allowing these substances to effectively store light energy in the form of very slowly degrading excited electron states. If the phosphorescent quantum yield is high, these substances will release significant amounts of light over long time scales, creating so-called "glow in the dark" materials.

Snell's law

Principle

Snell's law calculates the refraction of light when traveling between two media of differing refractive index n.

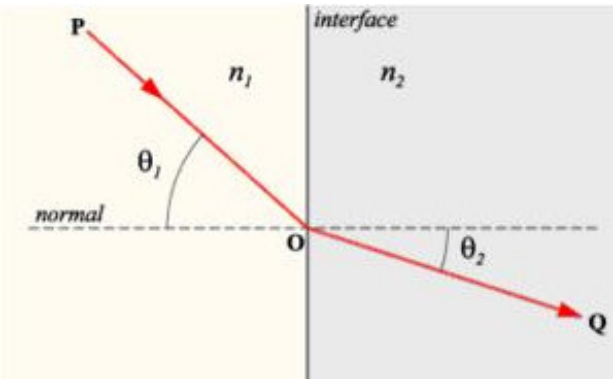


Fig. 1 Refraction of light at the interface between two media of different refractive indices with $n_2 > n_1$.

A ray entering a medium perpendicular to the surface is never bent.
Snell's law gives the relation between the angles θ_1 and θ_2:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2. \quad (1)$$

A qualitative rule for determining the direction of refraction is that the ray in the denser medium is always closer to the normal. A handy way to remember this is to visualize the ray as a car crossing the boundary between asphalt (the less dense medium) and mud (the denser medium). The car will swerve in the direction of that front wheel that first becomes into the mud.
Snell's law may be derived from *Fermat's principle*, which states that the light travels the path which takes the least time. In a classic analogy, the area of lower refractive index is replaced by a beach, the area of higher refractive index by the sea, and the fastest way for a rescuer on the beach to get to a drowning person in the sea is to run along a path that follows Snell's law.

Snell's law is only generally true for *isotropic* media (see **More Info**) such as glass and water.

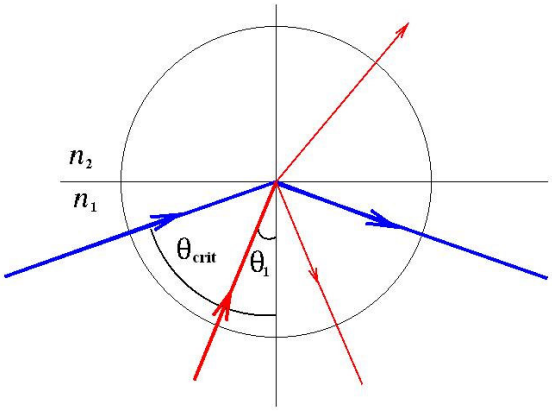


Fig. 2 Total internal reflection

Total internal reflection When moving from a dense to a less dense medium (i.e. $n_1 > n_2$), it is easily verified that the above equation (1) has no solution when θ_1 exceeds a value known as the critical angle:

$$\theta_{crit} = \arcsin (n_2/n_1) \quad (2)$$

When $\theta_1 > \theta_{crit}$, no refracted ray appears, and the incident ray undergoes *total internal reflection* from the interface (Fig. 2).
See for a more detailed description of refraction Light: refraction and for reflection Light: reflection.

Application

The law is applied innumerous in (medical and daily life) optical apparatus, like glasses, binoculars, microscopes etc.
Birefringence (see below) is widely used in optical devices, such as LCDs (liquid crystal displays) in notebooks, light modulators, color filters, wave plates, optical axis gratings, etc. It is also utilized in *medical diagnostics*. Needle biopsies of suspected gouty joints will be negatively birefringent if urate crystals are present.
In situ amyloidal deposits in Alzheimer's disease plaques also show birefringence.

More info

A medium is *anisotropic* when n is directionally dependent, such as occurs in some crystals. This causes birefringence, or double refraction. This is the decomposition of a ray of light into two rays (the *ordinary ray* and the *extraordinary ray*), depending on the polarization of the light (see Light: polarization). The plane of polarization of the ordinary and *extraordinary* or *e*-ray may are generally perpendicular to ach other. With a single axis of anisotropy, (i.e. it is uniaxial) birefringence can be formalized by assigning two different n's to the material for different polarizations. The birefringence magnitude is then defined by:

$$\Delta n = n_e - n_o, \qquad (3)$$

where n_o and n_e are the refractive indices for polarizations perpendicular (ordinary) and parallel (extraordinary) to the axis of anisotropy respectively. Fig. 3 gives an example.

Fig. 3 A calcite crystal laid upon a paper with some letters showing the double refraction.

Biaxial birefringence, also known as trirefringence describes an anisotropic material that has more than one axis of anisotropy

Radiation

Angiography and DSA

Principle

Angiography or arteriography is an imaging technique to visualize the inner volume of blood filled structures, including arteries, veins and the heart chambers. The X-ray image or video of blood vessels is called an angiograph, or more commonly, an angiogram.
Angiograms require the insertion of a catheter into a peripheral artery, e.g. the femoral artery.
As blood has the same radiodensity (see CT scan (dual energy)) as the surrounding tissues, a radiocontrast agent (which absorbs X-rays, see Spectroscopy) is added to the blood to make visualization possible. The image shows shadows of the inside of the vascular structures carrying blood with the radiocontrast agent. The tissue of the vessels or heart chambers themselves remain largely to totally invisible.
The images may be taken as either still images, displayed on a fluoroscope (see Fluoroscopy and Fluorescence) or film, useful for mapping an area. Alternatively, a video can be made, usually at 30 frames/s, which also show the speed of blood (i.e. the speed of radiocontrast).

Digital subtraction angiography (DSA) is a method to visualize blood vessels with contrast medium in a bony environment by subtracting the pre-contrast image (the mask) from the image with contrast medium (see **More Info**).
Intravenous DSA (IV-DSA) compares an X-ray image before and after a radiopaque iodine based dye has been injected intravenously. (Radiopacity is the ability of electromagnetic radiation to pass through a particular material.) Tissues and blood vessels on the first image are digitally subtracted from the second image, leaving a clear picture of the artery which can then be studied.

Application

With a catheter (in the groin or forearm) the radiocontrast agent is administrated at the desired area, allowing short lasting images of the inner size of the arteries. Presence or absence of atherosclerosis or atheroma within the walls of the arteries cannot be clearly determined. Angiogram of the coronary arteries is very common.
Angiography is also performed to identify vessel narrowing in patients with retinal vascular disorders, e.g. macular degeneration and diabetic retinopathy.
Other common clinical applications consider the cerebrum, extremities, lungs, liver, kidneys and the lymphatic system.

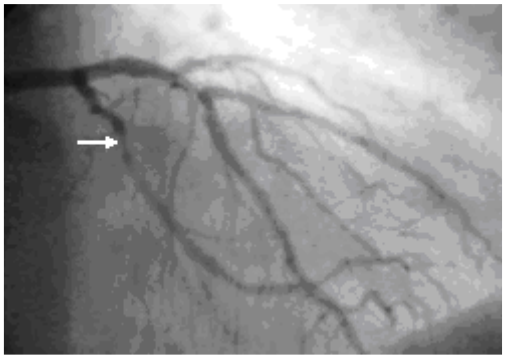


Fig. 1 Coronary angiogram with a stenosis (arrow) in the left branch (modified from Wikipedia/Angiography, digital subtraction).

DSA is useful in diagnosing arterial occlusion, including pulmonary artery thrombosis, carotid artery stenosis (Fig. 1) and in detecting renal vascular disease. After injecting contrast material in a vessel, fluoroscopic images are produced.

IV-DSA can be used for mapping cerebral blood flow and studying the vessels of the brain and heart, detecting carotid artery obstruction and lesions. IV-DSA has also been useful after coronary artery bypass surgery, in transplant operations and in assessing patients prior to surgery. However, IV-DSA is unsuitable for patients with diabetes or renal insufficiency (due to the high dose of the contrast dye).

More Info

The subtraction technique
By injecting contrast medium into the blood vessels the images thus obtained also show other structures besides blood vessels. In order to remove these distracting structures, we need to acquire a reference or mask image (image of the same area without contrast administration) to be *subtraction* from the contrast image. Now, clear pictures of blood vessels are obtained and a near-instantaneous film shows the blood vessels alone.
Nowadays, DSA is increasingly more replaced by CTA, Computed Tomographic Angiography, which produces 3D images and is less invasive.

Bioelectromagnetics

Principle

Bioelectromagnetics is the study of how electromagnetic fields interact with and influence biological processes; almost the same as radiobiology of non-ionizing radiation. Common areas of investigation include the mechanism of animal navigation and migration using the geomagnetic field, studying the potential effects of man-made sources of electromagnetic fields, such as those produced by the power distribution system and mobile phones (see Mobile phones radiation hazards), and developing novel therapies to treat various disorders.

Thermal versus nonthermal nature
Most of the molecules that make up the human body interact only weakly with electromagnetic fields that are in the radiofrequency or extremely low frequency bands. One basic interaction is the absorption of energy from the electromagnetic fields, which can cause tissue to heat up; more intense field exposures will produce greater heating. This heat deposition can lead to biological effects ranging from discomfort to protein denaturation and to burns. Many nations and regulatory bodies have established safety guidelines to limit the electromagnetic fields exposure to a non-thermal level (see Magnetobiology), which can either be defined as heating only to the point where the excess heat can be dissipated/radiated away, or as some small temperature increase(<0.1 ℃) that is difficult detectable with standard instruments. However, in thermotherapy (see **Application**) local heating is the aim of the therapy.
Biological effects of weak electromagnetic fields are the subject of study in Magnetobiology.
Bioelectromagnetics is not to be confused with bioelectromagnetism (also simply called Bioelectricity), which deals with the ability of living creatures to produce its own electromagnetism.

Application

While several treatments based on the use of magnetic fields have been reported, the only ones that have been approved by the US FDA are the use of pulsed magnetic fields to aid non-union bone fractures.

Transcranial magnetic stimulation (TMS)
A well known medical application is the various types of TMS (Transcranial magnetic stimulation) which attempts to affect brain functioning in psychiatric patients (depression, epilepsy) and consequently affect behavior.
TMS is an approved technique for transfering power and meaningful (e.g. sensory) signals from an outsite small box with a coil to an implanted secondary coil mounted in a processing unit. It is the common technique for commercial cochlear implants and experimental retinal and cortical implants to aid vision.

Thermotherapy
A completely different application is thermotherapy, the induction of local hyperthermia with temperatures > 45 ^{o}C. It is used to kill or weaken tumor cells with negligible effects on healthy cells. Tumor cells, with a disorganized and compact vascular structure poorly dissipate heat and hence are heated. As a result hyperthermia may cause cancerous cells to undergo apoptosis, while healthy cells can more easily maintain a normal temperature. Even if the cancerous cells do not die outright, they may become more susceptible to ionizing radiation treatments or to certain chemotherapies, allowing such therapy to be given in smaller doses. It is often used for prostate cancer. A non-cancerous application is pain relief, relaxation stiff muscles, rheumatoid arthritis and treatment of atrial fibrillation. Penetration depth are a few mm to a few cm, depending on the applied frequencies, which are from tens of MHz up to about 6 GHz (the higher the frequency, the smaller the depth).
High-energy thermotherapy applies temperatures > 60 ^{o}C (ref. 1). The radiation is delivered by a probe with a microwave power of some 40 W, a penetration depth of 2-5 mm and a delivery time of some 10-40 s. In prostate cancer treatment, the urethra is cooled at the same time to prevent damage.

Reference
1. Koliosyx MC, A Worthingtony AE, et al., Experimental evaluation of two simple thermal models using transient temperature analysis. Phys. Med. Biol. 43 (1998) 3325–3340.

CT scan (dual energy)

Principle

.

Fig. 1 Philips MX 6000 Dual CT apparatus with 0.8 mm slices and 0.8 s rotation speed (From ref.))

Computed tomography (CT), originally known as computed axial tomography (CAT or CT scan) and body section röntgenography (see X-ray machine), is a method (techniques and processes used to create images of parts of the human body for clinical purposes). It employs tomography (imaging by sections) where digital geometry processing (design of algorithms for 3D-modeling) is used to generate a 3D image of the internals of an object from a large series of 2D X-ray (electromagnetic radiation with 10-0.01 nm wavelengths, see also Spectroscopy) scans. The data of CT scans taken around a single axis of rotation can be manipulated by a process known as windowing (see Imaging: windowing). Standard CT generates images in the axial or transverse plane which can be reformatted in various planes and seen in (3D).

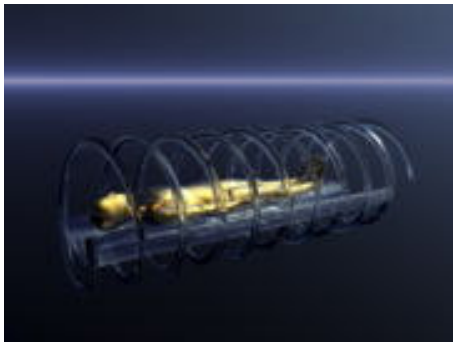


Fig. 2 Principle of spiral CT (From Wikipedia)

The principle of operation is an X-ray source that rotates around the object with X-ray sensors positioned on the opposite side of the circle from the X-ray source. X-ray slice scans are taken one after the other as the object is gradually passed through the gantry. The slice data are combined by a mathematical procedure known as tomographic reconstruction, for instance to generate 3D information (3D-CT scan). The images are viewable from multiple different perspectives on CT workstation monitors. This can also be done with spiral CT machines, which generate high-resolution 3D images from the data of the moving individual slices.
Dual energy CT machines use two X-ray frequencies, so combine the two images. A recent (experimental) application is spectral CT, where a broadband X-ray spectrum and a multi-energy photon counting detection systems is used (see ref. 1).

Application

Although most common in medicine, CT is also used in for example in experimental earth sciences, material sciences and archeology.

CT has become an important tool in medical imaging to supplement radiography (see X-ray machine) and ultrasonography or Echography. It is also used in preventive medicine or screening. Some important fields of application are described below.

Neurology CT is frequently used in neurology and related disciplines: cerebrovascular accidents, intracranial hemorrhage (stroke). increased *intracranial pressure*, facial and skull *fractures*, surgical planning for craniofacial and dentofacial deformities, ophthalmology (orbita), ear, trough and nose medicine (ETN), brain *tumors* with IV contrast (but less sensitive than MRI).
Pulmonology Non-contrast CT scans are excellent for acute and chronic changes in the lung parenchyma (pneumonia, cancer). Thin section CT scans are used for emphysema, fibrosis, etc., IV-contrast for the mediastinum and hilar regions. CT angiography (CTPA) is applied for pulmonary embolism and aortic dissection (helical scanners).
Cardiology Dual Source CT scanners and high resolution (multi-slice) and high speed (subsecond rotation) CT scanners are used for imaging of the coronary arteries.
Abdominal and pelvic region Cancer, acute abdominal pain, organ injury, disorders resulting in morphological changes of internal organs, often with contrast ($BaSO_4$ for Fluoroscopy, (see also Fluorescence), iodinated contrast for pelvic fractures).
Orthopedics Further CT is useful for (complicated) fractures in the *extremities*.
CT is inadequate for osteoporosis (radiation doses, costs) compared to DXA scanning for assessing bone mineral density (BMD), which is used to indicate bone strength.

More Info

Working principle

In conventional CT machines, an X-ray tube (see X-ray machine) and detector are physically rotated behind a circular shroud. In electron beam tomography (EBT) the tube is far larger and higher power to support the high temporal resolution. The electron beam is deflected in a hollow funnel shaped vacuum chamber. X-rays are generated when the beam hits the stationary target. The detector is also stationary. Contrast materials are used to highlight structures such as blood vessels and to obtain functional information.
The data stream, representing the varying radiographic sensed intensity, reaches the detectors on the opposite side of the circle during each sweep. Then, it is computer-processed to calculate cross-sectional estimations of the radiographic density, expressed in Hounsfield units (HU).

Hounsfield unit
This unit is defined as:

$$HU_{material} = 1000(\mu_{material} - \mu_{water})/\mu_{water},$$

where μ is the attenuation coefficient, which is analogue to the parameter A (absorbance or extinction) of the Lambert-Beer law. HU is dependent of the beam intensity of the X-ray source (mostly given in kilo-electron volt, keV).
Since μ_{water} is by definition zero, the *radiodensity* of distilled water at STPD conditions (see Gas volume units, STPD, BTPS and ATPS) is defined as 0 HU. At 80 keV, that of air at STPD is ca.–1000 HU, cancellous bone amounts to 400 HU and cranial bone to 2000 HU. The attenuation of metallic implants (dental, extremities) depends on the element's atomic number. Titanium has ca. 9000 HU (at 80 keV) and iron steel ca. 24500, which can completely extinguish the X-ray and is therefore responsible for well-known line-artifacts in computed tomograms.
Pixels in an image obtained by CT scanning are displayed in terms of HUs (from –1024 to +3071). As in all 3 imaging techniques the CT slice thickness determines the voxel (cubical pixel), the volumetric unit. The phenomenon that one part of the detector cannot differ between different tissues is called the Partial Volume Effect. That means that a big amount of cartilage and a thin layer of compact bone can cause the same attenuation in a voxel as hyperdense cartilage alone. New techniques (dual and spectral energy CT) is the answer to this problem.
Dual source (or energy) scanning offers the potential of differentiating materials beyond the direct visualization of morphology – for example, direct subtraction of either vessels or bone during scanning. Dual Source CT scanners, allow higher temporal resolution so reduce motion blurring at high heart rates, and potentially allow a shorter breath-hold time. The 256-slice MSCT (multi-slice CT) improved cardiac scanning performance.

Advantages and hazards

Advantages over Projection Radiography (see Radiography)
- CT completely eliminates the superposition of images of structures outside the area of interest.

- Because of the inherent high-contrast resolution of CT, differences between tissues that differ in physical density by less than 1% can be distinguished.
- Data from a single CT imaging procedure consisting of either multiple contiguous or one helical scan can be viewed as images in the axial, coronal, or sagittal planes. This is referred to as multiplanar reformatted imaging.

Radiation exposure
CT is regarded as a moderate to high radiation diagnostic technique. Unfortunately the newer CT technology requires higher doses for better resolution. For instance, CT angiography avoids the invasive insertion of an arterial catheter and guide wire and CT colonography may be as good as barium enema (see X-ray machine) for detection of tumors in the large intestines, but at the cost of a higher dose. Cardiac MSCT is equivalent of 500 chest X-rays in terms of radiation. The risk on breast cancer is not well established. The positive predictive value is approximately 82% while the negative prediction is ca. 93%. Sensitivity is ca. 80% and the specificity is about 95%. The real benefit in the test is the high negative predictive value.
The radiation dose for a particular study depends on multiple factors: volume scanned, patient build, number and type of scan sequences, and desired image quality and resolution.

Table 1 Typical CT scan doses

Examination	**Typical effective dose (mSv*)**
Head	1.5
Chest	6
Cardiac angiogram	7 - 13
Abdomen, colon	5 - 9
colongraphy	4 - 9

*The sievert (Sv) has the same dimensions as the gray (i.e. 1 Sv = 1 J/kg = 1 $m^2 \cdot s^{-2}$), but the former measures the biological effect and the later the radiated dose. For comparison: a Chest X-ray uses 0.02 mSv.

Because CT scans often rely on IV-contrast agents, there is a low but non-negligible level of risk associated with the contrast agents themselves, like (life-threatening) allergic reactions to the contrast dye (e.g. causing kidney damage).

References
1. http://www.medical.philips.com/in/products/ct/products/spectral_ct/

Diaphanography and optical mammography

Principle

Diaphanography or transillumination is a method to detect breast tumors by compressing the breast between glass plates and using red light to shine through the tissue (see Fig. 3a in **More Info**). Nowadays, diaphanography or better, optical mammography is based on ultra-short (10^{-12} s) Laser pulses, which are used as an alternative to traditional X-ray.

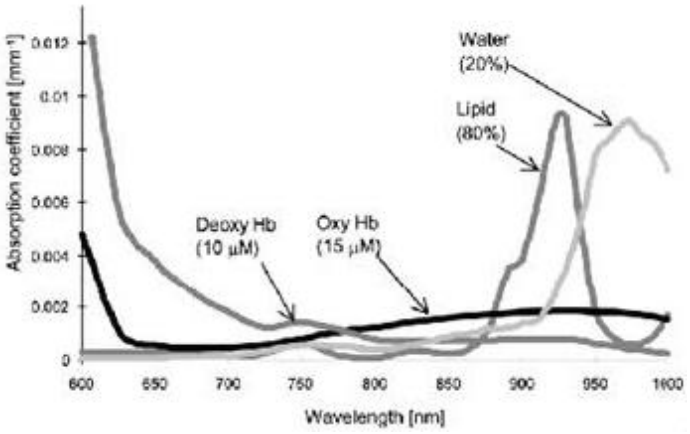


Fig. 1 Absorption coefficients (see Lambert-Beer law) of hemoglobin (Hb) (see also Pulse oximetry), water and fat (lipid) at concentrations typical for female breasts.

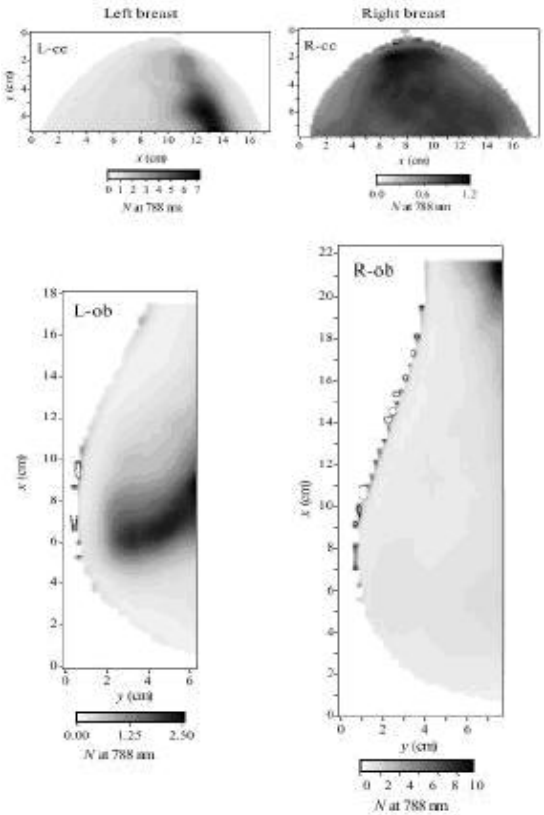


Fig. 2 Images obtained with 788 nm. These mammograms refer to a 58 year old woman affected by breast cancer (tumor size: 30 mm) in the left breast.

Tissue optical properties

The main absorbers in breasts at visible and near-IR light are hemoglobin, water and fat (see for absorption spectra Fig. 1). The 650-950 nm region is most suited for deep-structure detection, and the penetration depth is a few cm. The scattering is caused by small differences in refractive index (see Snell's Law and Light: refraction) at the microscopic level, mainly of fat droplets and structures inside the cell nucleus. The scattering increases for shorter wavelengths (λ), approximately obeying the λ law of Mie scattering theory (see Light: scattering). The scattering, however, is strong even for λ > 1µm.

The main advantage of using short-pulse lasers for deep-structure is that by using time-resolved detection it is possible to select light, which arrives at the detector at different time slots. The light that arrives early has traveled a shorter and straighter path than the late light. The object of the new generation of optical tomography is to reconstruct an image of both the absorption (direct pathway) and the scattering properties inside the tissue, based on measurements at several light-source and detector position on the skin. Actually, the whole shape of the detected pulses can be of use.

Oxygen as a key
On the basis of 3 different wavelengths the kind of tissues can be detected. A tumor wants to grow and therefore induces growth of a lot of little blood vessels around itself, (particularly an invasive one). But vessels growth is delayed and therefore the tumor gets more oxygen out of the blood than normally. So the volume of blood increase, but the oxygenation level of the blood goes down. By detecting both at the same time one can distinguish between benign and malignant tumors.

Application

Diaphanography has been developed to overcome the draw back of missing tumors in classical screening mammography due to a too low resolution. Fig. 2 shows an example of a scan.
Optical mammography is not yet full-grown commercial and in many western countries X-ray screening is still the practice, the more since today doses are very low and seldom causes the metastasizing of an existing cancer. Despite this, nowadays, retrorespective detection rate is about 90% of histological validated malign tumors.

More Info

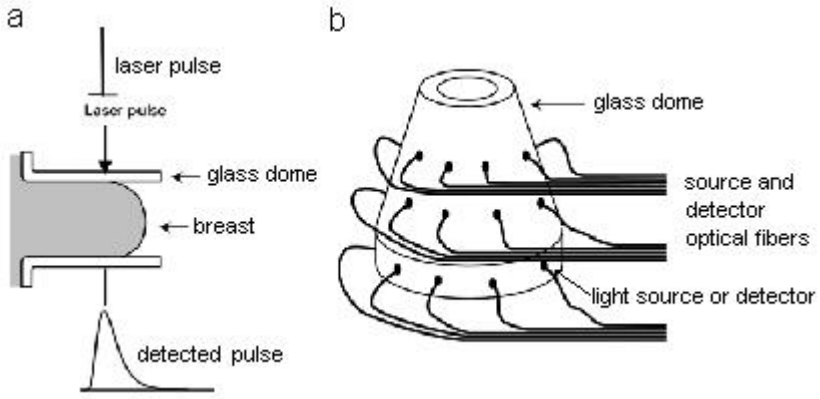


Fig. 3 a Principle of operation. b Principle of the design of the scanner.

In addition to the use of the whole detected light pulse (see Fig. 3a), the lasers may be chosen at λ's that enhance the contrast between the tumor and surrounding tissue, as in classical spectroscopy, where different compositions of the absorbers give rise to different absorption at various wavelengths. The scattering can also vary between tissue types and wavelengths, which can also be used for contrast enhancement. The choice of wavelengths and knowledge of tissue composition is crucial.

Pulsed laser diodes with so called time-correlated single-photon-counting detection are used to get time-resolved measurements with a sub-ns resolution. The small light sources and detector are placed in a conical configuration, more resembling the conical geometry of conventional tomography (Fig. 3b). The breast is compressed, and the source-detector optical fibers are scanned, for a series of point-wise measurements. Acquiring data for a whole breast with about 1000 points takes a few minutes. Because of this relatively long time, the breast is not compressed as hard as for X-ray mammography.

Finding the best geometry is related to the second problem, development of reconstruction algorithms. For example, should the breast be compressed or not, and it is possible to get better results using fluids

to match the optical properties between the skin and the detectors, are some of the questions under investigation.

The principle of the reconstruction algorithm is to divide the problem into two parts, the forward and the inverse problem. The *forward problem* deals with computation of what the detected signal would be, given that the absorption and scattering are known. The *inverse problem* is the most likely anatomy which matches best the signals measured with the detectors. Since the problem is highly non-linear, the reconstruction is based on iterations. The solution strongly depends on contrast function that discriminates the tumors. The relative compositions of water, fat and hemoglobin vary not only between tumors and healthy tissue, but also with for example age and hormonal cycles. Furthermore, there is not a single type of breast cancer. Tumors vary a lot in composition and structure. All these parameters have to be understood and quantified.

Literature

Dehghani H. et al. Multiwavelength Three-Dimensional Near-Infrared Tomography of the Breast: Initial Simulation, Phantom, and Clinical Results, Applied Optics, 2003, 42, 135-145.
Paola Taroni P. et al. Time-resolved optical mammography between 637 and 985 nm: clinical study on the detection and identification of breast lesions Phys. Med. Biol. 2005, 50, 2469-2488.
Rinneberg H. et al. Scanning time-domain optical mammography: detection and characterization of breast tumors in vivo. Technol Cancer Res Treat. 2005, 4, 483-96.

Electron diffraction

Principle

Electron diffraction is the basic principle of electron microscopy (EM). Its theory is similar as that of diffraction of light, where the wavelength is also the key-parameter.
Since electrons are charged particles they interact with matter through Coulomb forces. (The magnitude of the force on a charge q_1, due to the presence of a second charge q_1 at distance r, is given by Coulomb's law. Its scalar version is:

$$F = (4\pi\varepsilon_0)^{-1} q_1 q_2 r^{-2}, \quad (1)$$

where ε_0 the dielectric constant in vacuum, being $9.0 \cdot 10^9$ Nm^2/F^{-2}). This means that the incident electrons are influenced by both the positively charged atomic nuclei and the surrounding electrons. In comparison, X-rays interact with the spatial distribution of the valence electrons. In contrast, proton beams hardly interact with atoms. They hardly scatter. Because of these different forms of interaction, these three types of radiation are suitable for different application: EM, X-ray diagnostics and therapy, and proto therapy.

Electron diffraction techniques are increasingly important for life sciences and medicine since nowadays even single crystals can be studies. After diffraction by e.g. a crystal an interference pattern results. This phenomenon occurs due to the wave-particle duality, which states that a particle, in this case the incident electron, can be described as a wave. Consequently, this technique is similar to X-ray or light diffraction (see Light: diffraction).
Crystallographic experiments with electron diffraction are usually performed in TEM or SEM as so called electron backscatter diffraction (a microstructural-crystallographic technique).
The periodic structure of a crystalline solid acts as a diffraction grating scattering the electrons in a predictable manner. In principle, from the observed diffraction pattern (Fig. 2 of Electron microscopy), it is possible to calulate the structure of the crystal. Although this technique is limited by the so called phase problem (the loss of phase informatio) several techniques can reconstruct the phase in order to obtain an image.

Application

Electron diffraction is widely used in solid state physics and chemistry, in particular crystallography. Since many proteins behave as crystalline structures (e.g. in viruses), they are frequently studied with electron diffraction. And since diffraction is the basic principle of EM, implicitly it has a wide filed of applications in life sciences and experimental medicine.

More info

Intensity of diffracted beams
In the kinematical approximation of electron diffraction, the intensity of a diffracted beam is given by:

$$I_{\mathbf{g}} = |\psi_g|^2 \propto |F_g|^2 \quad (2)$$

Here ψ_g is the wave function of the diffracted beam and $F_{\mathbf{g}}$ is the so called structure factor of crystallography, which comprises the scattering vector of the diffracted beam, the scattering powers of the atoms (also called the atomic form factor) and the atomic positions. More explanation can be found in specialisic texbooks.
Wavelength of electrons
The wavelength of any particle given by the de Broglie equation:

$$\lambda = h/p. \quad (3)$$

Here h is Planck's constant and p the momentum (here electron mass m_0 times velocity v) of the electron. The electrons are accelerated with an electric potential U to the desired velocity:

$$v = (2eU/m_0)^{0.5}. \quad (4)$$

with e is the elementary charge. Hence, $p = (2eUm_0)^{0.5}$ and so λ is then given by:

$$\lambda = h/(2m_0eU)^{0.5}. \quad (5)$$

In EM, U is usually several kV causing the electron to travel at an appreciable fraction of the speed of light (c). A scanning EM (SEM, see Electron microscopy) may typically operate at 10 kV, giving an electron velocity approximately 20% of c. A typical transmission EM (TEM, see Electron microscopy: transmission EM) can operate at 200 kV raising the velocity to 70% of c, which needs a relativistic correction:

$$\lambda = h/(2m_0eU)^{0.5}(1 + eU/2m_0c^2)^{-0.5} \quad (6)$$

The last term is the relativistic correction factor that reduces λ.
The wavelength of the electrons in a 10 kV SEM is then 12.3×10^{-12} m (= 12.3 pm), while in a 200 kV TEM λ is 2.5 pm. This suggests a $2.5 \cdot 10^5$ better resolution than the light microscope (ca. 0.6 µm). However, due to electromagnetic lens errors the resolution is much less. With apertures in the mm-range to reduce the errors, a factor 1000 is possible, which gives a best resolution of about 0.5 nm. The wavelength of X-rays usually used in X-ray diffraction is 10 -100 pm, yielding slightly worse resolutions.

Electron diffraction in a TEM
There are several specific techniques of electron diffraction in TEM. For instance, by performing a diffraction experiment over several incident angles simultaneously, a technique called Convergent Beam Electron Diffraction (CBED), one can reveal the full 3D-structure of a crystal.
A technique to study electronic structure and bonding is so called electron energy loss spectroscopy (EELS). In EELS a material is exposed to a beam of electrons with a known, narrow range of kinetic energies. Finally one can study the inner potential through electron holography (see Holography).

Limitations
Electron diffraction in TEM is subject to several important limitations. First, sample thickness must be of the order of 100 nm or less to be electron transparent. Furthermore, many biological samples are vulnerable to radiation damage caused by the incident electrons.
The main limitation of electron diffraction in TEM remains the comparatively high level of practical skill and theoretical knowledge of the investigator. This is in contrast to e.g. the execution of powder X-ray (and neutron) diffraction experiments with its data analysis being highly automated and routinely performed.

Electron microscopy

Principle

Electron microscopy (EM) uses electrons as a way to create a visible image of a specimen. It has much higher angular magnification and resolving power (i.e. angular separation) than light microscopy, with magnifications up to about two million times, thousand times more than with visible light.
Unlike light microscopy EM uses electrostatic and electromagnetic lenses to control the “illumination” and imaging of the specimen.
Basically for the image formation is the diffraction of the electrons. To reach a good resolution the wavelength should be small. See Electron diffraction for electron diffraction theory, beam intensity and wavelength calculations.

Transmission electron microscopy (TEM)
TEM is the original form of electron microscopy. TEM involves a high voltage electron beam emitted by a cathode, usually a tungsten (W) filament and focused by electrostatic and electromagnetic lenses. The electron beam is transmitted through a specimen that is partly transparent to electrons. As the electrons pass through the sample, they are scattered by the electrostatic potential caused by the constituent elements of the molecules. After leaving the sample the electrons pass through the electromagnetic objective lens (Fig. 1). This lens focuses all electrons scattered from one point of the sample in one point on a fluorescent screen, a photographic plate, or a light sensitive sensor such as a CCD camera, constructing an image of the sample. The image detected by the CCD camera may be displayed in real time on a monitor or computer. At the dashed line in the Fig. 1, electrons scattered in the same direction by the sample are collected into a single point. This is the back focal plane, the plane where the diffraction pattern is formed. By manipulating the magnetic lenses of the microscope, the diffraction pattern may be observed by projecting it onto the screen instead of the image. This pattern is used for structural studies in crystals (see Electron diffraction).

See for a further description Electron Microscopy: transmission EM.

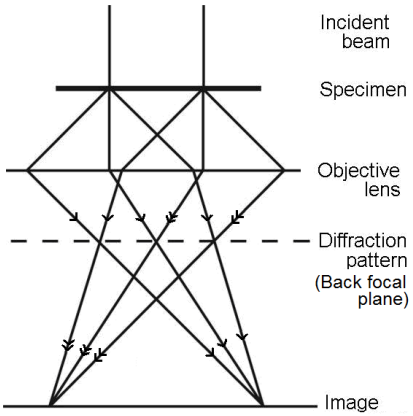


Fig. 1. Principle of the path of a parallel beam of electrons in a TEM from just above the sample up to the fluorescent screen.

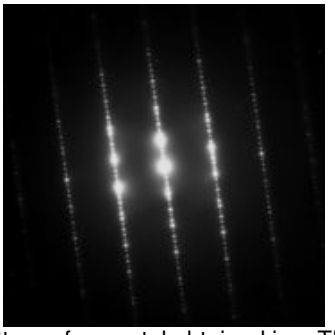


Fig. 2 Typical electron diffraction pattern of a crystal obtained in a TEM with a parallel incident electron beam.

Application

Except for all kind of industrial, technical and science applications, there are numerous ones in life sciences and experimental medicine such as cryobiology, electron and cellular tomography, 3D-tissue imaging, forensic medicine, toxicology, virology, microbiology and protein localization.

More info

In addition to the oldest EM technique, TEM, there exist various other EM techniques.

Scanning Electron Microscopy (SEM)
SEM produces images by detecting low energy secondary electrons which are emitted from the surface of the specimen due to excitation by the primary electron beam. In the SEM, the electron beam is scanned across the sample, with detectors building up an image by mapping the detected signals with beam positioning.
Generally, the TEM resolution is about an order of magnitude greater than SEM resolution, because the SEM image relies on surface processes rather than transmission. It is well able to image bulk samples and it has a much greater depth of view, and so can produce images that are a good representation of the 3D-structure.

Reflection Electron Microscopy (REM)
Like TEM, REM involves electron beams incident on a surface, but does not use transmission or secondary emission but reflected electrons. This technique is typically coupled with Reflection high-energy electron diffraction (RHEED) to study especially crystalline surfaces and with Reflection high-energy loss spectrum (RHELS) technique.

Scanning Transmission Electron Microscopy (STEM)

The STEM technique combines the high resolution of TEM with the sample scanning of SEM by using a small beam, focused on the sample. It allows a range of analytical techniques not possible with conventional TEM.

Sample preparation
For conventional SEM, samples must be electrically conductive, at least at the surface, and electrically grounded to prevent the accumulation of electrostatic charge at the surface. For sample preparation various steps and techniques are needed, depending on the sample and the analysis required:
Cryofixation – freezing a specimen so rapidly in liquid N_2 or He that the water forms vitreous (non-crystalline) ice. This preserves the specimen in a snapshot of its natural state. This technique has evolved to electron cryomicroscopy. It is also called cryo-EM, although it is the microscope and specimen and not the electrons that are cold. So, the vitreous sections of sample that are studied are at cryogenic temperatures (mostly liquid N_2).
Dehydration – replacing water with organic solvents such as ethanol and acetone.
Embedding – infiltration of the tissue with a resin (araldite, epoxy) and polished to a mirror-like finish using ultra-fine abrasives.
Sectioning –with an ultramicrotome with a diamond knife or glass knife (2^{nd} choice) to produce slices of about 100 nm.
Staining – uses heavy metals such as Pb, Ur, W, to scatter imaging electrons and thus give contrast between different structures, since many (especially biological) materials are nearly "transparent" to electrons.
Freeze-fracture or freeze-etch – a method particularly useful for examining lipid membranes and their incorporated proteins in "face on" view. The fresh tissue or cell suspension is cryofixed, then fractured by simply breaking or by using a microtome while maintained at liquid N_2 temperature. The cold fractured surface (sometimes "etched" by increasing the temperature to about -100°C for several minutes to let some ice sublime) is then shadowed with evaporated Pt or Au at an angle of 45° in a high vacuum evaporator. A second coating of carbon is often performed to improve stability of the replica coating. The specimen is returned to room temperature and pressure, then the extremely fragile "pre-shadowed" metal replica of the fracture surface is released from the underlying biological material by careful chemical digestion with acids, hypochlorite solution or sodium dodecyl sulfate detergent. The still-floating replica is washed from residual chemicals, fished up on EM grids, dried and then viewed in the TEM.
Ion Beam Milling – thins samples until they are transparent to electrons by firing ions (typically Ar containing ions) at the surface from an angle and sputtering material from the surface. A subclass of this is focused ion beam milling, where Ga ions are used to produce an electron transparent membrane in a specific region of the sample, for example through a device within a microprocessor. Ion beam milling may also be used for cross-section polishing prior to SEM analysis of materials that are difficult to prepare using mechanical polishing.
Conductive Coating – An ultrathin coating of electrically-conducting material (by high vacuum evaporation or by low vacuum sputter) to prevent the accumulation of static electric fields at the specimen due to the electron irradiation required during imaging. Such coatings include AU, AU/Pa, Pt, W, graphite etc., especially important for SEM. Os yields the thinnest coating. Another reason for coating, even when there is more than enough conductivity, is to improve contrast, a situation more common with the operation of FESEM (field emission SEM).

Electron microscopes requiring extremely stable high-voltage supplies and currents to each electromagnetic lens, continuously-pumped (ultra-)high-vacuum systems (up to 10^{-9} Pa), and a cooling water supply circulation through the lenses and pumps. As they are very sensitive to vibration and external magnetic fields, EMs aimed at achieving high resolutions must be housed in buildings (sometimes underground). Newer generations of TEM operating at lower voltages (around 5 kV) do not have stringent voltage supply, lens coil current, cooling water or vibration isolation requirements and as such are much less expensive to buy and far easier to install and maintain.
The samples have to be viewed in vacuum, as the molecules that make up air would scatter the electrons. Recent advances have allowed hydrated samples to be imaged using a so called environmental SEM (observed in low-pressure, 1 – 100 mbar). SEM usually image conductive or semi-conductive materials best. Non-conductive materials can be imaged by an environmental SEM.
To be sure that no artifacts corrupt the sample, various EM techniques are compared with the same sample. Also, high resolution X-ray crystallography is used for comparison.

Electron microscopy: transmission EM

Principle

Transmission electron microscopy (TEM) is an imaging technique whereby a beam of electronsnafter transmission through a sample forms an image (see Electron microscopy). Fig. 1 compares the basic pathways of radiation in TEM and light microscopy and in TEM.

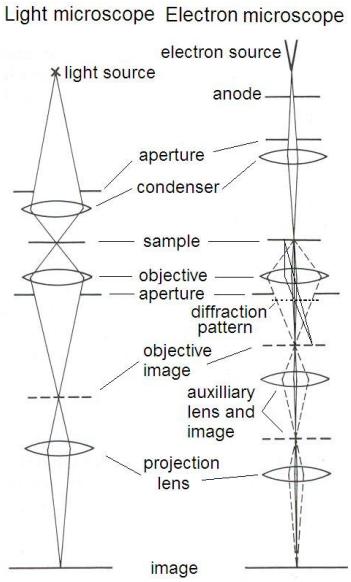


Fig. 1 Comparison of optical microscopy and TEM.

Formely, with microscopy nothing smaller than the wavelength being used could be resolved, whereas nowadays subdiffraction techniques (see Optical microscopy: super-resolution techniques) and the method of RESOLFT (REversible Saturable Optical Fluorescence Transitions) sets the limit for light microscopy. (The RESOLFT concept increases the classical resolution (see Optical microscopy) by raising the beam intensity.)
The wave-particle nature of all matter allows considering the electron beam as a beam of electromagnetic radiation. Its wavelength is dependent on the kinetic energy (equations 3-6 of Electron diffraction), and so can be tuned by adjustment of the acceleration of the electrons. The wavelength can be much smaller than that of light, yet the electrons can still heavily interact with the sample due to their electrical charge.
After emitting and accelerating the electrons, they are focused by electrical and magnetic fields (lenses) onto the sample. Contrast can be enhanced by “staining” the sample with heavy metals such as Os, Pb, U. The dense nuclei of the atoms scatter the electrons out of the path. The areas where electrons are scattered appear dark on the screen and the electrons that remain generate the image.

Resolution of the TEM is limited primarily by spherical aberration (see Light: the ideal and non-ideal lens), but a new generation of aberration correctors reduce the amount of distortion in the image and consequently this limitation.
A monochromator may also be used which reduce the energy spread of the incident electron beam to less than 0.15 eV. (A monochromator transmits a narrow band of wavelengths of radiation chosen from a wider range of wavelengths available at the input.)

Application

TEM is used frequently in material science/metallurgy and biological sciences. For biological specimens, the maximum thickness is roughly 1 µm, but ideally some 0.1 µm. To withstand the instrument vacuum, biological specimens are typically held at liquid N_2 temperatures after embedding in vitreous ice, or fixated using a negative staining material such as uranyl acetate or by plastic embedding. Typical biological applications include tomographic cell reconstructions and 3D reconstructions of individual molecules via Single Particle Reconstruction. This is a technique in which large numbers of images (10,000 - 1,000,000) of ostensibly identical individual molecules or macromolecular assemblies are combined to produce a 3D reconstruction.

More Info

Electrons are generated by a process called thermionic emission. This occurs with a cathode or field emission (a form of quantum tunneling in which electrons pass through a barrier in the presence of a high electric field.) Thermionic emission is the flow of charge carriers (electrons or ions), from a surface or over some other kind of electrical potential barrier, caused by thermal vibrational energy overcoming the electrostatic forces restraining the charge carriers. (A hot metal cathode emitting into a vacuum is the classical example (the Edison effect).

High Resolution TEM
With the most powerful diffraction contrast, TEMs equipped with aberration correctors can elucidate, crystal structure. This can also by performed with High Resolution TEM (HRTEM), also known as phase contrast imaging as the images are formed due to differences in phase of electron waves scattered through a thin specimen. (See for the phase contrast principle Optical microscopy: phase contrast.)
Software correction of spherical aberration for the HRTEM has allowed the production of images with sufficient resolution to show carbon atoms in diamond separated by only 0.089 nm. The ability to determine the positions of atoms within materials has made the HRTEM an important tool for nano-technologies research, metallurgic studies and semiconductor development.

TEM can be modified into scanning TEM (STEM) by the addition of a system that rasters the beam across the sample to form the image, combined with suitable detectors. The rastering makes these EMs suitable for mapping by energy dispersive X-ray spectroscopy, electron energy loss spectroscopy (EELS, see Electron microscopy) and annular dark-field imaging. These signals can be obtained simultaneously, allowing direct correlation of image and quantitative data.
By using a STEM and a high-angle detector, it is possible to obtain atomic resolution images where the contrast is directly related to the atomic number. This is in contrast to the conventional HRTEM that uses phase-contrast.
An analytical TEM is equipped with detectors that can determine the elemental composition of the specimen by analyzing its X-ray spectrum or the energy-loss spectrum of the transmitted electrons.

The TEM contrast is not like the contrast in a light microscopic image. A crystalline material (e.g. proteins) interacts with the electron beam mostly by diffraction rather than absorption, although the intensity of the transmitted beam is still affected by the volume and density of the sample material. The intensity of the diffraction depends highly on the orientation of the planes of atoms in a crystal relative to the electron beam. Modern TEMs are equipped with sample holders that allow the user to tilt the sample to a range of angles in order to obtain specific diffraction conditions, and apertures placed below the sample allow the user to select electrons diffracted in a particular direction.
A high-contrast image can therefore be formed by blocking electrons deflected away from the optical axis of the microscope by placing the aperture to allow only unscattered electrons through. This produces a variation in the electron intensity that reveals information on the crystal structure. This technique is known as *Bright Field* or *Light Field*.
It is also possible to produce an image from electrons deflected by a particular crystal plane. By either moving the aperture to the position of the deflected electrons, or tilting the electron beam so that the deflected electrons pass through the centred aperture, an image can be formed of only deflected electrons, known as a *Dark Field* image.

There are a number of drawbacks to the TEM technique:
- sample preparation is time consuming;
- sample preparation has the risk of introducing artifacts;
- there is a risk that the sample may be damaged by the electron beam;
- the small field of view, which raises the possibility that the region analyzed may not be characteristic of the whole sample.

Electron Spin Resonance (ESR)

Principle

Electron spin resonance (ESR) or electron paramagnetic resonance (EPR) is a spectroscopic technique (see Spectroscopy) which detects atoms that have unpaired electrons. It means that the molecule in question is a free radical if it is an organic molecule. Because most stable molecules have a closed-shell configuration of electrons without a suitable unpaired spin, the technique is less widely used than Nuclear magnetic resonance (NMR). (Spin refers to rotating of a particle around some axis trough the particle.)
The basic physical concepts of the technique are analogous to those of NMR, but instead of the spins of the atom's nuclei, electron spins are *excited*. Because of the difference in mass between nuclei and electrons, some 10 times weaker magnetic fields and higher frequencies are used, compared to NMR. For electrons in a magnetic flux field of 0.3 T (tesla), spin resonance occurs at around 10 GHz. (1 T = 1 N/(Am) = 10.000 Gauss = 10^9 gammas).

Application

ESR is used in solid-state physics for the identification and quantification of radicals (i.e., molecules with unpaired electrons), in (bio)chemistry to identify reaction pathways, e.g. the cytochrom electron transfer chain. Its use in biology and medicine is more complicated. A radical can be used to couple the probe to another molecule, e.g. a bio-molecule).
Since radicals are very reactive, they do not normally occur in high concentrations in living organisms. With the help of specially designed non-reactive, so stable free radical molecules carrying a functional group that attach to specific sites in e.g. a mitochondrion, it is possible to quantify structures comprising these sites with these so-called spin-label or spin-probe molecules.
To detect some subtle details of some systems, high-field-high-frequency electron spin resonance spectroscopy is required. While ESR is affordable for a ordinary-sized academic laboratory, there are few scientific centers in the world offering high-field-high-frequency electron spin resonance spectroscopy.

More Info

An electron has a magnetic moment (is torque/magnetic field strength, in Am^2) and spin quantum number s = 1/2, with magnetic components $m_s = +1/2$ and $m_s = -1/2$. When placed in an external magnetic field of strength B_0, this magnetic moment can align itself either parallel ($m_s = -1/2$) or antiparallel ($m_s = +1/2$) to the external field. The former is a lower energy state than the latter (this is the Zeeman effect), and the energy separation between the two is:

$$\Delta E = g_e \mu_B B_0, \qquad (1)$$

where g_e is the gyromagnetic ratio or g-factor of the electron, a dimensionless quantity. It is the ratio of its magnetic dipole moment to its angular momentum. The magnetic moment in a magnetic field is a measure of the magnetic flux set up by the rotation (spin and orbital rotation) of an electric charge in a magnetic field. Further, μ_B is the Bohr magneton (a constant of magnetic moment $\mu_B = 9.27 \times 10^{-24}$ Am^2) and B_0 the magnetic field strength (N/A·m).

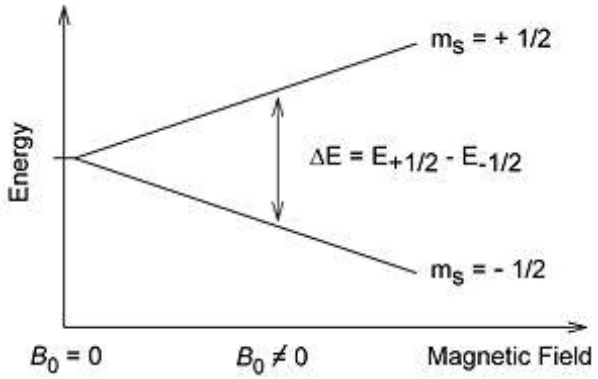


Fig. 1 (From Wikipedia/Electron Spin Resonnance)

An unpaired electron can move between the two energy levels by either absorbing or emitting electromagnetic radiation of energy ε = *hv* (Planck's constant times frequency) such that the resonance condition, ε = ΔE, is obeyed. Substituting in ΔE = *hv* in (1) gives:

$$hv = g_e \mu_B B_0. \quad (2)$$

This is the fundamental equation of ESR spectroscopy.
This equation implies that the splitting of the energy levels is directly proportional to the magnetic field's strength, as shown in Fig. 1.

A collection of paramagnetic molecules (molecules with the property to align in a magnetic field, see Diamagnetism and paramagnetism), such as free radicals, is exposed to microwaves at a fixed frequency. By increasing an external magnetic field, the gap between the $m_s = +1/2$ and $m_s = -1/2$ energy states is widened until it matches the energy of the microwaves, as represented by the double-arrow in Fig. 1. At this point the unpaired electrons can move between their two spin states. Since there typically are more electrons in the lower state, there is a net absorption of energy, and it is this absorption which is monitored and converted into a spectrum.

A free electron (on its own) has a *g* value of about 2.0023 (which is g_e, *the electronic* g factor). This means that for radiation at the commonly used frequency of 9.75 GHz (known as X-band microwave radiation, and thus giving rise to X-band spectra), resonance occurs at a magnetic field of about 0.34 T.

ESR signals can be generated by resonant energy absorption measurements made at different electromagnetic radiation frequencies *v* in a constant external magnetic field (*i.e.* scanning with a range of different frequency radiation whilst holding the field constant, like in an NMR experiment). Conversely, measurements can be provided by changing the magnetic field *B* and using a constant frequency radiation. Due to technical considerations, this second way is more common. This means that an ESR spectrum is normally plotted with the magnetic field along the horizontal axis, with peaks at the field that cause resonance (whereas an NMR spectrum has peaks at the frequencies that cause resonance).

For more information see e.g. the Wikipedia chapter on ESR.

References

Wikipedia Electron paramagnetic resonace; http://en.wikipedia.org/wiki/Electron_spin_resonance.

Gamma camera

Principle

A gamma camera is an imaging device most commonly used in nuclear medicine. It produces images of the distribution of gamma rays emitted by metastable radionuclides (isotopes), also called metastable nuclear isomers. A gamma ray comprises gamma photons, which are high energetic photons (at least 5000 times those of visible light). They are produced from sub-atomic particle interaction, such as radioactive decay and electron-positron annihilation. (Annihilation is the collision of a positron with an electron, followed by vanishing of both. Two and sometimes more photons are produced moving in almost opposite directions. A radionuclide can also produce subatomic particles (which give ionization). Excited metastable isomers de-excite with sending a gamma photon mostly within much less than one ps, but some isomers are far much slower. These are for example the technetium isomers ^{99m}Tc (here indicated without atom number; m indicates metastable; half-life 6.01 hours) and ^{95m}Tc (half-life of 61 days).

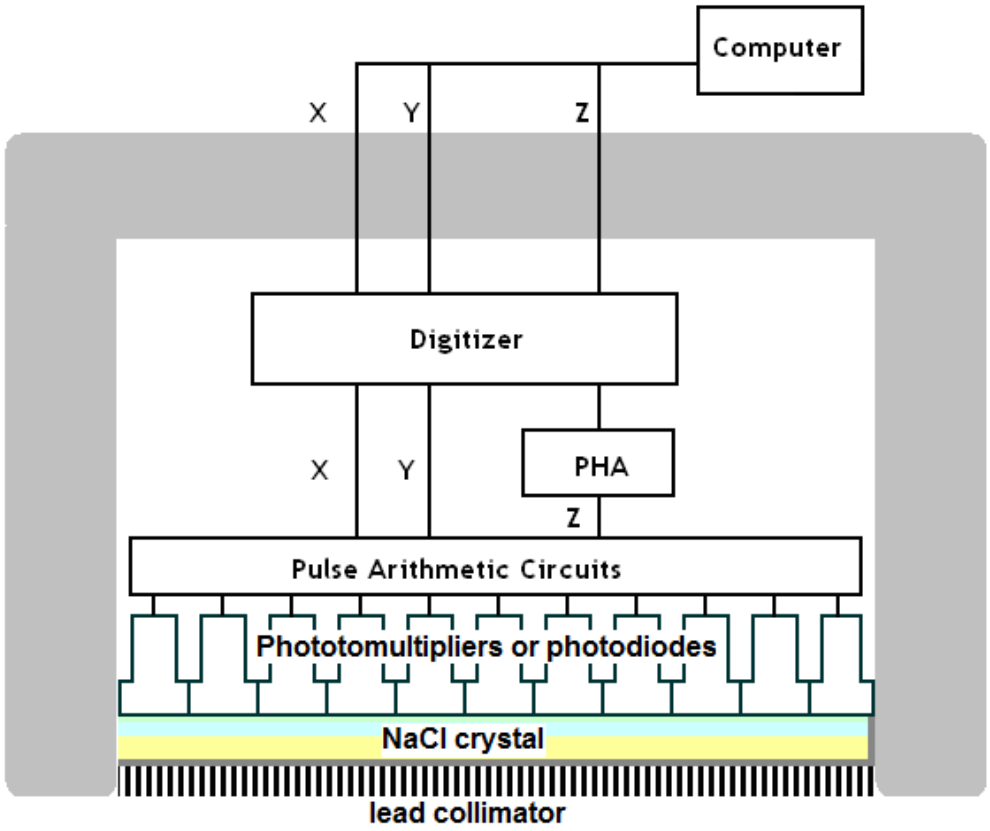


Fig. 1 Diagrammatic cross section of a gamma camera detector.

A gamma camera consists of one or more detectors mounted on a gantry. It is connected to an acquisition system for operating the camera and for storing the images. The system counts gamma photons that are absorbed by usually a large flat crystal of NaI with thallium doping in a light-sealed housing. The crystal scintillates in response to incident gamma radiation: when a gamma photon collides an electron loose from an iodine atom in the crystal, a faint flash of light is produced when the electron again finds a minimal energy state. The initial phenomenon of the excited electron is similar to the photoelectric effect (an electron hitting an atom, with as a result the emission of another electron and back-scatter of the electron with a lower speed) and (particularly with gamma rays) the Compton effect. This is generally an electron hitting an atom resulting the emission of a photon and back-scatter of the electron with a lower speed (see also SPECT, More info). So, actual it is Fluorescence, but here for impinging gamma rays. The flash of light must be detected. Photomultiplier tubes (extremely sensitive detectors of UV, near-IR and visible light) and nowadays photodiodes behind the crystal detect the fluorescent flashes and a computer sums the fluorescent flashes. The computer in turn constructs and displays a 2D image of the relative spatial count density. This image is the visualization of the distribution and relative concentration of radioactive tracer elements present in the tissues imaged.

Application

SPECT (Single photon emission computed tomography) machines are based at gamma cameras. Multi-headed gamma cameras can also be used for PET (Positron emission tomography), provided that their hardware and software can be configured to detect 'coincidences' (near simultaneous events on 2 different heads). Gamma camera PET is markedly inferior to PET imaging with a purpose designed PET scanner, as the scintillator crystal has poor sensitivity for the high-energy annihilation photons, and the detector area is significantly smaller. However, given the low cost of a gamma camera and its flexibility compared to a dedicated PET scanner, this technique is useful when the expense of a PET scanner cannot be justified.
In order to obtain spatial information about the gamma emissions from an imaging subject (e.g. a person's heart muscle cells, which have absorbed an intravenous injected radioactive, usually a ^{201}Th or ^{99m}Te based medicinal imaging agent) a method of correlating the detected photons with their point of origin is required.
The spatial resolution is a major limitation for heart muscle imaging systems. The thickest normal heart muscle in the left ventricle is about 12 mm and most of the left ventricle muscle is about 8 mm, always moving and much of it beyond 50 mm from the collimator face. Systems with scintillation counting limited to a portion of the heart cycle, called gating, improve imaging. However this further limits system sensitivity (see also PET).

More Info

The electronic circuit connecting the photodiodes is wired so as to reflect the relative coincidence of light fluorescence as sensed by the members of the detector array. All the photodiodes which simultaneously detect the (presumed) same flash of light provide a pattern of voltages within the interconnecting circuit array. The location of the interaction between the γ-ray and the crystal can be determined by processing the voltage signals from the photodiodes. In simple terms, the location can be found by weighting the position of each photodiode by the strength of its signal, and then calculating a mean position from the weighted positions. The total sum of the voltages from each photodiode is proportional to the energy of the gamma ray interaction, thus allowing discrimination between different radio-isotopes or between scattered and direct photons.

The conventional method is to place a collimator (device that selects the parallel rays to go through) over the detection crystal/photodiodes array. The collimator essentially consists of a thick sheet of lead, typically 25-75 mm thick, with thousands of adjacent holes through it. The individual holes limit photons which can be detected by the crystal to a cone; the point of the cone is at the midline center of any given hole and extends from the collimator surface outward. However, the collimator is also one of the sources of blurring within the image. Lead does not totally attenuate incident γ-photons. There can be some leakage from the one to the other hole.
Unlike a lens, as used in visible light cameras, the collimator attenuates most (>99%) of incident photons and thus greatly limits the sensitivity of the camera system. Large amounts of radiation must be present so as to provide enough exposure for the camera system to detect sufficient scintillation dots to form a picture.
Other methods of image localization (pinhole, rotating slat collimator with Cd, Zn, Te and others) have been proposed and tested; however, none have entered widespread routine clinical use.
High resolution camera systems point source resolution of 8 mm at 50 mm away from the collimator. Spatial resolution decreases lineary at increasing distances from the camera face (about half at 150 mm). This limits the spatial accuracy of the computer image: it is a fuzzy image made up of many dots of detected but not precisely located scintillation.

References

1. Valastyán I., D. Bone D. et al. 2007 Validation of an iterative reconstruction for a mobile tomographic gamma camera system. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, Volume 580, 1097-1100

Image processing: 3D reconstruction

Principle

To apply one of the reconstruction techniques, a preliminary step is *windowing* of the 2D slices. It is the process of using the measured radiographic densities in Hounsfield units, (HU, see CT scan (dual energy)) to make an image. The various HU amplitudes are mapped to 256 gray-shades. They are distributed over a wide range of HU values when an overview of structures is needed. They can also be distributed over a narrow range (called a *narrow window*) centered over the average HU value of a particular structure in order to discern subtle structural details. This image processing technique is known as *contrast compression*.
There exist various 3D reconstruction techniques. Briefly some are mentioned. Of some of them, details can be found in. Important are

Because contemporary CT scanners offer (nearly) isotropic (the property of being independent of direction) resolution in space, a software program can 'stack' the individual slices one on top of each other to construct a volumetric or 3D image. The procedure is called *multiplanar reconstruction* (MPR).

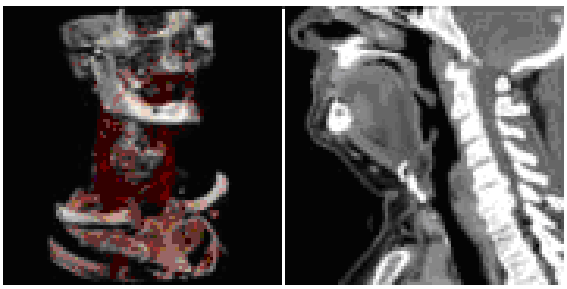


Fig. 1 Typical screen layout for diagnostic software, showing a 3D (left, torso) and a MPR (right, vertebrate column) view.

A special projection method *maximum-intensity projection* (MIP) or minimum-intensity projection (mIP) can be used to build the reconstructed slices. With MIP a transparent 3D view is visualized, which can be rotated.
With *rendering* techniques boundaries or volumes can be highlighted.
Segmentation is the technique to separate kinds of tissues from each other, for instance within an MRI image of the head skull, liquor, white and gray matter of the brain.

Application

3D-reconstruction techniques are basic tools for all imaging techniques in medicine, so it is not limited to CT (see CT scan (dual energy)), but also used for PET, SPECT, MRI (see MRI: general), fMRI (see MRI: functionl MRI) and optical imaging techniques such as Optical coherence tomography (OCT), Thermography, Diaphanography and optical mammography. They all rely on windowing as the preliminary step. To explain how windowing works in practice some examples are given.
Windowing liver and bones
For example, to evaluate the abdomen in order to find subtle masses in the liver, a good liver window is 70 HU as an average with the shades of gray distributed over a narrow window of ± 85 HU. Any HU value below -15 would be pure black, and any HU value above 155 HU would be pure white in this example. Using this same logic, bone windows would use a wide window since bone contains dense cortical bone as well as the low dense fatty marrow.
Visualization of spines
Axial images through the spine will only show one vertebral body at a time and cannot reliably show the intervertebral discs. By reformatting the volume with MPR it is possible to visualize the position of one vertebral body in relation to the others.

More Info

A maximum intensity projection (MIP) is a computer visualization method for 3D data that projects in the visualization plane the voxels (a voxel is the 3D analog of the 2D pixel) with maximum intensity that fall in the way of parallel rays traced from the viewpoint to the plane of projection. This implies that two MIP renderings from opposite viewpoints are symmetrical images.

This technique is computationally fast, but the 3D results do not provide a good sense of depth of the original data. To improve this, animations are usually rendered of several MIP frames in which the viewpoint is slightly changed from one to the other, thus creating the illusion of rotation. This helps the viewer's perception to find the relative 3D positions of the object components. However, since the projection is orthographic one cannot distinguish between right or left, front or back and even if the object is rotating clockwise or anti-clockwise.
Retrieved from "http://en.wikipedia.org/wiki/Maximum_intensity_projection" where also an animation can be found.

Modern software enables reconstruction in non-orthogonal (oblique) planes so that the optimal plane can be chosen to display an anatomical structure (e.g. for bronchi as these do not lie orthogonal to the direction of the scan).
For vascular imaging, *curved-plane reconstruction* can be performed. This allows bends in a vessel to be 'straightened' so that the entire length can be visualized on one image. After 'straightening', measurements of diameter and length can be made, e.g. for planning of surgery.
Since MIP reconstructions enhance areas of high radiodensity, they are useful for Angiography and DSA. mIP reconstructions enhance air spaces, so they are useful for assessing *lung structure*.

3D rendering techniques
Surface rendering A threshold value of radiodensity is set by the operator. A threshold is set, using edge detection image processing algorithms. In edge detection, small groups of pixels (e.g. forming a line piece) with strong intensity differences from neighboring ones are marked. In this way, finally 3D-surface is extracted and a 3D model with contour is constructed. (In fact, complicated mathematics is used to realize edge detection.) Multiple models can be constructed from various different thresholds, each with its own color to represent anatomical components such as bone, muscle, and cartilage. Surface rendering only displays surfaces meeting the threshold, and only displays the surface most close to the imaginary viewer. Therefore, the interior structure of each element is not visible.
Volume rendering In volume rendering colors in adition to transparency are used to allow a representation of the various structures within a volume, e.g. the bones could be displayed as semi-transparent, so that even at an oblique angle, one part of the image does not conceal another.

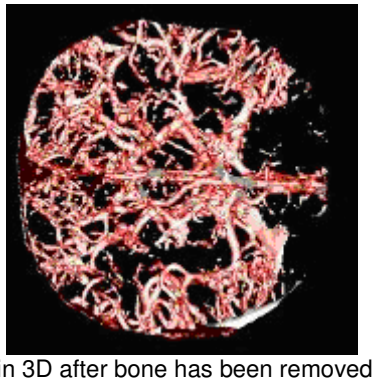


Fig.2 Brain vessels reconstructed in 3D after bone has been removed by segmentation

Image segmentation
Where different structures have similar radiodensity, it is impossible to separate them with volume rendering. The solution is called segmentation. This is a manual or automatic procedure that can remove the unwanted structures from the image.
After using a segmentation tool to remove the bone of the skull, now the previously concealed vessels can be demonstrated (Fig. 2).

Magnetobiology

Principle

Magnetobiology is an experimental approach in radiobiology of non-ionizing radiation, which respects biological effects of mainly weak, static and low-frequency magnetic fields, which do not cause heating of tissues. Magnetobiology is related to the more general term of Bioelectromagnetics.
It should be noted that the results of magnetobiological experiments are on the whole poorly reproducible. In most experiments, their success depended on a rare coincidence of suitable physiological and electromagnetic (EM) conditions. Many of the experiments await confirmation.
Since theory is lagging far behind experiment, the nature of biological effects of weak electromagnetic fields remains unclear as yet. On the one hand the ever growing level of the background electromagnetic exposure of human posed the question whether such fields are harmful. This question is addressed in Mobile phones radiation hazards. At the other hand it is questionable whether all kind of medical or non-medical treatments have any effect.
To be one the save side, safety standards have been developed by many national and international institutions (see Fig. 1).

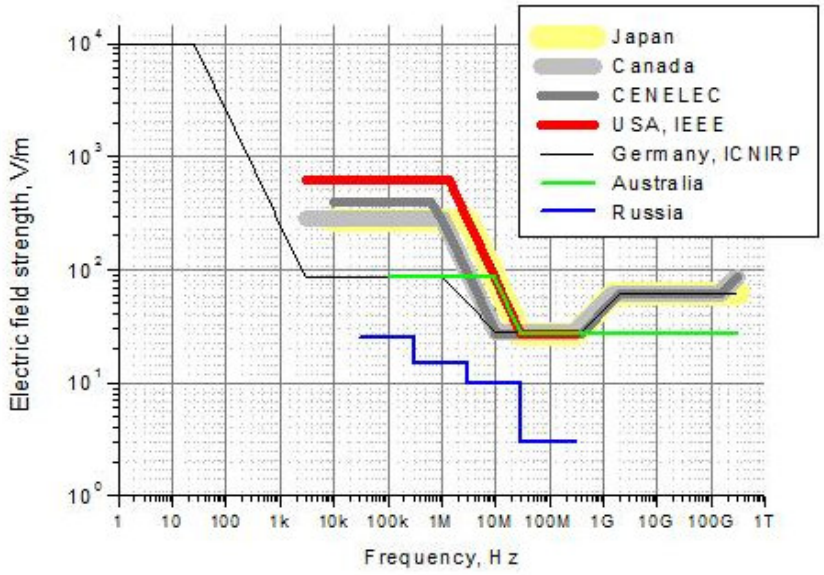


Fig. 1 Safety levels practices in various countries and by several institutions.

The fact that safety standards differ by tens and hundreds of times for certain EM field frequencies reflects the nearly complete lack of (electro)magnetobiological research. Today, most of the standards take into account biological effects just from heating effect and peripheral nerve stimulation from induced currents.

Application

New technologies are developed for medical treatment by the use of relatively weak electromagnetic fields that influence brain functioning (see Transcranial magnetic stimulation, (TMS)).
A complete different application, which explicitly aims to heat tissue, is thermotherapy (see Bioelectromagnetics) to destroy for instance cancer cells.

Mobile phone radiation risk

Principle

Health concerns about the use of mobile (cellular or cell) phone radiation and other wireless transmitting devices have been raised, because such apparatus use electromagnetic (EM) waves in the microwave range. These concerns have induced many studies (both experimental and epidemiological, in animals as well as in humans).

The WHO has concluded that serious health effects (e.g. cancer) of mobile phones or their base stations are very unlikely. However, some nations' radiation advisory authorities recommend their citizens to minimize radiation. Examples of recommendations are hands-free use, keeping the mobile phone away from the body and do not telephone in a car without an external antenna.

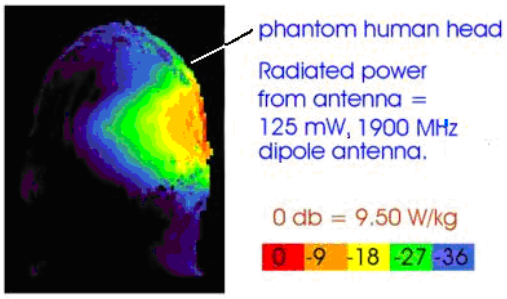


Fig. 1 Transmitted power from a typical mobile phone. Calculated specific absorbed radiation (SAR) distribution in an anatomical model of the head next to a 125 mW dipole antenna. Peak SAR is 9.5 W/kg in 1 mL tissue, the whole head average is 0.008 W/kg. (From ref.3.)

Part of the radio waves emitted by a mobile phone (digitally peak power of 2 W, analogue phone 3.6 W) are absorbed by the human head. Most mobile phone types and base stations (large antenna tower or small consumer device) check reception quality. Signal strength and power level is increased and decreased automatically, within a certain span, to adapt to different situations such as inside or outside of buildings and vehicles.
The rate at which radiation is absorbed by the human body is measured by the so called Specific Absorption Rate (SAR), and its maximum levels are governmental regulated (e.g. in USA 1.6 W/kg, averaged over a volume of 1 ml of head tissue).

Thermal effects
One effect of microwave radiation is dielectric heating (micro wave oven), in which any dielectric material (i.e. particles at molecular level comprising positive and negative charges, as holds in body tissues) is heated by rotations of the polar molecules induced by the electromagnetic field. Using a mobile phone, most of the heating effect will occur in the skin of the head, causing an irrelevant temperature increase by a fraction of 1 ^{o}C (see **More Info**) since the circulatory system is well capable of disposing of excess heat by increasing local blood flow. However, the cornea of the eye lacks this regulation mechanism. Premature cataracts are known as an occupational disease of engineers working on much high power at similar frequencies. Mobile phones do not have this hazard.
Microwave radiation can cause a leakage of albumin into the (rat) brain via a permeated blood-brain barrier. When this is a temperature effect, it will not occur in humans with their thicker skin and skull.

Non-thermal effects
Genotoxic effects These effects may occurr when exposed to very high powers, between 0.3 to 2 watts/kg whole-sample average (*in vitro* experiments).
Mobile phones and cancer Whether mobile phones do increase the risk of glioma or meningioma with long-term and frequent use is highly controversial (see **More Info**) for possible explanation).
Electromagnetic hypersensitivity? The report of users of mobile handsets of several unspecific symptoms during and after its use has never been confirmed scientifically. Probably this all is based on imagination; hypersensitivity of all these persons is very unlikely.

Health hazards of base stations
Another area of worry about effects on the population's health have been the radiation emitted by base stations (the antennas which communicate with the phones), because it is emitted continuously and is more powerful. On the other hand field intensity drops with the square of distance from the base of the antenna. The results of the various studies were diverging or even conflicting.

Studies of occupational health hazards of telecommunication workers show that the risk is totally negligible. Such workers spend time at a short distance from the active equipment, for the purposes of testing, maintenance, installation, etc. They may be at risk due to the much greater exposure than the general population. Many times base stations are not turned off during maintenance, because that would affect the network. A variety of studies over the past 50 years have been done on workers exposed to high radio frequent radiation levels. Studies including radar laboratory workers, electrical workers, military radar workers and amateur radio operators. Most studies found no increase in cancer rates over the general population or a control group. Many positive results may be attributed to other work environment conditions, and many negative results of reduced cancer rates also occurred.

Safety standards and licensing
In order to protect the population living around base stations and users of mobile handsets, governments and regulatory bodies adopt safety standards (often including licensing). They translate to limits on exposure levels below a certain value. The International Committee for Non-Ionizing Radiation Protection (ICNIRP), adopted so far by more than 80 countries, proposes two safety limits for radio stations: one for occupational and one for general exposure.

Precautionary principle
Often, authorities use the "precautionary principle". This means that as long as the ultimate conclusions of scientific research can not be drawn, recommendations are made. These are for instance the minimization of cellular phone usage, the limitation of use by at-risk population (children), avoiding mobile phones with maximal levels of radiation, the wider use of hands-off and earphone technologies such as Bluetooth headsets, the adoption of maximal standards of exposure, RF field intensity and distance of base stations antennas. Bluetooth is a wireless personal area network system to exchange information between devices such as mobile phones, laptops, PCs, printers and cameras.

More Info

Dielectric heating
Dielectric heating is a process, which increases directly the dielectric material's heat by externally forced dipole rotation. (A dielectric material poorly conducts electric current.)
The power density per volume, P_v, with an isotropic ohmic resistance, generated by dielectric heating is:

$$P_v = \varepsilon_0 E^2 \quad (W/m^3), \qquad (1)$$

where ε_0 is the permittivity of free space (ca. $8.86 \cdot 10^{-12}$ F/m) and E the electric field strength (V/m). With a field strength of 100 V/m, P_v, is about 10^{-7} W/cm^3. Supposing that a mobile phone produces such a field and that 10% of this power is produced in the 10 mL most close to the phone, then the dissipated energy in this volume is 10^{-3} W/mL, causing a temperature increase of 10^{-3} W/4.2 ≈ $0.25 \cdot 10^{-3}$ °C. The calculation can also be made with a known SAR. With a SAR of 1.6 W/kg (the maximal allowable US exposure), 1 mL of head tissue most close to the mobile phone raises about $(1.6/4.2) \cdot 10^{-3} = 0.4 \cdot 10^{-3}$ °C per second. The outcomes indicate that this dissipation is too small to heat even locally the head due to internal cooling (blood flow) and (mostly) to external cooling. However, using actual data about source strength, increases are about some 0.3 °C. Consequently, the observed increased risk after long lasting use (> 10 year) of acoustic neuroma seems to have another cause. Actual temperature measurements of the skin of the pinnae may go up to 1.5 °C, but this an effect caused by mechanical isolation of the pinnae and surrounding skin tissue. Summarizing, the cause of the increased risk remains enigmatic.

Sperm and eye lens heat hazard The testes are vulnerable to heating because of poor circulation and heat is known to have adverse effects on male fertility. However, that heavy mobile phone use (>4 hours per day) may significantly less viable sperm due to a radio frequent heat effect can be excluded definitely. A similar calculation as above, but now at a 10 times longer distance yields in the worst case an increase of only 0.003 °C/s, probably too small to yield an effect.
The eye lens is also sensitive for dissipating heat. Short duration exposure to *very high levels* of RF radiation can cause cataracts in rabbits. Mobile phones will not produce such an effect.

References
1. Kundi M, Mild K, Hardell L, Mattsson MO. Mobile telephones and cancer, a review of epidemiological evidence. J Toxicol Environ Health B Crit Rev. 2004, 7:351-84.
2. WHO, Online Q&A, "What are the health risks associated with mobile phones and their base stations?". 2005-12-05. http://www.who.int/features/qa/30/en. Retrieved 2008-01-19.
3. Wikipedia//Mobile phone radiation and health (main source of knowledge); http://en.wikipedia.org/wiki/Mobile_phone_radiation_and_health.

MRI: general

Principle

Magnetic resonance imaging (MRI), formerly referred to as *magnetic resonance tomography (MRT)* and alos called *nuclear magnetic resonance imaging (NMRI)*, is a non-invasive method used to render images of the inside of an object. From the physical, technological and computational point of view, MRI is by far the most complicated medical and biological imaging techniqie. It is primarily used in medical imaging to demonstrate pathological or physiological alterations of living, generally soft tissues. MRI is based on the principles of Nuclear magnetic resonance (NMR), a spectroscopic technique (see Spectroscopy). MRI, in contrast to CT, uses *non*-ionizing radio frequency (RF) signals to acquire its images and is best suited for non-calcified tissue.

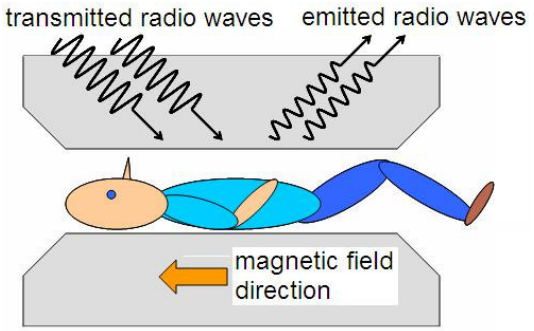


Fig.1 Basic set up MRI apparatus (Modified after Introduction to MRI, Erik M. Akkerman, 2001.)

Working principle MRI (mostly) measures the density of hydrogen, which is abundantly present in the (human) body in the form of water. The measurement is based on a specific magnetic behavior of the hydrogen proton (see Nuclear magnetic resonance (NMR)) and Paramagnetism, diamagnetism and magnetophoresis).
The patient is placed in a strong electromagnetic field (Fig. 1) of 0.3 to 4 T (specialized clinical MRIs up to 7 T (1 T = 1 $V{\cdot}s{\cdot}m^{-2}$ = 1 $N{\cdot}A^{-1}m^{-1}$ =1 $kg{\cdot}A^{-1}{\cdot}s^{-2}$) and experimental animal machines up to 9.4 T.
The billions of protons of the H-atoms (in the body), behaving als small magnetic dipoles, align themselves parallel with the magnetic field, either in the same direction or opposite to the direction of the field. At a particular "slice" level where it is desired to 'take a picture', a short, powerful radio signal (a form of electromagnetic energy) is sent through the part of the body to be examined, perpendicular to the main magnetic field. The H-protons, which can vibrate with the same frequency as the impinging radio wave, will become 'excited' (that is, they will be raised to a higher state of energy) and start to resonate with the frequency of the exciting wave.
When the radio signal is switched off, the H-atoms will, after some time, return to their original energy state. The gained excitation energy is now released in the form of radio waves, which are detected by sensors. The strength of the sensor signal is proportional to the proton density (PD). The time it takes for the excited H-atoms to return to their original energy level, the relaxation time, is measured and analyzed. There are two types of emitted radio waves, with relaxation times called T1 and T2. The former relies on the direct surrounding of the spinning proton and the latter on the mobility of the proton. One of the three (density, T1 or T2) can be used for the image, but a combination is also possible. For the contrast in the image and so for the interpretation of the image one should know which one of thyhe three is used. (For details see MRI: T1 and T2 relaxation.)
To make an image one should also know where the emitted RF signal arises, so one needs 3D-info. The RF coils (antennas) are not direction sensitive, so do not provide this information. The position is provided by the x, y and z gradient coils.

CT may be enhanced by use of contrast agents containing elements of a higher atomic number than the surrounding water-like tissues. Contrast agents for MRI must be strongly paramagnetic (becoming magnetic in a magnetic field, see Paramagnetism, diamagnetism and magnetophoresis). One example is gadolinium (Ga).
Unlike CT, MRI has many properties that may be used to generate image contrast. By variation of scanning parameters, tissue contrast can be altered and enhanced in various ways to detect different features. (See **Application** below.)

MRI can generate cross-sectional images in any plane (including oblique planes). For purposes of tumor detection and identification, MRI is generally superior. However, CT usually is more widely available, faster, less expensive, and seldom requires sedation or anesthesia of the patient.

Various specialized MRI machines or methods have been developed, such as fMRI (see MRI: functional MRI), diffusion MRI (MRI: diffusion MRI), MR Angiography and MR spectroscopy (see MRI: other specialized types).

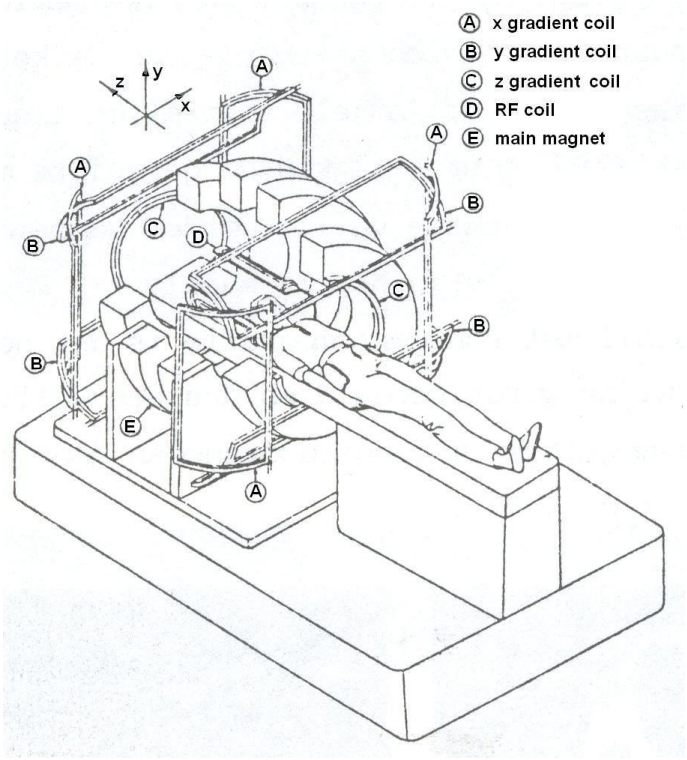


Fig. 2 The basic, but complete MRI instrument.

Application

In medicine
In clinical practice, MRI is used to distinguish pathologic tissue from healthy tissue. One advantage of an MRI scan is that it is harmless to the patient. It uses strong magnetic fields and non-ionizing radiation in the radio frequency range. CTs use traditional X-rays, which involve doses of ionizing radiation and may increase the risk of malignancy, especially in a fetus.
While CT provides good spatial resolution (the minimal distance to distinguish two points in a tissue from each other), MRI distinguishes better the differences between two similar but not identical tissues. The basis of this ability is the complicated library of *pulse sequences* that the modern clinical MRI scanner applies, each of which is optimized to provide *image contrast* based on the sensitivity of MRI for the paramagnetism of elements, H-atoms in particular.
Advantages:
- No ionizing radiation;
- Large contrast between different soft tissues caused by differences in T1, T2, etc. (see MRI: T1 and T2 relaxation), PD;
- Orientation of the selected cross section is free.

Disadvantages:
- In general taking more time than CT;
- Worse imaging of e.g. bone, lungs;
- Contra-indications: pacemaker, clips;
- Expensive.

Outside medicine MRI is used for e.g. detecting rock permeability to hydrocarbons and for non-destructive testing methods to characterize the quality of products such as fruit, vegetables and timber.

More Info

Gradients
Magnetic gradients are generated by three orthogonal coils, oriented in the x, y and z directions of the scanner. These are mostly resistive electromagnets powered by amplifiers, which permit rapid and precise adjustments to their field strength and direction. Typical gradient systems are capable of producing gradients from 20 to 100 mT/m (in a 1.5 T magnet). It is the magnetic gradient that

determines the imaging plane. Because the orthogonal gradients can be combined freely, any plane can be selected.
Scan speed is dependent on performance of the gradient system. Stronger gradients permit faster imaging and give higher resolution. Similarly, gradient systems capable of faster switching also permit faster scanning. However, gradient performance is limited in order to prevent nerve stimulation.

Scanner construction and operation
The static magnetic field, the RF transmitter and receiver, and the three orthogonal, controllable magnetic gradients form the major components of an MRI scanner.
Magnet The magnet is the largest and most expensive component, and the remainder of the scanner is built around the magnet. The straightness of flux lines within the centre or, as it is called the iso-centre of the magnet, need to be perfectly homogeneous (fluctuations and non-homogeneities < 0.0003 %). Present-day MRIs have *superconducting electromagnets.* When a Nb-Ti- alloy is cooled by liquid He at 4K (-269°C) it becomes superconducting where it loses all resistance to flow of electrical current. By building an electromagnet with a solenoid (coil) of superconducting wire, it is possible to develop extremely high field strengths, with very high stability.
The magnet wire is cooled directly by a so-called cryocooler instead of being immersed in liquid He, the 'cryogenless' scanner. Superconducting magnets are most frequently cylindrical.
Signal-to-noise ratio increases with field strength, permitting higher resolution or faster scanning. However, higher field strengths have increased safety concerns. 1.5 T field strengths are a good compromise between cost and performance for general medical use. However, for certain specialist uses (e.g., brain imaging) field strengths up to 4T may be desirable and for experimental clinical work 7-11 T.

RF system
The RF transmission system consists of a RF synthesizer, power amplifier and transmitting coil, usually built into the body of the scanner. The power of the transmitter is variable, but high-performance scanners may have a peak output power of up-to 35 kW, and are capable of sustaining an average power of 1 kW. The receiver consists of the coil, pre-amplifier and signal processing system. It is possible to use the integrated coil for transmitting and receiving. If a small region is imaged then better image quality is obtained by using a close-fitting smaller coil. Special coils are available which fit around parts of the body, e.g., the head, knee, wrist or abdomen.
A recent development is the development of sophisticated multi-element phased array coils, coils that outputs are together. This 'parallel imaging' technique uses unique acquisition schemes that allow for faster imaging.

Contrast enhancement
Both T1-weighted and T2-weighted images (MRI: T1 and T2 relaxation) are acquired for most clinical examinations. However, they do not always adequately show the anatomy or pathology. An improvement is to use a more sophisticated image acquisition technique such as fat suppression. Another is to administer a contrast agent to delineate areas of interest.
A contrast agent may be as simple as water, taken orally, for imaging the stomach and small bowel. In addition, substances with specific magnetic properties can be used. Mostly, a paramagnetic contrast agent (usually a Ga-compound) is given. This provides high sensitivity for detection of vascular tissues (e.g. tumors) and permits assessment of brain perfusion (e.g. in stroke). However, the toxicological safety of Ga-based contrast agents is questionable (kidney patients with e.g. hemodialysis following a contrast MRI).
Superparamagnetic contrast agents (e.g. Fe-oxide nanoparticles for liver imaging), often taken orally, appear very dark. (Superparamagnetism is a phenomenon, by which magnetic materials, which may exhibit a behavior similar to paramagnetism even when at very low temperatures.) They improve visualization of the gastrointestinal tract, and prevent water in the gastrointestinal tract from obscuring other organs (e.g. the pancreas).
Diamagnetic agents (repellence in a magnetic field, see Paramagnetism, diamagnetism and magnetophoresis) such as $BaSO_4$ have been studied for use in the gastrointestinal tract.

Image formation
In order to selectively image the voxels (3D-pixels), orthogonal magnetic gradients are applied. Generally, gradients applied in the principal axes of a patient (so that the patient is imaged in x, y, and z from head to toe). However, MRI allows completely flexible orientations for images. All spatial encoding is obtained by applying magnetic field gradients, which encode position within the phase of the signal. In one dimension, a linear phase with respect to position can be obtained by collecting data in the presence of a magnetic field gradient. In 3D, a plane can be defined by the “slice” selection, in which an RF pulse of defined bandwidth is applied in the presence of a magnetic field gradient in order to reduce spatial encoding to 2D. Spatial encoding can then be applied in 2D after slice selection, or in 3D without slice selection. Spatially encoded phases are recorded in a 2D or 3D matrix, the so-called k-space (see MRI: k-space formalism). This data represents the spatial frequencies of the image object. Images can

be created from the matrix using the discrete Fourier transform (DFT) (see Fourier analysis). Typical medical resolution is about 1 mm^3, while research models at cellular level can exceed 1 μm^3.

MRI Artifacts
The theoretical limit of the precision of images is determined by the so-called point spread function of the imaging device (see Light: the ideal and non-ideal lens, More info). In practice, however, here the physiological movements of a living subject (e.g. respiration, heartbeat, swallowing, twitching or tremor) and the finite thickness determine the limit. If the signals from different tissue compartments cannot be separated within separate voxels, then an artifact known as partial voluming occurs (see Image processing: 3D reconstruction). This uncertainty in the exact content of any voxel is an inherent property of the discretised image and would even exist if the contrast between tissues were infinite. Furthermore, chemical shift and susceptibility artifacts, magnetic field and RF non-uniformity, and slice thickness calibration inaccuracies can all compromise the accuracy. A detailed analysis of all these effects is, however, beyond the scope of this article.

MRI: Diffusion MRI

Principle

Diffusion MRI measures the diffusion (see Diffusion: general) of water molecules in the tissues. In an isotropic medium (e.g. water), the water molecules naturally move randomly according to Brownian motion. In tissues however, the diffusion may be anisotropic. For example, a molecule inside the axon of a neuron has a low probability of crossing the myelin membrane. Therefore, the molecule will move mainly along the axis of the axon. When the molecules in a particular voxel diffuse principally in one direction, we may assume that the majority of the fibers in this area are parallel to that direction.
A modification of regular diffusion MRI is diffusion-weighted imaging (DWI). DWI utilizes the measurement of Brownian (or random walk) motion of molecules. Since H-precession (see Nuclear magnetic resonance (NMR)) is proportional to the magnet strength, the protons begin to precess at different rates, resulting in dispersion of the phase and signal loss. Another gradient pulse is applied in the same direction but with opposite magnitude to refocus or rephase the spins. The rephasing will not be perfect for protons that have moved during the time interval between the pulses, and the signal measured by the MRI machine is reduced. This reduction in signal due to the application of the pulse gradient can be related to the strength of diffusion.

Application

Diffusion tensor imaging (DTI) enables diffusion to be measured in all directions and the fractional anisotropy in each direction can be calculated for each voxel. This enables researchers to make brain maps of fiber directions to examine the connectivity of different regions in the brain (using tractography, see ref. 1) and to examine areas of neural degeneration and demyelization in diseases like Multiple Sclerosis.

Following an ischemic stroke, DWI is very sensitive to the changes occurring in the lesion. It is speculated that increases in restriction (barriers) to water diffusion, resulting from cytotoxic edema (cellular swelling), is responsible for the increase in signal on a DWI scan. Other theories, including acute changes in cellular permeability and loss of energy-dependent (ATP) cytoplastic streaming, have been proposed to explain the phenomena. The DWI enhancement appears within 5-10 min after the onset of stroke symptoms and remains for up to two weeks. In contrast, CT often does not detect changes of acute infarct for up to 4-6 hours.

Further, coupling with scans sensitized to cerebral perfusion, one can highlight regions of "perfusion/diffusion mismatch" that may indicate regions capable of salvage by reperfusion therapy. Finally, it has been proposed that diffusion MRI may be able to detect minute changes in extracellular water diffusion and therefore could be used as a tool for fMRI. The nerve cell body enlarges when it conducts an action potential, hence restricting extracellular water molecules from diffusing naturally. However, this idea needs experimental validation.

More info
Diffusion weighted imaging (DWI) uses very fast scans with an additional series of gradients (diffusion gradients) rapidly turned on and off. Protons of water diffusing randomly within the brain, via Brownian

motion, lose phase coherence and thus signal during application of diffusion gradients. In a brain with an acute infarction water diffusion (see Diffusion: general) is impaired, and signal loss on DWI sequences is less than in normal brain. DWI is the most sensitive method of detecting cerebral infarction (stroke) and works within 30 minutes of the ictus (stroke, seizure or attack). See MRI: diffusion MRI for more information.
Like many other specialized applications, this technique is coupled with a fast image acquisition sequence, such as an echo planar imaging sequence (a single 90^0 excitation pulse and a single 180^0 relaxation pulse followed by a low and a high frequent pulse series). The whole sequence yields a whole series of spin echo's, however with a low SNR. Therefore, 2 or 3 sequences are applied for averaging).

References
1. Thurnher MM, Law M. Diffusion-weighted imaging, diffusion-tensor imaging, and fiber tractography of the spinal cord. Magn Reson Imaging Clin N Am. 2009;17:225-44.
2. Bammer R, Holdsworth SJ, Veldhuis WB, Skare ST. New methods in diffusion-weighted and diffusion tensor imaging. Magn Reson Imaging Clin N Am. 2009;17:175-204.

MRI: functional MRI

Principle

Functional magnetic resonance imaging (fMRI) is the use of MRI to measure the hemodynamic response related to neural activity in the brain or spinal cord of humans or other animals. It is one of the more recently developed forms of neuroimaging.
Changes in blood flow and blood oxygenation in the brain are closely related to neural activity. The more active a nerve cell is the more oxygen is consumed, resulting in a local increase in blood flow occurring after a delay of approximately 1-5 s. This hemodynamic response rises to a peak over 4-5 s, before falling back to baseline (and typically undershooting slightly). This leads to local changes in the relative concentration of oxy-Hb and deoxy-Hb and changes in local cerebral blood volume (CBV) and local cerebral blood flow (CBF).
Hb is diamagnetic when oxygenated but paramagnetic when deoxygenated (see Paramagnetism, diamagnetism and magnetophoresis). The blood MR signal is therefore slightly different depending on the amount of oxygenation. These signals can be detected using an appropriate MR pulse sequence as the so-called blood-oxygen-level-dependent (BOLD) contrast. Higher BOLD signals arise from an increased concentration of oxy-Hb. Then the blood magnetic susceptibility (the sensitivity to be magnetized in a magnetic field) more closely matches the tissue magnetic susceptibility.

Fig. 1 4T fMRI scanner.

Neural correlates of BOLD
The blood supply is regulated in space and time to provide the nutrients for brain metabolism. The BOLD signal is thought to be related with neural activity that can be measured electrophysiologically. This allows circuit models that may give insight in brain function.
Local field potentials (see Electrophysiology: general) an index of integrated electrical activity, form a better correlation with blood flow than the spiking action potentials that are most directly associated with neural communication. However, until now, no simple measure of electrical activity has provided an adequate correlation with metabolism and the blood supply across a wide dynamic range. Possibly, the increase in CBF following neural activity is not causally related to the metabolic demands of the brain region, but rather is driven by the availability of neurotransmitters.
The initial small, negative dip before the main positive BOLD signal is well localized (with resolution of about 0.5 mm) and seems to be correlated with measured local decreases in tissue O_2 concentration. In addition, this initial dip occurs within 1-2 seconds of stimulus initiation, which may not be captured when signals are recorded at long stimulus repetition times.

The BOLD signal is composed of CBF contributions from all types of blood vessels up to capillaries. The signal can be weighted to the smaller vessels, and hence closer to the active neurons, by using stronger magnetic fields. For example, whereas about 70% of the BOLD signal arises from larger vessels in a 1.5 T scanner, about 70% arises from smaller vessels in a 4 T scanner. Furthermore, the size of the BOLD signal increases roughly as the square of the magnetic field strength. Several 7 T commercial scanners have become operational, and experimental 8 and 9.4 T scanners are used for experimental work with small animals.

BOLD effects are measured using rapid volumetric acquisition of images with contrast weighted by T2 or T2* (see MRI: T1 and T2 excitation). Such images can be acquired with moderately good spatial resolution (3-6 mm voxels) with images usually taken every 1–4 s. Recently, the use of strong fields and advanced "multichannel" RF reception, have advanced spatial resolution to the mm-scale. The full time course of a BOLD response lasts about 15 s. This can be a sensory or sensory motor response, or a cognitive response, often related to a task.
An fMRI experiment usually lasts between 15 min and 2 hours. Depending on the aim, subjects may view movies, hear sounds, smell odors for sensory research or perform cognitive tasks such as memorization or imagination, press buttons, or perform other tasks.
As with MRI and MEG (Magnetoencephalography), subjects should hold their head immobile (at most motions of 3 mm). Even with short sessions, this is problematic with Alzheimer or schizophrenia patients, and young children.
Theoretically, temporal resolution is some seconds but in practice some tens of seconds due to the weak BOLD signal. This needs averaging of BOLD signals, which are each a response to a single stimulus lasting at most 1 s.

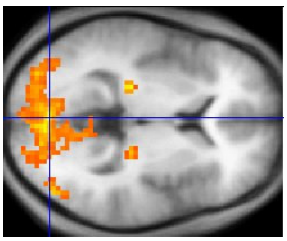


Fig. 2 fMRI data (yellow (or gray) squares) overlaid on the visual cortex.

Application

The response of the blood supply, which is the basis of fMRI, has a poor temporal resolution relative to e.g. the EEG and MEG technique the various types of electrical signals that can be recorded non-invasively or invasively. Therefore, fMRI is often combined with functional brain techniques such as EEG (see Electroencephalography) or other electrophysiological techniques.
Application of fMRI relays on the knowledge of various (neuroscience) disciplines: physics for the principles of MRI and fMRI, electrophysiology for applying EEGs, use of epidural and subdural electrodes (animal models) etc., neuroanatomy since the 2D visualized fMRI signals can be overlaid on an MRI image, and neuropsychology. Most fMRI studies are based on sensory, sensory-motor and cognitive or psychophysical paradigms. Finally, advantaged knowledge of spatiotemporal statistics is needed in order to correctly evaluate the data.
Since its introduction, fMRI has been strongly criticized, both as a research technique and in the way its results have been interpreted. A problem is that the BOLD signal is only an indirect measure of neural activity, and is therefore influenced by non-neural changes in the body.
BOLD signals are strongly associated with the input to a given area than with the output. Different brain areas may have different hemodynamic responses.

More Info

In an fMRI investigation, the BOLD signal is always recorded in rest (no typical brain activity of the subject) and during brain activity, such as sensory or cognitive processing, performing a task (sensory-motor) etc. Such activity should be recorded “time locked”, for instance by using the start of a sensory stimulus or pressing a button as a trigger (time reference) for the BOLD signal. Since a single response in the BOLD to stimulus presentation or task execution is weak, many BOLD signals are averaged. The actual response is superimposed on at the background (rest) BOLD signal. Therefore, the recorded rest BOLD signal is subtracted from the active BOLD signal, yielding the final fMRI response. An example is shown in Fig. 2. There is a large variation of experimental designs. Fig. 3 shows an example of a visual fMRI experiment with the BOLD signal of a particular set of voxels in the cortex.
For cognitive research, the so-called odd-ball paradigm is frequently used: an infrequently and randomly presented (sensory) stimulus (event or odd-ball) in a sequence of periodic stimuli (the frequent stimulus). The task is often to press a button to the odd-ball presentation or counting its occurrence in a

long sequence. (An EEG example is shown in Fig. 1 of Electroencephalography.) Every odd-ball stimulus yields one odd-ball (event) related fMRI response. These fMRI responses (regular or to an odd-ball etc.) are separately averaged.

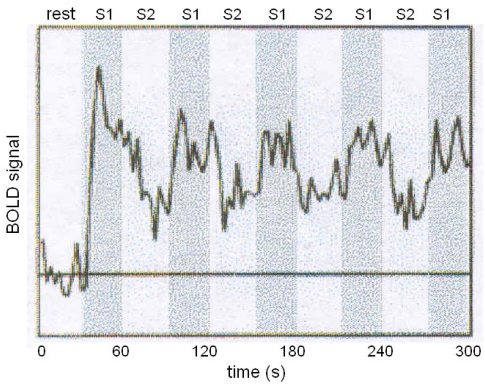


Fig. 3 Result of experiment with intermittent 30 s periods of a series of the visual stimulus S1 and S2.

MRI: other specialized types

Principle and application

Fig. 1 Magnetic resonance angiogram (from Wikipedia).

Magnetic Resonance Angiography
Magnetic resonance angiography (MRA) is used to generate pictures of the arteries in order to diagnose a stenosis or aneurysm. MRA is often used to evaluate the arteries of the neck and brain, the thoracic and abdominal aorta, the renal arteries, and the legs (called a "run-off"). Various techniques can be used to generate images, such as administration of a paramagnetic contrast agent (Ga) or using a technique known as "flow-related enhancement" (e.g. 2D and 3D time-of-flight sequences), where most of the signal on an image is due to blood which has recently moved into that plane, see also the description of FLASH MRI.
With magnetic resonance, venography (MRV) veins can be visualized. Now, the tissue is excited inferiorly while signal is gathered in the plane immediately superior to the excitation plane, and thus imaging the venous blood, which has recently moved from the excited plane.

FLASH MRI
The FLASH technique applies a gradient echo sequence, which combines a low-precession angle RF excitation with a rapid repetition of the basic sequence. The repetition time is usually much shorter than the typical T1 relaxation time. Only the combination of (i) a low-precession angle excitation which leaves unused longitudinal magnetization for an immediate next excitation with (ii) the acquisition of a gradient echo which does not need a further RF pulse that would affect the residual longitudinal magnetization, allows for the rapid repetition of the basic sequence interval and the resulting speed of the entire image acquisition. In fact, the FLASH sequence eliminated all waiting periods previously included to accommodate effects from T1 saturation. FLASH MRI reduced the typical sequence interval to what is

minimally required for imaging: a slice-selective radio-frequency pulse and gradient, a phase-encoding gradient, and a (reversed) frequency-encoding gradient generating the echo for data acquisition. Typical repetition times are on the order of 4-10 ms with image acquisition times of 64-256 repetitions (0.25 to 2.5 s/slice image.

Magnetic resonance spectroscopy
Magnetic resonance spectroscopy (MRS), also known as MRSI (MRS Imaging) and volume selective NMR spectroscopy, is a technique, which combines the spatially addressable nature of MRSI with the spectroscopically-rich information obtainable from NMR. Therefore, MRSI allows one to study a particular region within an organism or sample, but gives relatively little information about the chemical or physical nature of that region. Its main value is distinguishing the properties of that region relative to those of surrounding regions. MRSI provides a pleasure-garden of chemical information about that region, as would an NMR spectrum.

Interventional MRI
The lack of harmful effects on the patient and the operator make MRI well suited for intervention since the MRI images can be used to guide minimally invasive procedures.

Radiation therapy guidance
Because of MRI's superior imaging of soft tissues, it is a major tool to locate tumors within the body in preparation for radiation therapy (RT). For RT, a patient is placed in specific, reproducible, body position and scanned. The MRI system then computes the precise location, shape and orientation of the tumor mass. Next, the patient is marked or tattooed with points which, when combined with the specific body position, will permit precise positioning of the RT-beam. This is especially useful for stereotactic brain RT that requires very precise localization in order to reduce radiation damage...

Magnetic resonance guided focused ultrasound (MRgFUS)
In MRgFUS therapy, ultrasound beams are focused on a tissue, guided and controlled by the use of thermo-sensitive MRI contrast. Due to the high-energy deposition at the focus, temperature within the tissue rises to more than 65°C that destroys the issue. MR imaging provides a 3D-view of the target tissue, allowing for precise focusing of ultrasound energy. Thermal MRI provides quantitative, real-time, thermal images of the treated area. This allows the physician to ensure that the temperature generated during each cycle of ultrasound energy is sufficient to cause thermal ablation within the desired tissue.

Multinuclear imaging
H-based MRI is the common application due to the abundantly presence of H-atoms in biological tissues However, any nucleus which has a net nuclear spin could potentially be imaged with MRI. Such nuclei include ^{3}He, ^{13}C, ^{17}O, ^{23}Na, ^{31}P and ^{129}Xe. ^{23}Na and ^{31}P are naturally abundant in the body, so can be imaged directly. ^{17}O and ^{13}C can be administered in sufficient quantities in liquid form (e.g. ^{17}O-water, or ^{13}C-glucose solutions) that hyperpolarization (nuclear spin polarization far beyond thermal equilibrium conditions) is not a necessity.
Multinuclear imaging is primarily a research technique at present. However, potential applications include functional imaging and imaging of organs poorly seen on ^{1}H MRI (e.g. lungs and bones) or as alternative contrast agents. ^{31}P can potentially provide information on bone density and structure, as well as functional MR imaging of the brain. Inhaled hyperpolarized ^{3}He can be used to image the distribution of air spaces within the lungs. Gaseous isotopes (^{3}He and ^{129}Xe) must be hyperpolarized, as their nuclear density is too low to yield a useful signal under normal conditions. Injectable solutions containing ^{13}C or stabilized bubbles of hyperpolarized ^{129}Xe have been studied as contrast agents for angiography and perfusion imaging.

Experimental MRI techniques
Currently there is active research in several new MRI technologies like magnetization transfer MRI (MT-MRI), and Dendrimer-enhanced MRI as a diagnostic and prognostic biomarker of sepsis-induced acute renal failure.
Current density imaging (CDI) endeavors to use the phase information from images to reconstruct current densities within a subject. CDI works because electrical currents generate magnetic fields, which in turn affect the phase of the magnetic dipoles during an imaging sequence. However, application is music of the future.

MRI: safety technology

Principle

Safety guidelines concern especially avoidance of electric or metal implants and foreign bodies, hypothermia effects due to the RF field, peripheral nerve stimulation and acoustic noise damage. Further, they concern avoidance of moving metal objects close by, use of contrast agents, pregnancy and patient discomfort.

Application

Guidelines
Safety issues, including the potential for biostimulation device interference, movement of ferromagnetic bodies, and incidental localized heating, have been addressed in the American College of Radiology's *White Paper on MR Safety* which was in 2007 under the new title *ACR Guidance Document for Safe MR Practices*.The European Physical Agents (Electromagnetic Fields) Directive is European legislation that has been adopted in European legislature. By 2008, each individual state within the European Union must include this directive in its own law.

More info

Projectiles Because of the very high strength of the magnetic field, missile-effect accidents may happen when ferromagnetic objects are attracted to the center of the magnet. Therefore, ferrous objects and devices are prohibited in proximity to the MRI scanner.

Foreign bodies Ferromagnetic foreign bodies (e.g. shell fragments), or metallic implants (e.g. surgical prostheses, aneurysm clips) are also potential risks, as movement of the object, thermal injury, or failure of an implanted device. These objects are especially problematic when dealing with the eye. Most MRI centers require an orbital x-ray be performed on anyone who suspects they may have small metal fragments in their eyes (e.g. in metalworkers).
Because of its non-ferromagnetic nature and poor electrical conductivity, titanium and its alloys are useful for long-term implants and surgical instruments intended for use in image-guided surgery. Artifacts from metal often appear as regions of empty space around the implant - frequently called 'black-hole artifact' e.g. a 3 mm Ti-alloy coronary stent may appear as a 5 mm diameter region of empty space on MRI, whereas around a stainless steel stent, the artifact may extend for 10-20 mm.

Implants A device or implant that contain magnetic, electrically conductive or RF-reactive components can be considered safe, provided the conditions for safe operation are defined and observed (such as 'tested safe to 1.5 T').
However, cardiac pacemakers are considered an *absolute* contra-indication, though highly specialized protocols have been developed to permit scanning of select pacing devices. In the case of pacemakers, the risk is thought to be primarily RF induction in the pacing electrodes/wires causing inappropriate pacing of the heart, rather than the magnetic field affecting the pacemaker itself.
Other electronic implants have varying contra-indications, depending upon scanner technology, scanning protocols, implant properties and anatomy being imaged.
Scientists are working on a nano-coating for implants. This will screen the implants from RF waves and thus patients with future implants may be examined with MRI scanners.

Radio frequency energy This energy can heat the body significantly, with the risk of hyperthermia in patients, particularly obese patients or patients with thermoregulation disorders. Several countries have issued restrictions on the maximum absorption heat during a scan.

Cryogens Uncontrolled rapid boiling of liquid He (called spontaneous quenching) may result in massive release of He into the scanner room with the risk of hypoxia.

Contrast agents The most frequently used intravenous contrast agents are based on chelates of Ga, a rear earth metal. (A chelate is a macromolecule with a covalent binding of an organic compound with one or more centrally located metal ions.) In general, these agents have proved safer than the iodinated contrast agents used in X-ray radiography or CT. Anaphylactoid reactions are rare (ca. 0.03-0.1%).

Peripheral nerve stimulation The rapid switching (on and off) of the magnetic field gradients needed for imaging is capable of causing nerve stimulation. Twitching sensation when exposed to rapidly switched fields, particularly in extremities, have been reported. The reason the peripheral nerves are stimulated is that the changing field increases with distance from the center of the gradient coils (which more or less coincides with the center of the magnetic field). However, when imaging the head, the heart is far off-center and induction of even a tiny current into the heart must be avoided. The strong, rapidly switched gradients used in e.g. fMRI and diffusion MRI are capable of inducing peripheral nerve

stimulation. American and European regulatory agencies insist that t he gradients are below a specified number of dBs per unit of time.

Acoustic noise Loud noises are produced by minute expansions and contractions of the RF coils resulting from rapidly switched magnetic gradients interacting with the main magnetic field. This is most marked with high-field machines and rapid-imaging techniques in which sound intensity can reach 130 dB (beyond the auditory pain level). Appropriate use of ear protection is essential. Manufacturers are now incorporating noise insulation and active noise cancellation systems on their equipment.

Claustrophobia and discomfort Due to the tube-like construction of MRI scanners, they are unpleasant to lie in. The part of the body being imaged needs to lie at the center of the magnet. Because scan times may be up to one hour, people with even mild claustrophobia are often unable to tolerate an MRI scan. Solutions range from simple preparation (e.g., visiting the scanner to see the room and practice lying on the table), watching DVDs with a head-mounted display, the use of open-bore design scanners, upright MRIs, the use of sedation or even general anesthesia (children).

Pregnancy Harmful effects of MRI on the fetus are unknown, but as a precaution, pregnant women should undergo MRI only when essential, especially during the first trimester of pregnancy, as organogenesis takes place during this period. The fetus may be more sensitive to the effect, particularly to heating. Contrast agents with Ga are known to cross the placenta and enter the fetal bloodstream, and it is recommended that their use be avoided.
Despite these concerns, MRI examination of fetuses is rapidly growing because it can provide more diagnostic information than ultrasound or CT (ionizing radiation).

MRI: T1 and T2 relaxation

Principle

In order to understand MRI contrast, it is important to have knowledge of the time constants (see Halftime and time constant) involved in relaxation processes that establish equilibrium following radio frequent (RF) excitation. The excitation takes about 3 ms. As the high-energy nuclei relax and realign they emit energy at rates which are recorded to provide information about the material they are in. The realignment of nuclear spins along the magnetic field B_0 in the *longitudinal relaxation* and the time required for a fraction of 1/e (37%) of the tissue's nuclei to realign is termed "Time 1" or T1, typically 1 s. *T2-weighted imaging* relies upon local de-phasing of spins following the application of the transverse energy pulse; the *transverse* relaxation time is termed "Time 2" or T2, typically < 100 ms for tissue. De-phasing, which occurs in the transversal plane (x-y), has nothing to do with the longitudinal alignment.
A subtle but important variant of the T2 technique is called T2* imaging. T2 imaging employs a *spin echo* technique (see **More info**), in which spins are rephased (refocused) to compensate for local magnetic field inhomogeneities. T2* imaging is performed without refocusing. This sacrifices some image resolution but provides additional sensitivity to relaxation processes that cause incoherence of transverse magnetization. Applications of T2* imaging include fMRI (see MRI: functional MRI or evaluation of baseline vascular perfusion, e.g. cerebral blood flow (CBF) and cerebral blood volume (CBV) using injected agents. In these cases, there is an inherent trade-off between image quality and detection sensitivity. Because T2*-weighted sequences are sensitive to magnetic inhomogeneity (as can be caused by deposition of Fe-containing blood-degradation products), such sequences are utilized to detect tiny areas of recent or chronic intracranial hemorrhages ("Heme sequence").

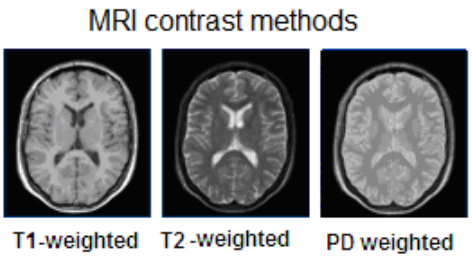


Fig. 1 Coronal brain image obtained with the various relaxation techniques resulting in different contrasts. (Modified after Introduction to MRI, Erik M. Akkerman, 2001.)

Contrast enhancement

Both T1-weighted and T2-weighted images are acquired for most clinical examinations. A more sophisticated image acquisition technique is for instance fat suppression. Another technique is to administer a contrast agent to delineate areas of interest.
A contrast agent may be as simple as water, taken orally, for imaging the stomach and small bowels. But also, substances with specific magnetic properties may be used. Most commonly, a paramagnetic contrast agent (usually a Ga-compound) is given. Ga-enhanced tissues and fluids appear extremely bright on T1-weighted images. This provides high sensitivity for detection of vascular tissues (e.g. tumors) and permits assessment of brain perfusion (e.g. in stroke). Ga-based contrast agents may possibly be toxic for persons with impaired kidney function. This asks for special actions, such as hemodialysis following a contrast MRI.
More recently, superparamagnetic (contrast agents, e.g. Fe oxide nanoparticles) have become available. (Superparamagnetism is the phenomenon that magnetic materials exhibit a behavior similar to paramagnetism even at very low temperatures.) These agents appear very dark on T2*-weighted images (liver). They can also be taken orally, to improve visualization of the gastrointestinal tract, and to prevent water in the tract from obscuring other organs (e.g. pancreas).

Application

Image contrast is created by using a selection of image acquisition parameters that weights signal by relaxation times T1, T2 or T2*, or uses PD, proton-density images. In the brain, T1-weighting causes the nerve connections of white matter to appear white, and the neurons of gray matter to appear gray, while cerebrospinal fluid appears dark. The contrast of white matter, gray matter and cerebrospinal fluid is reversed with T2 or T2* imaging, whereas PD imaging provides little contrast in normal subjects. Additionally, functional information (CBF, CBV, blood oxygenation) can be encoded within T1, T2, or T2*.

More Info

T1 relaxation
Following termination of an RF pulse, nuclei will dissipate their excess energy as heat to the surrounding environment (or lattice) and revert to their equilibrium position. Realignment of the nuclei along B_0, through a process known as recovery, leads to a gradual increase in the longitudinal magnetization. The time taken for a nucleus to relax back to its equilibrium state depends on the rate that excess energy is dissipated to the lattice. With $M_{0\text{-long}}$ the amount of magnetization parallel with B_0 before the 90^0 RF pulse and M_{long} the z component of M at time t, it follows that:

$$M_{long} = M_{0\text{-long}} (1 - e^{-t/T1}) \qquad (1).$$

Table 1

	T1 (ms)	T2 (ms)	PD (rel.)
Fat	260	60	-
White matter	510	67	0.61
Grey matter	760	77	0.69
Edema	900	126	0.86
CSF (Cerebro-Spinal Fluid)	2500	280	1

Values (as examples, they may vary strongly) for T1, T2 and PD. T1 dependents on field strength (here 1.5 T). (From Introduction to MRI, Erik M. Akkerman, 2001.)

T1 relaxation
Following termination of an RF pulse, nuclei will dissipate their excess energy as heat to the surrounding environment and revert to their equilibrium position. Realignment of the nuclei along B_0, through a process known as recovery, leads to a gradual increase in the longitudinal magnetization. The time taken for a nucleus to recover to its equilibrium state depends on the rate that excess energy is dissipated. With $M_{0\text{-long}}$ the amount of magnetization parallel with B_0 before the 90^0 RF pulse and M_{long} the z component of M at time t, it follows that:

$$M_{long} = M_{0\text{-long}} (1 - e^{-t/T1}). \qquad (1)$$

T2 relaxation
While nuclei dissipate their excess energy to the environment after switching off a 90^0 RF pulse, the magnetic moments interact with each other causing a fast decrease in transverse magnetization. This

effect is similar to that produced by magnet inhomogeneity, but on a smaller scale. The decay in transverse magnetization (which does not involve the emission of energy) has the time constant T2*. T2* characterizes de-phasing due to both B_0 inhomogeneity and transverse relaxation:

$$M_{trans} = M_{0\text{-}trans}\, e^{-t/T2^*} \qquad (2)$$

where $M_{0\text{-}trans}$ the amount of transverse (Mx-y) magnetization just after the end of the 90^0 RF pulse (left part Fig. 2).

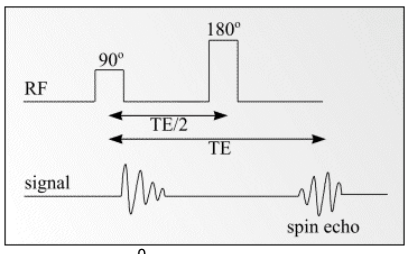


Fig. 2 Formation of the relaxation after the 90^0 RF pulse (left) and a spin echo at time TE after the 180^0 RF pulse (from ref. 1).

Spin echo

In order to obtain signal with T2 dependence rather than a T2* dependence, a pulse sequence known as the spin-echo has been deigned which reduces the effect of B_0 inhomogeneity on Mx-y. A pulse sequence is characterized by several parameters, of which the main ones are the repetition time (TR), the echo time (TE), flip angle, the number of excitations, bandwidth and acquisition matrix.

Fig. 2 and 3 show how the spin echo pulse sequence works. Fig. 2 is a graph of pulsed RF and received signal verses time, while Fig. 3 is a phase diagram of the magnetization vector M. After a 90^0 pulse, a signal is formed which decays with T2* characteristics. This is illustrated by the top right ellipse in Fig.3, which shows three spins at different phases due to their different precessional frequencies. The fastest spin is labeled f and the slowest s. At time TE/2, an 180^0 pulse is applied to the sample (see bottom left ellipse in Fig. 3) which causes the three spins to invert. After inversion, the order of the spins is reversed with the fastest lagging behind the others. At time TE, the spins become coherent again (re-phasing) so that a signal, the spin echo is produced.

If a next 180^0 pulse is applied at time TE/2 after the peak signal of the first 180^0 spin echo pulse, then a second spin echo signal will form at time TE after the first spin echo. The peak amplitude of each spin echo is reduced from its previous peak amplitude due to T2 dephasing, which cannot be rephased by the 180^0 pulses. Fig. 4 shows how the signal from a spin echo sequence decays over time. A line drawn through the peak amplitude of a large number of spin echoes describes the T2 decay, while individual spin echoes exhibit T2* decay.

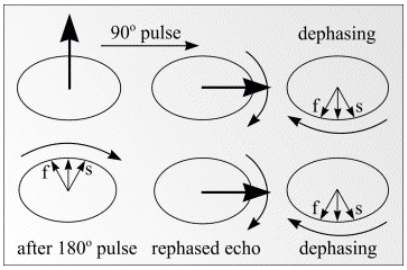


Fig. 3 Dephasing of the magnetization vector by T2* and rephasing by a 180 degree pulse to form a spin echo (from ref. 1).

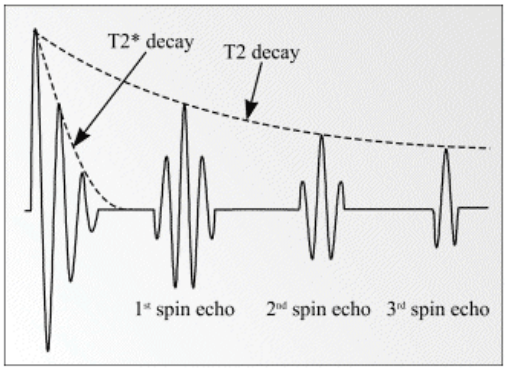


Fig. 4 Decay of signal with time in a spin echo sequence (from ref. 1).

Signal strength decays with time to varying degrees depending on the different materials in the sample. Different organs have different T1s and T2s (Table 1) and hence different time constants of the signals. When imaging anatomy, some degree of control of the contrast of different organs or parts of organs is possible by varying pulse repetition time TR and echo time TE.

MRI Sequences
Different pathologies can be selected by the correct choice of pulse sequence parameters. For a given type of nucleus in a given environment, TR determines the amount of T1 relaxation. The longer the TR, the more the longitudinal magnetization is recovered. Tissues with short T1 have greater signal intensity than tissues with a longer T1 at a given TR. A long TR allows more magnetization to recover and thus reduces differences in the T1 contribution in the image contrast. Echo time TE is the time from the application of an RF pulse to the measurement of the MR signal. TE determines how much decay of the transverse magnetization is allowed to occur before the signal is processed. It therefore controls the amount of T2 relaxation. The application of RF pulses at different TRs and the receiving of signals at different TEs produce variation in contrast in MR images.
Next, some common MRI sequences are described.

Spin Echo Pulse Sequence
The spin echo (SE) sequence is the most commonly used pulse sequence in clinical imaging. The sequence comprises two RF pulses, the 90^0 pulse that creates the detectable magnetization and the 180^0 pulse that rephases it at TE. The selection of TE and TR determines resulting image contrast. In T1-weighted images, tissues that have short T1 relaxation times (such as fat) present as bright signal. Tissues with long T1 relaxation times (such as cysts, cerebrospinal fluid and edema) show as dark. In T2-weighted images, tissues with long T2 relaxation times (such as fluids) are presented as bright in the image.
In cerebral tissue, differences in T1 relaxation times between white and grey matter permit the differentiation of these tissue. PD-weighted images also allow distinction of white and grey matter, but with tissue signal intensities mirroring those obtained on T2-weighted images. In general, T1-weighted images provide excellent anatomic detail, while T2-weighted images are better for detecting pathology. The typical MRI examination consists of 5-20 sequences, each of which is chosen to provide a particular type of information.

Gradient Recalled Echo Pulse Sequences
Gradient recalled echo (GRE) sequences, which are significantly faster than SE sequences, differ from SE sequences since there is no 180^0 refocusing RF pulse. In addition, the single RF pulse in a GRE sequence is usually switched on for less time than the 90^0 pulse used in SE sequences. The scan time can be reduced by using a shorter TR, but this is at the expense of a smaller signal to noise ratio (SNR). At the interface of bone and tissue or air and tissue, there is an apparent loss of signal that is heightened as TE is increased. Therefore, it is usually inappropriate to acquire T2-weighted images with the use of GRE sequences. Nevertheless, GRE sequences are widely used for obtaining T1-weighted images for a large number of slices or a volume of tissue in order to keep scanning times to a minimum. GRE sequences are often used to acquire T1-weighted 3D volume data that can be reformatted to display image sections in any plane.

On a T2-weighted scan, fat-, water- and fluid-containing tissues are bright (most modern T2 sequences are actually *fast* T2 sequences). T2-weighted sequence is sensitive for edema and so distinguishes pathologic tissue from normal tissue. With an additional RF pulse and additional manipulation of the magnetic gradients, a T2-weighted sequence can be converted to a FLAIR sequence (see below), in which free water is now dark, but edematous tissues remain bright. This sequence is in particular suitable for examining demyelinating diseases, e.g. multiple sclerosis.

FLAIR (Fluid Attenuated Inversion Recovery) is a specific pulse sequence, a so-called inversion recovery technique that nulls fluids. For example, it can be used in brain imaging to suppress cerebrospinal fluid in order to high-light periventricular hyperintense lesions, such as multiple sclerosis plaques. By carefully choosing the TI, the signal from any particular tissue (often fat tissue) can be nulled. TI is the inversion time, the time of sending a pulse preceding the excitation pulse. The appropriate TI depends on the tissue:

$$TI = \ln 2 * T1$$

One should typically yield a TI of 70% of T1. In the case of cerebrospinal fluid suppression, one aims for T2 weighted images.

Literature
1. http://www.easymeasure.co.uk/principlesmri.aspx.
2. Basic principles of MRI, Philips Medical systems.
3. Introduction to MRI, Erik M. Akkerman, AMC, Amsterdam, 2001

MUGA scan

Principle

A MUGA scan (Multiple Gated Acquisition Scan) is an examination of nuclear medicine to evaluate the function of the ventricles and so the health of the heart's major pumping chambers. It is more accurate than an echocardiogram and it is non-invasive.

At a high level, the MUGA test involves the introduction of a radioactive marker into the bloodstream, *in vivo* or *in vitro*. The patient is subsequently scanned to determine the circulation dynamics of the marker, and hence the blood. *In vivo* method, Sn (tin) ions are injected and a subsequent intravenous injection of the radioactive ^{99m}Tc-pertechnetate labels the red blood cells. With an administered activity of about 800 MBq (megabequerel; 1 Bq is one nuclear decay per second), the effective radiation dose is about 8 mSv (sievert, see CT scan (dual energy)). In the *in vitro* method, some the patient's blood is drawn and SnCl is injected into the drawn blood and then with the ^{99m}Tc. The Sn ions dilute the ^{99m}Tc and prevent it from leaking out of the red blood cells.
In vivo the patient is placed under a Gamma camera, which detects the low-level 140keV gamma radiation being given off by ^{99m}Tc. As the gamma camera images are acquired, the patient's heartbeat is used to 'gate' (line up) the acquisition. The final result is a series of images of the heart (usually 16), one at each stage of the cardiac cycle. The resulting images show the blood pool in the chambers of the heart and from the images, the left ejection fraction and other clinical can be calculated.

Application

MUGA's can be made in rest or during exercise (e.g. on a cycle ergometer) for suspected coronary artery disease. In some rare cases, a nitroglycerin MUGA may be performed, where nitroglycerin (a vasodilator) is administered prior to the scan.
An uneven distribution of ^{99m}Tc indicates coronary artery disease, a cardiomyopathy, or blood shunting. Abnormalities in a resting MUGA usually indicate a heart attack, while those that occur during exercise usually indicate ischemia. In a stress (exercise) MUGA, patients with coronary artery disease may exhibit a decrease in ejection fraction. For a patient that has had a heart attack, or is suspected of having another disease that affects the heart muscle, this scan can help to examine the degree of damage of the heart. MUGA scans are also used to evaluate heart function prior to and while receiving certain chemotherapies with known effects.

More Info

MUGA is typically indicated for the following patients:
- with known or suspected coronary artery disease (diagnosis and predict outcomes);
- with lesions in the heart valves;
- with a recent heart attack (damage and recidivous risk);
- with congestive heart failure;
- with low cardiac output after open-heart surgery;
- who have undergone percutaneous transluminal coronary angioplasty, coronary artery bypass graft surgery, or medical therapy, to assess the efficacy of the treatment;
- who are undergoing chemotherapy.

PET

Principle

Positron emission tomography (PET) is a nuclear imaging technique, which produces a 3D image or map of functional molecular processes in the body. PET is a clinical and a biomedical research tool.

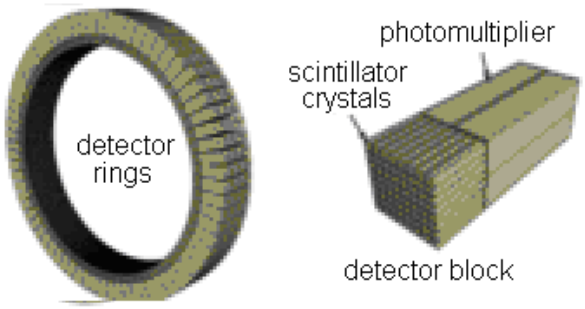


Fig. 1 Schematic view of a detector block and ring of a PET scanner.

A short-lived radioactive tracer isotope (radionuclide), which decays by emitting a positron (positive electron, the antimatter counterpart of an electron), also called positive β-decay, is chemically incorporated into a metabolically active molecule (the radiotracer). This molecule is injected into the living subject (usually into blood of a patient). The molecule most commonly used is the sugar ^{18}F-deoxyglucose (FDG). There is a waiting period while the metabolically active molecule becomes concentrated in tissues of interest; then the patient is placed in the imaging scanner. Radionuclides used in PET scanning are typically isotopes with short half-lives such as ^{11}C (20 min), ^{13}N (10 min), ^{15}O (2 min), and ^{18}F (110 min). Due to their short half-lives, the radionuclides must be produced in a nearby cyclotron (a circular charged-particle accelerator).
After annihilation (see Gamma camera) two (sometimes more) gamma photons are produced moving in almost opposite directions. These are detected when they reach a scintillator (see Gamma camera) material in the scanning device, creating a burst of light which is detected by photomultiplier tubes or nowadays often by silicon photodiodes. The technique depends on the simultaneous or coincident detection of the photon-pair. Photons, which do not arrive in pairs (i.e., within a few ns) are ignored.

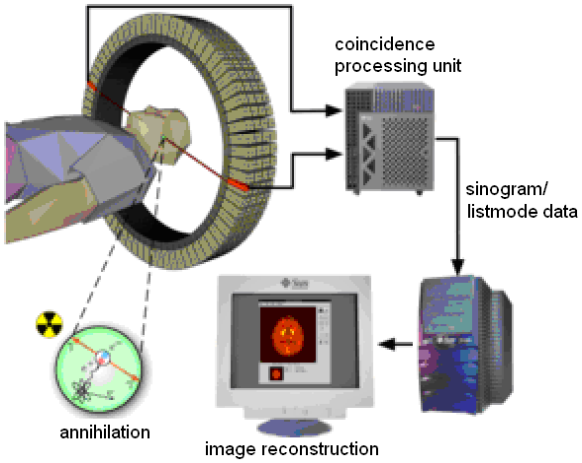


Fig. 2 Schema of a PET acquisition process

The PET scanner examines detailed molecular biological functions via the use of radiolabelled molecular probes that have different rates of uptake. This depends on the type and function of tissue involved. The changing of regional blood flow in various anatomic structures (as a measure of the injected positron emitter) can be visualized and relatively quantified with a PET scan.
PET imaging is best performed using a dedicated PET scanner. However, it is possible to acquire PET images using a conventional dual-head Gamma camera fitted with a coincidence detector (lower quality, slower).
Image reconstruction The raw data are a list of the annihilation photons by a pair of detectors. Each coincidence event represents a line in space connecting the two detectors along which the positron emission occurred.

Coincidence events can be grouped into projections images, called sinograms. The sinograms are sorted by the angle of each view and tilt. The sinogram images are analogous to the projections captured by CT scanners, and can be reconstructed similarly. However, the statistics of the data is much worse than those obtained through transmission tomography. A normal PET data set has millions of counts for the whole acquisition, while the CT reaches a few billion counts. Furthermore, PET data suffer from scatter and random events much more than CT data.
In practice, considerable pre-processing of the data is required - correction for random coincidences, estimation and subtraction of scattered photons, detector dead-time correction (after the detection of a photon, the detector must "cool down" again) and detector-sensitivity correction (for both inherent detector sensitivity and changes in sensitivity due to angle of incidence). For more details, see **More Info**.

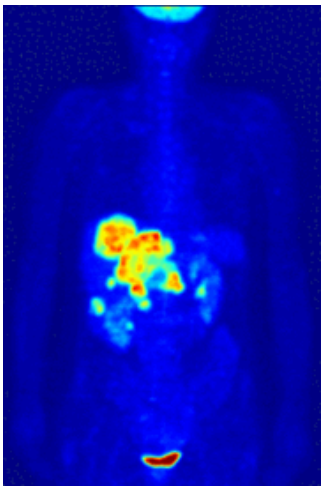


Fig. 3 Maximum intensity projection (MIP, see Image processing: 3D reconstruction) of a typical ^{18}FDG nearly whole body PET acquisition.

Limitations Limitations of PET are ethical since radioactive material is injected. Most radionuclides require a cyclotron and automated chemistry lab for radiopharmaceutical production in the hospital. This limitation restricts clinical PET primarily to the use of tracers labeled with ^{18}F or ^{82}Rb, due to their longer half-lifes. The latter is used for myocardial perfusion studies. Mobile cyclotrons with hot labs are a new development that also reduces cost.
Safety PET scanning is non-invasive, but it involves exposure to ionizing radiation. The total dose is small, usually ca. 7 mSv. This can be compared to about 2.2 mSv average annual background radiation, 0.02 mSv for a chest X-ray, up to 8 mSv for a CT scan of the chest and 2-6 mSv per annum for aircrew (data from UK National Radiological Protection Board).

Application

PET is a worthwhile technique for several diseases and disorders, because it is possible to target the radio-chemicals used for particular functions and processes.
Neurology/neuroscience PET neuro-imaging is based on areas of higher radioactivity. What is actually measured indirectly is the flow of blood to different parts of the brain by a ^{15}O tracer. However, because of its 2-minute half-life, ^{15}O must be piped directly from a medical cyclotron, and this is difficult and expensive. Therefore, standard FDG-PET can also be successfully used, e.g. to differentiate Alzheimer's disease (via visualization of amyloid plaques) from other dementing processes.
Specific radiotracers (i.e. radioligands) have been developed for PET that are ligands for neuroreceptor subtypes (for e.g. dopamine D2, serotonin 5-HT1A, etc.), and in addition transporters (such as for serotonin in this case), or enzyme substrates. These agents permit the visualization of neuroreceptor pools in the context of neuropsychiatric and neurological illnesses.
Psychiatry Numerous compounds that bind selectively to neuroreceptors have been radiolabeled with ^{11}C or ^{18}F. Radioligands that bind to serotonin receptors (5HT1A, 5HT2A, reuptake transporter), dopamine receptors (D1, D2, reuptake transporter), opioid receptors and other receptors have been used successfully.
Oncology FDG-PET scanning with ^{18}FDG is performed with a typical dose of 200-400 MBq. It results in intense radiolabeling of tissues with high glucose uptake, such as the brain, the liver, and most cancers.
Pharmacology In pre-clinical trials, it is possible to radiolabel a new drug, inject it into animals and monitored in vivo.

Cardiovascular system So-called "hibernating myocardium", and imaging atherosclerosis to detect patients at risk of stroke.
PET is increasingly used together with CT or MRI. Modern PET scanners are integrated high-end multi-detector-row CT scanners. The two types of scans can be performed during the same session, with the patient not changing position. This is important with respect of overlying both 3D images.

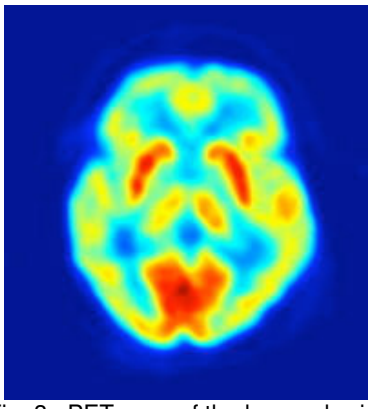


Fig. 3 PET scan of the human brain.

More Info

The largest fraction of electron-positron decays result in two 511 keV gamma photons being emitted at almost 180 ^{o} to each other, the "line of response" or LOR. In practice, the LOR has a small imperfection, as the emitted photons are not exactly 180^{o} apart. If the recovery time of detectors is in the ps-range rather than the 10's of ns range, it is possible to calculate the single point on the LOR at which an annihilation event originated. This is done by measuring the "time of flight" of the two photons. Nowadays, this technology is available on some new scanners. More commonly, a technique very like the reconstruction of CT and SPECT data is used, although more difficult (see below). Using statistics collected from tens-of-thousands of coincidence events, a set of simultaneous equations for the total activity of each voxel of tissue along many LORs can be solved by a number of techniques. Thus, a map of radioactivities as a function of location of voxels may be constructed and imaged. The resulting map shows the tissues in which the molecular probe has become concentrated, and can be interpreted by nuclear clinician or radiologist in the context of the patient's diagnosis and treatment plan.

Imaging
Preprocessing Filtered back projection (FBP) has been frequently used (simple but sensitive to low-level discrete noise (shot noise) forming streaks across the image). Iterative expectation-maximization algorithms are the preferred method of reconstruction, although requiring higher computer power. (For these techniques see imaging textbooks and for instance Wikipedia.)
Attenuation correction As different LORs must traverse different thicknesses of tissue, the photons are attenuated differentially. The result is that deep structures are reconstructed as having falsely low tracer uptake. Contemporary scanners can estimate attenuation using integrated CT equipment. Since correction is itself susceptible to significant artifacts both corrected and uncorrected images are always interpreted together.
2D/3D reconstruction Modern scanners have a cylinder of scanning rings. There are two approaches: individual reconstruction per ring (2D reconstruction) and the same extended with coincidence detection between rings (3D), and next reconstruction of the entire volume together (3D). 3D coincidence techniques have better sensitivity (because more coincidences are detected and used) and therefore less noise, but are more sensitive to the effects of scatter and random coincidences, as well as requiring correspondingly greater computer resources.

Planck's law

Principle

Planck's law describes the spectral radiance of electromagnetic radiation at all wavelengths from a black body at temperature T (in K). As a function of frequency v (in Hz), Planck's law is written as:

$$I(\nu,T)=\frac{2\nu^2}{c^2}\frac{h\nu}{e^{\frac{h\nu}{kT}}-1} \qquad (1)$$

where
I is spectral radiance ($W\cdot m^{-2}\cdot s^{-1}\cdot Hz^{-1}$; sr is steradian, the SI unit of solid angle. A sphere comprises 4π steradians, since its surface is $4\pi r^2$ (see Radian and steradian and Light: units of measure);
h is Planck's constant (J/Hz);
c is speed of light (m/s);
k is Boltzmann's constant (J/K).
The probability to emit a photon strongly decreases with the frequency v. The reason is that excitation is more rare the higher the exited orbit. This is expressed in the denominator of the second term. In the classical theory (Rayleigh-Jeans law), the second term is kT. Both laws agree well as long as hv << kT.
The shape of the spectrum in wavelength (λ) notation (with I in $W\cdot m^{-3}\cdot sr^{-1}$), or more commion $W\cdot m^{-2}\cdot sr^{-1}\cdot nm^{-1}$) is:

$$I(\lambda,T)=\frac{2hc^2}{\lambda^5}\frac{1}{e^{\frac{hc}{\lambda kT}}-1}. \qquad (2)$$

See for a graphic presentation of I as function of λ and T Fig. 1 of Wien's displacement law.
Although Planck's formula predicts that a black body will radiate energy at all frequencies, the formula is only applicable when many photons are being measured. For example, a black body at room temperature (300 K) with 1 m^2 of surface area will emit a photon in the visible range once about every thousand years or so, meaning that for most practical purposes, a black body at room temperature does not emit in the visible range.

Application

In medicine the application, together with Planck's law of black body radiation (see Planck's law) and Wien's displacement law is calculating heat transfer of body radiation (see Heat dissipation) in e.g. space and environmental medicine, and medical thermographic imaging (see Thermography).
In the above applications, the body generates the radiation. As human donor of radiation there are several fields of medical applications (see Wien's displacement law).

More info

Note that in literature the equations (1) and (2) can vary by a constant factor (often π) depending on the derivation. The two functions have different units; the first is radiance per unit frequency interval while the second is radiance per unit wavelength interval. Hence, the quantities I(vT) and I(λ,T) are not equivalent to each other. To derive one from the other, they cannot simply be set equal to each other (i.e. the expression for λ in terms of v cannot just be substituted into the first equation to get the second). However, the two equations are related through:

$$I(v,T)dv = -I(\lambda,T)d\lambda. \qquad (3)$$

The law is sometimes written in terms of the spectral energy density:

$$u(\nu,T)=\frac{4\pi}{c}I(\nu,T) \qquad (4)$$

which is in units of $J\cdot m^{-3}Hz^{-1}$. Integrated over frequency, this expression yields the total energy density. The spectral energy density can also be expressed as a function of wavelength:

$$u(\lambda,T)=\frac{4\pi}{c}I(\lambda,T). \qquad (5)$$

The radiation field of a black body may be thought of as a photon gas, in which case this energy density would be one of the thermodynamic parameters of that gas. (In physics, a photon gas is a gas-like collection of photons, which has many of the same properties of a conventional gas like H_2 or He, including pressure, temperature, and entropy.(see Thermodynamics: entropy)) The most common example of a photon gas in equilibrium is black body radiation.

Raman spectroscopy

Principle

Raman Spectroscopy is a technique used in (mainly) solid-state physics and in chemistry to study vibrational, rotational, and other low-frequency energy modes in a system. Raman techniques are frequently applied to biological material (cells, proteins etc.). It relies on inelastic scattering, or Raman scattering of monochromatic light, usually from a Laser in the visible, near-IR, or near-UV range. IR spectroscopy yields similar, but complementary information. Raman scatter is proportional to the 4^{th} power of wavelength.

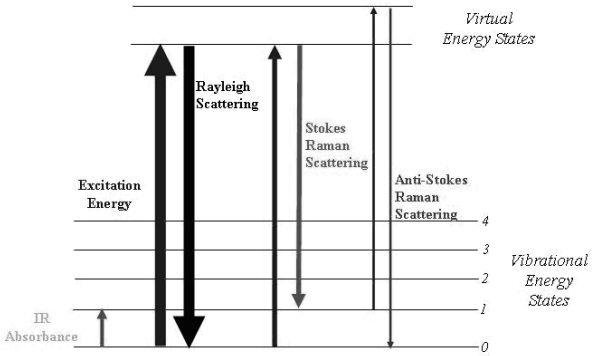


Fig. 1 Rayleigh and Raman scattering visualized with the energy states of the atom.

When light is scattered from an atom or molecule, most photons are elastically scattered (Rayleigh scattering, see Light: scattering) as illustrated in Fig. 1. The scattered photons have the same energy (frequency) and, therefore, wavelength, as the incident photons. However, a small fraction of scattered light (approximately 1 in 1000 photons) is scattered from excitations with optical frequencies different from, and usually lower (the arrow in Fig. 1 labeled with Stokes Raman Scattering) than the frequency of the incident photons. This type of scattering is shifted towards longer wavelengths. In a gas, Raman scattering can occur with a change, mostly an increase, in vibrational, rotational or electronic energy of a molecule. The following is amore precise description of the process.
The Raman effect occurs when light impinges upon a molecule and interacts with the electron cloud of the chemical bonds of that molecule. A molecular polarizability change or amount of deformation of the electron cloud, with respect to the vibrational coordinate is required for the molecule to exhibit the Raman effect. The amount of the polarizability change will determine the intensity, whereas the Raman shift, the shift in wavelength, is equal to the vibrational level that is involved. The incident photon (light quantum), excites one of the electrons into a virtual state. For the spontaneous Raman effect, the molecule will be excited from the ground state (0) to a virtual energy state, and relax into a vibrational excited state (e.g. 1), which generates (Stokes) Raman scattering (Fig. 1). If the molecule was already in an elevated vibrational energy state and relaxes in a lower state, the Raman scattering is then called anti-Stokes Raman scattering (shift to shorter wavelengths).

Application

In science
Raman spectroscopy is used in chemistry, since vibrational information is very specific for the chemical bonds in molecules. It therefore provides a molecule-specific spectrum by which the molecule can be identified. The spectral region of organic molecules e ranges 5 to 20 µm. Another way that the technique is used is to study changes in chemical bonding, e.g. when a substrate is added to an enzyme.
In solid-state physics, spontaneous Raman spectroscopy is used to, among other things, characterize materials and measure temperature. Raman scattering by an anisotropic crystal gives information on the

crystal orientation. The polarization of the Raman scattered light can be used to find the orientation of the crystal.

In medicine
Raman gas analyzers are for instance used for real-time monitoring of anesthetic and respiratory gas mixtures during *surgery*.
Spatially Offset Raman Spectroscopy (SORS), which is less sensitive to surface layers than conventional Raman, can be used for non-invasive monitoring of biological tissue.
Raman microspectroscopy
The advantage of Raman spectroscopy, since it is a scattering technique, is that samples do not need to be fixed or sectioned. Raman spectra can be collected from a very small volume (< 1µm in diameter); these spectra allow the identification of molecules present in that volume. Water does not interfere very strongly. Therefore, it is suitable for the microscopic examination of cells and proteins.
In *direct imaging*, the whole field of view is examined for scattering over a small number of wavelengths (Raman shifts). For instance, a wavelength characteristic for cholesterol gives info about the distribution of cholesterol within a cell culture.
The other approach is *hyperspectral imaging*, in which thousands of Raman spectra are acquired from all over the field of view. The data can then be used to generate images showing the location and amount of different components. The image could show the distribution of cholesterol, as well as proteins, nucleic acids, and fatty acids. Sophisticated signal- and image-processing techniques can be used to ignore the presence of water, culture media, buffers, and other interferents.
Raman microscopy
Lateral and depth resolutions with confocal Raman microscopy (see Microscopy: confocal) may be as low as 250 nm and 1.7 µm, respectively. Since the objective lenses of microscopes focuses the laser beam to several microns in diameter, the resulting photon flux is much higher than achieved in conventional Raman setups. Since the high photon flux can also cause sample degradation, a thermally conducting substrate for heat absorption is achieved.
In vivo, time- and space-resolved, Raman spectra of microscopic regions of samples can be measured. Now, the fluorescence of water, media, and buffers can be removed. Consequently, time- and space-resolved Raman spectroscopy is suitable to measure tissues, cells, e.g. erythrocytes and proteins.
Raman microscopy for biological and medical specimens generally uses near-IR lasers. Due to the low efficacy of Raman scatter, the laser needs a high intensity. To prevent tissue damage and improve efficacy, near-IR is used ($E \sim 1/\lambda$ and the intensity of the scattered beam is proportional with λ^4).

More Info

Measurement principle
Typically, a sample is illuminated with a laser beam. Light from the illuminated spot is collected with a lens and sent through a monochromator. Wavelengths close to the laser line (due to elastic Rayleigh scattering) are filtered out and those in a certain spectral window away from the laser line are dispersed onto a detector.
Molecular structure analysis
Phonons or other excitations in the system are absorbed or emitted by the laser light, resulting in the energy of the laser photons being shifted up or down. (A phonon is a quantized mode of vibration occurring in a rigid crystal lattice, such as the atomic lattice of a solid.) The shift in energy gives information about the phonon modes in the system.
Raman microscopy
A Raman microscope comprises a standard optical microscope, an excitation laser, a monochromator and a sensitive detector (such as a charge-coupled device (CCD) or photomultiplier tube).
Spontaneous Raman scattering is typically very weak, and as a result the main difficulty of Raman spectroscopy is separating the weak inelastically scattered light from the intense Rayleigh scattered laser light. Raman spectrometers typically use holographic diffraction gratings (see Holography) and multiple dispersion stages to achieve a high degree of laser rejection. A photon-counting photomultiplier tube or, more commonly, a CCD camera is used to detect the Raman scattered light. In the past, photomultipliers were the detectors of choice for dispersive Raman setups, which resulted in long acquisition times. However, the recent uses of CCD detectors have made dispersive Raman spectral acquisition much more rapid.
Raman spectroscopy has a stimulated version, analogous to stimulated emission, called stimulated Raman scattering.
In addition to the discussed types of Raman spectroscopy there exist some other seven types (see Wikipediia).

References
1. Wikipedia/Raman Spectroscopy (main source of knowledge);
http://en.wikipedia.org/wiki/Raman_spectroscopy

SPECT

Principle

Single photon emission computed tomography (SPECT) is a nuclear medicine tomographic imaging technique using gamma rays. It is very similar to conventional nuclear medicine planar imaging using a Gamma camera. However, it is able to provide true 3D information. This information is typically presented as cross-sectional slices through the patient, but can be freely reformatted or manipulated as required.
In the same way that a plain X-ray is a 2D view of a 3D structure, the image obtained by a gamma camera image is a 2D view of 3D distribution of a radionuclide. This is often ^{99m}Tc, a metastable (indicated by "m") nuclear isomer radionuclide (radioisotope, see Gamma camera) which emits gamma rays which can be detected by a gamma camera. It is coupled to an organ specific tracer molecule.

Fig. 1 Philips Precedence SPECT-CT machine (from Philips Healthcare systems).

SPECT imaging is performed by using a gamma camera to acquire multiple 2D images (also called projections, some 60-120), from multiple angles. Next, a tomographic reconstruction algorithm applied to the multiple projections yields a 3D dataset. Processing of this dataset shows thin slices along any chosen axis of the body, similar to those obtained from other tomographic techniques, such as MRI, CT (see CT scan (dual energy)), and PET.
Because SPECT acquisition is very similar to planar gamma camera imaging, the same radiopharmaceuticals may be used. If a patient is examined in another type of nuclear medicine scan but the images are non-diagnostic, it may be possible to proceed straight to SPECT by moving the patient to a SPECT instrument, or even by simply reconfiguring the camera for SPECT image acquisition while the patient remains on the table.
To acquire SPECT images, the gamma camera is rotated around the patient. Projections are acquired at defined points during the rotation, typically every 3-6^{o}. In most cases, a full 360^{o} rotation (otherwise 180^{o}) is used to obtain an optimal reconstruction. The time taken to obtain each projection is also variable, but 15 – 20 s is typical. This gives a total scan time of 15-20 min.
Multi-headed gamma cameras can provide accelerated acquisition. For example, a dual headed camera can be used with heads spaced 180^{o} apart, allowing 2 projections to be acquired simultaneously, with each head requiring 180^{o} of rotation. Triple-head cameras with 120^{o} spacing are also used.

Application

SPECT can be used to complement any gamma imaging study, where a true 3D representation can be helpful (tumor, infection, thyroid, bone). With accurate 3D localization, it can be used for functional imaging, e.g. myocardial perfusion imaging
Myocardial perfusion imaging (MPI) MPI used for the diagnosis of ischemic heart disease, is a type of cardiac stress test (restricted myocardium blood flow). A cardiac specific radiopharmaceutical is e.g. a technetium derivate, ^{99m}Tc-tetrofosmin.
Gated myocardial SPECT Triggered by ECG to obtain differential information about the heart in various parts of its cycle, gated myocardial SPECT can be used to obtain quantitative information about myocardial perfusion, thickness, and contractility of the myocardium during various parts of the cardiac cycle; and also to allow calculation of left ventricular ejection fraction, stroke volume, and cardiac output. The Technique is similar as a MUGA scan.

Functional brain imaging Usually the gamma-emitting tracer used in functional brain imaging is ^{99m}Tc-HMPAO (hexamethylpropylene amine oxime). When it is attached to HMPAO, this allows ^{99m}Tc to be taken up by brain tissue in a manner proportional to brain blood flow, in turn allowing brain blood flow to be assessed with the nuclear gamma camera.
Because blood flow in the brain is tightly coupled to local brain metabolism and energy use, the ^{99m}Tc-HMPAO tracer (as well as the similar ^{99m}Tc-EC tracer) is used to assess brain metabolism regionally, in an attempt to diagnose and differentiate the different causal pathologies of dementia. SPECT is able to differentiate Alzheimer's disease from vascular dementias. ^{99m}Tc-HMPAO SPECT scanning competes with FDG (2-fluoro-2-deoxy-D-glucose) PET scanning of the brain, which works to assess regional brain glucose metabolism, to provide very similar information about local brain damage from many processes. SPECT is more widely available, for the basic reason that the radioisotope generation technology is longer lasting and far less expensive in SPECT, and the gamma scanning equipment is less expensive as well. The reason for this is that ^{99m}Tc is extracted from relatively simple ^{99m}Tc generators, which are delivered to hospitals, and scanning centers weekly, to supply fresh radioisotope.
Ictal-interictal SPECT Analysis by SPM (ISAS) is a specific application of functional brain imaging with SPECT. SPM is statistical parametric mapping, a technique for examining differences in brain activity recorded during functional neuro-imaging experiments using neuro-imaging technologies such as SPECT, fMRI or PET. The goal of ISAS is to localize the region of seizure onset for epilepsy surgery planning. ISAS is an objective tool for analyzing ictal versus interictal SPECT scans.

More Info

Limitations
Scanning is time consuming, and it is essential that there is no patient movement during the scan time. Movement can cause significant degradation of the reconstructed images, although movement compensation reconstruction techniques can help. A highly uneven distribution of radiopharmaceutical also has the potential to cause artifacts. A very intense area of activity (e.g. the bladder) can cause extensive streaking of the images and obscure neighboring areas of activity.
Attenuation of the gamma rays within the patient can lead to significant underestimation of activity in deep tissues, compared to superficial tissues. Approximate correction is possible, based on relative position of the activity. However, optimal correction is obtained with measured attenuation values.

Typical SPECT acquisition protocols

Study	*Radio-isotope*	*Emission energy (keV)*	*Half-life (h)*	*Radio-pharmaceutical*
Bone	^{99m}Tc	140	6	Phosphon-ates / Bisphos phonates
Myo-cardial perfusion	^{99m}Tc	140	6	tetrofosmin; Sestamibi
Brain	^{99m}Tc	140	6	HMPAO; ECD
Tumor	^{123m}I	159	13	MIBG
White cell	^{111m}In	171 & 245	67	*in vitro* labelled leucocytes

Reconstruction
Reconstructed images typically have resolutions of 64x64 or 128x128 pixels, with the pixel sizes ranging from 3-6 mm. The number of projections acquired is chosen to be approximately equal to the width of the resulting images. In general, the resulting reconstructed images will be of lower resolution, have increased noise than planar images, and be sensitive to artifacts.
SPECT-CT
Modern SPECT equipment is available with an integrated X-ray CT scanner (Fig. 1). As X-ray CT images are an attenuation map of the tissues, this data can correct the SPECT map for attenuation. It also provides a precisely registered CT image, which can provide additional anatomical information.
Compton SPECT
This is a new development in which solely the Compton effect (increase in wavelength of scattered photon when an X- gamma ray photon interacts with matter) is used to generate the photons.

References
Elhendy *et al*., Dobutamine Stress Myocardial Perfusion Imaging in Coronary Artery Disease, J Nucl Med 2002 43: 1634-1646 (review).
W. Gordon et al. (2005). "Neuroreceptor Imaging in Psychiatry: Theory and Applications". International Review of Neurobiology, 67: 385-44.

Spectroscopy

Basic principle

Spectroscopy is the study of spectra of electro-magnetic radiation (Fig. 1) in the quality of quanta or particles, performed with a spectrometer.

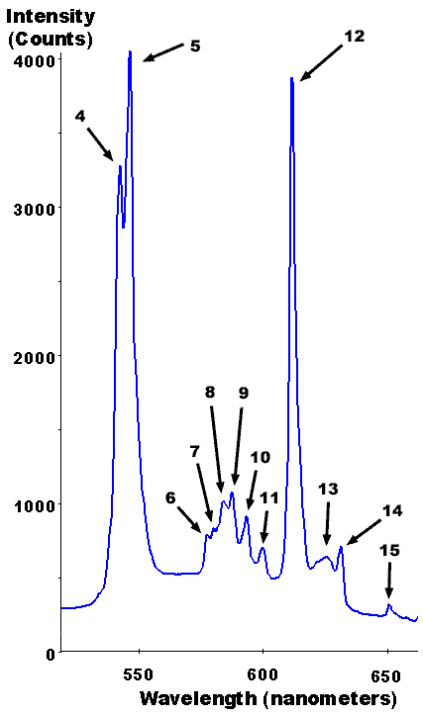


Fig.1 Fluorescence spectrum showing prominent mercury peaks

Spectroscopy can be classified according to the method of measurement. The general principle is that a (biological) sample is bombarded by radiation or by particles and that, the output is radiation or particles and that holds for both types of input. The physical process is absorbance and/or emittance. The electro-magnetic spectra are basically line spectra, but in practice, they are often a continuous spectrum with many peaks (Fig. 1). In the context of medicine it is mostly performed with visible light (Vis), infrared (IR) light and UV light. Spectroscopy is often used for the identification of molecular substances by emission or absorption. Fig. 2 shows the basic set up.

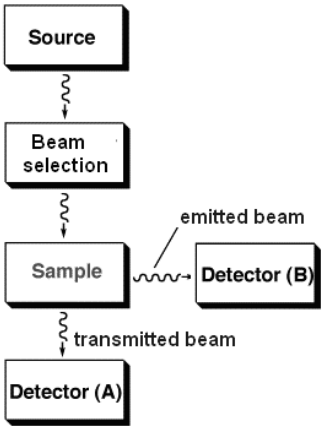


Fig. 2 Note that the Beam selection device becomes unnecessary if a Laser is used as light source (laser light is monochromatic).

Physical quantity measured

The type of spectroscopy depends on the physical quantity measured; normally the amount emitted or absorbed electro-magnetic radiation, but also the amplitude of small vibrations (acoustic spectroscopy). Spectroscopically useful radiation covers the whole electromagnetic spectrum from high energy X-rays to low energy radio frequencies. With nuclear magnetic and electron spin, resonance the signal is the change of the spin. The microwave principle is the interaction between the oscillating electric or magnetic field and molecules that can undergo a change in dipole moment. Vibration or bending of a bond gives rise to vibrational spectra (IR- and Raman spectra). Visible and UV spectroscopy are based on absorption and emission due to a change of the energy states of outer electron orbitals. With X-ray (emission) spectroscopy, the energy state of inner electron orbitals provide the signal of interest. Gamma spectroscopy is a radiochemistry measurement method that determines the energy and count rate of gamma rays emitted by radioactive substances. Fig. 3 visualizes the physics of these atomic or molecular effects. A vast number of different radiation sources have accordingly been developed, in many cases for the specific purpose of generating radiation for spectroscopic applications.

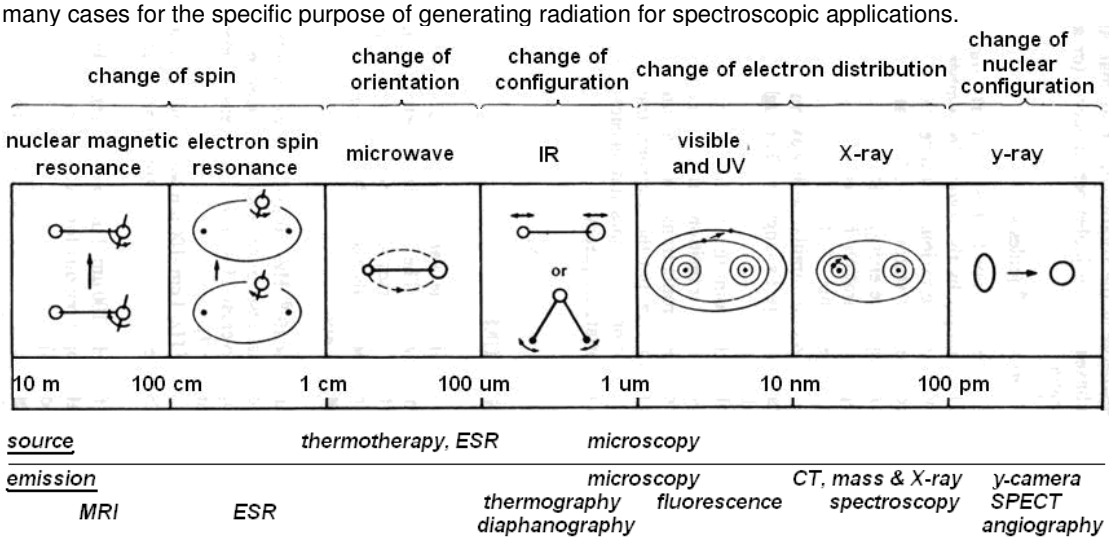


Fig. 3 The electro-magnetic spectroscopic methods with their wavelengths. Below the panel, important medical applications are mentioned, indicated by the radiation *source* and the type of *emission*.

Application

Applications are many in basic medical research and (bio)chemistry (see Fig. 3).
There are many spectroscopic techniques for element analysis, which are also useful for biomaterials, like atomic absorption, atomic fluorescence, Vis and UV spectroscopy (the latter two often combined), which causes often photo-ionization. IR and NMR spectroscopy are typically used for structure-analysis. However, the most relevant application of NMR is medical imagery (MRI).

More Info

Main types of spectroscopy

Absorption spectroscopy uses the range of electromagnetic spectra in which a substance absorbs. In atomic absorption spectroscopy, the sample is atomized and then light of a particular frequency is passed through the vapor. The amount of absorption can be related to the concentrations of various metal ions through the Lambert-Beer law, such as Na, K and Ca in blood. UV/Vis absorption spectroscopy is performed on liquid samples to detect molecular content and IR spectroscopy is performed on liquid, semi-liquid (paste or grease), dried, or solid samples to determine molecular information, including structural information.

Emission spectroscopy uses the range of electromagnetic spectra in which a substance radiates. The substance first absorbs energy and then radiates this energy as light, evoked by for instance chemical reactions (e.g. Bioluminescence) or light of a lower wavelength (Fluorescence).

Scattering spectroscopy measures certain physical properties by measuring the amount of light that a substance scatters (see Light: scatter) at certain wavelengths, incident angles, and polarization (see: Light: polarization) angles. One of the most useful applications of light scattering spectroscopy is Raman spectroscopy.

Table 1 Characteristics of the techniques (After I.R. Beattie, *Chem. Soc. Rev.*, 1975, *4*, 107)

Technique	Nature of the effect	Information	Potential Problems
X-ray diffraction	Scattering, mainly by electrons, followed by interference (λ = 0.01-1 nm)	Electron density map of crystal	Difficult to locate light atoms in presence of heavy atoms. Difficult to distinguish atoms with similar number of electrons
Neutron diffraction	Scattering, mainly by nuclei, followed by interference (λ = 0.1 nm)	oriented internuclear distances	Best method to locate hydrogen atoms. Allows determination of magnetic ordering. Requires large crystals (1mm and up)
Electron diffraction	Diffraction mainly by nuclei, but also by electrons (λ = 0.01-0.1 nm)	Scalar distances due to random orientation	Thermal motions cause blurring of distances. Molecular model required. Gas phase and surface only
Microwave Spectroscopy	Absorption of radiation due to dipole change during rotation (λ = 0.1-30 cm; 300-1 GHz in frequency)	Mean value of r^{-2} terms; potential function	Mean value of r-2 does not occur at r_e even for harmonic motion. Dipole moment necessary. Only one component may be detected. Analysis difficult for large molecules of low symmetry
IR Spectroscopy	Absorption of radiation due to dipole change during vibration ($\lambda = 10^{-1}$-10^{-4} cm)	Symmetry of molecule Functional groups	Useful of characterization. Some structural information from number of bands, position and possibly isotope effects. All states of matter
Raman Spectroscopy	Scattering of radiation with changed frequency due to polarizability change during a vibration (λ = visible usually)	Symmetry of mole-cule. Functional groups	Useful for characterization. Some structural information from number of bands, position, depolarization ratios, and possibly isotope effects. All states of matter
Electronic Spectroscopy (UV-Vis)	Absorption of radiation due to dipole change during an electronic transition ($\lambda = 10$-10^2 nm)	Qualitative for large molecules	All states of matter
Nuclear Magnetic Resonance Spectro-scopy	Interaction of radiation with a nuclear transition in a magnetic field ($\lambda = 10^2$-10^7 cm; 3 KHz to 300 MHz	Symmetry of molecule through number of magnetically equivalent nuclei Many many others	Applicable to solutions and gases. In conjunction with molecular weight measurements may be possible to choose one from several possible models
Mass Spectroscopy	Detection of fragments by charge/mass	Mass number,frag-mentation patterns	Gas phase only. Fragmentation pattern changes with energy of excitation
Extended X-ray absorption fine structure (EXAFS)	Back scattering of photoelectrons off ligands	Radial distances, number, and types of bonded atoms	Widely used for metallo enzymes and heterogeneously supported catalysts

Other common types of spectroscopy (see also Table 1)

Fluorescence spectroscopy uses higher energy photons to excite a sample, which will then emit lower energy photons. This technique has become popular for its biochemical and medical applications, and can be used for Confocal scanning microscopy.
Mass spectrometry, see Mass spectrography.
X-rays spectroscopy and X-ray crystallography are specialized types to unravel structures of e.g. biomolecules.
Flame Spectroscopy Liquid solution samples are aspirated into a burner or nebulizer/burner combination, desolvated, atomized, and sometimes excited to a higher energy electron-state. The use of a flame during analysis requires fuel and oxidant, typically in the form of gases. These methods analyses metallic element in the part per million or billion (ppm or ppb), or even lower concentration ranges. Light detectors are needed to detect light with the analysis information coming from the flame.
Atomic Emission Spectroscopy This method uses flame excitation that excites atoms from the heat of a very hot flame to emit light. A high resolution polychromator (a device to disperse light into different directions to isolate parts of the spectrum), can be used to produce an emission intensity versus wavelength showing multiple element excitation lines. So, multiple elements can be detected in one run.

Alternatively, a monochromator (device that transmits a selectable narrow band of wavelengths of light chosen from a wider range of wavelengths available at the input) can be adjusted to one wavelength for analysis of a single element based on its emission line. Plasma emission spectroscopy is modern version of this method.
Nuclear Magnetic Resonance spectroscopy NMR spectroscopy analyzes certain atomic nuclei to determine different local environments of H, C, O_2 or other atoms in the molecule of an (organic) compound, often for structure-analysis at a microscopic or macroscopic scale (medical NMR).

These are various modifications of the above basic types.

Literature
http://131.104.156.23/Lectures/CHEM_207/structure_and_spectroscopy.htm

Stefan-Boltzmann law

Principle

The Stefan-Boltzmann law, also known as Stefan's law, states that the total radiated power per unit surface area of a black body, j* (in J/(s·m^2 or W/m^2) is directly proportional to the fourth power of the black body's absolute temperature *T*:

$$j^* = \sigma T^4, \quad (1)$$

where σ, the constant of proportionality, the Stefan-Boltzmann constant or Stefan's constant. This power is also known variously as the black-body irradiance, energy flux density, radiant flux, or the emissive power.

Since most radiating bodies are not black, a more general expression is:

$$j^* = \sigma T^4, \quad (2)$$

with ε the emissivity of the source. It is the fraction of emitted energy compared to that of a black body.

Application

In medicine the application, together with Planck's law of black body radiation (see Planck's law) and Wien's displacement law is calculating heat transfer of radiation of a biological body (see Heat dissipation) in e.g. space and environmental medicine, and medical thermographic imaging (see Thermography).
In the above applications, the body generates the radiation. With the human body as donor there are several fields of applications in cancer therapy, including dermatology (see Wien's displacement law).

More Info

The Stefan-Boltzmann constant is non-fundamental in the sense that it derives from other known constants of nature:

$$\sigma = \frac{2\pi^5 \cdot k^4}{15 \cdot c^2 \cdot h^3} \approx 5.6704 \cdot 10^{-8} \text{ (W/m}^2\text{)}, \quad (3)$$

where
k is Boltzmann's constant (J/K);
c is speed of light (m/s);
h is Planck's constant (J/Hz).

Thermography

Principle

Thermography, or digital infra red (IR) thermal imaging (DITI), is a type of IR imaging. Thermographic cameras detect IR radiation (roughly 0.9-14µm) and produce images of the radiating body. IR radiation is emitted by all objects based on their temperature, according to the laws of (black body) electromagnetic radiation (see Stefan-Boltzmann law and Wien's displacement law) . Thermography makes it possible to "see" one's environment with or without visible illumination. The amount of radiation emitted by an object increases with temperature. Therefore, thermography allows one to see variations in temperature. An IR scanning device is used to convert IR radiation emitted from the object surface into electrical impulses that are visualized in color on a monitor. This visual image graphically maps the surface temperature and is referred to as a thermogram. With a thermographic camera, warm objects stand out well against cooler backgrounds. Humans and other warm-blooded animals become easily visible against the environment day or night; hence, historically its extensive use can be ascribed to military and security services.

Medical DITI is a noninvasive diagnostic technique that allows the examiner to visualize and quantify changes in skin surface temperature. Since there is a high degree of thermal symmetry in the normal body, subtle abnormal temperature asymmetry's can be easily identified.

Application

Medical DITI's major clinical value is in its high sensitivity to pathology in the vascular, muscular, neural and skeletal systems and as such can contribute to the pathogenesis and diagnosis made by the clinician. Attractive is its completely non-invasive nature and the use of a body-generated signal.

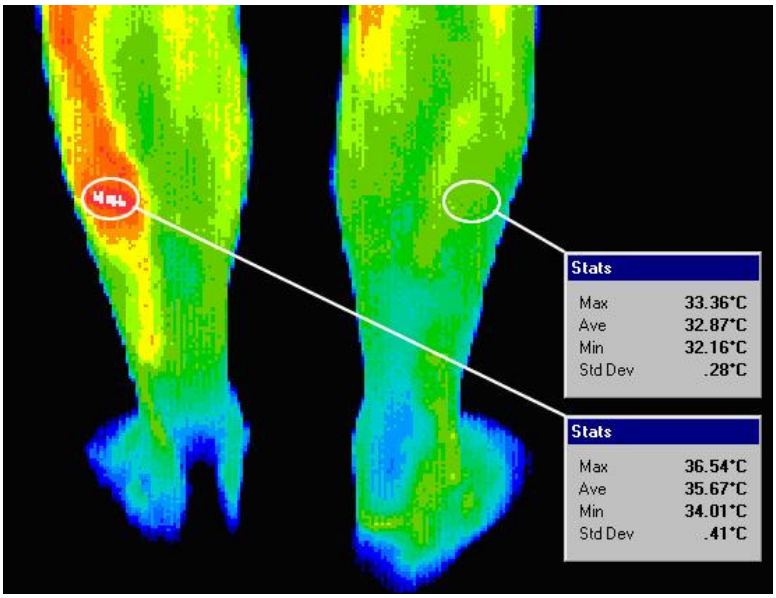


Fig. 1 A left sural muscle injury

Skin blood flow is under the control of the sympathetic nervous system. In healthy people, there is a symmetrical dermal pattern, which is consistent and reproducible for any individual. This is recorded in precise detail with a temperature sensitivity of 0.1°C by DITI. The neuro-thermography application of DITI measures the somatic component of the sympathetic response by assessing dermal blood flow. This response appears on DITI as a localized area of altered temperature, up to even 10° C with specific features for each anatomical lesion.

Regular imaging applications are in dermatology and sports medicine, e.g. early lesions before they are clinically evident and monitor the healing process before the patient is returned to work or training. A limitation is the restricted depth of imaging. Yet, there are techniques to measure deep venous thrombosis.
Further, applications in rheumatology

Another application is early diagnostics of breast cancer (IR mammography) and superficial neuro-musculo-skeletal examinations).

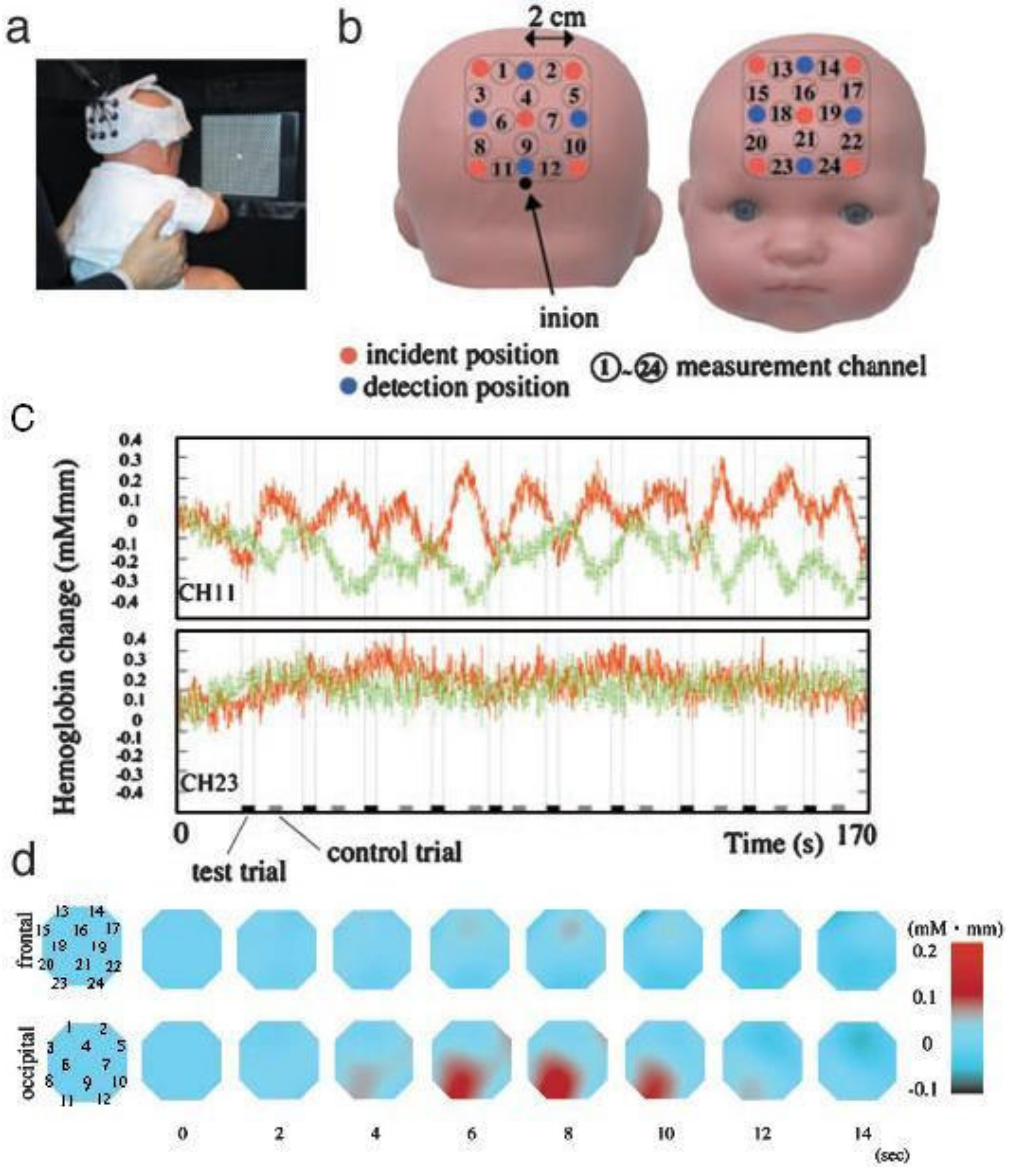


Fig. 2 Functional near-IR optical topography a). Child seated for a PC screen. b) channel configuration with emitting and receiving channels. c) Time series of relative changes in [oxy-Hb] (red or dark gray signal) and [deoxy-Hb] (green or light gray signal) at channel 11 over the occipital cortex and at channel 23 over the frontal cortex of a 2-month-old infant. d) response to visual stimulation. Epoch-averaged images of [oxy-Hb] over the occipital and frontal cortex of a 4-month-old infant are illustrated at 2-s intervals. Scale is relative [oxy-Hb] change from an arbitrary zero baseline. From ref. 1.

Neurophysiological activity can also be monitored with thermographic imaging, for instance brain activity, especially in infants as an alternative of the expensive fMRI technique. Actually it is an application in the field of spectroscopy, more precisely a multichannel near-IR optical topography measuring time courses of the levels of oxy-Hb (780 nm) and deoxy-Hb (830 nm) (Fig. 2).

Literature

1) http://www.pnas.org/cgi/reprint/100/19/10722
2) http://www.meditherm.com/therm_page1.htm
3) Maldague XPV et al. "Chapter 2: Fundamentals of Infrared and Thermal Testing: Part 1. Principles of Infrared and Thermal Testing," in Nondestructive Handbook, Infrared and Thermal Testing, Volume 3, X. Maldague technical ed., PO Moore ed., 3rd edition, Columbus, Ohio, ASNT Press, 2001, 718 p.

Transcranial magnetic stimulation

Principle

Transcranial magnetic stimulation (TMS) is a noninvasive method to excite neurons in the brain by weak electric currents, electromagnetically induced in the tissue by rapidly changing strong magnetic fields. This is possible since tissue (which is mainly water) is about as transparent for a magnetic field as air. In this way, the brain can be stimulated without the need for surgical implantation of intracranial electrodes (see Electrophysiology: general) or the use of external scalp electrodes, which give unspecific stimulation.
It is used to study the normal (experimental research) as well as pathological (clinical examination) global circuitry and connectivity of the cortex.

The generated currents will depolarize neurons in a selected part of the brain, leading to changes in the generation of action potentials (see Bioelectricity: action potential). Essentially, the effect of TMS is to change the synaptic currents... There are no morphological or heating effects that may damage the tissue
Scientists and clinicians are attempting to use TMS to replace electroconvulsive therapy (ECT, also known as electroshock, a controversial method) to treat disorders such as severe depression. Instead of one strong electric shock through the head as in ECT, a large number of relatively weak pulses are delivered in TMS treatment, typically at the rate of about 10 Hz. If very strong pulses at a rapid rate are delivered to the brain, the induced currents can cause convulsions (e.g. to treat depression). TMS is painless because the induced current does not pass through the skin, where most of the pain fibre nerve endings are located. Additionally, the currents induced by magnetic stimulation are relatively diffuse and hence the high current densities that occur underneath electrodes used for electrical stimulation do not occur.

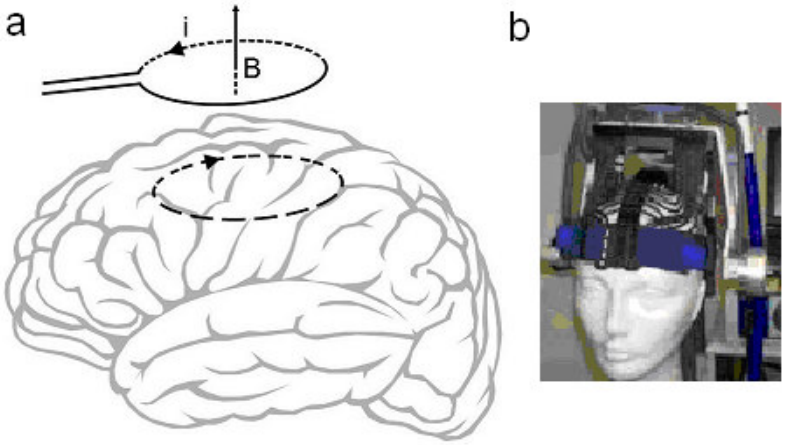


Fig. 1 a. Principle of the technique. b. A commercial device.

The TMS device is basically a coil of wire, encased in plastic, which is placed on the scalp (Fig. 1). When the coil is energized by the fast discharge of a large capacitor, a rapidly changing current flows in its windings. This produces a magnetic field oriented orthogonally to the plane of the coil. The magnetic field passes unimpeded through the skin and skull, inducing a current in the brain in the oppositely direction to that of the loops in the coil. In this way, the current flows tangentially with respect to skull (Fig. 1). The current induced in the structure of the brain activates nearby nerve cells in much the same way as currents applied directly to the cortical surface. The path of this current is complicated to model because the brain is a non-uniform conductor (white and gray matter, ventricles) with an irregular shape due to the gyri and sulci. With stereotactic MRI-based control (see MRI: general), the precision of targeting TMS can be approximated to a few mm. Some typical parameters of the stimulation are:

- magnetic field often about 2 Tesla on the coil surface and 0.5 T in the cortex,
- capacitor voltage some kV and discharge coil peak current some kA ,
- current rise time from zero to peak around 0.07-0.1 ms,
- wave form monophasic or biphasic with a magnetic field energy of some hundred J,
- repetition rate for slow TMS < 1 Hz, but typically some kHz, applied as pulse trains.

Depending on the application TMS coils can be round, double-conical for deep (>1.5 cm) TMS and two coils placed next to each other (figure-eight or butterfly), resulting in a more focal pattern of activation below the midline were both coils come together such that both equally directed magnetic fields add.

By stimulating different points of the cerebral cortex and recording responses, e.g. from muscles, one can obtain maps of functional brain areas. By measuring Electroencephalography (EEG) in between stimulation periods, information is obtained about the effect on cortical functioning and about area-to-area connections.

The effects of TMS can be divided into two types depending on the mode of stimulation:
Single or paired pulse TMS The pulse(s) causes a population of neurons in the neocortex to depolarize and discharge an action potential. If used in the primary motor cortex, it produces a motor-evoked potential (MEP), which can be recorded with Electromyography (EMG). If used on the occipital cortex, 'phosphenes' (imaginary flashes of light) might be detected by the subject. In most other areas of the cortex, the participant does not consciously experience any effect, but his or her behavior may be slightly altered (e.g. slower reaction time on a cognitive task), or changes in brain activity may be detected using PET or fMRI (see MRI: functional MRI). These effects do not outlast the period of stimulation.
Repetitive TMS (rTMS) rTMS produces effects which last longer than the period of stimulation. rTMS can increase or decrease the excitability of corticospinal or corticocortical pathways depending on the intensity of stimulation, coil orientation and frequency of stimulation. The mechanism of these effects is not clear. Possibly it changes the synaptic efficacy (see Bioelectricity: chemical synapse) akin to long-term potentiation (long-lasting enhancement in communication between two neurons that results from repetitive and simultaneous stimulating them) and long-term depression (weakening of a synapse that lasts from hours to days). TMS and rTMS are used for different purposes.

Application

TMS for diagnostic purposes
TMS is used clinically to measure activity and function of specific brain circuits in humans. The most robust and widely-accepted use is in measuring the connection between the primary motor cortex and a muscle by measuring amplitude and latency of the magnetic evoked potential (MEP).
This is most useful in stroke, spinal cord injury, multiple sclerosis and motor neuron disease. Success for other disorders is doubtful
TMS for therapeutic purposes
There is hardly evidence for the use of TMS and rTMS for therapy of any disorder.
Only few studies for particular disorders have been confirmed and most show very modest effects. TMS is particularly attractive as a potential treatment for drug resistant mental illness, particularly as an alternative to electroconvulsive therapy.

Risks
As it induces an electrical current in the human brain, TMS and rTMS can produce a seizure. The risk is very low with TMS except in patients with epilepsy and patients on medications. The risk is significantly higher, but still very low, in rTMS especially when given at rates >5Hz at high intensity.
The only other effects of TMS, which are reported in most subjects, are discomfort or pain from the stimulation of the scalp and associated nerves and muscles on the overlying skin. Further, TMS pulses can generate loud clicks.

More Info

One reason TMS is important in cognitive and emotional neuropsychology/ (e.g. brain plasticity) is that it can demonstrate causality.
A noninvasive mapping technique such as fMRI allows researchers to see what regions of the brain are activated when a subject performs a certain task. However, this is not proof that those regions are actually used for the task. It merely shows that a region is *associated* with a task. If activity in the associated region is suppressed (i.e. 'knocked out') with TMS stimulation and a subject then performs worse on a task, this is much stronger evidence that the region is *used* in performing the task.
This 'knock-out' technique (also known as virtual lesioning) can be done in two ways:
Online TMS: where subjects perform the task and at a specific time (usually 100-200 ms after starting) of the task, a TMS pulse is given to a particular part of the brain. This should affect the performance of the task specifically, and thus demonstrate that this task involves this part of the brain at this particular time point. The advantage of this technique is that any positive result can provide a lot of information about how and when the brain processes a task, and there is no time for a placebo effect or other brain areas to compensate. The disadvantages of this technique is that in addition to the location of stimulation, one also has to know roughly when the part of the brain is responsible for the task so lack of effect is not conclusive.
Offline repetitive TMS: where performance at a task is measured initially and then repetitive TMS is given over a few minutes, and the performance is measured again. This technique has the advantage of not requiring knowledge of the timescale of how the brain processes. However, repetitive TMS is very

susceptible to the placebo effect. Additionally, the effects of repetitive TMS are variable between subjects and also for the same subject.

A highly speculative cause of the claimed effects of TMS is the possible existence of long-lived rotational states of some molecules inside protein structures.

References

1. Fitzgerald PB et al. (2006). "A comprehensive review of the effects of rTMS on motor cortical excitability and inhibition". Clinical Neurophysiology **117** (12): 2584–96.
2. Marangell LB, Martinez M, Jurdi RA, Zboyan H (September 2007). "Neurostimulation therapies in depression: A review of new modalities". Acta Psychiatrica Scandinavica **116** (3): 174–81

Wien's displacement law

Principle

Wien's displacement law states that there is an inverse relationship between the wavelength of the peak of the emission of a black body and its temperature. A black body is an object that absorbs all electromagnetic radiation that falls onto it. No radiation passes through it and none is reflected. Despite the name, black bodies are not actually black as they radiate energy as well since they have a temperature larger than 0 K. The law states:

$$\lambda_{max} = b/T, \quad (1)$$

where
λ_{max} is the peak wavelength (nm),
T is the temperature of the blackbody (K) and
b is Wien's displacement constant, $2.898 \cdot 10^6$ (nm·K).
Wien's law states that the hotter an object is, the shorter the wavelength at which it will emit most of its radiation. For example, the surface temperature of the sun is on average 5780 K. Using Wien's law, this temperature corresponds to a peak emission at λ = 500 nm. Due to a temperature gradient in the surface boundary layer and local differences the spectrum widens to white light. Due to the Rayleigh scattering (see Light: scattering) of blue light by the atmosphere this white light is slightly separated, resulting in a blue sky and a yellow sun.

Application

In medicine the application, together with Planck's law of black body radiation (see below) and Stefan-Boltzmann law is calculating the heat loss due to human body radiation (see Heat dissipation) in e.g. space and environmental medicine, and medical thermographic imaging (see Thermography). Stefan-Boltzmann law calculates the total radiated power per unit surface area of a black body. In the above applications the body generates the radiation.
As human acceptor of thermal radiation, applications are thermo radiotherapy and thermo chemotherapy of cancers, microwave/thermo therapies of tumors, low level laser therapy (LLLT, Photobiomodulation), and laser surgery (see Laser). With these applications, energy transfer and wavelength dependency are the primary interests. Further the technology of IR physical therapy equipment and application is (medical) cosmetic equipment (skin) and costumer's applications (IR and UV apparatus, sunglasses for UV A and B).

More Info

A light bulb has a glowing wire with a somewhat lower temperature, resulting in yellow light, and something that is "red hot" is again a little less hot.

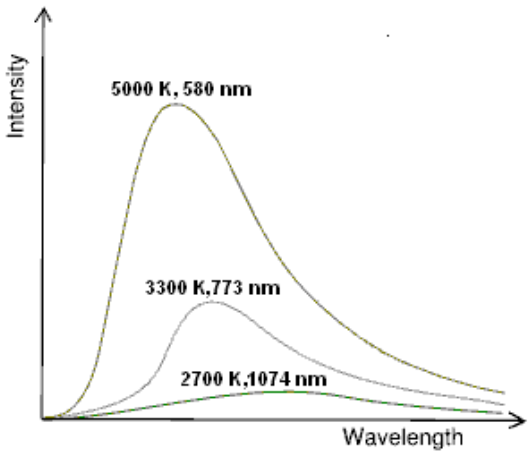


Fig. 1 As the temperature decreases, the peak of the black body radiation curve moves to lower intensities and longer wavelengths.

The shape of the spectrum in λ-notation is given by Planck's law of black body radiation (see there).

In terms of frequency f (Hz), Wien's displacement law becomes:

$v_{max} = \alpha \cdot k \cdot T/h \approx 58.79 \cdot 10^9 \cdot T$ (Hz/K), (3)

where v_{max} is the peak frequency (Hz) and $\alpha \approx 2.821$ is a constant resulting from the numerical solution of the maximization equation.
Because the spectrum resulting from Planck's law of black body radiation takes a different shape in the frequency domain from that of the wavelength domain, the frequency location of the peak emission does **not** correspond to the peak wavelength using the simple relationship between frequency, wavelength, and the speed of light.
For more info, for instance equations in the frequency domain and Planck's law of black body radiation see by Planck's law and e.g. Wikipedia.

X-ray machine

Principle

An X-ray imaging system consists of an X-ray source, an X-ray tube, a detector system and an imaging system. When a classical photographic film (see More Info) is used, the detector system is at the same the imaging system, e.g. in mammography this is an X-ray photograph of the breasts. Nowadays, the imaging system is a PC.

First, an electron beam is emitted from a heated cathode filament. This beam is focused and accelerated towards an angled anode target, called the focal spot. Only around 1% of the kinetic energy of the electrons is converted into X-ray photons and the excess heat is dissipated via a heat sink.
At the focal spot, X-ray photons are emitted in all directions from the target surface with the highest intensity between 60 ^{o} and 90^{o} due to the angle of the anode target. There is a small window in the X-ray tube directly above the angled target, allowing the X-ray beam to exit the tube with little attenuation.
The beam is projected on a target, in medicine a part of the body. The object absorbs part of the beam and a part passes through the object. The latter part will produce the image. The more photons pass the blacker the image on a classical transparent cassette film, the X-ray photograph or radiograph.
Areas where radiation is absorbed show up as lighter shades of grey.

Fig. 1 Philips BV Libra X-ray machine with fluoroscopy system (from http://www.medical.philips.com/main/products/index.wpd).

Applications

In medicine
Diagnostics
X-ray machines are used in radiology for visualizing bone structures (e.g. for osteoporosis and bone fractures) and other dense tissues such as teeth and tumors. Surgical mobiles can produce images continuously.
The two main diagnostic fields in which X-ray machines are used are radiography and dentistry. Radiography is used for fast, highly penetrating images, and is mostly used in areas with high bone content. Some forms of radiography include the so-called orthopantomogram (a panoramic X-ray of the jaw showing all teeth at once), mammography (breast tissue and tomography (see CT scanners).
Specific applications are:

- Fluoroscopy This is used when real-time visualization is necessary. In fluoroscopy, imaging of the digestive tract is done with the help of a radio-contrast agent, e.g. $BaSO_4$, which is opaque to X-rays;
- Angiography The examination of blood vessels in real time, see Angiography and DSA;
- Barium enema The examination of colon and lower gastrointestinal tract with $BaSO_4$;
- barium swallow The examination of the upper gastrointestinal tract with $BaSO_4$;
- Biopsy The removal of tissue for examination.

Therapeutics
Radiotherapy This is the use of X-ray radiation to treat malignant cancer cells, a non-imaging application.

Non-medical

These applications include X-ray crystallography, material science, food inspection, archeology and security inspections (luggage at airports and e.g. museums and student baggage at schools) and material analysis.

More info

Detection
A photographic film with rare earth metal compounds detects the X-ray beam with photons with an energy generally less than 450 kV by semiconductor detectors, or by X-ray image intensifiers. The latter are used in a design called a C-arm. The C-arm has an image intensifier system and an X-ray tube positioned directly opposite from each other. The C-arm can be positioned in many angles. It allows live images on a TV or PC screen. (See for more details Wikipedia, X-ray image intensifiers.)
In the clinic, a typical stationary radiographic X-ray machine also includes an ion chamber and grid. The ion chamber is basically a hollow plate located between the source and the patient. It determines the level of exposure by measuring the amount of X-rays that have passed through the electrically charged, gas-filled gap inside the plate. This allows for minimization of patient radiation exposure but also that an image is not underdeveloped. The grid is usually located between the ion chamber and patient and consists of many aluminum slats stacked next to each other (resembling a Polaroid lens). In this manner, the grid allows straight X-rays to pass through to the detection medium but absorbs reflected X-rays. This improves image quality by preventing scattered (non-diagnostic) X-rays from reaching the detection medium. However, using a grid results in a higher overall radiation dose.

Photographic film
The X-ray film comprises a mounting layer with at both sides an emulsion layer of 10-30 µm with silver bromide (SiBr) grains of a mean diameter of 1 µm (Fig. 2). The AgBr is composed of its ions Ag^+ and Br^-. By an X-ray photon, the Ag-ion is excited. This holds generally only for a few ions in the grain of 10^9 – 10^{10} AgBr molecules. During developing process, the Ag^+ ions of the grain together with all the Ag atoms, which were not excited, are converted to the black metallic Ag. During the subsequent fixation process the AgBr molecules of the grains without excites Ag-ions are washed out to prevent false blackening. One photon gives blackening of a constant umber of grains. Consequently, the blackening is proportional with the intensity on the film (with constant exposure time). The absorbing power of the film is small since the density of the grains is poor. Therefore, only some percent of the radiation is absorbed. Thickening the emulsion layers does not help since then the fixation liquid does not reach the deeper part of the layer and more grains affect the balance. However, during many decades the image quality is optimized by adjusting carefully all parameters of the process.

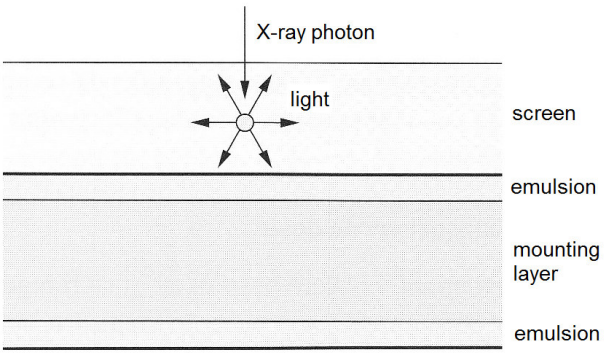


Fig. 2 X-ray film.

Backscatter X-ray
This technology is a new imaging system, which detects the radiation which comes back from the target. It has potential applications in almost every situation in which non-destructive examination is required, but only one side of the object (e.g. a human) is available for examination. One application currently under exploration is to perform a security scan of airline passengers.

X-ray microscopy

Principle

X-ray microscopy is a relatively new technique in biology and medicine. It uses electromagnetic radiation (see Spectroscopy) in the soft X-ray band to produce images of very small objects. As resolution depends on wavelength the resolution of X-ray microscopy lies between that of Optical microscopy and Electron microscopy (EM).
Unlike light in optical microscopes, X-rays do not reflect or refract easily, and they are invisible to the human eye. Therefore, the basic process of an X-ray microscope is to expose a photographic film or use a charge coupled device (CCD, see CDD camera) detector to detect X-rays that pass through the object, rather than light, which bounces off the object. It is a contrast imaging technology using the difference in absorption of soft X-ray in the water window region (wavelength region: 2.3 - 4.4 nm, photon energy region: 0.28 - 0.53 keV) by the carbon atom (main specific element of biomolecules) and the O atom (in water).
Early X-ray microscopes used reflective optics to micro-focus the X-rays, which grazed X-rays off parabolic curved mirrors at a very high angle of incidence. The present method of focusing X-rays is to use in addition to reflective optics refractive (compound refractive lenses) optics and a tiny Fresnel-zone plate of concentric gold or nickel rings on a Si-dioxide substrate. A Fresnel zone-plate, consisting of a set of radial symmetric rings alternating between opaque and transparent, is used to focus light. Unlike lenses, which use refraction, zone plates use diffraction (see Light: diffraction).

Sources of soft X-rays, such as synchrotron radiation sources, have fairly low brightness of the required wavelengths, so an alternative method of image formation is scanning transmission soft X-ray microscopy. A synchrotron is a particular type of cyclic particle accelerometer in which the magnetic field - to force the particles to circulate - and the electric field - to accelerate the particles - is carefully synchronized with the traveling particle beam. In X-ray microscopy, the X-rays are focused to a point and the sample is mechanically scanned through the produced focal spot. At each point, the transmitted X-rays are recorded with a detector such as a photodiode (a diode that functions as a photodetector).

Application

The resolution allows X-ray tomography of whole cells.
In most materials, X-rays cause Fluorescence and these emissions can be analyzed to determine the chemical elements of the object. Another use is to generate diffraction patterns, a process used in X-ray crystallography. By analyzing the internal reflections of a diffraction pattern, the 3D structure of a crystal can be determined down to the placement of individual atoms within its molecules. X-ray microscopes are sometimes used for these analyses when samples are too small to be analyzed in any other way.
Synchrotron X-ray fluorescence microscopy (SXRF) is a micro-analytical technique for the quantitative mapping of elemental distributions. Among currently available imaging modalities, SXRF is the only technique that is compatible with fully hydrated biological samples such as whole cells or tissue sections, while simultaneously offering trace element sensitivity and submicron spatial resolution. Combined with the ability to provide information regarding the oxidation state and coordination environment of metal cations, SXRF is ideally suited to study the intracellular distribution and speciation of trace elements, toxic heavy metals and therapeutic or diagnostic metal complexes.
X-ray microscopy can be combined with (cryo-)electron microscopy, for instance for viruses with their crystalline structure of the proteins. Another combination is with atomic force microscopy (see Scanning probe microscopy).

More Info

It has an advantage over conventional EM with nm-level resolution in that it can view biological samples in their natural state. EM needs thin dried sliced of the much thicker living cell (although it should be mentioned that cryo-electron microscopy allows the observation of biological specimens in their hydrated natural state). Until now, resolutions of 30 nm are possible using the Fresnel zone plate lens that forms the image using the soft X-rays emitted from a synchrotron. Recently, the use of soft X-rays emitted from *laser-produced plasma* (a plasma is an ionized gas with one or more free electrons not bound to an atom or molecule) ousts the technique with synchrotron radiation.

Sound and ultrasound

Acoustic impedance

Principles

Acoustic impedance Z (or sound impedance) is the ratio of sound pressure p to particle velocity v. Also it is the product of the density of air ρ (rho) and the speed of sound c. The acoustic impedance Z is expressed in rayl (from Rayleigh, in $N{\cdot}s{\cdot}m^{-3}=Pa{\cdot}s/m$):

$$Z = p/v = J/v^2 = p^2/J = \rho c \qquad (1)$$

with p = sound pressure, in N/m^2 = Pa = Pascal,
v = particle velocity in m/s,
J = sound intensity in W/m^2,
ρ (rho) = density of the medium (air) air in kg/m^3,
c = speed of sound (the acoustic wave velocity) in m/s.

v is the acoustic analogue of electric current, and p the analogue of voltage. Table 3 gives the densities (ρ), sound velocities (c) and acoustic impedance (Z) is of some bio-materials.

Table 1

material	density ρ (kg/m^3)	speed c (m/s)	impedance Z (rayl)
air (20 °C)	1.20	343	412
water (37 °C)	$1{\cdot}10^3$	1525	$1.52{\cdot}10^6$
brain	$1.02{\cdot}10^3$	1530	$1.56{\cdot}10^6$
muscle	$1.04{\cdot}10^3$	1580	$1.64{\cdot}10^6$
fat	$0.92{\cdot}10^3$	1450	$1.33{\cdot}10^6$
trabecular bone	$0.9{\cdot}10^3$	1540	$1.39{\cdot}10^6$
cortical bone	$1.9{\cdot}10^3$	4040	$7.68{\cdot}10^6$

In dry air at 20℃ (68℉) the speed of sound is approximately 343 m/s, about 1 m each 3 ms. The velocity in brain, muscle and other bio-materials with a high water content, and trabecular bone is slightly higher. Fat has a slightly lower value of 1450 m/s and cortical bone with its much higher density ($1.9{\cdot}10^3$ kg/m3) 4040 m/s.

Application

Ultrasound Sound speed is a basic parameter in ultrasound applications, especially for moving objects (like the valves of the heart) which movement is determined by the Doppler effect (see Doppler principle).

More info

Hearing For an equal sound pressure in two materials, v is reciprocally with Z. For instance $Z_{water} \approx 4000\ Z_{air}$, and so the particle velocity in water is 4000 smaller than that in air. Therefore, also the particle velocity of a sound impinging from a source in the air onto the head is 4000 times smaller than in the air. The resulting vibration of the head gives rise to bone conduction, but with respect to the sound sensation evoked by the pressure impinging onto the eardrum it is irrelevant.
Speed of sound Under normal conditions air is nearly a perfect gas, and so its speed does hardly depend on air pressure and on humidity.

Sound travels slower with an increased altitude, primarily as a result of temperature and humidity changes. The approximate speed (in m/s) is:

$$c_{air} = (331.5 + 0.6 \cdot t_C) \quad (2)$$

where t_C is the temperature in ^{o} C. A more accurate expression is

$$c = \sqrt{\gamma \cdot R \cdot T} \quad (3)$$

where γ is the adiabatic index or c_p/c_V ratio, the ratio of heat capacity of the gas (c_p) with constant p and the specific heat capacity of the gas (c_V) with constant volume, T the absolute temperature (K), and R (287.05 J/(kg·K) for air) the universal gas constant (see Gas laws) In this case, the gas constant R, which normally has units of J/(mol·K), is divided by the molar mass of air. The derivation can be found in various textbooks. For air γ = 1.402.

See further Sound and acoustics, Ultrasound and the Doppler principle.

Contrast enhanced ultrasound, CEU

Principle

CEU is administrating *gas-filled microbubbles* intravenously to the systemic circulation in echography. Hence, the mcrobubbles in the blood subjected to ultrasound show compressions and rarefactions, so they *oscillate* strongly and consequently reflect the waves. They are highly echogenetic, due to the large acoustic impedance difference between gas and liquid. There characteristic echo generates the strong and unique echograms of CEU. CEU can be used to image blood flow rate in organs.

Gas bubbles in blood are thought to be covered by a *surfactant* of blood-macromolecules. However, often bubbles with an artificial skin (coating, e.g. with phospholipids) are injected. This coating also improves stability and prevents shrinkage.

The gas bubbles evoke the most important part of the ultrasound contrast signal and determines the echogenicity. Surfactants lower the sensitivity for cavitation (the collapse of the bubble during rarefaction). The reflected signal as well as the signal emitted during cavitation can be used.
Common practiced gases are N_2, or heavy gases like sulfurhexafluoride (F_6S), perfuorocarbon and octafluoropropane, which are all inert. Heavy gases are less water-soluble so they diffuse less into the medium, guarantying long lasting echogenicity. Regardless of the shell or gas core composition, bubble size ranges 1-4 µm in diameter. F_6S bubbles have a mean of 2,5 µm and 90%<6 µm. With such sizes, they flow easily through the microcirculation.
Selection of surfactant material determines how easily the bubble is taken up and how long the bubbles survive.

Applications

Targeting ligands that bind to receptors characteristic of intravascular diseases can be conjugated to the bubble skin, enabling the bubble complex to accumulate selectively in areas of interest. However, the targeting technique has not yet been approved for clinical use.
Genetic drugs can be incorporated into ultrasound contrast agents. Gene-bearing bubbles can be injected IV and ultrasound energy applied to the target region. As the bubbles enter the region of insonation, the bubbles cavitate, locally releasing DNA. Cavitation likely causes a local shockwave that improves cellular uptake of DNA. By manipulating ultrasound energy, cavitation and hence delivery can be visualized in the vessels with bubbles. Imaging can be performed before IV, just after IV and during cavitation, each with a different energy, to control the process of delivery.

Untargeted CEU is applied in Echocardiography. Bubbles can enhance the contrast at the interface between the tissue and blood. When used in conjunction with Doppler (see Doppler principle) ultrasound, bubbles can measure myocardial flow to diagnose valve problems. The relative intensity of the bubble echoes can also provide a quantitative estimate on blood volume. In vascular medicine, bubbles visualize perfusion and so this technique is crucial for the diagnosis of a stenosis.

Targeted CEU is being developed for a variety of applications. Bubbles targeted with ligands are injected systemically in a small bolus. The ligands bind to certain molecular markers that are expressed by the area of interest. Bubbles theoretically travel through the circulatory system, eventually finding their respective targets and binding specifically. Sound waves can then be directed on the area of interest.
Specific applications are to visualize e.g. inflammatory organs (arteriosclerosis, heart attacks, Crohn's disease).
Bubbles-targeted ligands can bind receptors like VEGF to depress angiogenesis in areas of cancer. Detection of bound targeted bubbles may show the area of expression. This can be indicative of a certain disease state, or may identify particular cells in the area of interest.
Drugs can be incorporated into the bubble's lipid skin. The bubble's large size relative to other drug delivery vehicles like liposomes may allow a greater amount of drug to be delivered per vehicle. By targeted the drug-loaded bubbles with ligands that bind to a specific cell type, bubble will not only deliver the drug specifically, but can also provide verification that the drug is delivered if the area is imaged using ultrasound. Local drug delivery is used for angiogenesis, vascular remodeling and destruction of the tumor.
The force needed for bursting may temporarily permeabilize surrounding tissues and allow the DNA to more easily enter the cells. This can further be facilitated by coating materials of the bubble surfactant skin, e.g. liposomes, positively charged polymers, and viruses (as they do already for millions of years for delivering genetic materials into living cells).

More info

Gas bubbles in a liquid are characterized by a resonance frequency f, which is directly related to their diameter R_0. f can by approximated (liquid surface tension not included; S. De, 1987) by:

$$f = 0.5 \cdot \pi^{-1} \cdot R_0^{-1}\left[(3\gamma / \rho_l)P_0\right]^{0.5} \qquad (1)$$

with γ the specific heat ratio of the gas (=1.40 for N_2; see Gas laws), ρ_l the liquid density (=1050 kg/m^3 for blood) and P_0 the ambient pressure (here 114.7 kPa). With R_0 = 2.5 µm, f is about 1.36 MHz. With R_0 > 100 µm (1) is accurate but with $R_0 \leq 10$ µm f is about 3.5% too small. Implying the surface tension of blood (= 0.058 N/m) would increase f with 7.7%, but bubbles of this size will be surrounded by a surfactant skin, which counteracts the effect of the surface tension (see **More info** of Surface tension). In pure water, bubbles have an extremely sharp Resonance peak (quality Q about 70) but in blood and with the skin surfactant, this is much lower due to the skin-shear and skin-stiffness. With the heavy multi-atomic gases f is smaller since γ is smaller (for F_6S γ =1.10) and ρ_l considerably larger.

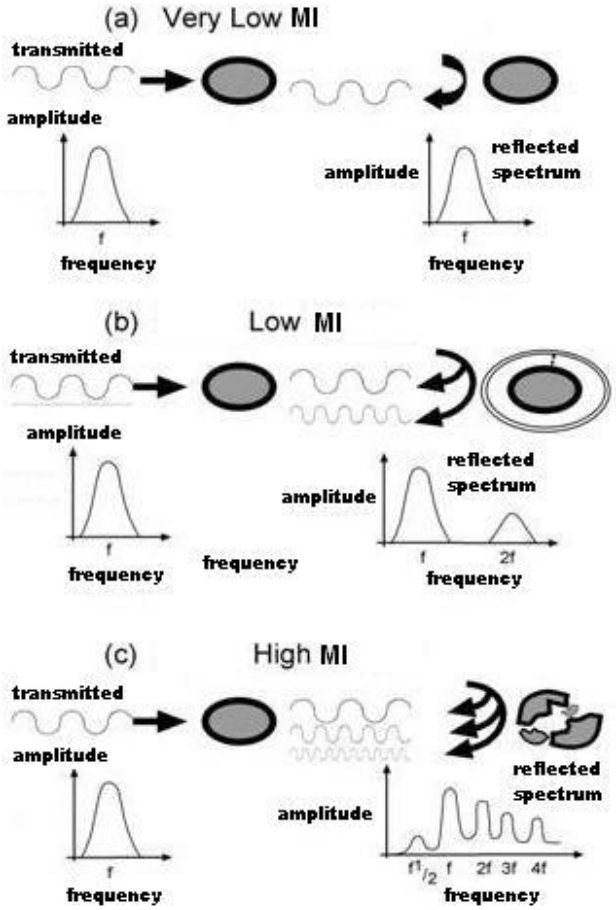


Fig. 1 Different MIs produce different reflected spectra. (From ref. 5.)

Bubbles oscillate (expand and contract) in the ultrasound field. The nature and pattern of their oscillation, and thus the nature of the backscatter signal, differs, depending on the transmitted acoustic power. The power of the ultrasound field is expressed as the mechanical index (MI; see Ultrasound). With *very low* MI (< 0.1), bubbles demonstrate linear oscillation (reflected frequency equals impeding frequency). *Low* MIs (0.1 – 0.6), generates nonlinear oscillation of the bubble whereby expansion is greater than contraction. In addition to the backscatter of the fundamental frequency, the bubbles also produce backscatter of harmonics (see Signal analysis and Fourier). When exposed to *high* MI (> 0.6, i.e. the MI used for standard imaging) the bubbles oscillate strongly and burst. Upon destruction, the bubbles produce a brief, high amplitude signal, with good resolution, which is rich in harmonic signals, containing backscatter at the second third and fourth harmonics etc (Fig. 1).
The most important limitation of this technique is the artifact caused by tissue motion, because tissue motion lookes like bubble destruction (a false negative). If Doppler frequency is increased, pulse

separation is reduced, so tissue movement between pulses can be minimized. However, if the pulses are too close, not all the gas within the bubble will have dissipated before arrival of the next pulse, so the difference between pulses is reduced, possibly leading to false positive perfusion defects. Air-filled bubbles are optimal for this technique because of rapid dissipation of the gas, allowing closely spaced pulses.
There are several high MI techniques, some developed for the moving myocard.

Triggered harmonic imaging
Intermittent high power imaging can improve imaging, with the best opacification with intermittent harmonic imaging. With this technique, high energy ultrasound is transmitted at specified intermittent intervals, triggered to the ECG (e.g. 1 of 4 cardiac cycles). The time between destructive pulses allows the bubbles to replenish the myocardium. With each destructive pulse, high amplitude backscatter rich in harmonics is returned to the transducer, enabling static images of myocardial perfusion.

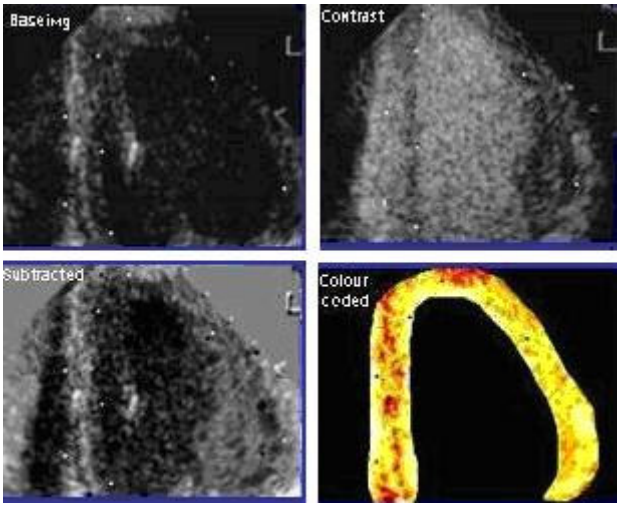


Fig. 2 From top left to bottom right: Base imaging, no bubbles; contrast, bubbles; subtracted, echo's subtracted; gray-scale, color coded. (From ref. 5.)

Pulse Inversion Doppler
Another grey scale high MI technique is pulse-inversion imaging whereby two synchronized beams of pulses impinging onto the myocardium. The second pulse sent has a 180° phase shift. The scanner processes the echo's of the two types of pulses by adding them together. When the bubbles generate a linear echo, the addition of one pulse to the other should cancel out to zero and no signal is generated. However, micro-bubbles produce non-linear echo signals at high MI and the summation of returning pulses will not equal zero.
With this technique, processing can theoretically be limited only to signals generated by bubbles and not by other objects. However, tissue motion artefacts are a major limitation, as movement of tissue also creates non-linear signals.
More specific applications can be found in ref. 5.

Physical advantages of CEU
- Ultrasonic molecular imaging is safer than molecular imaging modalities, e.g. CT with radionuclides.
- Since bubbles can generate such strong signals, a lower intravenous dosage is possible; micrograms compared to milligrams for other molecular imaging modalities such as MRI contrast agents.

Physical disadvantages of CEU
- Ultrasound produces more heat as f increases, so f must be carefully monitored.
- Equipment settings are subjected to safety indexes (see Ultrasound). Increasing ultrasound energy increases image quality, but bubbles can be destructed, resulting in microvasculature ruptures and hemolysis.
- Low targeted bubble adhesion efficiency, so a small fraction of bubbles bind to the area of interest.

References
1 De S. On the oscillations of bubbles in body fluids. J Acoust Soc. Amer, 1987, 81, 56-567.
2 Postma, M., Bouakaz A., Versluis M. and de Jong, N. IEEE T Ultrason Ferr, 2005, 52.
3 http://en.wikipedia.org/wiki/Contrast_enhanced_ultrasound
4 http://e-collection.ethbib.ethz.ch/ecol-pool/diss/fulltext/eth15572.pdf
5 Stuart Moir S and Thomas H Marwick TH, Combination of contrast with stress echocardiography: A practical guide to methods and interpretation. *Cardiovascular Ultrasound* 2004, 2:15. http://www.cardiovascularultrasound.com/content/2/1/15

Doppler echocardiography

Principle

The echocardiogram is the most common application of echography imaging and the Doppler velocity measurement (see Doppler principle). It allows assessment of cardiac valve function, leaking of blood through the valves (valvular regurgitation), left-right shunt (e.g. open oval foramen), and calculation of the cardiac output.
The echo-image is depicted in black and white, and the direction and velocity of the blood are depicted in red for approaching the probe and blue for removing away. The more blue or red the higher the velocity. See Echography and Contrast enhanced ultrasound for further explanation and examples of images in figures.

Application

Contrast echocardiography (CE, e.g. for detection of the right heart Doppler signals, intra-cardiac shunts). It uses intravenously administered micro-bubbles to traverse the myocardial microcirculation in order to outline myocardial viability and perfusion, see Contrast enhanced ultrasound.

Echography

Principle

Echography**,** also called medical (ultra)sonography is an Ultrasound-based diagnostic imaging technique.
Echography uses a probe containing acoustic transducers (generally of piezo crystals, see Piezoelectricity) to send sound into a material (here tissue). Whenever a sound wave encounters a material with a different Acoustical impedance, part of the sound wave is *reflected* which the probe detects as an *echo*, see Fig. 1.

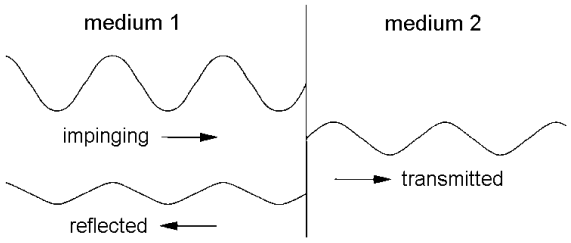


Fig. 1 Sound wave reflection and transmission.

The time it takes for the echo to travel back to the probe is measured and used to calculate the depth of the tissue interface causing the echo, provided that the sound speed in known. The greater the difference between acoustic impedances, the larger the echo is. The reflection coefficient R is:

$$R = \frac{A_r}{A_i} = \left| \frac{Z_1 - Z_2}{Z_1 + Z_2} \right| \qquad (1)$$

with A_r and A_i the amplitudes of reflected and impinging wave and Z_1 and Z_2 the acoustic impedences of medium 1 and 2 respectively. When these media are respectively water (Z = 152.000 rayl) and air (Z=400 rayl) then R=0.999474, which is equivalent to a transmission loss of 20log(1-R) = 65.6 dB. When medium 1 is air and 2 is water, the same holds (see (1)). Consequently: *with a large ratio of the both impedances, the reflection is large*.
The above consideration does not take in account scatter from the object, which diminishes the reflectance and disturbs imaging. Taking water as substitute for blood, R of blood-muscle interface is only 0.034 (see for tissue sound speed values Acoustic impedance), which asks for highly sofisticated hardware and software to obtain a good image (noise reduction). When bone is involved, R is some 2-20 times higher.
A water-based gel ensures good acoustic coupling between skin and the ultrasound scan head.

A 2D-image can be obtained by a probe with many transducers and a 3D images can be constructed with a specialized probe.
From the amount of energy in each echo, the difference in acoustic impedance can be calculated and a color is then assigned accordingly.

The echographic modes
In the *A-mode* the strength, i.e. **a**mplitude, of the reflected wave is indicated on the vertical axis and time at the horizontal one, with time zero at the origin.
In the *B-mode* the strength is indicated by the **b**rightness (grayscale) of a dot. With the B-scan, along the vertical axis the penetration depth is indicated. The beam of the ultrasound changes slightly its angle of incidence every time a new sound pulse is emitted. In this way a kind of section of the anatomical object is obtained. However, the less depth, so the closer to the sound source, the more compressed is the image in the horizontal direction (so parallel to the surface). Such a scan is made many times a second and in this way a kind of movie is made. This is very helpful for moving structures, especially the heart (Fig. 2).

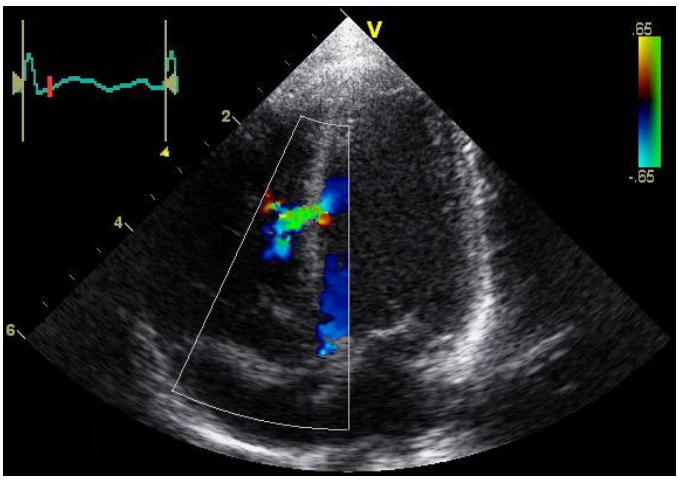


Fig. 2 Abnormal echocardiogram showing a mid-muscular ventricular septum defect in a new-born child. The red line in the ECG mark the time instant that the image was captured. Colors indicate blood velocity measured with the combined Doppler apparatus.

In the *M-mode*, **m**ovement is visualized by displaying time along the horizontal axis and an image is made for a single beam direction. When for instance this beam impinges on a mitralis valve, the image shows opening and closing of the valve (Fig. 3).

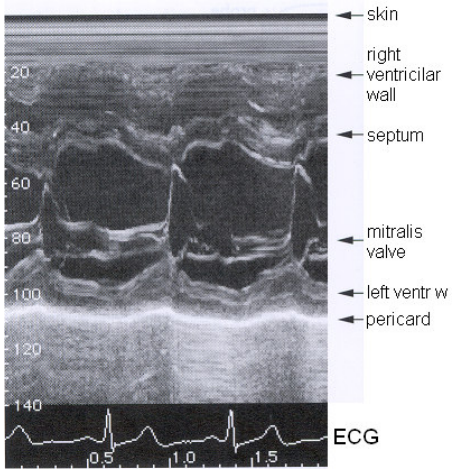


Fig. 3 Echocardiogram in M-mode

Technical strengths
- It shows the structure as well as functional aspects.
- It images soft tissues very well and is particularly useful for delineating the interfaces between solid and fluid-filled spaces.
- It renders "live" images

- Widely available and comparatively flexible.
- Small, cheap, easily carried scanners (bedside) available.

Mechanical weaknesses

- Classical ultrasound devices penetrating bone poorly but is expected that in the future ultrasound bone imaging is possible with dedicated devices.
- Performs very poorly when there is a gas between the scan head and the organ of interest, due to the extreme differences in acoustical impedance.
- Even in the absence of bone or air, the depth penetration of ultrasound is limited and frequency dependent (see Ultrasound). This makes it difficult to image structures that are far removed from the body surface, especially in obese patients.
- The method is operator-dependent. A high level of skill and experience is needed.

Applications

Echography is widely utilized, often with a hand-held probe. It is especially practiced in cardiology, gynecology and obstetrics, urology (kidney, prostate), vascular medicine (e.g. (deep) thrombosis, lower-limb veins for incompetence, phlebology), gastroenterology (also liver), endocrinology, and ophthalmology.

More info

Limitations

The spatial resolution in the axial direction (the depth) is directly related to the wavelength λ of a pure ultrasound frequency. In the lateral direction the resolution is determined by the width (the aperture angle) of the beam due to divergence.

Further, there is an ambiguity in depth position. This occurs when the time lapse between sending and receiving a wave is larger than the period time t_{per}, The reflected wave of the objects at all distances n times 0.5λ are superimposed at the reflected pulse of the object itself. Mathematically:

$$\text{depth} = (0.5t_{recieve}/t_{per}, - \text{integer}\{0.5t_{recieve}/t_{per}, -\})c + n\lambda. \qquad (2)$$

This problem is solved by sending short pulses and adjusting the pulse interval time such that any reflection from boundaries to at least the depth of interest are arriving within the pulse interval time. Since a pulse comprises many frequencies (see Signal Analysis and Fourier) the received signal needs some complicated computation (deconvolution) to reconstruct the echo-image.

Doppler echography

There exist several types of echography, nowadays combined with using the Doppler effect (see Doppler principle): the Doppler echography. This technique assess whether structures (usually blood) are moving towards or away from the probe, and its relative velocity. By calculating the frequency shift of a particular sample volume, for example a jet of blood flow over a heart valve, its velocity and direction can be determined and visualized. This is particularly useful in cardiovascular studies and vascular examinations of other organs (e.g. liver portal system, lower limb viens). The Doppler signal is often presented audibly using stereo speakers: this produces a very distinctive, although synthetic, sound. Doppler echography can by distinguished in several modifications. The most common ones are here discussed (ref. 2).

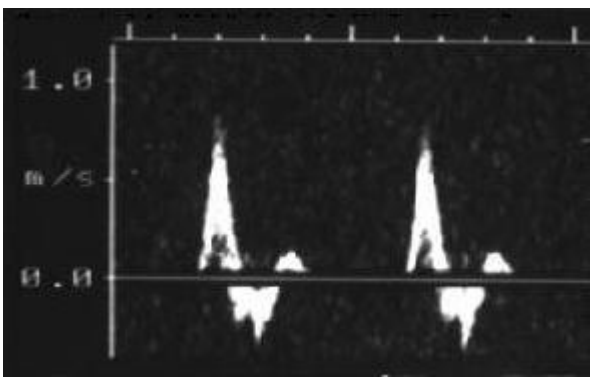


Fig. 4 Duplex scan of superficial femoral artery with a triphasic blood velocity waveform Horizontal axis: time (large ticks 1 s); vertical axis velocity. (From ref. 2.)

The duplex (Doppler) scanner

The duplex scanner detects in a selected part of the image the moving blood by using the Doppler effect. The scanner calculates the actual velocity of the blood provided the angle between the direction of the ultrasonic beam and the direction of movement is known. The operator therefore aligns a marker along the direction of flow in the blood vessel and positions a cursor at the height of the peak systolic blood velocity.

In the common and superficial femoral arteries, the waveform normally has a forward (i.e. upward) component followed by a reverse component and a second smaller forward component. This is called a triphasic waveform because of the three phases. More distally in the superficial femoral artery, the second forward component may be absent, giving a biphasic waveform.

The frequency shift is normally in the audio range (the difference of the ultrasound frequencies), so most duplex scanners send the signal to a pair of audio speakers, and this enables the operator to hear the signal in addition to seeing the display (as in Doppler echocardiography).

Color Doppler scanners

Color Doppler (CD) scanners detect and display moving structures by superimposing color onto the grey-scale image. Color is superimposed wherever the scanner detects a moving structure, usually blood (Fig. 2). A colour scale codes the direction and magnitude of the blood velocity. In this image, red and yellow indicate flow away from the probe, with dark red representing low velocities and orange and yellow indicating higher velocities. Flow towards the probe is indicated in blue and green, with green indicating higher velocities. The color can therefore be used to identify sites where the artery becomes narrower and the blood has to move faster to achieve the same volume flow rate. When the blood velocity exceeds the limit of the color scale, aliasing occurs (high velocity in one direction interprets as lower velocity in the other, wrong, direction). Color Doppler can also be used to display venous blood flow.

CD and Duplex sonography are often combined to Duplex/CD sonography, especially for assessing stenoses.

Power Doppler (PD)

Duplex/CD sonography is not an effective technique when the artery under study is almost perpendicular to the ultrasonic beam or by other poor conditions as bowel gas, breathing movements and obesity. Power Doppler (PD) has improved diagnostic capabilities of vascular Duplex/CD sonography, mainly because it is independent from the angle of insonation and has more sensitivity. PD generates an intravascular color map reflecting the integrated power in the Doppler signal, which essentially depends on the density of red blood cells within the sample volume. However, due to its intrinsic limitations, PD cannot replace conventional sonographic techniques and especially CD. So, PD is used as an adjunctive tool in vascular sonography.

Tissue Doppler Imaging (TDI)

Tissue Doppler Imaging (TDI) measures and displays peak velocities with high temporal resolution (ca. 8 ms) at any point of the ventricular wall during the cardiac cycle. The mean velocities can be calculated with time velocity maps and displayed as color encoded velocity maps in either an M-mode or a 2D format.

References

1. http://en.wikipedia.org/wiki/Ultrasonography

2. Lunt MJ, Review of duplex and colour Doppler imaging of lower-limb arteries and veins. *Journal of Tissue Viability* 1999, **Vol 9**, No 2, pages 45-55.Sept 2000. http://www.worldwidewounds.com/2000/sept/Michael-Lunt/Doppler-Imaging.html#Colour%20Doppler%20scanners

Optoacoustic imaging

Principle

Central to optoacoustic (or photoacoustic) imaging is the optoacoustic effect whereby pulsed Laser energy is absorbed by a medium causing a local temperature increase followed by the generation of pressure transients (acoustic waves).
In optoacoustic imaging, short laser pulses irradiate sound scattering (see Waves) biological tissue and adiabatically (see Compression and expansion) heat absorbing structures, such as blood vessels, to generate ultrasonic pressure transients by means of the thermo elastic effect. These acoustic transients propagate to the tissue surface and are recorded with transducers (Ultrasound or electromagnetic) to reconstruct high contrast images of the initial absorbed energy distribution exactly resembling the absorbing structures. They can be recorded using high frequency pressure sensors (piezoelectric or optical). The speed of sound in tissue (~1500 m/s) allows for the time resolved detection of these pressure waves and the determination of depth from where these pressure waves originated. By using an array of sensors the temporal delay of these incoming pressure wave fronts can be combined into an ultrasound image. The optoacoustic technique combines the accuracy of spectroscopy with the depth resolution of ultrasound.

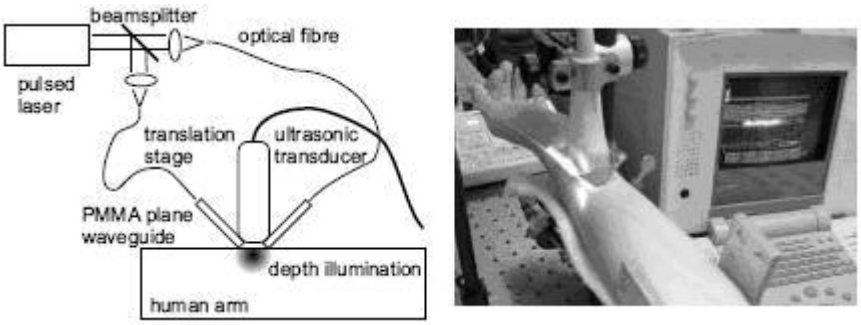


Fig. 1 Left principle of set up. Right the realized apparatus. From ref. 1

Application

Angiography of hypodermal vessels can be performed with optoacoustics. The in vivo optoacoustic images acquired from human finger, arm and legs show high contrast detailed blood vessel structures, which are hard to see on the corresponding ultrasound echography images. This is performed with an photoacoustic microscope. Its resolution is 4 µm.
The optoacoustic imaging and ultrasound techniques extract complementary information, which strongly suggests a combination (overlay) of the two techniques in a single device.
There are acoustic imaging devices in development for breast cancer detection, also utilizing the change in the optical properties of blood in respect to oxygen saturation and the strong optical contrast between hemoglobin and surrounding tissue.
In the last years, optoacoustic imaging was developing vast. Its field of application has been extended to intravascular imaging, oncology (breast cancer), tomographic imaging and experimental interferometric surface displacement,

More Info

Two real-time optoacoustic imaging systems have been developed recently:
(1) Laser excitation combined with classical medical ultrasound system for comparison of the two complementary techniques. In medical applications real-time imaging avoids motion artifacts (heartbeat and breath), facilitates imaging procedure and allows instant diagnosis.
(2) Optical systems detecting the Schlieren effect and equipped with an acoustic lens system for 3D imaging (Fig. 1). Schlieren are optical inhomogeneities in transparent material, often liquid, mostly not visible to the human eye. They are applied to visualize flow, based on the deflection of light by a refractive index (see Light) gradient (resulting e.g. from temperature of salt gradients). The index gradient is directly related to flow density gradient.) This method is still experimentally.
Both methods can be combined with ultrasound Echography.

In the 2nd optical system, the Schlieren transducer images the pressure transient in a fluid filled cuvette below the tissue with an ns-flash lamp and reconstructs images on a computer for visualization.

The first optical system uses an acoustic lens to directly reconstruct a 3D image of the original pressure distribution. This copied pressure image is optically dark field imaged at two angles to provide a stereo image pair of the original absorbing structures.
Both optical systems operate at real-time frame rates of 10-20 Hz and provide high resolutions up to 30-100 µm.
Medical ultrasound is limited by low acoustic contrast, which particularly deteriorates or inhibits imaging of smaller structures in near skin regions. The two systems combining laser excitation and commercial ultrasound are provide high optical contrast and sub-millimeter acoustical spatial resolution for in vivo biomedical imaging. Variation of the laser wavelength allows spectroscopy and functional imaging of blood oxygenation level based on oxygen dependent Hb absorption spectra.

References

1. Niederhauser, JJ. Real-time biomedical optoacoustic imaging. ETH PH.D. thesis, 2004. http://e-collection.ethbib.ethz.ch/ecol-pool/diss/fulltext/eth15572.pdf
2. Ermilov SA et al. Laser optoacoustic imaging system for detection of breast cancer. J Biomed Opt. 2009.

Phonocardiography

Principle

For auscultation of the heart, i.e. listening to the sounds generated by the beating heart, the classical acoustic stethoscope and more and more frequently the electronic stethoscope is used (see Stethoscope). Performed with the latter is phonocardiography. It is a method not solely to record the sounds produced by the heart valves, but also to visualize them on a PC and to analyze them with a computer algorithm in order to discover dysfunction of a valve (see ref. 1 and 2).

Heart sounds
In healthy adults, there are two normal heart sounds often described as a *lub* and a *dub* (or *dup*), that occur in sequence with each heart beat. These are the first heart sound (S1) and second heart sound (S2).
S1 is caused by the sudden block of reverse blood flow due to closure of the atrioventricular valves, mitral and tricuspid, at the beginning of ventricular contraction, or systole.
S2 is caused by the sudden block of reversing blood flow due to closure of the aortic valve and pulmonary valve at the end of ventricular systole, i.e beginning of ventricular diastole.
S1 and S2 can by split in various components. The extra sounds S3 and S4 are rarely audible. In addition, murmurs, clicks and rubs can be audible. They are especially of importance for diagnostics.

Analysis
The analysis is generally a spectral analysis (Fast Fourier Transform, see Fourier Transform) of the whole signal or of one of the sounds. Very suitable is a frequency (vertical) versus time plot with the amplitude (or power) depicted in a color scale, just as done in sonograms (see Fig.2 of Sound Spectra).

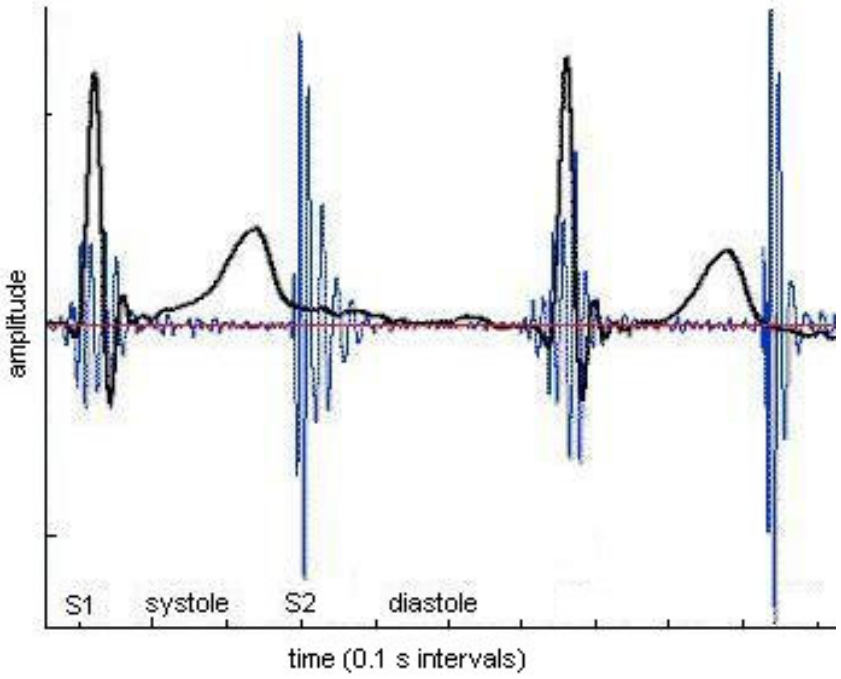


Fig. 1 Phonocardiogram (blue) with superimposed ECG (black) of nearly two heart cycles. Note the correspondence between S1 and QR, and S2 and end of T wave.

Application

Routinely by physicians of many disciplines and cardiologists.
With phonocardiography, pressure and sound manifestations of the pulse wave spreading in the cardiovascular system can be recorded. In addition to standard diagnostic application of phonocardiography can be used to measure pulse wave velocity. Then two signals are needed at minimum for an estimation of the pulse wave velocity (by cross-correlation). These signals have to be measured simultaneously from different points on a human body.

Nowadays, phonocardiography is evolving in acoustic cardiography (ref. 3) that is not only used for valve function.

There is also an enginering application in which stethoscopes are used to isolate sounds of a particular moving part of an engine for diagnosis.

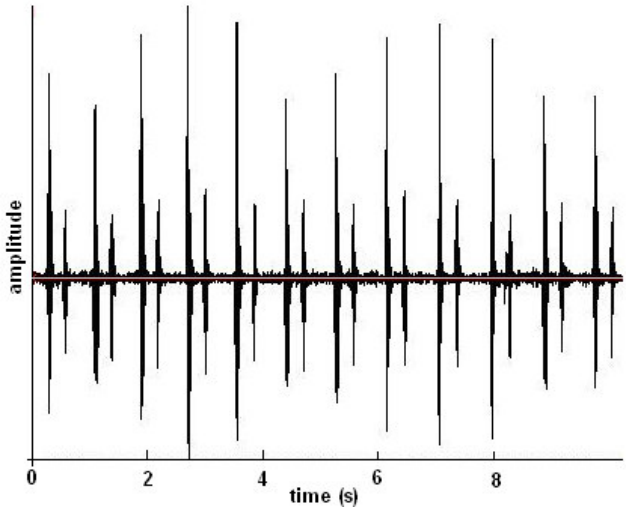


Fig. 2 Amplitude variation due to respiration typical for the healthy heart.

References

1. Crit Rev Biomed Eng. 1995; 23, 63-219 and also
2. Jelinek, M et al. Using a phonocardiography in a pulse wave velocity measurement. Proceedings of the 3rd IEEE International Symposium on phonocardiography, 14-17 Dec. 2003, 491-493.
3. Peacock WF et al. The utility of heart sounds and systolic intervals across the care continuum. Congest Heart Fail. 2006;12 Suppl 1:2-7.

Sound and Acoustics

Basic Principles

In any elastic medium, gas, liquid or solid, an *acoustic wave* can be described as mechanical *compressions and rarefactions*. To be propagated, in liquids and solids, the elasticity guarantees some compressibility, which is necessary to evoke and propagate sound pressure. The wave comprises a *pressure variation* within the medium and as well as a *to-and-fro oscillation* of the component particles. The oscillations of the particles can be expressed as displacement in time (and space). The simplest case is a standing (not traveling) sinusoidal wave. Such a wave occurs when a tout string of a violin is plucked to make a pure tone. However, this is a transversal wave. *Longitudinal standing waves* are produced by organ pipes. With a standing wave, the sound pressure within a single sound cycle is maximal when the displacement of particles of the medium is maximal. At that instant, the particles move towards each other and 0.5λ further away from each other, causing a compression and rarefaction respectively.
In formula, the standing wave is described as:

$$y(t) = y_0 \sin(2\pi f t) \text{ with } y_0 = \sin(2\pi x'/\lambda), \quad (1a)$$

with f the frequency, t time, λ the wavelength, x' the distance from the place of maximal excursion (the crest). However, most waves are traveling or propagating and then the equation comprises a single sine wave which is a function of t and the distance from the sound source x:

$$y(x,t) = y_0 \sin(\omega(t-x/c)) \quad (1b)$$

with y(x,t) the displacement of particles from the equilibrium position, y_0 the maximal amplitude of the particle displacement, ω the angular frequency (= 2πf), c the sound propagation velocity. x/c is the time that the wave needs to travel the path x.
When the particle oscillations are in the direction of propagation, the wave is longitudinal. When the direction is perpendicular on the propagation direction it is a transversal wave, like a taut string of a violin. Equation (1) holds for both types. However, generally sound is a longitudinal wave phenomenon (in the so called far-field, see **More Info**). Fig. 1 illustrates the propagation of a longitudinal sound wave. The oscillatory displacements of the particles is denoted with the vector **ξ** (vectors are denoted in bold), its time derivative, the particle velocity with **v**. The particle motion around an equilibrium position causes the pressure increase and decrease. This sound pressure p is superimposed at the ambient (the atmospheric) pressure. Sound pressures are generally much smaller than atmospheric pressure. Even at the human pain limit (depending on the definition) some 5000 smaller (Table 1 Hearing and Audiometry). With a longitudinal displacement wave the pressure wave is also longitudinal (see Fig. 1).

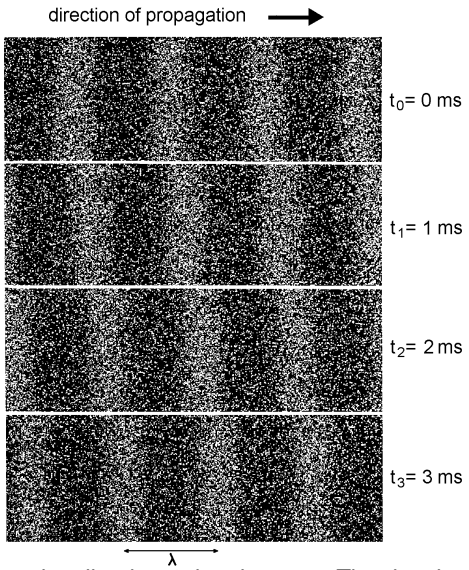


Fig 1 Longitudinal sound wave visualized at 4 time instants. The density of the white specks represents the particle density in the medium and so the sound pressure. The propagation direction is arbitrarily chosen to the right. The distance between to compression peaks is the wavelength λ. Shifting a peak over the distance of one wavelength λ takes about 6 ms. So, the frequency is ca. 167 Hz.

With a sound wave travelling through a medium, the particle displacement is the physical movement of a particle (e.g. a molecule of air). Far from the source the direction of the oscillations is the same as that of the sound wave. Close to the source the direction of the oscillations is dependent on the type of sound source (see **More Info**, near and far field). For a free propagating wave with a plane wave front, the sound pressure (the scalar p) and the particle velocity **v** are mutually related such that:

$$p = \rho c \mathbf{v}, \tag{1c}$$

with ρ being the (characteristic) density of the medium. The acoustic impedance Z is:

$$Z = \rho c, \text{ in } N \cdot s \cdot m^{-3} \text{ (rayl, from Rayleigh).} \tag{1d}$$

v is the acoustic analogue of electric current, and p the analogue of voltage. So, Z is also p/**v** and consequently, p**v** or p^2/Z gives the intensity J (in W/cm^2) of the sound. When the wave acts on surface A we obtain the power J·A (W).
Notice that in a gas the particle motion is random and many 100 m/s (see Gas laws), whereas the motion due to sound is coherent for all particles and in the nm to mm range. This motion is superimposed on the random motion.

Sound pressure
The amplitude or strength is expressed as the effective amplitude (= 0.5√2 of the maximal amplitude of a sinusoidal pressure wave) of the sound pressure. SPL or sound level L_p is a logarithmic measure of the energy of a particular sound relative to a reference sound source.

$$L_p = 20\log(p_1 p_2) = 10\log((p_1^2/p_2^2)) \text{ dB SPL,} \tag{2}$$

where p_0 is the reference sound pressure and p_1 is the sound pressure being measured.

Application

The use of sound in medical practice is enormous. In much classical application the physician is the listener to evaluate the sound. The sound can be evoked by him (palpation) or is generated by the subject (patient) and listened by an instrument (stethoscope etc.). Many modern applications generate sound as a probe and record the sound evoked by the probe. Examples are found in echography, audiology, cardiology etc.

More Info

Sound intensity
The sound intensity, I, (acoustic intensity, W/m^2) is defined as the sound power P_{ac} per unit area A. It is also equal to P_{eff}^2/Z and to p**v**. The usual context is the measurement of sound intensity (W/m) in the air at a listener's location.

$$I = \frac{1}{T} \cdot \int_0^T p(t) \cdot v(t) \cdot dt \tag{3}$$

For a spherical sound source, the intensity as a function of distance r is:

$$I_r = P_{ac}/A = P_{ac}/(4\pi r^2) \tag{4}$$

The sound intensity I in W/m^2 of a plane progressive wave is:

$$L_J = 10\log(J_1/J_0) \text{ dB SPL} \tag{5}$$

where J_1 and J_0 are the intensities. If J_0 is the standard reference sound intensity, where $J_0 = 10^{-12}$ W/m^2. Then instead of "dB" we use "dB SIL". (SIL = sound intensity level).
Our ears as sensors cannot convert sound intensities and powers. In most conditions for the sense of hearing, the auditory system only utilizes the sound pressure as input signal (see Hearing and Audiometry).
With all these parameters defined, now particle displacement can be represented in terms of many combinations of the other parameters.

$$\xi = \frac{v}{2 \cdot \pi \cdot f} = \frac{p}{Z \cdot \omega} = \frac{1}{\omega}\sqrt{J/Z} \tag{6}$$

Sound velocity
Sound's propagation speed c (from Latin celeritas, velocity) depends on the type of the medium (a gas, fluid or solid) through which it propagates, and the ambient temperature and pressure. See Sound impedance for the sound velocity of some (biomaterials).

Near-field and far-field
The near- and far-field are concepts to indicate whether an object (or listener) is close or remote from a sound source in relation to the emitted frequency (see ref. 1). The border between near-field and far-field is defined as $\lambda/2\pi$. The behavior of pressure and displacement is different in the near- and far-field, the actual reason why this distinction is made.
For hearing in the air with air borne sound, only the far-field is of importance and sound is perceived by the sound pressure received by the ear drum. Within the near field particle displacement also plays a role. Now, sound is also transmitted by particle displacement and perceived via conduction by the tissues of the head (e.g. by bone conduction). Listening with headphones is based on a transmission mixture of pressure and displacement, with the ear drum and the head-tissues as receiver.

Sound sources
The easiest sound source to deal with mathematically is the monopole source, a pulsating sphere that produces spherical waves. At distances from the source that are large with respect to λ, the amplitudes of **v** and p decrease linearly with the distance from the sound source (R), as does the time integral ξ and derivative of **v** (i.e., the particle displacement, and the acceleration, respectively). It is clear that intensity of the sound decreases with R^2. For $R << \lambda$, p decreases linearly but ξ decreases with the R^2. The phase difference between p and v depends upon R. When $R >> \lambda$, p and **v** are in phase with one another. However, for $R < \lambda$, p leads **v** for at least 45°, up to a maximum of 90°. For distances approximately $\lambda/2\pi$, there is a gradual transition from the near- to the far-field. In addition to the near- and far-field effects of λ, there is also an effect of the frequency f (where $f=c/\lambda$). Under the condition that **v** at the interface of the pulsating source is the same for all frequencies, p is proportional to f, irrespective of R. In the near-field, displacement is proportional to 1/f, but in the far-field displacement is independent of f.
Many (biological) sound sources are dipoles (vibrating spheres) rather than monopoles.
Dipole sources produce complicated sound fields that are composed of a radial and a tangential component. The tangential component, in contrast to the radial, is very small in the far-field. All frequency effects for the radial component are a factor of f (in Hz) times stronger in a dipole field than in monopole field.
In the far-field of a dipole source, distance effects are the same as for a monopole source, aside from the effect of the angle between the direction of oscillation of the dipole and the direction of observation (multiplication by a cosine function). The near-field of a dipole is very complicated. A more complete discussion of the fields of monopole and dipole sources can be found in Kalmijn (1988).

References
1. Kalmijn, A. J. (1988). Hydrodynamic and Acoustic Field Detection. In Sensory Biology of Aquatic Animals (ed. J. Atema, R. R. Fay, A. N. Popper and W. N. Tavolga), pp. 83-130. New York, Berlin, Heidelberg, London, Paris, Tokyo: Springer.

Sound Spectra

Basic Principles

Types of sounds
Sound is defined as *mechanical oscillations* perceivable for the human ear, generally from 16 to 16.000 Hz.
Sounds that are sine waves with fixed frequency and amplitude are perceived as pure tones. While sound waves are usually visualized as sine waves, sound waves can have arbitrary shapes and frequency content. In fact, most sounds and so the waves are composed of many sine waves of different frequencies, according to the principle of *Fourier* (see Fourier analysis). Waveforms commonly used to approximate harmonic sounds in nature include saw-tooth waves, square waves and triangle waves. The frequency of these periodic sounds is called the *fundamental frequency* f_1 and the higher frequencies in the sound are the overtones or *harmonics*. A harmonic with frequency f_k is an integer multiple (k) of the frequency f_1, so $f_k = k f_1$. Table 1 presents a number of sounds, indicated by its range of composing frequencies or by f_1. Tones produced by a music instrument are generally periodic (not with drums), but *speech* is *a-periodic*.

Table 1

type of sound	frequency (Hz)
central c of piano (C^{I}) :	262
'high c' (c^{3})	1046
range of concertgrand	27-4186
range of singing-voice (bass)	82-330
range of singing-voice (soprano)	262-1046
speech fundamental (man)	163
speech fundamental (woman)	262

Instruments can play a nearly (pure) tone, but mostly notes have many harmonics, accounting for the *timbre*. *Noises* (strictly speaking) are irregular and disordered vibrations including all possible frequencies. They are a-periodic, i.e. the wave shape does not repeat in time.

Amplitude and phase spectra

With Fourier analysis the spectrum of the signal can be calculated. Spectral analysis yields the *amplitude spectrum* (amplitude versus frequency) and the *phase spectrum* (phase versus frequency). They present the frequencies of which the signal is composed. With the amplitudes and phases of all harmonics the signal can be composed uniquely (Fourier synthesis). In the analysis of sounds, generally only the amplitude spectrum (or power spectrum) is calculated.
The emitted spectra are limited by the technology of the emitting apparatus. For instance, low frequencies (< 100 Hz) are hard to produce by loudspeakers, such that they are not contaminated by distorting higher harmonics. Another limitation is that the emitted spectrum is filtered by the medium in between generator and receiver. In air, high frequencies are strongly diminished (a train listened at a long distance produces a dull rumble). Finally, the receiver should by capable to sense all emitted frequency and, moreover, with the same sensitivity.
Fig. 1 presents two sounds as a function of time, together with their spectra, evoked by singing a Dutch aa (top) and a Dutch oo (bottom).

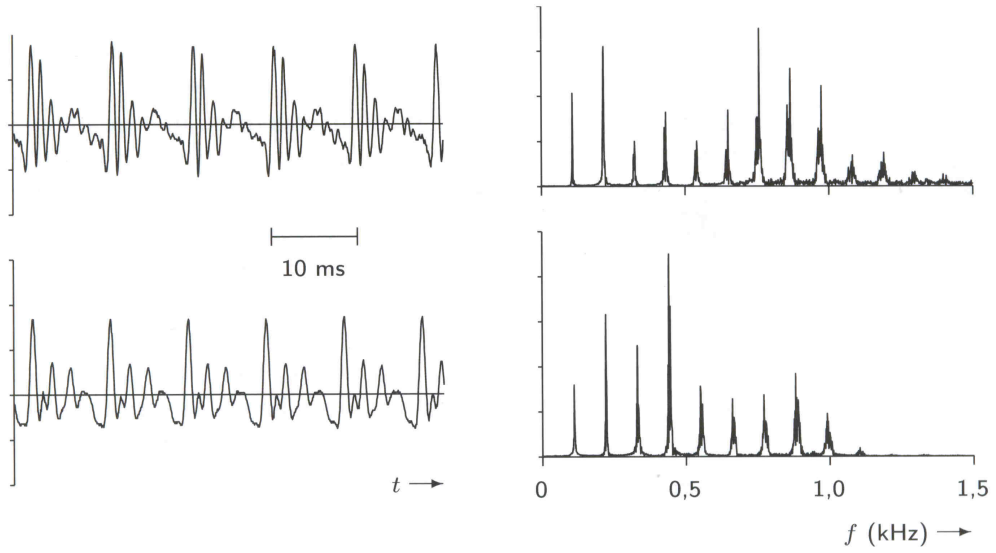


Fig. 1 Wave shapes (left) and amplitude spectra (right) of two vowels.

The periodicity of the signal is characteristic for *vowels*. The fundamentals of vowels are the same but the higher harmonics quite different. The fundamental can be adjusted by changing the tension of the *vocal cords*. This can be done within about two octaves (factor 2x2). The length of the vocal cords determines the lowest fundamental. The harmonics are determined by the mouth opening and by the shape of the mouth cavity, which acts as a resonator. Their spectrum is called the *formant*. It is specific for the vowel. *Consonants* are generated in a similar way but they contain more noise, produced by pressing air though a constriction. Their wave shapes are not periodic. This hampers a straightforward Fourier analysis. To get round this problem, the analysis is made in a short span of time and this time window glides over the signal, producing a spectrum at each time sample point. In this way a *phonogram* can be constructed with the frequency versus time. The amplitude of each frequency is depicted by a gray-scale. Fig. 2 presents the phonogram of two subjects, which pronounced the same

Dutch sentence. Comparison of both panels shows that the speech-phonogram is highly subjects specific.

Applications

Speech analysis is widely used in clinical speech-training examinations, in the technology of hearing aids, cochlear implants etc. Such technologies will develop fast, since *hearing-loss* and disorders will become *eudimic* due to hair cell damage (too frequently too long exposure to too high sound levels).

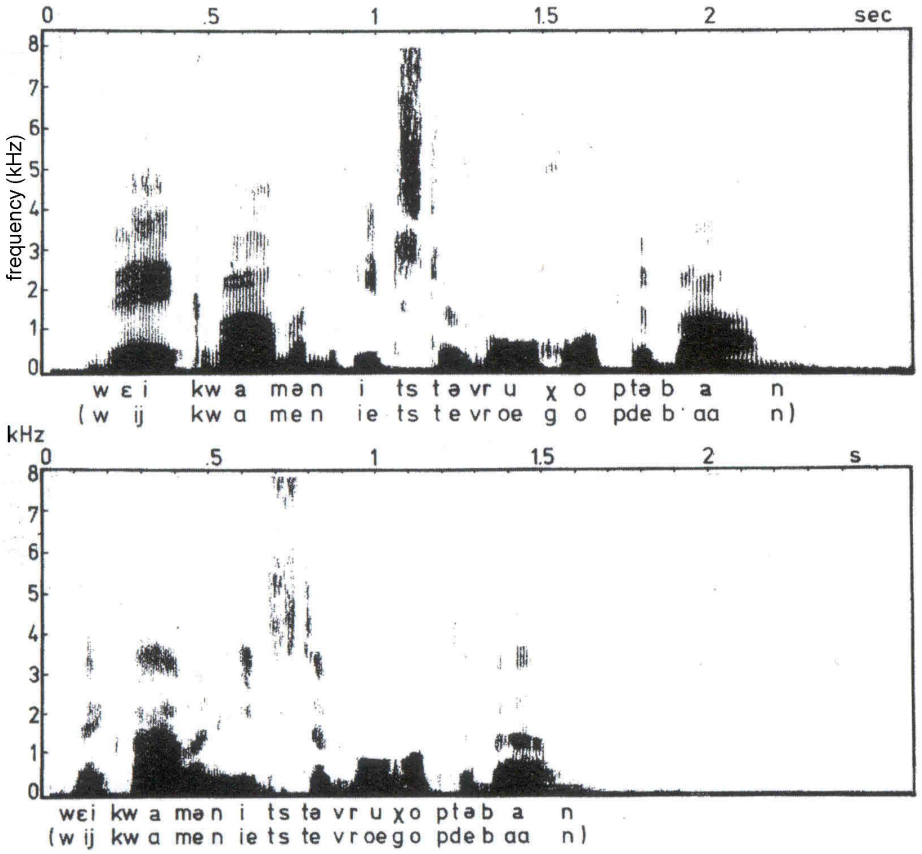


Fig. 2 Phonogram of two different subjects pronouncing the same sentence.

Other applications are *artificial speech* understanding/interpretation and producing artificial speech. Sound analysis is applied in many medical equipment and apparatus, e.g. in Echography .

Stethoscope

Principle

The classical acoustic stethoscope operates on the transmission of sound from the two-sided chest piece (a diaphragm and a bell) via air-filled hollow tubes to the listener's ears. If the diaphragm is placed on the skin, body sounds vibrate the diaphragm, creating acoustic pressure waves (see Sound and Acoustics) which travel up the tubing to the listener's ears. If the bell is placed on the patient, the vibrations of the skin directly produce acoustic pressure waves. The bell transmits low frequency sounds, while the diaphragm transmits higher frequency sounds. The problem with acoustic stethoscope is that the sound level is extremely low.
For children small size stethoscopes (also electronic) have been developed and another familiar stethoscope is the trumpet-shaped fetal stethoscope (fetoscope).

Electronic stethoscopes (or stetophones) overcome the low sound levels by amplification. Unlike acoustic stethoscopes, which are all based on the same physics, transducers in electronic stethoscopes vary widely. Microphones and accelerometers behave poorly due to ambient noise interference. Piezoelectric crystals (see Piezoelectricity) and a capacitive sensor attached to the diaphragm are suitable. With the latter, the diaphragm responds to sound waves identically to a conventional acoustic stethoscope, with changes in an electric field replacing changes in air pressure. This preserves the sound of an acoustic stethoscope with the benefits of amplification.
A more recent development is the electronic phonendoscope, an endoscopic stethoscope.
Ambient noise filtering is nowadays available in electronic stethoscopes.
The analysis is generally a spectral analysis (Fast Fourier Transform) of the whole signal or of one of the sounds.

Application

Application comprises routine physical examination by physicians of many disciplines and specifically by cardiologists.
For the examination of the heart sounds, the electronic stethoscope is very suitable with digital analysis advancing. Most helpful for evaluation of the heart sounds (see Phonocardiography) is a frequency versus time plot with the amplitude (or power) depicted in a gray or color scale, just as done in sonograms (see Fig.2 of Sound Spectra).

References

1. Wikipedia Stetoscope; http://en.wikipedia.org/wiki/Stetoscope

Ultrasound

Principle

Ultrasound comprises frequencies > 20 kHz, the upper limit of human hearing. For frequencies between 1 and 20 kHz this is not at a one-to-one basis; one spike is fired during many stimulus periods, but there is statistical still more or less a phase relation to the stimulus phase (phase locking). Above this limit the phase of spike firing in the axons of the acoustic nerve are not longer related to the phase of the impinging sound frequency. In other words, spike firing, if any, is random in time. Many mammals such as dogs can hear the most lower range of ultrasound. Echolocating bats and sea mammals (dolphins) have audiograms (see Hearing and Audiometry) up to 150 kHz and occasionally up to 200 KHz.
Ultrasound, that goes up to ca. a GHz (Acoustic microscopes), has many industrial and medical applications. Typical ultrasound frequencies f are 1 to 10 MHz (see Table 1). A higher frequency results in a higher resolution of the image, just as holds for the frequency (color) of the light in Optical microscopy. In ophthalmology with its fine structures, 5-20 MHz is applied, and in obstetrics, cardiology and urology 2-5 MHz. In the same way as with Doppler applications, ultrasound can by emitted continuously or pulsed.

Producing an ultrasound wave In medical echography (or ultrasonography) a sound wave is produced by creating continuous or short pulses of sound from an array of piezoelectric transducers (see Piezoelectricity). Wiring and transducers are encased in a probe. The electrical pulses vibrate the piezo crystals to create a series of sound pulses from each, which in turn produce together a single focused arc-shaped sound wave from the sum of all the individual emitted pulses. To make sure the sound is transmitted efficiently into the body (a form of impedance matching, see Acoustic impedance) between probe and skin a thin layer of a special gel is administrated. The sound wave, which is able to penetrate bodily fluids, but not (hardly) solids, bounces off the solid object and returns to the transducer. This rebound is an echo. This also happens with large differences between impedances of soft tissues.
A completely different method is to produce ultrasound with a Laser (ref. 1), see Optoacoustic imaging.
Receiving the echo's The piezo-transducer is in addition to sending also capable to receive the echo. Just in reverse, the echo-produced vibrations give rise to electrical pulses in the probe. These signals are analyzed in a PC and transformed into a digital image.

	f	**half-distance**
	(MHz)	mm
material		
water	1.0	14000
muscle	1.0	27
fat	0.8	69
brain	1.0	32
bone	0.6	9.5
bone	0.8	3.4
bone	1.2	2.1
bone	1.6	1.1
bone	1.8	0.8
bone	2.25	0.6
bone	3.5	0.45

Table 1 Half-distances of sound intensity for various tissues at the indicated frequencies.

Dangers of ultrasound can basically not be excluded since ultrasound is energy and energy is always partly transformed to heat. Heat may damage tissue directly or may give rise to bubble formation from bubble nuclei due to supersaturation of dissolved nitrogen and with heating at 100 °C (boiling). Bubbles may result in micro-traumata (as with decompression illness). Damage can be limited by reducing f. Integrated over a closed surface (e.g. a sphere for a monopole and a spheroid for a dipole source, see Sound and acoustics), the sound energy remains at any distance from the source the same as long as there is no absorption. Since the tissue is in the far-field (see Sound and acoustics) of the source, which can be approximated by an acoustic dipole, the sound energy increases with f^3 (for a monopole the increase is proportional with f). Therefore, f should be limited. This prevents tissue damage by heating. There is another reason to limit f: the absorption (this is dissipation: sound energy transformed to heat) increases with f (see **More Info**). Consequently, the penetration depth strongly decreases with f. From Table 1 it can be seen that this is progressively with f.
The maximal beam-intensity I_{max}, the mean beam intensity I_{mean} and the total administrated dose D all have their maximal allowable values. With constant I, $D = I_{max}t$, with t the application time. With pulsed ultrasound, pulse period t_{per} and pulse duration t_{pulse} the dose is $(t_{per}/t_{pulse})I_{max}t$. Fig. 1 shows the I_{max} versus t safety relation. The protocols of the application (Echography) are such that damage is actually impossible and (systematic) damage has not been reported in literature.

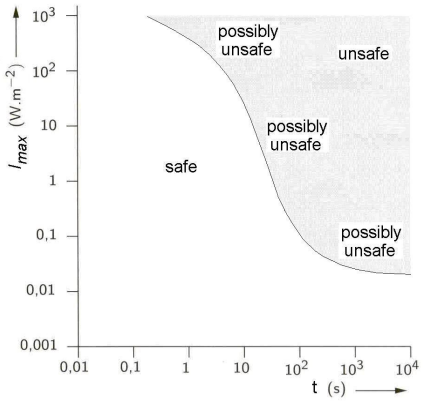


Fig. 1. Principle of safety norms of ultrasound with power/area (=intensity/(area·time) on the vertical axis.

Application

Echography, also called (ultra)sonography, can visualize muscle and soft tissue, making them useful for scanning the organs, also of the unborn child. The use of microbubble contrast media to improve ultrasound signal backscatter is known as Contrast enhanced ultrasound (CEU) as e.g. applied in echocardiography and recent applications in molecular imaging and drug delivery.
Diagnostic echography scanners operate at 0.5 -15 MHz. More powerful ultrasound sources are used in non-imaging applications, such as to generate local heating in biological tissue (physical therapy, with powers of 3W/cm^2). Focused ultrasound sources may be used to break up kidney stones or for cataract treatment, cancer treatment and cleaning teeth. The latter applications rely directly on the generated mechanical vibrations.

A well-know application is the ultrasonic cleaner (f is 20-40 kHz). In medicine (and other disciplines), they are used for cleaning of surgical and dental instruments, lenses and other optical parts. The cleaner generate millions of microbubbles in the liquid of the cleaner. The main mechanism for the cleaning action is actually the energy released from the collapse of the cavitation of the bubbles. Ultrasound when applied in specific configurations can produce exotic phenomena such as sonoluminesence (the emission of short bursts of light from imploding bubbles in a liquid when excited by ultrasound frequencies).

More Info

Sound absorption
Sound absorption in watery liquids is dependent on various variables, but most strongly on frequency. Other variables are the concentration of salts and other compounds, temperature, hydrostatic pressure (ref. 1 and 2) and seemingly also its (solution and colloidal nature).
Sound absorption behaves accordingly the Lambert-Beer law for light and is similarly characterized by the absorption coefficient α that gives the logarithmic decrement of the sound per unit distance. For sound is has the dimension frequency (f) squared per distance. Since sound intensity is generally given in dB, the logarithmic measure of α is given as α', with dimension f^{2}dB/km. In the literature, for pure water several values of α and α' are given and sometimes deviations are substantial (compare ref. 1 and 2). In pure water the absorption of the sound intensity, according to ref. 2 is by approximation:

$$\alpha = 0.00049 f^2 e^{-(T/27+p/170)} \text{ dB/km, or } \alpha' = 11.28\, f^2 e^{-(T/27+p/170)} \text{ /cm} \qquad (1)$$

since α=α'/4.343 (10log(e) = 4.343),
where f is frequency (Hz), T temperature in ^{o}C and p hydrostatic pressure (bar). For high frequencies α is strongly dependent on dissolved salts, the reason why in seawater and watery tissues the sound absorption is much higher as Table 1 indicates. The equation also indicates that with higher ultrasound frequencies penetration depth reduces with f^2.

The safety index
Exposures > 120 dB SPL (see Audiology and audiometry) may result in hearing loss and > 155 dB SPL can give burning.
For practical safety of an echo-apparatus the thermal index (TI) and the mechanical index (MI) is used. TI expresses the warming of the tissue, with different measures for soft tissue, bone and cranial bone. In general, the thermal index (TI) is defined by the relationship:

$$TI = W_p/W_{deg}, \qquad (2)$$

where W_p is the power of the impinging beam and W_{deg} is the estimated power necessary to raise the target tissue 1 ºC. As an absolute maximum TI < 4 is considered for 15 min sessions.
MI expresses the risk of cavitational effects of bubbles. (Cavitation is the phenomenon where small cavities of partial vacuum form in fluid, then rapidly collapse, producing a sharp sound. Cavitation occurs in pumps, propellers etc.). MI is defined as the quotient of the peak negative pressure P^- and the square root of the ultrasound frequency, i.e.:

$$MI = P^-/\sqrt{f_0} \qquad \text{(in MPa MHz}^{-1/2}\text{)} \qquad (3).$$

MI<1.9 is safe. (4 MPa is also used). This is more conservative for echography in fine structures). When exposed to high MI (> 0.6, i.e. the MI often used for standard imaging) the bubbles oscillate wildly and burst. Upon destruction, micro-bubbles produce a brief, high amplitude signal, with good resolution. Caviation, in principle, can result in microtraumata. However, in specific applications, cavitation is required and without risk of damage, as in some versions of Contrast enhanced ultrasound. When the bubbles are filed with air, after bursting the air is rapidly dissolved in the liquid phase.
The safety limits holds also for the Doppler mode (see Echography).

Constructing the image

The computer processes three things from each electrical impulse received: 1.) sorting out the impulse of the tens of transducers; 2.) determining the impulse amplitudes; 3.) determining the time interval between sending and receiving the impulses. Now, the image (grey scale) can be constructed. Transforming the electrical signal into a digital image looks like the use of a spreadsheet. The transducer receiving the impulse determines the 'Column', the time that it took to receive the impulse determines the 'Row', and the amplitude of the impulse determines the gray scale that the spreadsheet cell should change too (white for a strong pulse).

References

1. Fisher FH and Simmons VP Sound absorption in sea water. J. Acoust. Soc. Amer, 1977:62;558-564.
2. http://www.npl.co.uk/acoustics/techguides/seaabsorption/physics.html.
3. Wikipedia Ultrasound

Hearing and Audiometry

Audiology and audiometry

Principles

Audiology is the study of hearing, psychophysically as well as the physiological underlying mechanisms, peripherally and centrally. Vestibuology studies the sense of balance which sensory organs are the semicircular canals and the otoconial system (see Vestibular mechanics).
Audiometry is the testing of hearing ability. Typically, audiometric tests determine a subject's hearing levels, but may also measure the ability to discriminate between different sound intensities, recognize pitch, or distinguish speech from background noise. The tympanogram (see Tympanometry), the acoustic reflex or Stapedius reflex and Otoacoustic emissions may also be measured. Results of audiometric tests are used to diagnose hearing loss or diseases of the ear.

Sound pressure level and hearing level
Although pressure is measured in Pascals (Pa; see Sound and Acoustics), its amplitude is generally referred to as sound *p*ressure *l*evel L_p and measured in dB, abbreviated as dB SPL. 1 dB SPL ≡ 20 µPa, the reference sound pressure p_0, which is the standard human threshold at 1000 Hz. L_p is defined as the logarithmic ratio of the energy. This is the same as the log-ratio of the squared sound pressures of the sound p_1 and the reference p_0. In formula:

$$L_p = 20\log(p_1 p_0) = 10\log((p_1^2/p_0^2)) \text{ dB SPL} \quad (1)$$

In *underwateracoustics* a reference level of 1 µPa is used.
Table 1 presents values of some basic characteristics of various sounds.

Table 1 Typical sounds with their characteristics

sound	distance	sound level	pressure P_{eff}	particle velocity	displacement eardrum
	m	dB SPL	µPa	µm/s	µm
1000 Hz threshold		0	20	0.05	$0.1\cdot10^{-6}$
leave rustle	5	10	63	0.16	$0.3\cdot10^{-6}$
whispering	2	20	200	0.5	$1\cdot10^{-6}$
traffic	107	60	$20\cdot10^{3}$	50	$0.1\cdot10^{-3}$
normal speech	2	65	$36\cdot10^{3}$	89	$0.2\cdot10^{-3}$
pneumatic hammer	2	90	$0.63\cdot10^{6}$	$1.6\cdot10^{3}$	$3.1\cdot10^{-3}$
airplane	50	100	$2\cdot10^{6}$	$5\cdot10^{3}$	0.01
pain limit		120	$20\cdot10^{6}$	$50\cdot10^{3}$	0.1

For particle velocity ($p = \rho c\mathbf{v}$)see Sound and Acoustics. (From ref. 1.)

The ear drum as pressure-to-displacement transducer
Table 1 shows that at the 1000 Hz threshold particle velocity and eardrum displacement is extremely small, smaller than the diameter of a H_2O molecule. From here, a lot of amplification is needed to reach a hearing sensation. The acoustic impedance ($Z=\rho c$, with ρ the specific density of the medium and c the sound speed, see Acoustic impedance) of water, and so of the watery tissues in the cochlea, is about 4000 times that of water. Therefore, air born sound hardly penetrates water (loss ca. 70 dB). The mismatch in impedance is solved by the eardrum, which comes in mechanical vibration due to the sound pressure. In most conditions, the auditory system only utilizes the sound pressure and not particle displacement (as in bone-conduction) as input signal.
The *bones in the middle ear* amplify the eardrum displacements and transmit them to the *oval window*, which has a much smaller area (some 40 times). The oscillations of the oval window evoke a kind of displacement wave in the scala media.

Perceived sound strength: phon and sone
Since the perceived loudness correlates roughly logarithmically to sound pressure, a new measure is introduced, the phon. This is a psychophysical measure of perceived loudness. At 1 kHz, 1 phon is 1 dB above the nominal threshold of hearing, the sound pressure level expressed in dB SPL (rel. 20 µPa). By definition, two pure tones that have equal phones are equally loud. An equal-loudness contour, also called a loudness level contour or a Fletcher-Munson curve, is a measure in dB SPL versus frequency for which a listener perceives a constant loudness. Fig. 1 gives the isophon curves of human hearing
The loudness levels of each curve are expressed in phones, given by L_N.

Another psychophysical measure is the sone: 1 sone equals 40 phon. Since at 1 kHz the level in phone is roughly proportional with the level in dB SPL, L_N and the level in sones N is as follows related:

$$N = 2^{(LN\ -40)/10}. \qquad (2a)$$

The other way around it holds that:

$$L_N = 40 + 10^2 \log N. \qquad (2b)$$

In this way, 1, 2 and 4 sone is 40, 50 and 60 phone and so on.

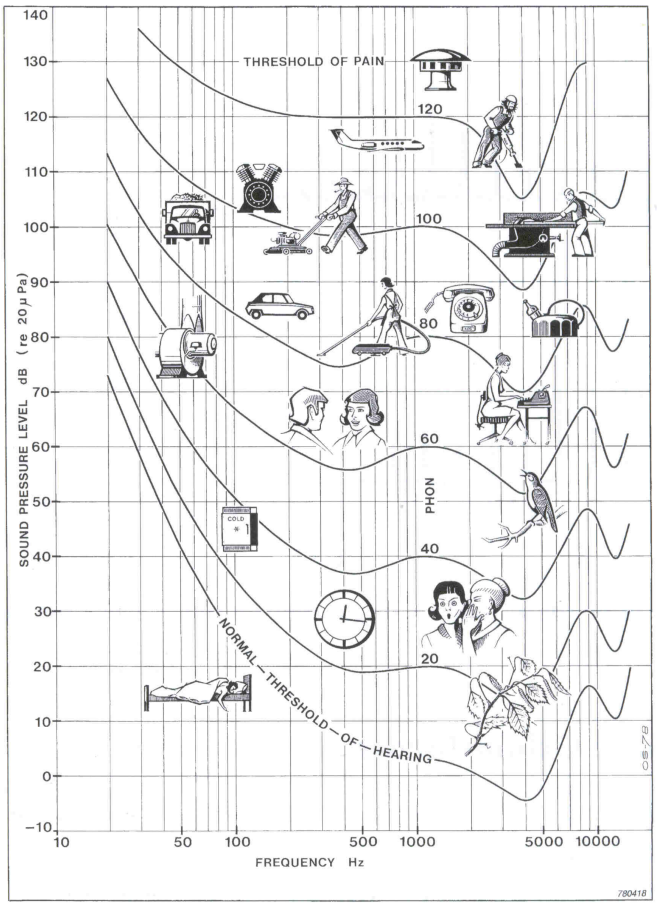


Fig. 1 Isophones of human hearing

Applications

In case of hearing, loss the clinician likes to know where in the pathway from outer ear to cortex there is a disorder. To find this out there are a number of psychoacoustic tests and measurements of the performance of the 'hardware' of the auditory system, from tympanum to auditory cortex.

Psychophysical tests

An *audiogram* is a graphical representation of how well sound frequencies, i.e. pure tones, are perceived. It is the normalized threshold curve (the most lower curve of Fig. 1), converted from dBSPL to dBHL, with HL hearing level. This curve has a level of 0 dBHL from 250Hz to 8kHz and represents the standard of adult hearing. Normal (healthy) hearing is between -10dBHL and 15dBHL.

Audiograms are generally measured with headphones. This gives a slightly different results as compared in the free field (performed in an anechoic, i.e. an echo-free room). Sometimes the audiogram is not recorded with pure tones but with small band noise (e.g. 1/3 octave noise) or with a pure tone masked by background noise.

In addition, all kind of audiological tests can by performed, for instance focused on a just audible frequency or intensity difference etc, time resolution (gap detection), whether or not with a noise mask. This mask can be presented just before the probe signal (forward masking), during the probe signal or after it (backward masking). Further, there are tests for speech perception (repeating one-syllable words

by the patient), for amplitude and frequency modulation, for directional hearing, also with masking by noise (*cocktail party effect*) etc. Some of these tests may have a more experimental rather than clinical character.
A *bone conduction* hearing test is administered to figure out whether the hearing loss is conductive (caused in the outer or middle ear) or sensorineural (cochlear). In this test, a vibrator is placed on the head behind the ear or on the forehead. The vibrator emits frequencies, and the person being tested hears tones or beeps just like in the test with earphones.
A routine audiological examination comprises a pure tone audiogram, often an audiogram with a noise mask and tests for speech perception.
Measurements of the 'hardware' of the auditory system
A tympanogram (see Tympanometry) is performed in order to examine the function of the tympanum and the middle ear. The acoustic reflex test (Stapedius Reflex) measures the reflexive contraction of the stapedius muscle, which is important in protecting the ear from loud noises. A brain stem potential tests whether the sound is processed by the various auditory nuclei of the brainstem. This response comprises many peaks and negativities that are characteristic for a certain auditory structure. Finally evoked potentials, measured by Electroencephalography) or evoked fields, measured by Magnetoencephalography, reveal whether central processing is free from disorders.

References
1. Van Oosterom, A and Oostendorp, T.F. Medische Fysica, 2nd edition, Elsevier gezondheidszorg, Maarssen, 2001.

Bone conduction

Principle

There exist two mechanical acoustic pathways to the inner ear:
- the regular one via the ear canal and middle ear
- the other via the body tissues, especially those in the head, the so-called "bone" conduction pathway.

The sound vibrations in the air evoke to and fro vibrations of the whole body. Self-evident, those of the head are most important. They are transmitted via the skin to the deeper tissues. Inside the head, due to differences in mechanical and acoustic properties between the head tissues, the tissues vibrate with different amplitudes and phases. This gives rise to the second pathway, the "bone" conduction pathways. The vibrations of the tissues of the head, such as the skull bones, stimulate the middle ear bones and at the same time, small vibratory deformations of the bony encapsulation of the cochlea emit waves in the basilar membrane and tectorial membrane. However, which anatomical structures, soft tissues and skeletal structures under what conditions are precisely involved and to what extent is still enigmatic.
"bone" conduction tends to amplify the lower frequencies, and so most people hear their own voice with a lower pitch than it actually is.

Application

Some hearing aids and non-clinical devices employ "bone" conduction to hearing directly by means of the ears. In these cases, a headset is ergonomically positioned on the temple and cheek and the electromechanical transducer, which converts electric signals into mechanical vibrations, sends sound to the internal ear through the cranial bones and other tissues. Likewise, a microphone can be used to record spoken sounds via "bone" conduction.

Most common are ears-free headsets or headphones. An example is the "bonephone" that is designed for stereo presentation of sounds, rather the traditional "bone"-conduction vibrator used in clinical audiology settings.
A bone-anchored hearing aid uses a surgically implanted abutment to transmit sound by direct bone conduction to the inner ear, bypassing the not-functional middle ear. A titanium "plinth" is surgically embedded into the skull with a small abutment exposed outside the skin on which a sound processor is attached. This transmits sound vibrations, finally vibrating the skull and so the inner ear.
Further, there exist specialized communication products (i.e. for underwater & high-noise environments). An example is a bone conduction speaker with a rubber over-molded piezo-electric flexing disc about 40

mm across and 6 mm thick used by SCUBA divers. The connecting cable is molded into the disc, resulting in a tough, waterproof assembly. In use, the speaker is strapped against one of the dome-shaped bone protrusion behind the ear. The sound produced seems to come from inside the user's head, but can be surprisingly clear and crisp.
Under water, “bone” conduction transmission can be used with individuals with normal or impaired hearing. Under water, there are several transmission pathways for sound, but which one is used is not clear. Some of the products that exploit “bone” conduction include the AquaFM and the SwiMP3, which are devices for transmitting sound to swimmers.

“Bone” conduction hearing devices have the following advantages over traditional headphones:
- Ears-free, thus providing more comfort and safety.
- No electromagnetic waves, eliminating the potential effect of such waves on the cerebrum, if any.
- High sound clarity in very noisy environments can be used with hearing protection.

Disadvantages of some implementations are the requirement of more power than headphones and the reduced frequency bandwidth.

More Info

“Bone” conduction is a highly complicated process (e.g. ref. 5). Despite this, it is certain that the pathway of the stimulation with a tuning-fork on the scalp is different from the pathway when for instance, the head is submerged and an underwater source generates the stimulus. For instance, the air filled head cavities are irrelevant with the tuning-fork, but they play a role when hearing with a submerged head, since then the cavities reflect the sound waves. In air, with air-born sound, the “bone” conduction pathway is again different but now its contribution in hearing is strongly overruled by the canal-tympanic pathway. For frequencies up to about 1 kHz and at 4 kHz, the tympanic pathway is ca. 55 dB more sensitive, and at 2 and 8 kHz about 45 dB (ref. 3), rather well in accordance with earlier measurements up to 2.5 kHz (ref. 1).

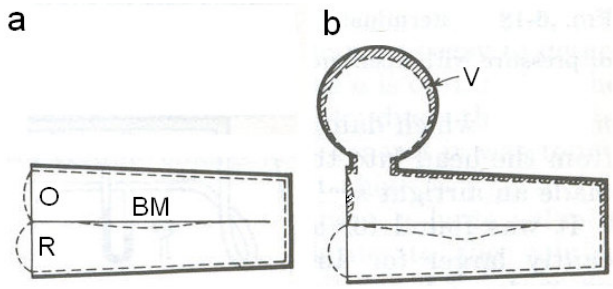


Fig. 1. Basilar membrane (BM) excursion due to translatatory (a) and compressional (b) vibrations. O oval window, R round window, V vestibule. See for further explanation the narrative text of **More Info**. (Modified from ref. 1

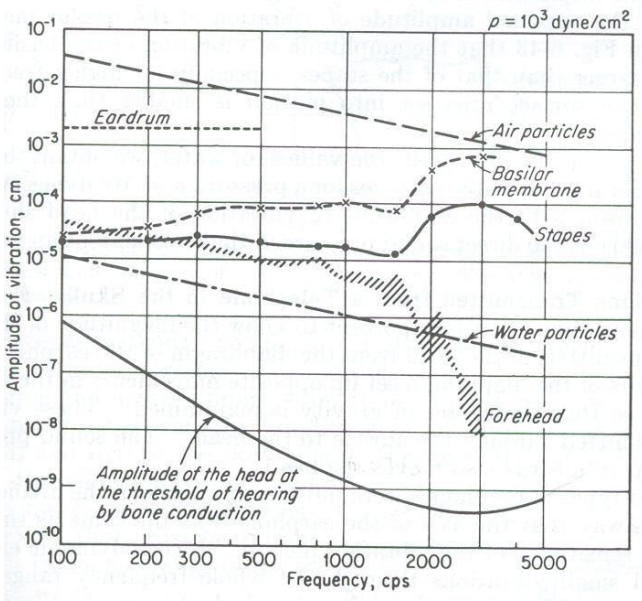


Fig. 2. From von Békésy (1960). See text for explanation.

The impact of sound waves on the head results in three types of “bone” conduction:

- the direct movement of the whole head producing an inertial reaction of the auditory ossicles;
- the penetration of sound waves into the head (with a theoretical loss of some 66 dB in air);
- the deformation of the skull by the mass reaction acting during the vibrations.

For frequencies up to about 1 kHz (1) dominates and the theoretical calculated loss is about 50 dB (ref. 1. For higher frequencies skull deformations become important. They result in translatatory and compressional vibrations indicated by panel (a) and (b) respectively of Fig. 1, which causes excursion of the basilar membrane. In (a) it is due to unequal mobility of the both windows and in (b) mainly by compressibility of the vestibule and semicircular canals.
Fig. 2 gives the result of measurements of the vibration of the basilar membrane, stapes and forehead with air born sound and the head in the free field with open ear canals and a strong air-born source of 100 Pa (= 1000 dyne/cm^2, equivalent to 134 SPL. The lower curve is measured at the threshold of “bone” conduction.
Recent research suggest that the 'classical' “bone” conduction theory with a vibrating tuning-fork as stimulator should be modified to include a major pathway for cochlear excitation which is non-osseous. The bone vibrations may induce audio-frequency sound pressures in the skull content (brain and cerebro-spinal fluid) which are then communicated by liquid-like channels to the liquids of the inner ear (ref. 2). A similar conclusion was drawn from results of experiments with the bullfrog (ref. 4).
We may conclude that the classical theory that the “bone-conduction” pathway is dominated by conduction via skull bones is questionable. A fluid pathway is probably at least as important.

References

1. G. von Békésy. Experiments in hearing, McGraw-Hill Book Company, 1960.
2. Freeman S, Sichel JY, Sohmer H. Bone conduction experiments in animals - evidence for a non-osseous mechanism. Hear Res. 2000;146:72-80.
3. Reinfeldt S, Stenfelt S, Good T, Håkansson B. Examination of bone-conducted transmission from sound field excitation measured by thresholds, ear-canal sound pressure, and skull vibrations. J Acoust Soc Am. 2007;121:1576-87.
4. Seaman RL. Non-osseous sound transmission to the inner ear, Hearing Research 2002;166:214-215.
5. Taschke H and Hudde H. A finite element model of the human head for auditory bone conduction simulation. ORL J Otorhinolaryngol Relat Spec. 2006;68: 319-323.

Echolocation by marine mammals

Principle

Echolocation, also called biosonar, is the biological sonar used by mammals e.g. dolphins, most whales, shrews and some bats. Technical sonar (Sound Navigation And Ranging) is a technique that uses sound (ranging from infra- to ultrasound) propagation (usually underwater) to navigate, communicate or to detect underwater objects. Two bird groups also employ biosonar for navigating through caves.
The physical principle is the same as applied in medical echography: sender (small-angle sound beam), reflecting object, receiver and analysis. The latter is basically the measurement of the time delay between sending and receiving to calculate the object distance and other characteristics. To obtain a well defined time delay echolocating animals emit very short calls (a few ms) out to the environment. The calls can be constant frequency (single or composed) or for instance frequency modulated (FM). They use these echoes to locate, range, and identify the objects. Echolocation is used for navigation and for foraging (or hunting) in various environments.

Animal echolocation relies on both ears which are stimulated by the echoes at different times and loudness levels, depending on the source position. These differences are used to perceive direction and distance, object size and other features (e.g. kind of animal).

Toothed whales (suborder Odontoceti), including dolphins, porpoises, river dolphins, orcas and sperm whales, use biosonar that is very helpful when visibility is poor, e.g. in rivers and estuaries due to light absorption and especially turbidity (see Light: scattering).
Echoes are received using the lower jaw and especially its filling of waxy tissue as the primary reception path, from where they are transmitted to the inner ear (Fig. 1). Lateral sound may be received though fatty lobes surrounding the ears with a similar acoustic density to bone.

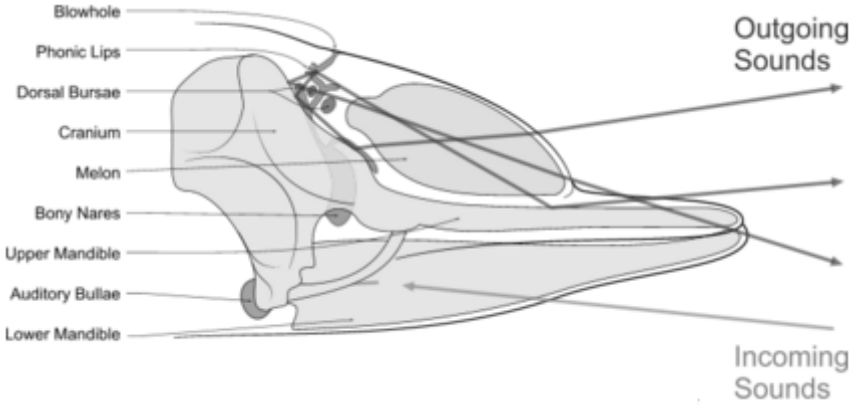


Fig. 1 Diagram illustrating sound generation, propagation and reception in a toothed whale. Outgoing sounds are drak gray (red) and incoming ones are light grey (green). (From ref. 5.)

Toothed whales emit a focused beam of high-frequency clicks in the direction that their head is pointing. Sounds are generated by passing air from the bony nares through the phonic lips. These sounds are reflected by the dense concave bone of the cranium and an air sac at its base. The focused beam is modulated by a large fatty organ, the 'melon'. This acts like an acoustic lens because it is composed of lipids of differing densities. Most toothed whales use a click train for echolocation, while the sperm whale may produce clicks individually. Toothed whale whistles do not appear to be used in echolocation. Different rates of click production in a click train give rise to the familiar barks, squeals and growls of the bottlenose dolphin. In bottlenose dolphins, the auditory EEG response resolves individual clicks up to 600 Hz trains, but yields a graded response for higher repetition rates.
Some smaller toothed whales may have their tooth arrangement suited to aid in echolocation. The placement of teeth in the jaw of a bottlenose dolphin, as an example, are not symmetrical when seen from a vertical plane. This asymmetry could possibly be helpful if echoes from the dolphins biosonar are coming from above or below. A similar anatomical asymmetry (ear position) for elevation localization has been found in the barn owl.

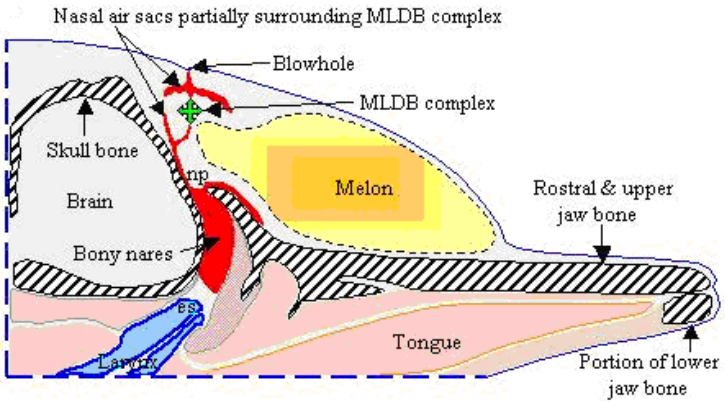


Fig. 2 Parasagittal slice in a plane lying slightly to the right of the midline of the common dolphin forehead. This diagram includes the skull and jaw bone (gray hatched), the nasal air sacs (black or in colour red), the right MLDB (monkey lip/dorsal bursae) complex, the nasal plug (np), the epiglottic spout (es) of the larynx, and the melon tissue. Other tissues, including the muscle and connective tissue surrounding many of the labeled forehead tissues, are not shown.

Table 1 summarizes the most important differences between cetaceans and human.

	Cetaceans (Fully Aquatic Ears)	**Human (Aerial Ears)**
Outer ear	Temporal bone is not a part of skull	Temporal bone is a part of the skull
	No pinnae	Pinnae
	Auditory meatus is plugged	Air-filled auditory meatus
Middle Ear	Middle ear is filled with air and vascular tissue	Middle ear completely air-filled
	Basilar membrane thin and broad at apex of odontocetes	Basilar membrane thick and narrow at the basal end
	Mysticetes basilar membrane thinner and wider than odontocetes and humans	
Inner Ear	Strong support of basilar membrane for odontocetes, less for mysticetes	Little support of basilar membrane
	Long basilar membrane length	Short basilar membrane length
	Semi-circular canals are small	Semi-circular canals are average to large
	Large number of auditory nerve fibers	Average number of auditory nerve fibers

From ref. 3.

Dolphins use a narrowly focused sound beams which mainly emanates from the forehead and rostrum during echolocation as illustrated in the simulation of Fig. 3. The beam formation results primarily from reflection from the skull and the skull-supported air sac surfaces. For the frequencies tested, beam angles best approximate those measured by experimental methods for a source located in a region of the MLDB complex. The results suggest that: 1) the skull and air sacs play the central role in beam formation; 2) the geometry of reflective tissue is more important than the exact acoustical properties assigned; 3) a melon velocity profile of the magnitude tested is capable of mild focusing effects; and 4) experimentally observed beam patterns are best approximated at all frequencies simulated when the sound source is placed in the vicinity of the MLDB complex.
Examples of vocalizations are given by ref. 4.

Application

Dolphins are used as reconnoiterers in navies of various counties. Also their echolocation system was of interest for technical military applications.

More Info

In echolocating mammals, the cerebral analysis of the echo's may be complicated since the animal has to correct for the speed of self-motion and that of the (living) object. Suppose that a marine animal is at a distance of 10 m from a prey. Then the echo returns after 2x10/1437 = 13.9 ms. Now suppose that the predator has a speed of 2 m/s (=7.2 km/h) in the direction of the prey and the prey swims with the same

speed towards to animal. In 14 ms the both animals are by approximation 4 m/s x 14 ms = 5.6 cm closer together, which makes the localization still accurate enough. This example shows that for marine animals self-motion and prey-motion are irrelevant. For terrestrial animals the outcome is however different. Due to the nearly five times lower sound speed and about 5 times higher velocities (10 m/s) of flying predators and prey (bats, insects) flying speeds become relevant. Again at a distance of 10 m both animals are now about 1.18 m closer. For prey catching this is not negligible (supposing that the trajectory of the insect is partly sideward). It is also obvious that with these high speeds the predator has to cope with the Doppler effect. Bats have cerebral mechanisms to compensate for that and also to estimate the speed and speed-direction of the prey.

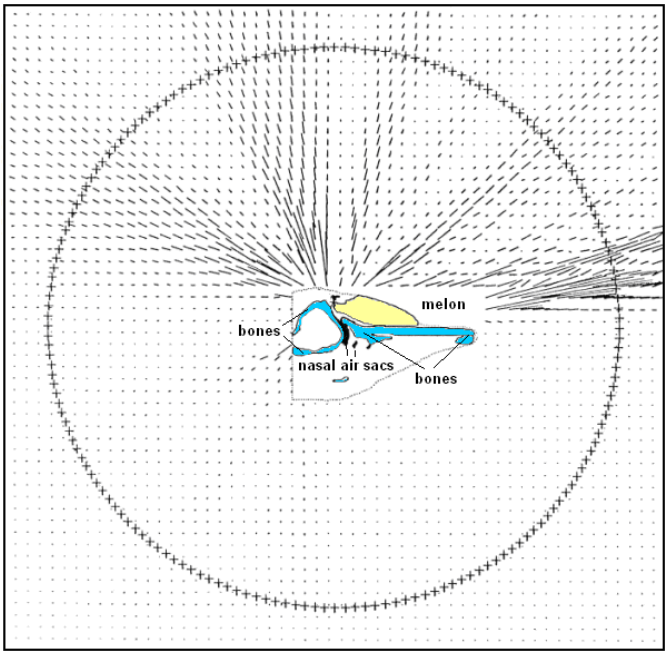


Fig. 3 A 2D computer simulation of sound production in the common dolphin. The dolphin's head (based at parasagittal CTs) is indicated by a dotted outline. Sonar pulses from a spot (just beneath the uppermost air sac) below the dolphin's blowhole reflect and refract through these structures. The lines around the dolphin's head represent the direction and intensity of sound waves emitted from the model. Most of the acoustic energy is emitted in a forward and slightly upward-directed beam in this 100 kHz simulation. (from ref. 1.)

References

1. Aroyan JL, McDonald MA, Webb SC, Hildebrand JA, Clark D, Laitman JT, Reidenberg JS (2000) Acoustic Models of Sound Production and Propagation. In: Au WWL, Popper AN, Fay RR (eds), Hearing by Whales and Dolphins. New York: Springer-Verlag, pp. 409-469.
2. Ketten DR. The marine mammal ear: specializations for aquatic audition and echolocation. In The evolutionary biology of hearing, Webster DB, Fay RR and Popper AN (eds.), 1992, pp 717-750. Springer, New York.
3. http://www.dosits.org/animals/produce/ceta.htm.
4. http://neptune.atlantis-intl.com/dolphins/sounds.html.
5. Wikipedia – the free encyclopedia.

Otoacoustic Emission

Basic principle

An otoacoustic emission (OAE) is a sound generated by an inner ear excitation by a number of different cellular mechanisms. OAEs disappears with inner ear damaged, so OAEs are often used as a measure of inner ear health. There are two types of otoacoustic emissions: Spontaneous Otoacoustic Emissions (SOAEs) and Evoked Otoacoustic Emissions (EOAEs).
OAE is considered to be related to the amplification function of the cochlea. OAEs occur when the cochlear amplifier is too strong. Possibly outer hair cells enhance cochlear sensitivity and frequency selectivity and provide the energy. The outer hair cells have few afferent fibers but receive extensive efferent innervation, whose activation enhances cochlear sensitivity and frequency discrimination.

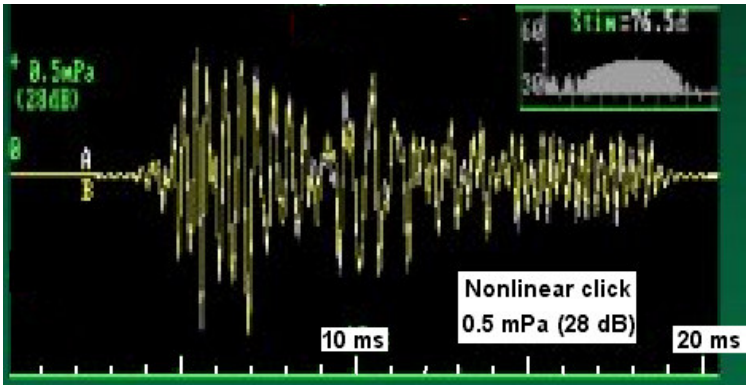


Fig. 1 Transient EOAE of a healthy ear

Applications

Last decade, OAEs became increasingly important in the clinical audiological practice for the diagnostics of middle and inner ear disorders, retrocochlear disorders, and tinnitus.

More Info

EOAEs are currently evoked using two different methodologies. Transient EOAEs (TEOAE or TrEOAE) are evoked using a click stimulus with a repetition frequency of about 20 Hz. 98% of the healthy ears have TEOAEs (0.5-4 kHz), but above 60 year only 35%. The evoked response from this type of stimulus covers the frequency range up to ca. 4 kHz. Originalla, only transient potentials were used. Distortion Product OAEs (DPOAE) are evoked using a pair of tones with particular intensities (usually either 65 - 55 dB or 65 for both) and frequency ratio's (F2/F1). The evoked response in the EEG from these stimuli occurs at a third frequency. This distortion product frequency is calculated based on the original F1 and F2. Inner ear damage diminishes the distortion product.

It is still uncertain if there is an interaction via the auditory brainstem nuclei between the outer hair cells of both inner ears regarding OAEs. In general, contralateral stimulation does not elicit EOAEs. The frequency resolution, e.g. for speech, depends on very fast modulation of the incoming signal. Due to the propagation time between both inner ears in case interaction, this modulation would lag behind if otoacoustic emissions in one ear would effect the opposite one.

Literature

1. "http://en.wikipedia.org/wiki/Otoacoustic_emission"
2. Probst R, Lonsbury-Martin BL, Martin GK. A review of otoacoustic emissions. J Acoust Soc Am. 1991, 89: 2027-67.

Physics of outer ear

Principle

The auricle (pinna) and auditory canal (meatus externus) together form the outer ear. The pinna's most important function is the reflection and channeling of the incoming pressure-waves to the opening of the auditory canal. In humans, channeling is a minor factor, but for certain animals (e.g. rabbits, owls) it contributes significantly to hearing. This channeling is of major importance for improvement of directional hearing in vertical directions. The most important auditory effect of the hearing canal is the increase in sensitivity around 3½ kHz.
Audiograms are preferably measured in anechoic chambers with loudspeakers on a large distance of the head. Remaining reflections, despite the absorbing structures, can be cancelled by electronic-acoustic feedback. This is an imitation of the free field measurement (no vegetation). However, routinely the measurements are performed with headphones or with earphones. It has been proven that these measurements give little difference with the measurements with loudspeakers. I.e. for the sensitivity of pure tones (or 1/8 octaves tones) the presence of a pinna is without function. However, this does not apply to source localization (direction and distance).

Source location in the horizontal plane
Source location is a complicated cerebral process since for pre-processing only the tonotopic organization of both basilar membranes is available. For localization in the horizontal plane the interaural phase difference (IPD) becomes available as a result of the difference in time of arrival caused by the interaural distance difference with regard to the source. This applies up to 1500 Hz. In addition, the interaural intensity difference (IID) is of importance, for example caused by a distance difference or by head shade. This applies above 1500 Hz, because waves of lower frequencies deflect around the head what reduces the IID. IID's amounts many dB's, up to a maximum of 35 dB. IPD and IID can concern a pure tone (as in a clinical test or laboratory experiment), but also a complete spectrum. The later applies in daily practice.

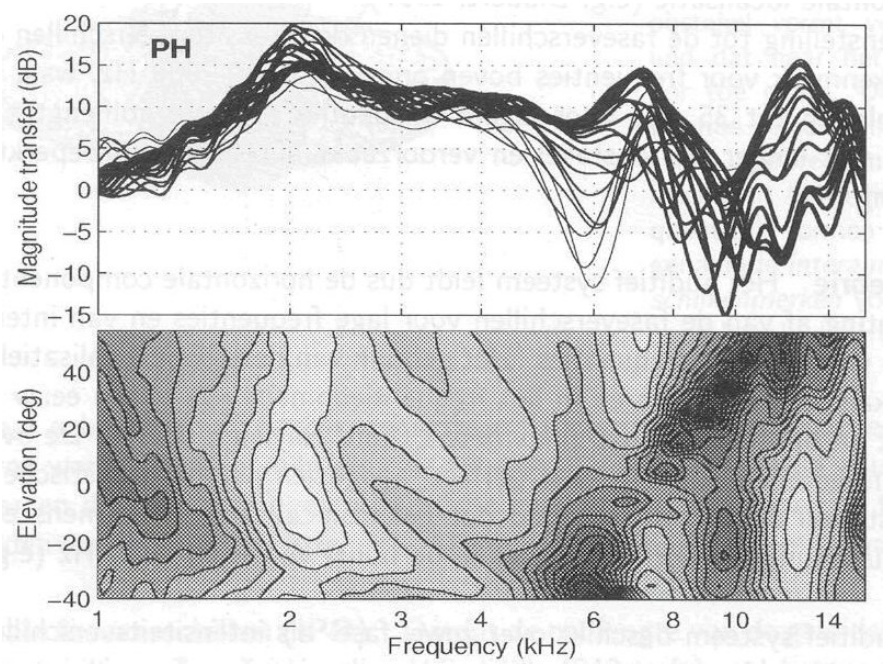


Fig. 1 Transfer of the pinna as a function of frequency. Upper panel: the lines give the angle of the elevation in the median plane of -40, -35,…, +55 degrees. The line thickness increases with the angle. Lower panel: the same data, but now in the elevation/frequency plane. More blackening represents a larger gain as a result of the pinna-acoustics. (From ref. 1.)

Pinna and source elevation
For source localization in the *vertical plane*, the pinnae are of importance. The signal impinging upon the tympanum is the sum of the sound which comes straight forward from the source, and the signal which reaches the tympanum by means of reflections of the ear flap. Generally these signals reinforce each

other, but can also (partly) cancel each other. The ear flap adds an acoustic “pinna image” which is dependent on the direction (elevation and azimuth) of the sound source. For the azimuth this is not relevant since then IID’s and IPD’s are the cues for localization, except for the case to solve the in front – behind (180^o) ambiguity (symmetrical front – behind locations give the same IPD and IID). Below 4 kHz the transfer characteristic of the auricle is little dependent on the elevation, although it can differ 15 dB from frequency to frequency. Without pinna there is an average loss of 10 dB. Above 4 kHz the contribution of the auricle is strongly elevation dependent, up to 20 dB, and moreover more frequency-dependent than below 4 kHz (Fig. 1). Besides the contribution of the pinna, the head and the shoulder contribute to some extent.

The ear canal (meatus)
Sound travels down through the (external) meatus and sound pressure wave causes the tympanum to vibrate. Frequencies will resonate when they can perform standing wave behaviour. The auditory canal can be considered as a vibrating cavity with the pinna as open end (ventral segment) and the tympanum as a closed end (node), analogue to the open organ pipe. Resonance occurs when ¼λ has the length (L) of the meatus, with λ the wavelength of the frequency. The average ear has a meatus of 2.5 cm in length, resulting in a resonance frequency of 3.4 kHz (with c_{sound} = 343 m/s). This frequency causes the dip in the audiogram of about 6 dB SPL between 3 and 4 kHz.

Application

The physics of the ear canal has been examined thoroughly for development of audiological equipment (such as an earphone, see **More Info**) and hearing aids.

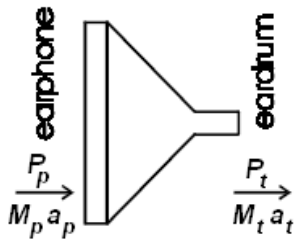
More Info

There are more frequencies which fulfil the condition of having a node at the tympanum and a ventral segment at the pinna. Their wavelengths and frequencies are:

$$\lambda = 4L/(2n+1) \text{ and } f = c(2n+1)/(4L). \quad (1)$$

Consequently, the next resonance frequency (for n=1) is 3x3.4 ≈ 10 kHz. Its effect is generally too small (depending on the experimental technique) to be revealed in the threshold audiogram.
The amplitude variation in transversal direction caused by a planar wave at the entrance of the meatus is irrelevant up to 4 kHz and beyond it slowly increases to almost 5 dB at 15 kHz. In practice it will be something more (no planar wave). The transfer in the canal can be calculated along the central axis, given the length along the curved central axis and the variation in diameters along this axis. More precise calculations have been done with a numerical 3D computer model (ref. 2).

Acoustic coupling between an earphone and the tympanum



The model of the acoustic coupling between the pinna and the eardrum is based on a funnel-shaped device. The air in the small tube on one end (the ear canal) of the coupler is treated as a lumped mass. A force $M_t a_t$ (with a the acceleration) is exerted on this mass (M) by the pressure at the earphone and the pressure at the eardrum. The air in the conical cavity is treated as a spring which couples the air in the tube with the diaphragm of the eardrum. Damping is included to represent losses in the conical cavity and the tube.

References

1. Hofman P. On the role of spectral pinna cues in human sound localization. Ph.D. thesis, Nijmegen, 2000.
2. Stinson MR, Daigle GA. Comparison of an analytic horn equation approach and a boundary element method for the calculation of sound fields in the human ear canal. J Acoust Soc Am. 2005;118:2405-11.

Physics of middle ear

Principle

The vibration of the tympanic is transmitted by the three middle ear ossicles (consecutively the malleus or hammer, incus (anvil) and stapes or footplate; Fig. 1). The middle ear with the three bones acts nearly perfectly as a pressure to displacement to pressure transducer with a sensitivity that is at quantum mechanical level. The acoustic power P, which enters the ear through the eardrum at the threshold of hearing (12 μPa at 4 kHz) is very small, about 12 aW (1 attowatt = 10^{-18} watt) and equivalent to an IR photon of 1660 nm in one second or the fluorescence of one chromophore transition/s from 527 to 400 nm.
This can be calculated from the relation between the intensity I (W/m^2) and the pressure p:

$$I = p\mathbf{v} = p(p/Z) = p^2/Z \qquad (1a)$$

And since power is intensity times area, it follows that:

$$P = AI = \pi r^2 p^2/Z. \qquad (1b)$$

Completing yields 11.7 aW. For 1 kHz and 20 μP 32.5 aW is found.

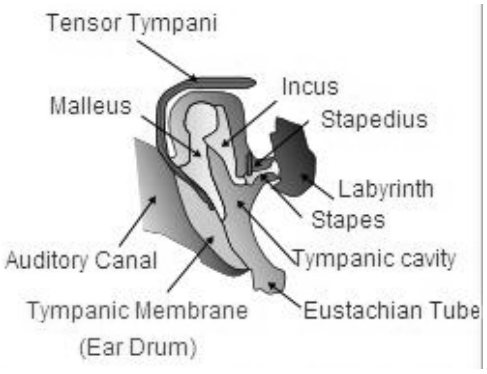


Fig. 1 Middle ear (from ref. 1).

The tympanic membrane is characterized by a large surface and low impedance (at both sides air), the oval window by a small surface and high impedance (from air to liquid). Generally speaking, going from a low to high impedance, there would not be hardly any transfer at all: the sound wave reflects. However, the ear has a very high sensitivity, up to the level that is set by the physics of quantum mechanics as described above. The problem is solved by impedance adjustment, which must take place in the middle ear. Impedance adjustment might be realized by increasing the pressure acting on the membrane of the oval window. Here, this pressure is transformed to displacement in the cochlear partitions.
Three principles are used to adjust the impedance in the middle ear:
1) The surface of the tympanic membrane is much larger than that of the oval window, which is practically the surface of the stapes. The force (F) produced by sound pressure (p) working on an area (A) is given by:

$$F = pA, \qquad (2a)$$

and assuming no losses, the sound pressure on the oval window (ow) of the scala vestibuli p_{ow} is:

$$p_{ow} = F/A_{ow}. \qquad (2b)$$

The result is that p_{ow} is considerably larger than the pressure at the tympanic p_t since $A_{ow} < A_t$. Since F is preserved, the gain in pressure is:

$$g = p_{ow}/p_t = A_t/A_{ow}. \qquad (2c)$$

The pressure at the tympanic membrane must be multiplied by a factor g in order to obtain the value for the pressure at the cochlea. Finally, g, the “gain” in pressure on the window is ca. 28 dB.

The gain is in pressure rather than in displacement since the ossicles can enlarge displacements only marginally. Simulations with a physical model showed that the displacement of the malleus and the stapes are rather the same, be it that for different frequencies with the same pressure at the tympanum the displacements differ more than a factor 10 (ref. 2) .

2) A second factor is the lever action of the middle ear bones. The arm of the malleus is longer than the arm of the incus. Normally a lever is in balance, i.e. the sum of the moment (M = F·L) of both arms is zero:

$$\Sigma M_n = 0, \text{ or} \quad (3a)$$
$$\left| F_{malleus} \cdot L_{malleus} = \right| F_{incus} \cdot L_{incus} \left|, \right. \quad (3b)$$

where L is length.
This causes a gain with a factor 1.3 (1 dB) at the stapes.

3) The third factor depends upon the conical shape of the tympanic membrane. Less surface moves due to the elasticity of the membrane when the stapes pushes the oval window in and out. This also increases the pressure, finally transformed to displacement of the oval window.

The first system is by far the most important one. This contributes to a gain of approximately 28 dB to the overall displacement gain of about 30 dB.

Another approach is to study qualitatively the action of the middle ear with a mechanical model (Fig. 2a), which can also be calculated through, although many parameters are numerically not known. Therefore, its action is hard to evaluate and has hardly practical value. However, we learn from it that the implicit assumption of the above description that all is frequency independent cannot hold. Systems with a mass or spring dictate that there is frequency dependency. That the transfer is frequency dependent is obvious from the model presented in Fig. 2a and also from the experimental results given in Fig. 2b. The later shows that only for a frequency of nearly 1 kHz the theoretical gain of about 30 dB is approximated.

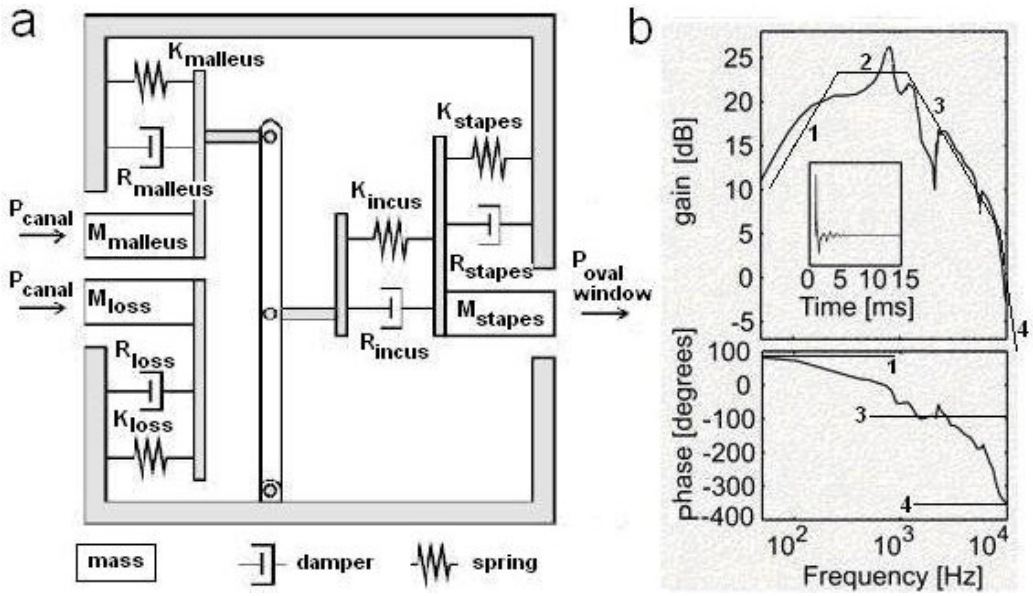


Fig. 2 (a) Mechanical model of the middle ear. The incus is fixed to the stapes, such that there is no cantilever function. The ineffective or unused motion is also modeled by the components indicated by "loss". (b) Top: Ratio of fluid pressure in the scala vestibuli to pressure in the ear canal in decibels (dB). The inset shows the click response. Bottom: phase characteristic. Numbers 1-4 denote asymptotes (see **More Info**). (Modified from ref. 3.)

Application

Knowledge of the biomechanics of the middle ear is of importance for e.g. optimizing replacement of ossicles (ossiculoplasty).
In terrestrial mammals, the high-frequency hearing limit (f_H) is roughly inversely proportional to the cubic root of the mass of the vibrating structures, i.e. $f_H \sim W^{-3}$, where W is the summed mass of malleus and incus (ref. 4). This means that an ossiculoplasty should be done with material as light as possible.

More Info

Frequency dependency of middle ear transduction
The gain characteristic of Fig. 2b can very roughly, as a first approximation, be interpreted as being caused by a band-pass second order filter, and indicated by the three asymptotes 1 to 3. At low frequencies, it gives rise to an asymptotic phase lead of 90° and at the high end of a 90 ° phase lag. However, above about 5 kHz three poles (and possible more) are added. Their action is indicated by the asymptote number 4, resulting in a final lag of 360° as the phase characteristic suggests.

Comparative aspects and a functional model of the mammalian outer and middle ear
Morphometric measurements have shown that there is a relation between the radius of the idealized tympanum r_t and the radius of the idealized middle ear volume r_v:

$$r_v = 0.05 + 1.685\ r_t. \quad (4a)$$

Similarly there is a relation between the wavelength λ of the ear canal resonance frequency f_d and r_t.:

$$\lambda = -\ 0.45 + 2.718\ r_t. \quad (4b)$$

The basic assumptions of a generalized model of the middle ear (ref. 4) are:

1. the middle ear behaves as a Helmholz resonator with resonance frequency f_b. This is a hollow solid sphere filled with a gas at much higher pressure than the ambient pressure. Via a small tube with a valve, the sphere can be depressurized. When the valve is opened very fast, the process of depressurizing is not continuous but intermittent with a certain frequency, i.e. the gas outflow resonates. This is based at *adiabatic* expansion (see Adiabatic compression and expansion)
2. both structures resonate with the same frequency, i.e. they are coupled:

$$f_d = f_b. \quad (5)$$

Both frequencies can be expressed by equations comprising anatomical parameters and material constants. With some assumptions about the tympanum (thin, uniform circular fixed at the rim, negligible stiff, ideal elastic and uniform tension) it is found that:

$$f_d = H \cdot M/(2\pi r_t) \text{ where } M = \sqrt{(T/m)} \quad (6)$$

H is a constant (= 2.405) holding for membranes like the one defined. M is a material constant derived from the tension T per unit length and mass m per unit area of the membrane.
For f_b of the Helmholtz resonator it holds that:

$$f_b = (c/2\pi)\sqrt{(S_m/L'V)},\ L' = 1.5 \cdot r_m + l_n, \quad (7)$$

where c is the sound velocity of air (352.9 m/s at 36 °C), r_m the radius of the neck opening, S_m the neck opening area into the meatus, L' the effective length of the resonator neck, l_n, the morphological neck length and V the middle ear volume.
By applying the model of Eq. 5 to a number of species, some general constants and species-specific parameters could be estimated. By comparing them with measured parameters, the model could be verified. For Homo sapiens it was found that f_d = 2.9 kHz, r_v = 1.93 mL, meatus cross area 44 mm^2, meatus length 22.5 mm and r_t = 3.22 mm. These values are rather close to the empirical values.

References

1. Wikipedia – the free encyclopedia.
2. Hudde H, Weistenhofer C. Key features of the human middle ear. ORL J Otorhinolaryngol Relat Spec. 2006;68:324-8. Review.
3. Mammano F. and Nobili R., http://147.162.36.50/cochlea/index.htm and http://147.162.36.50/cochlea/cochleapages/theory/index.htm
4. Plassmann W and Brändle K. A functional model of the auditory system in mammals and its evolutionary implications. In: The evolutionary biology of hearing, Webster DB, Fay RR and Popper AN (eds.), 1992, pp 637-653. Springer, New York.

Physics of cochlea

Principle

Functional anatomy
The cochlea is one of the parts of the inner ear (which also comprises the three semicircular canals and the two otoconia systems, see Fig. 1 of Vestibular systems). It is a spiraled, hollow, conical structure of three mechanically coupled canals, the scale vestibuli, scala media and scala tympani,and three separating membranes, the membrane of Reissner, the basilar membrane (BM), tectorial membrane (TM). Finally, there is the organ of Corti with the sensory hair cells (Fig. 1, 2).

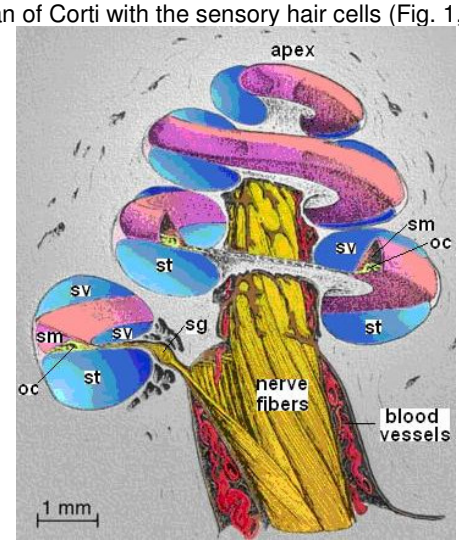


Fig. 1 Cochlea oc organ of Corti, sm scala media, st scala tympani, sv scala vestibule, sg spiral ganglion with the bipolar cell bodies of the auditory nerve. (After ref. 2.)

BM is nearly 35 mm long. The inner hair cells (IHCs) form a single row of ca. 5000 cells. The outer hair cells (OHCs, some 15000) occur in 3 or 4 rows. IHCs are innervated by numerous myelinated afferent nerve fibers and OHC by one afferent. Many unmyelinated efferent fibers innervate a single OHCs. OHCs make contact with TM, IHCs do not.

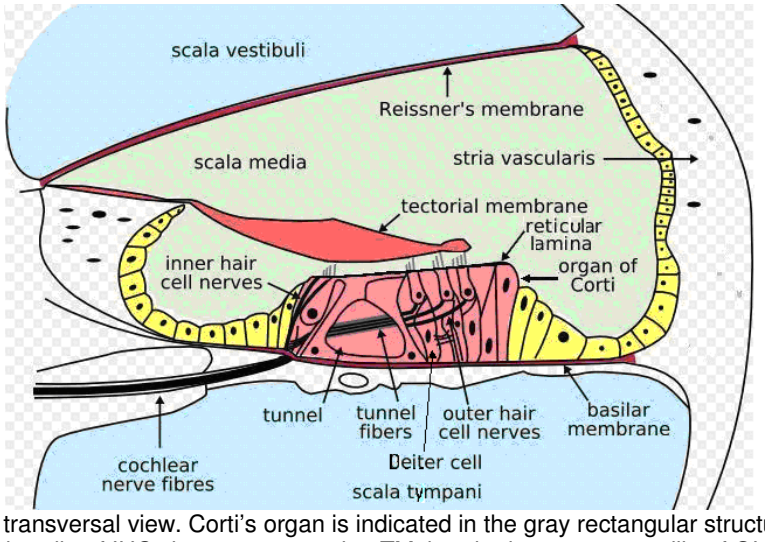


Fig. 2 Cochlear transversal view. Corti's organ is indicated in the gray rectangular structure in the middle. The hair bundle of IHCs is not connected to TM, but the longest stereocilia of OHCs do. (Modified from ref. 4.)

As a reaction to vibration of the oval window, TM and BM (basilar membrane) move up and down. This causes the hairs of the OHCs and IHCs to vibrate. The mechanical excitation of the IHCs, leads to electrical excitation, which results in IHC receptor potentials that gives rise to action potentials in the fibers of the auditory part of the VIIIth cranial nerve (the auditory-vestibular nerve).

Non-linear biomechanics and cochlear amplification

Hair cells convert sounds into receptor potentials when their stereocilia are deflected. IHCs act as the receptor cell. OHCs can act as a kind of motor cells for low and moderate sound levels, being their main function. They respond to variation in potential, and change length at rates unequalled by other motile cells. The forces generated by OHCs are capable of altering the delicate biomechanics of the cochlear partition, resulting in increased hearing sensitivity and frequency selectivity. The OHC electromotility goes far beyond the description of the cochlea as a simple frequency analyzer. It behaves as an active non-linear filter. In this view, frequency selectivity arises through the suppression of adjacent frequencies, a mechanical effect equivalent to lateral inhibition in neural structures. All these processes are explained by the interplay between the non-linear hydrodynamic interactions among different parts of the cochlear partition along the cochlea and the effective non-linear behavior of the electromotil OHCs.

Application

Precise knowledge of the tonotopical organization of OC is of importance for the placement of the electrode bundle of a cochlear implant, which is nowadays experimentally performed with 16 or 32 electrodes. Cochlear implants have routinely been applied more than 125,000 times (middle of 2008).

More Info

Non-linear biomechanics

The Reissner's membrane can be neglected in determining the stiffness of the endolymph sac. The major part of the stiffness of the organ of Corti is caused by BM, the reason why the biomechanics focuses at BM.

A traveling wave, starting at the oval window, runs along the BM with decreasing speed of propagation. The amplitude of vibration along BM first increases and then decreases quickly after reaching it maximum (Fig. 3). The point of maximal deflection is determined by the sound frequency. High frequencies resonate at a point near the windows, while low frequencies resonate more closely to the apex (also called helicotrema). Formerly, it was thought that practically the whole BM was mechanically stimulated by a pure tone. However, the mechanical properties of BM change along its length. From the windows with a membrane width of 0.04 mm and high stiffness (or tension), it changes to 0.5 mm width and low stiffness (ca. 2000 times less) near the apex. This gives rise to highly nonlinear behavior of BM resulting in narrowing of the BM excitation pattern.

The frequency dependency can be compared with that of length and tension of a violin string. It is denoted as the place theory or the tonotopical arrangement of frequencies. In this way, the nerve fibers can only be stimulated by a small frequency band with the most effective frequency, the best frequency, f_{best}, in the middle of the band.

The non-linearity is based on the mechanical coupling via the viscosity of the endolymph of adjacent segments of OC that act as oscillators. This gives a significant sharpening of the BM oscillations. Additional strong sharpening is provided by the active role of OHC cell bodies (see below).

Fig. 3 BM maximal response to a 250 Hz tone calculated with a model of the non-linear and active cochlear mechanics.

Active cochlear mechanics

In Fig. 4, the hair bundles of IHCs and OHCs are deflected by the shear displacement between the reticular membrane (RL) at the top of the OC and TM. As a result, the OHC cell bodies lengthen some 10%, driven by its receptor potential. The small box (in the upper right corner) shows the range of these displacements with OHC length change. The dot indicates the input-output relationship; its vertical motion is proportional to the input (BM displacement) and its horizontal motion is proportional to the output (IHC displacement). The dot goes beyond the limits of the box of Fig. 4. As a result, the oscillatory behavior of the BM-OHC-TM system is enhanced by the electrically driven length changes of the OHCs (called electromotility). This behavior was established by interferometric (see Interferometry) measurements based on laser beam reflections from the basilar membrane and RL. The changes in length of OHCs and Deiter cells enhance the resonance of BM by pumping energy into the mechanical system in the same way that one does when "pumping a swing". The energy contributed by OHCs will improve the sensitivity of the cochlea to low-level sounds, resulting in a larger stimulation of IHCs. This is the basis for the amplifier theory of cochlear mechanics.

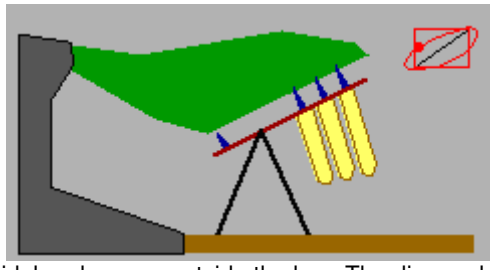


Fig. 4 The dot moves ellipsoidal and comes outside the box. The diagonal gives the relation without electromotility. See for animation of the movements ref. 1.

OHCs have evolved only in mammals. They improve the hearing sensitivity compared to other classes of vertebrates, which goes up to about 11 kHz (in some birds). However, their main function is to extend the hearing range above 11 kHz, until about 200 kHz (maximum in some marine mammals). They have also improved frequency selectivity of the TM-OHC-BM system. This allows better frequency discrimination.

Adaptation to sound level

With the Mossbauer technique measuring the Doppler shift in gamma radiation from a small radioactive source placed directly on BM it was showed that the amplitude of BM vibrations increased nonlinearly with increasing stimulus level. For example, Fig. 5 shows the level dependence of the ratio of malleus vibration to BM vibration. If the mechanics of the cochlea were linear, this ratio would be independent of the stimulus level. However, the response ratio near f_{best}, is largest when the signal level is smallest. This type of input/output relationship is referred to as a compressive nonlinearity. When the animal dies, or when OHCs are damaged, the magnitude of BM responses at low stimulus levels becomes smaller near f_{best}, and its growth with stimulus level becomes linear. The OHC electromotility is the source of the compressive nonlinearity of BM: the weaker the sound, the stronger the amplification of BM oscillations. This enables the matching of the enormous range of sound pressure levels, some 10^7 compressed to a 10^4 range of variation of BM oscillations. Further compression is provided by the steps in the processing sequence: BM oscillation →IHC bundle deflection →receptor potential → action potentials.

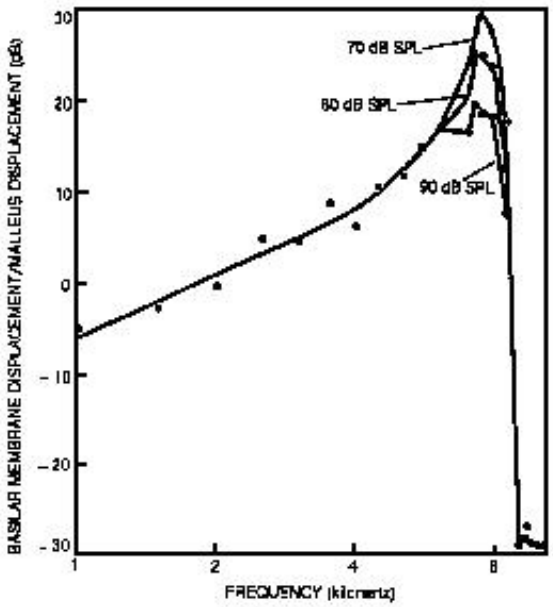


Fig. 5 Transfer characteristic of BM displacement/malleus (= tympanic membrane) displacement. Around f_{best} the ratio is strongly dependent on SPL, indicating the non-linear character of the motion transduction from TM to BM. Symbols present measurements and solid lines model fits. (From ref. 2.)

Otoacoustic emissions

Otoacoustic emissions are based at the electromotil activity of a restricted number of OHCs producing a too high cochlear amplification. Low-level sounds of cochlear origin can be recorded from the external auditory canal either spontaneously or evoked by sound stimuli. Emissions can also be detected when electric current is applied to the cochlea. It may possibly be caused by a reduced inhibitory efferent stimulation.

Efferent control

The mechanical feedback loop provided by the OHC must be finely regulated to guarantee optimal functioning of the cochlear amplifier. To cope with simultaneously occurring transient (sudden) and tonic (sustained) stimuli the regulatory mechanisms should at least control the operating point and gain of the OHC electromotility through a direct action of nerve efferents. Efferent stimulation suppresses afferent sound-elicited activity of the auditory nerve by the release of ACh on two time scales: a rapid action (tens of milliseconds) is responsible for modulating nerve responses to transient acoustic stimulation, whereas a slower action (tens of seconds) is thought to protect the ear from acoustic overstimulation.

References

1. http://www.boystownhospital.org/Research/Areas/Neurobiological/MoreInfoComLab/cochmech.asp (Research page Boys Town National Research Hospital)
2. Mammano F. and Nobili R., http://147.162.36.50/cochlea/index.htm and http://147.162.36.50/cochlea/cochleapages/theory/index.htm
3. Nobili R, Mammano F and Ashmore JF. How well do we understand the cochlea? *Trends in Neurosciences* 1998;21:159-167.

Sonotubometry

Principle

In sonotubometry, the acoustical features of the transmission from nose to outer ear are evaluated over time in order to evaluate the function of the Eustachian tube. Generally, in healthy subjects the tube is closed and opens when a swallowing movement, a Vasalva maneuver or an other pressure-equalizing maneuver is made. Dysfunction is, in gradations, in two directions: either always open or always closed. To test the tube function, an acoustical signal is emitted into the nose through a small loudspeaker and the system response is recorded in the ear canal by a miniature microphone. The signal strength measured during the swallowing movement is a measure for the tube opening.

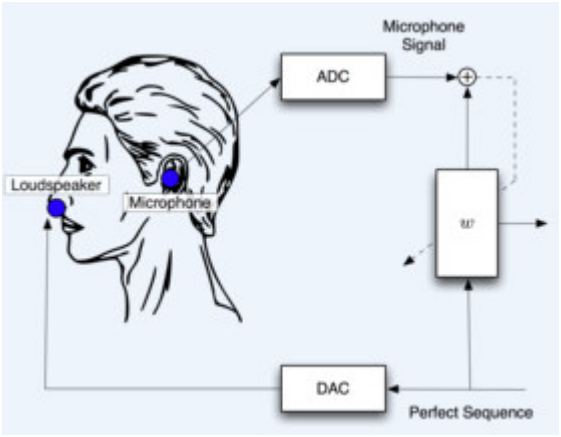


Fig. 1 Principle of sonotubometry

Application

Despite the fact that this test principle in theory is ideal and elegant, technical improvements and more research are required before sonotubometry can become a standard method in the ENT clinic. Recent developments for children seems promising

More Info

The results were very unreliable if the transmission link was excited with only a single frequency. This is because the resonant frequencies of the acoustical system nose-ear vary considerably between subjects. Another problem is that the signal received in the ear canal is confounded by the signal going outside the head from loudspeaker to microphone. Another one is produced by tissue propagation. Although the emitted sound hardly penetrates the head tissues (some 60 dB loss), the impedance of this transmission channel is not very high since there is a massive volume conduction compared to the very small air-transmission channel of the tube (some square mm in cross section). A compound signal (various frequencies) or a signal comprising noise band is applied to conquer this problems, followed by selective filtering. Further improvements are cross correlating the emitted and received signal and cross correlating the swallow movement with the microphone signal.

References

1. Lildholdt T, Brask T, Hvidegaard T. Interpretation of sonotubometry. A critical view of the acoustic measurement of the opening of the eustachian tube. Acta Otolaryngol (Stockh) 1984; 98: 250-254.
2. Avoort van der SJC, van Heerbeek N, Zielhuis GA, Cremers CWRJ. Sonotubometry: Eustachian Tube Ventilatory Function Test: A State-of-the-Art Review. Otology & Neurotology, 2005;6: ;38-543.
3. http://www.ind.rwth-aachen.de/en/research/medical-applications/acoustic-tube-endoscopy/sonotubometry/

Stapedius reflex

Basic principle

In Tympanometry, an acoustic reflex, or contraction of the stapedial and tensor tympani muscles, occurs under normal conditions when a loud acoustic stimulus (e.g. loud speech) is presented to the auditory system. The stapedius pulls the stapes (stirrup) of the middle ear away from the oval window and the tensor tympani muscle pulls the malleus (hammer) away from eardrum. These contractions cause a *stiffening of the ossicular chain*, which decreases the *compliance* (volume change over pressure) of the middle ear system. As in tympanometry, a probe tone is used. It detects the change of the reflected beam of the tone probe, which gives the change of compliance. Regardless of whether the acoustic stimulus is active in the left, in the right, or in both ears, the Stapedius reflex is always binaural, i.e. it occurs in both ears at the same time. The examination can be performed ipsilateral (probe and sensor at the same ear) or contralateral (different ears; Fig. 1).

Fig. 1 Ipsilateral and contralateral measurement of the Stapedius reflex.

The reflex measurement is expressed as the air pressure value where the compliance peak occurred during the tympanometric test. Stimulus tones of varying intensities at 500, 1000, 2000 or 4000 Hz are presented as short bursts. If the applied stimulus causes a decrease in compliance in the “probe ear”, a Stapedius reflex occurred. If a change greater than say 0.05 mL in the pressure/volume diagram (see Fig. 2 of Tympanometry) is detected, a reflex is considered to be present for the applied test level and frequency. Because this means an extremely small compliance change, any movement of the probe during the test may produce an artifact (false response). The smallest test level (in dBHL, dB hearing level) evoking the reflex is the reflex threshold. The Stapedius reflex is activated in normal-hearing adults with sound pressure levels between 70 and 105 dBHL.

Application

The reflex is measured when tympanometry is also done. The acoustic reflex is affected by the middle ear status as well as the amount of hearing loss.
The middle ear pressure should be equivalent to the ambient air pressure (0 daPa (decaPa) difference). Minor negative shifts of the peak compliance may occur when the patient is congested. Shifts are rarely to the positive side. A more negative pressure than -150 daPa generally indicates further examination. A perforation in the tympanic membrane will cause a seemingly high ear canal volume measurement because the volume of the entire middle ear space is added to the canal volume.
An extremely soft tympanic membrane or an ossicular chain discontinuity will yield a very high peak compliance in the presence of a normal middle ear pressure. Ear canal volume will be normal and the reflex may be absent.
A fixation of the ossicular chain, as in otosclerosis, will produce a tympanogram with very low compliance in the presence of normal middle ear air pressure. A resolving case or beginning case may produce a reduced peak in the presence of severe negative middle ear pressures. The ear canal volume is normal and the reflex is either absent or at an elevated level.
Eustachian tube dysfunction in the absence of fluid will show a normal compliance curve, but it will be displayed to the negative side of the tympanogram. Ear canal volume will be normal and the reflex may be present, depending on the degree of involvement.

Reference

1. Smyth SB. www.maico-diagnostics.com/eprise/ main/Maico/Products/Files/MI24/Tymp.Guide.pdf

Tympanometry

Basic principle

Tympanometry is a measurement of the Acoustic impedance of the middle ear. If a sound impinges the eardrum, part of the sound is absorbed and sent via the middle ear to the inner ear while the other part is reflected (see Utrasound). The stiffer the eardrum, the less sound reaches the inner ear. With fluid in the middle ear, the drum behaves very stiff (similarly as the elastic rebound of a tennis ball against a stone wall).

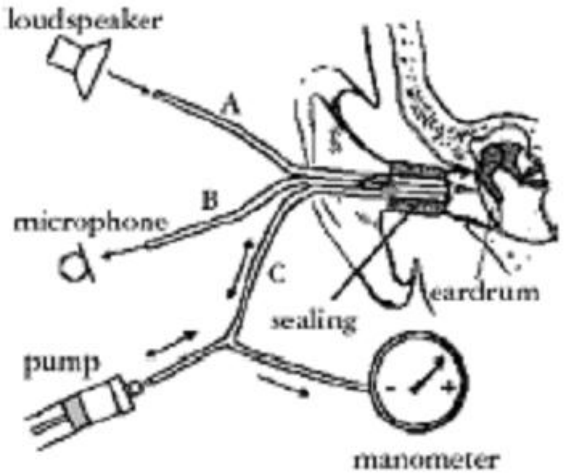


Fig. 1 Principle of the impedance measurement.

Inside the probe of the tympanometer are three small tubes. One contains a small loudspeaker, which emits a low frequency sound (Fig. 1: A). Another tube (B) is connected to a microphone. The third tube (C) contains a manometer and a pump, which produce positive and negative pressure. The probe, covered by a soft tip, is inserted airtight nearby the eardrum. Pressure is rapidly swept from positive to negative (50-200 daPa/s (deca-Pascals)). The highest compliance (the lowest impedance) is normally reached with an inner ear air pressure corresponding to the outside pressure. When performing tympanometry, the pressure in the canal varies continuously from +400 daPa to –200 daPa. Impedance is lowest (maximal compliance) when pressure in the canal equals pressure within the middle ear. Fig. 2 shows a tympanogram. Maximal compliance is at the peak and measured as an equivalent volume of air (in mL). The sharpness of the peak is expressed as its width at half peak amplitude and measured in daPa. The highest compliance is reached with normal pressure. With positive and negative pressure, the eardrum stiffens and the compliance decreases. The position of the peak on the horizontal axis and on the vertical axis provides diagnostic information regarding the function of the middle ear system.

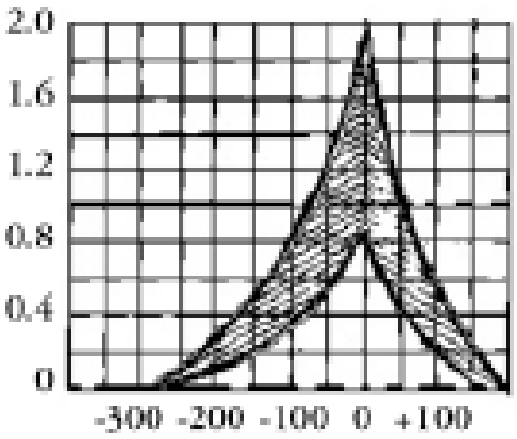


Fig. 2 Norm-values of the tympanogram curve indicated by the hatched area. Along the horizontal axis is pressure (daPa) and along the vertical axis the equivalent volume (mL).

Application

Tympanometry is the most utilized clinical test in audiology (see Hearing and audiometry). The test requires a clean ear canal. There are three types of tympanograms:
Type A. The peak compliance occurs at or near atmospheric pressure indicating normal pressure within the middle ear.
Type B. No sharp peak, little or no variation in impedance over a wide sweep range, usually secondary to (non-compressible) fluid within the middle ear space (otitis media, middle ear squeeze, or tympanic membrane perforation, the latter two often due to diving).
Type C. Peak compliance is significantly below zero, indicating negative pressure (sub-atmospheric) within the middle ear space. This finding is suggestive of Eustachian tube dysfunction or middle ear fluid.
The information derived from the tympanogram provides the physician with additional information regarding the patient's middle ear function (to document or rule out the presence of otitis media, tympanic membrane perforation or Eustachian tube dysfunction).
A compliance peak within the range 0.2-2.0 mL (children and adults) indicates normal mobility of the middle ear system.

More Info

More correctly, the compliance (the reciprocal of elastance, see Lung gas transport 2, pressure, volume and flow) is the slope of the tympanogram, e.g. measured in mL/daPa, which is maximal near the peak (see Fig. 2). However, in audiological practice, for convenience the difference referred to the healthy tympanogram is expressed as height of the peak in mL.

Reference

Smyth SB. www.maico-diagnostics.com/eprise/ main/Maico/Products/Files/MI24/Tymp.Guide.pdf

Vestibular mechanics

Basic Principle

The vestibular system, or balance system, is the sensory system that provides the dominant input about our locomotion, head movements and orientation in space. It is situated in the vestibulum of the inner ear (Fig. 1). As our movements consist of rotations and translations of the head or body, the vestibular system comprises two subsystems:

- the semicircular canals, which detects rotational accelerations to perceive rotational movements;
- the otoliths, which are sensitive for linear accelerations in order to perceive changes of linear motion.

The vestibular system sends signals to the neural structures that control the eye movements, and to the muscles that keep us upright. The projections to the former provide the anatomical basis of the vestibulo-ocular reflex (VOR), which is required for clear vision; and the projections to the muscles that control our posture are necessary to keep us upright.

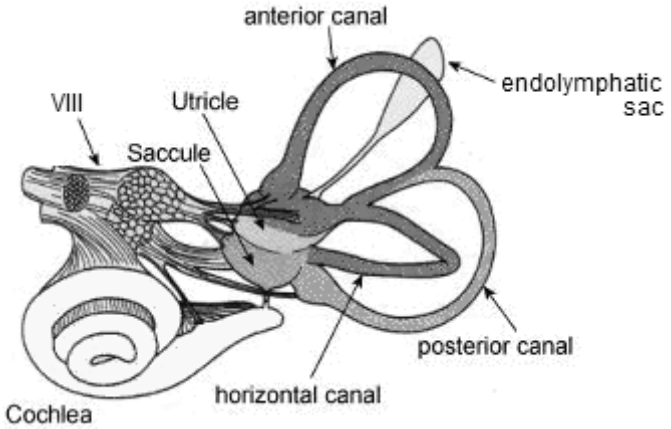


Fig. 1 Human labyrinth, from the left ear. VIII Eighth cranial nerve with auditory and vestibular branches.

Semicircular canals

Since we perceive the environment within 3D fashion, accordingly, our vestibular system contains three semicircular canals in each labyrinth. They are approximately orthogonal to each other, and are called horizontal (or lateral), anterior (or superior), and posterior (or inferior) canal.

The canals are arranged in such a way that each canal on the left side has an almost parallel counterpart on the right side. Each of these three pairs works in a push-pull fashion: when one canal is stimulated, its corresponding partner on the other side is inhibited, and vice versa.

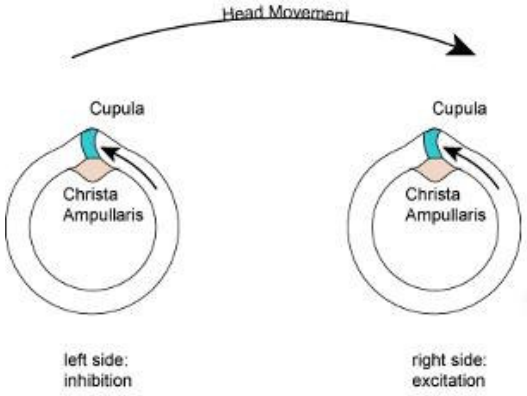


Fig. 2: Push-pull system of the semicircular canals, for a horizontal head movement to the right.

This push-pull system allows us to sense all directions of rotation: while the right horizontal canal gets stimulated during head rotations to the right (Fig 2), the left horizontal canal gets stimulated (and thus predominantly signals) by head rotations to the left.

Otolith organs

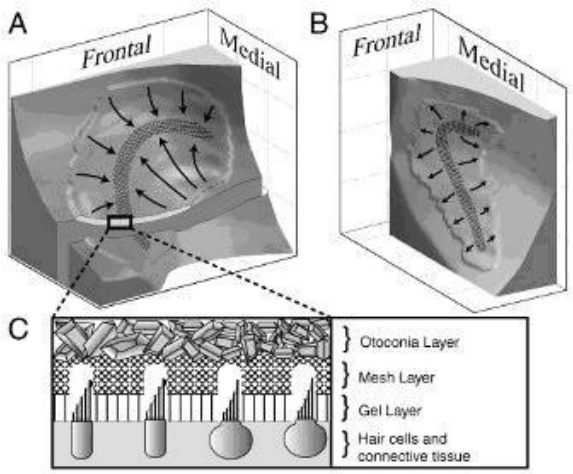


Fig. 3 Otoliths, left side. A) the utricle, and B) the saccule. C) Cross-section through the utricle: the Mesh Layer is fairly stiff, while the underlying Gel Layer is more viscous. When the Hair cells are bent in the directions indicated by the arrows in A) and B) they get excited, while a deflection in the opposite direction inhibits them. The upper part of the otoconia layer is embedded in the endolymph

While the semicircular canals respond to rotations, the otoliths sense linear accelerations. We have two on each side, one called utricle, and the other saccule. Fig. 3C shows a cross section through an otolith system. The otoconia crystals in the otoconia layer (Fig. 3C, top layer) rest on a viscous gel layer, and have a higher specific mass than their surroundings. Therefore, during linear acceleration of the head, they are displaced with respect of the hair cells, resulting in deflected of the hairs of the hair cells (Fig. 3C, bottom layer) and thus produce a sensory signal. Most of the utricular signals elicit eye movements, while the majority of the saccular signals projects to muscles that control our posture. Since gravity is equivalent to a constant linear acceleration, the brain has to correct the otolith signals for the constant action of gravity. This occurs quite well, but the neural mechanisms underlying this correction are not yet fully understood.

For a more extensive basic explanation see ref. 1 and for the more specialized physiology ref. 3.

Application

Vestibulo-ocular reflex (VOR)
The vestibular system needs to be fast since for clear vision the head movements need to be compensated almost immediately. Otherwise, our vision corresponds to a photograph taken with a moving hand. To achieve clear vision, signals from the semicircular canals are sent as directly as possible to the eye muscles. This direct neural pathway, a reflex arc, involves the sensory neurons, two higher order non-cortical neurons and the motor neuron, innervating the eye muscles (Fig. 4). Using this direct (cortex not involved) connection, eye movements lag the head movements by less than 10 ms, one of the fastest reflexes in the human body. The automatic generation of eye movements from movements of the head is the earlier mentioned VOR.
The VOR, combined with the push-pull principle described above, forms the physiological basis of the Rapid head impulse test or Halmagyi-Curthoys-test. When the function of the right balance system reduced by a disease or by an accident, quick head movements to the right cannot be sensed properly any more. Therefore, no compensatory eye movements are generated, and the patient cannot fixate a point in space during this rapid head movement.

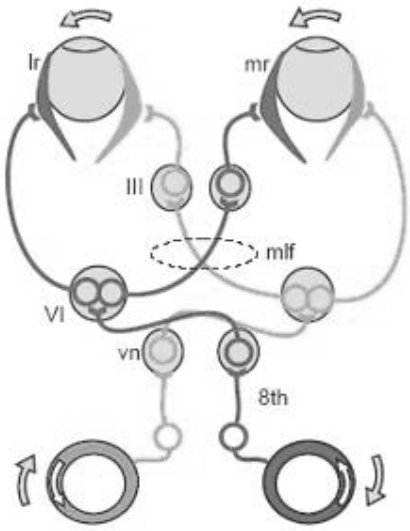


Fig. 4 Three-neuron arc, during a head movement to the right. 8th facial nerve, from the peripheral vestibular sensors to vn, the vestibular nuclei in the brainstem. VI abducens nucleus. The medial lateral fascicle (mlf) projects from the abducens nucleus to III, the oculomotor nucleus. The left lateral rectus muscle lr and the right medial rectus muscle mr get contracted, turning the eyes to the left. The dark gray objects are excited, the light gray ones inhibited. (From SensesWeb, Vilis, an educational website concerning al sensory systems).

Pathologies

Diseases of the vestibular system can take different forms, and usually induce vertigo and instability, often accompanied by nausea. In addition, the function of the vestibular system can be affected by tumors of the cochleo-vestibular nerve (vestibular schwannoma), an infarct in the brain stem or in cortical regions related to the processing of vestibular signals, and cerebellar atrophy. Less severe, is vertigo caused by the intake of large amounts of alcohol.

Benign Paroxysmal Positional Vertigo (BPPV) is probably caused by pieces that have broken off from the otoliths, and have slipped into one of the semicircular canals. In most cases, the posterior canal is affected. In certain head positions, these particles push on the cupula of the canal affected, which leads to dizziness and vertigo. This problem occurs rather frequently, often after hits to the head or after long bed rest. BPPV manifest itself by vertigo attacks, which repeatable appear when the head is brought into a specific orientation. In most cases, BPPV can be eliminated (for the patient in an almost miraculous way) by lying down, bringing the head in the right orientation, and sitting up quickly.

See for a more extensive description of clinical applications e.g. Brandt (2003).

More Info

Mechanics of the semicircular canals

The mechanics of the semicircular canals can be described by the Navier-Stokes equations. By applying its solution (Grant, 1995), the transfer function (see Linear first order system) of the semicircular canals can be found.

If we designate the deflection of the cupula with θ, and the head rotational velocity with ω, the cupula deflection in Laplace notation is approximately:

$$\frac{\theta}{\omega}(s) = \frac{\alpha s}{(s+1/\tau_1)\cdot(s+1/\tau_2)} \qquad (1)$$

with α a proportionality factor, and s corresponding to the frequency (s=iω). It appears to be a band-pass filter with a first order high frequency cut off and a first order low frequency cut off. For humans, the time constants τ_1 and τ_2 are approximately 3 ms and 21 s, respectively, or in the frequency domain 41 Hz (a rotation with 93^0/ms) and 0.0077 Hz (17^0/s).

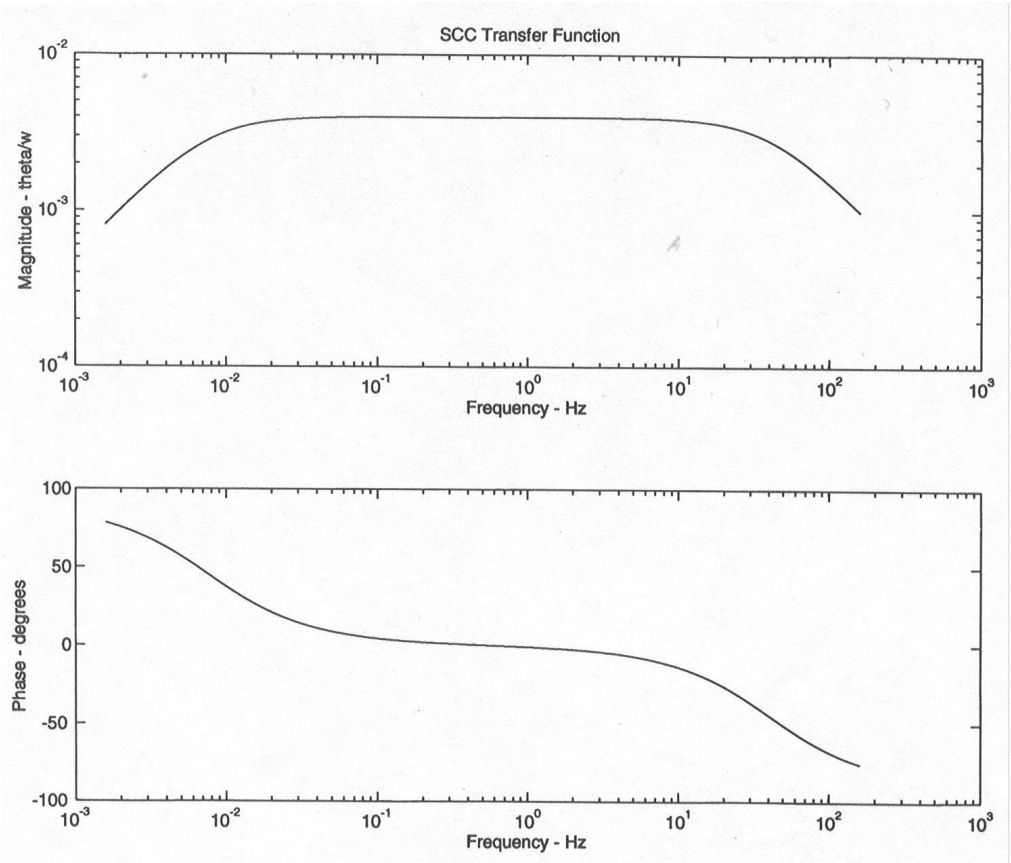


Fig. 5 Frequency response of semicircular canals for the transfer function of mean angular displacement of endolymph fluid θ as a function of angular head velocity ω (from ref. 2).

Hence, in the frequency domain it has a span of some 3.5 decades. As a result, for typical rotational head movements, which cover the frequency range of 0.1 Hz and 10 Hz, the deflection of the cupula is approximately proportional to the head-rotational velocity. This is very useful, since the velocity of the eyes must be opposite to the velocity of the head in order to have clear vision.

Mechanics of the otoliths

When the acceleration of the otoliths is plotted versus frequency, than they behave as an overdamped high-pass second order system. Using the acceleration and not velocity or displacement is obvious, since the otolith system behaves as an accelerometer. This means that any acceleration with a high frequency content (see Fourier analysis) is well perceived whereas very slow accelerations are not perceived, as can easily be confirmed from daily live experience.
The mechanics is dominated by the elasticity and viscosity of the gel (and mesh) layer. For the derivation of the transfer function, again the Navier-Stokes equations are basic, and in addition the motion equations are of importance. The derivations, which are more complicated than for the semicircular canals can be found in Grant (1995). For a moderate elasticity and a rather high viscosity ratio (gel over endolymph), the lower cut off frequency is at ca. 3 Hz and the upper one 3000 Hz. Higher elasticity results in less damping (with closer cut off frequencies) and the same holds for a lower viscosity ratio.

Literature

Dickman D. Vestibular Primer *http://vestibular.wustl.edu/vestibular.html*.
Grant W., Vestibular mechanics. In: The biomedical engineering handbook, Bronzino (ed), CRC Press, Boca Raton, pp 517-527, 1995.
Highstein S.M., Fay R.R. and Popper A.N. (eds), The vestibular system, Springer-Verlag, Berlin, 2004.
Brandt, T. Vertigo : its multisensory syndromes, Springer-Verlag, Berlin, 2003.
Vilis T. SensesWeb *http://www.med.uwo.ca/physiology/courses/sensesweb/*.

Electricity and Bioelectricity

Bioelectricity

Principle

Bioelectricity refers to the electrical and/or magnetic fields produced by living material (from cells to organisms). Bioelectricity should better be equated with bioelectromagnetism but this is easily confused with bioelectromagnetics, which deals with the effect on living matter from external electromagnetic fields, such as the diverse types of radiobiology, transcranial magnetic stimulation, but also animal migration and navigation by means of electroreception or magnetoreception. Some animals (sharks, rays, catfishes etc.) have bioelectric sensors, the sensory cells of their electroreceptive system, with sensitivities down to less than 1 µV/m (skates). Other animals, such as migratory birds, are believed to navigate in part by orienting with respect to the earth magnetic field.
Bioelectric phenomena are for example the cell membrane potential and the electric currents that flow in nerve axons and muscle fibers (as a result of action potentials) and dendritic currents. In addition to aims of communication (signal one another) biological cells use bioelectricity to store metabolic energy, to do work or trigger internal or external changes or events (e.g. the electrical field due to the electric discharge of an electric ray or eel).
Bioelectromagnetism is an aspect of all living things, including all plants and animals. It is studied predominantly through the techniques of electrophysiology in animals. It is based at the propagation of electrical activity along nerve cell or muscle cell membranes. The resulting evoked electrical fields can often been recorded non-invasively when enough cells are active (see Electroencephalography).
More recently, also physiological measurements of magnetic fields evoked by the electric activity of the brain and the heart can be measured (see Magnetoencephalography).

Many living cells of animals and micro-organisms have a potential difference over their cell membrane. For nerve cells including receptor cells and muscle cells this is basic for their performance. The membrane potential is maintained by the cell metabolism. The following holds for (most) nerve cells and muscle fibers.
The nerve potential is a consequence of the semi-permeability of the cell membrane, which is ca. 10 nm thick with a voltage of 0.1 V across it, yielding a field strength of 10 kV/mm, 10 times the break down voltage in air. The measured capacity of nerve membranes is large due to their thinness: ≈1 µF/cm^2, in accordance with the equation of the size of a capacitor, $C_m = \varepsilon_m \varepsilon_0 S/d$
where ε_m the relative dielectric constant (ca. 5), ε_0 the dielectric constant in vacuum, S the surface and d the thickness of the membrane. Nerve fiber diameters range from 0.1µm (unmyelinated sensory nerves) to I mm (giant unmyelinated axon of squid). Axon length ranges from mm to meters. Due to the squid giant axon's large size intracellular recordings can be made simultaneously at several places.
The charge of the membrane is caused by an unbalance in the ion concentrations between the cell interior and exterior, and as is later explained by the conductivities of these ions. For mammals, the concentrations of the ions are given in Table 1.

Table 1 Approximate concentrations of relevant small ion of mammalian nerve cells

ion	Intra-cellular mM	Extra-cellular mM	Plasma mM	Equilibrium potential E mV
Na^+	12.5	145	141	+60
K^+	135	4	4.2	− 90
Cl^-	9	115	104	− 80

Table 1 (which is slightly different for muscle cells and heart muscle cells) does not give the concentration of the large organic anions for which the membrane is impermeable. Their negative charge is compensated by especially K^+ for which, in rest, the membrane is well permeable. (In rest means that there is no disturbance of the membrane potential.) A disturbance can give rise to an action potential or dendritic potential which is propagated along the dendritic membrane to the cell body. Then the cell is active. The cell interior is electrically neutral. The charge of C_m at a potential difference of 0.1 V is due to only a fraction (10^{-3} - 10^{-5}) of the number of ions present in the cell's interior. This means that a nerve can generate many action potentials before the concentrations of the relevant ions change significantly.
For each ion, given the inside and outside concentrations, there exists an equilibrium potential according to the thermodynamical law of Nernst, which is based on Boltzmann's distribution:

$$\frac{n_1}{n_2} = \frac{V_1}{V_2} e^{\frac{-zFE}{kT}} \quad (1)$$

where n_1 and n_2 are the number of ions in the two volumes, V_1 and V_2, E the equilibrium potential, k the Boltzmann constant, T the absolute temperature, z the valence of the ion and F Faraday's constant. Rewriting (1) for K^+ yields:

$$E_{K+} = (RT/zF) \ln([K^+_{out}]/[K^+_{in}]) \quad (2)$$

where R the gas constant (see Gas laws). For the various ions this yields E as given in Table 1.
In reality, E is –80 mV meaning that the equilibrium for K^+ and Cl^- is nearly reached. In the following the influence of all other permeable anions (e.g. HCO_3^-) are taken together with Cl^-.
For E_{Na} which is strongly positive, so reversed, there is no equilibrium at all. From this follows that the concentration gradient of Na^+ as well as the actual E_m, mainly determined by the K^+ ions, will give an inflow of Na^+.
The ions pass the membrane via specific ionic pores or gates. Their passage is regulated by the membrane potential itself (see Bioelectricity: action potential). Although in rest the permeability for Na^+ is small, on the long run the nerve would lose its potential if the slowly intruding Na+ would not be ejected by an active process, the so-called sodium Na^+ pump. If this stops working, finally, due to osmotic laws (see Osmosis) dictating equal osmolarity at the two sides of the membrane, the cell will then swell and eventually burst.

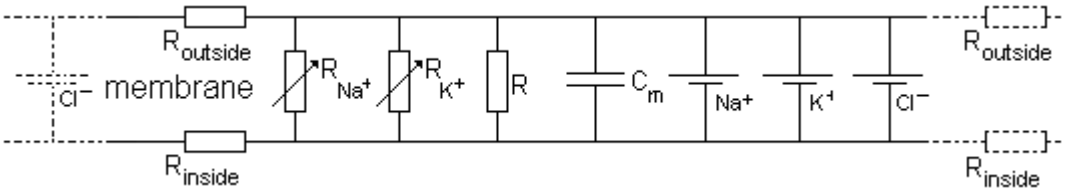


Fig. 1 Basic electric model of the nerve membrane. $R_{outside}$ is small and therefore often ignored. The RC-network drawn with solid lines is thought to be repeated many times to model a fiber.

Since the ionogenic sources (batteries) have a negligible resistance, the membrane conductance ($1/R_m$) equals:

$$1/R_m = 1/R_{Na} + 1/R_K + 1/R_{Cl}, \quad (3)$$

with the variable R_{Na+} and R_{K+} indicating the voltage controlled passage of these ions. They are constant in rest and variable during action. Sometimes the voltage source (the ionogenic batteries) is placed in series with their resistors, but this is less realistic since the sources are located at different sites in the membrane than the ionic gates. All other ions do practically not contribute to the membrane potential. Taking the action of the three ions together, E_m in rest can be found:

$$E_{m,rest} = (RT/F) \ln((g_K K^+_{out}] + g_{Na} Na^+_{out}] + g_{Cl}[Cl^-_{out}])/ (g_K K^+_{in}] + g_{Na} Na^+_{in}] + g_{Cl}[Cl^-_{in}])), \quad (4)$$

where the conductance g = 1/R. Obviously, the effect of the sources is weighted by their conductances. This is the reason why $E_{m,rest}$ is dominated by the K^+ equilibrium potential.

Application

Bioelectricity can be divided in:
Electrophysiology (see also Electrophysiology: general)

- Single cell physiology *in situ* and *in vitro* with isolated organs (e.g. retina), with brain slices or cell cultures (with phenomena such as membrane potential, resting potential, action potential, clamp potential, clamp current, excitatory postsynaptic potential, inhibitory postsynaptic potential, dendritic potential).
- Multi-unit recordings.
- Electrophysiology of small pieces of neuronal tissue *in situ* and *in vitro* such as brain slices (brain waves, field potentials, compound action potentials).
- Electro(magnetic)physiology of whole or great parts of an organ, used diagnostically *in situ*, for instance Electroencephalography, Electromyography, Electro-ocology, Electrocardiography (see ECG: basic electrocardiography), electroneurography (for recording nerve activity, such as brain the auditory stem potentials, the SEP (somatosensory potential of e.g. the ulnar or tibial nerve). Magnetography can be performed with humans (and animals)

and of human fetuses and also organs of tissues. Magnetoencephalography and cardiomagnetography are most known.

Electroreception and magnetoreception.

Predatory electric field generation.

Therapeutical equipement, for instance pacemakers for the heart, bladder, anus, heart defibrillators, generators for voluntary muscles, for pain release (spinal cord), cochlear and cortical implants (e.g. visual or auditory) and magnetic induction coils in neurology are formally bioelectromagnetics devices.

More Info

The cell membrane can be considered as a capacitor with the membrane as the 'non'-conductive medium between the two conduction 'plates', the outside and inside of the cell. The membrane charge is however so small, that the cell interior can be considered as electric neutral.

Bioelectricity: action potential

Principle

As a type of a propagating disturbance of the nerve (or muscle) membrane potential the electrotonic propagation has been described in Bioelectricity: electrotonic propagation. It happens in dendrites, axons and muscle fibers. These structures can be some mm long, but axons and muscle fibers are mostly some cm and axons may reach many meters (as in whales). However, mostly an electrotonic potential is strongly reduced along distances of centimeters, is propagating slow and, due to temporal filtering, the peak time at the nerve terminal will not be very well defined. Therefore it is not very appropriate to transmit information over long distances in living organisms. There exist a better way of propagating nerve information. This is via action potentials (spikes), an active way of propagation. With spikes information transport is faster and with higher temporal resolution. As a consequence, more information can be transmitted. Spikes arise from the somatic potential, the sum of the dendritic potentials, at the axon hillock between the axon and soma. Then they propagate along axons (sometimes certain types of dendrites), muscle fibers and also heart muscle fibers. Fig. 1 visualizes this propagation.

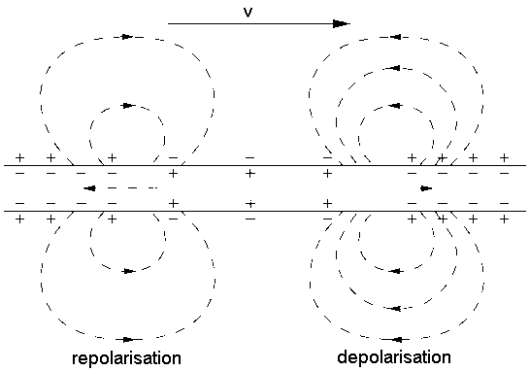


Fig. 1 Principle of propagation.

Fig. 2 depicts the 3 main types of action potentials. They are found in vertebrates and invertebrates, but also some plants (relying on K^+ and Ca^{++}, with the phloem as channels).

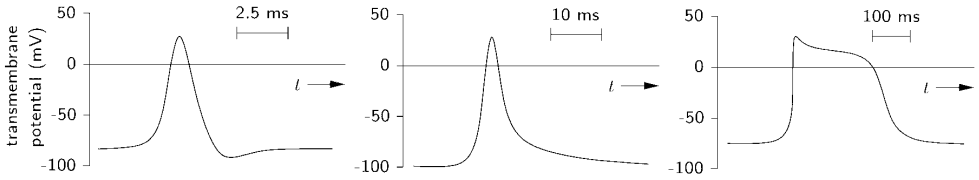


Fig. 2 From left to right action potential of axon, muscle fiber and heart muscle cell.

Depolarization and repolarization

Below the axonal spike, the 'common' one is described, first for an unmyelinated axon. Fig. 3 gives the various phases which can be distinguished during its time coarse. Often one dendritic potential can give rise to a couple of spikes, depending on its amplitude. A propagating spike generally maintains its waveform and amplitude. This is caused by the fact that the membrane conductance g_m (= $1/r_m$) of the axon is not constant. An excellent way to investigate the changes of the conductances is the voltage

clamp technique (holding the membrane potential at a constant value whatever injected current is needed, see Electrophysiology: clamping techniques). Essential for this technique is that there flows no axial current through the axon. In a thick axon (squid), this is achieved by inserting a fine silver wire longitudinally in an axon. By an intracellular electrode and an extracellular electrode current is injected into the axon to compensate for changes in current through the membrane. The current needed in the clamp technique compensates the ionic currents (and the initial capacitive current) and is measured as a function of time. According to Ohm's law the total ionic current at any time is proportional with the total membrane conductance since the membrane potential is kept constant. The Na^+ and K^+ conductances can be measured separately by applying certain drugs which make either the Na^+ or the K^+ conductance zero. The Na^+ and K^+ have different types of pores, i.e. selective channels. The permeability for Na^+ and K^+ appears to be a function of the membrane potential. In rest, most Na^+ channels are closed, but the K^+ channels open, causing a constant leaking out of K^+. This is way the rest potential is mainly determined by K^+ with its 75 times higher conductivity. The outflow of K^+ is constantly compensated by Na^+ inflow.
If now the membrane is stimulated and either depolarized or hyperpolarized the membrane potential will change with time, resulting in time-and-voltage dependent ionic resistances. Now r_m is composed of a resistance r_{Na} formed by the Na^+ channels, a resistance r_K formed by the K^+ channels and r_L formed by other passive ionic channels, mainly for Cl^-, each of them connected with the Nernst equilibrium potential (see Bioelectricity). The channel mechanisms involved in the generation of a spike are the rapid increase in the permeability for Na^+ and the delayed and slower increase in the permeability for K^+ if the cell is depolarized (so-called cathodal stimulation). Na^+ moves in under the influence of the *driving force*, the difference between the membrane potential and the equilibrium potential of Na^+. In the squid axon this occurs in a few ms. As soon as the Na^+ permeability (or conductance) increases, more Na^+ streams into the axon or soma, diminishing the membrane potential still further. This causes a further increase of the Na^+ permeability and at a certain critical value, the firing threshold (about 15 mV more positive than E_{rest}) this becomes so strong a positive feedback that the cell even reverses its potential to positive values in the direction of the Na^+ equilibrium potential. Actually, the threshold is reached when the inward Na^+ current exceeds the outward K^+ current. Very shortly afterwards the Na^+ permeability returns to its original value and the K^+ permeability increases temporarily. As with the Na^+ influx, it is not the movement of K^+ that changes E. It is the value for g_K rising above that for g_{Na}, dragging E back towards the equilibrium constant for K^+. Since the voltage-gated K^+ channels have a delayed response, such that K^+ continues to flow out of the cell even after the membrane has fully repolarized. This causes the undershoot (short hyperpolarization).
There is a common misconception that the Na^+/K^+ pump restores the resting potential during the spike falling phase by actively pumping Na^+ out and K^+ into the neuron. This (along with the misconception that sodium 'floods' the cell to cause the spike), is not correct. The Na^+/K^+ ATPase (the pump) does ultimately maintain the resting potential by maintaining the concentration gradients for Na^+ and K^+, but does so on a much slower time scale.

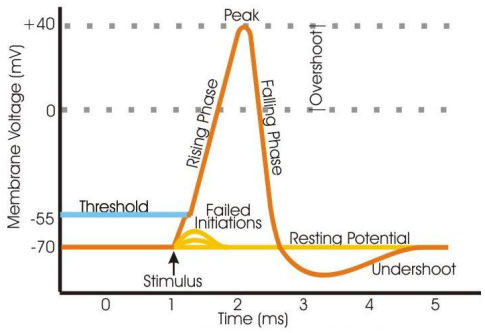


Fig. 3 Basic time coarse of action potential. (Fom ref. 4.)

Refractory period
During a short period after the occurrence of a spike the cell cannot be stimulated. This is the refractory (1-5 ms) period consisting of an absolute and relative phase. In the former, the Na^+ channels cannot be opened by a stimulus irrespective of applied voltage. In the subsequent relative phase, spikes can be initiated, since Na^+ channels are reactivated (in a stochastic manner) but the threshold is greater. This is caused by the slightly hyperpolarized state due to still higher than resting value for g_K, so more voltage is required to reach threshold, and also the threshold itself is higher than usual because some of the Na^+ channels will still be inactivated. (Note that Na^+ channel has at least three states: closed, open and inactivated - closed and not able to open). The refractory period is important because it ensures unidirectional (one way) propagation of the spike.
The basic theory of spike propagation, the Hodgkin-Huxley (HH) theory, is described in More Info.

Application

Spikes, mostly in the form of “spike trains”, are used most extensively by the nervous system for communication between neurons and for transmitting information from neurons to other body tissues such as muscles and glands (neurohypophysis).
Spikes are measured with the recording techniques of electrophysiology and more recently with neurochips containing EOSFETs (electrolyte-oxide-semiconductor field effect transistor). Such chips are applied in retinal and cortical implants to record and stimulate neuronal activity. (A cochlear implant is formally not a neurochip since it is only used for stimulation; it is a neuroprosthesis). An oscilloscope showing the membrane potential recording from a single point on an axon shows each stage of the spike as the wave passes. A speaker is very useful to listen to the elicited spike (trains).

Spikes in general cannot be measured at distance, since, due to its dipole nature, it diminishes with the third power of distance. The electrotonic potential changes caused by synaptic transmission which, if strong enough, give rise to the spike, have a less strong decay with distance. They also last longer. If there is enough geometrical coordination between a group of excited neurons, so-called slow or graded potentials can be recorded for instance on the skull of man. They are always sign of mass action. If they are spontaneous we speak of the EEG, if they are excited by light, sound or peripheral nerve stimulation we speak of visual, auditory or somatosensory evoked potentials (EPs) respectively. Also the electroretinography (ERG) reflects graded potentials and not the spikes of the optic nerve.

Some diseases reduce the speed of spike conductance. The most well-known of these diseases is multiple sclerosis, in which the breakdown of myelin impairs coordinated movement (see **More Info**, Myelinated axon).

More Info

The conductances for Na+ and K+ change according to:

$$g_{Na+} = \breve{g}_{Na+} m^3 h,$$
$$g_{K+} = \breve{g}_{K+} n^4, \qquad (1)$$

where $\breve{g}_{Na+}$ and $\breve{g}_{K+}$ are the maximal conductances. The variables n, m, and h have a value between 0 and 1. In the equation for the K^+ conductance n^4 denotes the fraction of the K^+ channels which are open.

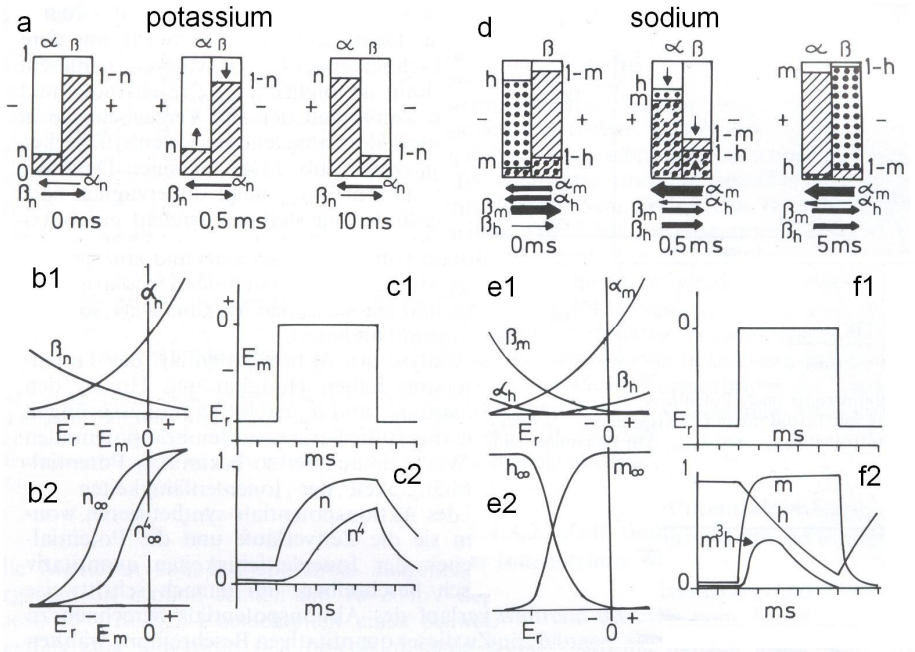


Fig. 4 The gating model for potassium and sodium. b2) and e2) depict the dependency of n and n^4, and m,h and hm^3, when a voltage step is applied very long (infinite).

If all channels are open then $n^4 \approx n \approx 1$. If all are closed n = 0. Apparently four events of equal nature have to coincide to open the four folded locked K^+ channel. For the Na^+ channel two kinds of keys (events) are used. Three identical keys are needed to open the three folded m lock. Another lock (h) is open in rest, but closes when the membrane is depolarized. These fractions m, h and n are voltage and consequently time dependent. They can be found by solving three experimentally found differential equations. For n this equation is:

$$dn/dt = \alpha_n (1-n) - \beta_n n \quad (2a)$$

where α_n denotes the open condition and β_n the closed condition. This is visualized in Fig. 4a. After solving, α_n appears to be a positive exponential function of the membrane potential E and β_n is a negative one (Fig. 4b1). Since during excitation E changes with time the two variables also change with time what finally results in the initially progressive increase of n^4, for a stepwise change of E (Fig. 4c1). Fig. 4c2 gives the final value of n and n^4 measured during voltage clamp.
For Na^+ we have to do with a change of m and h. Both can be calculated. The differential equations are:

$$dm/dt = \alpha_m (1-m) - \beta_m m \text{ and} \quad (2b)$$
$$dh/dt = \alpha_h (1-h) - \beta_h h. \quad (2c)$$

The time course of m looks like that of n but is faster. However, h behaves differently. It decreases instead of increases due to the decrease of β_h and increase of α_h when the membrane is depolarized (Fig. 4e2). Therefore, also for long lasting depolarization (voltage clamp) the Na^+ conductance restores to its originally low rest value within about 5 ms (Fig. 4e2). The opposite behavior of h and m during long lasting depolarization is clearly shown.
The currents which flow through the membrane are composed of the capacitive current i_c and three ionic currents of Na^+ (through the variable g_{Na}) of K^+ (through the variable g_K) and the anion current (mainly Cl^-, through the fixed g_l), together i_i. Fig. 5 gives the time coarse of a spike, together with i_c, i_i, there sum i_m, and also the underlying g_K and g_{Na}. For the 4 composing currents together the relation is (see equation (2) and (3) of Bioelectricity: electrotonic propagation):

$$i_m = (1/r_a)\partial^2 E/\partial x^2 = c_m \partial E/\partial t + n^4 \breve{g}_K(E-E_K) + m^3 h \breve{g}_{Na}(E-E_{Na}) + \breve{g}_L (E-E_L). \quad (3)$$

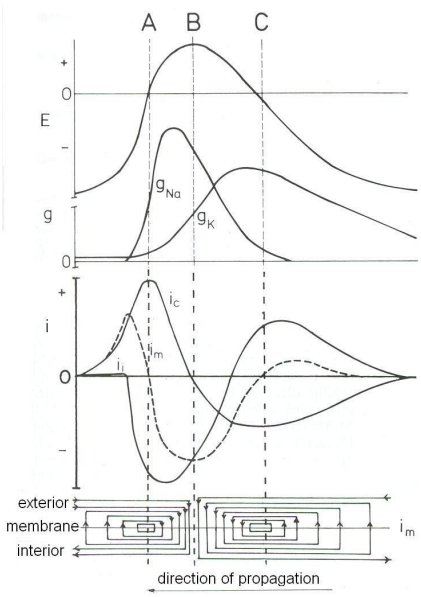


Fig. 5 Time coarse of conductances and currents during an action potential. At time A and C the slopes of the action potential are maximal and i_m zero. At B both reach their extrema.

Suppose that by summation of dendritic and somatic potentials at the axon hillock a spike arises. At that site the local currents become so strong that also the next part of the (axonal) membrane becomes enough depolarized to be excited and a propagated spike without decrement runs along the axon. The refractory period makes that the spike will not reverse and occur only once for a short lasting stimulus. If a nerve is stimulated in the middle of an axon the impulse will propagate to both sides. Longer lasting stimuli may cause trains of spikes. Just like for the propagation of a dendritic potential it can be shown that the conduction velocity Θ is a parameter of the equation:

$$\Theta^{-2} \cdot d^2E/dt^2 = c_m \, dE/dt + n^4 \breve{g}_K(E-E_K) + m^3 h \, \breve{g}_{Na}(E-E_{Na}) + \breve{g}_L(E-E_{L\text{-}}). \quad (4)$$

The numerical solution of this non-linear differential equation gives a quite complicated expression of Θ. Spikes recorded close to the soma (or axon hillock) are biphasic, as the one of Fig. 3, but when recorded in the vicinity of an axon they are triphasic.

One could think that a nerve impulse which reverses the nerve potential would bring about an important depletion of K^+ which leaves the cell because there is no potential gradient anymore keeping it in the

interior, and that the inflow of Na^+ would cause a permanent disturbance of the membrane potential. However, the amounts of ions displaced are small compared to the actual number present. Even in very small nerves several thousands of spikes can be generated without a significantly increased metabolism to expel Na^+.

The behavior of the channels has extensively been studied by clamping techniques (see Electrophysiology: clamping techniques), in which the i/E (so conductance) is influenced by administrating all kind of drugs. A discussion of these phenomena is beyond the scope of this chapter.

Myelinated axons

Propagation speed Θ can be increased by increasing the axon diameter. Taking for simplicity equation (8) of Bioelectricity: electrotonic propagation :

$$\Theta = 2\lambda/\tau = 0.5 C_m^{-1} (d/(R_m \cdot R_i))^{0.5} \qquad (5)$$

this speed is proportional with the square root of diameter (d).

However, for metabolic reasons, the diameter is limited (only the cold blooded squid reaches a value of 1 mm).

Unmyelinated fibers (about 2 µm) are generally found in the autonomic nervous system of vertebrates where speeds of about 1 m/s are sufficient.

In vertebrates, sensory and motor ones are generally myelinated. This is more effective to increase Θ. The effect of myelin can also be evaluated since all above considerations can be applied in principle to the myelinated nerve. Myelin can be considered as the dielectricum between two condenser plates. It decreases membrane capacitance, since myelin has a lower relative dielectric constant ε_m than interstitial fluid (with ε_m close to ε_m of water, being about 80; $C_m \sim \varepsilon_m/d$). Now, C_m is only about 4 nF/cm^2. and R_m is about 10^5 Ωcm^2. Increasing the effective membrane thickness by using myelin (leaving the inner fiber diameter constant) also decreases C_m. When myelin thickness and inner diameter increase with the same factor, then Θ increases about linear with this factor as follows from (5). This shows the efficiency of diameter increase. Experimentally this has been found indeed for vertebrate peripheral fibers.

However, a myelinated fiber longer than some 100 µm does not work properly. Myelin allows the rapid (essentially instantaneous) conduction of ions, but prevents the regeneration of spikes. Therefore, the cylindrical shape of the myelin sheath is interrupted every 0.01 – 0.1 mm by a node of Ranvier, a naked piece of ca. 0.5 µm of axon. Their C_m is about 4 F/cm^2 and R_m is only about 15 Ωcm^2. An abundance of voltage-gated Na^+ channels on these bare segments (up to 10^4 more than their density in unmyelinated axons) allows spikes to be efficiently regenerated at the nodes of Ranvier. The excitation jumps from one node to the other, which is a passive, so electrotonic transmission (see Bioelectricity: electrotonic propagation) implying some decrement. Basically, this can go in either directions, but the spike travels unidirectional because the node behind the propagating spike is *refractory*. This way of propagation of the spike is called saltatory conduction: at the myelinated segments the propagation is very fast (due to the insulation), whereas at the nodes there is a small delay of 0.01 to 0.1 ms. The length of the internodal segments are such that one, or sometimes even two nodes can be passed and that the amplitude is still sufficient to reach the threshold for restoring the amplitude of the spike. Thus, the safety factor of saltatory conduction is high, allowing transmission to bypass nodes in case of injury. Mammalian myelinated motor neurons can reach 100m/s with a restricted increase in diameter. Saltatory conduction increases nerve conduction velocity. Without saltatory conduction, the same conduction velocity would need large increases in axon diameter, resulting in organisms with nervous systems too large for their bodies.

Alternative models

A few observations are not easily reconciled with the model. A signal traveling along a neuron is accompanied by a slight local thickening of the membrane and a force acting outwards.

Also, a spike traveling along a neuron results in a slight increase in temperature followed by a decrease in temperature, whereas electrical charges traveling through a resistor always *produce* heat.

The recent soliton model explains the above observations and possibly all properties of the HH-model. A soliton is a self-reinforcing solitary wave (a wave packet or pulse) that maintains its shape while it travels at constant speed; solitons are caused by a cancelation of nonlinear and dispersive effects in the medium. This theory attempts to explain signals in neurons as pressure (or sound) solitons traveling along the membrane, accompanied by electrical field changes resulting from Piezoelectricity.

Literature

1. Noble, Physiol. Review, 46, 1-50, 1966. (HH-model.)
2.Heimburg T, Jackson AD. On soliton propagation in biomembranes and nerves. (Soliton model) PNAS **2005;**102:**9790-9795**
3. Heimburg T, Jackson AD. The thermodynamics of general anesthesia. Biophysical Journal, 92:3159 - 3165. (Soliton model)
4. Wikipedia Action potential.

Bioelectricity: chemical synapse

Principle

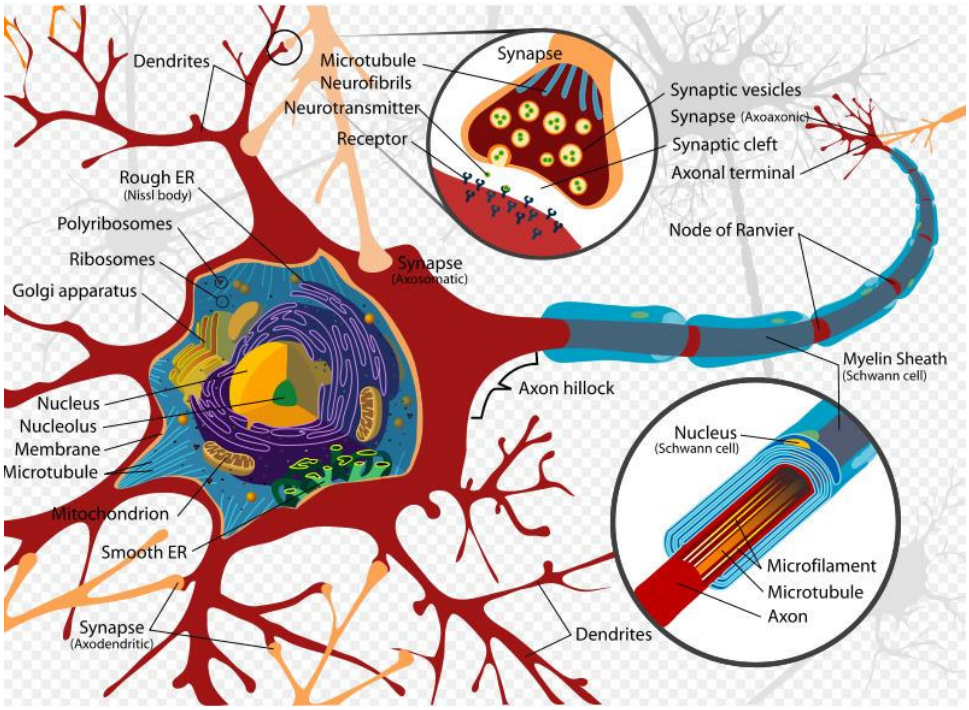


Fig. 1 Drawing of a typical neuron with the majority of the organelles indicated (from ref. 1).

Because an action potential (nerve impulses, also called spike) propagates only along contiguous membrane, another mechanism is necessary to transmit action potentials between cells. Neurons communicate with each other via synapses (Fig. 1).
Interneuron synapses are generally chemical synapses (Fig. 2). Other cell types, such as cardiac muscle cells, can communicate via electrical synapses.
The chemical synapse is a very small gap between neurons that allows one-way communication. As the presynaptic neuron undergoes an action potential, voltage-dependent Ca^{2+} selective channels open and cause finally the release of neurotransmitters into the synapse. Therefore, synaptic transmission is also called neurotransmission.
Besides neuron-neuron synapses there are also neuron-muscle fiber synapses, the neuromuscular junctions, and neuron-gland cell contacts (e.g. in the neurohypophysis).

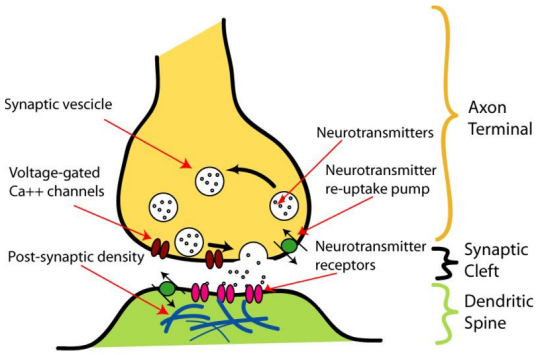


Fig. 2 Chemical synapse (from ref. 1).

Neurons form networks through which nerve impulses and the smaller graded potentials, such as the post synaptic potentials (PSP's) travel. Each neuron receives as many as 5000-15000 connections from other neurons. When a spike arrives at the synapse, it releases neurotransmitters. Receptors on the opposite side of the synaptic gap bind neurotransmitter molecules and respond by opening nearby ion channels in the post-synaptic cell membrane, causing ions to move in or out and changing the local membrane potential (Fig. 2). The resulting change in voltage is called an excitatory post synaptic

potential (EPSP) when excitatory (depolarization) or an inhibitory post synaptic potential (IPSP) when inhibitory (hyperpolarization or diminishing depolarization). Whether a synapse is excitatory or inhibitory depends on what type(s) of ion channel conduct the post-synaptic current display(s), which in turn is a function of the type of receptors and neurotransmitter employed at the synapse.
For instance the inhibitory neurotransmitter GABA causes an IPSP, which decreases the excitability and therefore decreases the neuron's likelihood to "fire" an action potential. In this way the output of a neuron may depend on the input of many others, each of which may have a different degree of influence, depending on the strength of its synapse with that neuron. The post synaptic neuron is connected to many more neurons, and if the total of excitatory influences is more than the inhibitory influences, it will also fire a spike.

Since each neuron is connected with numerous other neurons, it generates generally many PSP's. By dendritic summation of all PSP's (excitory and inhibitory) the membrane potential at the axon hillock may reach the firing threshold and so a spike in this second order neuron is generated.

Application

Various neurological disorders and diseases are caused by dysfunction of synapses, for instance since neurotransmitters are not synthesized, the vesicles in which they are stored dysfunction, the release fails or the deactivation by breaking down (e.g. acetylcholine) or re-uptake (recycling) fails, etc. Dysfunction of these synapses gives neuromuscular disorders, such as Myasthenia gravis (blockage of postsynaptic acetylcholine receptor of the neuromuscular junction) and the myasthenic syndrome of Lambert-Eaton where the presynaptic Ca^{++} channel of the neuromuscular junction, which regulate acetylcholine release, do not function well. A well known neurochemical disorder is Parkinson's disease caused by the loss of dopamine-secretion due to the death of dopaminergic neurons in the Substantia nigra.

More info

Of the more than 30 neurotransmitters is acetylcholine found in many synapses. Liberation of acetylcholine in the synaptic gap will lead to an increase of the sodium conductance g_{Na} and therefore to a depolarization of the membrane of the next nerve or muscle fiber.
Pharmacological studies have brought important contributions to our knowledge. Curare blocks the neuromuscular transmission of acetylcholine. TTX, tetradotoxine, another classical inhibitor, blocks the increase in g_{Na} at depolarization and is therefore a potent deadly drug. Tetraethyl ammonia (TEA), Cs and Rb block the K^+ channels.
Synaptic transmission can be modulated by e.g. desensitization; homotropic modulation of the presynaptic neuron by its own neurotransmitters (mostly inhibitory) and heterotropic modulation, a modulation of presynaptic terminals of nearby neurons.

Changes in synaptic strength can be short–term and without permanent structural changes in the neurons themselves, lasting seconds to minutes or long-term (long-term potentiation, LTP), in which repeated or continuous synaptic activation can result in second messenger molecules initiating protein synthesis, resulting in alteration of the structure of the synapse itself. Learning and memory are believed to result from long-term changes in synaptic strength, via a mechanism known as synaptic plasticity.

References
1. Wikipedia Chemical synapse.

Bioelectricity: electrical synapse

Principle

An electrical synapse is a mechanical and electrically conductive link, in general two-way, between two abutting neurons. It is formed at a narrow gap between the pre- and postsynaptic cells known as a gap junction. At gap junctions, the intercellular distance of about 25 nm narrows to about 3.5 nm of each other, a much shorter distance than the 20 to 40 nm distance that separates cells at chemical synapses. As opposed to chemical synapses, the postsynaptic potential in electrical synapses is not caused by the opening of ion channels by chemical transmitters, but by direct electrical coupling between both neurons. Electrical synapses are therefore faster and more reliable than chemical synapses. Electrical synapses are found throughout the nervous system, yet are less common than chemical synapses.

Electrical synapses are abundant in the retina (e.g. horizontal cells) and the cerebral cortex (e.g. in the vestibular and the trigeminal nucleus). Brain astrocytes show experimentally strong gap junctions (as do horizontal cells). There seems also to be weak neuron to glial cell coupling in some areas.
In organisms, electrical synapse-based systems co-exist with chemical-based, but are limited to systems that require the fastest possible response, for instance for escape mechanisms. The relative speed of electrical synapses also allows for many neurons to fire synchronously.
There exists another cell-cell junction, the tight junction, which has no particular electric properties. With respect to transport it behaves in some sense opposite to the gap-junction. By coupling cells with a dense network of tight junctions transport in the intercellular space is prevented. See **More Info** for a further description.
.

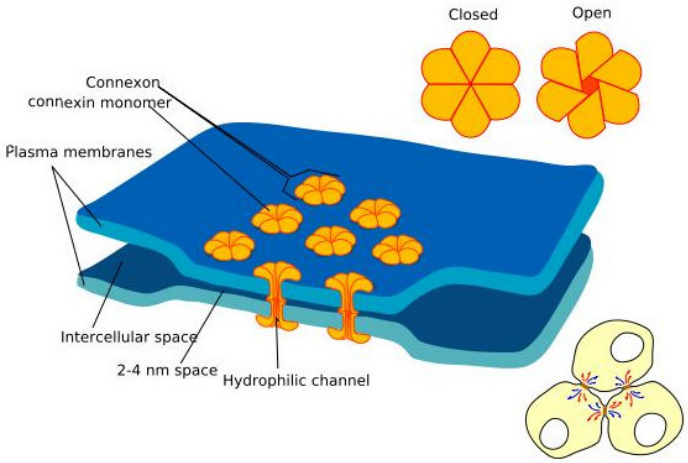


Fig. 1 Electrical synapse. Right under are three neurons connected by 3 electrical synapses (from ref. 1).

Application

Disorders caused by dysfunction of electrical synapses are scarce. Mutations in some gap junction genes cause white matter degeneration similar to that observed in Pelizaeus-Merzbacher disease (inhibiting myelin grow) and multiple sclerosis (disintegration of myelin sheaths).
More info

Gap junction
General Each gap junction (or nexus junction) contains many channels which cross the membranes of both cells. With a lumen diameter of about 1.2 to 2.0 nm, the pore of a gap junction channel is wide enough to allow ions and even medium sized molecules like signaling molecules to flow from one cell to the next thereby connecting the two cells' cytoplasm. Thus, when the voltage of one cell changes, ions may move through from one cell to the next carrying positive charge with them and depolarize the postsynaptic cell.
Morphology Gap junction channels are composed of two hemi-channels called connexons in vertebrates, one embedded in each of the membranes. Connexons are formed by six 7.5 nm long, two-pass membrane-spanning protein subunits, called connexins. However, in some cells, the hemichannel itself is active as a conduit between the cytoplasm and the extracellular space. Several gap junctions (hundreds) assemble into a macromolecular complex called a plaque. They are analogous to the plasmodesmata that join plant cells.
Delay and direction Without the need for receptors to recognize chemical messengers, signaling at electrical synapses is, with a synaptic delay of about 0.2 ms, more rapid than that across chemical synapses (about 2 ms). In cold-blooded animals the difference in speed is important since the chemical synapse is relatively slower due to the lower temperature.
Normally current carried by ions could travel in either direction through this type of synapse. However, sometimes the junctions are rectifying synapses containing voltage-dependent gates that open in response to a depolarization and prevent current from traveling in one of the two directions. Some channels may also close in response to increase the Ca^{++} or H^{+} ion concentration.
Long term adaptation There is evidence for "plasticity" at some of these synapses, i.e. that the electrical connection they establish can strengthen or weaken as a result of long lasting activity.

Function Gap junctions allows for direct electrical communication between cells, although with different connexin subunits can have different single channel conductances from about 30 pS (picoSiemens; 1 S = 1 mho = 1 Ω^{-1}) to 500 pS.
They allow also for chemical communication between cells through the transmission of small second messengers, such as IP_3 and Ca^{++}. Generally, molecules smaller than 1,000 Daltons can pass through.

Different connexin subunits can impart different pore sizes and different charge selectivity which results in selectivity for particular small molecules. Medium sized and large biomolecules (e.g. nucleid acid and protein) do not pass. Gap junctions ensure that molecules passing through the gap junction do not get leaked into the intercellular space.
Areas of electrical coupling In the myocardium the signal to contract is passed efficiently through the gap junctions, allowing the heart muscle cells to contract in tandem. Gap junctions occur in virtually all tissues of the body, with the exception of mobile cell types such as sperm and blood cells. Most disorders now associated with mutations in gap junction genes affect the skin, because this tissue is heavily dependent upon gap junction communication for the regulation of differentiation and proliferation.

Tight junctions
General Tight junctions, or zonula occludens, are the closely associated areas of two cells whose membranes join together forming a virtual impermeable barrier to fluid. It is only present in vertebrates. The corresponding junctions that occur in invertebrates are septate junctions (junctions with a septum).
Structure Tight junctions are composed of a branching network of sealing, independently acting strands, each strand from the others. Therefore, the efficiency of the junction in preventing ion passage increases exponentially with the number of strands. Their proteins are associated with membrane proteins located on the intracellular side of the plasma membrane which anchors the strands to the actin cytoskeleton. Thus, tight junctions join together the cytoskeletons of adjacent cells.
Functions They perform three vital functions. They hold cells together. They block the movement of integral membrane proteins between the apical and basolateral surfaces of the cell, allowing the specialized functions of each surface to be preserved. This aims to preserve transcellular transport. Finally, they prevent the passage of molecules and ions through the intracellular space between cells. So, materials must actually enter the cells (by diffusion or active transport) in order to pass through the tissue. Active transport is at the expenditure of metabolic energy, often in the form of ATP, to move molecules "uphill" against a concentration gradient or electric potential.
This pathway controls what substances are allowed to pass. Tight junctions play this role in maintaining the blood-brain barrier and blood-retina barrier. For example, L-DOPA, the precursor of dopamine, can cross the blood-brain barrier, whereas dopamine itself cannot. Therefore, L-DOPA is administered for dopamine deficiencies (e.g., Parkinson's disease) rather than dopamine).
Epithelia are classed as 'tight' or 'leaky' depending on the ability of the tight junctions to prevent liquid movement. Tight epithelia have tight junctions that prevent most movement between cells. An example of a tight epithelium is the distal convolutes tubule, part of the nephron in the kidney. Leaky epithelia do not have these tight junctions.

References and literature
1. Wikipedia Electrical synapse.
2. Gibson JR, Beierlein M, Connors BW. Functional properties of electrical synapses between inhibitory interneurons of neocortical layer 4. *J Neurophysiol*. 2005 93:467-80.
3. Hormuzdi SG et al. Electrical synapses: a dynamic signaling system that shapes the activity of neuronal networks. *Biochim Biophys Acta*. 20041662:113-37.
4. Kandel ER, Schwartz JH, Jessell TM. Principles of Neural Science, 4th ed., pp.178-180. McGraw-Hill, New York (2000), ISBN 0-8385-7701-6.

Bioelectricity: electrotonic propagation

Principle

If by (moderate) electrical stimulation a constant voltage is applied to the membrane of a nerve or muscle cell there is an exponential decay with a length constant (the lengthy over which the amplitude is diminished to the fraction 1/e), which can amount to some mm. Due to the considerable electrical resistance of the cell's interior $r_{interior}$ and the high capacitance of the membrane c_m there is a slow spread of the disturbance in time. Such phenomena are named electrotonic. They play an important role in the retina and in the central and autonomous nervous system especially in the dendrites and unmyelinated axons. It is the basis of the integrative action of the nervous system making one still more wonder why and how a simple mathematical model as treated below can have such an (even predictive) success.
The propagation of a disturbance of the rest potential E_{rest} is electrotonic with constant permeability's of the membrane for the various ions as long as no action potential (spike) is elicited. Electrotonic changes in membrane potential hold for dendritic, receptor and somatic potentials. They are also found in axons which axon hillock does not generate spikes. Similar potentials are found in synapses (pre- and postsynaptic potentials). All these potentials have a graded nature. During transmission they are filtered in time and place.
For the various dendrites and axons the membrane capacity per unit area and internal resistance are rather the same. Propagation speed is proportional with the square root of the diameter and reciprocal with the square root of membrane resistance. These two properties are highly variable and so speed varies a lot among dendrites and not-spiking axons. The decrement along the dendrite is proportional with the square root of membrane resistance. So, a high resistance means slow but with few losses over distance.
The electrotonic propagation is a passive phenomenon which only depends on the resistive and capacitive properties of the dendrite, axon or soma and is governed by the cable theory of physics. This is in contrast to the propagation of spikes along axons, an active process (see Bioelectricity: action potential).

More info

Suppose a nerve fiber or a dendrite has an axial membrane resistance r_m and capacitance c_m, both per unit length, and an internal axial, resistance of r_a (the outside resistance is neglected). The voltage difference between in- and outside is maintained by the three ionogenic equilibrium potentials (see Fig. 1 of Bioelectricity). Now we divide the dendrite in equal cylindrical pieces with a length x and we suppose that these pieces have the same properties.The current through the membrane per unit length i_m is the sum of the currents through r_m and c_m:

$$i_m = c_m \, dE/dt + E/r_m \qquad (1)$$

The axial current in the inside of the dendrite is proportional to the voltage difference E over a distance x. It is also proportional to the inverse of the resistance over a length x, i.e. $1/(r_a x)$. This means that the axial current is proportional with the second derivative of the voltage to the distance. Moreover, this current should be equal to i_m, since there cannot flow a net current in a closed circuit. So,

$$i_m = r_a^{-1} \cdot d^2E/dx^2. \qquad (2)$$

Combining (1) and (2) yields:

$$(1/r_a)\partial^2E/\partial x^2 = c_m \, \partial E/\partial t + E/r_m \qquad (3)$$

By scaling of the variable x with $X = x/\lambda$ where $\lambda = (r_m/r_a)^{0.5}$ (λ is the length constant) and t by $T = t/\tau$ were $\tau = r_m c_m$ (τ is the time constant, see Halftime and time constant), the equation becomes:

$$\partial^2E/\partial X^2 = \partial E/\partial T + E \qquad (4)$$

r_m, r_a and c_m can be found from the resistance and capacitance per unit of length, R_m, R_i and C_m. R_m is the specific membrane resistance times unit area, R_i the specific inner resistance, C_m is the specific capacitance per unit area, and d the diameter. So we obtain:

$$r_m = \pi d R_m, \; r_a = 4R_i/(\pi d^2), \; c_m = \pi d C_m. \qquad (5a)$$

And so λ is also equal to:

$$\lambda = (dR_m/4R_i)^{0.5}. \quad (5b)$$

Another important feature of the propagation of the electrotonic potential is the conduction velocity $\Theta = \partial X/\partial T$ of its peak potential.
By using the chain rule for $\partial^2 E/\partial x^2$.and the definition of θ it can finally be found that:

$$\partial^2 E/\partial x^2 = \theta^{-2}\cdot\partial^2 E/\partial T^2. \quad (6)$$

With (4) this yields the important expression:

$$\Theta^{-2}\cdot\partial^2 E/\partial T^2 .= \partial E/\partial T + E. \quad (7)$$

From the solution of this equation it appears that for the peak of the propagated wave the conduction velocity Θ is:

$$\Theta = \partial X/\partial T = 2\lambda/\tau = 0.5 C_m^{-1}(d/(R_m\cdot R_i))^{0.5} \quad (8)$$

For R_i = 36 Ω·cm, C_m = 1μF/cm^2, $R_m = 10^8$ Ω·cm^2 and d = 0.016 mm, θ is 3.3 mm/s.
From this example it is clear that passive conduction in membranes with a high R_m takes much time. But generally dendritic pathways are short. R_m is strongly dependent on the type of membrane. Membranes without ionic channels have an R_m of the order of 10^8 Ω·cm^2. This can become as small as 1000 Ω·cm^2, dependent on the channel density for the different ions (see Bioelectricity: action potential). For $R_m = 10^8$ Ω·cm^2 λ becomes 167 mm, so that 'the peak amplitude of the propagated wave hardly decreases within physiological distance. By applying a current pulse through an intracellular electrode τ can be estimated by measuring the voltage change with a second intracellular electrode (Fig. 1). An estimate of λ can be made by estimating the maximal potential (the asymptotic value) at various distances along the dendrite after a step function in current applied to the cell. Further r_a can be estimated. With λ, τ and d, r_m and c_m can be found (and also R_i, R_m, and C_m).
Of course the general solution of the wave equation is fairly complicated and will not be discussed here. Moreover it should be noted that the model is a strong simplification of the physiological reality. The model assumes that there is no intracellular radial resistance and that current density in the extracellular space is zero, i.e. the extracellular resistance is zero. Moreover the membrane model is not really linear, since also for small deviations from the rest potential the conductance's for Na^+ and K^+ change (see next paragraph), and consequently r_m is not constant.

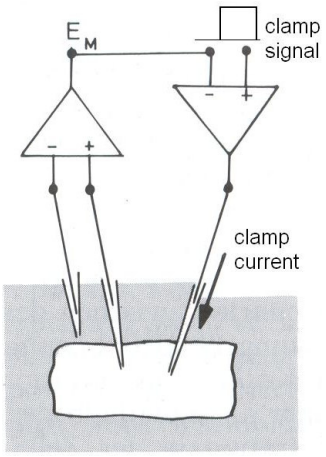


Fig. 1 Principle of current injection Clamping) and measuring membrane potential.

Many investigators tried to include in the model the neuronal geometry, especially the dendritic branches. Since the cell bodies are generally small and more or less globular of shape, the soma membrane is practically isopotential. At (passive) endings of dendritic and axonal branches o$\partial E/\partial x = 0$ since there the axial internal current is 0. A bifurcation of two branches, number 1 and 2, can be substituted by a single thicker branch, number 0, such that:

$$d_0^{3/2} = d_1^{3/2} + d_2^{3/2}. \quad (9)$$

If $d_1 = d_2$ then $d_0 = 2^{3/2} d_1 \approx 2.8\, d_1$. With this formalism a whole dendritic tree of a neuron can be substituted by a single piece of dendrite with a reasonable success.

ECG: augmented limb leads

Principle

For reading this chapter, it is advised to study the chapter ECG: basic electrocardiography in advance.

Unipolar leads and augmented limb leads
Basically a unipolar lead is an exploring electrode placed on a chosen site linked with an indifferent or reference electrode with an assumed very small potential.
To obtain a common reference with 'zero potential', two limb electrodes and an electrode placed at the sternum (called central terminal, CT, also Wilson's central terminal) are connected through 5 kΩ resistances to form the indifferent electrode, connected to the negative terminal of the ECG machine. The potential of the active electrode, connected to one the three limbs, creates one of the configurations of leads, being aVR, aVL or aVF. The other two electrodes of the standard leads are connected with CT. These are the *augmented limb leads*. So, the only difference with the bipolar leads I, II, and III is the choice of the common reference. As with the standard leads, the right leg is never connected for a lead, but it can be used to attach an earth electrode (see EEG: 12-lead ECG).

The augmented limb leads can be presented as vectors with a certain angle in some plane and time varying amplitude.

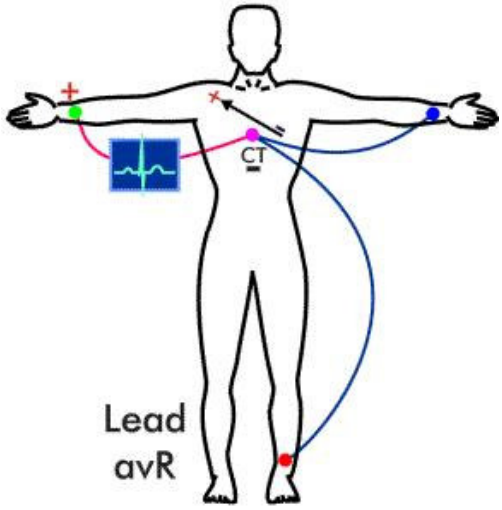


Fig. 1 Recording of aVR. Since the vector (above CT) is the source, the current which creates the voltage, flows in the body from negative to positive (as in a battery). The vector point to -150° (up are negative angles, just as in the hexaxial reference system (see ECG: Hexaxial reference system).

The three augmented leads are:
- Lead aVR or "augmented vector right" has the positive electrode (green in Fig. 1) on the right arm.
- Lead aVL or "augmented vector left" has the positive (blue) electrode on the left arm.
- Lead aVF or "augmented vector foot" has the positive (red) electrode on the left leg.

The augmented limb leads aVR, aVL, and aVF are amplified (augmented) compared to the signal when a unipolar limb lead is made with CT as the reference electrode. Such a lead is rather small to be useful. Together with leads I, II, and III, the augmented limb leads aVR, aVL, and aVF form the basis of the hexaxial reference system, which is used to calculate the heart's electrical axis in the *frontal plane*. See for an example of all six leads of a healthy subject Fig. 2 of EEG: 12-lead ECG.

Application

This three leads are frequently used in the clinic, for instance for intensive care and for routine non-cardial surgery, but then V5 (see EEG: 12-lead ECG) is used in addition.

More info

When the limb electrodes are connected one by one with CT the leads VR (right arm), VL (left arm) and VF (left leg) are obtained. aVR is equal to:

$$aVR = VR - (VL + VF)/2 \quad (1)$$

since (left arm) and VF (left leg) are connected, so the mean of both signals results. CT is the common reference and therefore it is not in the equation. Eithoven's triangle says that VR + VL + VF = 0 or VR = – (VL + VF). Substitution in (1) yields that aVR = 3VR/2. For the other two, aVL and aVF the same holds. So, recording the unipolar leads VR, VL and VF is a disadvantage. Therefore, the leads I, II and III are recorded together with the augmented leads.

ECG: basic electrocardiography

Principle

An electrocardiogram or ECG is a graphic representation of the electrical activity of the heart over time produced by an electrocardiograph. Understanding the various waves and normal vectors of depolarization and repolarization yields important diagnostic information.

Calibration
A typical electrocardiograph (or PC monitor) runs mostly at a paper speed of 25 mm/s. With a paper ECG, the finest division is a block of 1 mm^2 and 25 mm/s represents 1 mm/40 ms. A diagnostic quality 12-lead ECG is calibrated at 10 mm/mV.

Filter selection
Modern ECG monitors offer multiple filters for signal processing with a *monitor mode* and *diagnostic mode*. In monitor mode, the low frequency filter (also called the *high-pass filter*: signals above the threshold or cut-off frequency are allowed to pass) is set at either 0.5 Hz or 1 Hz. The high frequency filter (the *low-pass filter*) is set at 40 Hz (see Linear first order system). The *high-pass filter* reduces wandering of the baseline and the low pass filter reduces high frequency noise and 50 or 60 Hz power line hum. (Hum can also be suppressed by a selective band pass filter, a T-filter). In diagnostic mode, the high pass filter is set at 0.05 Hz, allowing accurate ST segments. The low pass filter is set to 40, 100, or 150 Hz. Consequently, the monitor mode ECG display is more filtered than diagnostic mode.

Waves and intervals of the ECG

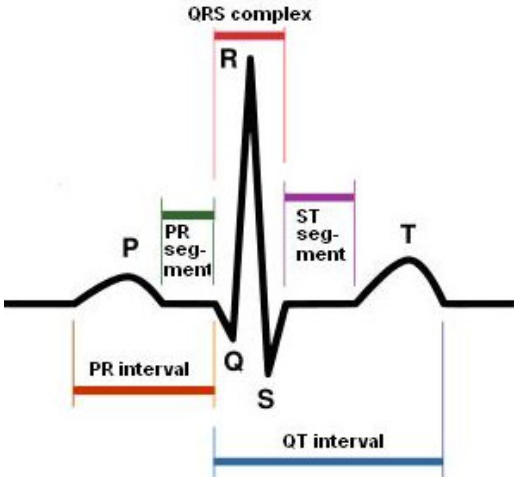


Fig. 1 Schematic representation of normal ECG

The baseline voltage of the electrocardiogram is known as the *isoelectric line*. A typical ECG tracing of a normal heartbeat (or cardiac cycle) consists of a P wave, a QRS complex and a T wave. A small *U wave* is normally visible in 50 to 75% of ECGs. For a detailed description the reader is referred to the textbooks of clinical physiology.

Working principle of electrodes

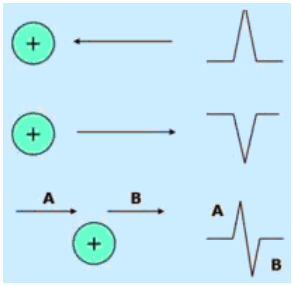


Fig. 2 Relationship between positive electrodes (green circles) and the propagating depolarization wavefronts (at the right).

An ECG is obtained by measuring electrical potentials between various points of the body using a biomedical instrumentation amplifier. A *lead*, records the electrical signals of the heart from a particular combination of recording electrodes which are placed at specific points on the body.
When a depolarization wavefront (or electrical vector) moves toward and away a positive electrode, it creates a *positive* and *negative* deflection in the corresponding lead respectively as depicted in Fig. 1.
When a depolarization wavefront (or electrical vector) moves perpendicular to a positive electrode, it creates an *equiphasic* (or *isoelectric*) complex. It will be positive as the depolarization wavefront (or mean electrical vector) approaches (A), and then become negative as it passes by (B).

Standard ECG leads

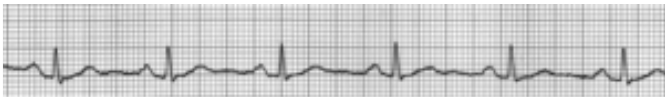


Fig. 3 Lead II

The basic three bipolar limb leads Leads I, II (Fig. 3) and III are the so-called bipolar *limb leads* because electrodes are attached to the arms and legs forming the *Einthoven's triangle*. All three electrodes are 'active' and there is no reference electrode. Therefore the leads are bipolar. Einthoven, who discovered the ECG, placed legs and arms in buckets of salt water (Fig. 4). They are the first three leads of the modern 12 lead-ECG (see ECG: 12-lead ECG).

Fig. 4 ECG as done by Willem Einthoven based on electromechanical technology with a wire galvanometer as sensor.

The bipolar (standard) leads
The electrodes are attached as follows:
• *lead I* = left arm positive minus right arm negative (LA–RA)
• *lead II* = left leg positive minus right arm negative (LL–RA).
• *lead III* = left leg positive minus left arm negative (LL–LA).

Application

- It is the gold standard for the evaluation of cardiac arrhythmias
- It guides therapy and risk stratification for patients with suspected acute myocardial infarction.
- It helps detect electrolyte disturbances (e.g. hyperkalemia and hypokalemia)
- It allows for the detection of conduction abnormalities (e.g. right and left bundle branch block)

- It is used as a screening tool for ischemic heart disease during a cardiac stress test
- It is occasionally helpful with non-cardiac diseases (e.g. pulmonary embolism or hypothermia)

However, the electrocardiogram does not assess the contractility of the heart.

More info

Lead II should be equal to the sum of leads I and III, so I + III = II. This is called *Einthoven's Law*. It is written this way (instead of I + II + III = 0) because Einthoven reversed the polarity of lead II in Einthoven's triangle. Then, QRS complexes are upright.

The position from which the heart is viewed by each of these leads is shown in Figure 5.

ECG: body surface mapping

Principle

Body surface mapping of the electric activity of the heart refers to the use of many recording sites (>64) arranged on the body so that isopotential surfaces could be computed and analyzed over time. Extensive computer algorithms, applied in experimental cardiologic research, calculate these isopotential surfaces.

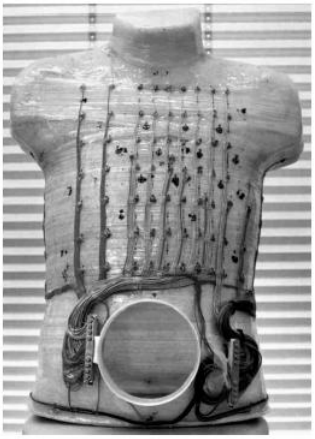


Fig. 1 Front view of a torso phantom with 138 electrodes (63 on the left anterior chest wall).

The main objective of body surface potential maps (BSPMs) is to calculate with complicated mathematics the underlying electric sources approximated by equivalent current dipoles (ECDs). There are many solutions of this calculation, in terms of number of dipoles, their time-varying positions, directions and magnitudes. This problem is called the *backward problem*. In contrast, the forward problem (calculating the potential distribution on the surface of the torso) has an unique solution. The theory of the forward and backward problem holds for any electrical source in the body and is, in addition to the heart, most advanced developed for brain activity (see Magnetoenceplalography).
The backward problem is piece wise solved. For instance, by focusing to a specific temporal window and restricting the area of location, one of the dipoles can be found. The solution can suppose electric isotropy of the heart and the surrounding torso tissues. A refinement is reached by supposing anisotropic heart models (Fig. 2). The myocardial anisotropy can be determined by a heterogeneous 3D matrix from segmented magnetic resonance images (see MRI: general).
The most relevant dipoles are located in the septum, apical area, left ventricular wall or right ventricular wall. The solution of the apical source is most dependent on the anisotropy of the torso tissues.

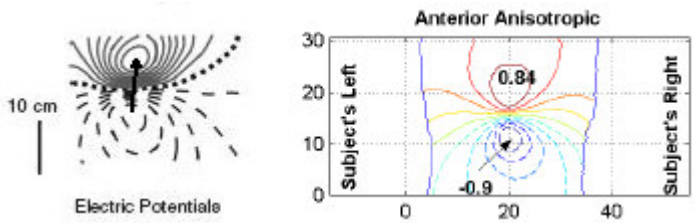


Fig. 2 Left Isocontour lines (0.5 mV resolution). Solid lines positivity, dashed lines negativity, zero line is dotted. The arrow indicates the ECD calculated with a BEM model. Right Contour plot of the BSPM in frontal view for a *z*-oriented (vertical) apical ECD (FEM). Potentials in mV and distances in cm.

More Info

The current and potential distributions in the torso can be computed using ECDs decomposed in the orthogonal x, y, z direction, or with a 3D algorithm. Solution can be based on the *finite element method* (FEM) or *boundary element method* (BEM). FEM is a volume-discretisation method of numeric calculus applied to a set of equations in a 3D-matrix (the elements). BEM is a numerical method of solving linear equations formulated in the *boundary integral* form. Conceptually, it works by constructing a "mesh" over the surface (with a small surface/volume ratio to be effective). The mesh is often a set non-uniform triangles (*triangular segmentation*), depending on the local curvature of the surface.
Often BSPM is combined with magnetocardiography. Together they give better solutions of the ECDs.

ECG: hexaxial reference system

Principle

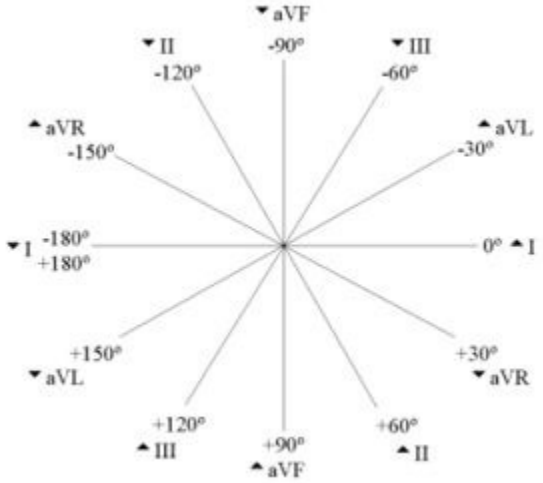


Fig. 1 Hexaxial reference system. Notice that negative is up. Triangle up denotes that with normal polarity the main peak of the QRS complex is positive. This holds for all 6 leads from 0 ^{o} to 150^{o} except AVR.

The hexaxial reference system is a diagram that is used to determine the heart's electrical axis in the frontal plane. The heart's *electrical axis* refers to the general direction of the heart's depolarization wavefront (or *mean electrical vector*) by using the polarity of the QRS complex in leads I, II, and III in the frontal plane. It is usually oriented in a right shoulder to left leg direction, which corresponds to the right inferior quadrant of the hexaxial reference system, although a slightly broader range, -30^{o} to +90^{o} is considered to be normal.

The diagram is based on the first six leads (I, II, III, aVR, aVL, and aVF) of the 12-lead ECG (see ECG: 12-lead ECG). To determine the heart's electrical axis, first the most isoelectric (or equiphasic) lead should be located on a diagnostic quality ECG with proper lead placement. Then the corresponding spoke on the hexaxial reference system should be found. The perpendicular spoke will point to the

heart's (principal) electrical axis. To determine which of the two opposite located numerical value (in degrees) should be used, use the polarity of the perpendicular lead on the ECG. For example, if the most isoelectric lead is aVL, the perpendicular lead, being lead II has an angle of +60°.

Application

The reference system is used for diagnosis of cordial disorders.
Normal and deviating directions are classified as:

- Normal axis: -30° to +90°
- Left axis deviation: -30° to -90°, may indicate left anterior fascicular block or Q waves from inferior myocardial infraction.
- Right axis deviation: +90° to +180° may indicate left posterior fascicular block, Q waves from high lateral myocardial infarction, or a right ventricular strain pattern.
- Extreme axis deviation: -90° to -180°

ECG: vectorcardiography

Principle

Vectorcardiography is the registration, by formation of a loop display on a PC, of direction and size (vector) of the moment-to-moment electric activity of the heart during one complete cycle. The electric activity is generated by one of the major event of the myogenic activity, the auricular depolarization (P waves), the ventricular depolarization (QRS complex) and waves of ventricular repolarization (T waves). The source of the activity are supposed to be a single *electric dipole* which position, direction and magnitude changes during the heart cycle. The position describes a single loop during one heart cycle. Because the resultant traces were all loops of variable shapes, the traces were referred to as P loop, QRS loop and T loop. The 3D-vector loops are represented by projections upon the frontal plane. Fig. 1 illustrates how the QRS vector is constructed in the frontal plane at the instant of the R peak.

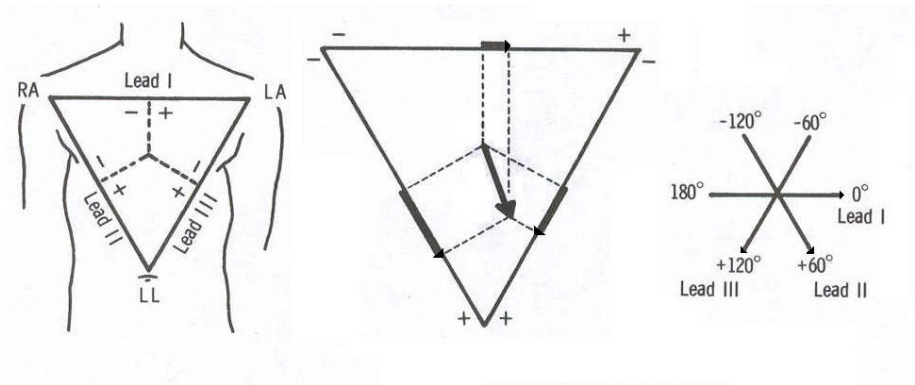


Fig. 1 Left: Configuration of Einthoven's leads. Middle: vector representation. Right: the 3 leads in the hexaxial presentation (see ECG: hexaxial reference system).

Application

Its application is mainly in the field of experimental cardiology. Clinical applications are limited. On-Line vectorcardiography has been applied during coronary angioplasty. Monitoring ST (segment) vector magnitude and QRS vector difference by vectorcardiography is used for identifying myocardial ischaemia during carotid endarterectomy.

More Info

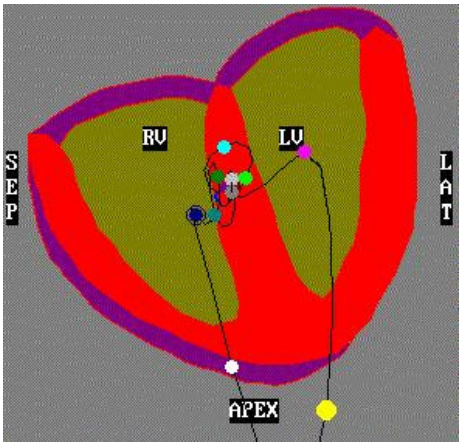


Fig. 2 High resolution vector cardiogram (HRVEC) with many loops.

The nomenclature and symbolic representation of the spatially oriented loops is standardized. The various letters denote the various vectors. One of the systems is:
q: the spatial direction of the vector of the QRS loop having the greatest magnitude,
p: the unit vector perpendicular to QRS and to the 'QRS plane' (the plane containing the QRS loop),
t: the spatial direction of the vector of the T loop having the greatest magnitude.

ECG: 12-lead ECG

Principle

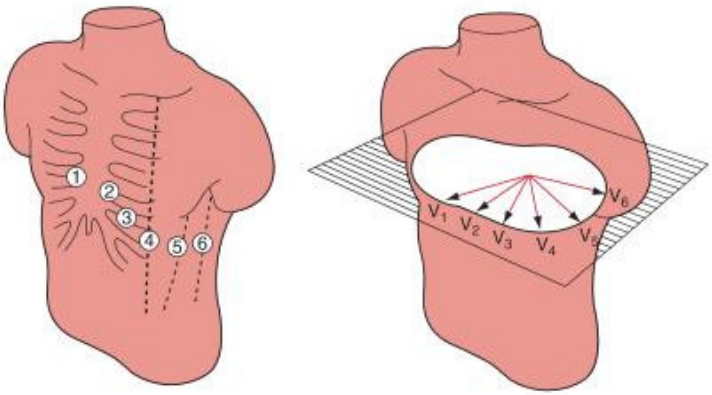


Fig. 1 Placement and vector representation of the precordial leads.

Unipolar chest leads, the precordial leads

When unipolar leads are recorded from the chest wall, the exploring electrode is connected to the positive pole of the ECG and the negative to the central terminal CT of Wilson (see ECG: augmented limb leads) at the sternum. By convention, the following sites are normally selected (Fig. 1):
• *V1*, the fourth intercostal space just to the right of the sternum
• *V2*, the fourth intercostal space just to the left of the sternum
• *V3*, midway between V2 and V4
• *V4*, the fifth intercostal space in the midclavicular line
• *V5*, the left anterior axillary line at the same horizontal level as V4
• *V6*, the left midaxillary line at the same horizontal level as V4.

Since the precordial leads V1, V2, V3 (the right precordial leads) and V4, V5, and V6 (left precordial leads) are placed directly on the chest, close to the heart, they do not require augmentation. CT is used as reference (negative input of the ECG machine), and consequently these leads are considered to be

unipolar. The precordial leads view the heart's electrical activity in the so-called horizontal plane, in which the electrical Z-axis is located.
The QRS complex should be negative in lead V1 and positive in lead V6. The QRS complex should show a gradual transition from negative to positive between leads V2 and V4.
The precordial leads are always recorded together with the three basic Einthoven leads I, II and III (see ECG: basic electrocardiography) and the augmented leads (see ECG: augmented limb leads). Fig.2 presents them all 12 for a healthy subject.

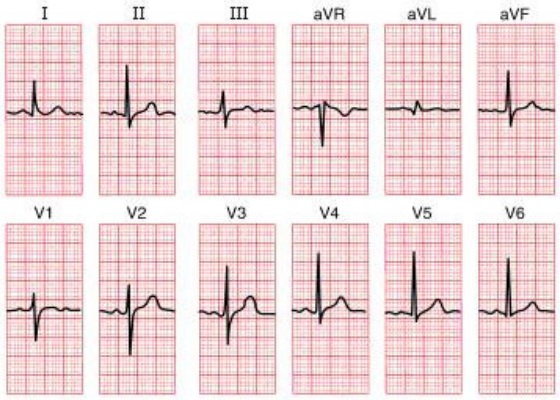


Fig.2. Normal 12-lead ECG

Ground electrode
In modern four-lead (the augmented leads and V5) and twelve-lead ECGs, an additional electrode is the ground electrode (usually green). This electrode is placed on the right leg by convention, although in theory it can be placed anywhere.
With a three-lead ECG, when one dipole is viewed, the remaining lead becomes the ground lead by default. Notice, that the ground lead is not the reference electrode (generally CT, see also ECG: augmented limb leads). The ground electrode suppresses 50 or 60 Hz hum of the mains.

Application

The equiphasic (or isolectric or biphasic, see bottom configuration Fig. 2 of ECG: basic electrocardiography) lead is referred to as the transition lead. When the transition occurs earlier than lead V3, it is referred to as an *early transition*. When it occurs later than lead V3, it is referred to as a *late transition*. There should also be a gradual increase in the amplitude of the R wave between leads V1 and V4. This is known as *R wave progression*. Poor R wave progression is a nonspecific finding. It can be caused by conduction abnormalities, myocardial infarction, cardiomyopathy, and other pathological conditions.

More info

The twelve leads, each recording the activity from a different perspective, which also correlates to the area of identifying acute coronary injury, are classified as follows.
The *inferior lead**s*** (leads II, III and aVF) look at electrical activity from the vantage point of the inferior or diaphragmatic wall of the left ventricle.
The *lateral leads* (I, aVL, V_5 and V_6) look at the electrical activity from the vantage point of the lateral wall of left ventricle. Because the positive electrode for leads I and aVL are located on the left shoulder, leads I and aVL are sometimes referred to as the high lateral leads. Because the positive electrodes for leads V5 and V6 are on the patient's chest, they are sometimes referred to as the low lateral leads.
The *septal leads*, V_1 and V_2 look at electrical activity from the vantage point of the septal wall of the left ventricle. They are often grouped together with the anterior leads.
The *anterior leads*, V_3 and V_4 look at electrical activity from the vantage point of the anterior wall of the left ventricle.
In addition, any two precordial leads that are next to one another are considered to be contiguous. In other words, even though V4 is an anterior lead and V5 is a lateral lead, they are contiguous because they are next to one another.
Lead aVR offers no specific view of the left ventricle, but views the endocardial wall from the right shoulder.

The modern ECG machine is completely integrated with an analog front end, a 12- to 16-bit analog-to-digital (A/D) converter, a computational microprocessor, and dedicated input-output (I/O) processors. These systems compute the 12 lead signals and analyze them with a set of rules. Fig. 3 shows the ECG

of a heartbeat and the types of measurements that might be made on each of the component waves of the ECG and used for classifying each beat type and the subsequent cardiac rhythm.

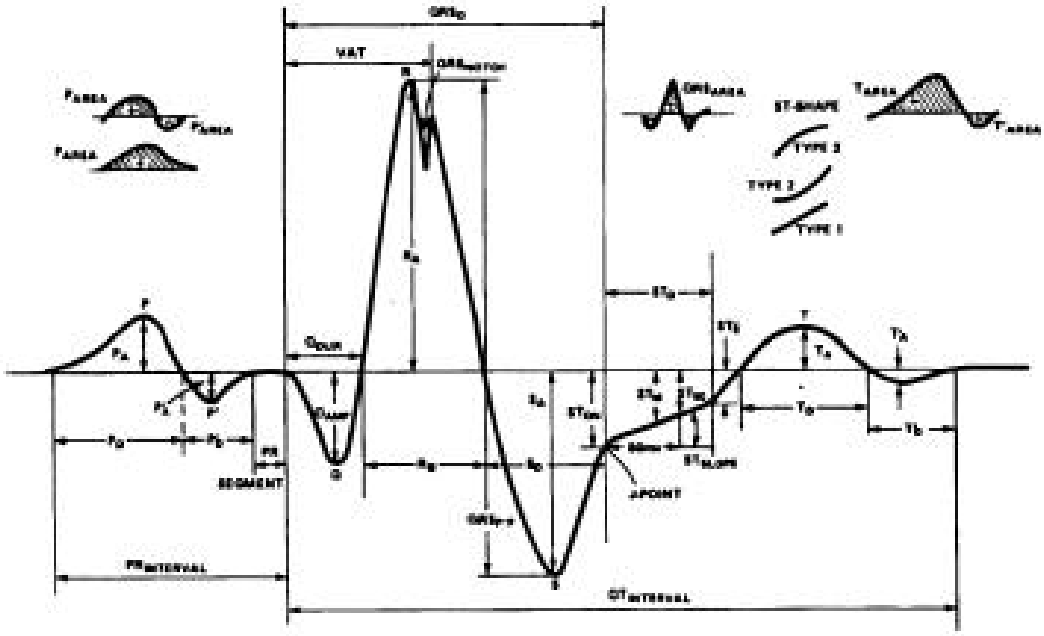


Fig. 3 The numerous ECG measurements that can be made with computer-based algorithms, with for instance artificial neural networks, wavelet analysis, component analysis. The measurements are primarily durations, amplitudes, and areas.

Electroencephalography

Basic Principles

Electroencephalography is the neurophysiologic measurement of the electrical activity of the brain by recording from electrodes placed to the scalp, or in special cases on the cortex (sub-dural). The resulting traces are known as an electroencephalogram (EEG).

The recording is obtained by placing electrodes on the scalp, usually after preparing the scalp area by light abrasion and application of a conductive gel to reduce impedance. Each electrode is connected to an input of a amplifier, which amplifies the voltage (typically 1,000–100,000 times, or 60–100 dB of voltage gain), and then displays it on a screen or inputs it to a computer. The amplitude of the spontaneous ongoing EEG is about 100 µV when measured on the scalp, and about 1-2 mV when measured on the surface of the brain. The amplitude of evoked potentials and pre-motor activity is up to some 15 µV. In general averaging is necessary to elucidate these signals from the spontaneous EEG.

Methods
The electrode-amplifier relationships are typically arranged in one of three ways:
Common reference derivation
One terminal of each amplifier is connected to the same electrode, and all other electrodes are measured relative to this single point. It is typical to use a reference electrode placed somewhere along the scalp midline, or a reference that links one or both earlobe electrodes.
Average reference derivation
The outputs of all of the amplifiers are summed and averaged, and this averaged signal is used as the common reference for each amplifier.
Bipolar derivation
The electrodes are connected in series to an equal number of amplifiers. For example, amplifier 1 measures the difference between electrodes A and B, amplifier 2 measures the difference between B and C, and so on.
This distinction has become void with the advent of computerized or paperless EEGs, which record all electrodes against an arbitrary reference and will calculate the above montages post hoc.
The choice of the reference is crucial for the resulting spontaneous or evoked EG and the activity maps.
EEG has several limitations. Scalp electrodes are not sensitive enough to pick out individual action potentials, the electric way signaling in the brain, resulting in releasing inhibitory, excitatory or modulatory neurotransmitters. Instead, the EEG picks up synchronization of neurons, which produces a greater voltage than the firing of an individual neuron. Secondly, EEG has limited anatomical specificity when compared with other functional brain imaging techniques such as functional MRI. With the use of a large number of electrodes EEG brain activity maps can be constructed (see Fig. 1) and the source of the activity can be estimated (see Magnetoencephalography).

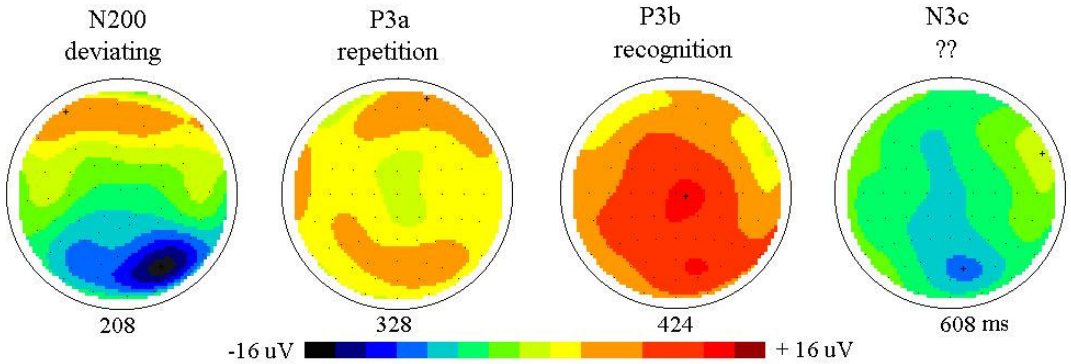


Fig. 1 Maps of the response components a 68 year old subject to the presentation of a rare course visual pattern within a continuous series of a fine pattern stimulus. The component names and their supposed functions are indicated at the top of the map. Nc3 is possibly related to anticipation, expectancy and action to a given stimulus. The numbers below the maps are the times after the start of the stimulus at which the maps are constructed from the responses of 64 channels.

EEG has several strong sides as a tool of exploring the brain activity. As other methods for researching brain activity have time resolution between seconds and minutes, the EEG has a resolution down to sub-millisecond. The brain is thought to work through its electric activity. EEG is in addition to the highly specialized MEG (Magnetoencephalography) method the only method to measure it directly. Other methods for exploring functions in the brain do rely on blood flow or metabolism which may be

decoupled from the brain electric activity. Newer research typically combines EEG or Magnetoencephalography with MRI, SPECT or PET to get high temporal and spatial resolution.

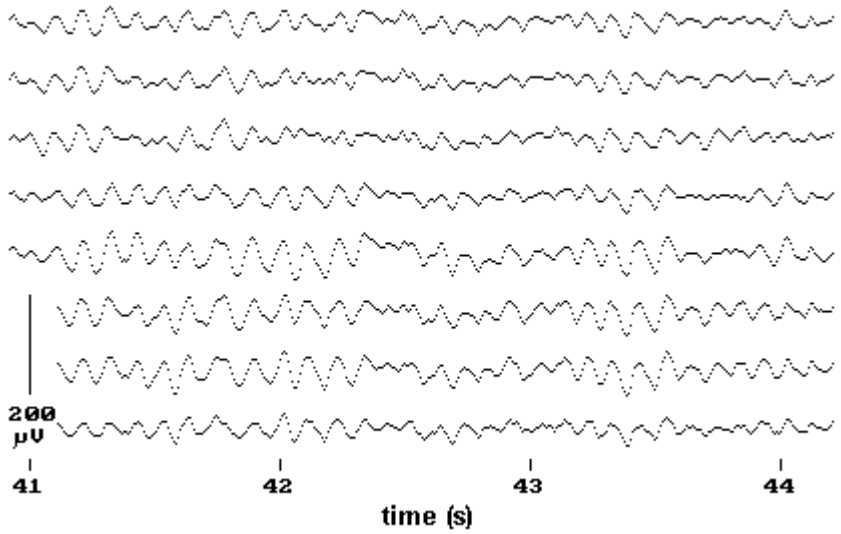


Fig. 2 Three seconds of spontaneous EEG recorded, shown for eight of the 64 channels. The sine-wave like signals represent the alpha activity. In the spectrum this activity gives rise to the so-called alpha-peak.

Wave types
Historically four major types of continuous (spontaneous) rhythmic sinus-like EEG waves are recognized: delta (up to 4 Hz), theta (4-8 Hz), alpha (8-13 Hz; Fig. 2) and beta (13-40 Hz). An alpha-like normal variant called mu is sometimes seen over the motor cortex (central scalp) and attenuates with movement, or rather with the intention to move. The sensorimotor rhythm (SMR) is a middle frequency (about 12-16Hz) associated with physical stillness and body presence. Gamma is the frequency range above 40 Hz (approximately 30-80 Hz to be precise).

Application

This device is used to assess brain damage, caused by tumors, CVA's and trauma's. It also applied for supporting the diagnosis of neurological disorders like dementia, etc. and psychiatric disorders. Neuroscientists and biological psychiatrists use EEGs to study the function of the brain by recording the spontaneous and evoked EEG during controlled behavior of human volunteers and animals in lab experiments. Theories to explain sleep often rely on EEG patterns recorded during sleep sessions. In addition, the procedure is routinely used clinically to assist in the diagnosis of epilepsy. In some jurisdictions it is used to assess brain death.
The EEG has a history of tens years and so there is a mass of journal publications and handbooks about EEG theory and applications.

Literature
Regan D. Human brain electrophysiology: evoked potentials and evoked magnetic fields in science and medicine. New York: Elsevier, 1989.

Electrophysiology: general

Principle

Electrophysiology is the study of the electrical properties of biological cells and tissues. It involves measurements of voltage change or electrical current flow by electrodes in various systems, from single ion channel proteins to single neurons (particularly action potentials) and whole tissues like the heart.

Classical electrophysiological techniques
The principal types of electrodes are:
1) metal discs (singles or arrays),
2) metal tracings on printed circuit boards.
2) needles (singles or arrays), diameter µm scale, inserted into or just outside a single cell.
4) small glass pipettes, also for extra- or intracellular recording.
The principal preparations include:
- living organisms,
- excised tissue (acute or cultured),
- dissociated cells from excised tissue (acute or cultured),
- artificially grown cells or tissues,
- hybrids of the above.

The pipette techniques has various versions:
- With a low impedance pipette an extracellular placement may pick up the activity of several nearby cells simultaneously, and this is termed multi-unit recording (also with a metal microelectrode).
- With the electrode tip (tip size ca. 1 µm, resistance some MΩ) more closely to the neuron, the recording is termed single unit recording (also with metal microelectrodes, resistance some 100 kΩ).
- With the pipette tip pressed against the cell membrane, to which it tightly adheres. This is a semi-intracellular recording with spikes of some 5-10 mV. (pipette resistance a few tens of MΩ).
- The electrolyte within the pipette may be brought into fluid continuity with the cytoplasm by delivering a pulse of pressure to the electrolyte in order to rupture the small patch of membrane encircled by the pipette rim (intracellular or whole cell recording, resistance some 50 MΩ). The electrolyte may contain a drug. When the pipette contains a tracer (e.g. HRP) then the resistance can be about 200 MΩ.
- Alternatively, ionic continuity may be established by "perforating" the patch by allowing exogenous ion channels within the electrolyte to insert themselves into the patch (perforated patch recording).
- The patch may be left intact (patch recording).
- Finally, the patch is so small that it comprises only one channel: a single channel recording.

As electrode size increases, the resolving power decreases. Larger electrodes are sensitive only to the net activity of many cells, the so called local field potentials (see **More Info**). Still larger electrodes, such as non insulated needles and surface electrodes (discs) used by clinical (EEG, and surgical neurophysiologists, are sensitive only to certain types of synchronous activity within populations of cells numbering in the millions.

Optical electrophysiological techniques
Optical electrophysiological techniques were created to overcome the limitations that electrical activity is recorded at approximately a single point within a volume of tissue. Interest in the spatial distribution of bioelectric activity prompted development of molecules capable of emitting light in response to their electrical or chemical environment. Examples are voltage sensitive dyes and fluorescent proteins. With one or more such compounds administrated via perfusion, injection or gene expression, the distribution of electrical activity may be observed and recorded.

Application

Many particular electrophysiological techniques and recordings have abundant clinical application:
Electrocardiography (ECG), for the heart;
Electroencephalography (EEG), for the brain and Electrocorticography for the cerebral cortex;
Electromyography (EMG) for the muscles
Electro-oculography, for the eyes
Electroretinography,for the retina
Electro-olfactography for the olfactory receptors;
Evoked potentials for auditory, visual, somatosensory assessment.

Bioelectric Recognition Assay (BERA)

BERA is a novel method for measuring changes in the membrane potential E_m of cells immobilized in a gel matrix. Apart from the increased stability of the electrode-cell interface, immobilization preserves the viability and physiological functions of the cells. BERA is primary used in biosensor applications in order to assay analytes which can change E_m in a characteristic, 'signature-like' way. BERA has been used for the detection for human viruses, veterinary disease agents and plants in a highly specific, rapid (1-2 minutes), reproducible and cost-efficient fashion. The method has also been used for the detection of environmental toxins, such as herbicides and the determination of very low concentrations of superoxide anion in clinical samples. A recent development of the BERA technology is the technique called *Molecular Identification through Membrane Engineering* (MIME). This technique allows for building cells with absolutely defined specificity against virtually any molecule of interest, by embedding thousand of artificial receptors into the cell membrane.

More info

Extracellular recording
A microelectrode (tip size ca. 1 µm), into the brain, nerve or sensory tissue (e.g. retina) of a living animal will usually detect the activity of one neuron or axon. The spikes of a "single unit" recording are very like the intracellular spikes, but they are much smaller (typically 0.1-1 mV).
Multi-unit recordings are often used in conscious animals to record changes in the activity of a discrete brain area during some behavior of the animal. Recordings from one or more such electrodes which are closely spaced can be used to identify the number of cells around it as well as which of the spikes come from which cell. This process is called spike sorting and is suitable in areas where there are identified types of cells with well defined spike characteristics.
If the electrode tip is bigger still, generally the activity of individual neurons cannot be distinguished but the electrode will still be able to record a field potential generated by the activity of many cells. Such an electrode inserted in or near the auditory nerve can record a compound action potential.

Intracellular recording
Intracellular recording involves measuring voltage and/or current across the membrane of a cell with (generally) a glass micropipette (< 1µm and a resistance of tens of MΩ) filled with an electrolyte and connected with a metal wire (generally an Ag wire coated with AgCl), to make a connection between the electrolyte and the amplifier. The coating reduces the galvanic potential between electrolyte and silver wire. This is necessary since the galvanic potential is very large (+1.98 V). Moreover, it reduces electrode noise which is higher the smaller the contact surface.
The pipette must be inserted inside the cell, so that the membrane potential can be measured (in rest −60 to −80 mV). Micropipettes are filled with a solution that has a similar ionic composition to the intracellular fluid of the cell. The voltage measured by the electrode is compared to the voltage of a reference electrode, usually an Ag-AgCl wire in contact with the extracellular fluid somewhere in the tissue.

Pipette impedance
In general, the smaller the electrode tip, the higher its electrical resistance so an electrode is a compromise between size (small enough to penetrate with minimum damage) and resistance (low enough so that small neuronal signals can be discerned from thermal noise in the electrode tip, see below). Pipettes have also a capacitance, as does the electrode cable and the amplifier has an input capacitance. With a capacitance poor cable (also short, and with a driven guard, i.e. the shield has the amplifier input voltage by using a buffer amplifier) and capacitance poor amplifier, these parallel capacitances together can be brought down to about 10 pF.
The pipette restistance R_{pip} is:

$$R_{pipette} = R_{shaft} + R_{tip} + R_{medium},$$

$$R_{pipette} = r_{electrolyte} L_{shaft}/(¼\pi d_{shaft}^2) + (2/d_{tip} - 2/d_{shaft}) r_{electrolyte} L_{tip}/(½\pi d_{shaft}) + r_{medium}/(2\pi d_{tip}), \quad (1)$$

where d is diameter and L is length. With $r_{electrolyte}$ = 75 Ωcm, = r_{medium} = 750 Ωcm, L_{shaft} = 5 cm. L_{tip} = 0.5 cm, d_{shaft} = 0.08 cm and $d_{tip} = 5\ 10^{-5}$ cm, then we obtain:

$$R_{pipette} = 75\ k\Omega + 11.9\ M\Omega + 2.4\ M\Omega.$$

The far majority of $R_{pipette}$ is due to the tip geometry. With d_{tip} fixed, $R_{pipette}$ can only be reduced by reducing L_{tip}. With an electrode resistance 100 MΩ the resulting low pass first order filter (see Linear first order system) has a cut off frequency of 160 Hz (time constant 1 ms), hardly enough to record intracellular spikes. Since the pipette capacitance is determined by the thickness (reciprocally) and the surface (linear) of the glass wall (the dielectricum), the geometry of the tip is of great importance.

Sharp electrode technique
For recording the potential inside the cell membrane with minimal effect on the ionic constitution of the intracellular fluid a sharp electrode can be used. They look like patch clamp pipettes but the pore is much smaller so that there is very little ion exchange between the intracellular fluid and the electrolyte in the pipette. The resistance of the electrode is tens to hundreds of MΩ. Often the tip of the electrode is filled with various kinds of voltage sensitive dyes like Lucifer yellow to be injected by Electrophoresis. A

positive or negative, DC or pulsed voltage is applied to the electrodes depending on the polarity of the dye. Later, this enables identification of the cell under a microscope.

Field potentials

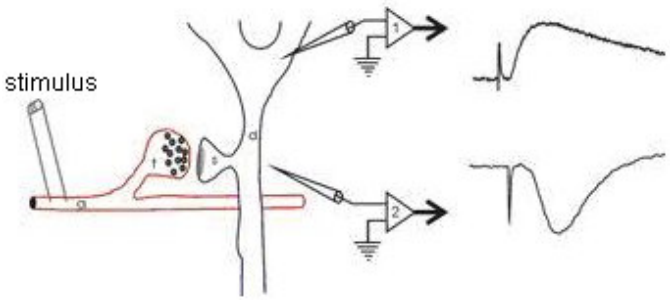


Fig. 1 A schematic diagram of a field potential recording from a rat hippocampus. When the synapse releases glutamate the postsynaptic membrane opens ionotropic glutamate receptor channels. The net flow of current is inward, so a current sink is generated. A nearby electrode (#2) detects this as negativity. An *intracellular* electrode placed inside the cell body (#1) records the change in membrane potential that the incoming current causes.

Extracellular field potential

Extracellular field potentials are local current sinks or sources that are generated by the collective activity of many cells. Usually a field potential is generated by the simultaneous activation of many neurons by synaptic transmission. Fig. 1 to the right shows a hippocampal synaptic field potential. The lower trace shows a negative wave that corresponds to a current sink caused by positive charges entering cells through postsynaptic glutamate receptors, while the upper trace shows a positive wave that is generated by the current that leaves the cell (at the cell body) to complete the circuit.

Local field potential

A local field potential (LFP) is recorded using a low impedance extracellular microelectrode, placed sufficiently far from individual local neurons to prevent any particular cell from dominating the signal. The unfiltered signal reflects the sum of action potentials from cells within approximately 50-350 µm from the tip of the electrode and slower ionic events from within 0.5-3 mm from the tip of the electrode. This signal is then low-pass filtered, cut off at ~100 Hz, to obtain the LFP. The low impedance and positioning of the electrode allows the activity of a very large number of neurons to contribute to the signal. The low-pass filter removes the spikes and the LFP remains.
The amplifier measures the electrical potential difference between the microelectrode and a reference electrode, placed somewhere else in the body or tissue with a similar extracellular medium (to cancel the both electrochemical polarization effects at the interface of both metal electrodes (or the Ag/AgCl wires in the pipette).

Synchronized Input

The LPF is believed to represent the synchronized input into the observed area. The quick fluctuations, caused by the short inward and outward currents of the action potential, are filtered out, leaving only the slower fluctuations. Therefore the spike plays no part in the LFP, which thus comprises the more sustained currents in the tissue, typical of somato-dendritic origin. The major slow current is the PSP. It was thought until recently that EPSP's and IPSP's were the exclusive constituents of LFP's. However, phenomena unrelated to synaptic events have been found to contribute to the LFP.

Geometrical Arrangement

Cells which contribute to the slow field variations are determined by the geometric configuration of the cells themselves (as with the pyramidal cells.)
When there is simultaneous activation of the dendrites a strong dipole is produced. His may even give rise to a signal recordable with EEG. In cells where the dendrites are arranged more radially, in a plane, the potentials between individual dendrites and the soma tend to cancel.

Thermal noise

The thermal noise generated in a conductor (e.g.a resistor), V_n (amplitude, measured as the effective (rms) value, in Volt) is:

$$V_n = 2(kTR \cdot \Delta f)^{0.5}, \qquad (2)$$

where k is Boltzmann's constant, T the absolute temperature (K), R the resistance (Ω), and Δf the bandwidth of the amplifier.
For extracellular recording V_n hardly plays a role since R is low. As example: R = 2.5 MΩ, T = 300 K, and Δf is 300 Hz, then V_n = 3.5 µV, whereas spikes are generally larger than 10 µV. For intracellular recordings the same holds. With R = 250 MΩ, V_n = 35 µV, just small enough to find the spikes

extracellular before impaling the cell. However, in addition there are other noise sources: amplifier noise; with a digital output the noise in the analog-to-digital converter; noise generated at the interface of the metal electrode (or the Ag/AgCl wires in the pipette); noise generated by the opening of an ion channel caused by the net flow of ions into the cell from the extracellular medium, or out of the cell into the extracellular medium. Which type of noise dominates depends on the precise conditions of the recording.

Electromyography

Principle

Electromyography (EMG) is a mainly clinical electrophysiological technique for evaluating and recording physiologic properties of muscles at rest and while contracting. EMG is performed with an electromyograph recording the potentials generated by the muscle cells.
The electrical source of the EMG is the muscle membrane potential of about -70mV. The resulting measured potentials range between about 50 µV and 30 mV. Typical repetition rate of muscle unit firing is about 7–20 Hz. Damage to motor units can be expected at ranges between 450 and 780 mV.
Muscle tissue at rest is normally electrically inactive. After the electrical activity caused by the irritation of needle insertion subsides, the electromyograph should detect no abnormal spontaneous activity (i.e. a muscle at rest should be electrically silent, with the exception of the area of the neuromuscular junction, which is normally electrically very spontaneously active). When the muscle is voluntarily contracted, action potentials begin to appear. As the strength of the muscle contraction is increased, more and more muscle fibers produce action potentials. When the muscle is fully contracted, there should appear a disorderly group of action potentials of varying rates and amplitudes (a complete recruitment and interference pattern).

EMG can be recorded invasively with an intramuscular needle electrode (see for electrodes Electrophysiology: general). Abnormal spontaneous activity might indicate nerve and/or muscle damage. Then the patient is asked to contract the muscle smoothly. The shape, size and frequency of the resulting motor unit action potentials (MUAPs) are judged. Because skeletal muscles differ in the inner structure, the electrode has to be placed at various locations to obtain an accurate study. So, the electrode is retracted a few mm, and again the activity is analyzed until at least 10-20 units have been collected.
The non-invasive technique with a surface electrode may be used to monitor the general picture of muscle activation. This technique is mostly used. Optimal electrode position and applied voltage are adjusted by auditory or visual inspection of the recorded signal (biofeedback).
The characteristics of the MUAP are determined by the number of muscle fibers per motor unit, the metabolic type of muscle fibers (red or pale) and other factors.
Nerve conduction testing (for conduction speed and amplitude decay) is also often done at the same time as an EMG in order to diagnose neurological diseases.
Of importance are amplitude (dependent on fibers/unit and MUAPs in the EMG, end the duration of the EMG (dependent on duration and synchronization of MUAPs).

Application

EMG is used to diagnose two general categories of disease: neuropathies (e.g. alcoholic neuropathy, peripheral neuropathy, sensorimotor polyneuropathy, denervation (reduced nervous stimulation), poliomyelitis, carpal tunnel syndrome) and myopathies (e.g. myasthenia gravis, Duchenne muscular dystrophy, and myopathy).
EMG is also applied in physiotherapy by biofeedback (auditory or visual inspection of the recorded signal) to learn to control muscles involved in backaches, headaches, neck pain, the breathing muscles and heart muscle.

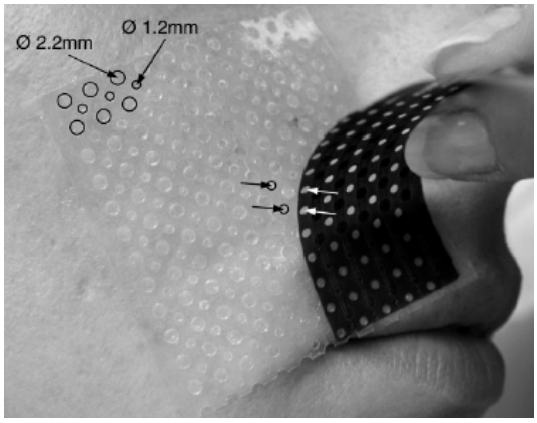


Fig. 1 Flexible multi-electrode grid for high-density surface EMG. The 1.2 mm circles are holes and the 2.2 mm circles the electrodes.

More Info

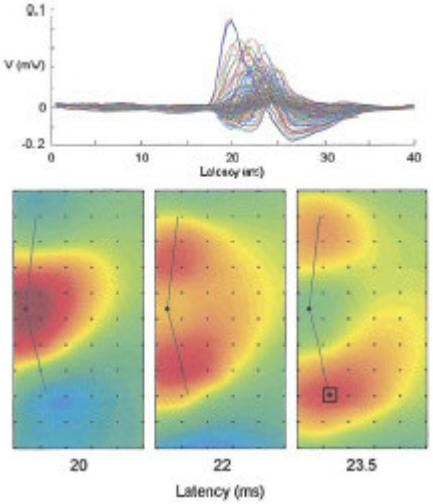


Fig. 2 Data from high-density multi-electrode grid recoding. Top: “butterfly” presentation of a 6x10 grid. Bottom: 3 activity maps of the 6*10 signals, constructed at the times indicated after stimulation.

For a thorough analysis, the measured EMG signals can be decomposed into their constituent MUAPs. MUAPs from different motor units tend to have different characteristic shapes, while MUAPs recorded by the same electrode from the same motor unit are typically similar. Notably MUAP size and shape depend on where the electrode is located with respect to the fibers and so can appear to be different if the electrode moves position. EMG decomposition is non-trivial, although many methods have been proposed.
A development of the last decade is high-density surface EMG, which can recently be performed with flexible multi-electrode grids.

Lorentz force

Principle

The Lorentz force is the force exerted on a charged particle in an electromagnetic field. The particle will experience a force due to electric field of *q*E, and due to the magnetic field *q*v × B. Combined they give the Lorentz force equation (or law):

$$F = q(E + v \times B), \quad (1)$$

where
F is the force (in N)
E is the electric field (in V/m)
B is the magnetic field (in Webers/m^2, or equivalently, Tesla's)
q is the electric charge of the particle (in C, Couloms)
v is the instantaneous velocity of the particle (in m/s)
and × is the cross product (see below).

Thus a positively charged particle will be accelerated in the *same* linear orientation as the *E* field, but will curve perpendicularly to the *B* field according to the right-hand rule ("kurketrekker regel").
The cross product is a vector operation in a three-dimensional Euclidian space (with generally Cartesian coordinates, so with orthogonal axes). It is also known as the vector product or outer product. It differs from the dot product (also called the inner or scalar product, i.e. in a one-dimensional space the common multiplication) in that it results in a vector rather than in a scalar. Its main use lies in the fact that the cross product of two vectors is orthogonal to both of them. Fig. 1 illustrates the vector product.

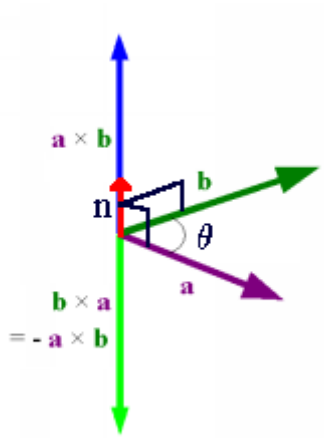


Fig. 1 Cross product of vectors a and b, with θ the measure of the angle between **a** and **b** ($0° \le \theta \le 180°$), on the plane defined by the span of the vectors. **n** is the unit vector perpendicular to both **a** and **b** (a vector of size 1 with the dimension of the both vectors taken together).

Equivalently to equation (1), we can express the Lorentz force law in terms of the electric charge density ρ and current density J as;

$$F = \int_V (\rho E + J \times B) \cdot dV \,. \quad (2)$$

Applications

The Lorentz force is a principle exploited in many physical devices including those also applied for medical research and in the clinic. A simple application is the electrophoretic trough and complicated one the mass spectrometer (see Mass spectrography) and the cyclotron and other circular path particle accelerators (necessary for PET, see Positron emission tomography).

Magnetoencephalography

Basic Principles

A magnetoencephalograph or MEG machine (Fig. 1) is the device to measure the magnetic field produced by the coherent electrical activity of assemblies of neurons in the brain.

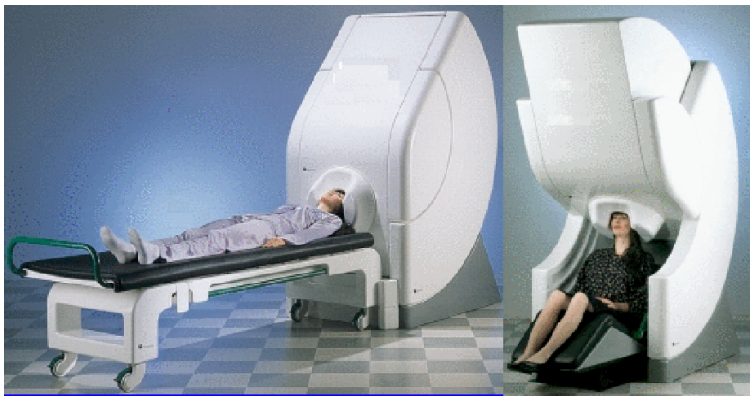


Fig. 1 MEG machine for operation in supine and sitting position of the subject

The magnetic signals themselves derive from currents flowing in the dendrites of neurons. The laws of electromagnetism govern the parameters of the field (see http://en.wikipedia.org/wiki/Magnetic_field). Given the direction of the current, the right-hand rule applies for the direction of the field. The fields are measured just outside the scalp, using extremely sensitive magnetic sensors (mostly gradiometers), including a so-called SQUID, (Super conducting Quantum Interference Device; see More info), mounted in a kind of helmet (Fig. 2), which is filled with liquid helium to provide super conducting.

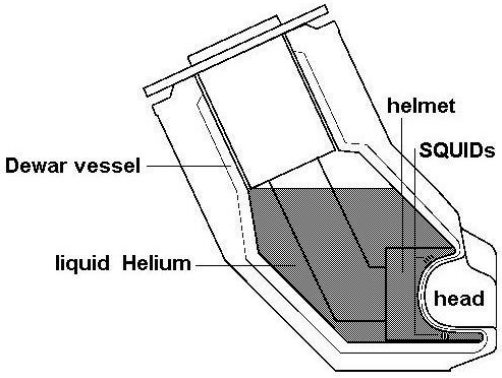


Fig. 2 Basic principle of operation of MEG machine

Because the magnetic signals emitted by the brain are on the order of a few femtotesla (1 fT = 10^{-15} T), the measurements are performed in a magnetically shielding room to exclude interference with external magnetic signals, including the 10^8 stronger earth's magnetic field (see Fig. 3). The room reduces the high-frequency noise, while noise cancellation algorithms reduce low-frequency signals. The signals of additional sensors mounted far from the helmet reduce low frequency spatial noise.
Modern whole-head systems have roughly 300 channels, and have a noise floor of around 5-7 fT (> 1 Hz). The ongoing spontaneously evoked magnetic field of the brain is typically around 100 to 1000 fT, while sensory responses and pre-motor signals are at least ten times smaller and often close or lower than the noise floor. These signals can only be made visible by applying signal-averaging techniques. Fig. 4 (top) gives an example of an analysis, a MEG map constructed 162 ms after applying a visual stimulus.

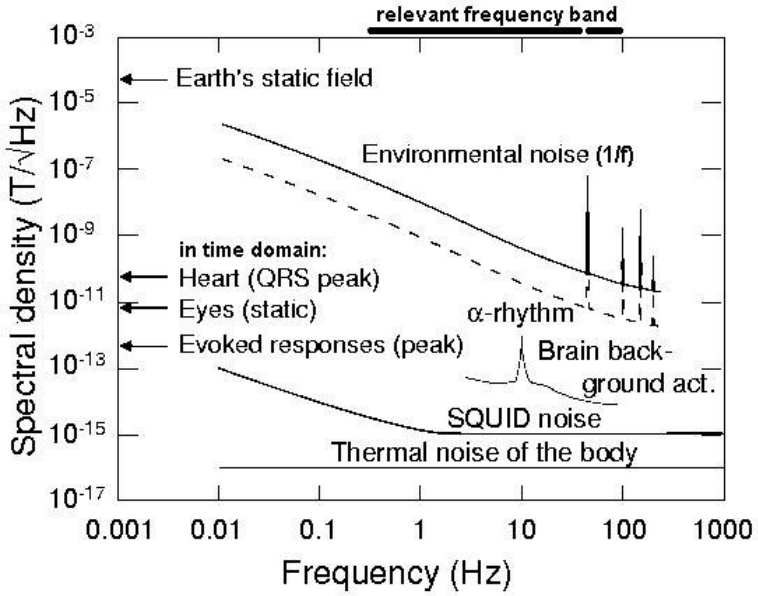


Fig. 3 Amplitudes of MEG signals and noise sources

MEG is a relatively new technique with good spatial and a temporal resolution, thus complementing other brain activity measurement techniques such as electroencephalography (EEG), positron emission tomography (PET), and functional Magnetic Resonance Imaging (fMRI). The primary technical difficulty with MEG is its high sensitivity for interfering signals evoked outside the brain by motion of charged objects (e.g. the eyeball, heart, limbs) and metal implants, and its vulnerability for non-brain generated signals (ECG, EMG, loudspeaker).

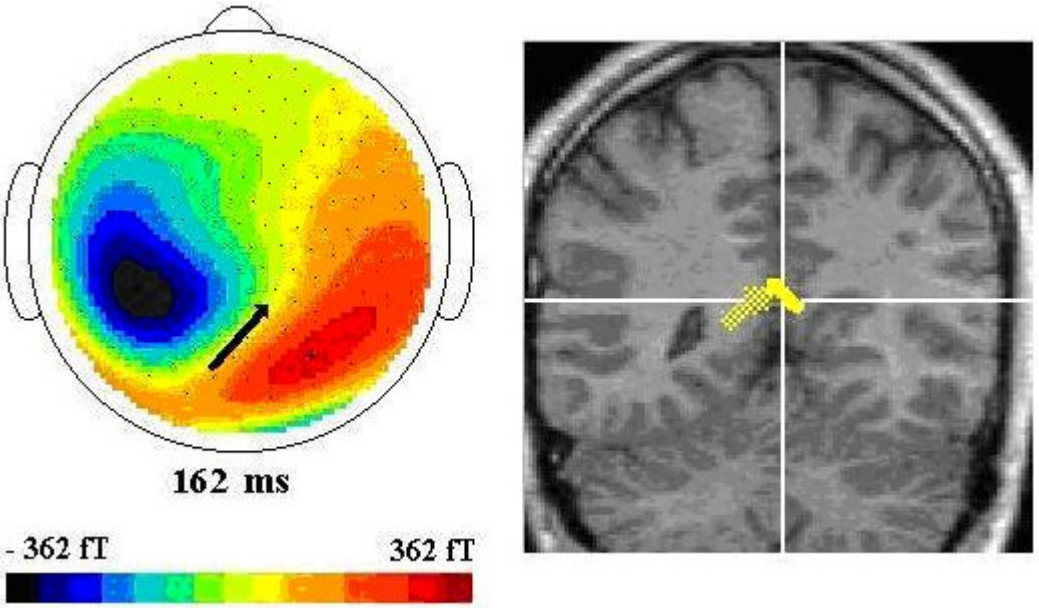


Fig. 4 Field and source analyses On top, the map is constructed 162 ms after applying a visual stimulus. The ECD is presented by the black arrow. An advanced analysis (bottom) of the same field, now with the MRI image, shows that this single ECD is comprised of two ECDs closely together. (Data provided by the author.)

Applications

Detecting and localizing epileptiform spiking activity in patients with epilepsy, and in localizing eloquent cortex for surgical planning in patients with brain tumors. (see More info). Localization of the source, source modeling, is performed with complicated mathematics by solving the 'inverse-problem', which is subject of intensive research (see **More info**). Diagnostics of cognitive and behavior disorders are often performed in combination with EEG or fMRI.

In the Netherlands the VUMC and UMC St Radboud have MEG machines. These machines are as expensive as fMRI machines and for its operation a highly specialized technical staff is necessary. Specialized commercially MEG machines are developed for examining the heart and the unborn human fetus.

More info

Localizing the source of activity

With mathematical physics it is possible to localize the source of electric brain activity. To obtain the source(s) the so-called inverse problem should be solved, a procedure of complicated numerical mathematics. A set of equations (the model) with numerically known constants is applied to an input-data set comprised of the signals of the sensors (the channels) and the anatomical data of the head (often provided by a MRI image), in order to estimate the source(s), the output of the model. This estimate is not unique, since there are many solutions. The solution, called an equivalent current dipole (ECD) is only unique if the system is noise-free, if there is only one source active and if the head is approximated by a magnetically homogeneous sphere. In reality these conditions certainly do not hold. In practice the simplest approach is mostly to consider a pair of ECDs, one in each hemisphere, and a spherical head. The best accuracy of localization is 2-3 mm, about two times better than with EEG. An advanced approach is to use MRI images of the head and supposing a distributed source. Now, the density (a cloud of sources of equal strength) of hypothetical sources is calculated and visualized similarly as fMRI and SPECT analyses.
Inverse problems are typically ill posed, as opposed to the well-posed problems The forward problem is well-posed: for a given source and system parameters, the magnetic field can be calculated and there is only one unique solution.

SQUID

There are various types of SQUID, but all have at least one super conducting Josephson junction (Fig. 5).

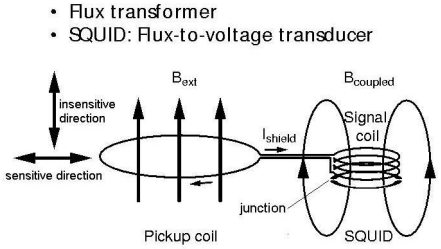


Fig. 5 Principle of operation of the magnetic sensor, in this case a magnetometer. B is the magnetic field density.

The basic principle of operation is to cancel the current though the junction with an adjustable current though the signal coil (Fig. 5). The strength and direction of the compensating current is the primary sensor signal. In MEG, often gradiometers are applied. They are characterized with two compensation coils with opposite winding.
SQUIDs are the most sensitive devices to measure extremely tiny magnetic fields with noise levels as low as 3 $fT \cdot Hz^{-½}$. Another application is the scanning SQUID microscope.

Literature

Sato S, et al. Principles of magnetoencephalography. J Clin Neurophysiol. 1991;8:144-56. Review.
Schellart NA, et al. Temporal and spatial congruence of components of motion-onset evoked responses investigated by whole-head magneto-electroencephalography. Vision Res. 2004 Jan;44(2):119-34.
Faugeras O, et al.Variational, geometric, and statistical methods for modeling brain anatomy and function. Neuroimage. 2004;23 Suppl 1:S46-55. Review.
Turner R, Jones T. Techniques for imaging neuroscience. Br Med Bull. 2003;65:3-20. Review.

Piezoelectricity

Basic principle

Piezoelectricity is the ability of certain quartz analogue crystals or ceramics to generate a voltage in response to applied mechanical stress. The word is derived from the Geek piezein, which means to squeeze or press. The piezoelectric effect is reversible in that piezoelectric crystals when subjected to an externally applied voltage, can change shape by a small amount. The deformation, about 0.1% of the original dimension, is of the order of nm, but nevertheless finds numerous applications.
In a piezoelectric crystal, the positive and negative electric charges are symmetrically separated, such that the crystal overall is electrically neutral. Each of these sites of charge forms an assembly of aligned electric dipoles. When a mechanical stress is applied, the symmetry is disturbed, and the charge asymmetry generates a voltage across the material. For example, a 1 cm cube of quartz with 2000 Newton of correctly applied force upon it, can produce a voltage of 12,500 V.
Piezoelectric materials also show the opposite effect, called converse piezoelectricity, where the application of an electrical field creates mechanical deformation in the crystal. The crystal is maximally excited by a sine wave signal at the resonance frequency (see Second order linear system) of the crystal. (See for a technical description of the piezoelectric effect http://en.wikipedia.org/wiki/Piezoelectric.)

Applications

As very high voltages correspond to only tiny changes in the width of the crystal, this width can be changed with better-than-µm precision, making piezo crystals the most important tool for positioning objects (e.g. by stepper motors) with extreme accuracy.
The applications of piezoelectricity in medicine, science, technology and daily life (e.g. electric lighter) is enormous. Examples are the production and detection of sound (e.g ultrasound in medicine, loudspeakers, microphones in medical (e.g. audiology) and daily live applications), ultra-fine focusing of optical assemblies (atomic force microscope and scanning tunneling microscope) employ converse piezoelectricity to keep the sensing needle close to the probe), fine-tuning a laser's frequency, generation of high voltages, electronic frequency generation, microbalances, laser mirror alignment, acousto-optic modulator, a device that vibrates a mirror to give the reflected beam a Doppler shift. This is useful for fine-tuning a laser's frequency.

More info

Ultrasonic transducers Piezoelectric materials are used as utrasonic transducers for imaging applications (e.g. medical imaging,) and high power applications e.g. in medical treatment and sonochemistry, e.i. the effect of sound waves on chemical systems, e.g in sonoluminesence (the emission of short bursts of light from imploding bubbles in a liquid when excited by ultrasound) and sonic cavitation. For imaging applications, the transducer can act as both a sensor and an actuator.
Piezoelectric motors Types of piezoelectric motor include the well-known travelling-wave motor used for auto-focus in cameras, inchworm motors (single cell electrophysiology) for linear motion, and rectangular four-quadrant motors with high power density (2.5 W/cm³) and speed ranging from 10 nm/s to 800 mm/s. All these motors work on the same principle. Driven by dual orthogonal vibration modes with a phase shift of 90°, the contact point between two surfaces vibrates in an elliptical path, producing a frictional force between the surfaces. Usually, one surface is fixed causing the other to move. In most piezoelectric motors the piezoelectric crystal is excited at the resonance frequency of the crystal.
Quartz clocks employ a tuning fork made from quartz that uses a combination of both direct and converse piezoelectricity to generate a regularly timed series of electrical pulses that is used to mark time. The quartz has a precisely defined natural frequency of oscillating and this is used to stabilize the frequency of a periodic voltage applied to the crystal. The same principle is critical in all radio transmitters and receivers, and in computers where it creates a clock pulse. Both of these usually use a frequency multiplier to reach the MHz ranges.

Vision

Ophthalmic Corrections

Principle

Optical corrections with lenses (spectacles or contact lenses) have primarily the aim to compensate for an unallowable deficit (hyperopia) or surplus (myopia) of the refractive power of the eye. Advanced measures like intraocular lenses and cornea reshaping are beyond our scope. With correction the image is no longer blurred, apparently increases the acuity. Actually acuity is not changed (see Visual acuity).

Application

Hyperopia
Hyperopia may be corrected with positive (convex) lenses. Mild to moderate hyperopia can be compensated through accommodation and may not need correction (or may not even be noticed). However, the perennial accommodation is a strain on the eye and may cause headache.
Myopia
Myopia causes problems already at 0.5D. Correction is with negative (concave) lenses.
Presbyopia
Against presbyopia, a lack of accommodation due to aging, some means of changing the correction according to distance is needed. This may be achieved using:

- Separate spectacles for long distance (weak positive or negative lens) and for reading (positive lens).
- Bifocal lenses, where most of the lens surface has a power suitable for vision at long distance, but a region in its lower part has a different power, for reading etc. There are also trifocal lenses.
- Progressive lenses, which use the same idea as bifocals, but there is a smooth grading between the upper and lower part of the lens, see Fig. 1.

Progressive lenses are adequate but not everybody can get used to having sharp sight only in a horizontal band across the field of vision. They require very precise fitting.

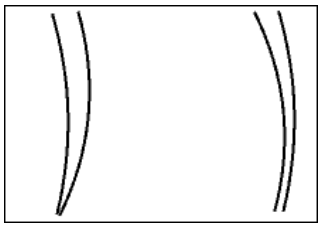


Fig. 1 Progressive lenses The left lens is intended for an emmetrope, the right one for a myope.

Astigmatism
Against astigmatism (i.e. different power of the eye for different sectors of the visual field), astigmatic lenses with the opposite orientation are needed, eventually as a progressive lens.
Contact lenses
Contact lenses, mostly chosen for cosmetic reasons, cause less distortion in the outer parts of the field of vision than spectacles do (since they are closer to the main plane of the eye optics).
The most important subdivision of contact lenses is that of soft (commonest) and hard lenses.
Hyperopia and myopia are corrected according to the same principles as with spectacles.
Presbyopia and astigmatism are more problematical, because contact lenses tend to rotate. Several devices can prevent this rotation mechanically; they are successful in some people and not in others.
There are several other ways to correct presbyopia with contact lenses, none of them entirely satisfactory:

- Different lenses on each eye - you use one eye for distant sight and the other for near sight. Stereoscopic vision and judgment of distance are impaired, but it works fairly well in practice.
- The centre of the lens has the power for distant sight and the periphery of the lens the power for near sight. The idea is that when you look downwards, the lower eyelid prevents the lens from following, so that you look through its peripheral part. This works only with hard lenses (soft ones stick too strongly to the cornea).
- The centre of the lens has the power for near sight and the periphery of the lens the power for distant sight (i.e. the opposite of the above). Here the idea is that when you accommodate, the pupil constricts, and the eye only receives light which has passed through the centre of the lens. Since this reflex weakens with age, the method is not too reliable.

Ophthalmoscopy

Principle

The ophthalmoscope, originally invented by the famous Herman von Helmholtz, is an instrument used in determining the health of the retina, the vitreous humor and other interior structures of the eye. It consisting essentially of a mirror that reflects light from a light source into the eye. Through a central hole in the mirror the physician can examine the patient's eye. Fig. 1 presents the principle with a the light pathway from the patient's retina to its image on the retina of the physician. b present a part of the pathway of illumination of the patent's retina. The physician actually watches the image (with height h', see c) at a distance of 250 mm from his eye.

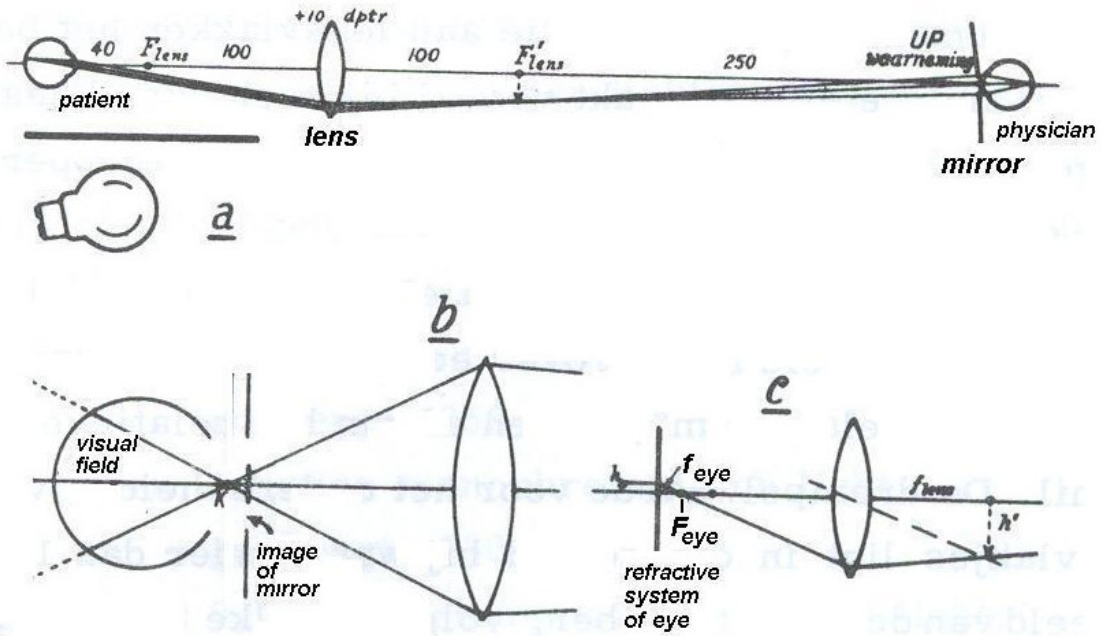


Fig. 1 Ophthalmoscope with image (a and c) and illumination (b) light pathways. All distances in mm.

Application

In addition to the mentioned general applications there are some specific ones.
In patients with headaches, the finding of swollen optic discs, or papilledema, on ophthalmoscopy is a key sign, as this indicates raised intracranial pressure which could have several causes. Cupped optic discs are seen in glaucoma.
In patients with diabetes, regular ophthalmoscopic eye examinations (once every 6 months to 1 year) is mandatory to screen for diabetic retinopathy as visual loss due to diabetes can be prevented by retinal laser treatment if retinopathy is spotted early.
In arterial hypertension, hypertensive changes of the retina closely mimic those in the brain, and may predict cerebrovascular accidents (strokes).

More info

There are several methods of types of ophthalmoscopy each performed with a different version of the opthalmoscope.

Direct ophthalmoscope
Direct ophthalmoscopy using a slit-lamp and negative auxiliary lenses can provide a very high level of magnification-even greater than that of the monocular hand held direct ophthalmoscope. Stereopsis is provided to a greater degree than all other examination techniques.

Binocular indirect ophthalmoscope
Binocular indirect ophthalmoscopy is a technique used to evaluate the entire ocular fundus. It provides for stereoscopic, wide-angled, high-resolution views of the entire fundus and overlying vitreous. Its optical principles and illumination options allow for visualization of the fundus regardless of high ametropia, hazy ocular media, or central opacities.

Monocular indirect ophthalmoscope
Monocular indirect ophthalmoscopy combines the advantages of increased field of view (indirect ophthalmoscopy) with erect real imaging (direct ophthalmoscopy). By collecting and redirecting peripheral fundus-reflected illumination rays, which cannot be accomplished with the direct ophthalmoscope, the indirect ophthalmoscope extends the observer's field of view approximately four to five times. An internal lens system then re-inverts the initially inverted image to a real erect one, which is then magnified. This image is focusable using the focusing lever/eyepiece lever.

Laser scanning ophthalmoscopy
This is a method of examination of the eye. It uses the technique of confocal laser scanning microscopy for diagnostic imaging of retina or cornea of the human eye. It is helpful in the diagnosis of glaucoma, macular degeneration, and other retina disorders. It has been combined with adaptive optics technology to provide sharper images of the retina. The principle of illumination and imaging is illustrated in Fig. 2 and 3.

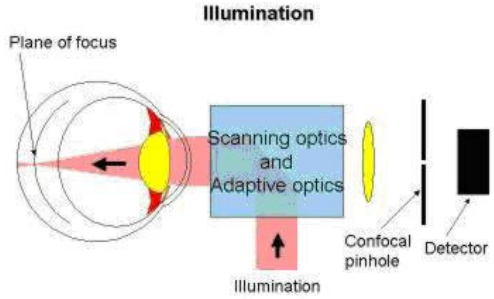


Fig. 2 Illumination The illumination beam uses the eye's optics to focus the illumination light to a point on the retina. The scanning optics move the focused spot across the retina in a raster pattern.

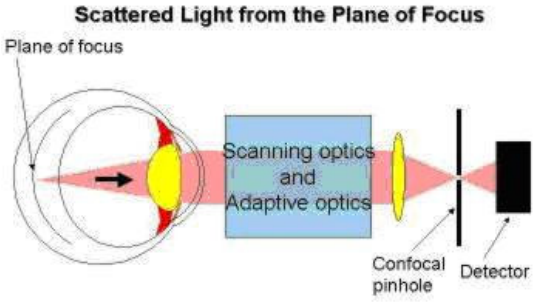


Fig. 3 Detection The confocal pinhole is conjugate to the plane of focus on the retina, so light that scatters from the plane of focus passes the through the pinhole and is detected by the photomultiplier.

Optics of the eye

Principle

There are a number of refracting surfaces in the eye, with the important ones the anterior surface of the cornea, and the lens.

Of these the cornea, because of the large difference in refractive index between air (1.0) and corneal tissue (1.37) is the more powerful, with a typical power of about 40 dioptre (dptr). The lens in the relaxed (not accommodated) state has a power of about 17 dptr. Accommodation may increase this, by about 14 dptr in children, less with increasing age.

These several components may, by approximation, be thought of as equivalent to a single ideal lens. The resulting simplification is the *reduced eye* (see Fig. 1).
Location and strength of the ideal lens vary in literature but the following is a good approximation. Its principal plane is situated just behind the iris and so its distance to the retina (the axis length) is 20 mm, and its power is 50 dptr.

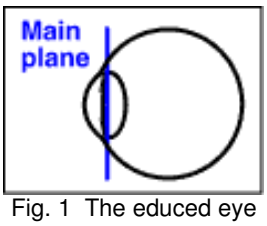


Fig. 1 The educed eye

Note that 50 dptr is less than the sum of the powers of cornea and lens. They are not close enough together for their powers to be additive. In the normal, resting reduced eye, somewhat arbitrarily, 36 dptr are due to the cornea and 14 dptr to the lens.
Both the axis length and the principal plane power are abstractions. However, the quantity power - 1/axis length (in m), the "net power" of the reduced eye, is concrete and measurable. The eye will see a sharp image when the image is focused on the retina, i.e. when the image distance is equal to the axis length. Replacing axis length by image distance and applying the lens formula (see Light: the ideal and non-ideal lens), under the condition of a sharply seen object, we find that the net power (in dptr, m^{-1}) is:
$\varphi = 1/f = 1/d_{object}$
The difference $1/d_{object}$ – 1/axis length directly gives the correction. Negativity means myopia and a negative lens equal to the absolute difference. With this correction the myopic eye will see clearly at infinite distance (the Foucault's principle, see Retinoscopy).

More info

More precisely the eye should by presented as a system with two principal planes (see The Ideal and non-ideal lens), as illustrated in Fig. 2. It appears that the two principal planes are only 0.3 mm from each other. This is caused by the small differences between most eye media and the small refractive contribution of the lens. Therefore, for most applications the reduced eye with one principal plane is adequate.

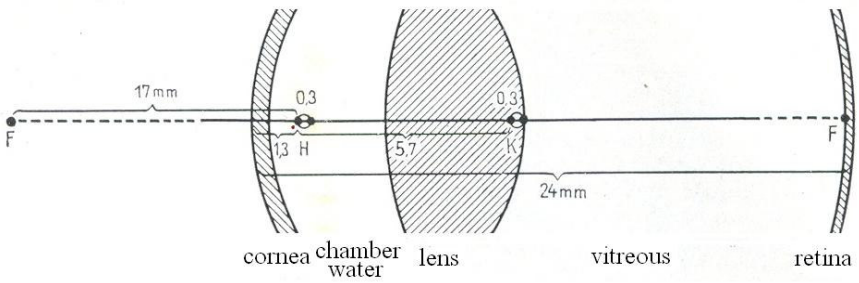


Fig. 2 The reduced eye with two principal planes.

The power of the reduced eye of Fig. 2 is:

$\varphi = 1/f = 1/0.017 = 58.8$ dptr or
$\varphi = 1.33/(0.024 - 0.0013 - 0.0003) = 59.4$ dptr, (1)

where 1.33 is the mean refractive index of the eye.

Optometry

Principle

Optometry is the health care profession concerned with examination, diagnosis, and treatment of the eyes and related structures with the objective to correct vision using lenses and other optical aids (see also Retinoscopy and Ophthalmic Corrections).

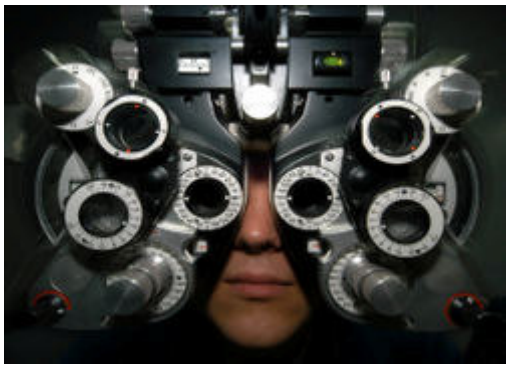


Fig. 1 An optical refractor, also called a phoropter in use.

More info

An optometrist (or opticians or optometric physicians) is a care practitioner for most vision and ocular health concerns, including, but not limited to, fitting and prescribing glasses and contact lenses, and in some countries diagnosing and treating (excluding surgery) muscular abnormalities, treating minor ocular injuries, diagnosing and treating diseases such as glaucoma and diagnosing others such as diabetic retinopathy.

Retinoscopy

Principle

Retinoscopy is a technique to obtain an objective measurement of the refractive condition of a patient's eye (see also Optics of the eye and Ophthalmic Corrections). The examiner uses a retinoscope to shine light into the patient's eye and observes the reflection off the patient's retina. While moving the streak or spot of light across the pupil the examiner observes the relative movement of the reflection. Next, he uses a phoropter (a large instrument to measure an individual's refractive error in order to determine the eyeglass prescription, see Optometry) or manually places lenses over the eye to "neutralize" the reflection.

Application

Retinoscopy is especially useful in prescribing corrective lenses for patients who are unable to undergo a subjective refraction that requires a judgment and response from the patient. It is also used to evaluate accommodative ability of the eye and detect latent hyperopia.

More info

Retinoscopy works on a principle called Foucault's principle. Basically it indicates that the examiner should simulate the infinity to obtain the correct refractive power. Hence a power corresponding to the working distance is subtracted from the gross retinoscope value.
Static retinoscopy is performed when the patient does not accommodate. Dynamic retinoscopy is performed when the patient has active accommodation from viewing a near target.

Stevens' power law

Principle

Whereas the classical Weber–Fechner law focuses on the just perceivable difference between two sensory stimuli, Stevens' power law is giving the relationship between the magnitude of a physical stimulus S and its perceived intensity or strength R.
Different sensory modalities varied too much in "steepness" to be fitted by the Webner-Fechner law. A better description of the experiments lead to the general equation:

$$R = k(S-S_0)^{\alpha}, \quad (1)$$

where S_0) is the threshold below which there is no sensation, k a constant and α a power (after which the low is mentioned), both dependent on the experimental paradigm. For $S-S_0 = 0$ the intersection with the vertical axis is gives by k. However, this is purely calculation. Actually the behavior for low S is asymptotic to R=1.
The equation is generally written and visualized with logarithms of R and S:

$$\log R = \alpha \log(S-S_0) + \log k, \quad (2)$$

.

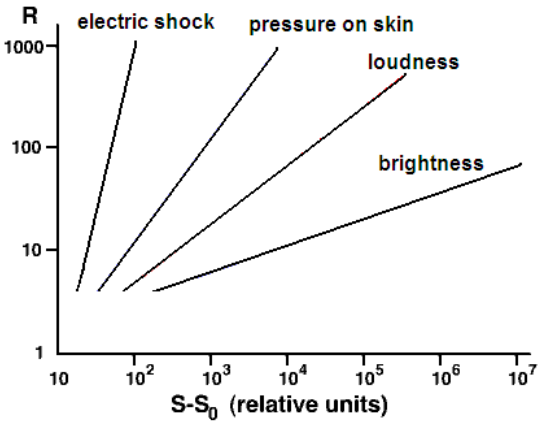


Fig. 1 $R/(S-S_0)$ relationship for various perceptions with α the slope and k the intersection with the vertical axis.

Fig. 1 gives a rough idea of the law. Pain by an electric shock at the finger tips has a high value of α (about 3.5), reflected in a steep curve. Once the shock is strong enough to elicit pain, the pain rapidly becomes stronger as the stimulus becomes stronger. Pain saturates fast. Brightness perception (in this case central vision in the dark) covers a scotopic range of about 6 log units. Therefore α should be small since the working range of the neural system is between 2 and 3 log units. The α for scotopic central brightness vision appears to be about 0.33.

The principal methods to measure the perceived intensity of a stimulus are *magnitude estimation* and *magnitude production*. In magnitude estimation with a standard, the experimenter presents a stimulus called a *standard* and assigns it a number called the *modulus*. For subsequent stimuli, subjects report numerically their perceived intensity relative to the standard so as to preserve the ratio between the sensations and the numerical estimates (e.g., a sound perceived twice as loud as the standard should be given a number twice the modulus). In magnitude estimation without a standard (usually just *magnitude estimation*), subjects are free to choose their own standard, assigning any number to the first stimulus and all subsequent ones with the only requirement being that the ratio between sensations and numbers is preserved. In magnitude production a number and a reference stimulus is given and subjects produce a stimulus that is perceived as that number times the reference. Also used is *cross-modality matching*, which generally involves subjects altering the magnitude of one physical quantity, such as the brightness of a light, so that its perceived intensity is equal to the perceived intensity of another type of quantity, such as warmth or pressure.

Application

This is mainly in experimental psychology, clinical psychology and psychophysics.

More Info

Hearing
The logarithmic perception of stimulus strength is expressed in the decibel (dB) scale of sound intensity (see Sound and Acoustics). And another is pitch, which, however, differs from the other cases in that the physical quantity involved is not a "strength".
Humans hear pitch in a logarithmic or geometric ratio-based fashion: for notes spaced equally apart to the human ear, the frequencies are related by a multiplicative factor. For instance, the frequencies of corresponding notes of adjacent octaves differ by a factor of 2. Similarly, the perceived difference in pitch between 100 Hz and 150 Hz is the same as between 1000 Hz and 1500 Hz. Musical scales are always based on geometric relationships for this reason. Notation and theory about music often refers to pitch intervals in an additive way, which makes sense if one considers the logarithms of the frequencies, as loudness. The Weber-Fechner law is a good approximation for higher loudness' but not for lower ones.
Vision
In man and other animals, brightness is perceived logarithmically. Brightness of stars is measured logarithmically.

Once Stevens' formula was established in psychophysics, it also got popular for describing results of neurophysiological experiments relating stimulus intensity to *objective* responses, i.e. neurophysiological responses such as frequency of actions potentials
Table 1 lists the exponents reported by Stevens.

Table 1 Exponents α as found by Stevens

Continuum	Exponent (***a***)	Stimulus condition
Loudness	0.67	Sound pressure of 3000 Hz tone
Vibration	0.95	Amplitude of 60 Hz on finger
Vibration	0.6	Amplitude of 250 Hz on finger
Brightness	0.33	5° target in dark
Brightness	0.5	Point source
Brightness	0.5	Brief flash
Brightness	1	Point source briefly flashed
Lightness	1.2	Reflectance of gray papers
Visual length	1	Projected line
Visual area	0.7	Projected square
Redness (saturation)	1.7	Red-gray mixture
Taste	1.3	Sucrose
Taste	1.4	Salt
Taste	0.8	Saccharine
Smell	0.6	Heptane
Cold	1	Metal contact on arm
Warmth	1.6	Metal contact on arm
Warmth	1.3	Irradiation of skin, small area
Warmth	0.7	Irradiation of skin, large area
Discomfort, cold	1.7	Whole body irradiation
Discomfort, warm	0.7	Whole body irradiation
Thermal pain	1	Radiant heat on skin
Tactual roughness	1.5	Rubbing emery cloths
Tactual hardness	0.8	Squeezing rubber
Finger span	1.3	Thickness of blocks

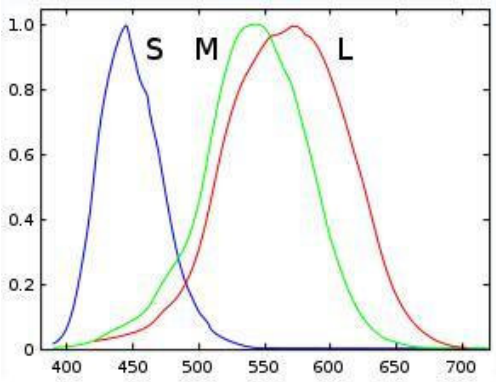
Pressure on palm	1.1	Static force on skin
Muscle force	1.7	Static contractions
Heaviness	1.45	Lifted weights
Viscosity	0.42	Stirring silicone fluids
Electric shock	3.5	Current through fingers
Vocal effort	1.1	Vocal sound pressure
Angular acceleration	1.4	5 s rotation
Duration	1.1	White noise stimuli

Vision of color

Principle

Color vision is the capacity to distinguish objects based on the wavelengths (or frequencies) of the light they reflect or emit. The nervous system derives color by comparing the responses to light from the several types of cone photoreceptors in the eye. These cone photoreceptors are sensitive to different portions of the visible spectrum, for humans from about 380 to 740 nm. The visible range and number of cone types differ between species.

The cone cells of vertebrates contain pigments with different spectral sensitivities. In most primates closely related to humans there are three types of cones, allowing trichromatic color vision. Hence, these primates, like humans, are known as trichromats. Other mammals are mostly dichromats, and many mammals have little or no color vision. Bony fish are di- tri or tetrachromats.
The cones are conventionally labeled according to the peak wavelengths of their spectral sensitivities: short (S or blue), medium (M or green), and long (L or red) cones, see Fig. 1.
The peak response of human color receptors varies, even amongst individuals with 'normal' color vision; in non-human species, this polymorphic variation is even greater, and it may well be even adaptive to the season as in several bony fish species. In human, within a cone type subtypes can be distinguished with different wavelengths. This is individually specific.



Cone cells in the human eye

Cone type	Name	Range	Peak wavelength
S	β	400–500 nm	420–440 nm
M	γ	450–630 nm	534–545 nm
L	ρ	500–700 nm	564–580 nm

Fig. 1 Left: normalized response spectra of human cones, S, M, and L types, to monochromatic spectral stimuli, with wavelength given in nm. Right: their range and peak wavelengths.

Scotopic (night), twilight) mesopic) and photopic (day) vision
In low light levels, scotopic levels, vision is mediated by rod cells en the visual percept is in grays, since only brightness remains. Human rods are maximally sensitive to 500 nm.
With daylight or photopic vision the cone cells of the retina mediate color perception. Now, the eye is most sensitive to wavelengths near 555 nm. The region between these regions is known as mesopic vision (range 0.05 – 8 cd/m^2), in which case both rods and cones are providing a meaningful signal. However, rods contribute a little up to about 200 cd/m^2.

Wavelengths, colors and white light
The characteristic colors are, in order from short to long wavelength: violet, blue, green, yellow, orange and red. Sufficient differences in wavelength give rise to a difference in perceived color. The just noticeable difference in wavelength varies from about 1 nm in the blue-green and orange-yellow (smallest) wavelengths, to 10 nm and more in the red and blue. Though the eye can distinguish up to a few hundred colors, when those pure spectral colors are mixed together or diluted with white light, the number of distinguishable hues (see for definition below) is some millions.

The perception of "white" is formed by seeing the entire spectrum of visible light, or by mixing a few monochromatic light beams, such as red, green, and blue, or even by mixing just a pair of such beams as blue and yellow. The pure colors of such a pair are called opponent or complementary colors. In paintings the use of the opponent pairs red-green and yellow-blue can make these paintings very attractive, as those of the Vincent van Gogh.

Theories of color vision
Two complementary theories of color vision are the trichromatic theory and the opponent color theory. The trichromatic theory, states that the retinal types of cones are preferentially sensitive to red, green,

and blue. The opponent process theory states that the visual system interprets color in an antagonistic way: red vs. green and blue vs. yellow. Both theories appeared to be correct. The first describes a very early stage of processing in the retina and the second a more central stage. A range of wavelengths stimulates each cone type to varying degrees. Yellowish-green light, for example, stimulates both L and M cones equally strongly, but only stimulates S-cones weakly and violet light stimulates almost exclusively S-cones. The brain combines the information coming from the retina to give rise to different perceptions of different wavelengths of light.
The pigments present in the L and M cones are encoded on the X chromosome. Defective encoding leads to one of the many types of color deficiencies or anomalies. The two most common forms of color blindness are protanopia (lacking the L cone) and deuteranopia (lacking the M cone).

Color vision in animals and evolution of color vision
Various vertebrates, such as birds and bony fish, have often more complex color vision systems than humans. In birds, tetrachromacy is achieved through up to four cone types. Brightly colored oil droplets inside the cones can produce a new cone type by shifting the spectral sensitivity of the cell. Eutherian mammals other than primates often are dichromats. Marine mammals have only a single cone type and are thus monochromats. Many invertebrates have color vision. Insects can have trichromacy with insensitivity to red but sensitivity to UV, and others are tetrachromats.
Color perception mechanisms are highly dependent on evolutionary factors, of which the most prominent is thought to be satisfactory recognition of food sources (for instance herbivorous primates look for mature fruit and leaves). The evolution of trichromatic color vision in primates occurred as the ancestors of modern monkeys, apes, and humans switched to diurnal (daytime) activity and began consuming fruits and leaves. Various species of birds, turtles, lizards, fish and insects have UV receptors in their retinas (or compound eyes). These animals can see the UV patterns of environmental light which may be used for orientation and homing. UV can also by used for recognizing prey and flowers.

Fig. 2 Color while for mixing paints. Brown is made by mixing green with red. The darker the more red paint is added. From ref. 1.

Application

The theory of color vision is of great importance for all kind of daily life application, such as lighting rooms, color television and all kind of electronic visual applications, making paints etc. For these purposes, one uses often the so called Color Wheel (see Fig. 2).

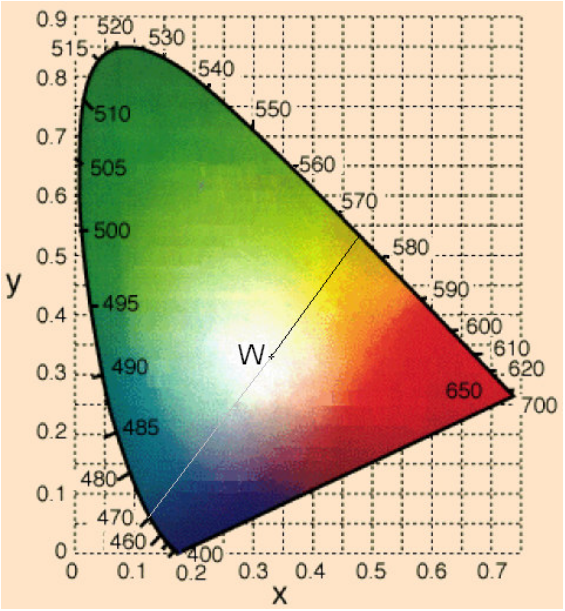


Fig. 3 The 1931 CIE colors on the chromaticity diagram. Modified after ref. 1. The black&gray line connects the complementary wavelengths 470 and 575 nm.

More Info

Wavelengths, colors and hues
Colors or better to say light of different wavelengths can be added but not subtracted. Adding is mixing. Now a new color, this means a new percept arises. E.g. mixing a red wavelength with a green one results in a yellow color. Adding two opponent wavelengths gives, by definition, a white percept. There are infinite opponent pairs of wavelengths. Such pairs can be found in the CIE color triangle or chromaticity diagram (Fig. 3). This gives, roughly speaking, along the x-axis the amount of a virtual purple like color and along the y-axis the amount of a virtual, nearly green of nearly 515 nm. The CIE color triangle is horseshoe-shaped, with its curved edge corresponding to all spectral wavelengths (the *spectral locus*), and the remaining straight edge corresponding to purple colors, which cannot be described with a single wavelength. It is mixtures of deep red and violet. In the middle, at the point (1/3,1/3) the white point W is found. Going in a straight line from W to the outer border, in the diagram to 575 nm, the color on the line remains the same but the hue differs. Hue is also given by the saturation. This goes from unsaturated (0%) to 100% saturated at the border. It can be made by taking 575 nm light and adding some 460 nm (the opponent color), which is located in at the end of the grey line being I line with the black one. , since that is. Finally hue is defines with a third parameter, the brightness. This can be seen as a third dimension perpendicular o the plane of the triangle.
Since the human eye has three types of cones, a full plot of all visible colors is a 3D figure. However, the concept of color can be divided into two parts: brightness and chromaticity. For example, the color white is a bright color, while the color grey is considered to be a less bright version of that same white. In other words, the chromaticity of white and grey are the same while their brightness differs.
Brown colors cannot be found in the triangle, since they arise not with mixing but with subtracting. Subtraction occurs when mixing paint. Adding beams with a red, green and blue wavelength can give white, but mixing paints of these colors will give some brownish gray paint. The explanation is that a red paint absorbs all wavelengths except some red wavelength band.
For a mathematical description of the color triangle and mixing colors, one should read Vision of color: mathematical description.

Chromatic adaptation
An object may be viewed under various conditions. For example, it may be illuminated by sunlight or electric light. In all situations of photopic vision, the object has the same color percept: On the other hand, a camera with no adjustment for light may register varying color. This feature of the visual system

is called chromatic adaptation, or color constancy; when the correction occurs in a camera it is referred to as white balance.
Chromatic adaptation is one aspect of vision that may fool someone into observing a color-based optical illusion (a perceived image that differs from objective reality).

References
http://hyperphysics.phy-astr.gsu.edu/hbase/vision/cie.html

Vision of color: mathematical description

Principle

For the phenomenology of color vision it holds that a color of a beam of light can always be imitated by mixing in the correct way three beams of light which are given as standard beams. Here we take as standards (according to Wright) the units beams:
1 unit red (R) as 0.65 lumen 650 nm light;
1 unit green (G) as 1 lumen 530 nm light;
1 unit red (B) as 0.044 lumen 460 nm light.

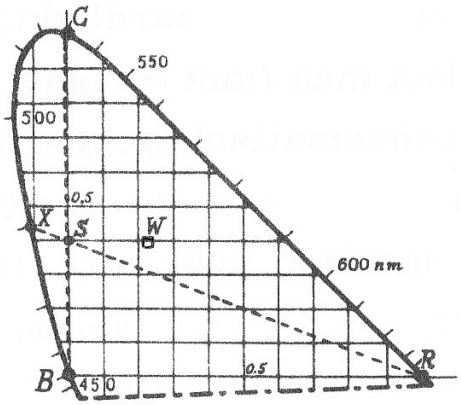


Fig. 1 Color 'triangle' of Wright. W is the white point.

It appeared that a white percept is achieved with the mix of W = 0.243R + 0.410G +0.347B The 3 numerical coefficients (we call them r, g and b) are the coordinates of W. The three units are such chosen that their sum is exactly 1. They yield one point in the 3D space RGB. When the coefficients are changed proportionally, the color remains the same. It only changes its brightness. So, actually color depends on two ratios of the three coefficients. Hence color can be presented in a plane (2D). The convention is to express along the horizontal axis (x) the ratio r/(r+g+b) and along the vertical axis (y) the ratio g/(r+g+b). Then, W is described with x = 0.243 and y = 0.410 and accordingly given in Fig. 1. (Notice that any set of three colors which when added in appropriate combination will yield white can be considered to be primary colors.)

In the color 'triangle' of Wright, B is in the origin and all wavelengths are located of a curved edge corresponding. By a mixture of R, G and B all points in the geometric triangle RGB can be obtained. Within the triangle RGB, by applying the rules of mixing of two colors, colors can be obtained which are all located at the line which connects both points of the two colors. However, the colors of the wavelengths, expect those of R, G and B, are located outside the rectangular triangle RGB. Hence, the wavelengths themselves cannot by obtained by adding. The following example explains that. With point X with 490, point S is described by S = 0.9X + 0.1R. Since S is also on the line GB it holds that S = 0.4G + 0.6B. Combining both yields X = -0.11R + 0.44G + 0.67B. Hence x = -0.11 and y = 0.44 and we allow negative coefficients.
To avoid negative coefficients a more appropriate diagram is the 1931 CIE (Commission Internationale d'Eclairage) chromaticity diagram of Fig. 2.

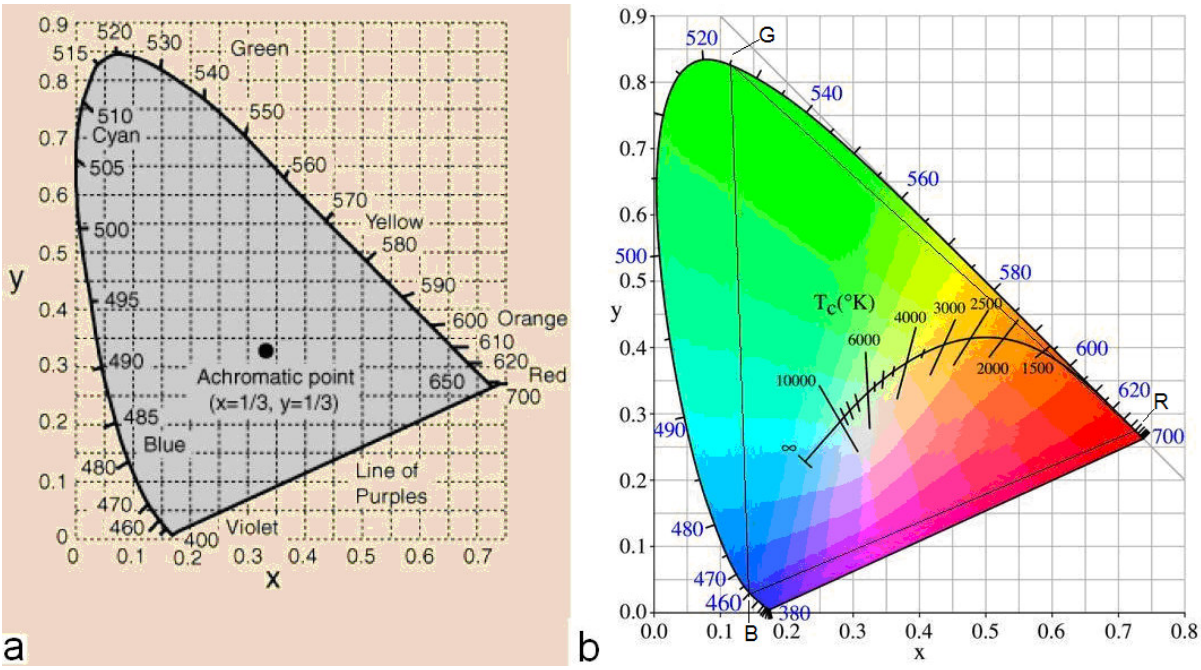


Fig. 2 a,b The 1931 CIE colors on the chromaticity diagram. b The curve gives the temperatures of black body radiation (see Planck's law), which crosses W at about 6000 K. R, G and B indicate the wavelengths of Wrights choice of the primary colors. Modified after ref. 1.

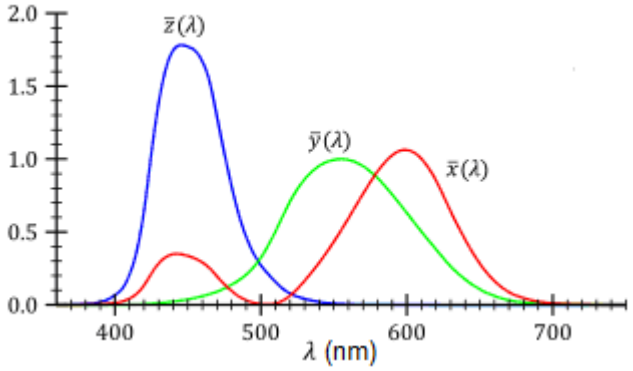


Fig. 3 The CIE standard observer color matching functions. Y is the human photopic luminosity function. Notice that the three functions look like the three cone spectral sensitivity curves. However, they are not the same.

Application

The CIE xyY color space (see below) is widely used to specify hues in practice.

More Info

The CIE 1931 system characterizes colors by a luminance parameter Y and two color coordinates x and y which specify the point on the chromaticity diagram. The CIE derived a new set of *color-matching functions* using the blue, green, and red primaries in both positive and negative combinations. This new set of primaries of the C.I.E. 1931 Standard Observer, called $\bar{x}(\lambda), \bar{y}(\lambda)$ and $\bar{z}(\lambda)$, depicted in Fig. 3, have the following properties:

They always produce positive tristimulus values X, Y and Z.
It is possible to represent any color or hue (x and y) in terms of these primaries.
They were derived so that equal values (being 1/3) of x and y produce white.
They were arranged so that a single parameter Y determines the luminance of the hue.

The three CIE *color matching functions* called $\bar{x}(\lambda), \bar{y}(\lambda)$ and $\bar{z}(\lambda)$ are considered as the spectral sensitivity curves of three linear light detectors that yield the CIE XYZ tristimulus values *X*, *Y*, and *Z*. X and Z comprises the color information (finally given by x and y) and Y luminance.
The tristimulus values for a color with a spectral power distribution *I*(λ) are given in terms of the standard observer by:

$$X = \int_0^\infty I(\lambda)\bar{x}(\lambda)d\lambda$$

$$Y = \int_0^\infty I(\lambda)\bar{y}(\lambda)d\lambda$$

$$Z = \int_0^\infty I(\lambda)\bar{z}(\lambda)d\lambda$$

where λ is the wavelength of the equivalent monochromatic light (in nm).
Other methods, hence with other standard observers, such as for the CIE RGB space and other RGB color spaces, are defined by other sets of three color-matching functions (e.g. CIE RGB with $\bar{r}(\lambda), \bar{g}(\lambda)$ and $\bar{b}(\lambda)$, and lead to tristimulus values in those other spaces. The former leads to the CIE rg chromaticity diagram and the CIE *rgb* color space

Since the human eye has three types of color sensors that respond to different ranges of wavelengths, a full plot of all visible colors is a three-dimensional figure. However, the concept of color can be divided into two parts: brightness and chromaticity. For example, the color white is a bright color, while the color grey is considered to be a less bright version of that same white. In other words, the chromaticity of white and grey are the same while their brightness differs.
The CIE XYZ color space was deliberately designed so that the *Y* parameter was a measure of the brightness or luminance of a color.
x = X/(X + Y + Z),
y = Y/(X + Y + Z),
z = Z/(X + Y + Z) = 1 –x - y,

The derived color space specified by x, y, and Y is known as the CIE xyY color space.
Note that the chromaticity diagram is a tool to specify how the human eye will experience light with a given spectrum. It cannot specify colors of objects (or printing inks), since the chromaticity observed while looking at an object depends on the light source as well.

The chromaticity diagram illustrates a number of interesting properties of the CIE XYZ color space:
It is seen that all visible hues (i.e. chromaticities) correspond to non-negative values of *x*, *y*, and *z* (and therefore to non-negative values of *X*, *Y*, and *Z*).
An equal mixture of two equally bright colors will not generally lie on the midpoint of that line segment. In more general terms, a distance on the xy chromaticity diagram does not correspond to the degree of difference between two colors.
It can be seen that, given three real sources, i.e. pure wavelenghts, these sources cannot cover the gamut of human vision. Geometrically stated, there are no three points within the gamut that form a triangle, as illustrated in the geometric triangle of Fig. 2b. This includes the entire gamut; or more simply, the gamut of human vision is not a triangle.

Literature

Kaiser, Peter K.; Boynton, R.M. (1996). Human Color Vision (2nd ed.). Washington, DC: Optical Society of America. ISBN 1-55752-461-0.
Vos JJ. From lower to higher colour metrics: a historical account. Clin Exp Optom. 2006 Nov;89(6):348-60. Review.
Wyszecki, Günther; Stiles, W.S. (1982). Color Science: Concepts and Methods, Quantitative Data and Formulae (2nd ed.). New York: Wiley Series in Pur and Applied Optics. ISBN 0-471-02106-7.

Vision of luminosity

Principle

The human luminosity function describes the average visual sensitivity of the human eye to light of different wavelengths. It is a standard function established by the Commission Internationale de

l'Éclairage (CIE) and may be used to convert radiant energy into luminous energy, which is a measure of visible sensitivity.

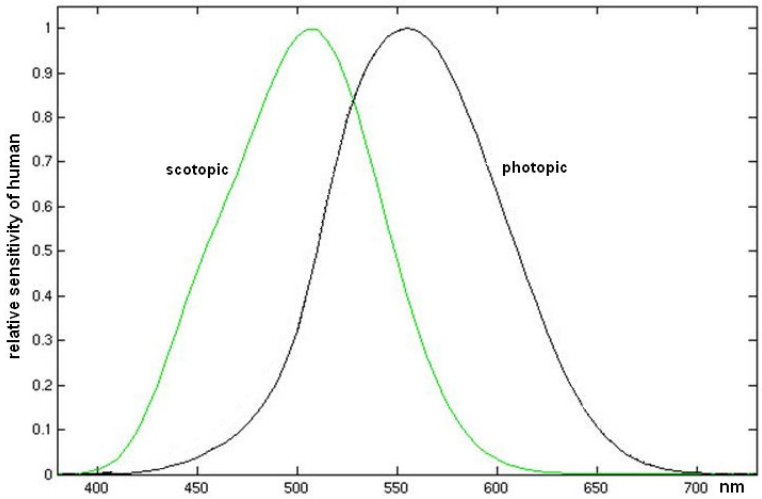


Fig. 1 The (dimensionless) photopic luminosity function (CIE 1931) and scotopic luminosity functions (CIE 1951).

There is a luminosity functions for day-light (photopic) levels, V(λ) or $\bar{y}(\lambda)$, and one for night vision (scotopic levels), V'(λ), see Fig. 1.

Application

The luminosity function (especially the photopic one) is of great importance for the lighting industry.

More info

The photopic luminosity curve also forms the middle color matching function $\bar{y}(\lambda)$ in the CIE 1931 color space (see Vision of color: mathematical description).

Photopic luminosity
The *luminous flux* (or visible energy) in a light source is defined by the photopic luminosity function. (See for photometric and radiometric units Light: units of measure.) The total luminous flux in a source of light is:

$$F = k\int_0^\infty V(\lambda)J(\lambda)d\lambda$$

Where:

- k is a proportionality constant, being 683 lm/W,
- F luminous flux (in lumens, lm),
- J (λ) the spectral power distribution of radiation (power per unit wavelength), in W/m).
- λ is wavelength (in m).

The number 683 comes from the 1979 definition of the candela, the SI unit of luminous intensity (See Light: units of measure.)

Scotopic luminosity
For scotopic vision the sensitivity of the eye is mediated by rods, and shifts toward about 507 nm for young eyes; the sensitivity is equivalent to 1700 lm/W at this peak wavelength.
The standard scotopic luminosity function, also called V'(λ) was adopted by the CIE in 1951.

Reference

Vos, J. J. (1978). "Colorimetric and photometric properties of a 2° fundamental observer". Color Research and Application 3 (3): 125–128. doi:10.1002/col.5080030309.

Visual acuity

Principle

Visual acuity (VA) is acuteness or clearness of vision, especially form vision, which is dependent on the sharpness of the retinal focus within the eye, the sensitivity of the nervous elements, and the interpretative faculty of the brain. It is also dependent on the stimulus and lightning conditions. Clinically, VA is a quantitative measure of the ability to identify black symbols on a white background at a standardised distance as the size of the symbols is varied. It represents the smallest size that can be reliably identified.
If the eye watches two objects, for example two black dots, it will see them under a certain angle. Obviously, if this angle gets too small, the eye cannot distinguish them from each other. The smallest angle that still allows the eye to see that there are two dots is the *resolution angle*. VA is defined as the 1/resolution angle (in minutes of arc. One minute, 1', of arc is 1/60^{o}). The "typical" human resolution angle is 1'. The VA-test is the most common clinical test of visual function since it easily detects a resolution impairment in normal daily vision.

Application

Traditional the Snellen chart (Fig. 1) is used for VA testing. This chart (and any other VA-chart) should be used under certain specified conditions of lighting etc. at a specific distance. (For more details about the method and other details of measurement see e.g. the VA chapter of Wikipedia.)

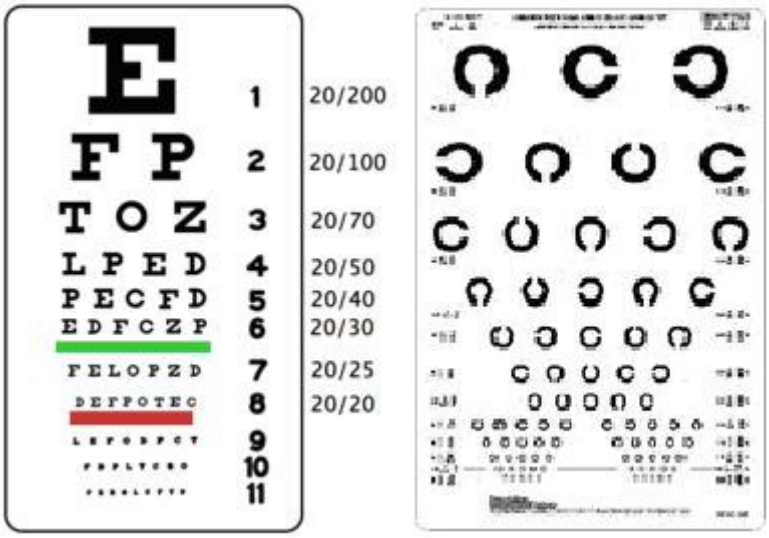


Fig. 1 Snellen (left) and Landolt C chart

Another frequently used and better chart is the Landolt C chart. The broken ring optotype is made with a "C" like figure in a 5 x 5 grid that subtends 5' (standard size) and an opening measuring 1'. The subjects have to indicate the orientation of the gap (left, right, up, down, see Fig. 2). The advantage is that all C's of the same size are equally recognisable. This chart is preferred for laboratory experiments and it can also be used for illiterates.
Other charts are the Sloan's chart and the Bailey&Lovie chart. The Lea chart (figures like apple, house, etc.) and Tumbling E chart (the letter E in normal, mirror, turned 90^{o} left en 90^{o} right) are also useful for illiterate people.

More Info

The eye media (tear film, cornea, anterior chamber, pupil, lens, vitreous, and finally the retina) affect the quality of the image. The retinal pigment epithelium is responsible for, among other things, absorbing light that crosses the retina to prevent backward scatter.
To resolve fine details, the eye's optical system has to project a focused image on the fovea. The fovea has the highest density of cone photoreceptors and no rods, thus having the highest resolution and best colour vision (but no night vision). VA and colour vision, both based on the same receptor cells, are

different but unrelated physiologic functions as also holds for other function such as reduced contrast, or inability to track fast moving objects.
Along the visual pathway, after the passing the lateral geniculate body, a relay station, the visual cortex, the posterior (occipital) part of the cortex, is the first major centre for visual processing. The central 10° of the visual field (approximately the extension of the macula) is represented by at least 60% of the visual cortex. Much of these neurones are believed to be involved directly into VA processing.

Table 1 Visual acuity scales

Foot	Metre	Decimal	LogMAR
20/200	6/60	0.10	1.0
20/160	6/48	0.13	0.9
20/125	6/37	0.16	0.8
20/100	6/30	0.20	0.7
20/80	6/24	0.25	0.6
20/63	6/18	0.32	0.5
20/50	6/15	0.40	0.4
20/40	6/12	0.50	0.3
20/32	6/9	0.63	0.2
20/25	6/7	0.80	0.1
20/20	6/6	1.00	0.0
20/16	6/4.8	1.25	-0.1
20/12.5	6/3.75	1.60	-0.2
20/10	6/3	2.00	-0.3

VA is expressed as a vulgar fraction (m/n) or as a decimal number (see Table 1 for conversions). Using the foot, VA is expressed relative to 20/20, the normal value. Otherwise, using the metre, the equivalent is expressed relative to 6/6. In the decimal system, the acuity is defined as the reciprocal value of the size of the gap (in arc minutes) of the smallest Landolt C that can be reliably identified. A value of 1.0 is equal to 20/20.
LogMAR is another commonly used scale, which is expressed as the logarithm of the resolution angle. It is a conversion to a linear scale. Positive values indicate vision loss, while negative values denote normal or better VA. This scale is mainly used in basic research.

With normal eyesight (VA is 6/6) one can see can see detail from 6 m away and with VA is 6/12 one can see detail from 6 m away as a subjects with VA is 6/6 sees details from 12 m away. So with 6/12 acuity is halved.
In humans, the maximum acuity of a healthy, emmetropic eye (and even ametropic eyes with correctors) is approximately 6/5 to 6/3.6 and6/6 is considered as the lower limit of normal. Maximum VA without visual aids (such as binoculars) is around 6/3. In case of hyperopia or myopia, the eye should be corrected for the viewing distance. Some birds, such as hawks, have acuity about 5 times maximal human VA.

VA is typically measured monocularly rather than binocularly. In some cases, binocular VA will be measured. Usually binocular VA is slightly better than monocular VA.
"Dynamic VA" is VA for a moving object.
VA can also be measured as a function of eccentricity in the various directions. It rapidly falls off with eccentricity.

References
http://www.neuro.uu.se/fysiologi/gu/nbb/lectures/VisAcuity.html

Weber-Fechner law

Principle

Subjective versus perceived intensity
Psychophysics tries to relate qualitatively the objective stimulus intensity (S) and the subjective perception of it (R). (Psychophysics is a subdiscipline of psychology dealing with the relationship between physical stimuli and their subjective correlates or percepts.) The stimulus is generally a sensory stimulus and the latter the response of the subject. The neurophysiological process in between is not taken into account. In addition to S and R, for the description we also need a threshold intensity S_0. Below S_0 the stimulus is not perceived at all because of intrinsic neural noise (the "dark light") in the system. In other words, S_0 is the absolute threshold.

A number of equations to describe the relation between S and R have been proposed on the basis of experiments in which people have been subjected to some kind of sensory stimulus and asked to somehow quantify their perception of it.
The Weber-Fechner law attempts to describe quantitatively the relationship between two physical magnitudes of stimuli with a just perceived difference between the two. Another psychophysical law, Stevens' power law, is especially aimed to quantify the extent of perception related to the physical magnitude of the stimulus.
With a Weber-Fechner test a subject is presented two nearly identical stimuli (for example, two weights) and tested whether he can notice a difference between them. The smallest noticeable difference seems to be roughly proportional to the intensity of the stimulus. For instance, if a person could consistently feel that a 110 g weight was heavier than a 100 g weight, he could also feel that 1100 g was more than 1000 g. Consequently, the mathematical description is:

$\Delta S/S = k$, or: (1a)
$\log\Delta S = \log k + \log S$, (1b)

where k is the smallest fraction to perceive a just noticeable difference (JND), also called the Weber fraction or Weber constant. The law is implicitly based on the concept that the magnitude of a subjective sensation increases proportional to the logarithm of the stimulus intensity S. This means that JNDs are additive.

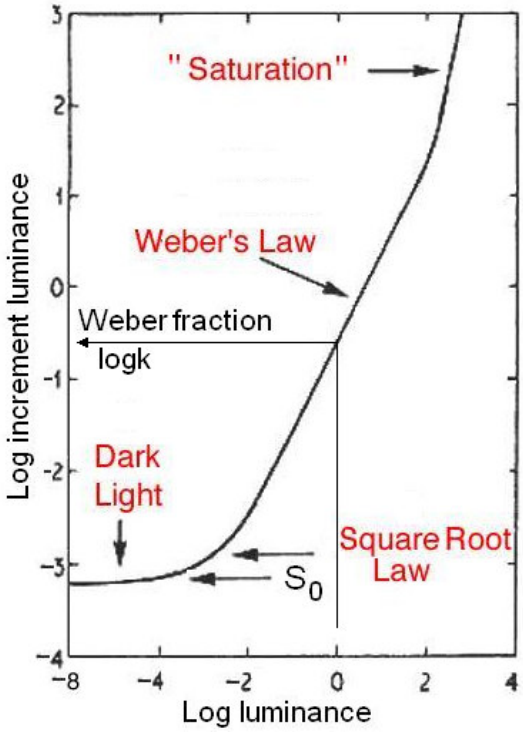


Fig. 2 The Weber-Fechner law for a JND of luminance for the scotopic system (rod vision) and its deviations for very small and very large stimuli.

The relationships are valid for all kind of sensory perceptions with k dependent on the stimulus paradigm.
The law only holds for a restricted range of S (see **More Info**), as is visualized in Fig. 1.

Application

The law is applied in experimental (clinical) psychophysical research of sensory systems. It is also practiced for particular clinical sensory tests.
ΔS/S (or ΔL/L when we talk about luminance) plots are most convenient to obtain insight in the regular or abnormal performance of a sensory system.

More info

Hearing
The JND of pitch of pure tones of equal loudness has a Weber fraction of 0.003. This specific relation is called Knudson's law. With an audible frequency range of 20 Hz to 20 kHz, there are 2000 tones whose frequency difference is audible.

Vision
As told, the law only holds for a restricted range of S. The strength of the perception increases with S. However, for very high S it does not increase anymore and consequently, k increases (theoretically) to infinite. This means that the curve of Fig. 1 becomes vertical: the system is saturated. Also for very low S, there is a strong deviation. Fig. 1 visualizes these deviations for a light stimulus (log luminance L; horizontal axis) which is incremented until the threshold for perception is reached (ΔL; vertical axis).

The JND in the horizontal part of the JND versus S curve is determined by S_0, since the luminance is relatively low and does not significantly affect JND. By taking the effect of S_0 into account Eq. (1b) becomes:

$$\log\Delta L = \log k + \log(L - L_0), \qquad (3a)$$

where L displaces S. This allows a slow transition from the horizontal part of the curve to the Weber part.
The second part of the JND/S curve is called the square root law or de Vries-Rose law region. This part of the curve is limited by quantal fluctuation in the background. The visual system is almost an ideal detector which can detect and encode each absorbed quantum of light. It is limited only by the noise due to quantal fluctuations in L. ΔL can only be detected when it sufficiently exceeds the fluctuations in L. With log ΔL plotted versus log L, the Weber law section ideally has a slope 1. This means that ΔL and L linearly plotted gives a straight line. For the rod pathway, a slope 0.8 or less is found. However, now the lin-lin plot of ΔL and L is downward concave like a square root. With the slope > 1, the curve is concave upward, like a quadratic function. The variability of slope (α) can be expressed in the equation as:

$$\log\Delta L = \log k + \alpha\log(L - L_0). \qquad (3b)$$

The slope is actually the extension of the law to Steven's power law.
The Weber section of the curve demonstrates an important aspect of our visual system which is designed to distinguish objects from its background. In the real world, objects have contrast, which is constant and independent of ambient luminance. The perception of contrast, which is governed by Weber's law leads to the concept of contrast constancy or contrast invariance. The Weber fraction for the rod and cone system is 0.14 and about 0.025 respectively. The last section of the curve shows "saturation": the rods are saturated at high luminance and now the cone system dominates the threshold mechanism. Real saturation has a slope of infinite.

Index