A

Absorption Coefficient

F. L. Galeener[†]

The absorption coefficient α measures the spatial decrease in intensity of a propagating beam of waves or particles due to progressive conversion of the beam into different forms of energy or matter. Absorption usually implies the creation of some form of internal energy in the traversed medium, e. g., the production of heat; however, it may also be associated with other inelastic scattering events, such as the ultimate conversion of incident particles into new types, or the change in frequency of waves from their incident values. Removal of intensity from the beam merely by diversion into new directions is called elastic scattering, and this process is not properly included in the absorption coefficient.

The extinction coefficient α_e measures the reduction in beam intensity due to *all* contributing processes and is often represented as a sum $\alpha_e = \alpha + \alpha_s$ where α_s is the coefficient associated with elastic scattering. Additional subdivisions of processes that remove intensity from the beam are possible and sometimes used.

These coefficients appear in the Bouguer or Lambert–Beer Law in the form $I(x) = I(0)e^{-\alpha_e x}$, where $I(x)$ is the beam intensity after it has traveled a distance x in the medium, while α_e, α and α_s have units of inverse length, often written cm^{-1}.

The absorption coefficient α appears frequently in discussions of the optical properties of homogeneous solids, liquids, and gases, where α may be a strong function of the wavelength of the light involved, the temperature, and various sample parameters. The theory of electromagnetic waves relates α to the complex permittivity ε and permeability μ of the medium.

See also: Electromagnetic Radiation

Bibliography

M. Born and E. Wolf, *Principles of Optics*, 7th ed. Pergamon, New York, 1999. (A)
F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 4th ed. McGraw-Hill, New York, 2001. (E)
R. B. Leighton, *Principles of Modern Physics*, Chapter l2. McGraw-Hill, New York, 1959. (I)

[†] deceased

Encyclopedia of Physics, Third Edition. Edited by George L. Trigg and Rita G. Lerner

ISBN: 3-527-40554-2

Accelerators, Linear

T. P. Wangler

Overview

It is common in the particle-accelerator field to use the term linear accelerator (linac) for a device in which a beam of charged subatomic particles moves on a straight path and is accelerated by time-varying electric fields. In a radiofrequency (RF) linac, a beam of electrons, protons, or heavier ions is accelerated by an RF electric field with a harmonic time dependence using a linear array of waveguides or cavities excited in a resonant electromagnetic mode. Another type of linear accelerator, related to the betatron and called a linear induction accelerator or an induction linac, uses pulsed nonharmonic electric fields for acceleration in a nonresonant device called an induction module. For both the RF and the induction linac, the accelerating voltages are related to a changing magnetic field through Faraday's law. In both cases the beam is synchronized with the applied time-varying accelerating voltages.

Since the end of World War II, the linac has undergone a remarkable development. Its technological base is a consequence of the science of both the nineteenth and twentieth centuries. Included are the discoveries of electromagnetism by Faraday, Maxwell, and Hertz in the nineteenth century and the discovery of superconductivity in the twentieth century. The linear accelerator has developed as a powerful tool for basic research. It provides beams of high quality and high energy, sufficient to resolve the internal structure of the nucleus and its constituent subnuclear particles. Like a microscope, it has been used to probe the internal structure of the nuclear constituents, the proton and neutron, and has given us our present picture that these constituents are themselves made of point-like particles called quarks. Electron linacs are used in hospitals throughout the world as a source of X-rays for radiation therapy to treat cancer, an application that may represent the most significant spin-off of high energy and nuclear physics research.

The main advantage of the linac relative to other accelerator types is its capability for producing high-energy high-intensity charged particle beams of high quality with small beam diameter and small energy spread. Other attractive characteristics of the linac include the following: (1) Strong focusing can be provided to confine a high-intensity beam. (2) The beam traverses the linac in a single pass eliminating conditions that cause destructive beam resonances in circular accelerators. (3) Injection and extraction are simple compared with circular accelerators, because the natural orbit of the linac is open at each end. (4) Because the beam travels in a straight line, there is no power loss from synchrotron radiation, which is a major limitation for high-energy electron beams in circular accelerators. (5) The linac can operate in a pulsed mode at any duty factor, up to and including 100%. High duty factor provides the capability for higher average beam currents and beam powers. (6) The linac is capable of simultaneous acceleration of multiple charge states of a given mass species resulting in higher intensities for heavy-ion linacs. The principal disadvantage of the linac is the multiplicity and cost of accelerating structures, RF equipment, and operating power needed to achieve a given final beam energy. The increasing application of RF superconductivity to linacs within recent years, promises to reduce costs and make linacs even more attractive for many new applications.

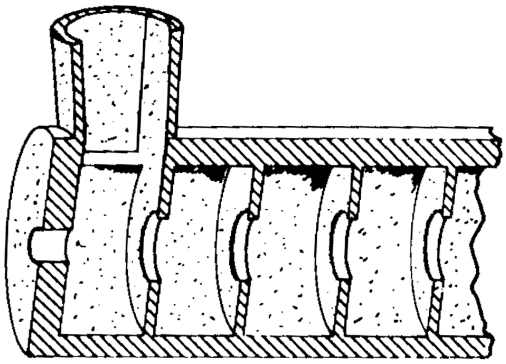

Fig. 1: Schematic drawing of the iris-loaded traveling-wave structure, showing the input waveguide through which the electromagnetic wave is injected into the structure at the end cell. The beam propagates along the central axis and is accelerated by the electric field of the traveling wave.

History

The first published proposal for a linac was made by Gustav Ising [1] in 1924. The first concept and experimental demonstration of an RF linac was due to Rolf Wideröe [2], who in 1928 showed that by applying a high-frequency voltage to a drift tube between two grounded electrodes, phased so that the beam experienced an accelerating voltage in both gaps, one could deliver an effective voltage gain to the beam of twice the applied voltage. This would not have been possible using a time-independent voltage, and it was clear that a multi-gap sequence of properly positioned drift tubes would allow repeated acceleration of the beam to arbitrarily high energies. The Wideröe result established the basis not only of RF linacs, but of the entire class of resonance RF accelerators that includes the cyclotron, and the synchrotron. But, linear accelerators that were useful for physics research were not feasible until after the developments of microwave technology stimulated by radar development during WWII.

Accelerating Structures

The modern RF linac uses high-Q resonant cavities or waveguides, either of which can be excited to high RF electromagnetic field levels at frequencies typically in the very high frequency (VHF) or ultrahigh frequency (UHF) ranges. These accelerating structures are tuned to resonance and are driven by external, high-power RF tubes such as klystrons, magnetrons, or gridded vacuum tubes. The accelerating structures are designed, through an optimized configuration of the internal geometry, for efficient transfer of electromagnetic energy to beam kinetic energy.

One can excite in a waveguide, a TM_{01}-like transverse magnetic traveling wave (having an axial electric accelerating field) that co-propagates with the beam. The waveguide uses a periodic array of conducting irises that reduces the phase velocity of the traveling wave to the velocity of the beam. Major developments of the iris-loaded traveling-wave linac (Fig. 1) for acceleration of relativistic electron beams began shortly after World War II at Stanford. These developments over two decades led to the design and construction of the 3-km-long Stanford Linear Accelerator (SLAC) linac [3].

Another method is to excite a standing wave in a multicell cavity, where the beam is accelerated by the longitudinal electric field in each of the cells. Various types of multicell structures have been invented for efficient acceleration over different ranges of beam velocity. One such structure is the Alvarez drift-tube linac (DTL), which is used to accelerate medium-velocity protons or other ions in the velocity range from about 0.04 to 0.4 times the speed of light. The first such accelerator was a 200-MHz DTL, proposed and built in 1946 at University of California by Luis Alvarez and co-workers [4]. The concept uses a quasi-periodic array of copper drift tubes enclosed in a high-Q cylindrical cavity. Figure 2 shows a DTL structure at the SNS facility at Oak Ridge National Laboratory. A TM_{010}-like transverse-magnetic mode is excited, which has an axial electric field in the gaps between drift tubes and zero field inside the drift tubes to shield the beam from deceleration when the field reverses. Unlike the original Wideröe structure, the fields in adjacent gaps are in phase, and the spacing between the centers of the gaps is equal to the distance the beam travels in a single RF period. Beam-focusing, necessary to confine the beam radially, is usually provided by installing magnetic-quadrupole lenses within the drift tubes. The DTL structure is not needed for electrons because electrons are so light that their velocity, as supplied from the electron source, is already above the applicable velocity range for a DTL.

Various types of standing-wave, coupled-cavity linac structures are used for acceleration of protons or electrons in the velocity range from about 0.4 to 1.0 times the speed of light. For example, a coupled-cavity linac operating at 805-MHz, called the side-coupled linac [5] (Fig. 3), invented at Los Alamos in the 1960s, is used at the Los Alamos Neutron Science Center (LANSCE) linac to accelerate a beam of protons or negative hydrogen ions from 100 to 800 MeV. The same structure type is used at the Spallation Neutron Source (SNS) now under construction at Oak Ridge.

The radiofrequency quadrupole (RFQ) linac, invented by I. M. Kapchinskiy and V. A. Tepliakov [6], is used for bunching and acceleration of low-velocity ion beams in the velocity range from about 0.01 to 0.1 times the speed of light. The RFQ bunches and captures most of the continuous beam injected from the ion source, and provides strong electric focusing to confine and accelerate high-current ion beams. The RFQ is a typically a few meters in length and is used as the first accelerating structure in a modern ion linac. A transverse electric-quadrupole mode is excited in a resonant RFQ cavity loaded with four equally-spaced conducting rods or vanes that are parallel to the beam axis (Fig. 4). The transverse focusing, provided by the electric-quadrupole field, is superior to magnetic focusing for these low-velocity ions. A longitudinal electric field for acceleration is obtained by machining a longitudinal modulation pattern onto the four vanetips, producing an array of accelerating cells.

Copper, because of its high electrical and thermal conductivity, is the most commonly used metal for the RF surfaces of room-temperature accelerating structures. The RF equipment and ac power for accelerator operation are major costs for a room-temperature linac. Reducing the fields to reduce the RF power dissipation means that more real estate is needed for the linac. For high-duty-factor room-temperature linacs, cooling the structures also becomes an important engineering requirement. The application of superconductivity to RF linacs has long been recognized as an important step toward better performance and lower costs. The use of superconducting niobium cavities reduces the Ohmic power dissipation to roughly 10^{-5} that of room-temperature copper. Even after the cryogenic refrigeration requirements are included, the net operating power for the superconducting linac is reduced typically by about two orders

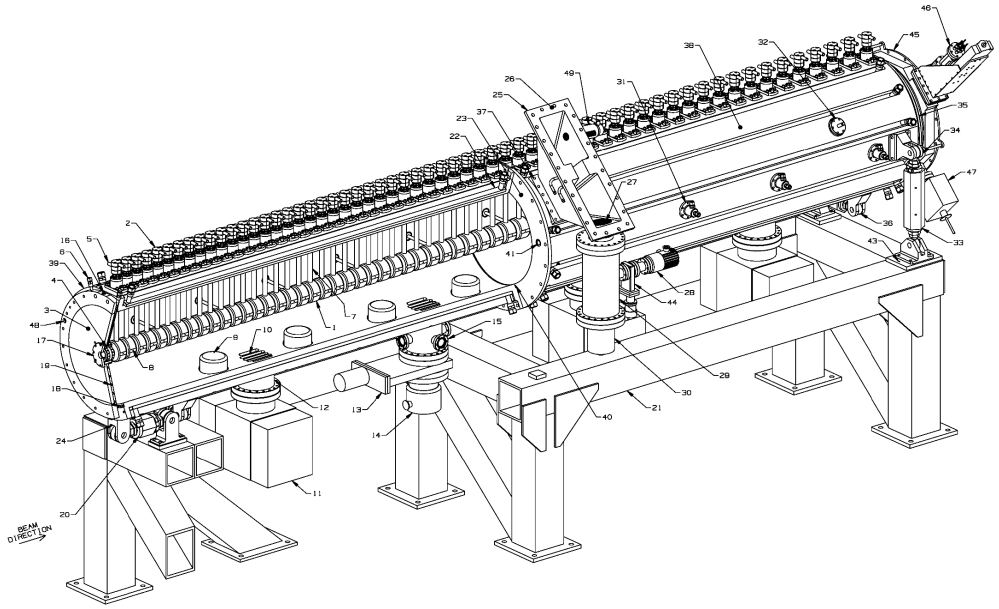

Fig. 2: Cut-away drawing of a 402.5-MHz 4.2-m 60-cell drift-tube-linac structure at the Spallation Neutron Source (SNS) at Oak Ridge National Laboratory. The photograph shows copper drift tubes and the drift-tube support stems.

of magnitude. For many applications the superconducting option is now superior to the room-temperature linac, resulting in reduced capital and operating costs, and a number of additional advantages including higher accelerating gradients and shorter linacs, and larger bore radii to reduce beam losses and to reduce wakefield effects in electron accelerators.

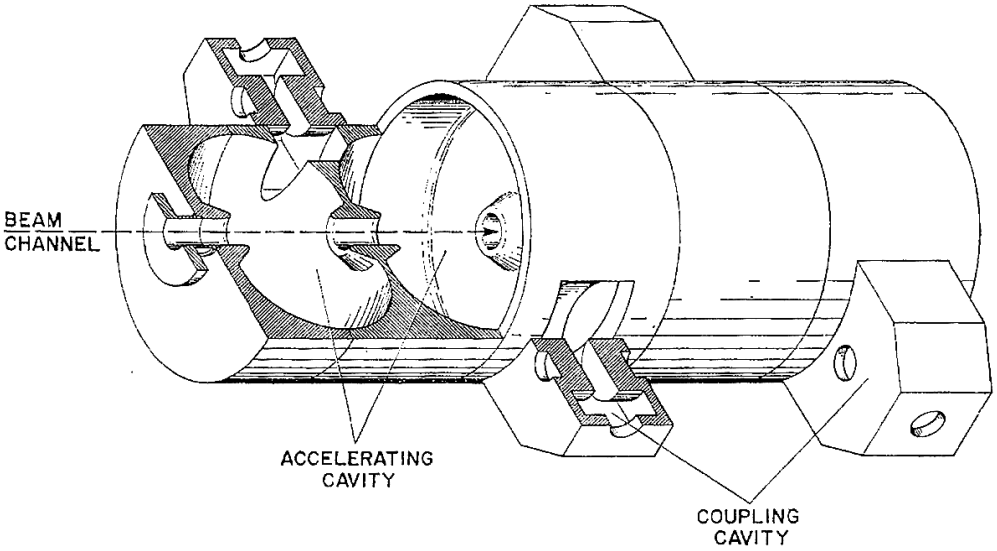

Fig. 3: The side-coupled linac structure was invented at Los Alamos in the 1960s. The cavities on the beam axis are the accelerating cavities. The coupling cavities on the side are nominally unexcited and stabilize the accelerating-cavity fields against perturbations from fabrication errors and beam loading.

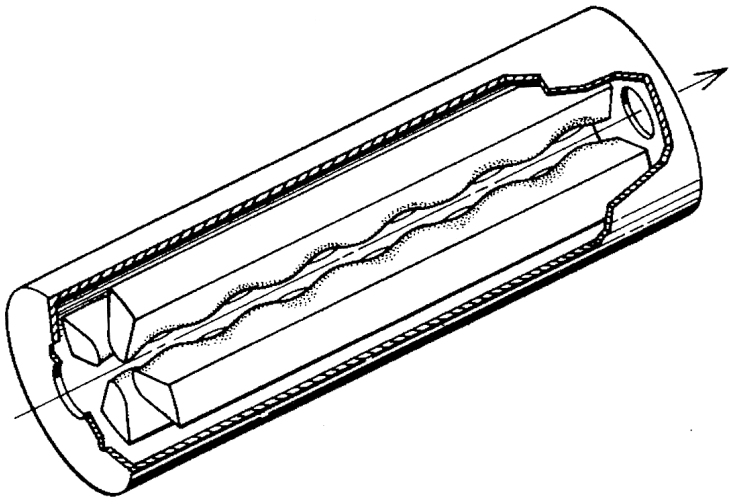

Fig. 4: Schematic drawing of a 4-vane RFQ accelerator. The four conducting vanes are excited with electric-quadrupole-mode RF voltages to focus the beam. The vanetips are modulated in the axial direction to produce longitudinal electric fields that bunch and accelerate low-velocity ions.

Beam Dynamics

A linac is designed for acceleration of a single on-axis particle, called the synchronous particle, which maintains exact synchronism with the accelerating field. Focusing forces must be provided to ensure stability for particles that deviate from the synchronous particle trajectory.

Longitudinal focusing is obtained when the synchronous particle is accelerated by an electric field that is increasing in time, i. e., before the peak of the sinusoidal waveform. Then, early particles experience a smaller field, which reduces their speed relative to the synchronous particle, whereas late particles experience a larger field, which increases their relative speed. This is known as the principle called phase stability, which results in stable phase and energy oscillations about the synchronous particle. Longitudinal stability is not a concern for relativistic electron linacs, where all particles travel at nearly the speed of light.

When off-axis particles experience an accelerating field that increases with time, as is required for longitudinal focusing, Maxwell's equations require that the particles also experience a radial RF defocusing force. In the case of relativistic electrons, the net radial force, comprised of oppositely directed electric and magnetic forces, is nearly zero, so that focusing may be unnecessary. For nonrelativistic ion beams, radial focusing must be provided to compensate for the RF defocusing force. With the exception of the RFQ, which provides its own transverse quadrupole RF electric focusing, the most effective solution to radial defocusing has been to include magnetic focusing lenses, usually magnetic quadrupoles, installed either between accelerating structures, or within the drift tubes, as is done in the DTL.

Controlling high-intensity effects can significantly influence the choices for the accelerator design, particularly for the beam-focusing requirements [7]. The main intensity limitation in nonrelativistic ion linacs is caused by the mutually repulsive Coulomb electric forces between the beam particles, commonly referred to as space charge. The space-charge force is important for lower velocity beams where the beam density is high, and where the attractive magnetic force from the moving beam is small. The nonlinearity of the space-charge force can also produce an extended halo surrounding the main core of the beam. This is of concern at high intensities, because halo particles may impact the accelerating structure, and induce undesired radioactivity.

In relativistic electron linacs the electric-space-charge force is nearly canceled by the oppositely directed magnetic force of the moving beam. But, Lorentz-compressed electromagnetic fields from the moving charges produce scattered radiation known as wakefields at discontinuities in the surrounding walls of the accelerator. Wakefields exert forces on trailing charges in both the same and later bunches, causing radial defocusing, energy loss, and energy spread in the beam. Another important intensity-limiting effect in electron linacs is known as the beam-breakup (BBU) instability. In this case, when a bunch travels off axis in an accelerating structure, it can excite a resonant mode that deflects trailing bunches. BBU is an instability that grows in time and distance along the accelerator, eventually leading to loss of the beam.

Linac Applications And Some Major Linac Facilities

A worldwide compendium of existing and planned scientific linacs published in 1996, listed 176 linacs distributed over the Americas, Europe, and Asia [8]. Contemporary electron linac applications include (1) a future TeV electron-positron international linear collider (ILC) proposed for high-energy physics research, (2) linacs for free-electron lasers (FEL), and (3) com-

pact electron linacs as X-ray sources for cancer therapy. The high intensity and excellent beam quality provided by the linac are important for achieving high luminosity in the collider, and high brilliance in the FEL. An important and very successful commercial application of RF linacs is the use of compact 10- to 20-MeV electron linacs for cancer therapy. Several thousand electron linacs are used worldwide for medical irradiations and this number is growing. In addition compact electron linacs are used for industrial applications including X-ray radiography, materials processing, and food sterilization.

Light-ion-linac applications include: (1) injectors to high-energy synchrotrons for physics research, (2) high-power proton linacs as drivers for spallation-neutron sources for condensed matter and material science research, nuclear material production for national defense, transmutation of nuclear wastes, and accelerator-driven fission reactors, (3) short-lived radioisotope production for medical diagnostics, (4) deuteron-linac-driven neutron sources for materials irradiation studies for fusion reactors, (5) multi-GeV linacs for heavy-ion inertial-confinement fusion, and (6) proton RFQ linacs for boron-neutron capture therapy. Important applications for heavy-ion linacs include (1) superconducting accelerators for a rare-isotope accelerator (RIA) facility to be used for nuclear-physics research with radioactive ion beams, and (2) ion implantation for commercial semiconductor fabrication.

The largest electron linac is the 3-km 50-GeV room-temperature traveling-wave structure at the Stanford Linear Accelerator Center (SLAC). This linac, which operates at 2856 MHz, has had a very productive physics history, beginning operation as a fixed-target accelerator facility in 1966, and as an electron-positron collider (SLC) beginning in 1989. The newest application of the SLAC linac is the Linac Coherent Light Source (LCLS), which will use electrons accelerated in the last kilometer of the linac for the world's first hard-X-ray free-electron laser, when it becomes operational in 2009.

At the Deutsches Electronen Synchrotron (DESY) laboratory in Hamburg, Germany, a superconducting standing-wave electron linac serves both as a state-of-the-art FEL facility, and as a pilot facility for the TESLA electron-positron-collider development project. The planned coherent X-ray FEL will use a 50-GeV electron linac built from superconducting niobium cavities operating at 1.3 GHz. The ILC collider concept, for which the site is not yet determined at the time this article is being written, uses an approximately 33-kilometer-long tunnel that will house two superconducting linear accelerators in which electrons and positrons will be accelerated and made to collide at a center-of-mass energy of 0.5 TeV. The linacs will use 21 000 superconducting niobium cavities (Fig. 5), cooled with liquid helium to an operating temperature of 2 K. The TESLA project has made significant advances in the performance of the superconducting accelerating cavities, achieving accelerating gradients larger than 25 MV/m.

The Continuous Electron Beam Accelerator Facility (CEBAF) at the Thomas Jefferson National Accelerator Facility (TJNAF) in Virginia is a 5-pass recirculating linear-accelerator facility that uses two superconducting electron linacs (320 5-cell superconducting cavities in 40 cryomodules) joined by two 180-degree magnet arcs through which the beam is recirculated and accelerated to a final energy of 6 GeV. CEBAF is used for nuclear physics research and began operation in 1994.

The first proton linac with beam energy near 1-GeV kinetic energy is the 800-m-long LANSCE linac at Los Alamos. This high-intensity standing-wave proton linac began operation in 1972 as a pion factory (Los Alamos Meson Physics Facility) for nuclear physics research,

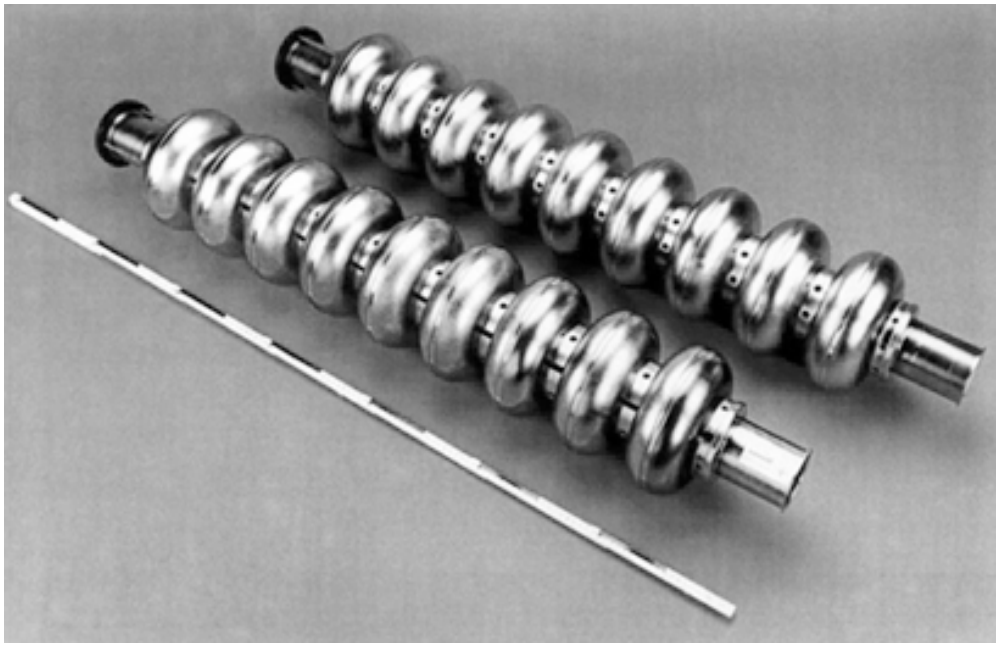

Fig. 5: Two 1.3-m 9-cell 1.3-GHz superconducting niobium elliptical cavities used for the X-ray FEL at the TESLA Test Facility and for the proposed linacs of the 0.5-TeV center-of-mass e^+–e^- International Linear Collider.

and is now a multipurpose facility used as a driver for a spallation-neutron source and for proton radiography. LANSCE can simultaneously accelerate H^+ and H^- ions to an energy of 800-MeV. The linac uses normal-conducting copper technology, and is comprised of a 201.25-MHz DTL (0.75 to 100 MeV) followed by an 805-MHz side-coupled linac (100 to 800 MeV).

The 1-GeV H^- Spallation Neutron Source (SNS) linac, now under construction at Oak Ridge, will incorporate the world's first proton superconducting linac. The room-temperature part of the linac consists of a 402.5-MHz RFQ and DTL, which accelerates the beam to 87 MeV, followed by an 805-MHz side-coupled linac that accelerates the beam to 186 MeV. The 805-MHz superconducting accelerator, comprised of 81 niobium superconducting cavities, continues the acceleration to the final beam kinetic energy of 1 GeV.

Hydrodynamic testing of mockups of nuclear-weapons implosion systems is an important application of induction linacs. At Los Alamos, a pair of electron induction linacs at the Dual Axis Radiographic Hydrodynamic Test (DARHT) facility is used for X-ray radiography. Energetic (20 MeV), intense (2 kA), short (60 ns) bursts of electrons produce X-rays in a tungsten target. The pair of linacs, oriented at 90°, provides time-resolved tomographic reconstructions of implosions driven by high explosives.

Heavy-ion linacs are used for study of the atomic nucleus. At Argonne National Laboratory the approximately 150-m-long superconducting ATLAS linac is capable of accelerating ions of any element to energies as high as 17 MeV per nucleon. ATLAS is the world's first heavy-ion accelerator to use superconducting cavities. ATLAS contains 62 independently

phased superconducting cavities; the independent phasing of these cavities provides adjustability needed for efficient acceleration over the wide range of ion species.

Heavy-ion superconducting linacs are the basis of the planned nuclear-physics facility called the Rare Isotope Accelerator (RIA), which is being designed for acceleration of radioactive ion beams. A powerful 1.4-GV RIA driver linac is used for acceleration of stable heavy-ion beams that bombard a target in which rare-isotopes are produced. The driver linac can simultaneously accelerate multiple charge states of ions of a given mass, yielding higher intensity beams than any other type of accelerator for this application. The rare isotope species produced in the target are then selected magnetically for acceleration by a second superconducting linac, providing energetic radioactive beams for a variety of nuclear-physics and astrophysics studies.

See also: Accelerators, Potential-Drop Linear; Microwaves and Microwave Circuitry

References

[1] G. Ising, *Ark. Mat. Fys.* **18** (no. 30), 1–4 (1924).
[2] R. Wideröe, *Arch. Electrotech.* **21**, 387 (1928).
[3] R. B. Neal, W. A. Benjamin (eds.), *The Stanford Two-Mile Accelerator*. New York, 1968.
[4] L. W. Alvarez *et al.*, *Rev. Sci. Instrum.* **26**, 111–133 (1955).
[5] E. A. Knapp, B. C. Knapp, and J. M. Potter, *Rev. Sci. Instrum.* **39**, 979–991 (1968); D. E. Nagel, E. A. Knapp, and B. C. Knapp, *Rev. Sci. Instrum.* **38**, 1583–1587 (1967).
[6] I. M. Kapchinskiy and V. A. Tepliakov, *Prib. Tekh. Eksp.* **2**, 19–22 (1970).
[7] Thomas P. Wangler, *Principles of RF Linear Accelerators*, Wiley, New York, 1998.
[8] J. Clendenin *et al.*, *Compendium of Scientific Linacs*. CERN report CERN/PS 96-32 (DI), Nov. 1996.

Accelerators, Potential-Drop Linear

R. G. Herb[†] *and G. M. Klody*

Introduction

Potential-drop accelerators were the first to open up nuclear physics to extensive experimentation. Despite the higher-energy beams that were soon available with the cyclotron, betatron, and later accelerators that provide very high energies through many small-energy increments, potential-drop accelerators have continued to play an important, and often dominant, role in nuclear physics. In addition, potential-drop accelerators have a broad range of other applications in many, diverse fields in industry, research, and medicine.

Potential-drop accelerators consist of a high-voltage terminal supported by an insulating column, a means of generating the high voltage, and an acceleration tube. To reduce their size, many potential-drop accelerators use high-pressure gas for electrical insulation. The

[†]deceased

electrostatic potential drop between the high-voltage terminal and ground accelerates charged particles to an energy equal to the charge times the terminal voltage ($E = qV$). A variety of ion and electron sources are available to produce the beams of charged particles for acceleration.

The acceleration tube is an insulating assembly, mounted between the high-voltage terminal and ground potential, through which the beam is accelerated. High vacuum inside the tube minimizes beam losses. Mechanical strength is also important for tubes in accelerators insulated with high-pressure gas. The tube design that maximizes voltage-holding capability and voltage stability under these conditions is a laminated series of insulating rings and metal electrodes. The electrodes are connected to a resistive voltage divider from the terminal to ground for a uniform voltage gradient (linear potential drop) along the tube.

Single-stage potential-drop accelerators can accelerate ions or electrons. Two-stage accelerators, called tandems, are for ion acceleration. Tandems accelerate negative ions from ground potential to a positively charged high-voltage terminal, change the ionic charge from negative to positive in the terminal, and accelerate the ions through a second acceleration tube, back to ground potential. For a given terminal voltage, tandems produce higher-energy ions than single-stage accelerators, but usually with less beam current.

Potential-drop accelerators are generally distinguished by the type of generator used to charge the high-voltage terminal. Each high-voltage generator design has advantages for specific applications. For precise measurements of nuclear-energy-level characteristics and many analytical techniques, the electrostatic accelerator is far superior. For certain neutron-induced reactions the Cockcroft–Walton-type accelerator may be most convenient. Needs for intense positive-ion or electron beams of a few MeV energy are most easily met by the Dynamitron.

Modern potential-drop accelerators are very reliable and simple to operate. This has greatly expanded their range of applicability in many fields. They are routinely used for ion implantation in the manufacture of semiconductor devices, nondestructive analysis of materials, surface hardening to reduce corrosion and wear, polymerization, sterilization, pollution monitoring, age determination of samples for archaeology, cosmology, and oceanography, and analysis of biomedical specimens.

Each of the accelerators described below is serving an important need in science and technology, and each is playing an a expanding role.

Cockcroft–Walton Voltage Multiplier Accelerator

This accelerator was the first to be used successfully for nuclear transmutation and gained wide recognition when results were published in 1932. It employs a circuit developed by H. Greinacher in 1920, as illustrated in Fig. 1, which utilizes two stacks of series connected capacitors. One capacitor stack is fixed in voltage except for voltage ripple with one terminal connected to ground and the other to the load which, in this case, is an evacuated accelerating tube.

One terminal of the second capacitor stack is connected to a transformer giving peak voltages of $\pm V$, and voltages at all points along the second capacitor stack oscillate over a voltage range of 2 V. The two capacitor stacks are linked by series-connected rectifiers. As voltage on the second stack oscillates, charge is transferred stepwise from ground to the high-voltage terminal. Voltage here is steady except for ripple caused by power drain and stray capacitance. Its value is approximately $2\,\mathrm{V}N$, where N is the number of stages. The power supply furnishes

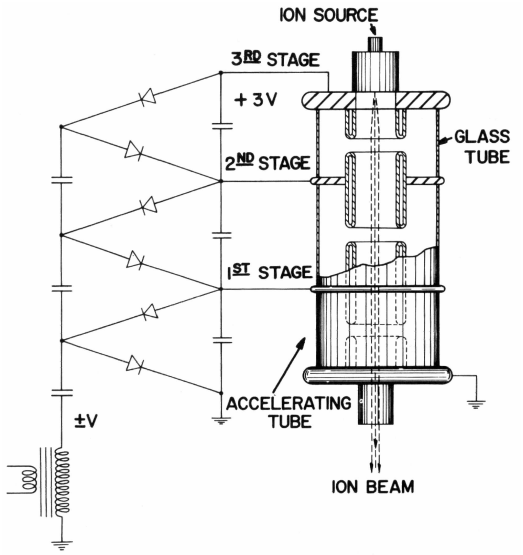

Fig. 1: Schematic drawing of a Cockcroft–Walton accelerator utilizing three stages.

power to an evacuated accelerating tube equipped with an ion source. Usually the tube and ion source are continually pumped and the multi-section tube may have one tube section per accelerator stage.

Cockcroft–Walton accelerators operating in open air at 1 million volts are large in size and must be housed in a very large room to avoid voltage flashover. Figure 2 shows an 850-keV accelerator of this type built by Emile Haefely & Co. Ltd. It serves to inject pulses of hydrogen ions into a high-intensity linear accelerator at the Los Alamos National Laboratory for production of intense meson beams.

Size and space requirements increase rapidly as voltage is extended above 1 million volts in open air and the practical upper voltage limit for these accelerators appears to be about 1.5 MV.

G. Reinhold of Emile Haefely & Co. Ltd. has shown that the terminal voltage developed by multipliers, using the circuit of Fig. 1, does not depart greatly from $2VN$ for multipliers going to a few hundred kilovolts. However, above about 500 kV, voltages achieved fall substantially below values given by this simple expression because of stray capacitances and other effects. He has developed another circuit called a symmetrical cascade rectifier in which the shortcomings of the simple circuit are largely eliminated. It employs two transformers and two capacitor stacks that oscillate in voltage. Both feed one fixed capacitor stack. Using this symmetric system, Emile Haefely & Co. Ltd. has built an open air rectifier without an accelerating tube going to 2.5 million volts.

The Philips Gloeilampenfabrieken has also manufactured accelerators with Cockcroft–Walton charging. Emile Haefely & Co. Ltd. has built accelerators utilizing voltage multipliers housed in tanks containing insulating gases such as SF_6 or a mixture of N_2 and CO_2. These machines have ranged in voltage from about 1 million up to 4 million volts. Nisshin-High

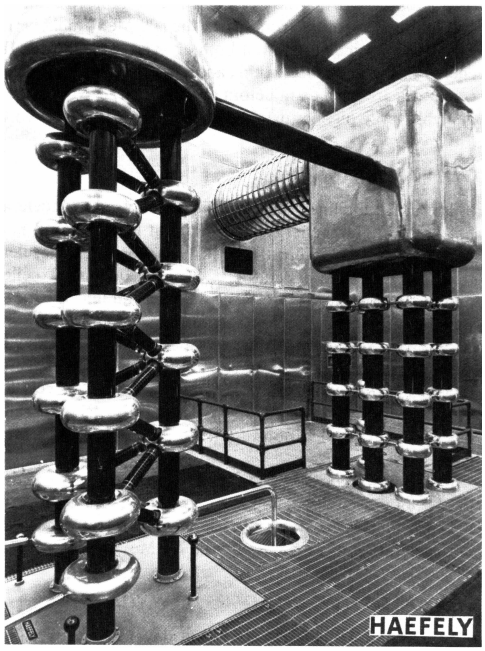

Fig. 2: 850-keV Haefely accelerator which serves to inject pulses of hydrogen ions into a linear accelerator at the Los Alamos National Laboratory.

Voltage Company also builds SF_6-insulated Cockcroft–Walton accelerators from 500 kV to 3 million volts for their industrial electron processing systems. The use of insulating gases at high pressure permits great savings in the size of equipment for a given voltage.

Dynamitron Accelerator

M. R. Cleland invented a cascaded rectifier system termed the Dynamitron in which series connected rectifiers are driven in parallel. The circuit is shown schematically in Fig. 3, and Fig. 4 is a photograph of a 4-million-volt positive ion Dynamitron accelerator.

Rectifiers connected in series between ground and the high-voltage terminal are positioned in two columns on opposite sides of the accelerating tube of the Dynamitron and the high-voltage column is enclosed by half rings that have a smooth exterior surface to inhibit corona and spark discharge. The half-rings serve as capacitor plates coupled capacitively to the large semicylindrical rf electrodes positioned between the walls of the tank and the high-voltage column of the accelerator. The rf electrodes form the tuning capacitance of an *LC* resonant circuit which is driven by a separate power supply through an oscillator tube.

Since ac power is fed in parallel to each of the series-connected rectifiers, the relatively large storage capacitors that are connected between successive stages of other cascaded rectifier systems are not required. Stored energy in the Dynamitron is low and does not differ greatly from that in electrostatic machines. This feature is important since damage due to discharge can be a serious problem in multimillion volt accelerators, especially for discharge in the accelerating tubes.

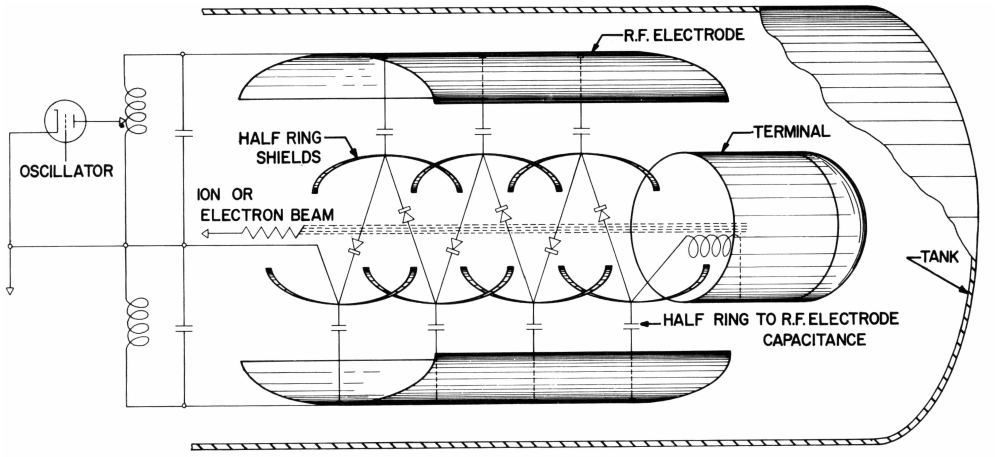

Fig. 3: Schematic diagram of a Dynamitron accelerator in a pressure tank.

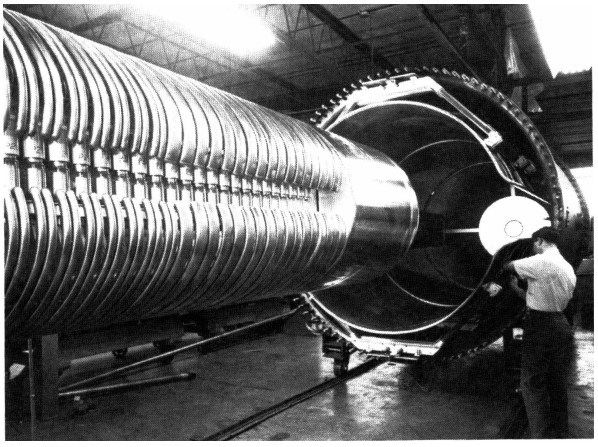

Fig. 4: A 4-MeV Dynamitron positive-ion accelerator with pressure tank rolled away from the high-voltage column.

These accelerators are enclosed in pressure tanks containing high-pressure SF_6 gas. At a pressure of 1 atm, this gas has a dielectric strength about 2.7 times that of air at the same pressure. Its dielectric strength rises approximately linearly with pressure up to a few atmospheres and at a pressure of about 7 atm it will sustain fields of about $200\,\mathrm{kV/cm}$.

Radiation Dynamics manufactured many single-stage and double-stage (tandem) positive-ion Dynamitrons operating at terminal potentials up to 4 MV. They are especially advantageous for applications requiring high currents.

A larger proportion of Dynamitrons manufactured have been electron accelerators for industrial applications such as polymerization of plastics and sterilization of disposable medical products. These applications require electron energies up to a few MeV and from 10 to 100 kilowatts of electron beam power.

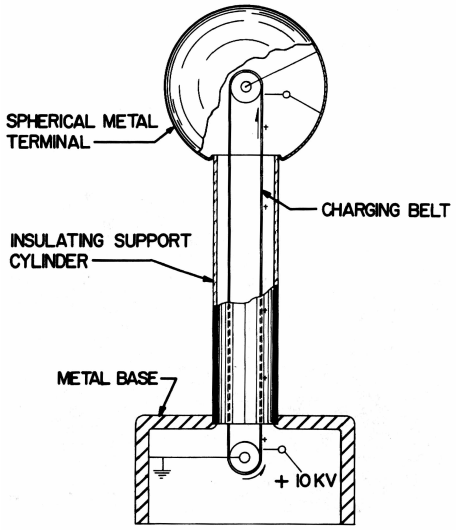

Fig. 5: Schematic drawing of a Van de Graaff generator for operation in atmospheric air.

The most powerful machines built by Radiation Dynamics include a 0.5-MV machine with an electron beam power of 50 kW, a 1-MV machine giving 100 kW of electron beam power, a 1.5-MV machine providing 75 kW, a 3-MV machine providing 150 kW, and a 5-MV machine providing electron beam power of 200 kW.

Van de Graaff Accelerator

In this accelerator, high voltage is generated by means of an insulating belt which carries charge from ground to the high-voltage terminal. Robert Van de Graaff built the first successful belt-charged high-voltage generator in 1929.

In this device, which is illustrated in Fig. 5, electrical charge is deposited on an insulating, motor-driven belt and is carried into a smooth, well-rounded metal shell which is shown in the figure as a sphere. Here charge is removed from the belt and passes to the sphere which rises in voltage until the sphere is discharged by a spark or until the charging current is balanced by a load current.

To charge the belt, a corona discharge is maintained between a series of points or a fine wire on one side of the belt and the grounded lower pulley or a well-rounded, grounded, inductor plate on the other side of the belt. If the corona needles are at a positive potential, the belt intercepts positive ions as they move from needle points toward the grounded pulley. Charge is carried into the sphere where it is removed by an array of needle points and passes to the outer surface of the sphere. The generator can provide a high negative voltage if the corona needles are operated at a negative voltage with respect to the grounded pulley.

Belt charging electrodes must be well shielded from the field of the high-voltage terminal and charge must be carried well within the sphere before removal is attempted. Charging current is then completely independent of voltage on the terminal and voltage will rise until

limited by corona or spark-over to ground, by leakage current along insulators, or by a load such as ion current through an accelerating tube.

The Van de Graaff belt-type generator was first used to accelerate ions in 1932 at the Department of Terrestrial Magnetism of the Carnegie Institution of Washington, D.C. A machine completed in 1934 at this institution was used extensively for research in nuclear physics and is now on display at the Smithsonian Institute of Washington, D.C.

Open-air belt-type accelerators for 1 million volts or more are large and require a very large enclosure. At Wisconsin this accelerator was adapted to a pressure tank and by 1940 a belt-charged accelerator was operated successfully up to 4.5 million volts.

These machines insulated by high-pressure gas were manufactured from 1946 by the High Voltage Engineering Corporation for accelerating of electrons and for positive ions. In 1988, Vivirad High Voltage Corporation purchased the Accelerator Division of High Voltage Engineering Corporation to build accelerators, for research and materials analysis. High Voltage Engineering–Europa also manufactures accelerators with Van de Graaff belt-charging to about 2 million volts, primarily for ion-implantation applications.

Two-Stage Accelerators

In 1958, the High Voltage Engineering Corporation completed a machine for acceleration of negative ions as illustrated schematically in Fig. 6. Negative hydrogen ions from a source developed for this purpose are accelerated as they pass from ground to the terminal which is

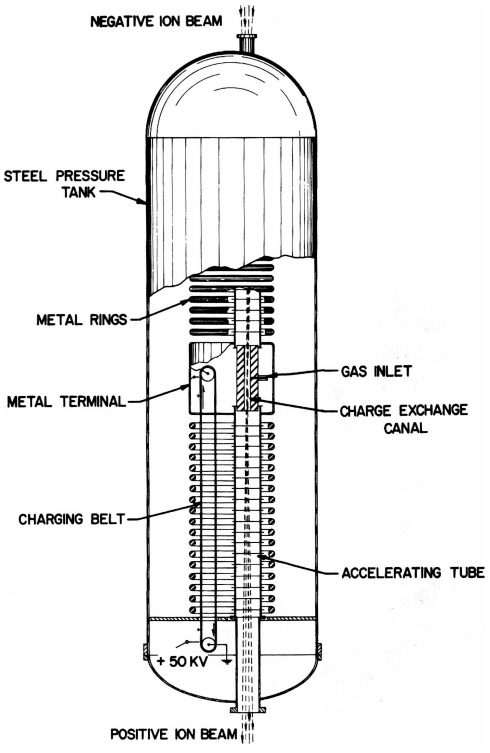

Fig. 6: Schematic drawing of a tandem electrostatic accelerator in a pressure tank.

at a high positive voltage V. Here they are stripped of both electrons as they pass through a very thin foil or through a small-diameter tube containing adequate gas. The protons are again accelerated as they pass from the terminal to ground and they emerge from the accelerator with an energy of $2Ve$. These machines, which give two stages of acceleration, are called tandems.

Atoms of a large proportion of the elements form stable negative ions. A negative oxygen ion gains an energy of 10 MeV as it goes from ground to the terminal of an accelerator operating at 10 million volts. Here the oxygen atoms may be stripped of all of their eight electrons. The oxygen nuclei gain 80 MeV as they pass to ground and they emerge from the machine with an energy of 90 MeV.

The High Voltage Engineering Corporation manufactured a large number of two-stage accelerators, which are in use in laboratories throughout the world. The largest reach terminal voltages of over 16 million volts. Many are used for acceleration of heavy ions for research and materials analysis applications.

Nisshin-High Voltage Company also manufactures tandem accelerators with Van de Graaff belt-charging to 3 million volts for basic research and materials modification and analysis.

Tandetron Accelerators

In 1980, K. H. Purser began development of the Tandetron, a line of compact tandem accelerators designed to meet the requirements for applications in materials analysis, accelerator mass spectrometry, high-energy implantation, and basic research. An MeV implanter which uses a Tandetron accelerator is shown schematically in Fig. 7.

The high-voltage generator in the Tandetron is a power supply with a parallel-driven cascaded rectifier circuit similar to that of the Dynamitron. Using silicon rectifiers and driven at high frequency (about 50 kHz), the power supply delivers several milliamperes of charging current at up to 3 million volts with high stability and negligible terminal voltage ripple. For accessibility, the high-voltage stack is mounted at right angles to the accelerator column, rather than being built into the column (compare Fig. 3). The accelerator tank contains SF_6 gas at about 500 kPa (5 atm) to insulate the accelerator and the power supply.

Through 1988, General Ionex Corporation manufactured a large number of Tandetron systems, many with specialized accessories that they developed for materials analysis and modification applications. Genus Corporation now manufactures Tandetron systems.

Pelletron™ Accelerators

Pelletron accelerators use a charging chain, rather than a belt or power supply, to generate high voltage on the accelerator terminal. The chain consists of steel cylinders (pellets) joined by links of solid insulating material such as nylon (Fig. 8). The chain is intrinsically spark-protected (undamaged in test sparks at over 30 million volts). Pellet-charging current is adjustable and highly uniform, so that terminal voltages are easily maintained at very precise values with very little voltage ripple. This is advantageous for many measurements in nuclear physics and is usually required in other applications. Many belt-charged machines have been converted to Pelletron charging.

The pellets are charged inductively, so there is no contact with the charging electrodes (inductors). For a positive terminal voltage, positive charge is induced on the pellets at a

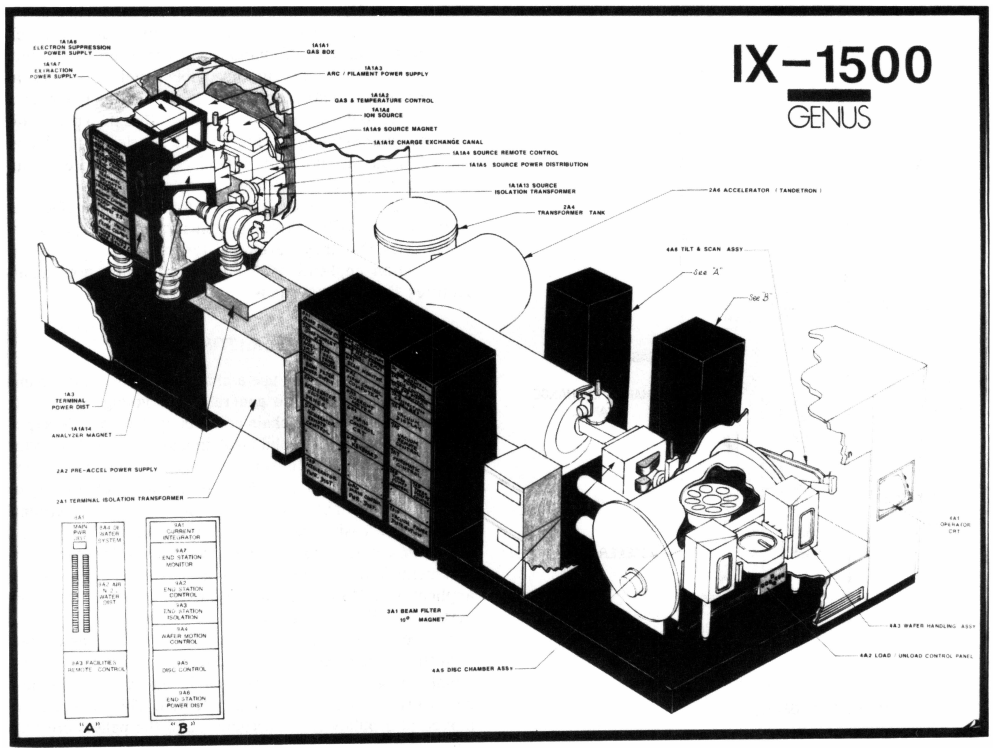

Fig. 7: Schematic diagram of a Tandetron accelerator in an ion-implant system. The high-voltage solid-state multiplier stack is at right angles to the tandem accelerator column, and the high-frequency driver is outside the tank for accessibility.

motor-driven pulley at ground potential. This charge is removed at a pulley in the terminal, and the pellets are then negatively charged. Thus, the up-going and down-going runs of the chain contribute equally to charging. Reversing the inductor voltage polarities gives a negative terminal voltage.

Pelletron accelerators are manufactured by the National Electrostatics Corporation. Single-stage Pelletrons have either an electron source or a positive-ion source in the terminal. Pelletrons manufactured by this company for industries and laboratories around the world range in voltage from 1 to 25 million volts, including the largest potential-drop accelerator, the 25-million-volt tandem at the Oak Ridge National Laboratory (described below).

Most Pelletrons above 4 million volts are vertically oriented for simplicity of support. The high-voltage column, which supports the terminal, charging chain, and acceleration tubes, is an assembly of standard column modules, each module holding 1 million volts. Most smaller Pelletrons, built for applications requiring only 4 million volts or less, use a simple, inexpensive cast acrylic support column.

Acceleration tubes in Pelletrons have ceramic insulators and titanium electrodes bonded together with a metal for organic-free, ultra-high-vacuum operation. This is important for

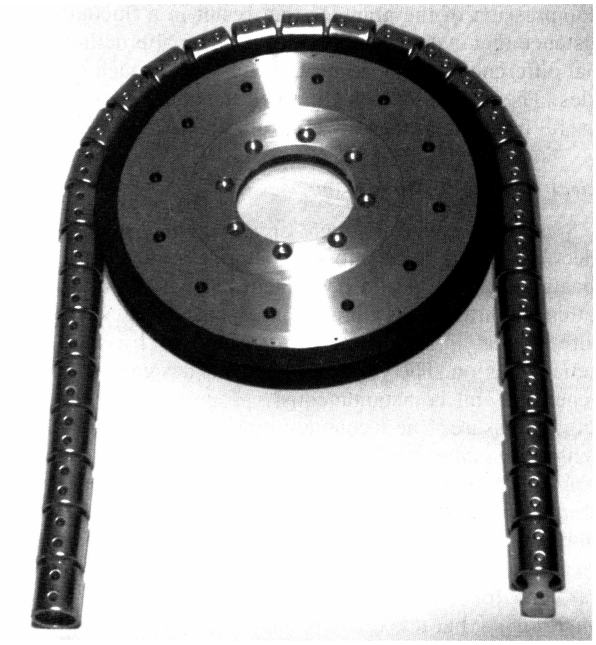

Fig. 8: Charging chain and pulley.

reliability at high voltage and for contaminant-free vacuum. Earlier tube designs have glass insulators which are bonded to the metal electrodes with organic glues.

Large Pelletrons (8 to 25 million volts) are in operation on five continents. They are used principally for basic research with heavy ions. Most Pelletrons are smaller, to 5 million volts, and are used primarily for applications such as production ion implantation, industrial materials analysis, accelerator mass spectroscopy, and biomedical analysis.

Folded Tandems

Folded tandems are like straight-through tandems (see Two-Stage Accelerators, above), except that both acceleration tubes are in a single insulating column with a 180° magnet in the terminal to steer the beam from one tube into the other. Figure 9 shows the folded 25-million-volt tandem at the Oak Ridge National Laboratory. The steel tank, which contains about 700 kPa (7 atm) of SF_6 gas for high-voltage insulation, is large (30 m high and 10 m in diameter). The straight-through design, however, would be much taller, because it has insulating columns above and below the high-voltage terminal plus an ion-beam injector above the tank (compare designs in Figs. 6 and 9).

The folded design, first successfully used by Naylor in New Zealand, reduces the costs for the tank, SF_6, and building for very high-voltage tandems. For higher-energy ion beams, Oxford University converted their single-stage accelerator to a folded tandem. General Ionex Corporation manufactured several 660-kilovolt folded tandem systems for materials analysis by Rutherford backscattering. Not only does the folded design give a compact system, but it also locates both the ion source and sample chamber near the system control panel.

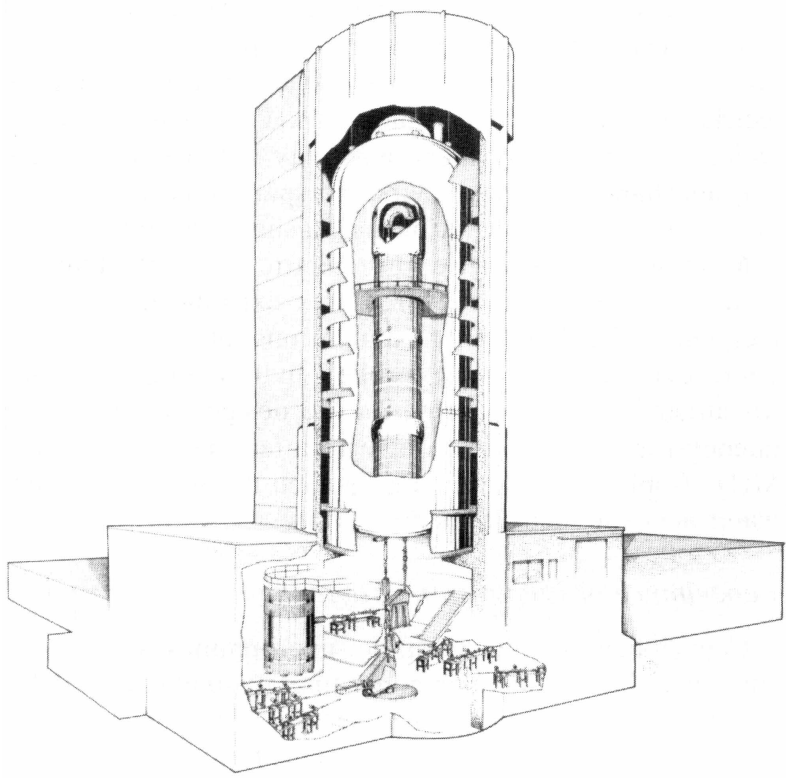

Fig. 9: Drawing of 25-MV folded tandem. Tank is approximately 30 m high and 10 m in diameter. Service platforms are shown in use inside and outside the column.

With +25 million volts on the terminal, negative ions injected from below the accelerator (see Fig. 9) are accelerated up to the terminal to an energy of 25 MeV, where they are stripped of some or all of their electrons. Stripping produces ions in a range of different charge states, and the energy gained in the second acceleration depends on which charge state is selected. If the magnet in the terminal is set for ions of charge +15, these ions gain $15 \times 25 = 375$ MeV in the second acceleration for a total energy of 400 MeV. A second stripper, one-third of the way down the second acceleration tube, can increase the charge to +34 to give ions with a final energy of over 700 MeV. The terminal magnet completely filters out all undesired components from the beam before the second acceleration, so there is no extra beam loading, and ion transmission is straightforward.

The 25-million-volt accelerator column consists of 1-million-volt modules which use the same components as in smaller Pelletrons. Six charging chains generate over 600 μA of current to the terminal. Service platforms inside and outside the column give complete access for maintenance without column disassembly. The computer-based control system uses light link telemetry to transmit over 200 control and monitoring signals to components at high-voltage locations in the accelerator column and terminal.

The ion beam injector at Oak Ridge (cylinder in lower left of Fig. 9) operates at $-500\,\text{kV}$ in air to preaccelerate ions from the negative-ion sources for injection into the accelerator. The 90° magnets below the accelerator provide high resolution of the injected ion mass and high resolution of the accelerated beam energy.

The other large folded tandem is the 20-million-volt Pelletron at the Japanese Atomic Energy Research Institute in Tokai-Mura.

Large Straight-Through Tandems

Argentina's Comisión Nacíonal de Energia Atómica operates a 20-million-volt Pelletron in their Tandar facility in Buenos Aires.

Daresbury Laboratory in England built a large tandem for scientific research. This accelerator is charged by a Laddertron, which looks like two parallel Pelletron chains with metal bars connecting the pellets of one chain to the adjacent pellets of the other chain.

At Strasbourg, M. Letournel is developing the Vivitron, a design for a very high-voltage belt-charged tandem accelerator. The design uses metal electrodes, called porticos, to modify the electric-field distribution between the accelerator column and the grounded steel tank for higher terminal voltages.

Bibliography

M. R. Cleland and P. Farrell, "Dynamitrons of the Future," *IEEE Trans. Nucl. Sci.* **NS-12**, 227 (1965).

Large Electrostatic Accelerators (D. Allan Bromley, ed.) [Reprinted from *Nuclear Instruments and Methods* **122** (1974)].

M. S. Livingston and J. P. Blewett, *Particle Accelerators*. McGraw-Hill, New York, 1962.

C. C. Thompson and M. R. Cleland, "Design Equations for Dynamitron Type Power Supplies in the Megavolt Range," *IEEE Trans. Nucl. Sci.* **NS-16**, 124 (1969).

Acoustical Measurements

M. Strasberg

Acoustics is concerned with fluctuations in the value of various mechanical quantities characterizing the state of matter – fluctuations of the pressure or other components of stress, or fluctuations of density, temperature, and the position of individual particles of matter. Primary acoustical measurements determine the magnitude and wave form of the oscillations of one of these quantities at one or more positions in space, whereas secondary measurements characterize the wavelike propagation of these oscillations through space by determining the speed of propagation, the intensity or rate of propagation of acoustic energy, and the absorption or rate of dissipation of acoustic energy.

Most of the early primary acoustical measurements used mechanical devices that would determine only the magnitude of the oscillatory particle displacement or velocity. The advent of electronic amplifiers led to development of electromechanical transducers which convert the oscillating mechanical quantities into an emf. Nowadays, primary acoustical measurements usually utilize linear electromechanical transducers to generate an oscillating emf that

is instantaneously proportional to the oscillating mechanical quantity. The magnitude and wave form of the oscillating mechanical quantity are deduced from measurements of the corresponding characteristics of the emf.

Several textbooks survey acoustical measurements. Wood [1], Stephens and Bate [2], and Meyer and Neuman [3], for example, describe mechanical techniques used before the present electronic age, and the latter two texts also cover contemporary instruments. The four-volume *Encyclopedia of Acoustics* has 150 pages discussing measurements [24]. The textbook *Acoustical measurements* [4] is devoted entirely to the subject, albeit primarily to audible sound. Various ultrasonic measurement techniques are described by Fry and Dunn [5] and are discussed in several chapters in the 18-volume collection edited by Mason [6]. Handbooks published by several manufacturers discuss the use of their instruments for sound measurements [7–9]. Standard procedures and apparatus for performing certain conventional measurements have been published by organizations concerned with standardization [10].

Measurement of Oscillating Pressure

Most present-day acoustic measurements in fluids utilize electromechanical transducers sensing the oscillating pressure. These are called *microphones* when used in gases and *hydrophones* when waterproofed for use in liquids.

Various physical effects are utilized for generating the alternating emf; a comprehensive survey is given in Chapter 8 of Olson [11]. The type chosen for a particular application depends on a compromise among conflicting desirable characteristics, e. g., small size, high sensitivity, constant sensitivity over the frequency range of interest, stability to varying temperature and humidity, and low cost.

Whatever type of transducer is used, the magnitude, wave form, and spectral characteristics of the generated alternating emf are measured with electronic instruments such as voltmeters, oscilloscopes, wave analyzers, and spectrum analyzers. The corresponding characteristics of the oscillating pressure are determined from the measured characteristics of the alternating emf by dividing the electrical magnitudes by a proportionality factor (called a sensitivity or calibration factor) that is equal to the generated emf per unit sound pressure, including the amplification in the electronic system. Ideally this factor is independent of the magnitude and frequency of the oscillating quantity, but in practical systems there is usually some variation in sensitivity with frequency within the range of interest that must be taken into account.

Carbon Microphone

The carbon microphone, an early pressure-sensing transducer still used in some telephones, depends on the piezoresistive property of compressed carbon granules. The oscillating sound pressure causes fluctuations in the compression of the granules that result in a fluctuating resistance that is detected electrically as an alternating potential difference developed by a dc current through the granules. The device has a high, albeit unstable, sensitivity; its useful frequency range covers from about 250 to 4000 Hz.

Electromagnetic Microphone

Sometimes called *dynamic* or *moving-coil* microphones, electromagnetic microphones comprise a light coil of wire located in the field of a permanent magnet and attached to a thin

diaphragm that vibrates in response to the sound pressure, so that an alternating emf is generated by electromagnetic induction. The vibrating system is designed so that the generated emf is instantaneously proportional to the oscillating pressure. The frequency range of relatively constant sensitivity can extend from perhaps 100 to 10 000 Hz.

Another type of electrodynamic microphone utilizes a limp metallic ribbon instead of the coil. If both sides of the ribbon are exposed to the sound, the ribbon tends to generate an alternating emf proportional to the particle velocity of the gas, and for this reason it is sometimes called a velocity microphone, but it is really responding to the gradient of the oscillating pressure.

Piezoelectric Microphones and Hydrophones

The piezoelectric microphones and hydrophones contain piezoelectric elements that develop an emf in response to mechanical stress. In the early days, thin disks cut from natural or synthetic single crystals were used. Rochelle salt crystals were used because of their high sensitivity, but they deteriorated easily. Lithium sulfate crystals were more stable but less sensitive. Disks cut from natural quartz and tourmaline were used if stability and ruggedness were important, particularly for underwater applications, but they were relatively insensitive. Most present-day piezoelectric microphones and hydrophones utilize polarized ceramic piezoelectric elements of barium titanate, lead zirconate, or others, which have relatively high sensitivity and are available in various shapes, e. g., cylinders and spheres as well as disks (see Chapter 3 in Vol. 1A of the Mason collection [6]). Thin sheets of polarized piezoelectric plastics, such as polyvinyladine fluoride, which can be adhered to curved surfaces, are also available.

Most hydrophones use piezoelectric elements that can withstand the high static pressures existing at deep submergences. The hydrophone dimensions can vary from small 1-mm cylinders, used as "probe" hydrophones, to 10-cm (or larger) units. Depending on the design, the useful frequency range can extend down to a few hertz and up to 1 MHz. Bobber [12] presents a detailed discussion of underwater acoustic measurements.

Capacitance or Electrostatic Microphones

Capacitance or electrostatic microphones comprise a small variable capacitance consisting of a metallic stretched membrane or thin diaphragm, separated from a rigid back plate by an air gap perhaps 10^{-3} cm thick. The oscillating sound pressure results in oscillating deflections of the diaphragm, which change the width of the air gap and the associated capacitance. The variations in capacitance are usually detected by placing a fixed electric charge onto the electrodes of the capacitance through a high-resistance leak from a source of several hundred volts dc, so that variations in capacitance result in a varying potential difference across the capacitor. For very low-frequency measurements, the capacitance can be placed in one arm of an ac bridge so that the capacitance variations result in an amplitude-modulated carrier output from the bridge; alternatively, the varying capacitance can be used to control the frequency generated by an oscillator and thus provide a frequency-modulated carrier.

To eliminate the need for dc polarization, electret capacitance microphones have been developed utilizing polarized electrets holding a permanent charge on the metal electrodes.

Calibration of Microphones and Hydrophones

The ratio of the generated emf to the oscillating pressure generating the emf is called the *sensitivity* of the microphone. The sensitivity is determined by a procedure called *calibration*. Most microphones are calibrated by comparing their generated emf to that of another primary standard microphone subjected to the same sound pressure at various frequencies.

Microphones used as primary standards can be calibrated by one of several absolute methods. The method usually used nowadays is called a *reciprocity calibration*. This procedure involves two sets of measurements, the first being a comparison of the emf generated by the microphone with that of a second microphone when both are responding to the same sound source, and the second being a measurement of the emf of the microphone when the second microphone is itself used as a source (see Chapter 4.2 of Beranek [4]).

Another absolute calibration uses a Rayleigh disk, to be described subsequently, to determine the oscillating particle velocity. A third method uses an oscillating piston driven by a rotating cam and forming one wall of a small gas cavity; the known volume change is used to calculate the oscillating pressure, taking into account departure from the adiabatic gas law because of heat conduction at the walls of the cavity. Microphones having flat diaphragms can be calibrated by placing an auxiliary electrode close to the diaphragm and applying an alternating potential difference between this electrode and the diaphragm so as to develop a calculable alternating electrostatic pressure on the diaphragm.

Hydrophones are calibrated by methods generally the same as those used for microphones. An absolute method suitable only for hydrophones is to hold the hydrophone in a vibrating container of liquid; the vibration results in an oscillating pressure that can be calculated in terms of the measured acceleration of the container [l3].

Measurement of Density and Temperature Oscillations

Measurements of fluctuating density are usually done optically, utilizing variations in index of refraction associated with the density variations. The Debye–Sears apparatus is suitable for plane waves of sound, the spatially periodic density variations acting as an optical diffraction grating. The diffraction angles depend on the wavelength of the sound, and the intensity of the diffracted light depends on the magnitude of the oscillating refraction index and density (see Sect. 6.81 of Meyer and Neumann [3]). Optical holographic techniques with laser light sources are now being developed to indicate density oscillations. All these optical methods require relatively large-amplitude sounds and are useful only at high frequencies.

Temperature fluctuations can be observed with small probe wires whose electrical resistance fluctuates with the temperature fluctuations [14].

Measurements of Oscillating Particle Motion

Direct Visual Observation

The oscillating motion of fluid particles associated with ordinary sound is usually too small to be observed visually, even with a microscope. For example, a sound in air having a frequency of 1000 Hz and an oscillating rms sound pressure of $1\,\mathrm{N/m^2}$ (94 dB re 20 μPa) may be loud enough to cause some damage to hearing, but the oscillatory particle displacement is only about 3.7×10^{-5} cm. However, the oscillatory motion can be observed if the sound is strong

enough and the frequency low enough. Photographs through a microscope showing streaks of smoke particles oscillating in an intense sound field are shown on p. 292 of Stephens and Bate [2].

Rayleigh Disk

The Rayleigh disk is an early device used for absolute measurement of the oscillating particle velocity associated with sound in gases. A Rayleigh disk consists of a small thin disk, perhaps 1 cm in diameter, suspended by a fine fiber (see [1–3] or Sec. 4.3A of Beranek [4]). The oscillating particle velocity results in a steady torque on the disk, proportional to the mean-square velocity, tending to turn the disk broadside to the direction of the oscillating velocity. The suspension fiber is used as a balance to indicate the magnitude of the steady torque. Since the torque has been calculated theoretically, the Rayleigh disk permits an absolute measurement of the particle velocity. It is used nowadays mainly to provide an absolute calibration of microphones. A difficulty is that the disc must be shielded from even the slightest steady wind.

Electronic Measurement of Fluid Particle Velocity

The oscillating particle velocity can be measured using the small-amplitude relation between the instantaneous pressure gradient and particle acceleration, viz., the particle acceleration is equal to the negative of the pressure gradient divided by the fluid density. The pressure gradient can be determined using two closely spaced pressure-sensing microphones or hydrophones, spaced much more closely than the wavelength of sound at the highest frequency of interest. The pressure gradient is equal to the instantaneous difference in the two sensed pressures divided by their separation (see Sec. 5.2.2 of Fahy [21]). Since both individual pressures usually are much larger than their difference, the two sensors must be accurately matched in both amplitude and phase response.

In a liquid, the particle velocity can be determined with a single vibration sensor attached to the interior wall of a small, rigid, and waterproof spherical or cylindrical shell immersed in the medium. If the shell is neutrally buoyant and much smaller than the wavelength of sound at the highest frequency of interest, its instantaneous vibration duplicates the vibration of the liquid it displaces (see Sec. 5.12.1 of Bobber [12]). The shell vibration is sensed by the attached vibration pickup. If an accelerometer is used, its electrical output can be converted from instantaneous acceleration to velocity by passing its electrical signal through an integrating circuit; if the acceleration spectral density is calculated digitally, the acceleration density can be converted to velocity by dividing each value by the square of its radian frequency.

Electromechanical Pickup of Surface Vibration

The most common method for measuring the vibration of a solid is to mount an electromechanical transducer on the surface to convert the vibratory motion into an alternating emf. These transducers are called displacement pickups, velocity pickups, or accelerometers, depending on whether they generate emfs instantaneously proportional to the vibratory displacement or to its first or second time derivatives [15, 16]. Their sizes range from units smaller than 1 cm and with masses of only a few grams, to units having dimensions of 5 cm or more. Although the larger units are more sensitive, their large mass can perturb the vibration they are intended to measure.

Displacement pickups are used for measuring relatively low-frequency vibrations covering the frequency range from zero to a few hundred hertz. Velocity pickups are usually electromagnetic and cover the frequency range from perhaps 5 to 1000 Hz. Accelerometers for measuring very low-frequency accelerations utilize active transducers, whereas measurements in the frequency range 10 to 10000 Hz and higher utilize piezoelectric elements. The sensitivity of vibration pickups can be determined in various ways, including an absolute reciprocity calibration (see Chapter 18 of Harris and Crede (15]).

Variable magnetic reluctance pickups are available which sense the vibratory velocity of a magnetic surface without any part of the pickup being in contact with the surface. Conventional phonograph pickups can be used as low-mass vibration pickups. If the two stereophonic outputs are combined in phase, the pickup is sensitive to vibrations parallel to the axis of the pickup stylus, whereas if combined out of phase, the pickup is sensitive to vibrations perpendicular to the stylus axis. The sensitivity of phonograph pickups can be determined as is done with other vibration pickups, or with a phonograph test record as described in Sec. 10.5 of Olson [11].

Piezoresistive Strain Gauges

The development of semiconductor strain gauges, which are some 50 times more sensitive than the older resistance-wire gauges, has made it possible to measure oscillating strains at the surface of a solid associated with sounds of ordinary magnitude. The fluctuating resistance associated with the oscillating strain can be measured electrically in various ways (see Vaughan [17] or Chapter 17 of Harris and Crede [15]).

Fiber Optics

Fine optical fibers can be used to measure oscillatory strain and displacement. Minute oscillations in the optical path length through the fiber can be sensed as oscillations in the phase of monochromatic light transmitted through the fiber. Various methods for converting the oscillatory phase into an electrical signal are available. To measure acoustic pressure, the optical fiber is bonded to a diaphragm so as to sense the strain in the diaphragm caused by acoustic pressure acting on it (see Chapter 7 of Vol. 16 of Mason [6]).

Optical Measurements of Surface Vibration

Optical interferometry has been used to determine the magnitude of the vibratory displacement of the surface of a solid or liquid. The laser vibrometer splits a beam of monochromatic light into two beams, one of which is reflected off the vibrating surface and then recombined with the other beam. Motion of the reflecting surface changes the path length and causes the phase of the light beams to change relative to each other, so that the intensity of the recombined beam oscillates in response to vibrations of the reflecting surface. The method can be used to measure the vibration of a small region of a surface [18] as well as to display the distribution of vibration amplitude over a large surface vibrating in a complicated vibration pattern. If a laser is used as the light source, holographic reconstructions of the vibration pattern are possible (see Sec. 6.3.1 of Meyer and Neumann [3], also Brown [25]). A laser-Doppler vibrometer senses the Doppler shift in the frequency of the light reflected by the vibrating surface to indicate its velocity of vibration.

Miscellaneous

Other methods for observing and measuring oscillatory displacements are used occasionally. Sensitive flames provided one of the early methods for detecting the oscillating particle velocity associated with sound in gases (see p. 409B of Wood [1]). Oscillating strains in the interior of a photoelastic solid can be measured using polarized light [19]. The hot wire, involving an electrically heated fine wire whose temperature and resistance fluctuate in response to the cooling effect of a fluctuating flow past the wire, is the conventional device for measuring the fluctuating velocity in turbulent flows in gases; it can also be used to measure the oscillatory particle velocity associated with sound (see p. 440 of Wood [1]). Fluctuating electrochemical potentials detected by a pair of probes inserted in an electrolyte have been used to measure the oscillating particle velocity associated with sound (see p. 439 of Wood [1]). Finally, mention should be made of the use of sand grains or small ball bearings to indicate when the vertical oscillatory acceleration of a solid surface exceeds $1g$ (see p. 210 of Meyer and Neumann [3]); or of the use of sand to provide a visual display of a complicated pattern of surface vibration (see discussion of Chladni figures on p. 172 of Wood [1]). 1

Secondary Measurements

The velocity of propagation of sound can be determined by direct measurement of the transit time of a transient sound, or by observing various phenomena that depend on the wavelength of continuous sounds of constant frequency. These methods are discussed in Chapters V and IX of Herzfeld and Litovitz [20]; also methods for measuring acoustic absorption.

Acoustic Intensity

Measurements of the distribution of acoustic intensity around a sound source have become increasingly commonplace since the advent of digital spectrum analyzers. The intensity is a vector quantity whose component in a specified direction is the acoustic power passing through unit cross section perpendicular to the specified direction, this being equal to the time average of the product of the instantaneous sound pressure and the instantaneous particle velocity component in that direction.

For frequencies below 10 kHz, the intensity can be measured using an intensity probe consisting of two small microphones spaced a small fraction of an acoustic wavelength apart. The intensity can be shown to be proportional to the time average of the instantaneous product of the sum of the sound pressures at the two microphone positions multiplied by the difference between the two pressures. An alternative procedure, especially convenient if a two-channel digital spectrum analyzer is available, is to measure the complex cross-spectral density of the sound pressures at the two-microphone positions; the spectral density of the intensity at any frequency can be shown to be proportional to the imaginary part of the cross-spectral density of the two pressures at that frequency (see Fahy [21]). Although the sum-and-difference method may be used for measurements covering any band of frequencies, the cross-spectral technique is suitable only for narrow bands.

Since the two microphones of intensity probes must be much less than a wavelength apart, they are not suitable for measurements at frequencies above about 10 kHz. For ultrasonic frequencies, radiometers are used which sense the radiation pressure on a small disc in the sound field; the force on the disc is proportional to the acoustic intensity (see Fry and Dunn [5], Sec. 6.7 of Beyer [22], or Sec. 2.8 of Bobber [12]).

Acoustical Holography

Acoustical holography, as distinct from acoustical imaging, is used to determine the distribution of sound pressure or vibration velocity over the surface of a sound source. The procedure involves measurements of the magnitude and phase of the sound pressure at many positions in the sound field and summation of these pressures with appropriate weighting functions to reconstruct the pressure or velocity distribution over the source surface. Only the radiating or nonevanescent portion of the pressure or velocity distribution can be determined from measurements made several wavelengths removed from the source; but the entire distribution can be reconstructed from measurements in the "near field" (see, e. g., Maynard [23]).

See also: Acoustics; Piezoelectric Effect; Sound, Underwater; Transducers; Waves.

References

[1] A. B. Wood, *A Textbook of Sound*. G. Bell and Sons Ltd., London, 1957.
[2] R. W. B. Stephens and A. E. Bate, *Acoustics and Vibrational Physics*. Edward Arnold Ltd., London, 1966.
[3] E. Meyer and E. G. Neumann, *Physical and Applied Acoustics*. Academic Press, New York, 1972.
[4] L. Beranek, *Acoustical Measurements*. Acoustical Society of America, New York, 1989.
[5] W. J. Fry and F. Dunn, "Ultrasound: Analytical and Experimental Methods in Biological Research," in *Physical Techniques in Biological Research* (W. L. Nastuk, ed.), Vol. 4, Chapter 6. Academic Press, New York, 1962.
[6] W. P. Mason, ed., *Physical Acoustics: Principles and Methods*, 18 vols. Academic Press, New York, 1964–1988.
[7] A. P. G. Peterson and E. E. Gross, Jr., *Handbook of Noise Measurement*. General Radio Co., Concord, MA, 1972.
[8] J. T. Brock, *Application of B&K Equipment to Acoustic Noise Measurements*. Bruel and Kjaer, Naerum, Denmark, 1971.
[9] *Acoustic Handbook*, Application Note 100. Hewlett Packard Co., Palo Alto, CA.
[10] *ASA National Standards Catalog 26-2004*. Acoustical Society of America, New York, 2004.
[11] H. F. Olson, *Acoustical Engineering*. Van Nostrand-Reinhold, Princeton, NJ, 1957.
[12] R. J. Bobber, *Underwater Electroacoustic Measurements*. U.S. Government Printing Office, Washington, DC, 1970.
[13] F. Schloss and M. Strasberg, *J. Acoust. Soc. Am.* **34**, 1958 (1962).
[14] J. Hojstrup, K. Rasmussen, and S. E. Larsen, "Dynamic Calibration of Temperature Wires in Still Air," published in *DISA Information*, No. 20, DISA Electronics, Franklin Lakes, NJ, 1975.
[15] C. M. Harris and C. F. Crede, eds., *Shock and Vibration Handbook*. McGraw-Hill, New York, 1976.
[16] J. T. Brock, *Application of B&K Equipment to Mechanical Vibration and Shock Measurement*. Bruel and Kjaer, Denmark, 1972.
[17] J. Vaughan, *Application of B&K Equipment to Strain Measurement*. Bruel and Kjaer, Denmark, 1975.
[18] F. J. Eberhart and F. A. Andrews, *J. Acoust. Soc. Am.* **48**, 603 (1970).
[19] R. C. Dove and P. H. Adams, *Experimental Stress Analysis and Motion Measurement*. Charles E. Merrill, Columbus, Ohio, 1964.
[20] K. F. Herzfeld and T. A. Litovitz, *Absorption and Dispersion of Ultrasonic Waves*. Academic Press, New York, 1959.

[21] F. J. Fahy, *Sound Intensity*. Elsevier, New York, 1989.
[22] R. T. Beyer, *Nonlinear Acoustics*. U.S. Government Printing Office, Washington, 1974.
[23] J. D. Maynard, E. G. Williams, and Y. Lee, *J. Acoust. Soc. Am.* **78**, 1395 (1985).
[24] M. J. Crocker (ed.), "Acoustical Measurements" in *Encyclopedia of Acoustics*, Vol. 4, Part XVII. Wiley, New York, 1997.
[25] G. M. Brown *et al.*, *J. Acoust. Soc. Am.* **45**, 1166 (1969).

Acoustics

R. T. Beyer

Acoustics is the science of sound. What is sound? When you open your mouth and utter speech you are said to produce a sound. Under normal conditions a nearby person with so-called normal hearing says that he hears the sound. In its study of the production and reception of sound and its transmission through material media, acoustics is a branch of physics, though speech and hearing obviously involve biological elements.

Let us examine the physics of what happens when a person speaks. A disturbance is produced in the air in front of the mouth, involving a slight compression of the air or, alternatively, an increase in the air pressure of the order of $0.1\,\mathrm{N/m^2}$. This amount is about one-millionth of the normal atmospheric pressure. Since air is an elastic medium, it does not stay compressed but tends to expand again and hence produces a disturbance in the neighboring air. This in turn passes on the disturbance to the air adjoining it, and the result is a pressure fluctuation that moves through the air in the form of a sound wave. When the wave reaches the ear of an observer, it produces a motion of the eardrum, which in turn moves the little bones in the middle ear, and this movement communicates motion to the hair cells in the cochlea in the inner ear. The ultimate result, by a rather complicated biophysical process that is not yet completely understood, is the hearing of the sound.

While acoustics, as defined above, applies only to audible sound in air, its scope has been gradually expanded until today it encompasses mechanical waves and vibrations in all material media – solids, liquids, and gases, and includes waves of any frequency, as well as aperiodic disturbances, such as shocks and noise. A perusal of the pages of the *Journal of the Acoustical Society of America* will easily demonstrate that this is the case.

It is convenient to divide this discussion of acoustics into three parts: the production, the transmission, and the reception of sound.

Production

Any change in stress or pressure leading to a local change in density or displacement from equilibrium in an elastic material medium can serve as a source of sound. We have already mentioned the human vocal mechanism as an example. All sound sources in practical use involve the vibrations of solids, liquids, or gases. Such a vibration is an oscillation of stress or pressure with a definite frequency. The unit of frequency used in acoustics is the hertz

(Hz), which is one complete cycle per second. The standard musical instruments are common sources of sound of more or less definite frequency. In sophisticated scientific and technological applications the sound source is called a transducer. Those in standard use are electroacoustic in character, that is, they depend on electrical action to produce mechanical vibrations. An example is the electrodynamic loudspeaker, in which an alternating electric current in a coil of wire placed in a magnetic field produces oscillatory motion in the coil. This motion is communicated to a membrane, whose resulting vibrations are radiated as sound. Another commonly used electroacoustic transducer is based on the piezoelectric effect, according to which certain crystals (e. g., quartz) can be made to vibrate when placed in an oscillating electric field.

Certain ceramic materials, such as barium titanate and lead zirconate, can be polarized by the use of an applied electric field and can thereafter serve as transducers in the same way as that described for quartz. In addition, the change in shape and dimensions of a piece of magnetic material, like nickel, when placed in a magnetic field, can also be used in the construction of transducers. This is known as the magnetostrictive effect. All these types of transducers are useful, not only in the production of sound but also in its reception.

The rapid flow of air over a rough surface or through a nozzle produces sound that can vary over a wide range of frequency and intensity. An example is the aerodynamic sound from a jet engine on an airplane. Sounds of this character are commonly known as noise, one of the chief problems of environmental acoustics in the late twentieth century.

An exciting new field has been the optoacoustic generation of sound by the thermoacoustic effect of absorption of the energy of a pulsed light source (laser beam) by a medium, which thereupon emits acoustic pulses. It is of interest to note that this effect was first observed by Alexander Graham Bell in 1881 (without the use of a laser!).

Transmission

Sound demands a material medium for its transmission from place to place. As previously mentioned, the propagation takes place by means of wave motion, a good visual example of which is provided by a wave on the surface of water. A sound wave travels with a definite velocity, depending on the type of wave, the physical nature of the medium, and the temperature. Through air at standard room temperature (20 °C) sound travels with the velocity 344 m/s, a value that increases as the temperature is raised. Under similar conditions, sound travels through water with a velocity somewhat more than four times as great. The velocity of sound in a highly elastic solid like steel is even greater, being as high as 6000 m/s.

A sound wave represents the transmission of mechanical energy through the medium in which the sound travels. The measure of this transmission is called the intensity of the sound wave. It is defined as the average flow of energy per unit time through a unit area of the surface through which the sound passes (with direction normal to the surface). The strictly scientific unit of sound intensity is the watt per square meter. In practice, however, this unit is replaced by a system in which 10 times the logarithm to the base 10 of the ratio of the given intensity to that corresponding to minimum audibility is taken as the number of decibels (abbreviated dB) represented by the sound in question. Ordinary conversational speech at a distance of 1 m from the speaker has an intensity of about 60 dB, whereas the intensity in the neighborhood of a jet airplane with the engine running can be as high as 140 dB. Sound of intensity in excess of

90 dB at the human ear can be harmful to that delicate and valuable organ of hearing. Sounds of very high intensity are called macrosonic and form the subject of what is termed nonlinear acoustics.

As sound spreads out from a source, its intensity decreases with distance. For a highly localized source from which the sound travels in all directions, the intensity varies inversely as the square of the distance from the source. Sound intensity also falls off with distance traveled through a process known as absorption, a dissipative action in which the energy being transmitted by the sound wave is gradually converted into heat. When a sound wave strikes an obstacle, it undergoes reflection and refraction, following the same laws as those that hold for light waves. The acoustic echo is a well-known phenomenon.

The simplest type of sound wave is the periodic one, in which at any given point in the medium being traversed by the wave the disturbance (e. g., the excess pressure for a wave in air) varies periodically between a minimum and a maximum value and back again a certain number of times a second. As indicated earlier in the case of sound source vibrations, this number is called the frequency of the wave and is measured in hertz. Another important quantity characterizing a periodic sound wave is the wavelength, or the distance between successive points in a wave train at which the disturbance has the same magnitude and is doing the same thing (i. e., is either increasing or decreasing in magnitude). The wavelength is related to the frequency by the simple but fundamental relation that the wavelength is equal to the velocity divided by the frequency. For a given velocity, a long wavelength means a low frequency and vice versa.

Reception

The frequency of a periodic sound wave determines, in a rather complicated way, the pitch at which it is heard. High pitch corresponds to high frequency. Sounds of frequency below 20 Hz are not normally heard by human beings even at very high intensity. These sound waves are called infrasonic. The upper frequency limit of audible sound varies markedly. For young people, this upper limit is at about 20 000 Hz. For older adults, this figure drops to 10 000 Hz or even lower. This phenomenon is known as presbycusis. Frequencies above the audible range are known as ultrasonic; they can be generated and detected into the gigahertz range (10^9 Hz). Ultrasound has many practical applications in such areas as sound signaling, metallurgy, and medicine.

The most important receiver of audible sound in human experience is, of course, the ear, a marvelously sensitive mechanism that can normally detect sound of intensity as low as $10^{-12}\,\mathrm{W/m^2}$ and can stand intensity as high as $1\,\mathrm{W/m^2}$ before pain ensues and possible ear damage develops. The normal ear is most sensitive at around 2000 Hz. Loudness, though of course related to intensity in such a way that it increases with intensity, is a subjective quantity connected with the biological response of the listener. The unit of loudness developed by the psychophysicists is the sone, defined as the loudness produced by a tone of 1000 Hz at 40 dB above the minimum audible threshold. A loudness scale in sones has been established by the statistical study of the hearing of a large number of individuals.

The principal artificial sound receiver in common use is the microphone, of which there are many varieties, mainly based on the electroacoustic effects used for transducers in general, as mentioned earlier. Their widespread employment in the audio industry – radio, television, public address systems, sound recording and reproduction, hearing aids, etc. – is well known.

More detailed information about the topics mentioned here involving the application of acoustics to fields like architectural acoustics, ultrasonics, vibration problems, and underwater sound, can be found in the relevant articles in the encyclopedia.

See also: Acoustical Measurements; Acoustics, Architectural; Acoustics, Linear and Nonlinear; Acoustics, Physiological; Acoustoelectric Effect; Musical Instruments; Noise, Acoustical; Sound, Underwater; Ultrasonic Biophysics; Ultrasonics; Vibrations, Mechanical; Waves.

Bibliography

L. L. Beranek, *Acoustics*. Reprint of the 1954 edition, with revisions. Acoustical Society of America, New York, 1986. (I)

F. V. Hunt, *Acoustics*. Reprint of the 1954 edition. Acoustical Society of America, New York, 1982. (I)

L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 3rd ed., Wiley, New York, 1982. (I)

James Lighthill, *Waves in Fluids*. Cambridge University Press, Cambridge, 1978. (A)

R. B. Lindsay, ed., *Acoustics: Historical and Philosophical Development*. Dowden, Hutchinson k Ross, Stroudsburg, PA., 1973. (E)

Iain G. Main, *Vibrations and Waves in Physics*. 2nd ed. Cambridge University Press, Cambridge, 1984. (I)

P. M. Morse and K. U. Ingard, *Theoretical Acoustics*. Reprint of the 1968 edition. Princeton University Press, Princeton, 1986. (A)

A. D. Pierce, *Acoustics, An Introduction to its Physical Principles and Applications*. 1981 text reprinted by the Acoustical Society of America, New York, 1989. (I)

J. R. Pierce, *Almost All About Waves*. MIT Press. Cambridge, MA, 1974. (E)

R. W. B. Stephens and A. E. Bate, *Acoustics and Vibrational Physics*, 2nd ed. St. Martin's Press, New York, 1966. (I)

J. W. Strutt (Lord Rayleigh). *The Theory of Sound*. 1877 (reprinted by Dover Publications, New York, 1945). (A)

For information on current progress in acoustical research the *Journal of the Acoustical Society of America* may be consulted. This is published monthly by the American Institute of Physics for the Acoustical Society of America.

Acoustics, Architectural

T. D. Northwood

Architectural acoustics may be defined as the science of acoustics applied to the design of buildings. It thus derives from such diverse disciplines as physics, psychology, the arts, architecture, and engineering, the techniques needed to produce acoustical environments acceptable to the building occupants. The objective is twofold: to bring to the occupants the sounds they wish to hear, and to protect them from the sounds they do not wish to hear. The first of these tasks is the more attractive and creative one, but the exercise will be successful only if the second task is handled with equal care.

The profession of architectural acoustics may be said to have been inaugurated by Wallace Clement Sabine (1868–1919), a professor of physics at Harvard University, who in 1895 was assigned the task of curing the acoustical defects of a new lecture theater at the university. In solving that problem he went on to evolve the first quantitative theory of reverberation processes in rooms, and applied it to the solution of problems in many theaters and concert halls, the most famous of which was Symphony Hall in Boston.

For studies of these rooms the instrumentation available to Sabine consisted of a set of organ pipes, a stop watch, and his two ears. He used these with great ingenuity to determine what is now known as the "reverberation time" of a room and, by extension, the "sound absorption coefficients" of typical room surfaces. From his collected data for many rooms he was able to develop the Sabine reverberation theory that is still used by acousticians everywhere.

Another technique first developed by Sabine was the use of two-dimensional models of complex surfaces, utilizing a spark technique and schlieren photography to observe the progression of waves in such models. Today, 80 years later, the same topics are dealt with in more sophisticated ways, but there is scarcely any problem in architectural acoustics on which Sabine did not make a useful contribution.

Most of the sounds of interest in architectural acoustics, such as speech or music, consist of a sequence of transient impulses, and these brief transients constitute the "message." Following the progression of one of these sounds in a room, we can identify three phases: first, the direct transmission of sound through the air from source to listener; then the first reflections from the various room surfaces; and finally, the reverberant field, composed of multiple reflections from the room surfaces.

The acoustical design of a room involves a consideration of these three phases. The direct transmission does not carry much energy very far, but it is an important reference, establishing the location of the source and the timing of the sequence of sounds. The first-order reflections, if they arrive soon enough, provide useful reinforcement of the direct transmission. Strong delayed reflections (echoes) tend to garble the sequence of transient sounds and thus interfere with the perception of speech. Even without forming discrete echoes, the ensemble of multiple reflections forms a persistent reverberation that also can interfere with perceptions of the original sound sequences.

In halls used for speech the design objective is to shape the room surfaces so as to provide short-delay reflections that reinforce the direct sounds and thus extend the range for perception of intelligible speech. For one-way communication the range can be further extended by electroacoustic reinforcement. At the same time the reverberation time must be limited so that it does not blur the sequence of sounds. For musical performances, on the other hand, the hall plays a more active role. A certain optimal amount of reverberation provides a desirable blending and smoothing of musical sounds. In addition, the presence of early reflections, reaching the listeners especially from lateral directions, are essential to the listeners' impression of the space enveloping them and the performers. For the performers an additional requirement is that each should hear his own part in relation to the whole and to what is heard in the hall. To achieve this sort of ambience for every listener and performer and every kind of music is a delicate task, which becomes increasingly difficult as halls and audiences become larger.

The reduction of unwanted or disturbing noises in an enclosure involves the opposite sort of techniques: surfaces are designed to absorb rather than reflect sounds, so that the first-order reflections and the reverberant portions of the sound are made negligible. The direct sound

is reduced by interposing partitions or screens between source and listener, or by a partial or complete enclosure around the source. These principles apply with minor variations to all large spaces containing noise sources: factory areas, open-plan offices, restaurants, air terminals, and so on. In many instances the noise may be intrusive speech, for which the annoyance increases rapidly with intelligibility. Reduction of intelligibility can be accomplished either by reducing the level of the transmitted speech or by increasing the level of "background" noise. The latter sometimes takes the form of background music, which is deemed a lesser evil than the intrusive speech. This question of detection of intrusive noises in the presence of acceptable "background" noise is implicit in most noise control problems.

Sound insulation between rooms is mainly a function of the separating wall or floor. In the design of such partition elements it is usual to distinguish between airborne sounds, which travel through the air to reach the partition, and structure-borne sounds, which begin as vibrations in the structure itself. With respect to airborne sounds the most important virtues of a simple partition are that it be impermeable and heavy. For a given total weight, however, it is found more effective to use a multiplicity of relatively independent layers.

The sound transmission loss of a partition increases systematically with frequency in the incident sound, and this must be taken into account in arriving at a representative performance rating for partitions. For sounds such as speech and music (live or by way of radio or television) the customary figure of merit is the sound transmission class (STC), which emphasizes the importance of middle- and high-frequency components.

With respect to structure-borne sounds, such as footsteps and machinery vibrations, the first requirement is structural discontinuity between the vibrating surface and the contiguous structure. A floor, for example, may be composed of a finished floor panel separated from the main structural slab by a soft layer. In the case of machinery it is usual to provide an individually designed mounting tuned to filter out the driving frequency of the machine. Other problems characteristic of modern buildings, especially office buildings, are noise propagation in ventilation ducts and in continuous plenum spaces over suspended acoustical ceilings, and noise produced by plumbing appliances. These are but special cases of the noise mechanisms already described.

See also: Noise, Acoustical.

Bibliography

L. L. Beranek, *Music, Acoustics and Architecture*. Wiley, New York, 1962. (I)

L. Cremer, H. A. Muller, and T. J. Schultz (translator, English edition), *Principles and Applications of Room Acoustics*. Applied Science Publishers, 1982.

Vern O. Knudsen, and Cyril M. Harris, *Acoustical Designing in Architecture*. Wiley, New York, 1950. (E)

Heinrich Kuttruff, *Room Acoustics*. Taylor & Francis, New York, 2000. (A)

Anita Lawrence, *Architectural Acoustics*. Elsevier, Amsterdam, 1970. (I)

T. D. Northwood, *Architectural Acoustics, Benchmark Papers in Acoustics, Vol. 10*. Dowden, Hutchinson & Ross, Stroudsburg, PA, 1977.

W. C. Sabine, *Collected Papers on Acoustics*. Peninsula, 1993. (E)

Acoustics, Linear and Nonlinear

J. E. Greenspon

Definitions and Nomenclature

Linear acoustics is the study of sounds of relatively small amplitude. Nonlinear acoustics is the study of sounds of relatively large amplitude. The physical phenomena associated with what is called relatively small and what is termed relatively large are, in part, the topics to be discussed in this article. The categories of linear and nonlinear acoustics have to be discussed independently of architectural acoustics, underwater acoustics, physical acoustics, and musical acoustics since linear and nonlinear refers to an amplitude characteristic of the sound whereas the other categories are associated with the acoustic environment and the mechanisms causing the sound. The study of linear acoustics is sometimes referred to as the study of infinitesimally small amplitude waves or just the study of ordinary sound waves.

Nonlinear acoustics is sometimes referred to as finite-amplitude acoustics, high-intensity acoustics, or macrosonics. There are many problems in fluid mechanics which are nonlinear because of the nature of the mathematics involved in their solution. Those physical problems in fluid mechanics which are primarily concerned with the sounds produced are acoustic problems. From a mathematical standpoint the problems involving the classical linear partial differential equation of sound waves (which will be discussed later in this article) are termed linear acoustic problems whereas the ones involving nonlinear partial differential equations (some examples being discussed later) are nonlinear acoustic problems.

Physical Phenomena in Acoustics

Sound Propagation in Air, Water, and Solids

Many practical problems are associated with the propagation of sound waves in air or water. Sound does not propagate in free space but must have a dense medium to propagate. Thus, for example, when a sound wave is produced by a voice, the air particles in front of the mouth are vibrated, and this vibration, in turn, produces a disturbance in the adjacent air particles, and so on.

If the wave travels in the same direction as the particles are being moved, it is called a longitudinal wave. This same phenomenon occurs whether the medium is air, water, or a solid. If the wave is moving perpendicular to the moving particles, it is called a transverse wave.

The rate at which a sound wave thins out, or attenuates, depends to a large extent on the medium through which it is propagating. For example, sound attenuates more rapidly in air than in water, which is the reason that sonar is used more extensively under water than in air. Conversely, radar (electromagnetic energy) attenuates much less in air than in water, so that it is more useful as a communication tool in air.

Sound waves travel in solid or fluid materials by elastic deformation of the material, which is called an elastic wave. In air (below a frequency of 20 kHz) and in water, a sound wave travels at constant speed without its shape being distorted. In solid material, the velocity of the wave changes, and the disturbance changes shape as it travels. This phenomenon in solids is called dispersion. Air and water are for the most part nondispersive media, whereas most solids are dispersive media.

Reflection, Refraction, Diffraction, Interference, and Scattering

Sound propagates undisturbed in a nondispersive medium until it reaches some obstacle. The obstacle, which can be a density change in the medium or a physical object, distorts the sound wave in various ways. (It is interesting to note that sound and light have many propagation characteristics in common: The phenomena of reflection, refraction, diffraction, interference, and scattering for sound are very similar to the phenomena for light.)

Reflection. When sound impinges on a rigid or elastic obstacle, part of it bounces off the obstacle, a characteristic that is called reflection. The reflection of sound back toward its source is called an echo. Echoes are used in sonar to locate objects under water. Most people have experienced echoes in air by calling out in an empty hall and hearing their words repeated as the sound bounces off the walls.

Refraction and Transmission. Refraction is the change of direction of a wave when it travels from a medium in which it has one velocity to a medium in which it has a different velocity. Refraction of sound occurs in the ocean because the temperature of the water changes with depth, which causes the velocity of sound also to change with depth. For simple ocean models, the layers of water at different temperatures act as though they are layers of different media. The following example explains refraction: Imagine a sound wave that is constant over a plane (i. e., a plane wave) in a given medium and a line drawn perpendicular to this plane (i. e., the normal to the plane) which indicates the travel direction of the wave. When the wave travels to a different medium, the normal bends, thus changing the direction of the sound wave. This normal line is called a ray.

When a sound wave impinges on a plate, part of the wave reflects and part goes through the plate. The part that goes through the plate is the transmitted wave. Reflection and transmission are related phenomena that are used extensively to describe the characteristics of sound baffles and absorbers.

Diffraction. Diffraction is associated with the bending of sound waves around or over barriers. A sound wave can often be heard on the other side of a barrier even if the listener cannot see the source of the sound. However, the barrier projects a shadow, called the shadow zone, within which the sound cannot be heard. This phenomenon is similar to that of a light that is blocked by a barrier.

Interference. Interference is the phenomenon that occurs when two sound waves converge. In linear acoustics the sound waves can be superimposed. When this occurs, the waves interfere with each other, and the resultant sound is the sum of the two waves, taking into consideration the magnitude and the phase of each wave.

Scattering. Sound scattering is related closely to reflection and transmission. It is the phenomenon that occurs when a sound wave envelops an obstacle and breaks up, producing a sound pattern around the obstacle. The sound travels off in all directions around the obstacle. The sound that travels back toward the source is called the backscattered sound, and the sound that travels away from the source is known as the forward-scattered field.

Standing Waves, Propagating Waves, and Reverberation

When a sound wave travels freely in a medium without obstacles, it continues to propagate unless it is attenuated by some characteristic of the medium, such as absorption. When sound waves propagate in an enclosed space, they reflect from the walls of the enclosure and travel in a different direction until they hit another wall. In a regular enclosure, such as a rectangular room, the waves reflect back and forth between the sound source and the wall, setting up a constant wave pattern that no longer shows the characteristics of a traveling wave. This wave pattern, called a standing wave, results from the superposition of two traveling waves propagating in opposite directions. The standing wave pattern exists as long as the source continues to emit sound waves. The continuous rebounding of the sound waves causes a reverberant field to be set up in the enclosure. If the walls of the enclosure are absorbent, the reverberant field is decreased. If the sound source stops emitting the waves, the reverberant standing wave field dies out because of the absorptive character of the walls. The time it takes for the reverberant field to decay is sometimes called the time constant of the room.

Sound Radiation

The interaction of a vibrating structure with a medium produces disturbances in the medium that propagate out from the structure. The sound field set up by these propagating disturbances is known as the sound radiation field. Whenever there is a disturbance in a sound medium, the waves propagate out from the disturbance, forming a radiation field.

Coupling and Interaction between Structures and the Surrounding Medium

A structure vibrating in air produces sound waves, which propagate out into the air. If the same vibrating structure is put into a vacuum, no sound is produced. However, whether the vibrating body is in a vacuum or air makes little difference in the vibration patterns, and the reaction of the structure to the medium is small. If the same vibrating body is put into water, the high density of water compared with air produces marked changes in the vibration and consequent radiation from the structure. The water, or any heavy liquid, produces two main effects on the structure. The first is an added mass effect, and the second is a damping effect known as radiation damping. The same type of phenomenon also occurs in air, but to a much smaller degree unless the body is traveling at high speed. The coupling phenomenon in air at these speeds is associated with flutter.

Deterministic (Single-Frequency) Versus Random Linear Acoustics

When the vibrations are not single frequency but are random, new concepts must be introduced. Instead of dealing with ordinary parameters such as pressure and velocity, it is necessary to use statistical concepts such as autocorrelation and cross-correlation of pressure in the time domain and auto- and cross-spectrum of pressure in the frequency domain. Frequency is a continuous variable in random systems, as opposed to a discrete variable in single-frequency systems. In some acoustic problems there is randomness in both space and time. Thus statistical concepts have to be applied to both time and spatial variables.

Some of the Practical Problems in Linear and Nonlinear Acoustics

The majority of problems in architectural and musical acoustics involve small-amplitude sounds and therefore come under the category of linear acoustics. In fact, as will be seen later, all sounds which are below the threshold of pain[1] are well within the linear acoustic region. Most problems in submarine and surface-ship sound radiation which involve interaction of a vibrating structure with water are also linear acoustic problems since only very small motions of both the structure and the medium in contact with it, are involved. The problem of propagation of explosive waves is a large-amplitude, nonlinear acoustic problem. The transition from subsonic to supersonic flow and the associated production of shock waves is also a problem in nonlinear acoustics.

References to Fundamental Developments in Linear and Nonlinear Acoustics

There are a number of books on acoustics which contain various degrees of mathematical sophistication. The only presentations on acoustics that the writer would classify as elementary are the accounts given in standard physics texts such as Duff *et al.* [1] and Sears and Zemansky [2]. There are many of these texts, and they usually give good elementary discussions of acoustics. The reader can almost pick any college physics text at random, and it will usually give a reasonably good presentation of elementary acoustics. Of the books which are devoted entirely to acoustics, the writer would rate as number one the treatise of Lord Rayleigh [3]. The writer would classify Rayleigh's two volumes as advanced. The reader who is reasonably familiar with elementary differential equations would do well to go through such books as Kinsler and Frey [4], Beranek [5], and Morse [6] before going to Rayleigh's work. Two very fine books of recent vintage which the writer would classify as advanced are the one written by Skudrzyk [7] and the treatise by Morse and Ingard [8]. The most recent reference is Beyer's excellent work on nonlinear acoustics [13].

The most complete mathematical treatment to date on linear and nonlinear wave motion is contained in a very recent book by G. B. Whitham [9] which is devoted entirely to this subject. A most easily readable account of the mathematics associated with both linear and nonlinear acoustics is given by R. B. Lindsay [10]. In this book Lindsay devotes an entire chapter to discussing sound waves in fluids, and he considers both small- and large-amplitude motions in a general way. Short but readily readable accounts of finite-amplitude waves in acoustics and their comparison to linear waves are given in both Rayleigh's [3] and Lamb's [11] texts on sound. Finally, the *Journal of the Acoustical Society of America*, which is a monthly technical publication published by the American Institute of Physics, has two of its sections devoted entirely to work in general linear acoustics and nonlinear acoustics or macrosonics. Finally, a review of the important aspects of linear acoustics is contained in a recent article [14].

Basic Governing Equations of Acoustics

The fundamental concepts of both linear and nonlinear acoustics can be covered for the one-dimensional case in a simple but general way. From these notions the reader will be able to

[1]The threshold of pain is the intensity at which an average person starts to feel pain in his ears and at which permanent damage to the ears could result if exposure to the sound is sustained.

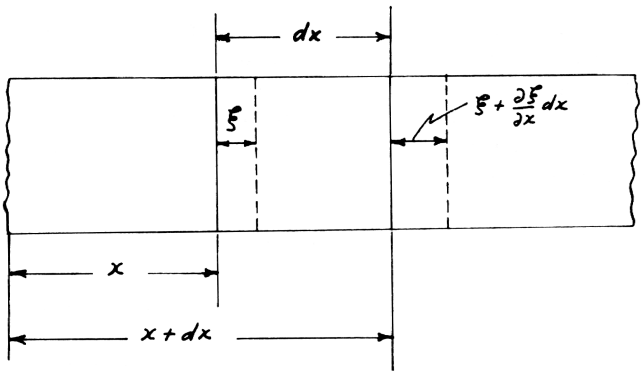

Fig. 1: Propagation of plane one-dimensional sound wave.

grasp a physical feeling concerning sounds produced in both the linear and nonlinear acoustic regimes. The extension to two and three dimensions is much more mathematically complicated, but involves no new physical concepts. For this discussion the writer will follow a combination of the presentation on plane sound waves by Kinsler and Frey [4] and Lamb [11]. Let

x = equilibrium coordinate of a given particle of the medium from some given origin (see Fig. 1);
ξ = particle displacement along the x axis from the equilibrium position;
u = velocity of the particle $= \partial\xi/\partial t$;
ρ = instantaneous density at any point;
ρ_0 = equilibrium density of the medium;
S = condensation at any point, which is defined as $S = (\rho - \rho_0)/\rho_0$ so that $\rho = \rho_0(1+S)$;
$\bar{p}$ = instantaneous pressure at any point;
p_0 = equilibrium pressure in the medium;
p = excess pressure (which is the acoustic pressure) at any point, thus $p = \bar{p} - p_0$;
C = velocity of propagation of the wave.

It will be assumed that the wave is plane, i. e., all particles on the plane at x have the same displacement, and that this displacement is a function only of space coordinate x and time t. We employ three basic concepts to derive three independent equations and then we will combine these equations into a single equation. The first of these concepts is the conservation of mass. We apply the principle of conservation of mass to a cross-sectional area $\bar{S}$ of the undisturbed fluid contained between planes positioned at x and $x + \mathrm{d}x$. The mass of this undisturbed fluid is $\rho_0 \bar{S}\,\mathrm{d}x$. Assume that upon passage of a sound wave the plane at x is displaced a distance ξ to the right (see Fig. 1) and that the plane at $x + \mathrm{d}x$ is displaced a distance $(\xi + \mathrm{d}\xi$ $[\mathrm{d}\xi = (\partial\xi/\partial x)\,\mathrm{d}x]$. The volume enclosed is therefore changed to $\bar{S}\,\mathrm{d}x(1 + \partial\xi/\partial x)$ and its mass is $\rho\bar{S}\,\mathrm{d}x(1 + \partial\xi/\partial x)$. Equating the original mass to the new mass (i. e., employing the conservation of mass) gives

$$\rho\bar{S}\,\mathrm{d}x = \left(1 + \frac{\partial\xi}{\partial x}\right) = \rho_0\bar{S}\,\mathrm{d}x\,; \tag{1}$$

noting that $\rho = \rho_0(1+S)$ we obtain

$$(1+S)\left(1+\frac{\partial\xi}{\partial x}\right) = 1 \; . \tag{2}$$

If both S and $\partial\xi/\partial x$ are *small* (i. e., $\ll 1$), then we can neglect the product of S and $\partial\xi/\partial x$ and Eq. (2) reduces to

$$S = -\frac{\partial\xi}{\partial x} \; . \tag{3}$$

This equation is known as the equation of continuity.

The second basic concept that shall be employed is Newton's equation of motion of the element. The resultant pressures on the two faces of the volume element $\bar{S}\,\mathrm{d}x$ will be slightly different from each other producing a net force which will accelerate the element. The external force acting on each face is equal to the product of the pressure and the area of the face. The net force acting upon $\bar{S}\,\mathrm{d}x$ in the positive x direction is

$$\mathrm{d}F_x = \left[p - \left(p + \frac{\partial p}{\partial x}\,\mathrm{d}x\right)\right]\bar{S} = -\frac{\partial p}{\partial x}\,\mathrm{d}x\bar{S} \; . \tag{4}$$

This net force is equal, by Newton's second law of motion, to the product of the element's mass and its acceleration, thus

$$-\frac{\partial p}{\partial x} = \rho_0 \frac{\partial^2\xi}{\partial t^2} \; . \tag{5}$$

One other relation is necessary to combine the equation of continuity (2) or (3) with the equation of motion (5), and this is a relation between the pressure and the density. If we assume that the process is isothermal, then the temperature does not change during the passage of the sound wave and the pressure density relation follows Boyle's law, i. e.,

$$\frac{\bar{p}}{p_0} = \frac{\rho}{\rho_0} \; . \tag{6}$$

However, in ordinary sound waves the condensation S changes sign so frequently and the temperature, consequently, rises and falls so rapidly, that there is no time for transfer of heat between adjacent portions of the fluid. The flow of heat has hardly gone from one element to another before its direction is reversed; therefore the conditions are close to being adiabatic, i. e., no heat transfer occurring, and the pressure density relation becomes

$$\frac{\bar{p}}{p_0} = \left(\frac{\rho}{\rho_0}\right)^{\gamma} \; , \tag{7}$$

where γ is the adiabatic constant having a value of about 1.4 for air.

For large-amplitude adiabatic waves we combine (2), (5), and (7) as follows:

From before we had, by definition

$$\rho = \rho_0(1+S) \; ; \tag{8}$$

Eq. (2) gives

$$(1+S) = \frac{1}{1+\partial\xi/\partial x} \, .$$

Thus

$$\rho = \frac{\rho_0}{1+\partial\xi/\partial x} \, ; \tag{9}$$

Eq. (7) gives

$$\frac{\bar{p}}{p_0} = \left(\frac{\rho}{\rho_0}\right)^{\gamma} \tag{10}$$

or

$$\frac{\bar{p}}{p_0} = \left(\frac{\rho_0}{\rho_0(1+\partial\xi/\partial x)}\right)^{\gamma} = \left(\frac{1}{1+\partial\xi/\partial x}\right)^{\gamma} \, . \tag{11}$$

but

$$\bar{p} = p_0 + p \, . \tag{12}$$

So

$$\bar{p} = \frac{p_0}{(1+\partial\xi/\partial x)^{\gamma}} \, . \tag{13}$$

So

$$\frac{\partial p}{\partial x} = \frac{\partial \bar{p}}{\partial x} = -p_0\gamma\left(1+\frac{\partial\xi}{\partial x}\right)^{-\gamma-1}\frac{\partial^2\xi}{\partial x^2} \, . \tag{14}$$

Thus (5) gives

$$\frac{p_0\gamma}{\rho_0}\frac{\partial^2\xi/\partial x^2}{(1+\partial\xi/\partial x)^{\gamma+1}} = \frac{\partial^2\xi}{\partial t^2} \, . \tag{15}$$

Let

$$C^2 = p_0\gamma/\rho_0 \, . \tag{16}$$

Then the final adiabatic equation of motion for large amplitude nonlinear plane waves is

$$C^2\frac{\partial^2\xi/\partial x^2}{(1+\partial\xi/\partial x)^{\gamma+1}} = \frac{\partial^2\xi}{\partial t^2} \, . \tag{17}$$

If the process is isothermal, $\gamma = 1$.

For very small motions the condensation S is very small. When we employ this assumption and eq. (3) the equation of motion (17) reduces to the linear equation of sound waves

$$C^2\frac{\partial^2\xi}{\partial x^2}=\frac{\partial^2\xi}{\partial t^2}\ . \tag{18}$$

The general solution of Eq. (18) can be written in the form

$$\xi=f_1(Ct-x)+f_2(Ct+x)\ . \tag{19}$$

This is easily verified by substituting (19) into (18) and performing the indicated differentiations. The solution (19) states that small displacements propagate with velocity C without change of shape. If we neglect any losses in the medium, a sound disturbance which is started at a given point will propagate with the velocity C without being distorted as it propagates. This is not true of large-amplitude waves which satisfy Eq. (17). To see this consider the transitional region between small- and large-amplitude waves in a long straight tube in which a piston (at $x=0$) is made to move in some arbitrary manner described by

$$\xi=f(t) \tag{20}$$

If we expand the denominator of Eq. (17) and neglect the terms greater than the second derivative of ξ, we obtain the following approximate equation for the transition region:

$$\frac{\partial^2\xi}{\partial t^2}=C^2\frac{\partial^2\xi}{\partial x^2}-(\gamma+1)C^2\frac{\partial\xi}{\partial x}\frac{\partial^2\xi}{\partial x^2}\ . \tag{21}$$

By means of a procedure adopted by Airy [11] an approximate solution can be constructed. First we note that the solution of (18) is

$$\xi=f(t-x/C)\ . \tag{22}$$

Substituting this value of ξ in the last term on the right-hand side of (21) we obtain

$$\frac{\partial^2\xi}{\partial t^2}=C^2\frac{\partial^2\xi}{\partial x^2}-\frac{1}{2}(\gamma+1)\frac{\partial}{\partial x}\left\{f'\left(t-\frac{x}{C}\right)\right\}^2\ . \tag{23}$$

The solution of this equation is

$$\xi=f\left(t-\frac{x}{C}\right)+\frac{\gamma+1}{4C^2}x\left\{f'\left(t-\frac{x}{C}\right)\right\}^2\ . \tag{24}$$

Now assume that the motion of the piston at one end of the tube is simple harmonic, i. e.,

$$f(t)=a\cos\omega t\ . \tag{25}$$

Formula (24) then gives

$$\xi=a\cos\omega\left(t-\frac{x}{C}\right)+\frac{(\gamma+1)\omega^2a^2}{8C^2}x\left\{1-\cos 2\omega\left(t-\frac{x}{C}\right)\right\}\ . \tag{26}$$

It is thus seen that the displacement of any particle is no longer simple harmonic at $x>0$ but is distorted from the original wave shape.

Sound Intensity and Pressure in Linear and Nonlinear Acoustics – Sine Waves (Siren-Type Sounds)

In order to give the reader a physical feeling of the relative sounds coming from various processes in both linear and nonlinear acoustics let us first compute the intensity of sine waves that might come from any physical source such as a siren or transducer of some sort. The sound intensity is defined as the time average of the power flow per unit area (or rate at which energy is transmitted per unit area [6]) as follows:

$$I = \frac{1}{T}\int_0^T p\dot{\xi}\,\mathrm{d}t \tag{27}$$

where I is the intensity, p is the excess pressure, and $\dot{\xi}$ is the velocity of the medium particle. The value of T is taken as some arbitrary value. In the case of sine waves we will take T to be a number of periods of the wave. In the last section it was found that the pressure p was (for adiabatic processes)

$$\bar{p} = p_0(\rho/\rho_0)^{\gamma}\,, \tag{28}$$

(for an isothermal process $\gamma = 1$) and that the density ρ was connected to the condensation S by the relation

$$S = \rho/\rho_0 - 1\,. \tag{29}$$

It has been found [11] that the velocity $\dot{\xi}$ has the following values for a general nonlinear wave:

$$\dot{\xi} = \pm C\log(1+S) \qquad \text{(for an isothermal process);} \tag{30}$$

$$\dot{\xi} = \pm\frac{2C}{\gamma-1}[1-(1+S)^{(\gamma-1)2}] \qquad \text{(for an adiabatic process).} \tag{31}$$

For both processes it is easily verified that for the linear case ($S \ll 1$)

$$\dot{\xi} = \pm CS\,. \tag{32}$$

For sine waves of small amplitude

$$I = \frac{1}{T}\int_0^T p_0\gamma C S^2\,\mathrm{d}t\,, \tag{33}$$

so

$$\frac{I}{p_0 C} = \gamma\frac{1}{T}\int_0^T S^2\,\mathrm{d}t\,,$$

$$S = S_0\cos\omega t\,,$$

therefore

$$\begin{aligned}\frac{I}{p_0 C} &= \gamma\frac{S_0^2}{2} \qquad \text{(for the adiabatic case),} \\ &= S_0^2/2 \qquad \text{(for the isothermal case).}\end{aligned} \tag{34}$$

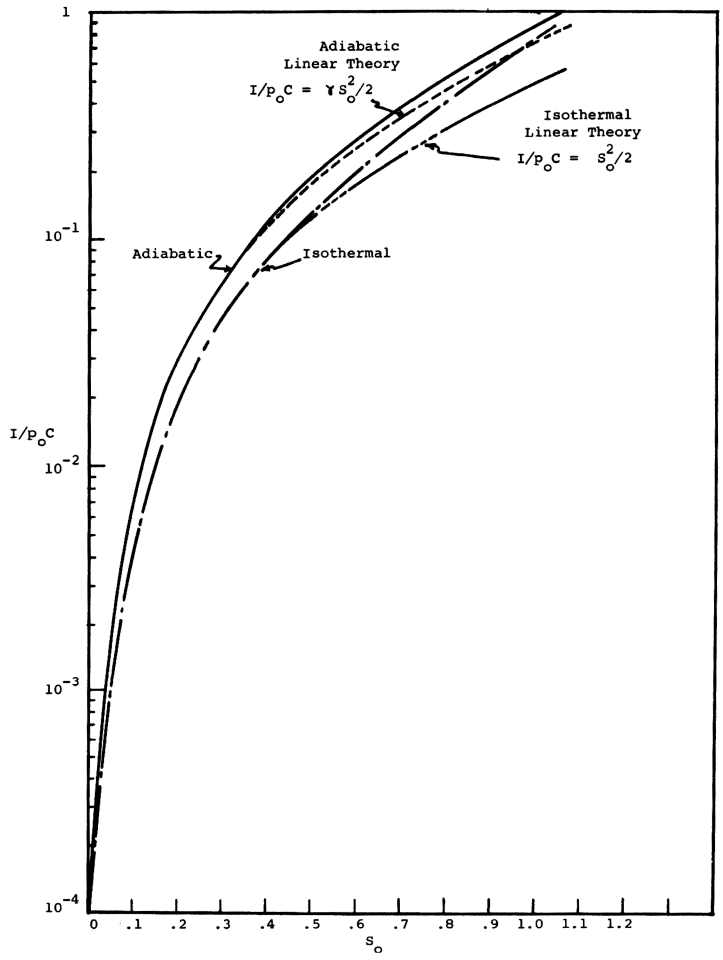

Fig. 2: Intensity as a function of condensation for sine waves.

For the nonlinear case we just substitute (28), (29), (30), or (31) into (27) and integrate numerically. The result is given in Fig. 2 along with the linear approximation. It is seen that the linear theory holds for values of S_0 less than 0.4.

The sound pressure generated for either the linear or nonlinear adiabatic case is

$$p = p_0(1+S)^{\gamma} - p_0 \ .$$

Assuming that the ambient pressure is sea level pressure (i. e., 1 atmosphere) the pressure level in dB relative to $0.0002\,\mathrm{dyn/cm^2}$, which is a standard measure for pressure levels in air, is as follows:
Sound Pressure Level = $2\log_{10} p + 74\,\mathrm{dB}$ relative to $0.0002\,\mathrm{dyn/cm^2}$ where p is expressed in $\mathrm{dynes/cm^2}$. Table 1 gives the sound pressure levels for the adiabatic case as a function of the condensation S.

Table 1: Sound pressure level (in dB relative to $0.0002\,\mathrm{dynes/cm^2}$)[a] as a function of condensation.

Condensation S	Sound Pressure Level(SPL)	Condensation S	Sound Pressure Level(SPL)
10^{-10}	-3	0.9	196
10^{-9}	17	1.0	197
10^{-8}	37	2.0	203
10^{-7}	57	3	207
10^{-6}	77	4	209
10^{-5}	97	5	211
10^{-4}	117	6	213
10^{-3}	137	7	214
10^{-2}	157	8	215
10^{-1}	177	9	216
0.2	183	10	217
0.3	186	12	219
0.4	189	14	220
0.5	191	16	221
0.6	193	18	222
0.7	194	20	223
0.8	195		

[a] $1\,\mathrm{dyn/cm^2}$ is called a microbar, written µbar.

In order to get a physical feeling of the order of magnitude of the sound as a function of condensation S_0, Fig. 3 illustrates the intensity as a function of frequency showing the threshold of hearing and the threshold of pain [6]. The vertical lines, which are not frequency dependent, illustrate the intensity values for various values of condensation amplitude S_0. The intensity level is defined as follows:

$$\text{Intensity Level} = 10\log_{10}(10^9 I) = 90 + 10\log_{10} I\,, \tag{35}$$

where I is expressed in ergs per square centimeter per second. Thus the intensity level is in dB relative to 10^{-10} microwatts per square centimeter per second. Note that at 3000 Hz (i. e., 3000 cyles per second) the threshold of hearing is about -6 dB relative to 10^{-10} microwatts/cm^2. This corresponds to a sine wave condensation of about 10^{-10} in air and would even correspond to a much smaller S_0 in water. This means that an average person could hear a sound in air at a frequency of 3000 Hz if the condensation amplitude was as small as 10^{-10}. The threshold of pain, which Fig. 3 shows is almost frequency independent, corresponds to condensations of the order of 5×10^{-4} in air. Thus the entire auditory area (i. e., from threshold of hearing to threshold of pain) is well within the region of linear acoustics. Comparing Table 1 with Fig. 3 it is seen that the sound pressure level corresponding to a condensation of 10^{-10}, which corresponds to the intensity at threshold of hearing, is about -3 dB relative to $0.0002\,\mu$bar and that the sound pressure level at threshold of pain is of the order of 125 dB relative to $0.0002\,\mu$bar.

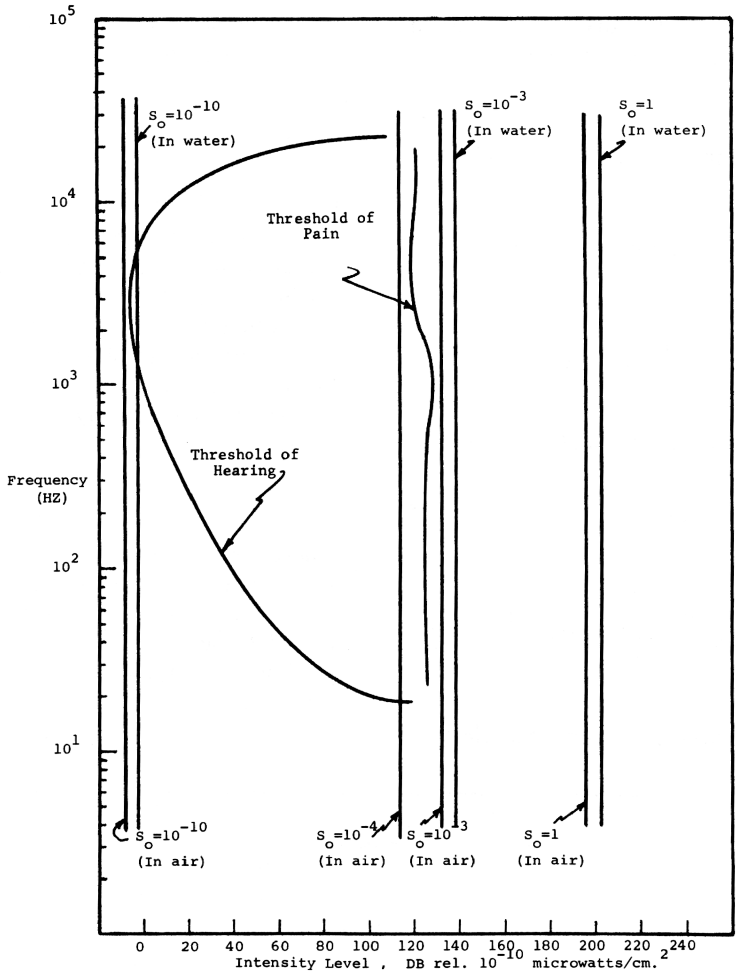

Fig. 3: Intensity of sine waves as a function of frequency.

Linearity and Nonlinearity as Related to Elasticity of the Medium

In any physical system which involves forces and displacements it is the usual understanding that if the force is a linear function of displacement, then the system is linear and when the force becomes a nonlinear function of displacement, the system becomes nonlinear. The same concept holds true in acoustics. In order to see this use Eqns. (28), (29), (30), (31) and write the pressure in terms of velocity as follows:

$$\frac{\bar{p}}{p_0} = \left(1+\frac{\gamma-1}{2C}\dot{\xi}\right)^{2\gamma/(\gamma-1)} \quad \text{(Adiabatic Case);} \tag{36}$$

$$\frac{\bar{p}}{p_0} = e^{\xi/C} \quad \text{(Isothermal Case).}$$

Table 2: Pressure and condensation for blast waves.

		dB relative to 0.0002 μbar			dB relative to 0.0002 μbar	
$\bar{R}$	P_1/P_0	P_1	S_1	P_2/P_0	P_2	S_1
0.10	67.9	231	6.0	585	249	32.2
0.2	20.4	220	3.9	146	237	17.1
0.3	7.3	211	2.6	37.7	225	9.0
0.4	3.5	205	1.7	15.3	218	5.1
0.5	2.1	200	1.1	9.4	213	3.2
0.6	1.4	197	0.8	6.1	210	2.1
0.8	0.77	192	0.5	2.6	202	1.1
1	0.51	188	0.33	1.3	196	0.66
2	0.16	178	0.11	0.36	185	0.22
3	0.089	173	0.063	0.19	180	0.12
4	0.062	170	0.044	0.13	176	0.087
5	0.047	167	0.033	0.095	174	0.066
6	0.037	165	0.027	0.077	172	0.053
8	0.026	162	0.019	0.054	169	0.039
10	0.020	160	0.014	0.040	166	0.028
20	0.0087	153	0.0062	0.018	159	0.012
30	0.0054	149	0.0039	0.011	155	0.0077
40	0.0039	146	0.0028	0.0079	152	0.0056
50	0.0030	144	0.0022	0.006l	150	0.0043
60	0.0025	142	0.0018	0.0050	148	0.0035
80	0.0018	139	0.0010	0.0036	145	0.0021
100	0.0014	137	0.00082	0.0028	143	0.0016
500	0.00024	122	0.00017	0.00049	128	0.00033
1000	0.00012	116	0.000082	0.00023	121	0.00017

For sine waves $\xi = \xi_0 \sin \omega t$, so

$$\left(\frac{\bar{p}}{p_0}\right)_{\max} = \left(1 + \frac{\gamma-1}{2}\frac{\omega\xi_0}{C}\dot{\xi}\right)^{2\gamma/(\gamma-1)} \quad \text{(Adiabatic Case);} \tag{37}$$

$$\left(\frac{\bar{p}}{p_0}\right)_{\max} = e^{\omega\xi_0/c} \quad \text{(Isothermal Case).}$$

but $p = \bar{p} - p_0$. In Fig. 4 the dimensionless pressure ratio, p/p_0 is plotted as a function of the dimensionless deformation parameter $\omega\xi_0/C$. It is seen in Fig. 4 that when the displacements are small the pressure is a linear function of displacement. As the displacements become larger the medium becomes stiffer and a small change in displacement gives a proportionally larger increase in pressure. The nonlinear region starts at $\omega\xi_0/C \approx 0.1$. This corresponds to an $S \approx 0.1$, which is a much better criterion for linearity than Fig. 2.

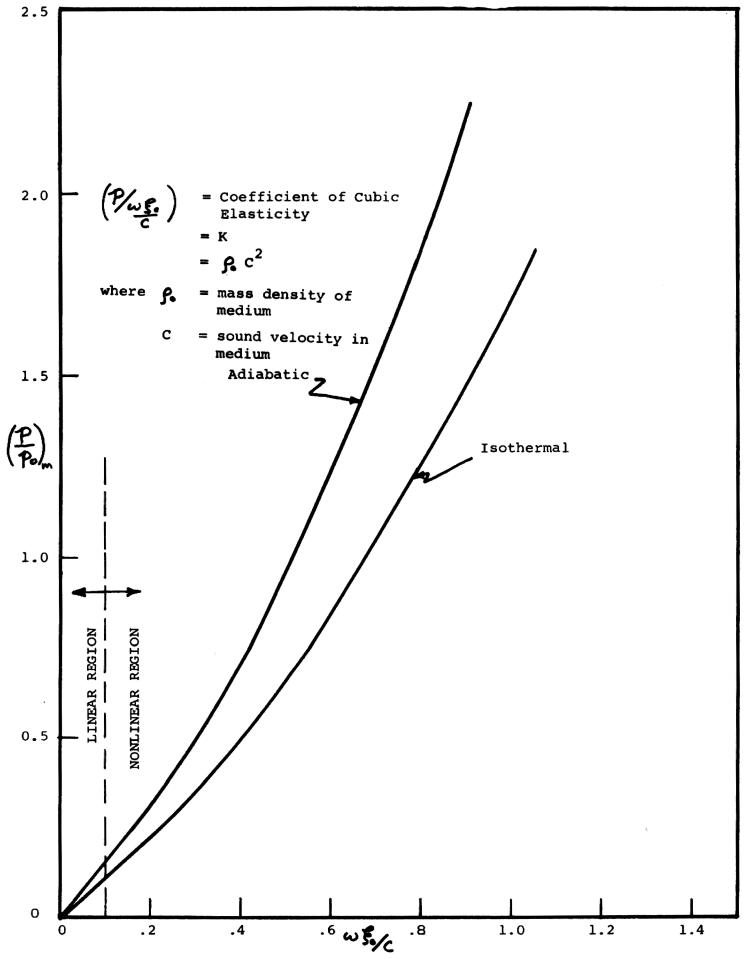

Fig. 4: Sound pressure as a function of displacement for sine waves.

Explosive Waves

Consider next the sounds generated by explosive-type waves. Unless the charge weight of the explosive is exceptionally small, the sounds generated by the explosion are in the nonlinear acoustic region at the point of the explosion, but as they die out away from the explosion, they become smaller and enter the linear region. In his book *Explosions in Air* [12], Baker gives the necessary information needed to estimate the sounds obtained from the explosions. The waves produced by the explosions are shock waves and have steep wave fronts. Therefore they have to be analyzed in a different manner from the relatively simple way that has been explained previously in this article. However, the results for shock waves will be given here in order to give the reader a feeling of the sounds produced by these waves as a function of the explosive weight, the distance from the explosion, and the type of explosive. Table 2 gives the pressure levels and condensation values for explosive waves.

Table 3: Characteristics of explosives.

Explosive	Specific Energy E/M (in./lb$_m$)
Pentolite	20.50×10^6
TNT	18.13×10^6
RDX	21.5×10^6

In Table 2 $\bar{R} = Rp_0^{1/3}/E^{1/3}$, where R is the distance from the explosion, p_0 is the ambient pressure, and E is the total energy in the explosive charge. P_1, S_1 correspond to the pressure and condensation in the incident blast wave (i. e., the wave coming directly from the explosion to the point at which the pressure is being measured) and P_2, S_2, correspond to the pressure and condensation in the reflected wave. The reflected wave parameters are determined from the assumption that the reflection occurs from a rigid wall.

Table 3 contains the energy characteristics for several of the more important explosives [12]. In Table 3 the symbol # represents a pound of force (the weight) while lb$_m$ represents a pound of mass. In order to obtain the energy for a given weight of explosive we use the following relation:

$$E = (E/M)(W/g) \; .$$

Thus the energy contained in 1000 # of TNT is

$$E = 18.13 \times \frac{1000}{386} = 47 \times 10^6 \# \text{ in.}$$

The value of R at 1 mile from an explosion of 1000 # of TNT would be

$$\bar{R} = 5280 \times 12 \times \left(\frac{14.7}{47 \times 10^6} \right) = 430$$

Examination of Table 2 indicates that for $\bar{R} = 430$ the incident blast pressure is of the order of 125 dB relative to 0.0002 µbar, and the reflected pressure of the order of 130 dB relative to 0.0002 µbar, both being around the threshold of pain for hearing. However, since the blast wave acts only for a very short time, the pain in the ear will undoubtedly not have time to develop for such an explosion.

The values of condensation S contained in Tables 1 and 2 compare very favorably for sound pressure levels less than 200 dB relative to 0.0002 µbar, i. e., for S values less than 1. For the larger S values the characteristics of the shock wave front enter the problem and there is no longer any correlation between the value contained in the two tables.

See also: Acoustics, Architectural; Fluid Physics; Nonlinear Wave Propagation; Shock Waves And Detonations; Sound, Underwater.

References

[1] A. W. Duff (ed.), *Physics for Students of Science and Engineering*. Blakiston Co., Philadelphia, PA, 1937. (E)

[2] F. W. Sears and M. W. Zemansky, *College Physics*. AddisonWesley, Reading, MA, 1960. (E)

[3] Lord Rayleigh, *The Theory of Sound*. Dover, New York, 1945. (A)

[4] L. E. Kinsler and A. R. Frey, *Fundamentals of Acoustics*. Wiley, New York, 1962. (I)

[5] L. L. Beranek, *Acoustics*. McGraw-Hill, New York, 1954. (I)

[6] P. M. Morse, *Vibration and Sound*. McGraw-Hill, New York, 1948. (I)

[7] E. Skudrzyk, *The Foundations of Acoustics*. Springer Verlag, Wein, 1971. (A)

[8] P. M. Morse and K. U. Ingard, *Theoretical Acoustics*. McGrawHill, New York, 1968. (A)

[9] G. B. Whitham, *Linear and Nonlinear Waves*. Wiley, New York, 1974. (A)

[10] R. B. Lindsay, *Mechanical Radiation*. McGraw-Hill, New York, 1960. (I)

[11] H. Lamb, *The Dynamical Theory of Sound*. Dover, New York, 1960. (I)

[12] W. E. Baker, *Explosions in Air*. University of Texas Press, Austin Texas, 1973. (I)

[13] R. T. Beyer, *Nonlinear Acoustics*. Naval Sea Systems Command, 1974. (I)

[14] J. E. Greenspon "Acoustics, Linear", *Encyclopedia of Physical Science and Technology, Vol. I*, p. 135. Academic Press, San Diego, California 1987.

Acoustics, Physiological

J. Tonndorf[†]

Physiological acoustics, an expression coined after Helmholtz's *Physiological Optics* (1856), concerns itself with analytical assessments of the reception of sound by the ear and of the further processing of the signals thus received at the various levels of the central auditory nervous system.

This work requires close cooperation between physiologists and anatomists. Formal analyses are made possible by inputs from fluid mechanics (inner-ear dynamics); biochemistry (chemical events underlying the responses of the sense organ); systems analysis, including electrical network theory, and advanced statistics (electrical responses in both the organ and the central nervous system, equivalent network analysis); and many others.

The following brief description of the current state of the art in physiological acoustics must include some anatomical remarks.

The ear is traditionally divided into three parts: the outer, middle, and inner ears (Fig. 1). Outer and middle ears help shape the acoustic signal for optimal reception by the inner ear and its receptor cells.

The outer ear consists of the pinna and ear canal. The ear canal, a continuation of the funnel-shaped pinna, is an open tube terminated at its inner end by the tympanic membrane. The latter seals off the middle ear from the outside. The middle ear, an air-filled cavity (with a volume of approximately $2\,\mathrm{cm}^3$), can be aerated via the Eustachian tube, a connection with the upper pharynx. The tympanic membrane, a thin elastic structure, vibrates in response to sound. It is connected by a mechanical transmission chain, consisting of a series of three small

[†] deceased

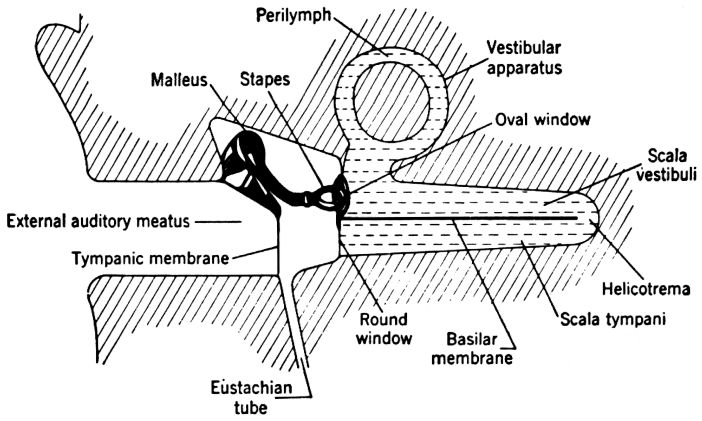

Fig. 1: Highly schematic outline of the ear (from van Békésy and Rosenblith, 1951).

leverlike ossicles, to one of the two "windows" of the inner ear, the oval window. The other one, the round window, looks likewise into the middle ear, but it is simply closed over by a membrane. The inner ear is deeply hidden in hard dense bone. Like that of all vertebrates, it is a fluid-filled cavity, very complex in shape, whence its classical name, otic labyrinth. In addition to the auditory receptor organ, the inner ear houses five other organs that have to do with spatial orientation and maintenance of equilibrium, the so-called vestibular system. (The "vestibule" is part of the inner ear.)

The functions of outer and middle ears are (a) protection of the inner ear (foremost by virtue of their position, which shields the inner ear against physical insults; then there are middle-ear nonlinearities occurring at high amplitudes; and finally there are the two small middle-ear muscles; their reflex contraction on sound exposure attenuates middle-ear transmission); (b) optimization of transmission of acoustic energy into the inner ear (the impedance of the fluid-filled inner ear is much higher than that of the air on the outside; thus, impedance matching is needed; this task is accomplished by a series of mechanical transformers that are completely integrated with one another, involving both the outer and middle ears).

In addition to the route just described, i. e., via tympanic membrane–ossicular chain–oval window (so-called air conduction), mechanoacoustic energy may also be brought into the ear when the bones of the head are set in vibration by contact with a vibrating object. This bone-conduction mode plays an important role in the clinical diagnosis of hearing disorders.

The auditory receptor organ is located in the lower, or cochlear portion of the inner ear. (The vestibular system occupies the upper portion.) The cochlea is a long (35 mm) but narrow bony chamber, coiled up $2\frac{1}{2}$ times like a snail shell (which is what cochlea means in Latin). It is subdivided by a number of membranes into a system of three compartments ("scalae") (Fig. 2). The receptor organ, the organ of Corti, lies in the middle scala and stretches over the whole length of the cochlea. It consists of about 16 000 to 20 000 receptor cells, the hair cells, distributed in a characteristic four-row pattern and held in place by a supporting-cell structure. Each hair cell carries a tuft of 80–100 sensory hairs on its top surface. Sensory cells of this kind are also found in other receptor organs; all of them are stimulated by a mechanical, sideways deflection of their hairs. In the cochlea, the necessary mechanism is provided

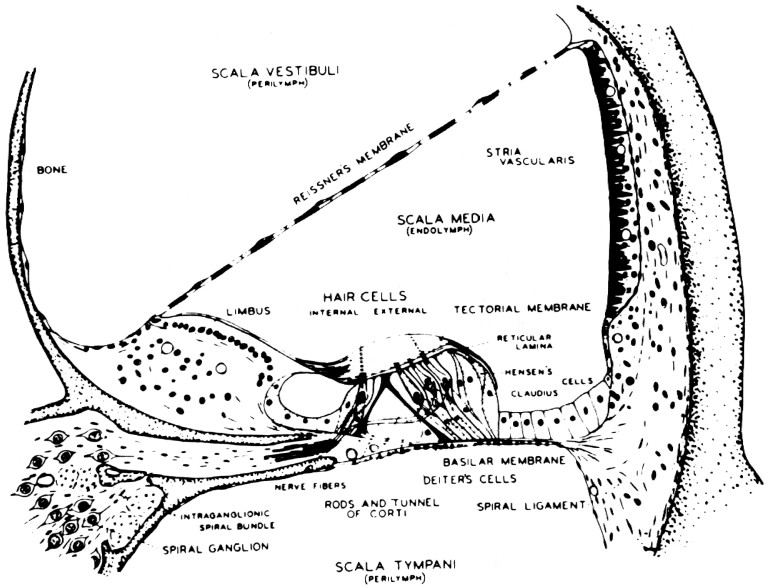

Fig. 2: Cross section of scala media (guinea pig) (from Davis *et al.*, 1953).

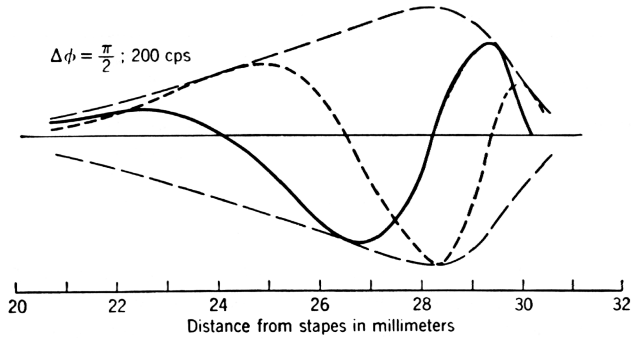

Fig. 3: Schematic outline of cochlear traveling wave at two instances, 90° apart in phase. Waves are moving from base to apex, i. e., left to right in the figure; amplitudes overstated (from von Békésy and Rosenblith, 1951).

by the connection of the sensory hairs with a membrane that covers the organ of Corti from side to side for its full length. This tectorial membrane (Fig. 2) executes a sliding ("shearing") movement across the organ that leads to the deflection of the sensory hairs. This constitutes the ultimate mechanical input to the hair cells; it is the final one in a series of interlinked mechanical events that are elicited when mechanoacoustic energy enters the cochlea, usually via the oval window. Such signals set up a series of displacements of the cochlear membranes, including the basilar membrane, on which the organ of Corti is situated (Fig. 2). These displacements progress along the basilar membrane in the manner of traveling waves (Fig. 3), invariably in the direction from the cochlear base to its apex. In this respect, the cochlea acts

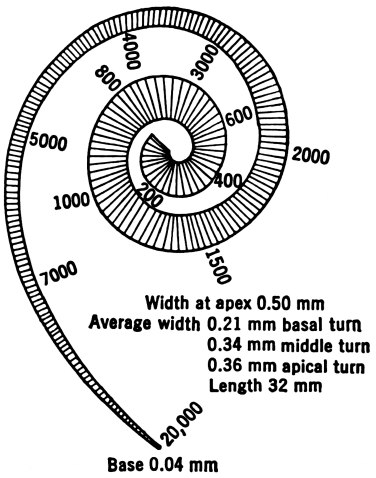

Fig. 4: Outline of the human basilar membrane. Note its width increasing with distance. Places of frequency maxima as indicated (from Stuhlmann, 1943).

like a mechanical delay line. For a given sine-wave input, the traveling-wave mechanism creates a displacement maximum at a distinct, frequency-dependent place along the membrane ("place principle"). For high frequencies, the maxima are formed near the cochlear base, and as frequency goes lower this place moves toward the apex in a systematic manner (Fig. 4). Therefore, given sine-wave signals stimulate only limited regions along the basilar membrane and hence distinct groups of hair cells. This tonotopic relation is maintained throughout the entire central auditory system. It enables the latter to process frequencies, by substituting place for frequency, up to approximately 20 kHz, while single fibers of the auditory nerve are capable of responding to frequencies not higher than $3-4$ kHz.

At its bottom end, each hair cell is supplied by a fiber (or fibers) of the cochlear (sensory) nerve. There are some 25 000 to 30 000 individual fibers, and their pattern of distribution to the hair cells is complex but systematic. The mechanoacoustic signals received by the hair cells elicit – in a multiple-step operation that bridges the hair-cell–nerve junction – nerve-action potentials (bursts of negative electrical impulses) of essentially the same kind as those observed in fibers of all other nerves. Their energy source is chemoelectric and inherent to the nervous system. These potentials represent a signal code particularly suited for neural transmission and processing.

After entering the part of the brain known as the brain stem, the auditory fibers run in well-defined tracts that go sequentially from one central station ("nucleus") to the next higher one (Fig. 5). This chain of nuclei finally terminates in the auditory portion of the cortex located in the temporal lobe of the brain. In each nucleus, the signal carried by the incoming fibers is switched onto a new set of outgoing fibers. The underlying networks are structurally very intricate, allowing for complex signal processing. Most, but by no means all, fibers in each tract cross over to the opposite side, so that the left brain receives primarily signals from the right ear, and vice versa. The left auditory cortex appears to handle mainly signals for which analytical processing is of importance (e. g., speech signals), while the right one is primarily concerned with signals of emotional importance (e. g., music). In the region where the two

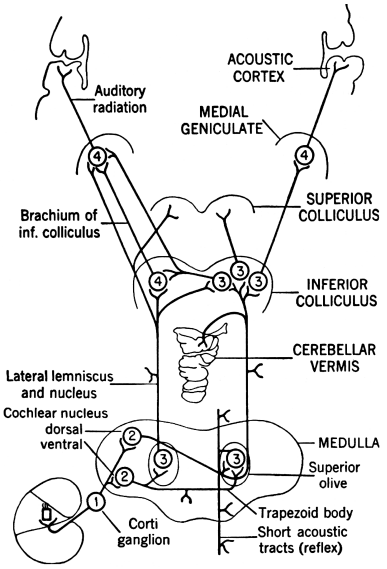

Fig. 5: Schematic outline of the afferent auditory system (from Davis, 1951).

tracts first cross over, signals received by the two ears are brought together in a set of special nuclei (superior olive; Fig. 5) initiating neural processing that concerns spatial hearing. These ascending ("afferent") fiber tracts are paralleled by a similar system of descending ("efferent") tracts that appear to exert a central (feedback) control upon the input at various levels, mainly at the hair-cell–nerve junction.

See also: Fluid Physics; Network Theory, Analysis and Synthesis; Statistics.

Bibliography

G. van Békésy, *Experiments in Hearing*. Krieger, 1980.

G. van Békésy and W. Rosenblith, in *Experimental Psychology*, (S. S. Stevens, ed.), Chapter 27. Wiley, New York, 1951.

P. Dallos, *The Auditory Periphery*. Academic, New York, 1974.

H. Davis, in *Experimental Psychology* (S. S. Stevens, ed.) Chapter 28. Wiley, New York, l95l.

O. Stuhlmann, Jr., *Introduction to Biophysics*, Wiley, New York, 1943.

Acoustoelectric Effect

E. M. Conwell

The acoustoelectric effect is the appearance of a dc electric field when an acoustic wave propagates in a medium containing mobile charges. It was first named and discussed, theoretically, by R. H. Parmenter in 1953. As pointed out by G. Weinreich, who first detected it experimentally, it is an example of the general phenomenon of "wave–particle drag," of which the

operation of a linear accelerator and the motion of driftwood toward a beach are other examples. Although first proposed for metals, it is only significant in semiconductors. It can be quite strong in piezoelectric semiconductors, such as CdS, ZnO and GaAs, but is also seen in nonpiezoelectric semiconductors. In the piezoelectric case, the wave and mobile charges interact through the electric field arising from the strain associated with the wave. This interaction is significant when the electric field is longitudinal, i. e., parallel to the wave propagation direction, and in what follows we assume that the type of wave (i. e., longitudinal or shear) and sample orientation have been chosen to make this the case. In nonpiezoelectric media the interaction is through the shift of carrier energy induced by the strain, the so-called deformation potential. At the frequencies ordinarily used, $\ll 1\,\mathrm{GHz}$, the latter interaction is much smaller than that in piezoelectric materials. In either case, however, because of the interaction the passage of an acoustic wave through the medium causes a periodic spatial variation of the potential energy of the charge carriers. If the mean free path l of the carriers is small compared to the acoustic wavelength λ, as is the case for most of the usual (ultrasonic) frequency range, this results in a bunching of the carriers in the potential-energy troughs. Since the wave is propagating, it drags the bunches along with it. This is the origin of the acoustoelectric field and clearly also causes attenuation of the wave. The effect is stronger, the stronger the bunching. Thus it is very weak in ordinary metals, where space-charge effects prevent appreciable bunching. It is enhanced in those nonpiezoelectric semiconductors where space-charge effects are minimized by either (1) the presence of both positively and negatively charged carriers or (2) the existence of different groups of carriers (many-valley band structure) whose energy is affected differently by the strain so that they bunch in different phases of the wave. Being sensitive to the number of free carriers, acoustoelectric interaction has proven to be a powerful tool in nondestructive testing of semiconductors, testing that does not even require contacts.

The foregoing discussion suggests that if an external electric field were applied to the sample to give the carriers a drift velocity, v_d, greater than the wave velocity, v_s, the carriers should drag the wave, i. e., the wave should be amplified. This conjecture was verified experimentally on CdS samples at frequencies of 15 and 45 MHz by Hutson, McFee, and White in 1962. They found acoustic gain for shear waves at fields greater than $700\,\mathrm{V/cm}$, at which field v_d equals the shear wave velocity. For not too large acoustic wave amplitudes it was possible to account quite well for the size of the gain and its variation with frequency, etc., with a linear phenomenological theory taking into account the currents and space charge produced by the piezoelectric fields that accompany the acoustic wave. The gain, called *acoustoelectric gain*, is found to be low at low frequencies (less than the conductivity relaxation frequency σ/ε), where the carriers can redistribute themselves quickly enough to essentially cancel out the piezoelectric field. It peaks at the frequency for which the acoustic wavelength is of the order of the Debye length, where the bunching is optimum. In a fairly strong piezoelectric like CdS acoustoelectric gains as high as $40\,\mathrm{dB/cm}$ have been found, leading to consideration of this effect for practical use as an amplifier (*acoustoelectric amplifier*). This type of amplification has been found particularly useful for amplification of surface acoustic waves (SAWs). Because the amplitude of a SAW decays exponentially with distance below a free surface, the surface acts as a waveguide for such a wave. SAWs at microwave frequencies are easily introduced into a piezoelectric material such as $LiNbO_3$ by coupling in microwaves through a suitable transducer. A similar transducer can reconvert the SAWs into microwaves. The SAW velocity being smaller by a factor of 10^5 than electromagnetic wave velocity, a short length

of a SAW-propagating material is useful as a delay line and for various types of signal processing. For maximum utility the losses of the SAWs in the guide are conveniently overcome by incorporating acoustoelectric gain. If the piezoelectric material is insulating, this may be accomplished by providing a conducting layer, e. g., Si, in contact with it. Alternatively, a conducting piezoelectric material, e. g., GaAs, may be used to support both the SAW and the electrons drifting in the electric field.

The theory and effects considered so far are linear in the sound amplitude provided it is small, i. e., small enough to bunch only a small fraction of the carriers. At large sound amplitude new effects appear. When two acoustic waves are present, the interaction of the piezoelectric field of one with the bunched carriers of the other may result in the generation of the difference or sum frequency of the two. As a special case of this, for a single large wave the interaction of its piezoelectric field with its own carrier bunches results in the generation of a dc current, called the *acoustoelectric current*, flowing in a direction opposite to the usual or Ohmic current, and may result in the generation of the second harmonic.

When an outside acoustic wave is not introduced, application to a highly conducting sample of CdS or ZnO, for example, of a field high enough to make $v_\mathrm{d} > v_s$ causes a large amplification of the thermal equilibrium acoustic waves or flux present in the sample. This gives rise to unusual behavior, the exact nature of which depends on the details of sample inhomogeneity and contacts. One possibility is that immediately after the high field is applied a dc current will flow of the expected magnitude for the Ohmic resistance, i. e., the resistance displayed for $v_\mathrm{d} < v_s$, but in a short time the current will drop to a much smaller value and remain there. The smaller value is due to the opposing acoustoelectric current arising from the flux amplification. Another frequently seen possibility is the onset of strong current oscillations with a period equal to the length of the sample divided by v_s. These oscillations are due to the creation close to the cathode of a domain or narrow region of high acoustic flux density, which typically moves down the sample with velocity v_s. When it exits at the anode, the current rises to its Ohmic value, strong flux generation begins again at the cathode, and the process is repeated.

The above discussion has been couched in terms appropriate for $l \ll \lambda$, since this is the case for most of the experimental situations studied. However, both linear and nonlinear processes have been studied with a microscopic theory that does not make this restriction. In this theory acoustic gain, for example, may be thought of as due to an excess of stimulated phonon emission by the carriers over absorption.

See also: Piezoelectric Effect; Semiconductors.

Bibliography

N. G. Einspruch, "Ultrasonic Effects in Semiconductors", in *Solid State Physics* (F. Seitz and D. Turnbull, eds.), Vol. 17, p. 217. Academic Press, New York, 1965.

J. H. McFee, "Transmission and Amplification of Acoustic Waves in Piezoelectric Semiconductors," in *Physical Acoustics* (W. Mason, ed.), Vol. IV, part A, p. l. Academic Press, New York, 1966.

H. N. Spector "Interaction of Acoustic Waves and Conduction Electrons", in *Solid State Physics* (F. Seitz and D. Turnbull, eds.) Vol. 19, p. 291. Academic Press, New York, 1966.

R. Bray, "A Perspective on Acoustoelectric Instabilities," *IBM J. Res. Devel.* **13**, 487 (1969). See also other articles in this volume., pp. 494–510.

N. I. Meyer and M. H. Jorgensen, "Acoustoelectric Effects in Piezoelectric Semiconductors with Main Emphasis on CdS and ZnO", in *Festkörper Probleme X* (0. Madelung, ed.), p. 21. Vieweg, Braunschweig, Germany, 1970.

E. M. Conwell and A. K. Ganguly, "Mixing of Acoustic Waves in Piezoelectric Semiconductors", *Phys. Rev.* **B4**, 2535 (1971).

Adsorption

J. G. Dash

All surfaces are typically coated with films of foreign molecules that are either specifically applied or unintentionally drawn from their environment, and these films can affect most of the properties of the interface. The very active field of surface science [1] is driven by interest in the fundamental properties of surfaces and films, and by their technical importance, for they are prime factors in many industrial areas, including adhesion, catalysis, corrosion, fracture, lubrication, and solid state electronics. The films are broadly classified by the nature of the forces binding them to the substrate; *chemisorption* when the bonding is primarily chemical, and *physisorption*, or simply *adsorption*, when there is little electron transfer. Most substrates are structurally and chemically heterogeneous, so that their films are highly disordered and difficult to analyze on a fundamental level. Steele [2] reviews many studies of heterogeneous adsorption, and Rudzinski and Everett [3] describe the characterization of heterogeneous adsorbents by vapor pressure isotherms. More uniform substrates, which are essential in modern electronic and optical devices, began to be produced and studied in the latter part of the 20th Century. Accounts of some of the crucial steps in the development of surface science are described in the collections edited by Duke [1], and Duke and Plummer [4]. In what follows we limit the discussion to films adsorbed on uniform solid surfaces, which have both contributed to and benefited from, modern surface science.

The forces of attraction that cause adsorption are the relatively weak and long-range interactions that exist between neutral atoms, molecules, and macroscopic objects. These *dispersion* forces are due to the attractions between the electric dipole moments induced in each body by the fluctuating fields of their neighbors [5]. Dispersion forces between neutral atoms and molecules are responsible for the condensation of vapors to liquid or solid phases at sufficiently low temperature.

The states of surface films depend on their thickness, temperature, and composition, as well as the structure, uniformity, and constitution of the substrate. As a result of this interplay films display a great variety of distinctive regimes. Monolayer films exhibit two-dimensional analogs of the familiar vapor, liquid, and solid phases of bulk matter, as well as others that have no three-dimensional equivalents [5–7]. In thicker films some layers may behave with distinctly different thermodynamic character while the remainder is diffuse. Experimental techniques for film studies include calorimetry, nuclear resonance, and electron, xray, neutron, gamma ray, molecular and optical spectroscopies. Special methods developed rapidly since the 1960s and are developing still.

Considerable attention has been focused on the properties of films in the context of general questions involving the physics of two-dimensional (2D) systems. Theories had indicated that there can be no perfectly ordered states or structures in 1D and 2D matter; that lower-dimensional crystals, magnets, superconductors, and superfluids cannot exist above absolute zero. However, more recent theories have shown that certain types of 2D long-range order may persist at finite temperatures, and several experiments bear out these newer ideas [8].

A crystal has long range *positional* order if it has periodicity of unlimited extent. In principle, a 3D crystal has long-range positional order, barring dislocations that span the entire structure. Peierls [9] showed that thermal excitations destroy the long range positional order of a 1D chain. At $T = 0$ the position of the nth atom, in a chain with interparticle spacing d, is predictably at nd even as n diverges, but at finite T the uncertainty in position becomes greater than d for $n \to \infty$. Mermin [10] extended Peierls' model to 2D, but went on to show that long-range *directional* order in a 2D crystal is much more robust. Directional order (a more general term is *topological order*) is preserved if a large closed loop can be traced stepwise from atom to atom, and as long as the bonds remain unbroken such a closed path will be intact at all temperatures. The distinction between positional and topological order is that of elasticity and rupture; an elastic net when undisturbed has both forms of order; stretching it may destroy positional order, but the net retains topologically order as long as it is not torn apart. Kosterlitz and Thoules [11] described how topological order in a 2D crystal could be destroyed by thermally excited dislocations, in a continuous melting transition.

Subsequent elaboration of the theory by several investigators [8] predicted that the development of complete liquid-like disorder may take place via two successive continuous transitions. However, up to the present time, there have been no unambiguous confirmations of continuous 2D melting in any experimental monolayer films; in contrast, several experimental films undergo first order melting. It is possible that the predicted continuous transitions would occur if they were not preempted by first-order melting. The theory neglects the effects of grain boundaries and the edges of films of finite extent, where the weakness of the solid allows premelting at relatively low temperature, and enables the first order phase change to proceed at a lower temperature than the theoretical transitions [12].

The new ideas have been much more successfully applied to superfluidity [8]. Helium films provide the simplest example of a topological phase transition. The order parameter, which is the condensate wave function, is a complex function of the 2D position, so that the system is equivalent to a 2D planar spin model. In this system the singularities that destroy long-range order are vortices. Kosterlitz and Thouless [11] predicted a continuous transition to superfluidity, and that the transition temperature would vary with the film coverage as a power law. The theory was confirmed in a very sensitive experiment by Bishop and Reppy [13]. Their method was based on the changes of period and dissipation of the torsional oscillations of a spiral of Mylar plastic, covered by a thin film of adsorbed helium. The experiments showed that the superfluid transition temperature followed a power law in film coverage, with the theoretical exponent of 1/2, over more than a decade in the temperature.

Several international conferences have explored the great variety of phases and phase transitions displayed by monolayer and multilayer films [14]. When a uniform surface is sparsely covered, the adsorbed atoms can act like a 2D gas [6]. The essential characteristics are adhesion to the substrate, which at sufficiently low temperature leads to a "freezing out" of higher states of motion normal to the surface, and low surface density. The atoms' surface mobility

depends on the substrate's lattice size and the amplitude of the corrugation of binding energy. On strongly corrugated surfaces the adatoms tend to be immobilized at surface sites for long dwell times. As T rises the dwell time decreases, due to more rapid hopping between sites. Surface mobility may be appreciable even at low T for low-mass adatoms on relatively smooth surfaces, due to quantum-mechanical tunneling between sites. Low-density helium and hydrogen films adsorbed on graphite exhibit 2D mobile gas-like heat capacities at low T, evidently as a result of rapid quantum tunneling. Interactions between the adsorbed atoms become important at low temperature and high surface density. Corrections to the equation of state can be in the form of a series expansion in the surface density, similar to the virial expansion for a 3D gas. The second virial coefficients of several films have been deduced from calorimetric measurements or vapor-pressure isotherms, and they agree well with theoretical coefficients calculated from atomic pair potentials [15].

Phase condensation of low coverage films to 2D liquid or solid phases occurs when they are cooled to low temperature. Critical temperatures of the 2D gas-liquid transition are somewhat less than half of the critical temperatures of the bulk phases for most of the noble gases and other simple molecular gases. The gas–liquid critical point is especially interesting, since it belongs to the universality class of the 2D Ising model. Measurements on several monolayer films show critical behavior in good agreement with theory [16].

Typical monolayer systems have dense spatially ordered phases at high density and low temperature, where the structure of the film is incommensurate with that of the substrate [6, 7, 14, 17]. Such "floating solid" monolayers are effectively 2D solids; exemplary systems are helium and hydrogen isotopes, neon and xenon on basal-plane graphite. Their heat capacities exhibit temperature dependence varying as T^2, the 2D analog of the well-known Debye T^3 law, and they melt at sharp triple points. In some films the substrate structure imposes a regularity on the atomic arrangement in the monolayer. In the simplest cases the adsorbed atoms have the same regularity as the substrate atoms, but they may have more complex structures due to the competition between substrate-atom and atom-atom interactions. Particular interest focuses on the order-disorder transition of the registered phases of helium and hydrogen isotopes on graphite, which exhibit strong heat capacity peaks with power-law temperature dependence. The transitions belong to the universality class of the three-state Potts model, and the experimental exponent is in excellent agreement with the theory [18].

Films of several layers thickness display a variety of habits. In some examples the one or two layers closest to the substrate behave as relatively distinct 2D solids while a topmost third layer is effectively a two-dimensional gas. In many systems there is a *wetting transition* [19] between layer formation and cluster growth, generally occurring at a thickness of several atomic layers. The instability of layer formation *vis a vis* cluster formation is an important phenomenon in all types of films. It can occur in systems that have strongly cohesive interactions as well as those having relatively weak dispersion forces. In strongly adsorbed solid films the substrate attraction tends to produce strained layers next to the substrate, which prevents the formation of very thick uniform deposits. This effect belongs to a complex of wetting phenomena, which are of fundamental interest and practical importance. Adsorbed multilayer films are also valuable as test systems for the study of phenomena that can occur on typical surfaces of bulk solid materials. Recent examples are surface roughening and surface melting, which have been observed in multilayer noble gas and light molecular films on carbon nanotubes [20].

See also: Catalysis; Ising Model; Order–Disorder Phenomena; Phase Transitions; Surfaces and Interfaces; Thin Films.

References

[1] C. B. Duke (ed.), *Surface Science, The First Thirty Years*. North-Holland, 1994.

[2] W. A. Steele, *The Interaction of Gases with Solid Surfaces*. Pergamon, Oxford, 1974.

[3] W. Rudzinski and D. H. Everett, *Adsorption of Gases on Heterogeneous Surfaces*. Academic Press, New York, 1992.

[4] C. B. Duke and E. W. Plummer (eds.), *Frontiers in Surface and Interface Science*. North-Holland, 2002.

[5] L. Bruch, M. W. Cole and E. Zaremba, *Adsorption: Thermodynamics and Structure*. Wiley, New Yrok, 1998.

[6] J. G. Dash, *Films on Solid Surfaces*. Academic Press, New York, 1975.

[7] I. Lyuksyutov, A. G. Naumovets and V. Pokrovsky, *Two-Dimensional Crystals*. Academic Press, New York, 1992.

[8] D. J. Thouless, *Topological Phase Transitions*. World Scientific, Singapore, 2000.

[9] R. E. Peierls, *Ann. Inst. H. Poincaré* **5**, 177 (1935).

[10] N. D. Mermin, *Phys. Rev.* **176**, 250 (1968).

[11] M. Kosterlitz and D. J. Thouless, *J. Phys. C* **5**, L124 (1971) and **6**, 1181 (1973).

[12] J. G. Dash, "Melting From One to Two to Three Dimensions", *Contemp. Phys.* **43**, 427 (2002).

[13] D. J. Bishop and J. D. Reppy, *Phys. Rev. Lett.* **40**, 1727 (1978).

[14] Coll. Int. du CNRS, "Phases Bidimensionelles Adsorbées", *J. Phys. (Paris)* **38**, C-4 (1977); S. K. Sinha (ed.), *Ordering in Two Dimensions*, North-Holland 1980; J. G. Dash and J. Ruvalds (eds.), *Phase Transitions in Surface Films*, Plenum, New York, 1980; H. Taub, G. Torzo, H. J. Lauter and S. C. Fain, Jr., *Phase Transitions in Surface Films 2*, Plenum, New York, 1991; M. Michailov and I. Gutzow, *Thin Films and Phase Transitions on Surfaces*, Inst. Phys. Chem. Bulgarian Acad. Sci., 1994.

[15] R. L. Siddon and M. Schick, *Phys. Rev. A* **9**, 907 (1974).

[16] R. B. Griffiths, in C. Domb and M. S. Green (eds.) *Phase Transitions and Critical Phenomena*, Vol. 1. Academic Press, New York 1972.

[17] E. Domany, M. Schick, J. S. Walker and R. B. Griffiths, *Phys. Rev. B* **18**, 2209 (1978).

[18] J. G. Dash, M. Schick and O. E. Vilches, *Surf. Sci.* **299/300**, 405 (1994).

[19] S. Dietrich, in C. Domb and J. Lebowitz (eds.) *Phase Transitions and Critical Phenomena*, Vol. 12, Academic Press, New York, 1987.

[20] M. M. Calbi, M. W. Cole, S. M. Gatica, M. J. Bojan and G. Stan, *Rev. Mod. Phys.* **73**, 857 (2000); T. Wilson, A. Tyburski, M. R. DePies, O. E. Vilches, D. Becquet and M. Bienfait, *J. Low Temp. Phys.* **126**, 403 (2002).

Aerosols

F. S. Harris, Jr.[†]

An aerosol is a suspension of liquid, solid, or mixed particles in a gas, usually air. The size of the particles ranges from about 10^{-9} m, just larger than molecules, to a radius of about 25 μm, as in cloud droplets and dusts with short-time stability due to gravitational settling. Examples are hazes, mists, fogs, clouds, smokes, and dusts, as well as living bacteria, viruses, and molds. Aerosols are important in atmospheric electricity, cloud formation, precipitation processes, atmospheric chemistry, air pollution, visibility, radiation transfer, and hence climate.

The smallest particles, called Aitken nuclei, are from molecular sizes up to about 0.05 μm in radius. They vary in concentration from a few particles per cubic centimeter over the South Pole Plateau to 300 in clean continental air, up to hundreds of thousands in a polluted city or downwind from a combustion source. The condensation nuclei serve as centers upon which cloud and fog droplets form. The large droplets, as in fags and clouds, ordinarily range from 1 to 25 μm with a concentration of 20–500/cm^3 and a liquid water content up to 1 g/m^3.

The tropospheric aerosol sources are (1) inorganic gas-to-particle conversion, primarily SO_2, NH_3, NO_x both natural and man-made; (2) mineral dust, primarily from arid zones and deserts; (3) sea salt; and (4) organic matter of apparently complex but still unidentified origins, but thought to have heavy contributions from plant-derived terpene compounds, forest fires, and oxygenated hydrocarbons. The residence time in the lower troposphere is about 3–6 days for the particles. Above the ocean the maritime aerosol is found only at the lower altitudes; at higher altitudes there is continental aerosol such as the Sahara Desert dust over the North Atlantic Ocean. The wind (aeolian) transport of aerosols is sometimes as far as 10000 km. By using elemental tracers regional pollution aerosols of both North America and Europe have been followed for several thousand kilometers downwind. Dusts from the central and eastern Asian deserts have been carried to Hawaii and the Marshall Islands in the Pacific Ocean. The continental atmospheric aerosol has approximately 60% of the total mass water-soluble material and 25–30% organic matter; 25% of the material is volatile at temperatures below 150 ℃. The hazes over remote areas may be due to photochemical transformation of terpenes from vegetation. Recent work has shown the importance of sulfate regionally distributed sources. Clean air background in remote areas of the earth is about 10 μm/m^3. The standard mass loading established by the U.S. Environmental Protection Agency, not to be exceeded appreciably, is 75 μm/m^3. The mass loading sometimes reaches to about 2000 μg/m^3. The number concentration varies from 10^2 to 10^7 /cm^3. In the stratosphere the total number above about 0.01 μm is about 10 particles/m^3, primarily sulfates. There is a maximum at an altitude of about 20 km, the amount depending on the length of time since a major volcanic eruption.

Often a simple function has been found to represent the natural aerosol from 0.1 to 20 μm when in equilibrium, the Junge power law, $dn/d\log r = Cr^{-b}$, where dn is the number of particles in a logarithmic size interval, C a constant, r the radius, and b usually a value of about 3. No simple size model can represent the wide variety of sources and complex interactions in the atmosphere. With differing sources, often a log normal or modified gamma distribution can be used for each, or a combination of distributions with one for each source, such as one for small particles from combustion and one for larger dust particles. The particle sizes

[†]deceased

important for health through retention in the human body are those retained in the breathing system after passing through the nose, in the range 4.5 μm down to 0.25 μm radius. The particle size distribution in the atmosphere is affected by the type of source, the changes due to gas-to-particle conversion, condensation, and removal through aggregation, precipitation formation and washout, and gravitational settling.

Experiments have shown that at 75–95% relative humidity (RH) 0.3–0.9 of the submicron aerosol mass can be liquid water, and that even at 50% RH 0.1–0.2 may be liquid water. Actual dry maritime aerosol particles collected over the Atlantic Ocean may increase in volume from 5 to 15 times when the RH is increased to 96%. In air pollution from combustion sources the particles are originally small or are formed by gas-to-particle conversion, and in the Los Angeles, California, basin, 50–80% of the submicron aerosol mass is volatile at 220 ℃, with primarily sulfate, nitrate, noncarbonate carbon, and liquid water.

The optical behavior of the particles is determined by the complex refractive index (which includes the wavelength-dependent real and absorption parts), the shape, the size relative to the wavelength of the radiation, and the size distribution. For particles small compared to the wavelength, the scattering intensity is proportional to the inverse fourth power of the wavelength. For spherical particles (many particles are not), in the range of the radiation wavelength size, the Lorenz–Mie theory must be used in which complicated functions describe the polarization parameters as a function of scattering angle, refractive index, and size distribution. The maximum scattering per unit volume for visible light is for particles 0.5 μm in radius. The amount of solar energy absorbed is comparable in amount with the absorption by atmospheric gases. The absorption part of the refractive index is the critical parameter in determining whether such particles on a world-wide basis will tend to cause cooling or warming of the earth and hence climatic change. For a variety of purposes aerosols are often produced by using a gas under pressure to disperse liquids or solids into the atmosphere. One of the propellants commonly used is a group of chlorofluoromethanes which are chemically quite stable and nontoxic. Currently, however, there is serious investigation of the possible accumulation in the stratosphere and by complex processes reducing the ozone, letting more solar ultraviolet reach the earth's surface.

See also: Atmospheric Physics.

Bibliography

G. Bouesbet and G. Brehan, eds. *Optical Particle Sizing*. Plenum, New York, 1988. (A)

Ardash Deepak, ed. *Atmospheric Aerosols, Their Formation, Optical Properties and Effects*. Deepak, Hampton, VA, 1982. (A)

S. K. Friedlander, *Smokes, Dust, and Hazes, Fundamentals of Aerosol Behavior*. Oxford University Press, New York, 2000. (I)

Peter V. Hobbs and M. P. McCormick, eds. *Aerosols and Climate*. Wiley, 1988. (A)

Kenneth Pye, *Aeolian Dust and Dust Deposits*. Academic Press, New York, 1987. (I)

Parker C. Reist, *Introduction to Aerosol Science*. Macmillan, New York, 2000. (E)

Allotropy and Polymorphism

F. J. DiSalvo

The equilibrium crystal structure of some solids changes when the external conditions, such as pressure or temperature, are varied. In addition, the structure of some compounds depends upon the preparation conditions. One of these structures may be the thermodynamically stable structure, while the remainder are metastable phases. This phenomenon is called allotropy when it occurs in an element, and polymorphism when it occurs in a compound. Three common examples of allotropy are presented below.

Sulfur

Solid sulfur consists of nearly flat S_8 molecular rings that are stacked on top of one another. When heated to 95 ℃ the molecules change orientation, forming a differently stacked structure. Sulfur at room temperature is called rhombic sulfur, and above 95 ℃, monoclinic sulfur (after the shapes of their respective crystallographic unit cells). Monoclinic sulfur melts at 120 ℃.

Sulfur can also exist in an amorphous form. When liquid sulfur is heated to several hundred degrees centigrade, most of the S_8 molecules break open and join with others to form long sulfur chains. If this liquid is rapidly cooled to room temperature, the chains remain intact and are randomly packed together to form a rubbery solid. At room temperature, amorphous sulfur will very slowly change back into rhombic sulfur. Rhombic sulfur is the stable, or equilibrium, form of sulfur at room temperature. By other preparation methods a number of other metastable forms of sulfur can be obtained at room temperature. Consequently, sulfur has a large number of allotropes; however, rhombic and monoclinic sulfur are the only equilibrium forms (in their respective temperature ranges of stability and at atmospheric pressure).

Iron

At room temperature, iron has a body-centered cubic (bcc) structure; the unit cell is shown in Fig. 1a. (The structure can be visualized by imagining space to be filled with closely packed cubes. At each corner, where eight cubes come together, place an iron atom and then put another in the center of each cube.) When iron is heated to 910 ℃ its structure changes to face-centered cubic (fcc); a unit cell is shown in Fig. 1b. (In this structure an iron atom is placed at each cube corner and one iron atom on each face of the cube, where two cubes touch.) Iron changes back to the bcc structure at 1390 ℃ and melts at 1536 ℃.

The allotropy of iron is very important for the production of steels. Carbon is moderately soluble in fcc iron, the carbon atoms occupying some of the holes between the iron atoms in this structure (at the center in Fig. 1b). However, the solubility of carbon is much lower in bcc iron. If iron containing several weight percent of carbon is cooled from 1100 ℃ (fcc phase) to room temperature, the carbon not soluble in bcc iron forms a compound, Fe_3C. The Fe_3C exists in small plate-like regions dispersed in bcc iron. Fe_3C is called cementite, since it makes the iron much stronger. Iron prepared in this manner is called carbon steel.

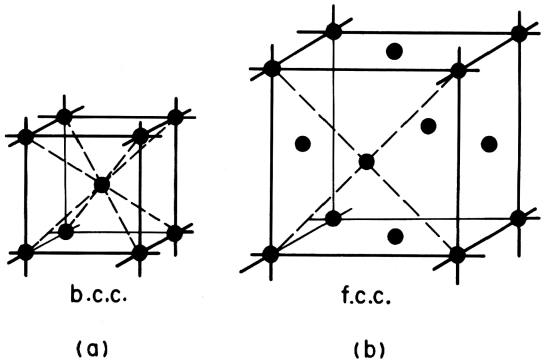

Fig. 1: (a) A unit cell of body-centered cubic (bcc) iron. Iron atoms are spheres that touch along the dotted lines (body center to cell edge). (b) A unit cell of face-centered cubic (fcc) iron. The cubic unit cell is larger than the bcc cell because the iron atoms now touch from the face center to the cell edge.

Carbon

Carbon exists in two structural forms: graphite and diamond. Graphite has a layered structure with weak interlayer bonding. Because of the weak interlayer bonds the layers slide easily over each other. Consequently graphite is used as a lubricant and in pencil lead. Graphite becomes diamond under high pressure (greater than 10 000 bar). With Ni as a catalyst, small diamonds can be manufactured at 1200 °C and high pressure. These diamonds are used in making grinding wheels and cutting tools, since diamond is a very hard material. Recently, diamond films up to millimeters thick have been prepared in the laboratory by a low-pressure plasma deposition technique using methane as a source gas.

Conclusion

Obviously the properties of some materials are quite affected by a change in their structure, and a knowledge of the allotropic or polymorphic forms of materials is important to the development of many technologies.

See also: Crystal Symmetry; Metallurgy.

Bibliography

B. Meyer, *Elemental Sulfur*. Wiley (Interscience), New York, 1965. (I)
A. L. Ruoff, *Introduction to Materials Science*. Prentice-Hall, Englewood Cliffs, NJ, 1972. (I)
W. J. Moore, *Physical Chemistry*, 3rd ed. Prentice-Hall, Englewood Cliffs, NJ 1963. (A)

Alloys

P. L. Leath

An alloy is a macroscopically homogeneous mixture (solution or compound) of metals or, as in the case of carbon steel, a metallic mixture of metals and nonmetals. If they are macroscopically inhomogeneous they would often be called *composites*. Most, but not all, alloys are metallic (for one exception, indium antimonide is a semiconductor). Most pairs of metals are miscible (i. e., form *binary alloys*) at some concentrations, although there are many notable exceptions (e. g., indium is insoluble in gallium). Since there are 70 elemental metals, the subject of alloys is immense, and an enormous variety of electronic and other physical properties is possible. The subject has now expanded even further with recent interest in ternary (three-component) alloys, tertiary or quaternary (four-component) alloys, etc. Several examples of alloys include cast iron, steel, stainless steel, brass, bronze, pewter, solder, intermetallics, aluminum alloys, stellite, chromel, mu-metals, nichrome, constantan, invar, alnico, sterling silver, electrum, and type metal.

Alloys are often classified into ordered (or stoichiometric) and disordered alloys. The ordered alloys have the symmetry of a Bravais lattice with a multiatomic unit cell. Their structure is specified by giving the location of each atom in the unit cell. Some alloys exist essentially only as ordered alloys over the corresponding very narrow ranges of composition necessary for stoichiometry; these alloys are called *intermetallic compounds*. Other alloys (such as β-brass) have ordered phases at stoichiometric concentrations when the temperature is below a phase transition temperature but are disordered otherwise. (That is, they undergo an order–disorder phase transition .) More recently *quasiperiodic* alloys or *quasicrystals* (most notably Al_4Mn) have been discovered which display sharp diffraction peaks that form three-dimensional icosahedral patterns with 5-fold symmetry axes and which thus are not Bravais lattices and do not have the translational symmetry of crystals but do have point symmetries.

Disordered alloys, also called *solid solutions*, occur usually over appreciable ranges of composition. Most common are substitutional alloys, where the various types of atoms randomly occupy the normal sites of a lattice. But there are also *interstitial* alloys (such as carbon in γ-iron), where the solute atoms are small enough to occupy randomly the interstices between the normal lattice sites of the host metal; and there are *amorphous* alloys, where the atoms are not on sites of a regular lattice but are randomly placed, as in a liquid or glass. The occupations of the sites in a disordered alloy by the atomic types may be purely random but generally there is some degree of short-range order; that is, the species occupation of a particular site may be dependent on the occupation of the neighboring sites (e. g., in the disordered phases of brass, the copper atoms are more likely to have zinc than copper nearest neighbors).

Generally, alloys do not have a single melting point, but a solidus temperature at which melting begins, and a liquidus at which melting is complete. In specially designed *eutectic* mixtures, these two temperatures merge into a single melting point.

Certain principal variables that qualitatively give the alloy structures and phases were pointed out in the classic work of Hume-Rothery and Jones. It is, however, only rarely possible to use these few variables to predict detailed behavior of alloy phases. Clearly the relative sizes of the atoms constitute a vital factor in alloys because the volume-dependent potentials in the cohesive energy are an order of magnitude larger than the interatomic rearrangement potentials. This size factor is especially important in interstitial alloys and certain intermetal-

lic compounds (e. g., interstitial alloys are generally not formed when the ratio of the radii of the solute atoms to those of the host atoms exceeds about 0.6). Electrochemical differences are such that generally we find only intermetallic compounds or very restricted ranges of solubility for elements widely separated in the electrochemical series. Particularly interesting are the interstitial alloys of hydrogen in metals.

When size and electrochemical factors allow solid solutions, the alloy structure can in some cases be directly related to the electron density or electron-to-atom ratio. According to the Hume-Rothery rules, in nearly free-electron alloys the stable crystal structure at a particular electron-to-atom ratio will be that which minimizes the energies of the electrons in the crystal potential; thus the position of the Fermi surface relative to the Brillouin zone faces is an essential ingredient. These rules seem to work qualitatively well for the d-band transition metal alloys (especially copper, silver, and gold alloys), but they fail for the more nearly free-electron alkali metal alloys because of electrochemical differences. Clearly the d bands play an important role.

Only recently have basic calculational methods been developed to predict the physical behavior of alloys accurately from first principles. The pseudopotential, orthogonalized plane wave (OPW), augmented plane wave (APW), and Korringa–Kohn–Rostoker (KKR) Green's function methods of calculating electronic energy-band structure of metals beginning with the atomic potentials (the atomic potentials in alloys look very much the same as those atomic potentials do in the respective pure metals) are now in many cases being directly applied successfully in the calculation of such physical properties of alloys as energy-band structure, crystal structure, lattice vibration spectra, electrical and thermal resistivity, and magnetic and superconducting properties. In those cases where the potentials of the constituent atoms do not vary greatly the average potential may be used; this is called the *virtual crystal* (or rigid band, or common band) *approximation*. In the cases of strong disorder, when the potentials differ greatly, the *average t-matrix approximation* (ATA) and the *coherent potential approximation* (CPA), which are capable of producing the separate energy bands for each atomic species, are used. Although these calculations have been somewhat successful, such effects as charge transfer between atoms and atomic cluster effects are often important but are not included in the simple approximations just mentioned. A fine review of the experimental electronic properties of alloys is given by Sellmyer (1978).

Disordered alloys are dramatically different from ordered alloys and pure metals in their electrical resistance at low temperatures. In ordered metals there is a striking decrease in resistance with decreasing temperature that is absent in disordered alloys. For example, in very pure disordered brass at liquid-helium temperatures the electrical resistance is about half its room-temperature value, in contrast to drops by factors of about 10^{-4} in comparable ordered metals. This phenomenon is caused by electronic scattering off of the disorder or of those regions where the periodicity is destroyed.

Finally, the physical properties of alloys are often greatly affected by heat and mechanical treatment, which may introduce or eliminate such defects as vacancies, dislocations, or grain boundaries. For example, wrought alloys, which have been hot or cold worked and hence are generally very anisotropic and fibrous in contrast to cast alloys, which are generally crystalline, are generally more ductile. And there are *shape memory alloys* (notably Ti–Ni) which under certain treatment will return to an original shape. The effect of such defects is only understood qualitatively, although progress is rapidly being made.

See also: Electron Energy States in Solids and Liquids; Metals.

Bibliography

G. Alefeld and J. Volkl, eds. *Hydrogen in Metals*, Vols. I & II. Springer-Verlag, Berlin, 1978.

R. Banks, *Shape Memory Effects in Alloys*. Plenum, New York, 1975.

C. S. Barrett and T. B. Massalski, *Structure of Metals*, 3rd ed. McGraw-Hill, New York, 1966. (I)

R. J. Elliott, J. A. Krumhansl, and P. L. Leath, "The Theory and Properties of Randomly Disordered Crystals and Related Physical Systems," *Rev. Mod. Phys.* **46**, 465–543 (1974). (A)

M. Hansen and K. Anderko, *Constitution of Binary Alloys*, 2nd ed. (1958); R. P. Elliott, 1st suppl. (1965); F. A. Shunk, 2nd suppl. (1969). McGraw-Hill, New York. (A compendium of data on specific alloys.)

V. Heine and D. Weaire, "Pseudopotential Theory of Cohesion and Structure," in *Solid State Physics* (F. Seitz, D. Turnbull, and H. Ehrenreich, eds.), Vol. 24, pp. 249–463. Academic, New York, 1970. (A)

J. Janssen, M. Fallon, and L. Delacy, *Strength of Metals and Alloys* (P. Haasen, ed.). Pergamon, London, 1979.

F.E. Luborsky, ed, *Amorphous Metallic Alloys*, Butterworths, London and Boston, 1983.

N. F. Mott and H. Jones, *The Theory and Properties of Metals and Allloys*. Oxford, London and New York, 1936. (E)

P. S. Rudman, J. Stringer, and R. I. Jaffee, eds., Phase Stability in Metals and Alloys. McGraw-Hill, New York, 1967. (I)

D. J. Sellmyer, "Electronic Structure of Metallic Compounds and Alloys," in *Solid State Physics* (H. Ehrenreich, F. Seitz, and D. Turnbull, eds.), Vol. 33, pp. 83–248. Academic, New York, 1978.

W.F. Smith, "Structure and Properties of Engineering Alloys", 2nd ed., McGraw-Hill, New York, 1993.

P. J. Steinhardt and S. Ostlund, *The Physics of Quasicrystals*. World Scientific, Singapore, 1987.

K. Tien and G.S. Ansell, eds., "Alloy and Microstructural Design", Academic, New York, 1976.

Alpha Decay

I. Ahmad

Soon after the discovery of radioactivity by Becquerel in 1896, it was established that three types of radiations are emitted by radioactive substances. The most easily absorbed radiations were named alpha (α) rays. In 1909, Rutherford and Royds obtained a direct experimental proof that α particles are doubly ionized helium atoms. Since α particles are electrically charged they are deflected in electric and magnetic fields and produce intense ionization in matter. The thickness of material required to stop an α particle is called its range and it depends on the kinetic energy of the α particle and on the nature of the stopping medium. The range of an α particle with the typical energy of 6.0 MeV is ~ 5 cm in normal air and ~ 0.05 mm in aluminum.

At present more than 500 α-emitting nuclides are known and most of these are produced artificially. The kinetic energies of α particles range from 1.83 MeV for ^{144}Nd to 11.65 MeV for ^{212m}Po; these energies correspond to α particle velocities of $(1–3)\times 10^9$ cm/s. The measured

half-lives of known α emitters vary from 3.0×10^{-7} s for ^{212}Po decay to 2.1×10^{15} years for ^{144}Nd decay. Normally α decay occurs from the ground state of the parent nucleus and several groups of α particles (each group contains monoenergetic α particles) are emitted leaving the daughter nucleus in its ground state or in an excited state. In a few cases α particles are also emitted from an excited state of the parent nucleus. These α particles have kinetic energies of 9–12 MeV and are called long-range α particles. Examples of such α emitters are ^{212}Po and ^{214}Po.

Alpha decay has recently been used to characterize newly produced transactinide elements. By following the decay chain down to a known nuclide, it has been possible to determine the atomic number of a new element. This procedure has been used to identify elements with $Z = 107$–112.

For a nucleus to be unstable toward α decay, its mass must be greater than the sum of the masses of the daughter nucleus and the α particle. If we write the α decay of a nucleus with mass number A and atomic number Z as

$$^{A}Z \rightarrow {}^{(A-4)}(Z-2) + {}^{4}_{2}\mathrm{He} \,, \tag{1}$$

then the α decay energy, also called Q value, is given by

$$Q = (M_A - M_{A-4} - M_{\mathrm{He}})c^2 \,. \tag{2}$$

In the above equation, M represents the atomic mass and c is the velocity of light. Q values have been calculated from known atomic masses. Calculations show that Q values are positive for all β-stable nuclei with $A >\sim 150$. Although such nuclei are thus unstable with regard to α emission, in many cases the half-life is too long for the α decay to have been detected. Experimentally α radioactivity has been detected in most translead and some rare-earth nuclei.

In order to conserve linear momentum, the decay energy Q is divided between the α particle and the daughter nucleus in inverse proportion to their masses. The energy imparted to the daughter nucleus is called recoil energy. The α particle energy E_α and the recoil energy E_{R} are given by the equations

$$E_\alpha = (M_{A-4}/M_A)Q \quad \text{and} \quad E_{\mathrm{R}} = (M_\alpha/M_A)Q \,. \tag{3}$$

The laws of conservation of angular momentum and of parity (even or odd character of the state wave function) plus the fact that the α particle has no intrinsic spin and even parity lead to simple selection rules for α decay. The orbital angular momentum L of the emitted α particle is restricted to integral values between the sum and the difference of the total spins of the initial and final nuclear states. If the parent and the daughter nuclear states have the same parity, only even values of L are permitted; if their parities are opposite, only odd L values are allowed.

The energies of α particles and the intensities of α groups are measured with gas ionization counters, solid-state detectors, or magnetic spectrographs. At present, Passivated Implanted Planar Silicon (PIPS) detectors are widely used in spectroscopic measurements. These silicon detectors, under the best conditions, have resolutions [full width at half-maximum (FWHM) of the α peak] of 9.0 keV and efficiencies of ∼30%. Magnetic spectrographs, on the other hand, have low transmission (∼0.1%) but can achieve resolution (FWHM) of less than 3.0 keV for

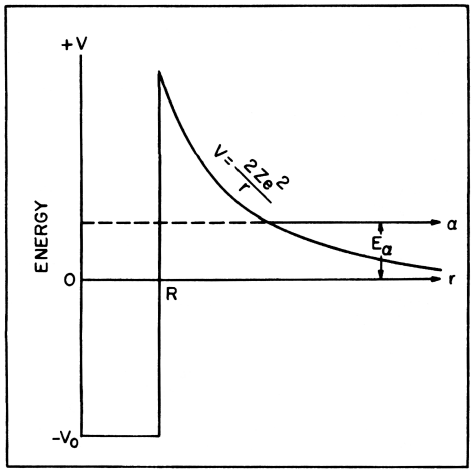

Fig. 1: Schematic representation of the potential energy of an α particle in the vicinity of a heavy nucleus. The potential energy is plotted against the distance r between the centers of the α particle and the residual nucleus.

6.0-MeV α particles. Very thin, essentially massless, sources are used in these measurements. The absolute energies of the α particles emitted by a few nuclides have been measured with high precision by Rytz using a magnetic spectrograph; energies of other α groups are measured relative to these standards. Because of the monoenergetic character of α groups plus the fact that their energies and intensities can be measured with high precision, α particle spectroscopy has been extensively used in nuclear-structure studies of heavy elements.

The systematic relationship between the α decay half-life and the decay energy was first discovered by Geiger and Nuttall in 1911. According to a modified version of their rule, when the logarithm of the α decay half-life is plotted against the inverse square root of Q, a straight line is obtained for each element; i. e.,

$$\log t_{1/2} = A/Q^{1/2} + B \,, \tag{4}$$

where A and B are constants and depend on the atomic number Z. This relationship applies only to α transitions between ground states of even–even nuclei. Because of the strong dependence of the α decay rate on the Q value, only states up to a few hundred kilovolts excitation are measurably populated.

The mechanism of α decay and the Geiger–Nuttall rule were first explained by Gamow and independently by Gurney and Condon in 1927. In the potential-energy diagram (see Fig. 1) the maximum occurs at R, where R is equal to the sum of the radii of the α particle and the residual nucleus. At distances $r < R$ the potential is attractive because of the short-range nuclear force and for $r > R$ electrostatic repulsion gives a positive potential which decreases with increasing r according to Coulomb's law. The typical energy of an α particle within the nucleus, with respect to the zero of energy at $r = \infty$, is 6.0 MeV and the height of the potential barrier at R (called the Coulomb barrier) for heavy elements is ~ 20 MeV. Since the kinetic energy of the α particle is less than the barrier height, according to classical mechanics the α particle can never leave the nucleus. However, the wave nature of matter permits a

6.0-MeV α particle occasionally, on the nuclear time scale, to "tunnel" through the 20-MeV potential barrier. Using a simplified shape for the potential and assuming that the α particle pre-exists as a clustered entity inside the nucleus and is constantly impinging on the barrier, Gamow derived an expression for the α decay rate which explains the observed exponential dependence of transition rates on Q values.

The measured partial half-lives of α groups in the decay of odd-mass and odd–odd nuclei and α transitions to the excited states of even-even nuclei are found to be longer than the partial half-life $(t_{1/2})_{\mathrm{e-e}}$ of an α group of the same energy emitted in the decay between ground states of even–even nuclei. The relative retardation of the former decay is called its hindrance factor and its reciprocal gives the relative reduced α transition probability. The values of $(t_{1/2})_{\mathrm{e-e}}$ are calculated either by Eq. (4) or by some other theory. In most recent publications the values of $(t_{1/2})_{\mathrm{e-e}}$ are computed with the one-body α decay theory of Preston. In this theory, as in Gamow's, the α particle is assumed to preexist inside the nucleus and is ejected with no orbital angular momentum. Radius parameters of even–even nuclei are obtained by normalizing to the measured transition rates between their ground states and the radius parameters of odd-mass nuclei are determined by interpolation between the values of adjacent even–even nuclei.

Hindrance factors for α transitions of odd-mass nuclei vary from unity to several thousands and these yield significant information on the parent and daughter states involved in the decay. Alpha transitions of odd-mass and odd–odd nuclei with hindrance factors of 1–4 are called favored transitions; in these decays the parent and the daughter states have similar wave functions. Since the α transitions are only mildly inhibited by angular momentum changes L, high hindrance factors give a clear indication that the α particle does not exist as a clustered entity all the time in the corresponding nucleus. Instead, the α particle is formed from four nucleons (two protons and two neutrons) at the nuclear surface at the time of its ejection. The probability for the formation of an α particle from four nucleons can be calculated theoretically. Alpha-decay rates for spherical nuclei in the lead region and spheroidal actinide nuclei have been calculated by Mang and Rasmussen and these reproduce the general trend in the observed α decay rates. Although in most cases the calculated and measured rates agree within a factor of 2, there are several unfavored transitions for which the calculations and measurements differ by a factor of $\sim$10. Despite these deficiencies, these calculations are extremely useful in nuclear structure studies.

See also: Nuclear Properties; Radioactivity.

Bibliography

R. D. Evans, in *McGraw-Hill Encyclopedia of Science and Technology*, Vol. 1, p. 305. McGraw-Hill, New York, 1971. (E)

I. Perlman and J. O. Rasmussen, in *Handbuch der Physik*, Vol. 42, p. 109. Springer-Verlag, Berlin, 1957. (I)

J. O. Rasmussen, in *Alpha-, Beta-, and Gamma-Ray Spectroscopy* (K. Siegbahn, ed.), Vol. 1, p. 701. North-Holland, Amsterdam, 1965. (A)

Ampère's Law

L. T. Klauder, Jr.

The name Ampère's law has been applied to several of the formulas that give magnetic effects of time-independent electric currents. (Formulas given here assume rationalized mks units.) The equation

$$\oint_C \mathbf{H}\cdot d\mathbf{l} = I \tag{1}$$

relating the magnetic intensity along a closed curve C and the current I linking C is commonly referred to as Ampère's law or Ampère's circuital law. This nomenclature has become popular in recent years because it associates a useful elementary formula with the most important of the original investigators. Historically, this law was discovered by Gauss with help from the theorem of Ampére stating that the field produced by a magnetic shell is the same as that due to a current flowing around the boundary of the shell. The modern form distinguishing between B and H was first given by Maxwell.

A few authors apply the term Ampère's law to both the integral relationship (1) and the corresponding differential equation

$$\nabla\times\mathbf{H} = \mathbf{j} \tag{2}$$

where $\mathbf{j}$ is the electric current density. (For the generalization to cases in which fields are time dependent, *see* Maxwell's equations.)

A number of authors apply the term Ampère's law to the formula for the force exerted by a current element $I_2\,d\mathbf{l}_2$ on another current element $I_1\,d\mathbf{l}_1$:

$$\begin{aligned} d\mathbf{F}_{12} &= \frac{\mu_0}{4\pi}\frac{I_1 I_2}{r_{12}^3}\,d\mathbf{l}_1\times(d\mathbf{l}_2\times\mathbf{r}_{12}) \\ &= \frac{\mu_0}{4\pi}\frac{I_1 I_2}{r_{12}^3}[(d\mathbf{l}_1\cdot\mathbf{r}_{12})d\mathbf{l}_2-(d\mathbf{l}_1\cdot d\mathbf{l}_2)\mathbf{r}_{12}] \end{aligned} \tag{3}$$

where $\mu_0 = 4\pi\times10^{-7}$ is the permeability of free space and $\mathbf{r}_{12}$ is the vector from current path element $d\mathbf{l}_1$ to $d\mathbf{l}_2$. This formula is the basis for the SI unit of electric current referred to as the absolute Ampère.

The existence of the interaction between electric currents was discovered by Ampère in 1820, and he subsequently carried out a remarkable program of experiments and analysis that led him to a formula related to Eq. (3). The reason for the difference is itself interesting. Ampère shared the general view that electrostatic and gravitational forces were cases of action at a distance and assumed that the same was true of the force between currents. Thus, in the interest of conservation of momentum, he assumed that the force between two current elements would have to be directed along the line between them. Accordingly, his result lacked the first term in the second line of Eq. (3) but included another term directed along $\mathbf{r}_{12}$ and causing the entire expression to conform to his experimental result that the force exerted by a closed electric circuit on a current element is perpendicular to the current element. When applied to complete circuits, the formula deduced by Ampère gives the same results as Eq. (3). It can be shown that Eq. (3) is consistent with conservation of momentum as long as the momentum of the electromagnetic field is taken into account.

Following a suggestion by Heaviside, some authors have applied the term Ampère's law to the related formula

$$d\mathbf{F} = I\, d\mathbf{l} \times \mathbf{B} \ , \tag{4}$$

giving the force exerted by the magnetic field $\mathbf{B}$ on the current element $I\, d\mathbf{l}$.

Finally, a number of authors apply the term Ampère's law to the formula

$$d\mathbf{H}_1 = \frac{1}{4\pi} \frac{I_2}{r_{12}^3}\, d\mathbf{l}_2 \times \mathbf{r}_{12} \ , \tag{5}$$

giving the contribution of a current element $I_2\, d\mathbf{l}_2$ to the magnetic intensity at location 1. However, this equation is a little more frequently referred to as the Biot–Savart law. By their experiments, Biot and Savart established the r^{-1} dependence of the force on a magnetic pole due to current in a long straight wire, and Biot credited Laplace with having inferred from their result that the field contribution from a current element must have the r^{-2} behavior exhibited in formula (5).

See also: Maxwell's Equations; Electrodynamics, Classical; Electromagnets.

Bibliography

For physical explanations of the formulas in this article see any college physics text.

For a discussion of Ampère's work from the point of view of the history of ideas, see the article on Ampère in the *Dictionary of Scientific Biography*, C. C. Gillespie (ed.). Scribners, New York, 1970.

Historical references for Eqns. (1) and (2) are C. F. Gauss's article "Allgemeine Theorie des Erdmagnetismus" in *Carl Friederich Gauss, Werke*, Vol. 5, pp. l70, 171; (Göttingen, 1867) and J. C. Maxwell's article "On Faraday's Lines of Force" in *Trans. Camb. Phil. Soc.* **10**, 27 (1856) [reprinted in Vol. 1 of *The Scientific Papers of J. C. Maxwell*. Cambridge, 1890].

Ampère's counterpart to Eq. (3) is discussed in E. T. Whittaker's *A History of the Theories of Aether and Electricity*, 2nd ed., Vol. 1, pp. 85–87, (London, 1951; reprinted by Harper, New York, 1960) and in J. C. Maxwell's *A Treatise on Electricity and Magnetism*. 3rd ed., Vol. 2, pp. 163–174 (Oxford, 1892). Translations of most of Ampère's papers are available in R. A. R. Tricker's *Early Electrodynamics: The First Law of Circulation* (Pergamon, New York, l965).

For a demonstration that formula (3) does not violate conservation of momentum when the role of the electromagnetic field is included, see the article by L. Page and N. E. Adams in *Am. J. Phys.* **13**, 141 (1945). For an extended treatment see F. Rohrlich, *Classical Charged Particles* (Addison-Wesley, Reading, Mass., 1965).

Anelasticity

A. S. Nowick

The term anelasticity, although once used loosely to refer to nonelastic behavior, was given a more specific meaning by C. Zener in 1946; this meaning has since been widely adopted. Anelasticity is a generalization of Hooke's law of elasticity, which allows for time-dependent

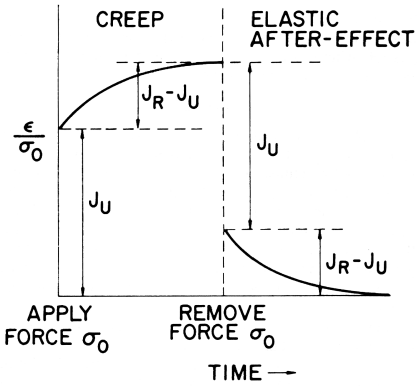

Fig. 1: Behavior of the standard anelastic solid upon application of a static stress σ_0 (creep), and upon the subsequent release of this stress (elastic aftereffect).

effects. Hooke's law may be stated as $\varepsilon = J\sigma$, where ε is strain, σ is stress, and J is the compliance constant. In anelasticity, the instantaneous response and single-valuedness inherent in Hooke's law are discarded. However, two restrictions are retained: (a) *linearity*, in the sense that doubling the stress doubles the strain at each instant of time; and (b) a *unique equilibrium relationship*, which means that to every value of stress there corresponds a unique value of strain that is attained if sufficient time is allowed. The simplest relation between stress and strain and their time derivatives that obeys these conditions is

$$J_R\sigma + \tau J_U\dot{\sigma} = \varepsilon + \tau\dot{\varepsilon}\ , \tag{1}$$

involving three constants σ, J_R, and J_U. Any material that obeys Eq. (1) is called a *standard anelastic solid*. Equation (1) can be solved under conditions of constant stress, say σ_0, (which constitutes a "creep" experiment), to give

$$\frac{\varepsilon(t)}{\sigma_0} = J_U + (J_R - J_U)\left[1 - \exp\left(\frac{-t}{\tau}\right)\right]\ . \tag{2}$$

From this equation the meaning of the constants becomes clear: J_U, called the unrelaxed compliance, corresponds to the instantaneous response, $\varepsilon(0)/\sigma_0$, at $t = 0$; J_R, the relaxed compliance, is ε/σ_0 as $t \to \infty$; τ is the relaxation time (at constant stress). Figure 1 shows this creep behavior, as well as the time-dependent recovery, or "elastic aftereffect," which takes place after the stress is removed. Equation (1) can also be solved under conditions of constant strain to obtain an exponentially decreasing stress, describing a "stress-relaxation experiment." The most important manifestation of anelasticity, however, occurs in the dynamical case, where stress and strain are both periodic, with the strain lagging behind the stress by a phase angle ϕ. Then we can express a complex compliance by $J^* = \varepsilon/\sigma \equiv J_1 - iJ_2$, where J_1 is the real part, which is in phase with the applied stress, and J_2 the imaginary part, which lags σ by $\pi/2$. Using Eq. (1), we can express J_1 and J_2 as functions of the angular frequency ω:

$$J_1(\omega) = J_U + (J_R - J_U)/(1 + \omega^2\tau^2)\ , \tag{3}$$

$$J_2(\omega) = (J_R - J_U)\omega\tau/(1 + \omega^2\tau^2)\ , \tag{4}$$

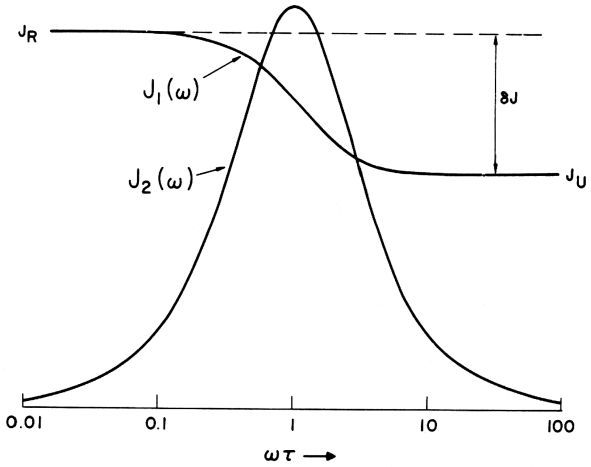

Fig. 2: Dependence of the dynamical functions $J_1(\omega)$ and $J_2(\omega)$ on $\omega\tau$ for the standard anelastic solid.

Equations (3) and (4) are the celebrated Debye equations. The function $J_2(\omega)$ when plotted versus $\log(\omega\tau)$ gives a symmetrical peak centered about $\log(\omega\tau) = 0$ (i. e., $\omega\tau = 1$), which is called a Debye peak. Figure 2 shows the variation of both J_1 and J_2 with the variable $\log(\omega\tau)$. The phase angle ϕ by which ε lags behind a is given by $\tan\phi = J_2/J_1$. This quantity, which also takes the form of a Debye peak, is often called the internal friction, since it is a measure of the energy dissipated per cycle.

Since the Debye peak depends on the variable $\omega\tau$, it can be traced out either by varying the frequency, ω, or by changing τ. Although a continuous variation of the vibration frequency is sometimes possible, the usual experimental methods make it preferable to work at one frequency. It is therefore quite valuable to have a method for tracing out a Debye peak by varying τ while keeping ω constant. This is possible when τ is controlled by a thermally activated process involving an Arrhenius-type relation

$$\tau = \tau_0 \exp(Q/kT) \tag{5}$$

where Q is the activation energy for the process, τ_0 is the preexponential constant, k is Boltzmann's constant, and T is the absolute temperature. In fact, it then turns out that a plot of $\tan\phi$ (or J_2) versus T^{-1} gives a symmetrical peak like that in Fig. 2 except for a change in scale factor. If the peak is then obtained at two or more different frequencies, the activation energy is readily obtained from the shift of the peak location with frequency. Figure 3 shows an example of this type of plot, which is most important in studying anelastic phenomena.

Many phenomena in solids are describable in terms of the equations of the standard anelastic solid. Often, however, this simple model is insufficient to describe the behavior of a material. For example, the internal friction may show a superposition of two or more Debye peaks instead of a single one, or in other cases a single peak may be obtained that is broader than a Debye peak. To treat such cases, it is necessary to introduce a spectrum of relaxation times in place of the single relaxation time of the standard anelastic solid. Thus, instead of a single exponential in the creep function of Eq. (2) there may be a summation of terms with different

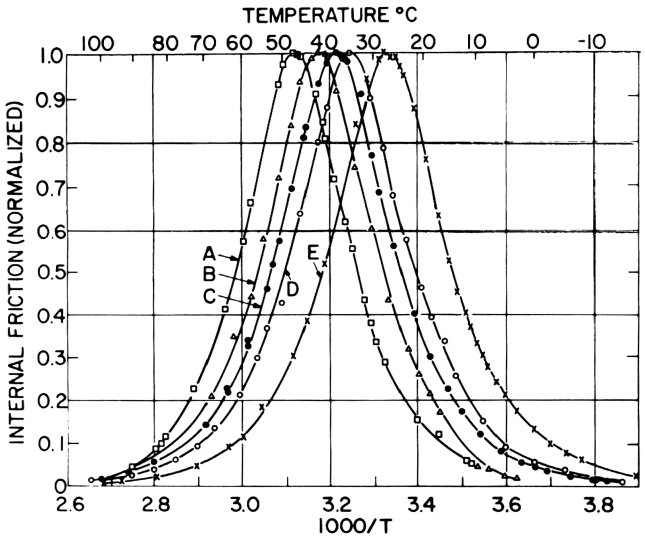

Fig. 3: A series of internal friction peaks for an Fe–C alloy as function of $1/T$ for five different frequencies: A, 2.1; B, 1.17; C, 0.86; D, 0.63; and E, 0.27 Hz. From C. Wert and C. Zener, *Phys. Rev.* **76**, 1169 (1949).

τ values and weighting factors. Correspondingly, Eq. (4) becomes a sum of Debye peaks. Such behavior is called a discrete relaxation spectrum. In more complex cases there may be a continuous variation in τ, a continuous spectrum, described by an appropriate distribution function. In either of these situations, the display of data in the form of a plot of internal friction ($\tan\phi$) versus T^{-1} is widely used and interpreted.

The physical origins of anelasticity are very varied and encompass almost all aspects of solid-state physics. In crystalline materials, anelastic behavior can result from any of the following mechanisms:

1. Point-defect relaxations: redistribution of point defects (whose symmetry is lower than that of the crystal) into sites that become preferential in the presence of a stress field.

2. Dislocation relaxations: motion of dislocation segments, present either from growth or from plastic deformation, in a variety of ways with the aid of jogs, kinks, and impurity atoms on the dislocation lines.

3. Grain-boundary relaxation: viscous sliding of one grain over another in a polycrystalline material.

4. Phonon relaxation: change in the frequency distribution of phonons (lattice vibrations) due to stress.

5. Magnetic relaxations: magnetoelastic coupling via magnetostriction of a ferromagnetic material, giving rise to a number of different relaxation processes.

6. Electronic relaxations: change in the energetics of the electronic configuration produced by stress leading to a redistribution of both free and bound electrons in various materials.

Glasses (including amorphous alloys) are also capable of showing a variety of relaxations, many of which are similar in origin to those found in crystalline materials. In amorphous polymers a major relaxation is associated with the glass transition and attributed to large-scale rearrangements of the main polymer chain. Secondary relaxations, at lower temperatures, are due to side groups that are capable of independent hindered rotations.

See also: Elasticity; Relaxation Phenomena; Rheology.

Bibliography

W. Benoit and G. Gremaud, eds., "Internal Friction and Ultrasonic Attenuation in Solids." *J. Phys. (Paris)* **42**, Colloque No. 5 (1981).

R. De Batist, *Internal Friction of Structural Defects in Crystalline Solids*. North-Holland, Amsterdam, 1972.

R. De Batist and J. Van Humbeeck, eds., "Internal Friction and Ultrasonic Attenuation in Solids." *J. Phys. (Paris)* **48**, Colloque C8 (1987).

J. D. Ferry, *Viscoelastic Properties of Polymers*. Wiley, New York, 1961.

A. V. Granato, G. Mozurkewich, and C. A. Wert (eds.), "Internal Friction and Ultrasonic Attenuation in Solids." *J. Phys. (Paris)* **46**, Colloque C10 (1985).

R. R. Hasiguti and N. Mikoshiba, eds., *Internal Friction and Ultrasonic Attenuation in Solids*. Univ. Tokyo Press, Tokyo, 1977.

D. Lenz and K. Lücke, eds., *Internal Friction and Ultrasonic Attenuation in Crystalline Solids*, Vols. I and II. Springer, Berlin and New York, 1975.

W. P. Mason and R. N. Thurston (eds.), *Physical Acoustics*, Vols. 1–18. Academic Press, New York, 1964–1988.

N. G. McCrum, B. E. Read, and G. Williams, *Anelastic and Dielectric Effects in Polymeric Solids*. Wiley, New York, 1967.

A. S. Nowick and B. S. Berry, *Anelastic Relaxation in Crystalline Solids*. Academic Press, New York, 1972.

R. Truell, C. Elbaum, and B. B. Chick, *Ultrasonic Methods in Solid State Physics*. Academic Press, New York, 1969.

C. Zener, *Elasticity and Anelasticity of Metals*. Univ. of Chicago Press, Chicago, 1948.

Angular Correlation of Nuclear Radiation

R. M. Steffen[†]

The probability of emission of a particle or a quantum by a decaying nucleus depends, in general, on the angle between the nuclear spin axis **I** and the direction of emission **k**. In most cases (e. g., ordinary radioactive sources) the total radiation is isotropic, because the nuclear spin axes are randomly oriented in space. An anisotropic intensity distribution of the radiation is only observed if it is emitted from an ensemble of nuclei that is *not* randomly oriented, i. e., in which the spin axes of the decaying nuclei show some preferred direction in space. Such an ensemble is called an *oriented ensemble*.

[†] deceased

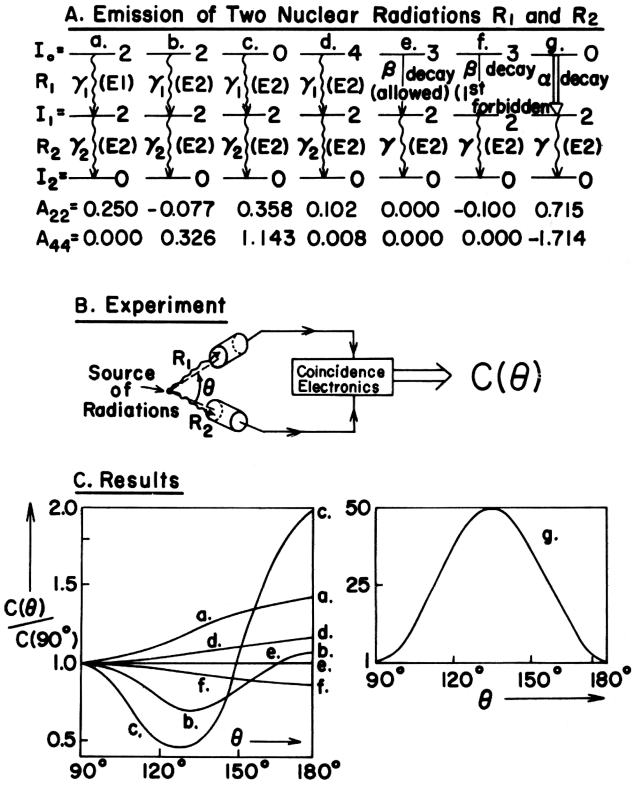

Fig. 1: Examples of directional correlations. (A) Typical gamma–gamma radiation cascades [(a)–(d)], beta–gamma radiation cascades [(e)–(f)], and an alpha–gamma cascade (g). Below each cascade are given the values of the directional correlation coefficients. (B) Experimental arrangements for measurement of directional correlations. (C) Directional correlations of the radiation cascades (a)–(g).

Oriented ensembles of nuclei can be prepared, e. g., by placing a radioactive sample at a very low temperature in strong magnetic or electrostatic gradient fields, thereby polarizing or aligning the nuclei by virtue of the interaction of the magnetic and electric moments of the nuclei with the external fields. The angular distribution of the radiation emitted by such an oriented source is then, in general, anisotropic with respect to the direction of the applied fields. Another method of preparing an oriented subensemble of nuclei is based on selecting only those nuclei whose spin axes happen to be in a preferred direction. Nuclear reactions or decay processes that lead to the formation of nuclei in a particular excited state of spin I_1 (the "intermediate" state) can be used in such a selection process.

Many nuclei decay through the *successive emissions of two radiations* R_1 and R_2 via a short-lived (lifetime $\tau \lesssim 10^{-9}$ s) intermediate nuclear state of spin I_1. Some examples of such cascade decays are depicted in Fig. 1A. The observation of RE in a fixed direction $\mathbf{k}_1$ selects from the originally random ensemble of nuclei with spin I_0 a subensemble of nuclei in the

intermediate state with spin I_1. This subensemble has, in general, a preferred direction of the spin axes $\mathbf{I}_1$ with respect to the observation direction $\mathbf{k}_1$ of R_1, because the radiation emission probability depends on the angle between $\mathbf{k}_1$ and $\mathbf{I}_0$. Since the second radiation R_2 is now emitted from this *oriented* subensemble of spin I_1, the intensity of R_2 observed in a direction $\mathbf{k}_2$, depends, in general, on the angle θ between $\mathbf{k}_1$ and $\mathbf{k}_2$, i. e., the second radiation R_2 has an anisotropic angular distribution with respect to the direction $\mathbf{k}_1$ in which R_1 has been observed. The angular distribution of R_2 with respect to $\mathbf{k}_1$ (or of R_1 with respect to $\mathbf{k}_2$) is called the *angular correlation* of the radiations R_1 and R_2.

If only the propagation directions (no polarization phenomena) of the two radiations are measured, the *directional* correlation is observed. If the linear or circular polarization of one or of both of the radiations is measured, a *polarization–directional correlation* or a *polarization–polarization correlation*, respectively, is observed. The term angular correlation comprises all three cases.

The observation of an angular correlation requires a coincidence experiment, i. e., the two radiations R_1 and R_2 must be recorded, each in one of two detectors that respond only if R_1 and R_2 strike the detectors simultaneously (actually within a very short time interval $\tau_0 \approx 10^{-9} - 10^{-8}$ s) in order to maximize the probability that the observed radiations R_1 and R_2 are emitted from the same nucleus. A directional correlation experiment consists thus simply of measuring the coincidence rate $C(\theta)$ of R_1 and R_2 as a function of the angle θ between the axes of the two detectors (Fig. 1B).

The relative probability $W(\theta)\,\mathrm{d}\Omega$ that R_2 is emitted into the solid angle $\mathrm{d}\Omega$ at an angle θ with respect to the propagation direction $\mathbf{k}_1$ of R_1 is characterized by the angular correlation function $W(\theta)$. For an *ordinary directional correlation*, $W(\theta)$ can be expressed in the general form

$$W(\theta) = 1 + A_{22}P_2(\cos\theta) + A_{44}P_4(\cos\theta)\ . \tag{1}$$

The angular functions $P_i(\cos\theta)$ are Legendre polynomials, i. e., $P_2(\cos\theta) = (3\cos^2\theta - 1)/2$ and $P4(\cos\theta) = (35\cos^4\theta - 30\cos^2\theta + 3)/8$. The directional correlation coefficients A_{ii} ($i = 2, 4$) can be expressed as the product, $A_{ii} = A_i(R_1, I_1; I_0) \cdot A_i(R_2, I_1; I_2)$, of two directional distribution coefficients $A_i(R_1, I_1; I_0)$ and $A_i(R_2, I_1; I_2)$, each being characteristic of one of the two emission processes R_1 and R_2 that make up the radiation cascade. The directional distribution coefficients $A_i(R, I; I')$ depend on the properties of the radiation R that is emitted in the transition $I \rightarrow I'$ and on the spins I and I' of the initial and final nuclear states, respectively. In particular, the distribution coefficients depend on the so-called multipolarity L of the emitted radiation R. A 2^L-pole radiation carries away an angular momentum of $L\hbar$ with respect to the center of the emitting nuclei. The directional distribution coefficients, however, do not depend on the reflection symmetry of the emitted radiation. For emission of gamma radiation, e. g., the directional distribution coefficients do not distinguish between electric 2^L-pole (*EL*) and magnetic 2^L-pole (*ML*) radiation. The observation of the linear polarization of the gamma radiation is required to distinguish between *EL* and *ML* radiation.

The directional distribution coefficients $A_i(\gamma, I; I')$ for gamma transitions do not depend on the energy of the gamma transitions. For alpha-particle emission the $A_i(\alpha, I; I')$ depend on the energy of the alpha particles only if two (or more) alpha-particle waves of different L interfere with each other. In beta emission two particles are emitted simultaneously, an electron (or positron) and an antineutrino (or neutrino), of which only the electron (or positron) is, in

general, observed in an angular correlation observation. The electrons (or positrons) have a continuous energy spectrum up to a maximum energy E_0 and the directional distribution coefficients for these electrons (or positrons) depend on the energy of the observed particle.

Theoretical expressions for the directional distribution coefficients (and for polarization distribution coefficients) are available for all types of radiations and for all cases of interest. For details see Refs. [1]–[4]. Four illustrative examples of various multipole gamma–gamma directional correlations are shown in Figs. 1A and 1C, (a)–(d). Beta–gamma directional correlations involving so-called allowed beta transitions are isotropic (e), first-forbidden beta–gamma directional correlations (f) are, in general, nonisotropic. Alpha–gamma directional correlations can show very large anisotropies (g).

Angular correlations are, in general, observed with the initial nuclear state of spin I_0, which emits R_1, randomly oriented (ordinary angular correlation). If R_1 itself is emitted from an oriented state, e. g., from a state produced in a nuclear reaction, the angular correlation from an oriented state (ACO) or the directional correlation from an oriented state (DCO) is observed.

An equivalent situation prevails in triple angular correlations where the angular correlation of the radiations R_1 and R_2 is observed with respect to the observation direction $\mathbf{k}_0$ of a preceding radiation R_0 that is emitted from a random ensemble I_{00} resulting in an oriented nuclear ensemble I_0 from which the radiation R_1 is emitted.

Directional correlations of two successively emitted gamma radiations R_1 and R_2 *emitted by an oriented nuclear ensemble* I_0 that is axially symmetric with respect to a direction $\mathbf{k}_0$ are characterized by a correlation function of the general form

$$W(\theta_1, \qquad (2)$$

$$theta_2; \varphi) = \sum_{i,k,l} B_l(I_0) A_l^{ki}(R_1, I_0, I_1) \times A_{ll}(R_2, I_1, I_2) H_{ikl}(\theta_1, \theta_2, \varphi) \qquad (3)$$

$$(i, k, l) = 0, 2, 4$$

where θ_1 and θ_2 are the polar angles, with respect to the orientation axis $\mathbf{k}_0$, of the directions $\mathbf{k}_1$ and $\mathbf{k}_2$ in which the radiations R_1 and R_2, respectively, are observed and the azimuthal angle φ is the angle between the planes determined by $\mathbf{k}_0\mathbf{k}_2$ and $\mathbf{k}_0\mathbf{k}_1$. The parameter $B_1(I_0)$ describes the state of orientation of the nuclear ensemble I_0 and $A_l^{ki}(R_1, I_0, I_1)$ is a generalized directional distribution coefficient. Expressions for the latter and for the angular function $H_{ikl}(\theta_1, \theta_2, \varphi)$ can be found in Ref. 5. DCO measurements are particularly useful in assigning spins to nuclear states that are produced in nuclear reactions and in exploring the multipole character of gamma radiations between such states.

In many experimental situations the time t elapsed between the formation of the intermediate oriented state I_1 by the radiation R_1 and the time moment of emission of the second radiation R_2 is long enough ($\sim 10^{-9} - 10^{-6}$ s) to cause an appreciable change of the orientation of the nuclear ensemble through the interactions of the electromagnetic nuclear moments (magnetic-dipole moment μ, electric-quadrupole moment Q) of the individual nuclei with external fields. In such cases the angular correlation can be influenced by the external fields and a perturbed angular correlation (PAC) is observed (see Refs. [6]–[8]).

A strong external (or internal atomic) magnetic field B, e. g., causes a precession of the magnetic moment μ of the nucleus in the intermediate state about the direction of B as axis (Fig. 2A) with a frequency ω_B that is proportional to B (Larmor precession). The angular

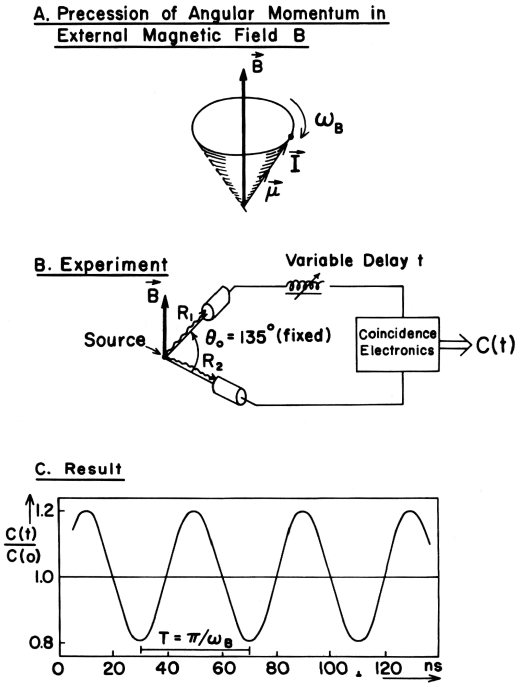

Fig. 2: Extranuclear perturbations by an external magnetic field. (A) Precession of spin about a magnetic field B (Larmor precession). (B) Delayed-coincidence observation. (C) Periodic variation with time t of the relative coincidence rate C(t)/C(0) reflecting the spin precession in the intermediate nuclear state.

distribution pattern of R_2 is then rotated about the direction of B by an angle $\Delta\theta = \omega_B t$. By observation of the angular shift $\Delta\theta$ of the angular correlation pattern, ω_B can be determined and thus either μ or B can be measured. For larger values of $\omega_B\tau$ (i. e., $\omega_B\tau \gtrsim 1$) and if $\tau \gg \tau_0$, the precession of the angular distribution pattern of R_2, i. e., the precession of nuclei in the intermediate state, can be directly observed by measuring the coincidence rate $C(t)$ in two fixed detectors as a function of the time during which the intermediate nuclear state is exposed to the magnetic field B. In practice this is done through delaying (electronically) the detector signal caused by R_1 by a time t before it reaches the coincidence circuit (Fig. 2B). The oscillating behavior of the observed coincidence rate as a function of t represents the spin precession of the nuclei in the intermediate state I_1 (Fig. 2C).

Angular correlation observations are a very important tool in nuclear spectroscopy for the determination of angular momenta and electromagnetic moments of excited nuclear states and for precise measurements of the multipolarities of nuclear radiations. Perturbed angular correlation experiments are also used to explore the magnetic and electric field gradients at the site of nuclei and thus can be applied to atomic, solid-state, and liquid-state problems.

See also: Alpha Decay; Beta Decay; Gamma Decay; Multipole Fields; Nuclear Polarization; Polarization.

References

[1] H. Frauenfelder and R. M. Steffen, "Angular Correlations," in *Alpha, Beta and Gamma Ray Spectroscopy* (K. Siegbahn, ed.), pp. 997–1198. North-Holland, Amsterdam, 1965. (I)

[2] R. M. Steffen and K. Alder, "Angular Distributions and Correlations of Gamma Radiation: Theoretical Basis," in *The Electromagnetic Interaction in Nuclear Spectroscopy* (W. D. Hamilton, ed.), pp. 505–581. North-Holland, Amsterdam, 1975. (A)

[3] S. Devons and L. J. B. Goldfarb, in *Handbuch der Physik* (S. Flügge, ed.), Vol. 42, p. 362. Springer-Verlag, Berlin and New York, 1957. (A)

[4] A. J. Ferguson, *Angular Correlation Methods in Gamma Ray Spectroscopy*. North-Holland, Amsterdam, 1965. (A)

[5] K. S. Krane, R. M. Steffen, and R. M. Wheeler, "Directional Correlations of Gamma Radiations Emitted from Nuclear States Oriented by Nuclear Reactions or Cryogenic Methods," *Atomic and Nuclear Data Tables, Vol. 11*, pp. 351–405. Academic Press, New York, 1975. (A)

[6] R. M. Steffen, "Extranuclear Effects on Angular Correlations of Nuclear Radiation," *Adv. Phys. (Phil. Mag. Suppl.)*, **4**, 293–362 (1955). (E)

[7] R. M. Steffen and H. Frauenfelder, "The Influence of Extranuclear Perturbations in Angular Correlations," in *Perturbed Angular Correlations* (E. Karlsson, E. Mathias, and U. Siegbahn, eds.), pp. 1–89. North-Holland, Amsterdam, 1964. (A)

[8] R. M. Steffen and K. Alder, "Extranuclear Perturbations and Angular Distributions and Correlations," in *The Electromagnetic Interaction in Nuclear Spectroscopy* (W. D. Hamilton, ed.), pp. 583–643. North-Holland, Amsterdam, 1975. (A)

Antimatter

G. Steigman

All quantum theories consistent with the special theory of relativity and the requirement of causality require that particles exist in pairs. Particles and their antiparticles have the same masses and lifetimes; electrically charged particles (e. g., electron, proton) have antiparticles (e. g., positron, antiproton) with equal but opposite electric charges; some electrically neutral particles (e. g., photon) are their own antiparticles (self-conjugate). Following the discovery of the positron (Anderson 1933) there was a hiatus of some 22 years before the antiproton was produced and detected at an accelerator (Chamberlain *et al.* 1955). Subsequent accelerator experiments have provided strong confirmation that all particles do, indeed, exist in pairs and, further, that particles carry certain quantum numbers (baryon number, lepton number, etc.) which *seem* to be conserved in all reactions. If these conservation "laws," inferred from experimental data, are exact, then matter is restricted to appear (creation) or disappear (annihilation) only as particle–antiparticle pairs. This apparent symmetry (about which, more later) in the laws of physics has stimulated serious speculation on the antimatter content of the Universe (Goldhaber 1956) and the possible astrophysical consequences of macroscopic amounts of antimatter (Burbidge and Hoyle 1956). In approaching the issue of the matter–antimatter symmetry (or, asymmetry) of the Universe, it is valuable to distinguish between two distinct questions: Is the Universe symmetric? Must the Universe be symmetric? The first question will be considered before a "modern" (a la "Grand Unified Theory") answer is given to the second question. For further details and references see Steigman (1976).

Searching for Antimatter in the Universe

Antimatter is, in principle, trivially easy to detect. You place your sample in a detector (the most rudimentary device will suffice) and, if the detector disappears (annihilates), the sample was made of antimatter. Unfortunately, only the solar system and the cosmic rays provide a sample of the Universe which may be subjected to such a direct test. Lunar landings and the Venus probes establish that the Moon and Venus are made of ordinary matter. The absence of annihilation gamma rays when the solar wind sweeps through the solar system establishes that the Solar System consists only of (ordinary) matter; were any of the planets made of antimatter, annihilation of the solar wind particles which strike their surfaces would have made them the strongest gamma ray sources in the sky.

Cosmic rays provide the only direct sample of extrasolar system material in the Universe. The cosmic rays, perhaps the debris of exploding stars (supernovae) or the accelerated nuclei of interstellar gas, bring information about the material in our Galaxy. As the cosmic rays traverse the Galaxy they collide with interstellar gas nuclei, occasionally producing (secondary) antiprotons. Therefore, antiprotons in the cosmic rays do not provide an unambiguous signal for the presence of "primary" sources of antimatter (e. g., antistars) in the Galaxy. In contrast, virtually no antihelium (antialpha) nuclei would be present as secondaries in the cosmic rays. The discovery of even one antialpha particle in the cosmic rays would provide compelling evidence for the existence of antimatter in the Galaxy. None has ever been found. In contrast, antiprotons have been observed in the cosmic rays (Golden *et al.*, 1979; Bogomolov *et al.*, 1979). However, the upper limits to the $\bar{p}$ flux at low energies (Ahlen *et al.*, 1988) are completely consistent with a secondary origin; these latest results are in conflict with – and cast doubt on the reality of – an earlier claim (Buffington *et al.*, 1981) of a positive detection.

Astronomy is an observational science, often relying on the interpretation of indirect evidence. When matter and antimatter meet, they annihilate producing, among the debris, gamma rays of energy from several tens to several hundred MeV. The annihilation gamma rays can provide an indirect probe for antimatter in the Universe. However, since gamma rays may be produced by other astrophysical processes (synchrotron radiation, Compton scattering, etc.), they are not an unambiguous signal; the best approach is to use the observed gamma ray flux to place upper limits on annihilation and, hence, on the amount of mixed matter and antimatter at various astrophysical sites (Steigman 1976).

Gamma ray observations of the Galaxy limit the antimatter fraction in the interstellar gas to less than 10^{-15}. Such a small limit is not surprising once it is realized that the lifetime – against annihilation – of an antiparticle in the interstellar gas is less than 300 years (the Galaxy is at least 10 billion years old). Our galaxy is clearly made entirely of ordinary matter. What of other galaxies?

Clusters of galaxies are the largest astrophysical entities which can be probed by gamma rays for a possible antimatter component. Many rich clusters shine in the x-ray part of the spectrum; the x-rays are the thermal bremsstrahlung emission from a hot intracluster gas. From the absence of gamma rays, less than one part in 10^5 of that hot gas could be made of antimatter. Thus, the observational data shows no evidence for astrophysically interesting amounts of antimatter in the Universe. If antimatter were present, it would have to be separated from ordinary matter on scales at least as large as clusters of galaxies ($\sim 10^{15}\,M_\odot \sim 10^{48}\,\mathrm{gm} \sim 10^{72}$ protons). Could matter and antimatter be separated in the Universe on such large scales?

Symmetric Cosmologies

In the context of the hot big bang model, particle–antiparticle pairs were present in great abundance during the early ($\lesssim 10\,\mu$s), hot ($T \gtrsim$ few hundred MeV) epochs in the evolution of the Universe. As the Universe expanded and cooled, however, these pairs annihilated. In a completely symmetric, perfectly mixed Universe the annihilation is so efficient that less than one nucleon–antinucleon pair remains for each 10^{18} microwave background photons in the Universe; this is less matter than is present in our Galaxy alone! Either the Universe at these early times ($\lesssim 10\,\mu$s) was asymmetric or, possibly, matter and antimatter were separated. However, although the Universe was very dense, it was also very small and only ~ 1 g of nucleons (antinucleons) could have been separated by any causal process (none is known which would effect such a separation). Inevitably, then, astrophysicists have led the way to the conclusion that the early Universe ($\lesssim 10\,\mu$s) was asymmetric.

Baryon Asymmetry and GUT

Inspired by the (then recent) discovery of *CP* violation in the $K^0 - -\bar{K}^0$ system, Sakharov (1967) outlined the recipe for generating a baryon asymmetry in an initially (matter–antimatter) symmetric Universe. First, there must be interactions which violate conservation of baryon number (matter–antimatter symmetry is a "broken" symmetry). Next, he noted a technical – but crucial – requirement that *CP* conservation be violated (e. g., the branching ratios for decays into certain channels for particles and antiparticles differ; note that the total lifetimes are still required to be equal). Finally, Sakharov (1967) pointed out that these *B*- and *CP*- violating processes must occur "out of equilibrium." The offspring of the marriage of particle physics (Grand Unified Theories) and cosmology (the expanding, hot big bang model) is endowed with the requisite properties (Yoshimura 1978; Dimopoulos and Susskind 1978; Ellis, Gaillard and Nanopoulos 1979; Toussaint *et al.*, 1979; Weinberg 1979). Baryon-number- and *CP*-violating interactions occurring very early ($\lesssim 10^{-35}$ s) in the evolution of the Universe would have, due to the expansion and cooling of the Universe, dropped out of equilibrium and left behind a small matter–antimatter asymmetry (the net baryon number is $\sim 10^{-10} - 10^{-9}$ of the photon number). This tiny relic, the legacy of the earliest epochs in the evolution of the Universe, is responsible for the presently observed, matter–antimatter asymmetric Universe.

See also: Cosmology; Elementary Particles; Positron.

Bibliography

S. P. Ahlen *et al.*, *Phys. Rev. Lett.* **61**, 145 (1988).
C. D. Anderson, *Phys. Rev.* **43**, 491; 44, 406 (1933).
E. A. Bogomolov *et al.*, *Proc. of the 16th Int. Cosmic Ray Conf.* **1**, 330 (1979).
A. Buffington, S. M. Schindler, and C. R. Pennypacker, *Astrophys. J.* **248**, 1179 (1981).
G. R. Burbidge and F. Hoyle, *Nuovo Cimento* **1**, 558 (1956).
O. Chamberlain, E. Segre, C. Wiegand, and T. Ypsilantis, *Phys. Rev.* **100**, 947 (1955).
S. Dimopoulos and L. Susskind, *Phys. Rev.* **D18**, 4500 (1978).
J. Ellis, M. K. Gaillard, and D. V. Nanopoulos, *Phys. Lett.* **B80**, 360 (1979).
R. L. Golden *et al.*, *Phys. Rev. Lett.* **43**, 1196 (1979).

M. Goldhaber, *Science* **124**, 218 (1956).
A. Sakharov, *JETP Len.* **5**, 24 (1967).
G. Steigman, *Ann. Rev. Astron. Astrophys.* **14**, 339 (1976).
D. Toussaint, S. Treiman, F. Wilczek, and A. Zee, *Phys. Rev.* **D19**, 1036 (1979).
S. Weinberg, *Phys. Rev. Lett.* **42**, 850 (1979).
M. Yoshimura, *Phys. Rev. Lett.* **41**, 381 (1978).

Arcs and Sparks

U. H. Bauder

General Properties

The electric arc is characterized by high current densities, low potential differences between the electrodes, and small differences compared to other discharges between the temperatures of the different particle species present in the column. In equilibrium temperatures exceeding 50 000 K have been reached [1]. Arcs can be operated over a wide pressure range: from several millibars gas pressure to 1000 bars and in all gaseous media. The initiation of an arc discharge may be achieved either by contacting the electrodes or by preionizing the gas in the discharge channel by breakdown processes following Paschen's law [2]. The unstationary discharge preceding a sustained arc discharge is the spark; besides duration time the main difference between arcs and sparks relates to electrode effects. Most stationary arc discharges operate with a combined thermal and field emission of electrons at a hot cathode [3], whereas duration times of sparks are too small to allow for a substantial increase of the bulk temperature of electrodes.

Depending on the external conditions under which the arc is operated one distinguishes between electrode-stabilized (short) arcs, flow-stabilized arcs (longitudinal or swirl flow), wall-stabilized arcs (operated in a segmented cascaded tube if high power levels have to be achieved), high pressure arcs, and vacuum arcs.

The High-Pressure Arc Column

Arcs operated at atmospheric pressure or above are high-pressure arcs. Once established by one of the ignition processes mentioned above, the arc column is sustained by ionizing processes whose energy is supplied by the electric field generated by the power supply. Free electrons coming from the cathode are accelerated in the cathode fall region and in the column during a free path length between two collisions. High-pressure arc columns are characterized by mean free path lengths which are much smaller than the geometrical dimensions of the arc column. Due to their small mass, electrons reach higher velocities in the field than ions; mainly collisions of electrons with molecules, atoms, and ions lead to further ionization. Ionization is possible from the ground state as well as starting from excited states. Depending on the gas species as well as on the plasma density and temperature, photoionization processes may also play a role. Carriers are lost by recombination processes. In some application plasmas such as SF_6 plasmas, electron attachment with the formation of negative ions may also reduce the

number density of free electrons in the column. The large difference between the masses of electrons and heavy particles frequently leads to thermal nonequilibrium in the plasma. It takes approximately 10^3–10^4 elastic collisions of electrons with atoms or ions until thermalization is achieved. Due to this fact local thermodynamic equilibrium (LTE) is only achieved in high-temperature arc plasmas with high collision rates (elevated pressures); in other arc columns partial local thermodynamic equilibrium (PLTE) prevails. In the case of PLTE only the higher excited states are populated in equilibrium with the electron temperature, the ground state being largely overpopulated.

Considering the differential energy balance of the cylindrical arc column without convection losses [4] as it is realized experimentally in the cascade arc [5],

$$\sigma E^2 + \frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}\left(r\kappa\frac{\mathrm{d}T}{\mathrm{d}r}\right) - e + a = 0\,. \tag{1}$$

it can be noted that the energy input to the volume element of the plasma column is mainly due to Ohmic heating (σ being the electrical conductivity and E the electric field strength). Absorption processes may also play a role; the quantity a is the volumetric absorption coefficient integrated over all frequencies. The energy is lost from the column by radiation (e = volumetric emission coefficient) and by thermal conduction. The thermal conductivity κ is composed of the contact conductivities of all species; in the temperature regions of dissociation of molecules and that of ionization a large contribution to κ may be due to the radial transport of dissociation and ionization energies. Due to this transport, temperature profiles of arc discharges operating in molecular gases exhibit a "shoulder" as shown in Fig. 1 [6] for the case of an atmospheric pressure nitrogen arc operated in a cascade channel of 5 mm diameter. By contrast, the $T(r)$ profiles of noble gas cascade arc columns at elevated pressures have a flat central portion and very steep gradients at the wall (Fig. 2 [7]). The energy balance of the central portion of such arcs is largely radiation dominated. It has been shown [8] that up to 80% of the energy of a high-pressure argon arc may be lost by radiation; this property of the column is used in high-energy radiation source applications.

The field-strength/current (E–I) characteristic of the high-pressure arc column exhibits a falling portion at small currents and a rising part in the higher current range. Stable arc operation at small currents (point C in Fig. 3 [9]) therefore requires an Ohmic resistance in series with the arc; thus the resulting total characteristic can be made to increase at the current of interest and stable operation is achieved. Introducing the heat flux potential $S = \int \kappa\,\mathrm{d}T$ into the transport function $\sigma(T) = \sigma(S)$ and neglecting radiation Eq. (1) may be rewritten

$$\sigma E^2 + \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial S}{\partial r}\right) = 0\,. \tag{2}$$

Defining the Ohmic heating per unit arc length as $L = IE$ and normalizing the arc radius $v = r/R$ (R being the total radius) Eq. (2) yields

$$\frac{\sigma L}{\pi\int_0^1 \sigma\,\mathrm{d}(v^2)} + \frac{1}{v}\frac{\partial}{\partial v}\left(v\frac{\partial S}{\partial v}\right) = 0\,. \tag{3}$$

Since Eq. (3) is independent of the total radius R of the column, equal values of L lead to identical $T(r)$-distributions. This independence on radius may be used to predict electrical arc data (E, I) for arcs of different diameters (similarity laws).

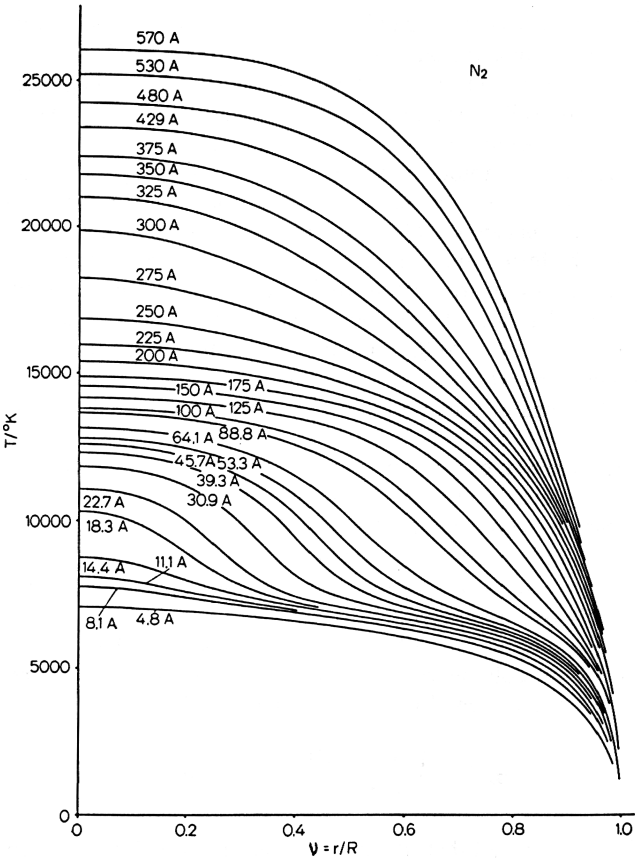

Fig. 1: Temperature distributions in a 5-mm nitrogen cascade arc (from [6], with permission).

Due to its importance for applications the interaction of flow fields and magnetic fields with the arc column has been studied extensively [10–12]. By solving the MHD conservation laws with the appropriate boundary conditions, predictions of arc motion, gas flow in the arc and arc stability may be obtained. The energy Eq. (1) has to be used in its more general form:

$$\rho\frac{\mathrm{d}h}{\mathrm{d}t}=\rho\frac{\partial h}{\partial t}+\rho\mathbf{v}\cdot\boldsymbol{\nabla}h=\sigma E^2+\boldsymbol{\nabla}\cdot\kappa\boldsymbol{\nabla}T-e+a\,, \tag{4}$$

where $\mathbf{v}$ is the velocity of the gas, ρ is its mass density, and h the enthalpy. In addition, the momentum and the mass balance equations (5) and (6) have to be considered:

$$\rho\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t}=\rho\frac{\partial\mathbf{v}}{\partial t}+\rho\mathbf{v}\cdot\boldsymbol{\nabla}\mathbf{v}=\mathbf{j}\times\mathbf{B}-\boldsymbol{\nabla}p+\boldsymbol{\nabla}\cdot\tau\,, \tag{5}$$

$$\frac{\partial\rho}{\partial t}+\boldsymbol{\nabla}\cdot(\rho\mathbf{v})=0 \tag{6}$$

with $\mathbf{v}$ being the friction tensor.

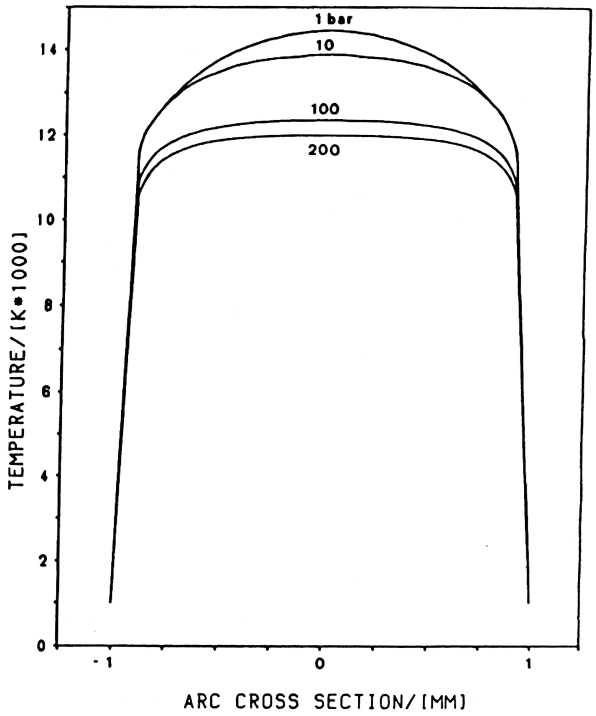

Fig. 2: Temperature distributions in an argon arc at elevated pressures. Arc current 60 A [7].

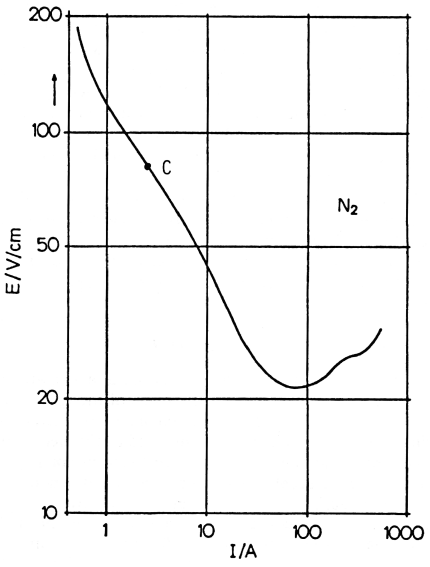

Fig. 3: Electrical arc characteristic (from [9], with permission).

Vacuum Arcs

This arc discharge is initiated in an evacuated environment. Since plasma can only be formed if ionizable gas is present, the vacuum arc has to generate its own gaseous environment. Vaporization of the electrodes, mainly that of the cathode material, provides the necessary atoms. Vacuum arcs therefore operate typically in a metal vapor atmosphere. Since the crater formation process which is necessary for the evaporation at the cathode ends after a certain crater size has been reached [13], vacuum arcs are not stationary in nature. The cathode spot only remains at a given crater for times shorter than 50 ns; thus a single (undivided) short vacuum arc is very similar in nature to a spark.

If magnetic fields are present, the vacuum arc moves against the $\mathbf{j} \times \mathbf{B}$ direction; this "retrograde motion" is due to cathode jet phenomena. At higher currents (above $50 - 100$ A) the single vacuum arc splits up; a multitude of cathode or anode spots may coexist in such arcs.

Applications

Arc discharges are used in a host of scientific and technical applications. Due to the temperature range which is achievable in the column ($T \leq 50\,000$ K) different ionization stages of the gas atoms are reached. Fundamental data such as collision cross sections, transition probabilities, index of refraction, and spectral line broadening parameters can be determined from the investigation of arc plasmas.

Technical applications also make use of the high plasma temperatures which lead to an excellent electrical conductivity, to high radiative power losses, to high enthalpies in the gas – to name only the most important properties. The high electrical conductivity of an arc column, together with the short recovery time of dielectric strength after current zero, is used in AC-circuit breakers. Very high voltage levels can be handled with gas filled-gas breakers (SF_6) where the interaction of flow fields and magnetic fields with the arc column plays an important role. Vacuum circuit breakers are much more compact, however, their voltage handling capability is reduced. At present, voltages of 50–60 kV can be handled. The high specific radiation of noble gases and rare earths is used in radiation source applications, while the high enthalpies reached in arc heaters allow for special chemical reactions and for surface treatment processes, such as spark etching and plasma coating.

See also: Corona Discharge; Ionization; Lightning; Photoionization.

Bibliography

H. Maecker and W. Finkelnburg, "Elektrische Bögen und thermisches Plasma", *Handbuch der Physik*, Bd. XXII. Springer-Verlag, Berlin, 1957.

J. M. Lafferty, ed., *Vacuum Arcs, Theory and Application*. Wiley, New York, 1980.

H. Raether, *Electron Avalanches and Breakdown in Gases*. Butterworths, London, 1964.

K. Günther and R. Radtke, *Electric Properties of Weakly Nonideal Plasmas*. Birkhaeuser, Basel, 1984.

K. Ragaller, ed., *Current Interruption in High-Voltage Networks*. Plenum, New York, 1978.

References

[1] F. Burhorn, H. Maecker, and Th. Peters, *Z. Phys.* **131**, 28 (1951).

[2] F. Paschen, *Ann. Phys. (Leipzig)* **37**, 69 (1889).

[3] G. Burkhard, PhD Thesis, Technische Hochschule Ilmenau (1971).
[4] W. Elenbaas, *Physica* **3**, 947 (1936).
[5] H. Maecker and S. Steinberger, *Z. Angew. Phys.* **23**, 456 (1967).
[6] E. Schade, *Z. Phys.* **233**, 53 (1970).
[7] H. Poisel, F. J. Landers, P. Höß, and U. H. Bauder, *IEEE Trans. Plasma Sci.* **PS-14**, 306 (1986).
[8] U. H. Bauder and P. Schreiber, *Proc. IEEE* **59**, 633 (1971).
[9] U. Plantikow, *Z. Phys.* **237**, 388 (1970).
[10] H. Maecker and H. G. Stablein, *IEEE Trans. Plasma Sci.* **PS-14**, 291 (1986).
[11] N. Sebald, *Proc. XIIth ICPIG*, North Holland/American Elsevier Eindhoven, p. 187 (1975).
[12] J. Blass and U. H. Bauder, *Proc. IIW Asian Pacific Regional Welding Congress*, p. 528. Hobart, Australia, 1988.
[13] J. Prock, *IEEE Trans. Plasma Sci.* **PS-14**, 482 (1986).

Astronomy, High-Energy Neutrino

E. Waxman

Most of the information we have on astronomical objects and systems is obtained by observing the electromagnetic radiation, i. e., the photons, they emit. In the past four decades a new type of astronomy has emerged, where information is carried to us from astronomical objects not by photons but rather by a different type of particle, the neutrino. Neutrinos are nearly massless particles traveling at essentially the speed of light (*see* Neutrino). The great strength of neutrino astronomy is related to the fact that neutrinos interact very weakly with matter. Looking at the Sun, for example, we can only observe photons emitted from the Sun's surface. Photons produced within the Sun can not reach us, since they can not propagate much through the dense solar plasma. Neutrinos produced in the depth of the Sun, on the other hand, can propagate almost unhindered through the Sun and reach our detectors on Earth, carrying information on the physical processes taking place in the core of the Sun.

The strength of neutrino astronomy is also its challenge. Since neutrinos interact only weakly with matter, large detector mass is required in order to detect them. The detection of solar neutrinos became possible with the construction of detectors with kilo-tons of detecting medium, *see* Solar Neutrinos [3]. The probability that a neutrino passing through kilo-tons of matter would be "captured", i. e., would interact within the detector, is still very small. However, the large flux of neutrinos from the Sun, some 100 billion neutrinos per square centimeter per second, allows hundreds of them to be detected every year. The detection of solar neutrinos enabled direct observations of nuclear reactions in the core of the Sun, confirming the hypothesis that the energy source of the Sun is nuclear fusion and demonstrating the validity of solar structure models. It has also taught us that the standard model of particle physics is incomplete. Neutrinos come in three types, or "flavors": electron-type, muon-type and tau-type. Solar neutrino detection demonstrated that neutrinos can change their flavor as they propagate in matter or in vacuum, a phenomenon not accounted for by the standard model. This flavor change also indicates that neutrinos are not massless, as assumed in the standard model, but rather have finite, albeit small, masses (*see* Neutrino; Solar Neutrinos).

The characteristic energy of neutrinos produced in the Sun is Mega (million) electron-Volt, MeV. An eV is the energy typically required to detach an electron from an atom, while MeV is the characteristic energy released in the fusion or fission of atomic nuclei. Solar neutrino detectors, "telescopes", are capable of detecting MeV neutrinos from supernova explosions in our "local" Galactic neighborhood, at distances smaller than 100 000 light years. Since the rate of such explosions is one per few decades, only one supernova, the famous SN 1987A, has so far been detected. Much like the detection of solar neutrinos, the detection of neutrinos from SN 1987A provided a direct observation of the physical process powering the explosions, the collapse of the core of a massive star to a neutron star, as well as constraints on fundamental neutrino properties [14].

The detection of MeV neutrinos from sources well outside our local neighborhood, lying at distances ranging from several million light years, the typical distance between galaxies, to several billion light years, the size of the observable universe, is impossible using present techniques. In order to extend the distance accessible to neutrino astronomy to the edge of the observable universe, several high energy neutrino telescopes are currently being constructed [10]. These telescopes are designed for the detection of neutrinos with energies exceeding Terra-electron-Volt (TeV, equal to million-MeV).

The detection of astrophysical high energy neutrinos will allow to answer some of the most important open questions of high energy astrophysics [19], e. g., the identity and physics of the most powerful accelerators in the universe and the mechanisms for energy extraction from black-holes. It will also allow to study fundamental neutrino physics issues, e. g., neutrino coupling to gravity and the existence of weakly interacting massive particles (WIMPs). These issues are discussed at some length below. It should be kept in mind, however, that as the construction of high energy neutrino telescopes opens a new, unexplored window of observations on the universe, one should be ready for surprises. It may well be that the most important things that we will learn would be related to such surprises, rather than to the open questions discussed below.

High Energies for Large Distances

The neutrino flux from cosmologically distant sources is too low to be detectable by kiloton detectors of MeV neutrinos. The construction of orders of magnitude larger detectors, that would be required for the detection of extragalactic sources at this energy, is currently unfeasible. This situation changes at higher neutrino energy, due to two reasons. First, the interaction cross section increases with energy, that is, higher energy neutrinos are more likely to interact with matter than lower energy ones. This implies that smaller fluxes of higher energy neutrinos may be detectable with a given detector mass. Second, at TeV neutrino energy the construction of gigaton, rather than kiloton, telescopes becomes feasible.

Interactions of high energy, > 1 TeV, muon-type neutrinos with atomic nuclei on Earth produce muons (charged particles about hundred times heavier than the electron), which propagate at a straight line and at nearly the speed of light over more than a kilometer through rock, water or ice. While propagating at nearly the speed of light, the muon emits visible light, "Čerenkov radiation". Thus, if the muon propagates through transparent water or ice, its "track" may be identified by detecting the light it emits. Since the muon track is colinear to within one degree with the initial neutrino trajectory, the direction to the neutrino source may be determined.

The feasibility of detection of high energy muons in deep sea or lake water has been demonstrated by the DUMAND experiment off the coast of Hawaii and by the Lake Baikal experiment. The AMANDA collaboration has demonstrated that the deep ice of the south pole is also a suitable medium, and that construction of a cubic kilometer ice detector, with a gigaton of ice as detecting medium, is feasible. A schematic description of the AMANDA detector is presented in Fig. 1.

The reason for constructing neutrino telescopes in deep water or ice is twofold. First, the ice and water properties improve with depth. The scattering of light is weaker at large depth, thus allowing to place the photomultipliers, the detectors of the muon Čerenkov light, at larger spacing. This implies that at larger depth a smaller number of photomultipliers is required to instrument a given detector volume, and allows the instrumentation of a km^3 of ice or water at an acceptable cost. Second, the atmospheric muon background to the neutrino signal is reduced at larger depth. High energy muons are produced in the atmosphere of the Earth by interaction of cosmic rays with air. The cosmic rays which hit the atmosphere are mostly high energy nuclei, believed to be produced in Galactic supernovas. The muons produced by their interactions in the atmosphere constitute a background to the signal of neutrino-induced muons, that is, to the signal of muons produced by neutrino interactions. Since the muons penetrate a distance of order a km in ice or water, putting the detector several kilometers deep under water or ice strongly suppresses the atmospheric muon background. The large depth does not affect the neutrino induced muon signal, since neutrinos can easily penetrate the Earth and interact near or within the detector.

Figure 2 shows a neutrino event detected by the AMANDA telescope. The fact that the muon track is "upgoing", crossing the detector from bottom to top, allows to confidently identify it as a as neutrino-induced muon: Only neutrinos can cross the Earth to approach the detector from below and produce upgoing muons. This neutrino is, most likely, an "atmospheric neutrino". The interaction of cosmic rays in the atmosphere produce both muons and neutrinos. While the atmospheric muon flux is suppressed by going to large depth, the atmospheric neutrino flux is not. Atmospheric neutrinos constitute therefore an unavoidable background, which sets a lower limit to the flux of astronomical sources which are detectable (for a given detector size).

The AMANDA detector, which is roughly 0.1 gigaton in mass, is currently expanded in the IceCube project to a km^3, 1 gigaton, detector. In addition, several efforts are currently underway in the Mediterranean to construct gigaton-scale underwater detectors (the ANATRES, NESTOR and NEMO projects).

Cosmic Accelerators as Neutrino Sources

Nuclear fusion in stars generally does not lead to production of neutrinos at energy much higher than MeV. Should we expect, therefore, any sources of TeV neutrinos to be out there? If so, what is the mechanism by which they produce neutrinos, and what is the detector size that is required to detect them? The answers to these questions rely largely on observations of high energy cosmic rays.

The cosmic-ray spectrum extends to energies of 10^{20} eV, 100 million TeV. We have strong indications that ultrahigh-energy cosmic rays (UHECRs), cosmic rays of energy exceeding 10 million TeV, are produced by extragalactic sources, and that they are light nuclei, most

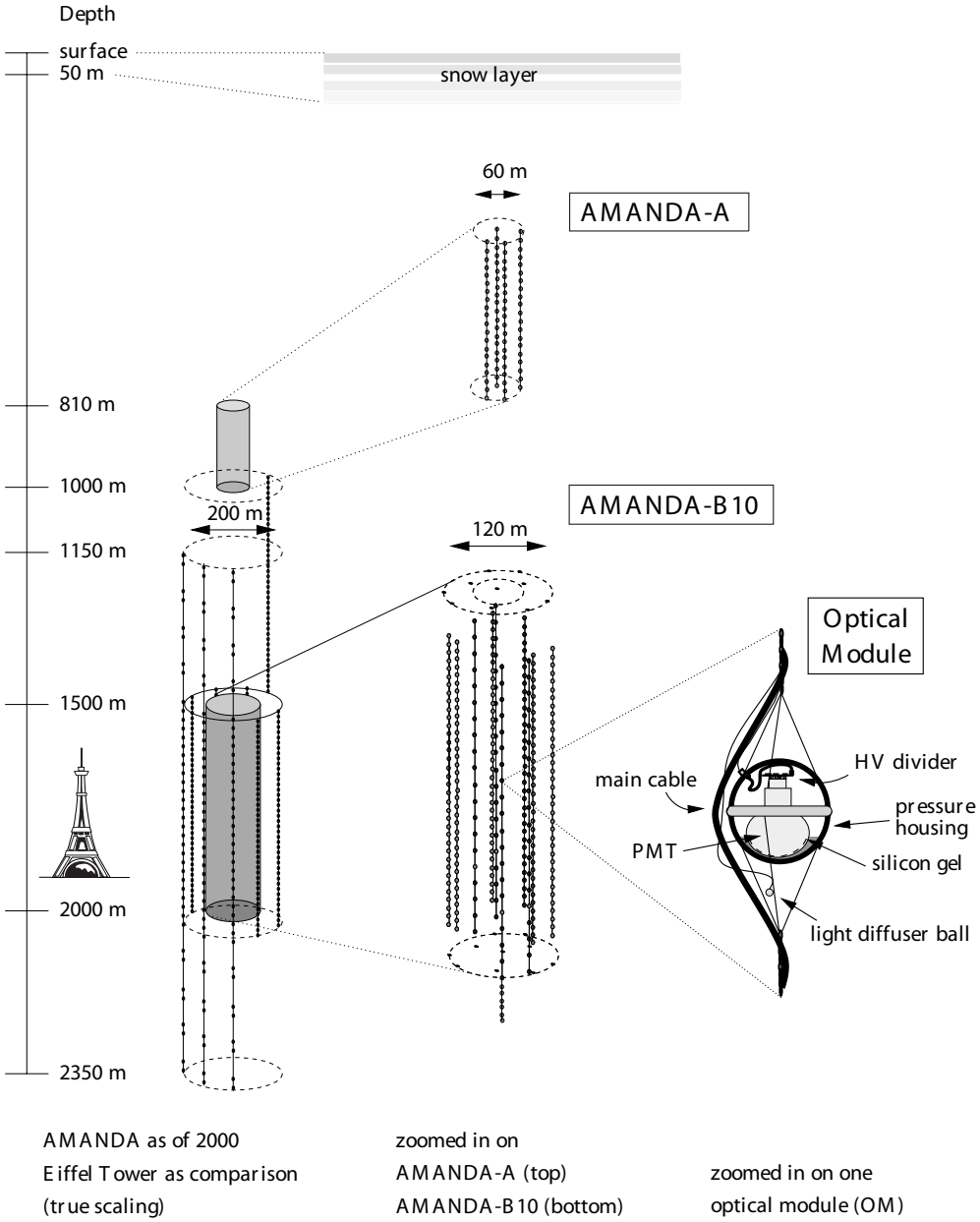

Fig. 1: The AMANDA experiment. > 1 TeV muons, produced by high-energy neutrinos interacting with atomic nuclei near or within the detector, are identified by an array of photomultipliers (PMTs) deployed 2 km deep under the South-Pole surface. Detection of the muon emission of visible, Čerenkov, radiation allows to reconstruct the > 1 kilometer-long muon track. A neutrino event recorded in AMANDA is presented in Fig. 2.

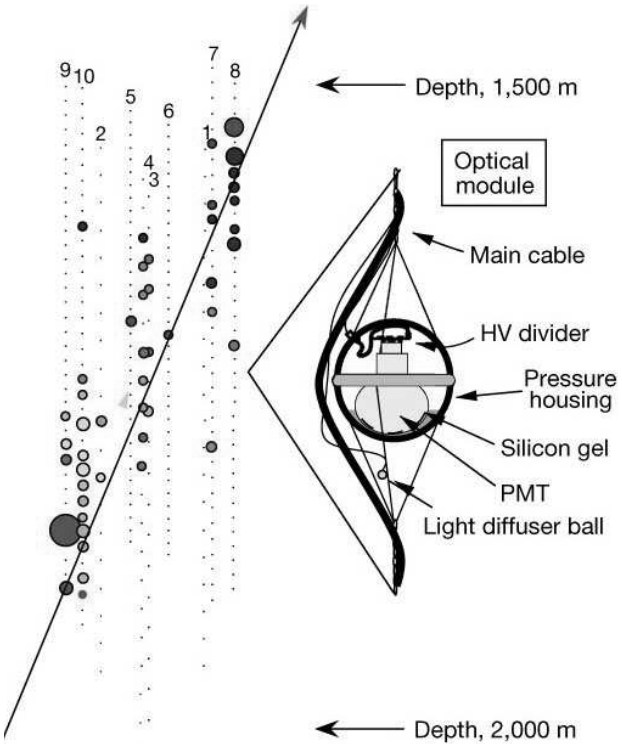

Fig. 2: A high-energy muon track reconstructed by the AMANDA detector [E. Andres *et al.*, *Nature* **410**, 441 (2001)]. Each dot represents an optical module, shown in detail on the right. The gray circles show pulses from the PMTs: The size of the circle indicates the pulse amplitude. The arrow indicates the upward-moving muon track.

likely protons [13]. The identity of the sources is yet unknown: The high energies, 100 million times larger than the highest energy achieved by man made accelerators, challenge all models proposed for particle acceleration. Moreover, the cosmic-ray arrival directions do not necessarily point back to their sources, since protons (being charged particles, unlike photons and neutrinos) are deflected by Galactic and extragalactic magnetic fields, and do not propagate along straight lines.

The essence of the challenge of accelerating to 10^{20} eV, or 100 million TeV, can be understood from Fig. 3. Most models involve the acceleration of charged particles, like protons, which are confined to the accelerator by magnetic fields. Magnetic confinement requires the product of accelerator magnetic field strength and accelerator size to exceed a value, which increases with particle energy. Only two types of astrophysical sources are known to be large enough and to contain magnetic fields strong enough to possibly allow proton acceleration to 10^{20} eV: Gamma-Ray Bursts (GRBs) [17] and Active Galactic Nuclei (AGN) [9, 12]. GRBs are the brightest transient sources known in the universe. Lying at cosmological distances, they produce short (typically 1 to 100 s long) flashes of γ-rays with luminosity exceeding that of the Sun by 19 orders of magnitude. AGN are the brightest known steady sources, with

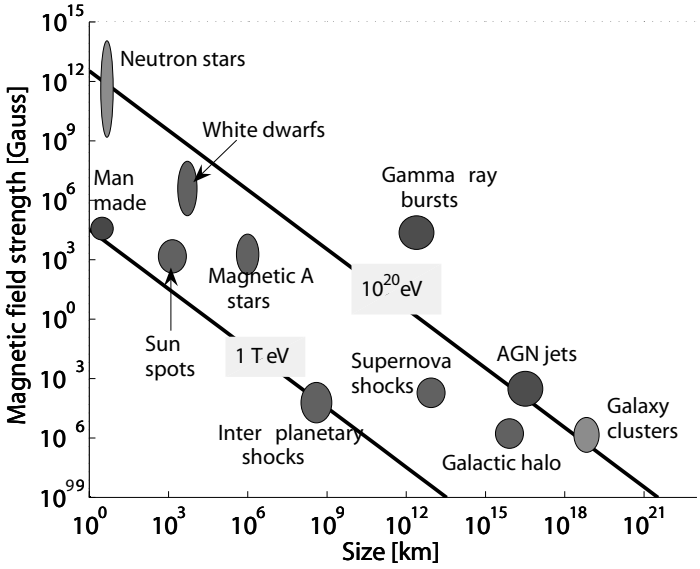

Fig. 3: Size and magnetic field strength of possible sites of particle acceleration (following [9]). Proton acceleration to 1 TeV or 10^{20} eV is possible only for sources lying above the appropriately marked lines. This is a necessary, but not sufficient requirement: Proton acceleration to 10^{20} eV is impossible in galaxy clusters, since the acceleration time in these objects is larger than the age of the universe, and unlikely in highly magnetized neutron stars, due to severe energy losses.

luminosity exceeding that of the Sun by 12 orders of magnitude. While GRBs and AGN are plausible candidates for UHECR production, we have no direct evidence for proton acceleration in these sources despite many years of photon observations. Furthermore, our theoretical models describing these sources are incomplete, a point to which we return below.

Irrespective of the nature of the UHECR sources, some fraction of their energy output is bound to be carried by high energy neutrinos. Protons, p's, of sufficiently high energy may interact with photons, γ's, to produce charged pions, π^+'s, particles that decay to muons and neutrinos. This interaction is represented symbolically as

$$\mathrm{p} + \gamma \rightarrow \mathrm{n} + \pi^+ \,, \tag{1}$$

where n stands for a neutron. The subsequent decay of the pion produces neutrinos:

$$\pi^+ \rightarrow \nu_\mu + \mu^+ \rightarrow \nu_\mu + \mathrm{e}^+ + \nu_\mathrm{e} + \bar{\nu}_\mu . \tag{2}$$

The positively charged pion decays to a positively charged muon (μ^+) and to a muon-type neutrino (ν_μ), and the positively charged muon decays to a positron (e^+, the anti-particle of the negatively charged electron), electron-type neutrino (ν_e) and anti-muon-type neutrino ($\bar{\nu}_\mu$, the antiparticle of ν_μ). Charged pions may be produced also in collisions of high energy protons with other, low energy, nucleons (protons and neutrons). In this case, both positively

and negatively charged pions (π^+ and π^-) may be produced. The decay of the negatively charged pions produces neutrinos in a manner similar to that described by Eq. (2),

$$\pi^- \rightarrow \bar{\nu}_\mu + \mu^- \rightarrow \bar{\nu}_\mu + e^- + \bar{\nu}_e + \nu_\mu \ . \tag{3}$$

The neutrinos produced by the decay of the pions carry a significant fraction of the energy of the parent proton. Neutrinos produced, for example, by the decay of pions produced by interaction with photons, Eq. (1), typically carry 5% of the proton energy. UHECR sources are expected therefore to be sources also of high energy neutrinos. The detection of high energy neutrinos emitted by extragalactic sources will provide the first direct evidence for acceleration of protons in such sources, and may resolve the mystery of the UHECR source identity.

The Waxman–Bahcall Bound and Gigaton Neutrino Telescopes

UHECR observations provide guidance to estimating the expected high energy neutrino signal and the detector size required to detect it. Assuming that UHECRs are protons produced by extragalactic sources, the observed flux of UHECRs determines the average rate, per unit time and volume, at which such high energy protons are produced in the universe. The inferred rate at which energy is injected into the universe in the form of protons with energies in the range of 10^{19} eV to 10^{20} eV is [20]

$$\dot{\varepsilon} = 1.5 \times 10^{44}\,\mathrm{erg\ Mpc^{-3}\ yr^{-1}} \ . \tag{4}$$

The distance unit used here, megaparsec (Mpc), equals approximately to 3-million light years. Observations are consistent with the rate of energy injection being independent of the proton energy decade, i. e., the rate of energy production in protons in the energy range of, e. g., 10^{18} eV to 10^{19} eV is also given by Eq. (4).

Waxman and Bahcall have shown that the observed UHECR production rate sets an upper bound to the neutrino flux produced by extragalactic sources. High energy protons produced in candidate UHECR accelerators, such as AGN and GRBs, are likely to escape the source with only few pion production interactions. This implies that protons do not lose a large fraction of their energy to pion production. For sources of this type, the energy generation rate of neutrinos must be smaller than the proton energy generation rate, given by Eq. (4). Assuming that neutrinos are produced at this rate over the age of the universe, the resulting upper bound (for muon and anti-muon neutrinos, neglecting propagation flavor changes) is [21]

$$E_\nu^2 \Phi_\nu < 5 \times 10^{-8}\,\mathrm{GeV\ cm^{-2}\ s^{-1}\ sr^{-1}} \ . \tag{5}$$

Here Φ_ν is the number flux of neutrinos (number of neutrinos per unit area, time and solid angle) per unit neutrino energy, E_ν is the neutrino energy, and GeV stands for Giga-eV (1000 MeV). $E_\nu^2 \Phi_\nu$ describes the energy flux of neutrinos (energy per unit area, time and solid angle) carried by neutrinos with energies spread over (approximately) half a decade. The upper bound, which came to be known as the "Waxman–Bahcall" (WB) bound, is compared in Fig. 4 with current experimental limits, and with the expected sensitivity of planned neutrino telescopes.

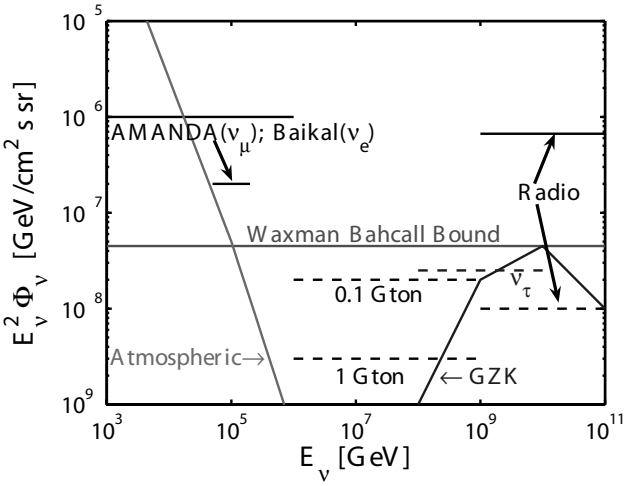

Fig. 4: The upper bound imposed by UHECR observations on the extragalactic high energy muon neutrino intensity, compared with the atmospheric neutrino background and with the experimental upper bounds (solid lines) of optical Čerenkov experiments, BAIKAL [4] and AMANDA [1, 10], and of coherent Čerenkov radio experiments (RICE [11], GLUE [7, 16]). The curve labelled "GZK" shows the intensity due to interaction with micro-wave background photons. Dashed curves show the expected sensitivity of 0.1 Gton (AMANDA, ANTARES, NESTOR) and 1 Gton (IceCube, NEMO) optical Čerenkov detectors [10], of the coherent radio Čerenkov (balloon) experiment ANITA [16] and of the Auger air-shower detector (sensitivity to ν_τ) [15]. Space air-shower detectors (OWL-AIRWATCH) may also achieve the sensitivity required to detect fluxes lower than the WB bound at energies $> 10^{18}$ eV [15].

The figure indicates that gigaton neutrino telescopes are needed to detect the expected extragalactic flux in the energy range of ~ 1 TeV to ~ 1000 TeV, and that much larger effective mass is required to detect the flux at higher energy. Few tens of ~ 100 TeV events per year are expected in a gigaton telescope if GRBs are the sources of ultra-high energy protons [18, 21]. These events will be correlated in time and direction with GRB photons, allowing for an essentially background free experiment. A lower rate is expected if AGN are the UHECR proton sources [2].

"GZK" Neutrinos

Protons of sufficiently high energy, exceeding $\sim 5 \times 10^{19}$ eV, may interact with photons of the cosmic microwave background, the big bang "relic" of 2.7 K radiation permeating the universe, to produce pions as described by Eq. (1) [8]. Protons of sufficiently high energy, $> 10^{20}$ eV, lose most of their energy over less than 300 million years, a time much shorter than the age of the universe, which is ~ 10 billion years. All the energy injected into the universe in the form of such protons is thus converted to pions which decay to neutrinos, producing a background neutrino intensity similar to the WB bound [5]. The expected neutrino intensity

is schematically shown in Fig. 4. It is denoted "GZK neutrinos", where "GZK" stands for Greisen, Zatsepin and Kuszmin, who were the first to point out the rapid energy loss of high energy protons due to interaction with the microwave background. The GZK neutrino flux peaks at $\sim 5 \times 10^{18}$ eV, since neutrinos produced by $p\gamma$ interactions typically carry 5% of the $\sim 10^{20}$ eV proton energy (The GZK intensity in Fig. 4 decreases at the highest energies since it was assumed that the maximum energy of protons produced by UHECR sources is 10^{21} eV).

The detection of GZK neutrinos will be a milestone in neutrino astronomy. Most important, neutrino detectors with sensitivity better than the WB bound at energies $> 10^{18}$ eV will test the hypothesis that the UHECR are protons (possibly somewhat heavier nuclei) of extragalactic origin. The large effective detector mass, much larger than gigaton, required to achieve this sensitivity may be obtained by detectors searching for radio, rather than optical, Čerenkov emission. "Air shower" detectors, which detect the "shower" of high-energy particles produced in the atmosphere following the interaction of a high-energy neutrino in the atmosphere, may also achieve sufficiently large effective mass at ultrahigh energy.

The challenge posed by the existence of UHECRs to models of particle acceleration, and the lack of direct evidence for proton acceleration in any extragalactic source, have led many to speculate that modifications of the basic laws of physics are required in order to account for the existence of UHECRs. Such "new physics" models commonly postulate the existence of very massive particles, the decay of which produces the observed UHECRs, and generally predict large fluxes of $\sim 10^{20}$ eV neutrinos [6], well above the WB bound. Measurements of the neutrino flux above $\sim 10^{19}$ eV would therefore allow to discriminate between "new physics" models for UHECR production and models where UHECRs are produced by "standard physics" acceleration in astrophysical objects, like GRBs and AGN.

Probing Astrophysical Accelerators with High-Energy Neutrino Telescopes

GRBs and AGN are believed to be powered by the accretion of mass onto black holes. GRBs are most likely powered by the accretion of a fraction of a Solar mass on second time scale onto a newly born Solar mass black hole. Recent observations strongly suggest that the formation of the black hole is associated with the collapse of the core of a very massive star. AGN are believed to be powered by accretion of mass onto massive, million to billion Solar mass, black holes residing at the centers of distant galaxies. As illustrated in Fig. 5, the gravitational energy released by the accretion of mass onto the black hole is assumed in both cases to drive a relativistic jet, which travels at nearly the speed of light and produces the observed radiation at a large distance away from the central black hole. The models describing the physics responsible for powering these objects, though successful in explaining most observations, are largely phenomenological. In particular, the answer to the question of whether or not the out flowing jet carries protons, which has major implications to our understanding of the mechanism by which gravitational energy is harnessed to power the jet, is not known despite many years of photon observations. This situation is common also to our understanding of Galactic micro-Quasars, which may be considered as a scaled down versions of AGN, with ~ 1 Solar mass black hole (or neutron star) "engines". Neutrino observations of GRBs, AGN and micro-Quasars will provide new information that can not be obtained using photon observations, and that may allow to answer the underlying open questions.

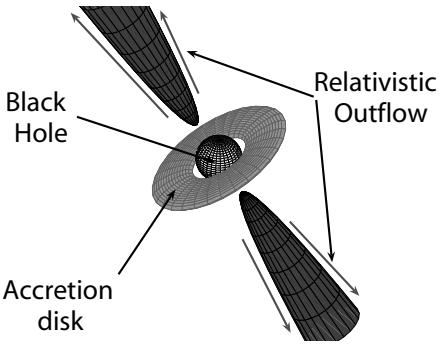

Fig. 5: GRBs and AGN are believed to be powered by black holes. The accretion of mass onto the black hole, through an accretion disk, releases large amounts of gravitational energy. If the black hole is rotating rapidly, another energy source becomes available: The rotational energy may be released by slowing the black hole down through interaction with the accretion disk. The energy released drives a jet-like relativistic outflow. The observed radiation is produced as part of the energy carried by the jets is converted, at large distance from the central black hole, to electromagnetic radiation.

Fundamental Neutrino Properties

Since neutrinos are expected to be produced in astrophysical sources via the decay of charged pions, production of high-energy muon and electron neutrinos with a 2:1 ratio, as described in Eqs. (2) and 3), is expected. Because of neutrino flavor changes during propagation, usually termed "neutrino oscillations", neutrinos that get to Earth are expected to be almost equally distributed between flavors. This implies that one should detect equal numbers of muon-type and tau-type neutrinos. Upgoing taus, rather than muons, would be a distinctive signature of such oscillations. Although the Čerenkov emission along the track of a high energy tau is similar to that produced by a muon, it may be possible to distinguish between taus and muons in a km^3 ice or water detector, since at 1000 TeV the tau decays after propagating ~ 1 km. This will allow a "tau appearance experiment". At present, the oscillation of muon-type neutrinos to tau-type neutrinos is inferred from the "disappearance" of muon-type neutrinos produced by cosmic-ray interactions in the atmosphere, without detecting the tau-type neutrinos that should be produced by such oscillation. The detection of taus in a neutrino telescope will provide direct confirmation of the oscillation hypothesis.

Detection of neutrinos from GRBs could be used to test the simultaneity of neutrino and photon arrival to an accuracy of ~ 1 s (~ 1 ms for short bursts), checking the assumption of special relativity that photons and neutrinos have the same limiting speed (The time delay due to the neutrino mass is negligible: for a neutrino of energy ~ 1 TeV with mass ~ 1 eV traveling 1000 Mpc, the delay is ~ 0.1 s.) These observations would also test the weak equivalence principle, according to which photons and neutrinos should suffer the same time delay as they pass through a gravitational potential. With 1 s accuracy, a burst at a distance of 1000 Mpc would reveal a fractional difference in limiting speed of 1 part in 10^{17}, and a fractional difference in gravitational time delay of order 1 in 10^6 (considering the Galactic potential alone). Previous applications of these ideas to supernova 1987A, where simultaneity could be checked only to

an accuracy of order several hours, yielded much weaker upper limits: of order 10^{-8} and 10^{-2} for fractional differences in the limiting speed and time delay, respectively.

Weakly Interacting Massive Particles

Most of the mass in the universe is currently believed to be in the form of "dark matter", composed of particles which were not detected in laboratories on Earth, and which interact with the normal matter that we know essentially only through gravitational forces. Weakly Interacting Massive Particles (WIMPs) are the leading dark matter particle candidates. If WIMPS populate the halo of our galaxy, the Sun or Earth would capture them, where they would annihilate occasionally into high-energy neutrinos. The annihilation rate depends on the details of the model. A widely discussed WIMP candidate is the lightest neutralino in minimal super-symmetric models. Gigaton neutrino telescopes currently under construction complement direct search detectors by reaching good sensitivity at high neutralino masses, typically in excess of a few hundred GeV, and by allowing us to probe regions of parameter space with large branching fractions to W and Z bosons.

References

[1] J. Ahrens *et al.*, *Phys. Rev. Lett.* **90**, 251101 (2003). The upper bound on high energy neutrino flux derived by the AMANDA experiment (A).

[2] A. Atoyan and C. D. Dermer, *Phys. Rev. Lett.* **87**, 221102 (2001); J. Alvarez-Muñiz and P. Meszaros, *Phys. Rev.* **D70**, 123001 (2004); J. K. Becker, P. L. Biermann and W. Rhode, `astro-ph/0502089`. Model predictions for neutrino emission from AGN (A).

[3] J. N. Bahcall, *Neutrino Astrophysics*. Cambridge University Press, New York, 1989. Textbook review of neutrino astrophysics, with focus on Solar neutrinos (I,A).

[4] V. Balkanov *et al.*, *Nucl. Phys. B (Proc. Suppl.)* **110**, 504 (2002). Description of the first operating optical Čerenkov detector of high energy neutrinos, the lake Baikal experiment (A).

[5] V. S. Berezinsky and G. T. Zatsepin, *Phys. Lett.* **28B**, 423 (1969). Prediction of the exitence of GZK neutrinos (A); R. Engel, D. Seckel and T. Stanev, *Phys. Rev.* **D64**, 093010 (2001). A calculation of the expected GZK neutrino flux (A).

[6] P. Bhattacharjee and G. Sigl, *Phys. Rep.* **327**, 109 (2000). Review of models of sources of ultra-high energy cosmic-rays, with detailed discussion of "new physics" models (A).

[7] P. W. Gorham *et al.*, *Phys. Rev. Lett.* **93**, 041101 (2004). Description of the upper bound on ultra-high energy neutrino flux derived by radio observations of the moon (A).

[8] K. Greisen, *Phys. Rev. Lett.* **16**, 748 (1966); G. T. Zatsepin, V. A. Kuzmin, *JETP* **4**, 78 (1966). The first papers to point out the rapid energy loss of high energy protons due to interaction with cosmic microwave background photons (A).

[9] A. M. Hillas, *ARA&A* **22**, 425 (1984). A review of the phenomenology of ultra high energy cosmic ray observations, and of the theoretical challenges facing models for particle acceleration (I).

[10] F. Halzen, *Proc. Nobel Symp. 129 "Neutrino Physics"* to appear in *Physica Scripta*, L. Bergstrom, O. Botner, P. Carlson, P. O. Hulth, and T. Ohlsson (eds.), `astro-ph/0501593`. Description of the current status of gigaton-scale optical Čerenkov neutrino detectors (I); F. Halzen and D. Hooper, *Rep. Prog. Phys.* **65**, 1025 (2002). Review of high energy neutrino astronomy (A).

[11] I. Kravchenko *et al.*, *Astropar. Phys.* **19**, 15 (2003). Description of the south pole coherent Čerenkov radio detector of high energy neutrinos (A).

[12] R. V. E. Lovelace, *Nature* **262**, 649 (1976). Phenomenological arguments suggesting that the brightest active galactic nuclei jets may accelerate protons to ultra high energy (I).

[13] M. Nagano and A. A. Watson, *Rev. Mod. Phys.* **72**, 689 (2000). Review of high energy cosmic-ray observations (I,A).

[14] G. Raffelt, *Stars as Laboratories for Fundamental Physics: The Astrophysics of Neutrinos, Axions, and Other Weakly Interacting Particles*. University of Chicago Press, Chicago, 1996. Textbook review of constraints imposed by stellar astrophysics on fundumental particle properties (I,A).

[15] D. Saltzberg, *Proc. Nobel Symp. 129 "Neutrino Physics"*, to appear in *Physica Scripta*, L. Bergstrom, O. Botner, P. Carlson, P. O. Hulth, and T. Ohlsson (eds.), `astro-ph/0501364`. Analysis of the sensitivity of large air-shower arrays as detectors of ultra-high energy neutrinos (A).

[16] A. Silvestri *et al.*, `astro-ph/0411007`. Description of the south-pole ballon radio experiment designed for detection of ultra-high energy neutrinos (A).

[17] E. Waxman, *Phys. Rev. Lett.* **75**, 386 (1995); M. Vietri, *Astrophys. J.* **453**, 883 (1995); M. Milgrom, and V. Usov, *Astrophys. J.* **449**, L37 (1995). First papers suggesting gamma-ray bursts to be ultrahigh energy cosmic-ray sources (A).

[18] E. Waxman and J. N. Bahcall, *Phys. Rev. Lett.* **78**, 2292 (1997). First calculations of high energy neutrino emission from gamma-ray bursts (A).

[19] E. Waxman, *Proc. Nobel Symp. 129 "Neutrino Physics"*, to appear in *Physica Scripta*, L. Bergstrom, O. Botner, P. Carlson, P. O. Hulth, and T. Ohlsson (eds.), `astro-ph/0502159`. Description of some major open questions of theoretical high energy astrophyscis, that may be addressed by high energy neutrino telescopes (I).

[20] E. Waxman, *Astrophys. J.* **452**, L1 (1995); J. N. Bahcall and E. Waxman, *Phys. Lett.* **B556**, 1 (2003). Derivation of the high energy cosmic-ray energy generation rate in the universe (A).

[21] E. Waxman and J. N. Bahcall, *Phys. Rev.* **D59**, 023002 (1999); J. N. Bahcall and E. Waxman, *Phys. Rev.* **D64**, 023002 (2001). Derivation of the upper bound on the extragalactic neutrino flux, implied by high-energy cosmic ray observations (A).

Astronomy, Optical

J. W. Fried

Scope

The purpose of astronomical observations is to provide information that can be used to explain the composition, structure, formation, dynamics and evolution of these objects by applying physical laws. The list of celestial objects includes the solar system, stars, star clusters, galaxies, clusters of galaxies, and the universe as a whole. Virtually all celestial objects can be studied by means of optical observations. In this article, optical astronomy includes observations using ultraviolet, visible, or near infrared radiation, since the technology used is nearly identical in these spectral regimes.

Telescopes

The purpose of a telescope is to collect the light from an object onto a detector. Obviously, the larger the telescope, the more light is collected and hence fainter and so more distant objects can be studied. All large telescopes use mirrors rather than lenses because mirrors are much easier to mount in such a way that they will not distort and degrade the images as the telescope

points in different directions. Telescopes are usually used with a small convex "secondary" mirror placed in the converging beam from the "primary" just before it comes to a focus. The beam is reflected by the secondary through a hole in the center of the primary and comes to a focus just behind the primary's mount. This is the "Cassegrain" focus and is a much more convenient place to mount instruments than the "prime" focus of the primary. The convex secondary magnifies the image, that is, it increases the effective focal length of the telescope by typically a factor of 3 or more. This design provides a long focal length in a compact size.

Telescopes are specified by the diameter of their primary. Many telescopes with diameters of 4–6 m existed already until the 1980's, and at the end of the last century, several new telescopes with diameters of 8–10 m have been built. Their mirrors are thinner and lighter than the older smaller ones, but since they are also more flexible their mounting requires actively moving supports to compensate changes in the gravitational force as the telescope moves. There are two ways to build large mirrors, a single monolithic design (used for example in the VLT of ESO) or a segmented mirror, assembled from many small segments. The 10-m mirrors of the Keck telescopes on Mauna Kea, Hawaii, for example, consist of 36 hexagonal segments. The segments have the surface shape appropriate to their location in the mirror which is approximately a parabola.

Instruments

Optical instruments range from simple cameras to complex computer-controlled multi purpose instruments. In the simplest case, a cameras just consists of a box to hold a 2-dimensional detector at the telescope's focus. Formerly this was generally a photographic plate, but since the 1980's the detector is almost exclusively a CCD (see below). Several of these detectors can be mounted side by side in the focal plane to cover a large field of view. Filters placed in front of the detector allow only one color band to pass. Images taken in several filters already give information on the physical state of the objects detected. For example, a hot star will appear brighter in a blue pass band than in a red pass band, i. e., measuring the brightness difference between blue and red gives the temperature of the star. An image of the sky is the most basic observation to be done, yielding position, brightness and form of the objects.

More information about the objects can be derived if several pass bands are used. Modern multi color surveys, carried out on telescopes of the 2–4 m class, use up to ~ 20 filters; from these data spectral types and red-shifts of the objects detected can be derived, albeit with moderate precision. These surveys detect objects which are up to 100 times fainter than those found in the famous "Palomar Sky Survey" which used the 48-in. Schmidt telescope on Mt. Palomar to photograph the entire northern sky on 935 pairs of 14-in. square plates (6×6 degrees each) in the blue and red pass bands.

In order to get more detailed information about an object, its spectrum – that is its intensity versus wavelength – has to be measured at much higher wavelength resolution. A diffraction-grating spectrograph indexspectrograph is normally used for these measurements. With such a spectrum from a star one can measure the star's temperature, composition, and surface gravity. From this information and a liberal amount of theory one can deduce the entire structure of the star right down to its center. Another important use is as a speedometer. An object's relative speed toward or away from the observer can be found by measuring the Doppler shift of its entire spectrum; the amount of shift is proportional to the speed. Among other things this is how the cosmological red-shift is measured.

There are many other, more special purpose instruments such as Fourier-transform spectrometer, Fabry–Perot interferometers, spectroheliographs, occultation photometers, speckle interferometers, etc.

Instruments in the infrared are very similar to those used at visible wavelengths. Observations from the ground are limited to a few spectral regions, since the Earth's atmosphere is not transparent outside these regions. At ultraviolet wavelengths, the Earth's atmosphere is opaque, too. Observations in this regime must be done from space.

Detectors

Photographic plates were the first recording detectors. The spatial resolution of a plate is 10–20 μm and so a 4×5-in. plate contains over 10 million independent measurements. Their main problem is that they are not very sensitive to light (low efficiency), and it is very difficult to obtain accurate intensities from them. The other main class of detectors are the photoelectric devices that convert incident light to an electrical signal. The photomultiplier tube is an ultraviolet and visible light detector. It uses the 'outer' photoelectric effect to convert an incident photon into an electron which is then amplified a million-fold within the tube.

While the photomultiplier makes only one measurement at a time, charged coupled devices (CCDs) contain up to 16 million individual detectors (pixel) and so can be used to take images. Making use of the 'inner' photoelectrical effect, an optical image is converted into an electrical one which is transmitted to a computer. CCDs are linear, record up to 90% of the incoming light, and have very little noise. Therefore, they are practically ideal detectors and are nowadays used almost exclusively in optical astronomy.

Infrared detectors are generally also semiconductor devices. These detectors can have over 80% efficiency but they are fairly noisy in most cases unless they are cooled down to under 4 K. The advent of these detectors opened up the field of infrared astronomy to the point where it is now a major branch of observational astronomy. The reason for this is that dust between the object under study and the observer blocks light at optical wavelengths, but much less at infrared wavelengths: infrared observations penetrate dust. One active field of infrared astronomy is the study of formation of stars which is usually going on at dusty places.

Observatories

There are many small and large observatories throughout the world. They are owned or operated by colleges, universities, research organizations and nations. Major observatories are located on Kitt Peak near Tucson, Arizona, Mauna Kea on Hawaii, La Silla, Cerro Tololo and Cerro Paranal in Chile. Recent site testing campaigns have shown that at least one location in Antarctica has outstanding characteristics and is possibly inferior only to space. Observatories in space are not hampered by the Earth's atmosphere and so are not restricted to certain atmospheric windows; the ultraviolet region is observable from space (or at least high flying balloons or rockets) only. Space observatories such as the Hubble Space Telescope also do not suffer from 'seeing', i. e., image smearing due to atmospheric turbulences, which degrades the resolution of ground-based telescopes. Seeing effects can be reduced in ground-based observations if a deformable mirror is placed in the light path and deformed in such a way that atmospheric distortions of the incoming wave front are compensated ('adaptive optics').

See also: Cosmology; Galaxies; Milky Way; Solar System; Sun.

Bibliography

M. A. Seeds, *Horizons: Exploring the Universe*. Belmont, Calif., Wadsworth, 1987. (E)
J. H. Robinson, *Astronomy Data Book*. Wiley, New York, 1972.
Astronomy (Astromedia Corp., Milwaukee, 1972). (E)
B. V. Barlow, *The Astronomical Telescope*. Wykeham Publ, Ltd., 1975. (I)
C. R. Kitchin, *Astrophysical Techniques*. Adam Hilger Ltd., 1984. (A)
W. H. Steel, *Interferometry*, 2nd ed. Cambridge University Press, Cambridge, 1983. (A)
R. N. Wilson, *Reflecting Telescope Optics I*. Springer, Berlin, Heidelberg, New York, 1996. (A)
R. N. Wilson, *Reflecting Telescope Optics II*. Springer, Berlin, Heidelberg, New York, 1996. (A)
P. L'ena, F. Lebrun, F. Mignard, *Observational Astrophysics*. Springer, Berlin, Heidelberg, New York, 1998. (A)
J. N. Hardy, *Adaptive Optics for Astronomical Telescopes*. Oxford University Press, 1998. (A)

Astronomy, Radio

A. A. Penzias and B. E. Turner

Radio astronomy differs from the other observational branches of astronomy in that the incident energy is coherently amplified. Unlike film, bolometers, or particle counters, the phase of the incident wave is preserved in the process. The terminals of the receiver are located at the focal point of the *antenna*, where the field in the aperture is coherently added. The most generally used antenna and the one which most nearly resembles its optical counterpart is the parabolic reflector (Fig. 1). The angular distribution of the antenna response, the "antenna pattern," has an angular width at half-intensity of $\sim \lambda/a$ radians, where a is the width of the antenna aperture.

Mechanical limitations on antenna size and precision limit the resolution (minimum beam width) of the largest single antennas to about 10 arc seconds. To obtain higher resolution, an array of antennas is connected together to form an *interferometer* (Fig. 2).

When the signals obtained by a pair of antennas are multiplied together, the product is proportional to the cosine of the difference between the phases of the received signals. The angular dependence of the response of such a system to a distant source of monochromatic radiation will have a one-dimensional periodicity in the direction which is coplanar with the line of sight between the two antennas. The angular frequency of the response is given by the projected distance, in wavelengths, between the antennas as viewed from the source.

By observing the source with a number of differently spaced antenna pairs one can obtain enough angular frequency components to construct a Fourier synthesis of the angular distribution of the source intensity. These different spacings can be obtained by use of a number of antennas with different spacing, by moving one or more of the antennas between observations, and by making use of the earth's rotation to change the projected spacing. Two-dimensional information can be obtained by arranging the antenna array in the form of a cross or "Y"

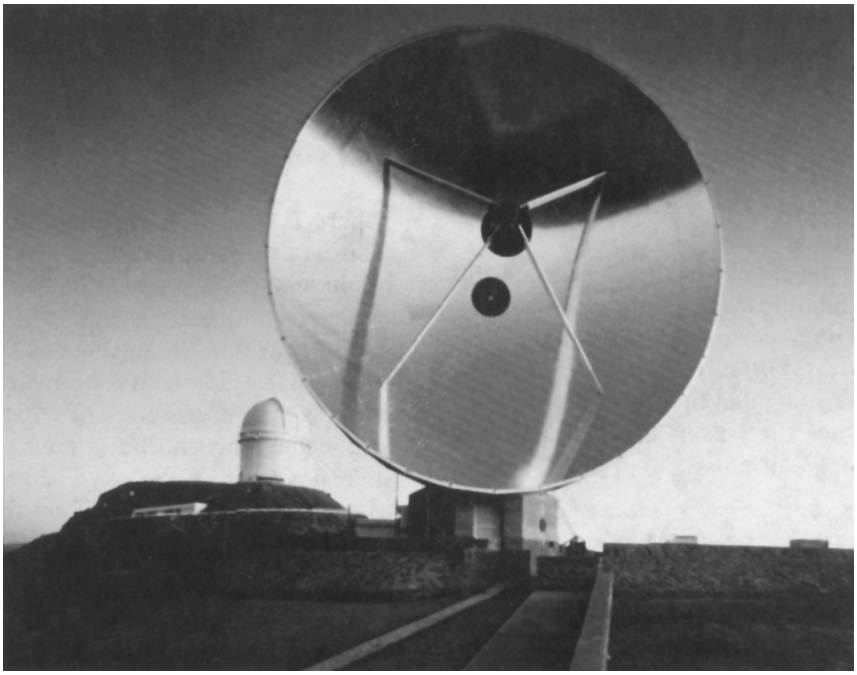

Fig. 1: The 15-m submillimeter telescope of the Swedish–European Southern Observatory consortium, at an elevation of 2350 m at La Silla, Chile. The surface accuracy (deviation from a perfect paraboloid) is 35 μm rms, allowing operation to wavelengths as short as 550 μm (Onsala Space Observatory photo).

(Fig. 2) or making use of the change in orientation of the array axis with respect to the source caused by the rotation of the earth. The resulting intensity map of the source extending over the entire area covered by the relatively broad beams of the individual antennas can have the resolution equivalent to that of a single antenna whose aperture encompasses the entire array; hence the name, *aperture synthesis*.

Radiometry

The intensity scale used in radio astronomy is antenna temperature. Consider a warm uniform opaque cloud extending over an angular area much larger than the antenna beam. The power radiated per unit area by this cloud is given by the usual blackbody formula

$$B = \frac{4\pi h\nu^3}{c^2(e^{h\nu/kT} - 1)}\Delta\nu \ ,$$

and, in the limit $h\nu/kT \ll 1$, the power radiated per unit solid angle becomes the familiar Rayleigh–Jeans formula

$$b = \frac{2kT}{\lambda^2}\Delta\nu \ .$$

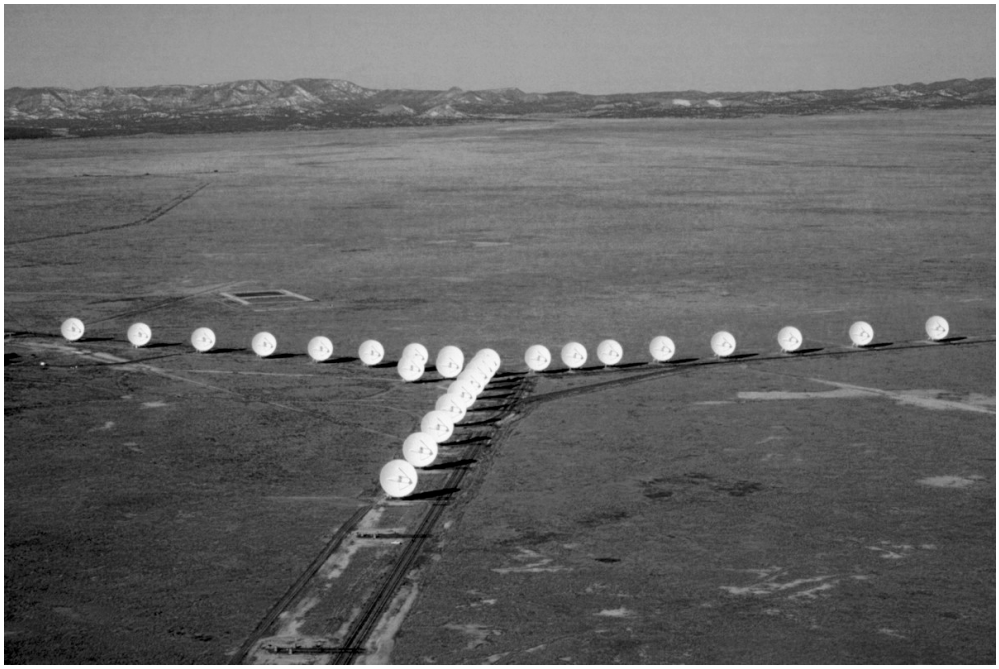

Fig. 2: The Very Large Array, consisting of 27 antennas, each 25 m in diameter, in a Y configuration. Shown is the most compact configuration (arm length 0.6 km). At its shortest wavelength, 1.3 cm, and largest configuration (arm length 21 km), the VLA provides a spatial resolution of 0.08 arcsec (NRAO photo).

At some distance R from the cloud we put an antenna with an aperture of width a pointed toward it. The antenna will receive energy only from that portion of the cloud intercepted by its beam, i. e., $\sim (\lambda/a)^2 R^2$. Furthermore, we must multiply by the solid angle subtended by the antenna at the cloud surface, $\sim (a/R)^2$. Thus the power incident upon the antenna terminals in one plane of polarization is

$$\frac{kT\Delta\nu}{\lambda^2} \times \frac{\lambda^2 R^2}{a^2} \times \frac{a^2}{R^2} = kT\Delta\nu \; .$$

This relation leads to the definition of antenna temperature as the power per unit bandwidth received at the antenna terminal divided by Boltzmann's constant. Note, therefore, that antenna temperature is an intensity which only corresponds to a thermodynamic temperature in the Rayleigh–Jeans limit.

Radio astronomy observations are simply the measurements of antenna temperature as a function of angle and frequency. To relate the observed quantity to the property of the emitting medium we introduce the concept of equivalent brightness temperature of an astronomical object. This is the physical temperature of a perfect absorber with the same angular dimensions as the object which would produce the observed antenna temperature. To obtain the equivalent brightness temperature from the antenna temperature we must take the response pattern of the

antenna into account. If the object is small compared to the antenna beam, we must determine its angular size by other means. Failing that, we can only assign a flux (incident power density per unit bandwidth per unit frequency interval) to the object by dividing the observed antenna temperature by Boltzmann's constant and the effective area of the antenna, and multiplying the result by 2. (The multiplication by 2 reflects the fact that a coherent receiver responds to only one polarization of the incident field.)

Antenna temperature is measured by means of a radiometer, which in its simplest form consists of an amplifier, frequency-selective filter, rectifier–detector, and an integration and voltage-measurement circuit. Incident power levels from astronomical sources upon even the largest antennas are extremely small and must be measured in the presence of the generally much larger noise power of the radiometer itself.

Only the portion of this power contained within the particular frequency interval selected is detected; it is thus useful to characterize the radiometer's sensitivity by a quantity proportional to its noise power per unit frequency interval, its *equivalent noise temperature*, T_N, defined as the temperature of a perfect absorber placed at the input which delivers the same output noise power for an equivalent noiseless amplifier.

The fluctuation in output due to receiver noise is expressed in terms of the equivalent change in noise power at its input in units of antenna temperature by the radiometer formula

$$\Delta T_{RMS} = T_N/\sqrt{B\tau}\,,$$

where B is the bandwidth and τ is the postdetection integration time. (A heuristic understanding of the origin of this relation may be obtained by thinking of $B\tau$ as the number of independent measurements one may make of the noise temperature within the integration time.)

Practical radiometry normally employs comparisons in making measurements. The most common such comparisons are made between: the antenna and a reference termination; two positions in the sky; or two different radiometer frequencies. A common method of periodic comparison used is called synchronous detection. The output of the receiver is inverted in synchronism with the switching of the input from the signal to reference condition. Thus, over a number of cycles the integration accumulates a voltage proportional to the difference between the two inputs. The great virtue of this arrangement can be illustrated by considering the effect of a small change in gain ΔG. In the unswitched case this would cause a change in output of ΔG times the total system temperature, whereas in the switched case the fluctuation in output is ΔG times the difference of two temperatures, which can be made as small as desired by proper selection of the reference temperature. The penalty paid is a decrease in the signal-to-noise ratio by approximately a factor of 2. An array of such radiometers, usually sharing a common amplifier and with their filters spaced adjacently in frequency, is the most generally used system for line studies, the multichannel line receiver. An alternative to the multichannel radiometer is the autocorrelation receiver, which uses the fact that the autocorrelation function of the time variation of the input signal is the Fourier transform of the power spectrum in frequency.

It is useful to compare the sensitivity of radiometers with that of incoherent devices as detectors of astronomical radiation. The radiometer contains active elements which amplify both the incident power as well as its own noise to levels much higher than the noise associated with its detector. Thus, it is the noise in the amplifier which limits the sensitivity of the device.

Since this noise is proportional to bandwidth, whereas that of the incoherent detector is not, a meaningful comparison of the two types of devices can only be made when the bandwidth of the observation is specified. We may relate the noise equivalent power, NEP, of an incoherent detector to the minimum detectable increment in antenna temperature:

$$\Delta T_{\mathrm{RMS}} = \mathrm{NEP}/kB\sqrt{\tau}\,.$$

Astrophysical Sources of Radio Emission

Free–free emission, often called "thermal emission" for historical reasons, arises from the interaction (acceleration) between unbound charged particles in an ionized gas (HII region). It has a characteristic spectrum whose brightness temperature decreases roughly as the inverse square of the frequency above a certain "turnover" frequency, below which it is equal to the temperature of the ionized gas. The turnover frequency marks where the gas becomes opaque to radiation and ranges from 100 MHz for large, diffuse HII regions to as high as 30 GHz for small, dense ones. Several hundred HII regions are known in the Milky Way; they serve as signposts for massive hot stars since the gas is ionized by uv photons at wavelengths shorter than 912 Å from the central stars. The balance between uv heating rate and cooling by emission from trace elements (O, N) endows most HII regions with a temperature of 10 000 K. Superimposed upon the broadband spectrum of the ionized gas are *recombination lines* caused by electronic transitions in the constituent atoms (mostly H, He) as given by the Rydberg formula ($n \sim 50$ to 500). The line-to-continuum intensity ratio is a sensitive indicator of physical conditions in HII regions.

Synchrotron emission is generated by high-energy electrons moving in a magnetic field. It is called "nonthermal emission" because its intensity is related not to the temperature of the emitter but to the strength of the field and the number and energy distribution of the electrons. It produces much of the radiation from supernova remnants, radio galaxies, and quasars. Radio galaxies and quasars comprise extragalactic radio astronomy. Their emission falls into two categories: the extended structure (which is transparent) and the compact structure (which is opaque to its own radiation). The extended emission is typically associated with galaxies, but in many cases with quasars with no visible optical extent. Most compact sources are identified with quasars or with active galactic nuclei. In less powerful radio galaxies, the radio emission is often confined to the region of optical emission (about 30 000 light-years in size), but in more powerful radio galaxies the emission comes from two well-separated regions hundreds of thousands of light-years across. Figure 3 shows a radio image of the powerful radio galaxy 3C175, made with the VLA at 6 cm wavelength at a spatial resolution of 0.35 arcsec. The unresolved bright spot near the center is a compact source coincident with a quasar at a redshift of 0.77. The long, narrow, one-sided radio jet is typical of powerful double-lobed (extended) sources. There are prominent hot spots in both lobes suggesting they have both been recently supplied with relativistic particles despite the appearance of only a single jet.

Our own galaxy radiates largely by synchrotron emission. An all-sky map of the galaxy's emission at 408 MHz is shown in Fig. 4; it was made at a single resolution of 0.85 degrees using three of the world's largest parabolic reflectors. The galactic center serves as the center of symmetry for the entire sky, with the low galactic latitude intensity dropping away steeply on each side out to longitudes ~ 60 and 280 degrees. Large-scale prominences, rising from low galactic latitudes and extending nearly to the poles in both hemispheres, distort the symmetry

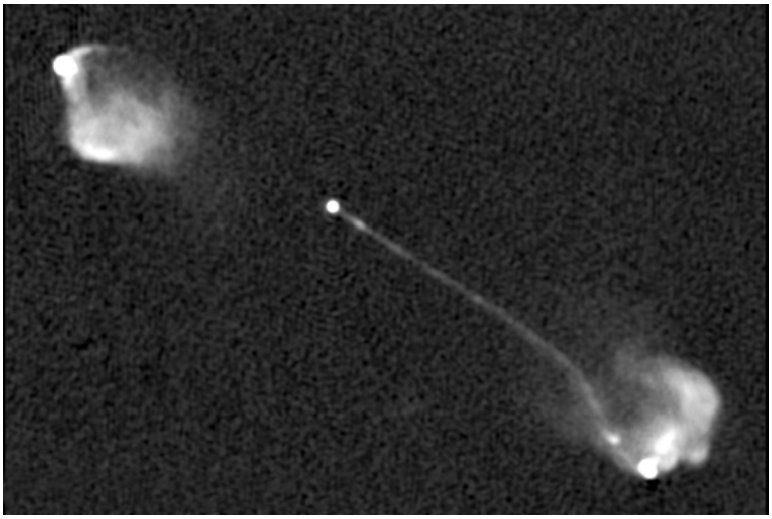

Fig. 3: The radio galaxy 3C175, observed with the VLA at a resolution of 0.35 arcsec at 6 cm wavelength (photo courtesy A. Bridle).

of the high-latitude emission. These loops and spurs trace large corresponding features in the magnetic field of the galaxy.

Line emission provides a powerful method for the study of neutral (and some ionized) interstellar matter. Hydrogen, the most abundant element, is studied in our own galaxy, in external galaxies, and on cosmological scales by means of its ground-state magnetic hyperfine transition at 1420 MHz (21 cm wavelength). Together with optical redshift data, HI data has established the existence of large-scale "voids" over cosmological distances, regions that are underpopulated at least in matter concentrated in normal galaxies. Current cosmological HI research is aimed at determining whether significant amounts of baryonic matter, capable of contributing importantly to the mass density distribution and therefore to the question whether the universe is closed or open, may reside in noncollapsed intergalactic clouds of HI.

HI studies of clusters of galaxies have shown gas deficiencies in cluster spiral galaxies – evidence that gas has been swept from at least their outer regions by passage through the center of the cluster. Several mechanisms may be at work (galaxy collisions, tidal interactions, ram pressure sweeping by the intracluster medium, evaporation) and may remove up to 90% of the initial HI mass. Reduced star formation rates are evident.

HI studies of noncluster galaxies, especially nearby ones for which good spatial resolution exists, serve to relate the structure of the interstellar medium with that of the stellar populations. Clearly defined HI disks correspond to optical disks but extend much further out, allowing galaxy mass determinations. Spiral arms seen in HI often reveal important differences from their optical counterparts that relate to star formation mechanisms. Strong warps are often seen in the outer regions of HI disks, which in many cases appear to be self-maintaining rather than resulting from perturbations such as tidal interactions. HI studies of velocity distributions in galaxies show in the case of spirals that the dynamical masses greatly exceed the total mass (by factors of 2 to 3) seen within the optical radius. HI studies also show that low-luminosity galaxies contain the same large mass-to-light ratio as luminous galaxies do,

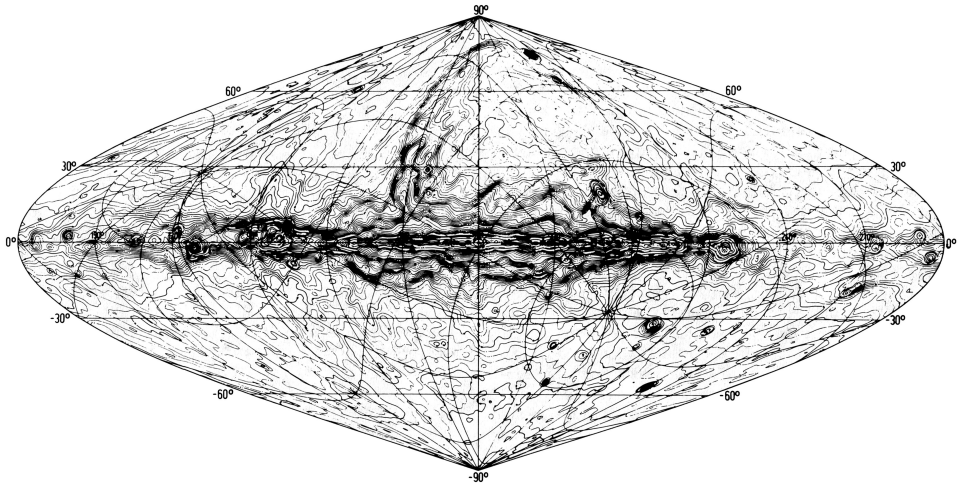

Fig. 4: An all-sky map of the Milky Way at 408 MHz (photo courtesy G. Haslam).

and that some types also contain dark halos. Since low-luminosity galaxies are by far the most numerous, these HI results are important to the question of the "missing mass" needed to close the universe.

Line emission is also observed from a large number of interstellar molecules, which reside in the denser regions of the interstellar medium. Together with studies of HI in the Milky Way, these lines have delineated the spiral structure and have established the existence of several thousand giant, massive molecular clouds (10^5 solar masses or greater) as well as many more smaller ones. These "GMCs" are the most massive entities in our galaxy and have been established as the sites of massive star formation. In other galaxies as well as our own, the GMCs are nor, known to trace the spiral arms. Molecular CO has now been observed at redshifts up to 0.16, and promises to become a cosmological tool as important as HI.

Around 130 interstellar molecular species have now been identified in dense molecular clouds, forming the current dicipline of astrochemistry. The interstellar chemistry is carbon rich, as on earth, but produces many exotic species not found on earth, as well as many familiar in the laboratory. These molecules play a major role in how stars form from the interstellar gas. The identification rate of new species is still about 2 per year. Recent discoveries have focused on larger organic molecules such as vinyl alcohol, propenal and propanal, the latter two being "sugars". Glycine, the smallest biologically important amino acid, has possibly been identified, but is highly controversial. Equally intriguing is the discovery of highly deuterated species such as doubly deuterated ammonia NHD_2, and formaldehyde (D_2CO), and most important triply deuterated protonated hydrogen (H_3^+), which does not occur on Earth but along with H_3^+ is the most basic progenitor of astrochemistry. At a cold 10 Kelvin deep inside thick molecular clouds, H_3^+ and D_3^+ are the only species not frozen out onto icy grains.

Blackbody radiation is characteristic of solar system objects. At long radio wavelengths the brightness temperature of the sun is very large, $\gtrsim 10^6$ K, because the observed emission

comes from the corona and material ejected during solar flares. This ionized matter becomes essentially transparent at shorter wavelengths, and the millimeter wavelength brightness temperature of the sun is essentially that of the photosphere, $\sim 6000\,\mathrm{K}$. Conversely, in the case of a *planet*, the long wavelengths are able to penetrate un-ionized atmosphere better than shorter wavelengths. Thus the long-wavelength brightness temperature is that of the surface, whereas at shorter wavelengths the brightness temperature corresponds more closely to that of the cooler atmosphere.

Blackbody radiation fills the entire sky and is the remnant of the big-bang fireball expanded and cooled to 3 K at the present epoch. It is unique in that it has the same brightness temperature, 3 K, at all wavelengths. The radio spectrum of "empty" sky is a superposition of galactic radiation, unresolved distant radio sources, and the cosmic background radiation.

See also: Astrophysics; Blackbody Radiation; Galaxies; Interferometers and Interferometry; Photosphere; Radiometry; Sun; Synchrotron Radiation; Transmission Lines and Antennas

Bibliography

J. D. Kraus, *Radio Astronomy*. Cygnus-Quasar Books, 1986. An intermediate level book emphasizing technical aspects, with advanced references.

G. L. Verschuur and K. I. Kellermann (eds.), *Galactic and Extragalactic Radio Astronomy*. Springer, New York, 1988. A graduate-level text covering 15 major areas of radio astrophysical research.

Astronomy, X-Ray

P. Gorenstein and W. Tucker

Introduction

X-ray astronomy involves photons of cosmic origin in the energy band 0.2 to 30 keV. The low-energy limit is determined by the opacity of the interstellar medium, the higher limit by the falling spectra of sources. Observations take place above the absorption of the Earth's atmosphere. X-ray production in a cosmic setting is associated with thermal radiation from plasmas with temperatures from 10^6 to 10^8 K, synchrotron radiation from highly energetic electrons in a magnetic field and inverse Compton scattering where electrons elevate lower energy photons to the x-ray band. Solar x-ray emission was detected in 1948 and a rocket flight in 1962 made the first positive detection of sources outside the solar system. A series of spinning satellites, initially bearing proportional counters beginning with UHURU in 1970 and continuing through ROSAT in 1990's with an imaging telescope, surveyed the sky and cataloged about 80 000 sources. Even more have been found serendipitously in the fields of pointed observations. Their celestial distribution shows two components: one distributed along the plane of the galaxy and another having the isotropy characteristic of extragalactic objects. In addition, aside from an irregular soft component associated with the local interstellar medium, there is an intense isotropic background which now appears to be fully explained as the sum of distant extragalactic point-like sources.

X-ray astronomy was transformed by the findings of two facility-class observatories with focusing telescopes launched in 1999 plus results from the large area counters of the Rossi X-Ray Timing Explorer (1995). NASA's Chandra X-Ray Observatory, with the ability to make images with sub-arc-second resolution is well complemented by the lower-resolution, but higher throughput European Space Agency's observatory, XMM-Newton. Both observatories are equipped with dispersive spectrometers that allow high- resolution spectra ($E/\delta E \sim 300$–1000). Chandra and XMM have studied planetary atmospheres, normal stars of every type, brown dwarfs, white dwarfs, neutron stars and black holes, the remnants of exploded stars, galaxies of every size, shape, and stage of evolution, galaxy clusters with thousands of galaxies embedded in vast clouds of hot gas and dark matter millions of light years across, and even larger streamers of hot gas between the clusters. These studies plus the fast temporal intensity variations measured by the Rossi X-Ray Timing Explorer have led to advances on such fundamental questions as the geometry of spacetime around a black hole, the ability of accreting black holes to produce highly collimated, relativistic jets a million light years in length, the distribution and properties of dark matter, and the existence of dark energy. Their findings have placed x-ray astronomy in realm of fundamental physics where several basic principles can be tested in domains that are not accessible on Earth.

X Rays from Planets and Comets

While the temperature of planets, satellites, asteroids and comets, is typically well below 1000 K they produce x rays through mechanisms that involve the Sun directly or indirectly. The x-ray power is weak, ranging from a few megawatts in the Martian atmosphere to a few gigawatts for Jupiter but provides information about chemical composition, mass, and radiation belts that are difficult to obtain with measurements at other wavelengths.

Charge exchange, primarily between heavy neutral atoms such as oxygen in the atmosphere of a comet or a planet, and fast ions in the solar wind, operates throughout the solar system. It is especially important for comets, which have extended atmospheres. By observing x rays from comets, it is possible to study the elements present in the solar wind, the structure of the comet's atmosphere, and cometary rotation.

Planetary atmospheres and in the absence of an atmosphere, the bare surface, emit fluorescence x rays when illuminated by solar x radiation. Fluorescent radiation from oxygen and other atoms has been detected in the Venusian atmosphere between 120 and 140 kilometers above the surface of the planet. In contrast, the optical light from Venus is caused by the reflection of sunlight from clouds at altitudes 50 to 70 kilometers.

Fluorescent x rays from oxygen atoms in the Martian atmosphere probe similar heights. The lack of variation of Martian x radiation during a dust storm on the surface shows that the storm did not reach into the upper atmosphere. The detection of a faint halo of x rays, presumably due to the solar wind charge–exchange process operating in the tenuous extreme upper atmosphere of Mars some 7 000 kilometers above the surface indicates that Mars is still losing its atmosphere into deep space.

Jupiter produces x rays as a result of its substantial magnetic field. X rays are emitted when high-energy particles from the Sun get trapped in its magnetic field and are accelerated toward the polar regions where they collide with atoms in Jupiter's atmosphere. Chandra's image of Jupiter shows strong concentrations of x rays near the north and south magnetic poles. Weaker

x-ray signals have been detected from two of Jupiter's moons, Io and Europa, and from the Io Plasma Torus, a doughnut-shaped ring of energetic particles that circles Jupiter. Gases, such as sulfur dioxide, are produced by Io's volcanoes, escape from Io and become trapped in an orbit around Jupiter, where they are accelerated to high energies. Collisions between the particles within the torus, and with the surfaces of Io and Europa can account for the observed x rays.

Chandra's observation of Saturn, which has a weaker magnetic field than Jupiter, revealed an increased x-ray brightness in the equatorial region. Furthermore, Saturn's x-ray spectrum, or the distribution of its x rays according to energy, was consistent with scattered solar x rays. The same process may be responsible for the weak equatorial x radiation observed from Jupiter.

Galactic X-Ray Sources

The bright galactic sources are associated with the final phases of stellar evolution. They are identified with remnants of supernova explosions or binary star systems containing a compact object such as a neutron star or a black hole. They are among the most luminous objects in the galaxy, radiating 10^{36} to 10^{38} erg/s in the x-ray band. In comparison, the Sun, an average star, radiates 4×10^{33} erg/s, principally at optical wavelengths. A much lower level of x-ray emission, 10^{28}–10^{32} erg/s, has been detected from many relatively nearby stars. The x-ray emission from normal stars is produced by processes that are analogous to solar coronal and flare activity, or shock waves in the winds of massive stars or colliding stellar winds of binary stars. Binary systems containing white dwarf stars, and isolated neutron stars with surface temperatures of the order of a million degrees have also been detected at these x-ray luminosities.

Young Star Clusters

About a thousand faint x-ray-emitting stars have been detected in the Orion star cluster by the Chandra X-Ray Observatory. Since flaring activity is more pronounced at x rays than at optical wavelengths, long-term monitoring of this sample of stars, all at the same distance, and approximately the same age is indicative of the rate and magnitude of flaring in young stars. This information on the high-energy activity of young stars is relevant to understanding the environment in which the planets were formed and evolved.

In a related discovery, non-thermal x rays have been detected from a cloud of high- energy electrons enveloping the young star cluster RCW 38. These extremely high-energy particles could cause dramatic changes in the chemistry of the disks that will eventually form planets around stars in the cluster.

Compact X-Ray Binaries

All of the bright luminous galactic x-ray sources that have not been identified with supernova remnants fall into a class designated as "compact x-ray binaries." The essential features of these sources are (i) a 1–10-keV luminosity in the range 10^{36}–10^{38} erg/s; (ii) membership in binary systems, as evidenced by eclipses or periodic variations on a time scale of days in the radiation of the compact object or its companion star; (iii) a spectrum similar to that produced

by radiation from a hot gas having a temperature in the range 50–500 million degrees; (iv) fast, and in some cases quasi-periodic oscillations on a time scale of seconds or less.

The model consists of matter lost from the primary star in a close binary system accreting onto a neutron star or black hole companion. Gravitational potential energy heats a gas which radiates the input energy as x rays. If the secondary has a mass M_x and a radius R, the gravitational energy released per gram would be on the order of GM_x/R. For a mass accretion rate $\dot{m}$, the energy released per second is

$$L \sim \frac{GM_x\dot{m}}{R} \sim 10^{41}\left(\frac{M_x}{M_\odot}\right)\left(\frac{R_\odot}{R}\right)\dot{m} \quad \text{erg/s} \tag{1}$$

where $M_\odot$ and $R_\odot$ are the mass and radius of the sun and $\dot{m}$ is the accretion rate in units of solar masses per year. For both neutron stars and black holes $R_\odot/R \sim 10^5$, we have, therefore,

$$L \sim 10^{46}\dot{m} \quad \text{erg/s} \tag{2}$$

Equation (2) shows that mass accretion rates in the range 10^{-10}–10^{-8} solar masses per year can produce x-ray luminosities in the range 10^{36}–$10^{38}\,\text{erg/s}$.

For comparable accretion rates, the expected luminosity for white dwarf companion stars ($R/R_\odot \sim 10^{-2}$) is three orders of magnitude smaller. Accreting white dwarf binaries are perhaps the most common type of binary x-ray source in the galaxy.

Compact binary systems also differ with respect to the nature of the noncompact primary star. In young binary systems the primary is a giant blue star that has a mass more than 10 times that of the sun and whose age is less than 10 million years. They are found in regions of active star formation, such as the galactic spiral arms. In old or low-mass x-ray binaries, the noncompact star is often less massive than the sun. These systems have existed for at least a hundred million years and show no preference for spiral arms.

Young neutron stars have intense magnetic fields which modify the accretion flow and produce an asymmetric radiation pattern that, when coupled with the rotation of the neutron star, appears to a distant observer as a series of pulses with eclipses and Doppler variations from the binary motion. For old x-ray binaries, two physical effects change the nature of their radiation. First, the accreting plasma transfers angular momentum to the neutron star, causing it to spin faster over time. Second, the magnetic field of the neutron star has weakened to the point where it can no longer effectively channel the accreting plasma. Consequently, these sources are characterized, not by stable periodic pulses on a time scale of seconds, but by quasi-periodic oscillations (QPO's) down to a scale of milliseconds. QPO's were discovered in neutron star binaries in 1996 by investigators analyzing data from the large area counters of the Rossi X-Ray Timing Explorer. Relativistic dragging of inertial frames, known as the Lense–Thirring effect, around fast rotating collapsed stars, has been proposed by some theoreticians as a mechanism responsible for QPO's.

Similar behavior is observed from an accreting black hole. Strong magnetic fields cannot exist in the vicinity of black holes and the period of the last stable orbit around a black hole having the mass of a few solar masses is on the order of milliseconds. The similarity in the radiation patterns of accreting black holes and old neutron stars has made it impossible to identify black holes conclusively on the basis of x-ray data alone. X-ray novae are a clue. Some of these sources exhibit a faint x-ray luminosity in quiescence, whereas others are undetectable. The absence of x radiation in quiescence has been interpreted as evidence for an

event horizon in these sources, since the presence of a solid surface would lead to low-level x-ray emission, or x-ray bursts due to thermonuclear burning of the accreted matter.

In general, by far the strongest evidence for a black hole consists of optical observations of the primary star indicating that the primary has an invisible companion with a mass greater than $3M_{\odot}$ the theoretical upper limit for the mass of a neutron star. To date, about 20 such systems have been discovered, with estimated black hole masses ranging from 4 to 16 solar masses.

With most physicists and astrophysicists in agreement that black holes do indeed exist, intense theoretical and observational efforts are underway to understand the detailed properties of these systems. The K-α fluorescent line of iron provides an especially useful probe of the region within a few gravitational radii of the event horizon of a black hole. The gravitational, or Schwarzschild radius is defined as

$$R_{\mathrm{S}} = \frac{2GM}{c^2} \tag{3}$$

Detailed x-ray spectroscopy of broadened iron line features has been used to study Doppler and gravitational redshifts, thereby providing key information on the location and kinematics of the cold material. Observations of both stellar mass black holes and supermassive black holes have provided intriguing and impressive evidence for the gravitational red-shift and the effects of black hole spin on the spacetime around black holes.

The orbit of a particle near a black hole depends on the curvature of space around the black hole, which also depends on how fast the black hole is spinning. A spinning black hole drags space around with it and allows atoms to orbit closer to the black hole than is possible for a non-spinning black hole. For example, x-ray observations of the K-α line from the stellar black hole Cygnus X-1 show that the profile of the iron line is skewed to lower energies in a manner consistent with a slowly rotating or non-rotating black hole, whereas data from the black hole, XTE J1650-500, show a much larger skewing to low energies, consistent with a rapidly spinning black hole. Previous observations of some supermassive black holes by Japan's ASCA satellite, XMM-Newton and Chandra have indicated that they may also be rotating rapidly.

In recent years, evidence has been found for a third class of black holes intermediate in mass between stellar mass black holes and supermassive black holes at the center of galaxies. If their emission is not beamed towards us, preferentially, these x-ray sources would have a significantly higher luminosity than normal stellar mass black holes which suggests that they may be black holes with masses in the range of a few hundred solar masses. Current possible explanations for the formation of intermediate mass black holes include the mergers of scores of stellar black holes, or the collapse of a extremely massive star. If their x-ray emission is beamed toward the Earth it would reduce the overall output of x rays from the source, and reduce the estimate of their mass to a value consistent with stellar mass black holes but their higher flux at the Earth would still place them in a separate class.

Supernova Remnants & Pulsars

Over eighty sources in our galaxy and numerous sources in nearby galaxies have been identified with supernova remnants (SNR). Nearly all are characterized by a fragmentary shell with a diameter that gets larger and a spectrum that gets softer with age. The evolution of lumi-

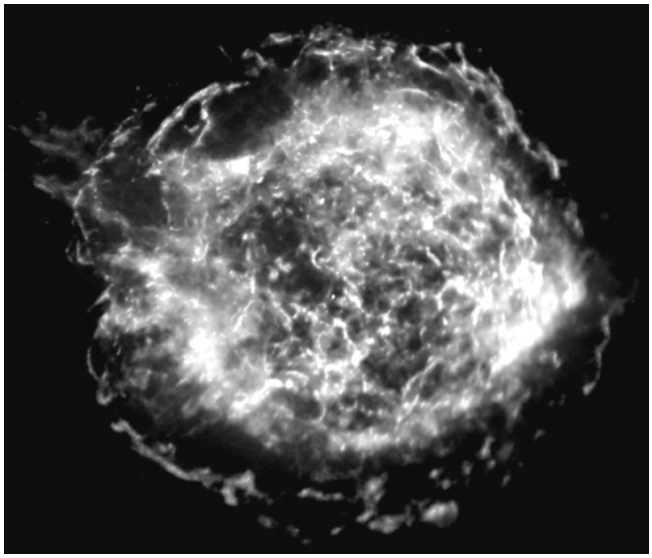

Fig. 1: X-ray image of the supernova remnant Cas A taken with the Chandra X-ray Observatory. There is evidence for a point source at the center, which could be a hot neutron star. The image scale is 7 arcminutes on a side.

nosity is more complex. Some twenty SNR show evidence for the existence of a neutron star. Four of these are fast pulsars that are losing rotational energy at a rate sufficient or more than sufficient to explain their luminosity. Figures 1 and 2 show x-ray images of two relatively young SNR: Cas A, and the Crab Nebula which contains a rotationally powered pulsar. Cas A was born in 1680 and the Crab in 1054. The distance to both is about 2 kpc. Their x-ray spectra are shown in Figure 3. Cas A contains x-ray lines of highly ionized silicon, sulfur, calcium, and argon, indicative of radiation from a hot plasma ($\sim 10^7$ K) with enriched elemental abundances expected to be associated with supernova ejecta. Two shock waves are visible: the outward-moving shock that is moving into the interstellar medium, and the reverse shock that is moving into the stellar ejecta. In a high quality image a point source is seen near the center which may be a neutron star that appeared when the core of a massive star collapsed to initiate the supernova explosion.

The Crab Nebula presents a dramatically different picture. It is atypical in that there is no indication of a shell of ejecta. Its appearance is dominated by the effects of the rapidly rotating neutron star, or pulsar at its center. The spectrum of the continuum radiation of the Crab Nebula is due to synchrotron radiation from relativistic electrons in the magnetic field of the nebula. It is intrinsically featureless but absorption in the interstellar medium produces a low energy cutoff and elemental absorption edges. The rapidly rotating neutron star is seen directly as a point source twice during each 33-ms period when the radiation is beamed in our direction. Enormous electrical voltages generated by the rotating, highly magnetized neutron star accelerate particles outward along its equator to produce the pulsar wind. These pulsar voltages also produce the polar jets seen spewing x-ray emitting matter and antimatter electrons perpendicular to the rings.

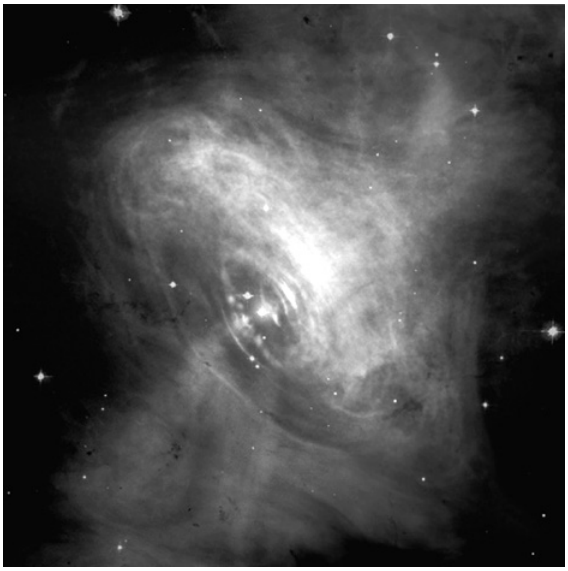

Fig. 2: Chandra's x-ray image of the Crab Nebula (SN 1054) with a 33-ms period pulsar at the center of the rings. The pulsar brightens twice during each period when its rotating beams are pointing toward the Earth. The Crab Nebula is unusual among supernova remnants in that there is no shell produced by supernova ejecta interacting with interstellar material. The image scale is 1.7 arcminutes on a side

As this relativistic wind of matter and antimatter particles from the pulsar plows into the surrounding nebula, it creates a shock wave and forms the inner ring. Energetic shocked particles move outward to brighten the outer ring and produce an extended x-ray glow called a pulsar wind nebula. Observations with the Chandra X-Ray Observatory have led to the discovery of about 50 pulsar wind nebulae in various stages of evolution, both inside supernova remnants and outside. Some of them show spectacular bow shock waves due to their supersonic motion through the interstellar gas.

Extragalactic Sources

Galaxies, Starburst Galaxies

The x-ray emission from galaxies consists of three basic components: (i) individual sources of the type discussed above, (ii) x rays from a hot interstellar medium, and (iii) for at least some galaxies x rays connected with a supermassive black hole in the center. Their neutron star and black hole x-ray binary populations provide clues to their history. For example, the Chandra image of the elliptical galaxy NGC 4261 reveals dozens of black holes and neutron stars strung out across tens of thousands of light years like beads on a necklace. The structure, which is not apparent from the optical image of the galaxy, is thought to be the remains of a collision between galaxies a few billion years ago. According to this interpretation, a smaller galaxy was captured and pulled apart by the gravitational tidal forces of NGC 4261. As it fell into the larger galaxy, large streams of gas were pulled out into long tidal tails.

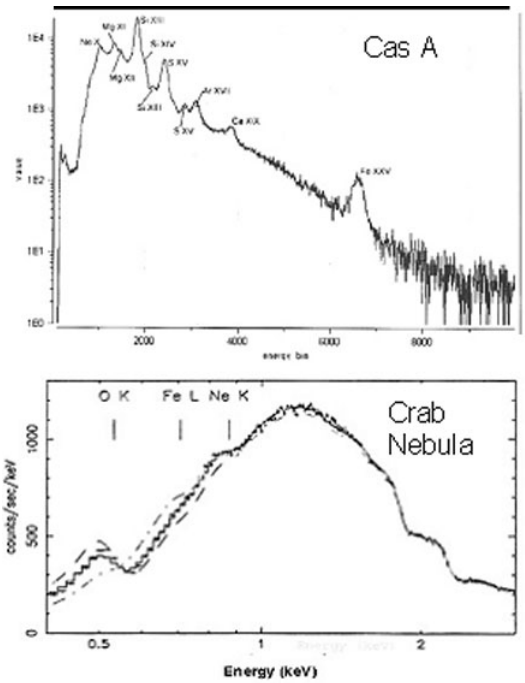

Fig. 3: X-ray spectra of Cas A and the Crab nebula. Strong elemental lines in Cas A's spectrum are indicative of thermal radiation from a hot plasma containing enriched material released from the exploded star and shocked heated matter from the circumstellar medium. The absence of lines in the Crab's spectrum and the presence of polarization is consistent with synchrotron radiation. The low energy cutoff and the absorption edges are caused by the detector response and absorption in the interstellar medium

Shock waves induced by these tidal tails triggered the formation of many massive stars causing it to become a "starburst" galaxy. The shock waves push on giant clouds of gas and dust, causing them to collapse and form a few hundred stars. The massive stars use up their fuel quickly and explode as supernovas, which produce more shock waves and more star formation. The process continues until the gas is consumed or blown away by the explosions to end the starburst phase. During a starburst, stars can form at tens, even hundreds of times greater rates than the star formation rate in normal galaxies. Over the course of a few million years, these stars evolved into neutron stars or black holes. A few of these collapsed stars had companion stars, and became bright x-ray sources as gas from the companions was captured by their intense gravitational fields. Starburst activity typically lasts for ten million years or more, a small fraction of the ten billion year age of the galaxy. The optical evidence for starburst activity fades rather quickly into the stellar background of the galaxy, whereas the x-ray signature such as Chandra's image of the many point sources in NGC 4261 lingers for hundreds of millions of years and may be the best means of identifying the ancient remains of mergers between galaxies.

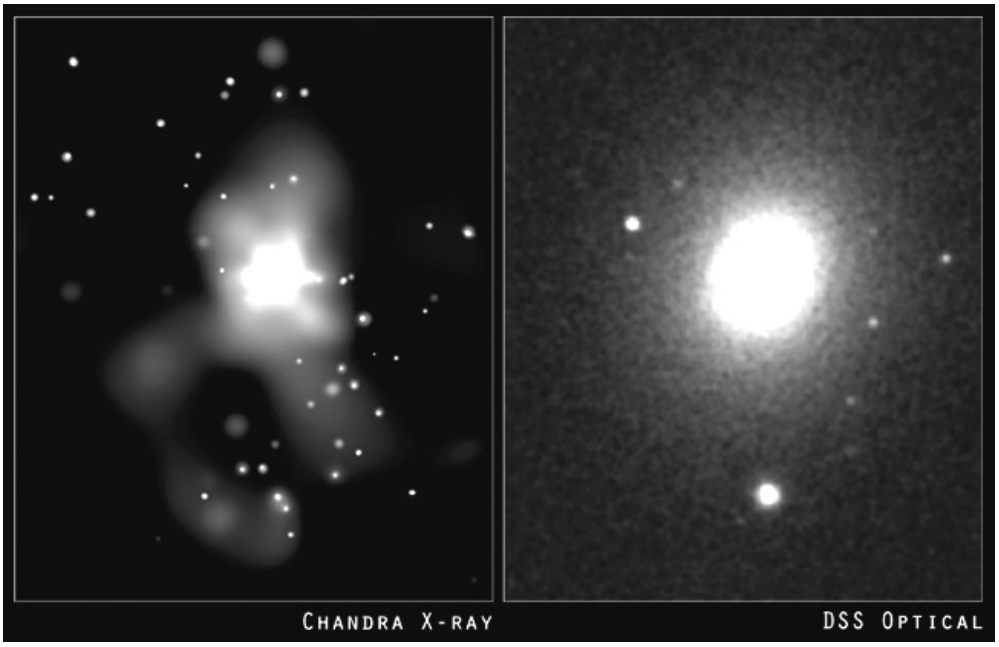

Fig. 4: X-ray image obtained by Chandra (left panel) and optical image of the galaxy NGC 4261. The x-ray image contains many neutron star and black hole binary systems, the residue of an intense period of star formation that was probably induced by the collision between and the merger of a large and smaller galaxy. Each panel is 4×5 arcminutes.

Active Galactic Nuclei

Thousands of x-ray sources have been identified with quasars, Seyfert galaxies, radio galaxies, and BL Lacertae objects, known collectively as active galactic nuclei (AGN). Some differences among them may be due to our observing direction relative to their galactic plane or a jet emanating from a supermassive black hole at the center. The x-ray luminosity of AGN is an appreciable fraction of their total radiative output, in some cases accounting for most of it. Because most supermassive black holes produce x radiation that can penetrate the clouds of dust and gas that surround them, Chandra and XMM have been able to show that most galaxies harbor a supermassive black hole at their centers.

The Chandra Deep Sky surveys showed that the x-ray background radiation is not diffuse. Rather, it resolves into numerous point-like sources due to supermassive black holes in active galactic nuclei. An important, as yet unanswered, question is whether the first supermassive black holes were formed before, at the same time as, or after the first galaxies were formed. At present we know only that some supermassive black holes already existed about a billion years after the Big Bang, about the same time as the most distant observed galaxies.

Larger black holes and/or larger supplies of gas produce higher luminosities, which suggests that the typical galactic nucleus was once a quasar and that the typical quasar will eventually become the nucleus of a normal galaxy when all the matter susceptible to accretion is consumed. This process might take a few million years, or a few hundredths of a percent of

the age of a galaxy, consistent with the observed ratio of quasars to normal galaxies and their ages.

The estimated masses of supermassive black holes range from about a million solar masses to more than a billion solar masses. For example, the Milky Way has a small bulge and a supermassive black hole with a mass of only about 3 million solar masses, whereas the giant elliptical galaxy M87 contains a black hole with a mass in excess of a billion solar masses. An empirical relation between the mass of the black hole, and the mass of the spherical bulge containing the black hole (as deduced from the magnitude of the average random velocity of the stars in the bulge) suggests that the initial rotation rate, or collisions with other galaxies may play a role.

X-ray and radio observations have shown that the influence of supermassive black hole continues well beyond the central regions of the galaxy. High-energy jets extend away from some supermassive black holes moving at nearly the speed of light in tight beams that travel hundreds of thousands of light years. These jets are thought to originate in the accretion disk around the black hole, where the twisting of magnetic field lines creates large electromagnetic fields that launch and collimate the jet.

X-radiation from the jets near the nucleus of the galaxy is produced most likely by synchrotron radiation from relativistic electrons in the jet, or possibly by Compton upscattering of optical, infrared or radio synchrotron photons by high energy electrons in the jet. Outside the galaxy, the X-radiation of the jets is probably due to Compton upscattering of cosmic microwave background photons by the electrons in the jet. The number and energy density of the cosmic microwave background photons increases with increasing redshift (or look-back time) and offsets the effect of increasing distance, so the large jets should be detectable at very large distances, and can provide a useful probe of the environment of young galaxies.

The energy content of the large x-ray jets is enormous, suggesting that the conversion of the jet's kinetic energy into heat could play a role in regulating the growth of galaxies. A striking example of episodic outbursts from the central regions of a galaxy is provided by the Chandra x-ray image of the galaxy Perseus A, which shows wavelike features produced by explosions occurring at intervals of about ten million years.

Clusters of Galaxies

Clusters of galaxies contain hundreds to thousands of galaxies immersed in an enormous cloud of hot gas held together by dark matter. They are the largest gravitationally bound systems in the universe. The hot gas clouds in clusters are detected as x-ray sources extending over a region several million light years in diameter. The emission is characteristic of radiation from a hot, optically thin plasma with temperatures in the range 10–100 MK, central densities of $\sim 10^{-3}\,\mathrm{cm}^{-3}$, and relative elemental abundances, based mainly upon the measurement of x-ray lines from Fe ions with one and two electrons, that are between one-quarter and one-half of the solar abundance (Fe/H $\sim 3 \times 10^{-4}$).

The mass of the intracluster gas exceeds the mass of all the stars, gas and dust in all the galaxies in the cluster. The gas was likely heated primarily by shock waves generated as the system collapsed. Large N-body computer simulations indicate that clusters are formed gradually over the eons through the mergers of groups and subclusters of galaxies. X-ray images of clusters provide dramatic confirmation of this picture, showing many examples of clusters with rich substructure due to merging subclusters.

Warm-hot Intergalactic Medium and Missing Baryons

However, the mass of the intracluster gas plus all the stars and dust in galaxies is insufficient by a factor of two to account for all the baryons that must have existed in the early universe. Therefore theorists have predicted that the missing baryons must reside in the intergalactic medium where the atoms are too hot and therefore too highly ionized to be detectable by absorbing the radio or visible light emissions of objects along the line of sight but not so highly ionized that they cannot absorb x rays. Indeed measurements of x-ray absorption lines in the spectrum of an AGN have confirmed the hypothesis that there exists a diaphanous filamentary warm-hot intergalactic medium (WHIM) outside of galaxies and clusters. If the points sampled are typical then the number of baryons that exists in the WHIM is sufficient to resolve the factor of two discrepancy.

Dark Matter

Because the cluster gas is in hydrostatic equilibrium in the cluster's gravitational potential, the morphology of the x-ray source reflects the distribution of mass in the cluster. Hence, the underlying mass of subcluster components and extended dark halos around individual galaxies can be studied by measuring the distribution of gas density and temperature. X-ray studies of numerous clusters have shown that 70 to 90 percent of the mass of a typical cluster consists of dark matter, i. e., matter which does emit any radiation - mysterious particles left over from the dense early universe that interact with each other and "normal" matter only through gravity. The favored type of dark matter is cold dark matter, which gets its name from the assumption that its constituent particles were moving slowly when galaxies and galaxy clusters began to form.

The exact nature of cold dark matter remains a mystery, but x-ray observations have provided a constraint. For example, the Chandra image of the galaxy Abell 2029 shows a smooth increase in the intensity of x rays all the way into the central galaxy of the cluster. By precisely measuring the temperature and intensity distribution of the x rays, astronomers were able to make the best map yet of the distribution of dark matter in the inner region of the galaxy cluster. The x-ray data imply that the density of dark matter increases smoothly all the way into the central galaxy of the cluster. This discovery agrees with the predictions of cold dark matter models, and is contrary to other dark matter models, such as self-interacting dark matter, that predict a leveling off of the amount of dark matter in the center of the cluster.

Dark Energy

While an explanation for dark matter is still lacking an effect that is even more enigmatic has been discovered. Optical astronomers have observed that the visible light from Type-1a supernovas, which act as standard candles, is fainter than expected in distant galaxies. The best explanation is that they are more distant than originally thought, which implies that the expansion of the universe must be accelerating. Chandra's measurements of the dark matter content of clusters of galaxies corroborates this astounding result by a method that is both independent of and complementary to the Type 1a supernova findings.

X-ray observations have determined the gas fraction of the total mass, i. e., the ratio of the mass of the hot gas and the mass of the dark matter for a number of cluster of galaxies. The observed values of the gas fraction depend on the distance scale adopted, which in turn

depends on the expansion rate of the universe. Because galaxy clusters are the largest bound structures in the Universe, they are thought to represent a fair sample of the matter content in the universe. If so, the ratio of hot gas and dark matter should be the same for every cluster. Using this assumption, the parameters in the distance scale can be adjusted to determine which one fits the data best. The best fit parameters are consistent with a model in which the expansion of the Universe was first decelerating until about six billion years ago, and then began to accelerate.

The driving force behind cosmic acceleration is being attributed to a new entity known as dark energy. Accounting for the existence of dark energy requires either a refinement of Einstein's theory of general relativity or a major revision of some other area of fundamental physics. Assuming that dark energy is responsible for the acceleration, combining the x-ray results with observations of Type 1a supernovas and the cosmic microwave background radiation indicates that dark energy makes up about 75% of the Universe, dark matter about 21%, and visible matter about 4%.

Concluding Remarks

The success of the x-ray astronomy observatories launched at the turn of the century has placed x-ray astronomy on a par with optical astronomy as an area of research into the structure and evolution of the universe. Furthermore, the observation of strong gravity effects around black holes plus the constraints placed on dark matter and dark energy have demonstrated that x-ray astronomy has an important role to play in fundamental physics research. For the future we require much larger area telescopes to provide more photons for spectroscopy and to extend our reach further back in time. International collaboration, already a prominent feature of the current generation of observatories will be even more essential for the larger instruments of the next generation.

See also: Black Holes; Galaxies; Interstellar Medium; Neutron Stars; Pulsars; Quasars; Synchrotron Radiation.

Bibliography

R. Giacconi, H. Gursky, F. Paolini, and B. Rossi, *Phys. Rev. Lett.* **9**, 439 (1962).

P. A. Charles and F. D. Seward, *Exploring the X-Ray Universe*. Cambridge University Press, Cambridge, 1995. (New edition in few years.)

M. C. Weisskopf, B. Brinkman, C. Canizares, G. Garmire, S. Murray, and L. P. van Spreybroeck, *PASP* **1**, 114 (2002).

W. H. Lewin and M. van der Klis (eds.), *Compact Stellar X-Ray Sources*. Cambridge University Press, Cambridge, 2005.

Web Sites that are resources for current information

Chandra X-Ray Observatory: `chandra.harvard.edu`

NASA's High Energy Astrophysics Science Archive Research Center: `heasarc.gsfc.gov`

Astrophysics

M. Bartelmann

Astrophysics is unique among the physical disciplines for two reasons. First, experiments with astrophysical objects are generally impossible, except perhaps for the rare occasions when extraterrestrial material like lunar rocks or meteorites can be studied under laboratory conditions. Instead, astrophysics relies on information transmitted mostly by light from radio waves to γ rays, but also elementary particles (e. g., neutrinos) or gravitational waves (in future). Second, there is hardly any area of physics which does not play a role in astrophysics, from quantum field theory to the general theory of relativity.

Physics was first applied to astronomy by Galileo, Kepler and Newton, who succeeded in explaining the motion of bodies in the solar system. Modern astrophysics emerged from the study of stars, and the discovery that stellar spectra exhibit characteristic absorption lines. Questions as to the origin of these spectra, the internal constitution of the stars and their energy source led to the vigorous subsequent development of astrophysics.

Of the four fundamental interactions known in physics, only gravity is relevant on the largest scales, because the strong and weak interactions are restricted to subatomic scales, and the electromagnetic interaction can be shielded by opposite charges. A possible exception are magnetic fields, which can be important even on cosmological scales. The current theory of gravity is Einstein's general theory of relativity, which allows simple, symmetric models for the universe as a whole to be constructed. Perhaps one of the most surprising recent successes of astrophysics was the discovery that these models indeed seem to describe our universe extremely well. According to them, the universe originated 14 billion years ago in a hot, dense state, called the Big Bang, from which it subsequently expanded and cooled.

At times very close to the Big Bang, quantum physics must surely be used, but it is yet unknown how general relativity and quantum theory are to be combined. It appears that a particular class of quantum fields, so-called scalar fields, are crucially important for the appearance of the universe and for the origin of the structures contained in it. The cooling of the universe is described by the thermodynamics of (relativistic) quantum gases. When the universe was about three minutes old, helium and a few light elements could be formed from hydrogen by nuclear fusion. About 400 000 years after the Big Bang, atoms formed from nuclei and electrons. The sudden disappearance of charged particles allowed the ubiquitous background of electromagnetic radiation to decouple from matter and stream almost freely through the universe since.

Cosmic structures as we see them today were already laid out at that time, and they left their imprint in the cosmic background radiation, which we can observe today. The corresponding fluctuation patterns in the radiation density can be predicted using kinetic theory and Compton scattering. Although they are at the level of 10 parts per million only, they have been observed and accurately mapped, thereby confirming and tightly constraining the generally-relativistic cosmological models.

Three questions immediately arise from there. First, what was the origin of the structures we see in the density of the cosmic background radiation? Second, starting from these tiny fluctuations, how could the cosmic structures be built that we see today? And third, how could the rich diversity of cosmic objects form?

The speculative answer to the first question, which has become plausible recently, is that vacuum fluctuations of quantum fields very early in the universe were stretched to cosmic scales in a process called cosmological inflation. Structures as large as galaxies and galaxy clusters could then be traced back to quantum fluctuations which are an inevitable consequence of Heisenberg's uncertainty principle.

The second question leads us to assume that the majority of matter in the universe does not interact with light, because otherwise the fluctuations in the cosmic radiation background would have to be much larger. Such a hypothetic form of "dark matter" has not been observed yet, but apparently we have to accept that ordinary matter composed of protons, neutrons and electrons is the exception rather than the rule. Assuming that there is dark matter, and that it consists of weakly-interacting, massive elementary particles, the appearance of cosmic structures from scales smaller than galaxies to much larger than galaxy clusters can well be explained. According to general relativity, any matter inhomogeneities deflect light. This gives rise to the gravitational lensing effect, which allows even dark structures to be traced.

The third question is extremely complicated in detail. Accepting that the skeleton of cosmic structures is provided by dark matter, we have to study when and how visible entities like stars, galaxies and others could originate. Almost all ordinary gas in the universe became neutral 400 000 years after the Big Bang when the cosmic radiation background was released. In the spectra of the bright, distant quasars, we see that the gas must have been ionized again when the universe was about a billion years old. The picture is plausible that gas fell into the structures provided by the dark matter, heated up, and formed the first stars, whose energetic radiation could re-ionize the cosmic gas. Much of the cosmic gas is not bound in stars and galaxies, but forms the diffusely distributed intergalactic medium.

It is equally plausible that the first galaxies formed at about the same time from gas falling into "halos" of dark matter. The most obvious observational facts on galaxies which need to be explained are that they broadly fall into two morphological classes, ellipticals and spirals, whose parameters exhibit tight correlations. How galaxies form and develop, and how they change in response to the density of their environment, is an area of active research. Galaxy clusters, which are assemblies of hundreds or even thousands of galaxies held together by the gravity of dark matter, formed much later in cosmic history than the galaxies, when the universe had approximately reached half of its present age.

We have no consistent theory of how stars form. Necessary conditions are that sufficient amounts of gas are concentrated sufficiently so that they can cool, contract and reach central densities and temperatures necessary for nuclear fusion to set in. For the first generation of stars, the cooling seems to be the most eminent problem. Efficient cooling agents are traces of heavy elements, broadly called "metals" in astrophysics, which can radiate energy away by line emission. Primordial gas, as it was available shortly after the Big Bang, did not contain any elements heavier than Lithium in appreciable quantities. All heavier elements had to be produced by nuclear fusion in stellar interiors, and then released by stellar winds and supernova explosions. That way, the gas could be enriched with metals, which allowed for progressively more efficient cooling, and the formation of new generations of stars.

It marked one of the most fundamental achievements of astrophysics when it was realized that stars produce their energy by nuclear fusion. A firm proof of nuclear energy production in the sun was given by the detection of solar neutrinos, which are emitted in the β decay accompanying many nuclear reactions. The energy produced in stellar cores has to travel

a long way until it is finally released, mostly as visible light, at the stellar surfaces. The energy transport depends on the macroscopic constitution of the stars, as described by their density, temperature, and pressure profiles, and on microscopic quantities such as the opacity of the stellar material. While the stellar structure is determined by hydrodynamics and energy-transport mechanisms, opacities have to be calculated from the quantum mechanics of atoms. The radiation that we receive from stellar "surfaces", or photospheres, shows spectral lines caused by atomic absorption. Those "Fraunhofer lines" allow the chemical composition of outer stellar layers and many physical properties of the stars to be deduced. Their detection marks the onset of modern astrophysics.

Observations show that stars are not randomly distributed in the plane spanned by temperature and luminosity, but fall within well-defined, sharp regions on that plane. It was one of the breakthroughs of astrophysics when the theory of stellar structure, based on the assumption of nuclear fusion in stellar cores, was able to explain those patterns. The most prominent of those are the "main sequence", on which stars fall while they produce helium from hydrogen in their cores, and the "giant branch", to which stars move when their central hydrogen supply is exhausted and they proceed to burning helium.

As isolated, gravitationally-bound systems, stars can oscillate in a multitude of modes. Such oscillations can be inferred from velocity patterns on the solar surface, and their wave lengths and frequencies can be compared to those predicted from stellar models. The overall excellent agreement between theory and observations of solar oscillations provides further support for the theory of stellar structure and energy production.

It was thus perceived as a fundamental physical problem when neutrino detectors kept finding substantially fewer solar neutrinos than predicted by the otherwise well-established solar model. This problem was solved when it was proven that neutrinos can change flavour in a process called neutrino oscillations, which requires the neutrinos to have a small, but non-vanishing mass. This is a prototypical example for astrophysics driving theoretical physics beyond the standard model of elementary-particle theory.

Stars end in different ways depending on their mass. When their nuclear fuel is exhausted, low-mass stars cool and contract to form white dwarfs, which keep radiating until their internal energy is lost. White dwarfs are stabilized by the electron degeneracy pressure in their interiors, which follows from Pauli's exclusion principle. Masses higher than 1.4 solar masses (the Chandrasekhar limit) cannot be stabilized against gravity that way. They collapse further until electrons and protons are converted to neutrons by inverse β decay. Thus neutron stars are formed, which are stabilized by the degeneracy pressure of the neutrons. Yet more massive objects can collapse to form black holes, i. e., singularities in spacetime from which even light cannot escape. Stellar collapse is accompanied by explosions giving rise to supernovae. Despite intense research, the physics of core-collapse supernovae is not yet fully understood. Neutrino transport seems to be crucially important. It is likely that the mysterious γ-ray bursts are also related to the collapse of (very) massive stars.

Another type of supernova (type Ia) occurs in binary systems in which a white dwarf accretes mass from an overflowing companion star. When the Chandrasekhar mass limit is reached, the white dwarf collapses and explodes. Since the exploding mass is approximately fixed, so is the luminosity. Thus, supernovae of type Ia can be used as "standard candles" in cosmology. From their apparent brightness, compared to their known luminosity, their distance can be inferred. Since looking at large distances means looking back in time, observa-

tions of supernovae of type Ia allow the expansion history of the universe to be reconstructed, yielding the surprising result that the universe turned over from decelerated to accelerated expansion when it was about half of its present age. This is interpreted as evidence for a "dark energy" which can drive accelerated cosmic expansion due to its negative pressure. Besides dark matter, dark energy is one of the most fundamental enigmas in current astrophysical research, which may find its solution in the physics of quantum fields.

Black holes are not only possible end products of stellar evolution. Detailed measurements of stellar dynamics in the vicinity of the center of our Galaxy has revealed the presence of a black hole with about a million solar masses. Similar, but necessarily much less well-resolved observations indicate that black holes exist in the cores of most, if not all, galaxies. A surprising recent discovery showed that the masses of galactic black holes are correlated with the masses of their host galaxies, which indicates that their formation processes may be closely linked. Black holes in galactic cores are believed to be the central machines of the so-called quasars, objects which appear point-like as stars, but have luminosities well exceeding those of ordinary galaxies. It was realized soon after quasars were first discovered that nuclear fusion is insufficient to release such amounts of energy. The only viable energy-production mechanism is the conversion of gravitational potential energy into radiation in an accretion process. When gas streams into a black hole, its angular momentum forces it to orbit around the black hole in a disk. The friction in the disk allows the gas to lose its angular momentum and to gradually flow inward. At the same time, it heats up the gas and causes it to radiate in a broad wavelength range. Quasars were abundant in the young universe and apparently ceased being effective when the gas supply was exhausted.

A widespread phenomenon in astrophysical objects are magnetic fields. They may be produced in stars and accretion disks by battery or dynamo mechanisms, but possibly also during phase transitions in the early universe. Stellar magnetic fields can be blown into the ambient medium by winds or supernova explosions. Contraction processes like those leading to the formation of white dwarfs, neutron stars, or accretion disks can then produce ordered magnetic fields of considerable strength. Even galaxy clusters are found to be permeated by large-scale magnetic fields, whose origin is yet unclear. Magnetic fields can collimate jets streaming away along the symmetry axes of accretion disks. Relativistic particles, possibly accelerated in shock fronts, emit synchrotron radiation in the radio waveband when gyrating along magnetic-field lines.

On the smallest scales, accretion disks are also crucially important for the formation of planets around stars. The material in dusty disks can efficiently cool, fragment, and coagulate to form planetesimals and progressively larger objects. A large number of extra-solar planets was recently discovered, almost exclusively through their gravitational pull on their host stars. The distribution of these planets in mass and orbital parameters is much different from that seen in our solar system, illustrating that the theory of planet formation is in an early state.

Astrophysical observations started in the visible light, then expanded to the radio, infrared, X-ray, γ-ray and ultraviolet wave bands. The last remaining wide gap in the observed electromagnetic spectrum will be closed by submillimeter observations which have already begun and will be intensified in the near future. Gravitational-wave detectors are being built which promise to provide insight into a broad variety of astrophysical processes, from supernova explosions to the physics of the early universe.

See also: Black Holes; Cosmology; Interstellar Medium; Neutron Stars; Nucleosynthesis; Pulsars; Quasars; Stellar Energy Sources And Evolution; Sun; Universe.

Bibliography

The best references on astrophysics are current journals, review volumes, and reports of symposia.

Journals

Astrophysical Journal, published by the University of Chicago Press, Chicago, Ill.
Astrophysics and Space Science, published by Reidel Publishers, Dordrecht, The Netherlands.
Astronomy and Astrophysics, published by Springer-Verlag, Berlin.
Monthly Notices of the Royal Astronomical Society, published by Blackwell Scientific Publications, Oxford, England.

Review Volumes

Annual Review of Astronomy and Astrophysics. Annual Reviews Inc., Palo Alto, Calif.

Reports of Symposia

International Astronomical Union Symposia Proceedings, published by Reidel Publishers, Dordrecht, The Netherlands. Examples pertaining to the *Sun*: Vols. 35, 43, 56, 68, 71, 86; pertaining to *galactic structure, galaxies, quasars*, etc.: Vols. 38, 44, 58, 60, 63, 64, 69, 74, 79, 84, 92, 97, 104, 106, 108, 116, 117, 119, 124, 126, 127, 130; pertaining to *planets and meteors*: Vols. 33, 40, 47, 48, 62, 65, 89, 90; pertaining to *gaseous nebulae and the interstellar medium*: Vols. 34, 39, 46, 52, 76, 87, 103, 120, 131; pertaining to *stars, stellar spectra, and stellar evolution*: Vols. 42, 50, 52, 54, 55, 59, 66, 67, 70, 72, 75, 83, 95, 98, 99, 101, 105, 113, 115, 122, 123, 125.

Atmospheric Physics

Eric P. Shettle

The atmosphere is the gaseous shell that surrounds the earth and atmospheric physics is the discipline describing the physical processes occurring in the atmosphere. The atmosphere is hundreds of kilometers thick, providing the air we breathe. It presence is most notable through the phenomena that we know as weather.

Atmospheric Composition

The composition of the earth's atmosphere is dominated by two gases, nitrogen and oxygen which together make up about 99% of clean dry air, with most of the remaining 1% argon. These are nearly homogeneous up to altitudes of 80 to 90 km. Water vapor is highly variable; values near the surface can range from 3% in the tropics to a few tenths of a percent at mid-latitudes during the winter. The concentration of water vapor decreases rapidly above the surface and only makes up a few parts per million of the upper atmosphere. The concentrations of these and the other principal atmospheric gases are summarized in Table 1.

Table 1: Composition of the atmosphere.

Substance	Vol. % in dry air	Mol. weight
Total atmosphere		
Dry air	100.000	28.97
Nitrogen	78.083	28.02
Oxygen	20.946	32.00
Argon	0.934	39.88
Carbon dioxide	0.037	44.00
Neon	0.0018	20.0
Helium	0.00052	4.00
Ozone	Variable	48.00
Water vapor	0 to 3.0	18.02

Vertical Structure of the Atmosphere

As noted in the previous section the primary atmospheric species are uniformly mixed up to about 80 to 90 kilometers. At higher altitudes molecular diffusion begins to separate the different gases by their molecular weight. So the concentration of oxygen decreases relative to nitrogen. Also at these altitudes the solar ultraviolet radiation dissociates the oxygen molecules providing a source of atomic oxygen, and the photodissociation of water vapor at lower altitudes produces atomic hydrogen. The atmosphere is nearly in hydrostatic equilibrium which means that the pressure at any altitude is equal to the weight of the atmosphere above that altitude. The result of this and the ideal gas law is that both the atmospheric pressure and the density decrease exponentially with height. Half of the total mass of the atmosphere is below about 5.5 km and nearly 90% is below 16 km. By 100 km the density of the atmosphere has dropped by a factor of three million. Satellite orbits are generally above at least 300 km to minimize atmospheric drag. Satellite orbits at 700 to 800 km are still modified by drag effects over several years.

The atmospheric temperature tends to decrease up to an altitude of about 8 km in the polar regions to about 16 km in the tropics. This region of generally decreasing temperatures is known as the *troposphere* and the local temperature minimum the *tropopause*. The tropopause temperatures range from 180 to 220 K. The *lapse rate*, which means the decrease of temperature with altitude, is about 6.5 K/km on a global average, but can vary considerably and there can be regions where there is a temperature inversion where the temperature increases with altitude over a shallow layer. The troposphere is the part of the atmosphere where most of our weather occurs. Large-scale turbulence and mixing play a significant role in the distribution of the atmospheric properties. Above the tropopause there is a marked change in the lapse rate in the region known as the *stratosphere*. In the first few kilometers of the stratosphere the temperature is nearly constant and then the temperature increases with altitude up to a height of about 50 km where the *stratopause* is reached. This increase in temperature is caused by the absorption of the incident solar radiation by ozone which reaches in maximum concentrations in the stratosphere. Above stratopause the temperature again decreases with height, in the region known as the *mesosphere*. This region ends with the *mesopause*, where the lowest temperature in the atmosphere is about 180 K and can be as cold as 120 to 130 K during the

polar summers. Above the mesopause the atmospheric temperatures once again increase due the absorption the solar ultraviolet radiation for wavelengths less than 185 nm. This region is known as the *thermosphere*.

The incident ultraviolet radiation and energetic particles from the sun produces dissociation and ionization of the atmospheric constituents, in an atmospheric region known as the *ionosphere*, which overlaps with the mesosphere and thermosphere. This ionization produces a free electron and a positively charged ion. The ionosphere is divided into three regions or layers based on the number of free electrons and the dominant ion. The D layer extends from about 50 to 90 km, the E layer from roughly 90 to 150 km, and the F layer above about 150 km.

Atmospheric Radiation

The atmospheric structure discussed in the previous section is largely controlled by the absorption of the incident solar radiation and the long wave radiation emitted by the earth. This absorbed radiation provides the energy required to drive the atmospheric motions. The intensity of the solar radiation incident on a perpendicular plane at the top of the atmosphere is about 1370 watt/m^2. The daily solar flux per unit horizontal incident on the top of the atmosphere depends on the angle of incidence and the length of daylight. It ranges a maximum at poles at summer equinox, decreasing slowly towards the equator and then more rapidly in the winter hemisphere going to zero at latitudes where there are 24 hours of darkness. About half (51%) of the sunlight is absorbed at the earth's surface. Another 30% is reflected back into space by the surface, clouds, or atmospheric scattering. The remaining 19% is absorbed by the atmosphere. Essentially all of the ultraviolet radiation at wavelengths less than 300 nm is absorbed before it reaches the surface. By contrast much of the visible and some of the infrared solar radiation reaches the earth's surface. Significant portions of the infrared solar radiation are absorbed by water vapor and carbon dioxide, as well other gases.

The earth emits radiation with a characteristic blackbody temperature of about 270 K which peaks around 10 μm. However much of these is absorbed in the atmosphere predominately by water vapor and carbon dioxide plus ozone and other species. The atmosphere re-emits the radiation, and some of the upwelling radiation is absorbed in turn at higher levels in the atmosphere where it is again re-emitted. Much of the surface radiation in the *atmospheric window* between about 8 to 12 μm reaches space without being absorbed. The characteristic blackbody temperature of the long wavelength radiation emitted into space is a weighted average temperature of where in the earth/atmosphere system the radiation was emitted from. This is called the *terrestrial* (or *earth*) *radiation*.

On an annual basis the global average the solar radiation absorbed by the earth/atmosphere system equals the terrestrial radiation emitted into space. However on a regional basis this is not true. On an annual basis the equatorial latitudes out to about 35° absorb more solar radiation than terrestrial radiation is emitted back into space, and at higher latitudes the reverse ins true. Over most of the winter hemisphere, the terrestrial radiation lost to space exceeds the absorbed solar radiation with the reverse holding in much of the summer hemisphere. These energy imbalances drive the large scale atmospheric circulation.

The Greenhouse Effect

The atmospheric radiation emitted downward heats the lower atmosphere and the earth's surface causing them to be warmer than they would be in the absence of an atmosphere. This is called the *greenhouse effect* by analogy with how greenhouses were thought to be heated. The glass transmits most of the incident solar radiation but absorbs most of the outgoing infrared radiation from inside the greenhouse and radiating half of it back down. However this is a misnomer, since R. W. Wood demonstrated that primary mechanism heating greenhouses is that the glass blocks the removal of heat by convection. He compared the temperature inside two small greenhouses one of which used rock salt, which is transparent in the infrared instead of glass.

There is a concern that the increase in carbon dioxide (and other gases such as methane) since the industrial revolution could increase the atmospheric absorption of the outgoing terrestrial radiation. Carbon dioxide has increased from less than 290 parts per million by volume (ppmv) of the atmosphere before 1900, to 375 ppmv in 2003. The combustion of fossil fuels such as coal or oil produces carbon dioxide. The global average temperature has increased over the last century by about half a degree Centigrade, with geographic and seasonal variations. This is consistent with the predictions of global climates on the impact of such an increase of the greenhouse gases.

Remote Sensing

The radiation emitted, scattered or transmitted by the atmosphere is used increasingly to remotely measure the properties of the atmosphere from satellites, ground based, or airborne instruments. Measurements of the radiation emitted by carbon dioxide at different wavelengths where the strength of the absorption varies can be used to measure the atmospheric temperature as function of altitude. Similar measurements as function of wavelength across the water vapor absorption bands can be used to determine the vertical distribution of water vapor. Satellite measurements of the amount of sunlight backscattered at different ultraviolet wavelengths can be used to determine the vertical distribution of ozone.

Moisture in the Atmosphere

Water plays a unique role in the atmosphere because at the range of atmospheric temperatures and pressures it can be present as a solid, as a liquid, or as gas. The phase changes between these different states require that heat be absorbed or released. The absorbed heat energy is known as latent heat and is released back into the atmosphere when the phase transition is reversed. This provides one of the mechanisms for the energy transport required to compensate for the regional imbalances between the absorbed solar radiation and emitted terrestrial radiation. Water which evaporates from the tropical oceans absorbs latent heat which is when the air containing that water vapor is cooled sufficiently for the water vapor to condense forming cloud droplets. If it cools further so that the cloud droplets freeze, forming ice particles additional latent heat is released.

See also: Aurora; Corona Discharge; Ionosphere; Lightning; Magnetosphere; Meteorology; Rayleigh Scattering; Refraction.

Bibliography

R. G. Fleagle and J. A. Businger, *An Introduction to Atmospheric Physics*. Academic Press, San Diego, CA, 1980.

A. S. Jursa (ed.), *Handbook of Geophysics and the Space Environment*. National Technical Information Service, Springfield, Virginia, 1985.

F. K. Lutgens and E. J. Tarbuck, *The Atmosphere – An Introduction to Meteorology*. Prentice Hall, Englewood Cliffs, NJ, 1995.

M. L. Salby, *Fundamentals of Atmospheric Physics*. Academic Press, San Diego, CA, 1996.

Atomic Spectroscopy

A. Lurio[†] *and A. F. Starace*

Introduction

The progressively more detailed understanding of the emission and absorption spectra of atoms has led from the Bohr theory of the hydrogen atom (1913), to the discovery of electron spin by Uhlenbeck and Goudsmit (1925), to the Pauli exclusion principle (1926), and ultimately to the nonrelativistic quantum-mechanical Schrödinger equation (1926).

From these developments we find the following principles apply to the interpretation of all atomic spectra.

1. An atomic system can exist only in discrete stationary states corresponding to a well-defined energy E (within the Heisenberg uncertainty limit $\Delta E \Delta t \sim \hbar$, where Δt is the lifetime of the state). Transitions between these states, including the emission and absorption of radiation, require the complete transfer of an amount of energy equal to the difference in energy between these states.

2. The frequency of the emitted or absorbed radiation in going from state 2 to state 1 is given by $\omega = (E_2 - E_1)/\hbar$ (ω negative is absorption).

3. Each stationary state has associated with it a definite quantized angular momentum J and a definite parity (defined later). The projection of the angular momentum on any chosen direction in space is quantized with allowed values $m_J = J, J-1, \ldots, -J$.

Each atom has its own unique spectrum. The interpretation of this spectrum has led to the classification of many of the stationary-state energy levels of neutral and several-times-ionized atoms. These results are tabulated in a classic three-volume NBS publication by Charlotte Moore. We shall attempt here to give a simplified treatment of the physical basis of this classification.

[†]deceased

Table 1: Designation of electron states.

	$l=0$	$l=1$	$l=2$	$l=3$	$l=4$
	s	p	d	f	g
$n=l$	1s				
$n=2$	2s	2p			
$n=3$	3s	3p	3d		
$n=4$	4s	4p	4d	4f	
$n=5$	5s	5p	5d	5f	5g

Electronic Configurations

The specification of the stationary states of an N-electron atom is given by the solution of the time-independent Schrödinger equation

$$H\Psi = E\Psi \,, \tag{1}$$

where the dominant terms contributing to the Hamiltonian H are

$$H = \sum_{i=1}^{N} \frac{p_i^2}{2m} - \sum_{i=1}^{N} \frac{Ze^2}{r_i} + \sum_{i>j=1}^{N} \frac{e^2}{r_{ij}} \,. \tag{2}$$

The first term is the kinetic energy of the electrons, and the other terms are the potential energies of Coulomb interaction of the electrons with the nucleus and with each other. A good starting approximation to H is the "central-field approximation" in which one replaces the second and third terms of Eq. (2) by $\sum_{i=1}^{N} V(r_i)$, where $V(r_i)$ is the spherically symmetric average potential seen by the ith electron due to all other electrons. Equation (1) is now solvable with a wave function which is the product of N single-electron wave functions. The Schrödinger equation for each of these single-electron wave functions is like that for hydrogen (see Burke) except that the hydrogenic potential-energy term e^2/r is replaced by $V(r)$. Consequently, the same set of quantum numbers n, l, s, m_l, m_s, used to describe the hydrogenic electron apply here. These are respectively, the principal quantum number n, the orbital and spin angular momentum quantum numbers, and their projections on the quantization axis.

Complete specification of an N-electron state requires N sets of these quantum numbers with the Pauli restriction that no two sets can be identical. For a given n and l if all possible m_l and m_s states are occupied, we have a closed subshell; if all $2n^2$ states ($l = 0, 1, \ldots, n-1$) are filled, we have a closed shell. Closed shells and subshells have exactly spherically symmetric charge distributions and zero net angular momentum.

The standard notation for designating the individual electron configurations or orbitals (n, l values) which are combined to form an N-electron product state is shown in Table 1. To illustrate, the ground-state configuration of sodium ($Z = 11$) is written $(1s)^2(2s)^2(2p)^6 3s$, where the superscripts indicate the number of electrons of a given n,l type.

Spectroscopic Notation

To completely specify an atomic state, besides listing as above the individual electron orbitals, we must also specify the coupling of the angular momenta for all unfilled subshells. The closed shells couple to give zero angular momentum.

The Hamiltonian of Eq. (2) commutes with $\mathbf{L} = \sum_{i=1}^{N} \mathbf{l}_i$, the total orbital angular momentum, and with $\mathbf{S} = \sum_{i=1}^{N} \mathbf{s}_i$, the total spin angular momentum, and thus with $\mathbf{J} = \mathbf{L} + \mathbf{S}$, the total electronic angular momentum of the atom. Within a given configuration, therefore, one may take linear combinations of products of the central-field-approximation orbitals to form states having exact values of the quantum numbers L^2, S^2, J^2, and M_J. All these states are degenerate in energy in the central-field approximation. When we use perturbation theory to take into account the difference between the central-field-approximation potential energy and the potential-energy terms of Eq. (2) we find that states with different L and S have significantly different energies. This treatment, which works well for many atoms (especially the light ones), is called the *LS* or Russell–Saunders coupling scheme. In the *LS* coupling scheme we couple vectorially all open-shell electron spins to obtain a number of different S values and couple all the open shell L_i to obtain a number of different L values. In early atomic spectroscopy the *vector model* was used to visualize these different coupling schemes. White gives an extensive discussion of this model. The number of permitted L and S values is discussed in detail by Condon and Shortley. $\mathbf{L}$ and $\mathbf{S}$ are now coupled vectorially to form the total angular momentum $\mathbf{J} = \mathbf{L} + \mathbf{S}$. J takes values from $|L+S|$ to $|L-S|$.

At this point we add the spin–orbit interaction $H_{\text{so}} = \xi \mathbf{L} \cdot \mathbf{S}$ (*see* Fine and Hyperfine Spectra and Interactions) which removes the degeneracy between the different J values of the same L and S. The spectroscopic notation to designate these *LS* coupled states is ${}^{2S+1}L_{2J+1}$, where, similar to Table 1, for $L = 0, 1, 2, 3$ we use the capital letters S, P, D, F, etc. To illustrate, if $S = 1$, $L = 2$, and $J = 1$ we would write ${}^3D_1$. The levels of a given J have $2J+1$ magnetic sublevels which in the presence of an external magnetic field split apart (*see* Zeeman and Stark Effects).

Transition Rates

Any isolated atom in an excited state will decay spontaneously to lower energy states emitting radiation and ultimately ending in the ground state. Also, in the presence of an external radiation field of the proper frequency, an atom can absorb radiation and make a transition to an excited state.

We will discuss only allowed electric dipole transitions. Classically the time-averaged power radiated per unit solid angle in direction $\mathbf{n}$ by a harmonically varying charge distribution $\rho(r)$ with an electric dipole moment $\mathbf{P} = \int_0^\infty \mathbf{r}\rho(\mathbf{r})\, \mathrm{d}v$ is (in Gaussian units)

$$\frac{\mathrm{d}w}{\mathrm{d}t} = \frac{ck^4}{8\pi} |\mathbf{n} \times (\mathbf{n} \times \mathbf{P})|^2 ,$$

which for a linearly oscillating dipole moment parallel to the z axis reduces to $\mathrm{d}w/\mathrm{d}t = (ck^4/8\pi)|\mathbf{P}|^2 \sin\theta$, where $k = \omega/c$ and θ is the angle between the z axis and the direction of radiation n. For a charge distribution rotating about the z axis, $\mathrm{d}w/\mathrm{d}t = (ck^4/8\pi)|\mathbf{P}|^2(1+\cos\theta)$. We make the connection with the quantum-mechanical description of the atomic system by $\mathbf{P} = 2\mathbf{P}_{ij}$ where $\mathbf{P}_{ij} = \int \bar{\Psi}_i \mathbf{P} \Psi_j \, \mathrm{d}v$, and by setting the harmonic frequency ω equal to $\omega = (E_i - E_j)/\hbar$, where i and j refer to the initial and final states of the transition.

Electric-Dipole Selection Rules

Electric dipole transitions are possible only between certain atomic energy levels. Rules that tell us which transitions are allowed are called selection rules. We will consider only electric dipole transitions, i. e., "allowed" transitions. If electric dipole transitions are forbidden, transitions can still occur by other radiation processes such as higher-order multipole radiation (see Garstang) but these transitions are much weaker.

The parity operation $P\mathbf{r_i} = -\mathbf{r_i}$ commutes with H (i.e., with the complete atomic Hamiltonian, not just our approximation) and for a product wave function yields $P\Psi(\mathbf{r_i}) = (-1)^{\sum l_i}\Psi(\mathbf{r_i})$. If $\sum_{l_i}$ is (even/odd) we say the state is an (even/odd) parity state. In spectroscopic notation an upper righthand superscript "o" is sometimes used to indicate explicitly states having odd parity. Electronic dipole transitions *only* take place between states of *different* parity so that for the common case of a one-electron jump, $\Delta l = \pm 1$. No condition on Δn is required.

By arguments similar to those given to explain hydrogenic selection rules, we also find

$$\Delta S = 0 \text{ (no spin change)}, \quad \Delta J = 0, \pm 1 \ (0 \rightarrow 0 \text{ forbidden}),$$
$$\Delta L = 0, \pm 1, \quad \Delta M_J = 0, \pm 1.$$

Radiation from a $\Delta M_J = 0$ transition is linearly polarized ($\sin^2\theta$ dependence); radiation from a $\Delta M_j = \pm 1$ transition is circularly polarized [$(1 + \cos^2\theta)$ dependence] when viewed along the axis of space quantization.

The strength of a transition depends on the magnitude of the radial part of the $\mathbf{P}_{ij}$ integral, which is quite sensitive to the approximations used in finding ψ and is difficult to calculate accurately.

Examples of Simple Spectra

Alkali-like Spectra

Alkalis and ions with a single electron outside of closed subshells have energy levels and spectra very similar to hydrogen because the core electrons, not taking any role in optical transitions, act principally to screen the nuclear charge. The larger the n of the valence electron, the larger its orbit and so the more complete is the screening. The energy level diagram of sodium is shown in Fig. 1. The configurations responsible for these levels are:

$$(1s)^2(2s)^2(2p)^6 n\mathrm{s}\ {}^2\mathrm{S}_{1/2} \qquad n = 3, 4, \ldots,$$
$$(1s)^2(2s)^2(2p)^6 n\mathrm{p}\ {}^2\mathrm{P}_{1/2,3/2} \qquad n = 3, 4, \ldots,$$
$$(1s)^2(2s)^2(2p)^6 n\mathrm{d}\ {}^2\mathrm{D}_{3/2,5/2} \qquad n = 3, 4, \ldots,$$

Early spectroscopists called the $n\mathrm{d}\ {}^2\mathrm{D} \rightarrow 3\mathrm{p}\ {}^2\mathrm{P}$ transitions the diffuse series because in low resolution the three allowed lines, ${}^2\mathrm{D}_{5/2} \rightarrow {}^2\mathrm{P}_{3/2}$ and ${}^2\mathrm{D}_{3/2} \rightarrow {}^2\mathrm{P}_{3/2,1/2}$ appeared as a blend.

Two-Electron Spectra

The alkaline-earth elements Be, Mg, Ca, Sr, and Ba are representative of atoms with two electrons outside of closed shells. We shall consider Be, whose term diagram is shown in Fig. 2. The low-lying excited states arise from excitation of one of the ground state $(2s)^2$ electrons. The low-lying excited configurations are

$$(1s)^2 2s\, n\mathrm{s}\ {}^1\mathrm{S}_0\ {}^3\mathrm{S}_1, \qquad n = 3, 4, \ldots,$$
$$(1s)^2 2s\, n\mathrm{p}\ {}^1\mathrm{P}_0\ {}^3\mathrm{P}_{2,1,0}, \qquad n = 2, 3, 4,$$
$$(1s)^2 2s\, n\mathrm{d}\ {}^1\mathrm{D}_0\ {}^3\mathrm{D}_{3,2,1}, \qquad n = 3, 4, 5, \ldots$$

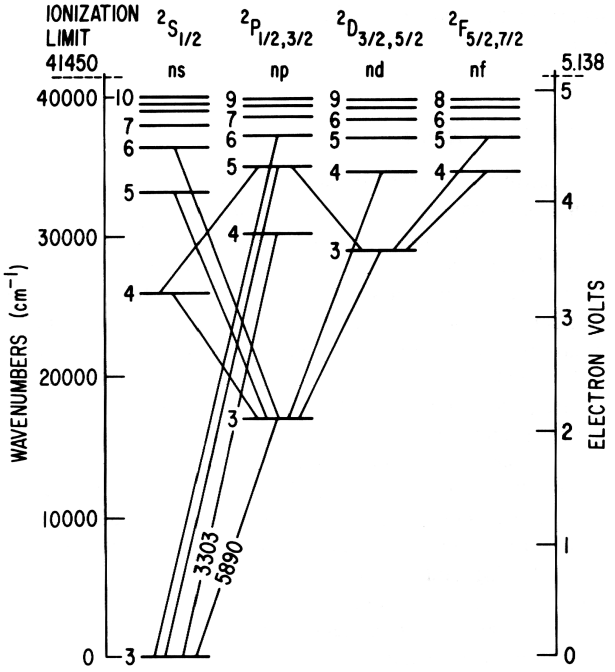

Fig. 1: Term diagram of sodium.

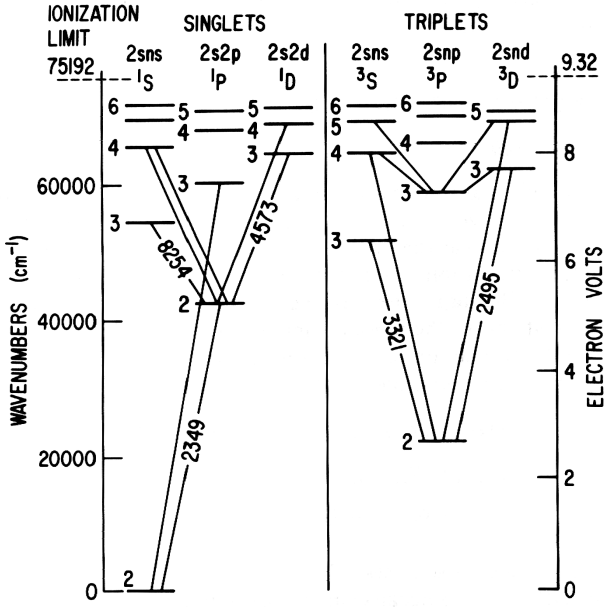

Fig. 2: Term diagram of beryllium.

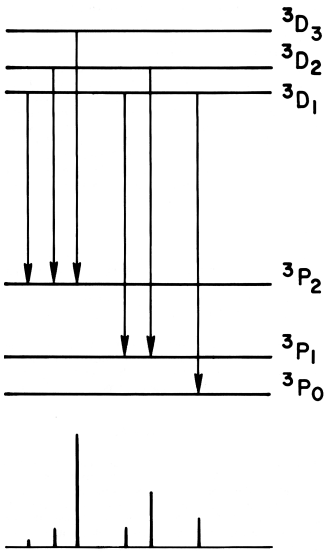

Fig. 3: Allowed ^{3}D–^{3}P transitions and relative line strengths.

In each configuration the $s = \frac{1}{2}$ spins of the two unpaired electrons are coupled to form the resultant total spin $S = 0, 1$. The L value is that of the excited electron. To illustrate: for the D states, S and L are combined to form the resultant J as follows: (for the singlet) $L = 2$, $S = 0$, $J = 2$; (for the triplets) $L = 2$, $S = 1$, $J = L+S, L+S-1, L-S = 3, 2, 1$. The $\Delta S = 0$ selection rule prohibits singlet to triplet transitions. The allowed ^{3}D to ^{3}P triplet transitions are shown in Fig. 3. The same methods can be applied to more complicated spectra (see White and Kuhn).

Quantum Defect Theory

A key feature of the attractive Coulomb field in which atomic electrons move is that it supports an infinite number of bound states which converge in energy to a particular ionization threshold. These states may be grouped in series. In simple cases each series may be identified by the term level of the ion to whose threshold the series converges, the orbital angular momentum of the excited electron, and the coupling of the electron to the atomic core. In general, the various series of states may interact with each other, thereby complicating the analysis in the very region where the number of levels is becoming infinite.

The quantum defect theory (QDT) is a method of using the analytically known properties of excited electrons moving in a pure Coulomb field to describe such atomic spectra in terms of a few parameters. These parameters may be determined either from experimental data or from *ab initio* theoretical calculations. In addition, they are usually nearly independent of energy in the threshold energy region (i. e., within a few electron volts of the atomic ionization threshold). Thus the determination of these parameters at any *single* energy suffices to predict the *variation with energy* of numerous atomic properties in the threshold energy region such as total and partial photoionization cross sections, photoelectron angular distributions, discrete line strengths, autoionization profiles, etc. These properties are often very strongly energy

dependent and difficult to measure or to calculate by other methods. Yet all these phenomena, according to the QDT, depend on only a few essential parameters which represent the proper interface between theory and experiment.

The QDT assumes that the configuration space for an excited atomic electron can be divided into two regions: an *inner region*, $0 \leq r \leq r_0$, where electron correlations are strong and difficult to treat, and an *outer region*, $r_0 \leq r \leq \infty$, where the electron–ion interaction potential is assumed to be purely Coulombic and where the form of the electron wave function is known analytically. The boundary radius r_0 between the two regions is typically of the order of the atomic radius.

Consider the simple problem of an excited electron of angular momentum l in an alkali atom: the electron sees a Coulomb field for $r \geq r_0$, where r_0 is roughly the ionic radius. We measure the energy ε of the excited electron relative to the ionization threshold as $\varepsilon = -0.5\nu^{-2}$, where the parameter ν is our measure of energy. The Schrödinger equation for $r \geq r_0$ has two solutions, one regular and one irregular for small values of r:

$$f(\nu, r) \sim r^{l+1} \quad \text{as} \quad r \to 0 \,, \tag{3a}$$

$$g(\nu, r) \sim r^{-l} \quad \text{as} \quad r \to 0 \,. \tag{3b}$$

A general solution of the Schrödinger equation for $r \geq R_0$ is a linear combination of $f(\nu, r)$ and $g(\nu, r)$ with coefficients to be determined by application of boundary conditions at infinity and at r_0. This general solution may be written as

$$\psi(\nu, r) = N_\nu \{ f(\nu, r) \cos \pi\mu - g(\nu, r) \sin \pi\mu \} \quad \text{for } r \geq r_0 \,, \tag{4}$$

where N_ν is a normalization factor that is determined by the behavior of $\psi(\nu, r)$ at large r. μ on the other hand, is the relative phase with which the regular and irregular solutions are superimposed. Its value is determined by the behavior of $\psi(\nu, r)$ in the core region, $0 \leq r \leq r_0$, where the effective potential is non-Coulombic: i. e., μ has that value which allows the analytically determined $\psi(\nu, r)$ given by Eq. (4) for $r \geq r_0$ to be joined smoothly at $r = r_0$ onto the numerically determined portion of $\psi(\nu, r)$ that obtains in the inner core region, $0 \leq r \leq r_0$.

Alternatively, μ may be determined semiempirically from atomic spectral data on energy levels, as we show here. Consider the asymptotic behavior of $\psi(\nu, r)$ in the case of excited electron energies below threshold, i. e., $\psi(\nu, r)$ must tend toward zero. The asymptotic forms of the regular and irregular Coulomb functions are

$$f(\nu, r) \to u(\nu, r) \sin \pi\nu - v(\nu, r) \exp[\mathrm{i}\pi\nu] \quad \text{as} \quad r \to \infty \,, \tag{5a}$$

$$g(\nu, r) \to -u(\nu, r) \cos \pi\nu + v(\nu, r) \exp\left[\mathrm{i}\pi\left(\nu + \tfrac{1}{2}\right)\right] \quad \text{as} \quad r \to \infty \,, \tag{5b}$$

where $u(\nu, r)$ is an exponentially increasing function of r and $v(\nu, r)$ is an exponentially decreasing function of r. Substituting Eq. (5) in Eq. (4) gives

$$\psi(\nu, r) \to N_\nu \{ u(\nu, r) \sin \pi(\nu + \mu) - v(\nu, r) \exp[\mathrm{i}\pi(\nu + \mu)] \} \quad \text{as} \quad r \to \infty \,. \tag{6}$$

In order that $\psi(\nu, r)$ tend toward zero at large values of r, the coefficient of $u(\nu, r)$ must be zero; i. e., $\sin \pi(\nu + \mu) = 0$ or $\nu + \mu = n$, where n is an integer. Substituting $\nu = n - \mu$, in the expression for the electron's energy gives (in atomic units)

$$\varepsilon_n = -\frac{1}{2\nu^2} = \frac{1}{2(n-\mu)^2} \,. \tag{7}$$

μ is thus the quantum defect of spectroscopy and may be determined directly from Rydberg energy level data for the alkalis.

It is well known empirically in atomic spectroscopy that the quantum defect μ is a nearly constant function of energy near the ionization threshold. Theoretically, μ *should* be only a slowly varying function of energy since it is determined from the wave function in the inner core region, where the electron's large instantaneous kinetic energy makes it insensitive to the relatively small energy differences between the energy levels near threshold. Knowledge of the parameter μ therefore enables one to predict the energies of a whole series of atomic energy levels according to Eq. (7), thereby illustrating the ability of QDT to describe large amounts of atomic spectral information in a very compact way.

QDT may also be used to understand the variation with energy of the intensities of atomic spectral levels. From Eq. (4) we can see directly that for small $r \geq r_0$, the wavefunction $\psi(\nu, r)$ depends on energy mainly through the normalization factor N_ν since μ is weakly energy dependent and so are $f(\nu, r)$ and $g(\nu, r)$ at small r [cf. Eq. (3)]. The normalization factor N_ν is determined by the asymptotic behavior of $\psi(\nu, r)$ and may be very energy dependent. The point of this discussion thus is that at small radii $\psi(\nu, r)/N_\nu$ is likely to be quite insensitive to energy in the threshold energy region. Yet it is in this small r region that spectral transitions between the ground and the excited states occur. We conclude that the intensities of these transitions along a series of lines converging to the ionization threshold should depend on the energies of the excited levels in proportion to N_ν^2. QDT shows that N_ν^2 is proportional to $\nu^{-3} = (n-\mu)^3$ for discrete (i. e., negative) electron energies. This implies that multiplication of the measured intensity of the nth level by $(n-\mu)^3$ will produce a renormalized intensity that is only a slowly varying function of n. Hence, accurate measurements of only a few level intensities allows one to determine this slowly varying function and therefore to predict the intensities of all other levels in the series.

The QDT may also be used to describe atomic spectra more complicated than those of the alkali metals as well as to relate an atom's discrete spectrum to collision processes occurring at energies above the atom's ionization threshold. These topics, however, are beyond the scope of this article. The interested reader is referred to review articles on QDT by Fano and by Seaton.

See also: Bohr Theory Of Atomic Structure; Hamiltonian Function; Rotation And Angular Momentum; Schrödinger Equation; Zeeman And Stark Effects.

Bibliography

E. U. Condon and G. H. Shortley, *Theory of Atomic Spectra*. Cambridge University Press, Cambridge, 1953. (A)

U. Fano, "Unified Treatment of Perturbed Series, Continuous Spectra, and Collisions," *J. Opt. Soc. Am.* **65**, 979 (1975).

R. H. Garstang, "Forbidden Transitions," in *Atomic and Molecular Processes*. Academic Press, New York, 1962.

G. Herzberg, *Atomic Spectra and Atomic Structure*, 2nd ed. Dover, New York, 1944. (E)

H. G. Kuhn, *Atomic Spectra*. Academic Press, New York and London, 1962. (E)

C. E. Moore, *Atomic Energy Levels*, Circular 467, Vols. I, II, and III. National Bureau of Standards, Washington, DC, 1949, 1952, 1958.

M. J. Seaton, "Quantum Defect Theory," *Rept. Prog. Phys.* **46**, 167 (1983).
B. W. Shore and D. H. Menzel, *Principles of Atomic Spectra*. Wiley, New York, 1968. (A)
J. C. Slater, *Quantum Theory of Atomic Structure*, Vols. 1 and 2. McGraw-Hill, New York and London, 1960. (Vol. 1, I; Vol. 2, A)
I. I. Sobelman, *Introduction to the Theory of Atomic Spectra*. Pergamon Press, Oxford, 1972. (I)
H. E. White, *Introduction to Atomic Spectra*. McGraw-Hill, New York and London, 1934. (E)

Atomic Structure Calculations, Electronic Correlation

R. K. Nesbet

The electronic wave function for noninteracting electrons would be a single Slater determinant, an antisymmetrized product of one-electron "orbital" wave functions. The Hartree–Fock model of interacting electrons [1] optimizes this model function by solving variational equations for the occupied orbital functions. A simple linear combination of Slater determinants is used when the state in question is not invariant under rotation. The Hartree–Fock equations are analogous to noninteracting Schrödinger equations, but contain a nonlocal "self-consistent-field" potential, due to Coulomb and exchange interactions between electrons. This is an independent-particle model, in which electron quasiparticles interact only indirectly through a self-consistent mean field. The electrostatic energy depends on correlations between the locations of the electrons and cannot be computed exactly in the Hartree–Fock approximation [2].

The expression "electronic correlation" refers generally to corrections to the Hartree–Fock approximation. For nonspherical atomic states, the self-consistent potential function is usually spherically averaged, resulting in one-electron "correlation" associated with spin or rotational symmetry-breaking [3]. Otherwise the principal correlation effect in atoms is short-range Coulomb-cusp relaxation due to the electronic Coulomb repulsion. In molecules and solids this is supplemented by the long-range correlation effect of multipole polarization response.

One-electron mean-value properties computed in the closed-shell Hartree–Fock approximation are subject only to second-order correlation corrections [2, 3]. The practical effect of this is that the total electronic density distribution is well described for many purposes in the Hartree–Fock approximation, but reliable and consistent theoretical results for atomic properties sensitive to open-shell structure or to the response to external perturbations require a quantitative treatment of electronic correlation. Such properties include hyperfine structure, polarizabilities, oscillator strengths, and electron scattering cross sections [4, 5].

While some physical properties are described more or less accurately in the Hartree–Fock approximation, others are not described at all. For example, the van der Waals or dispersion potential energy of two spatially separated electronic systems is simply the long-range limit of the correlation energy of spatially separated electrons [6]. Orbital functional theory (OFT) [7] extends Hartree–Fock to a mean-field model that incorporates correlation. The polarization potential that dominates low-energy electron scattering is the asymptotic form of the nonlocal OFT correlation potential. In the usual Hartree–Fock approximation, the magnetic hyperfine structure constant is zero by symmetry in the ground states of nitrogen and phosphorus. Observed values differ from zero because of a combination of spin symmetry-breaking and Coulombic correlation effects [4].

Theoretical methods for the computation of correlation effects are often based on a preliminary Hartree–Fock calculation, or on a mean-field model such as density functional theory (DFT) [8], which includes an approximation to both exchange and correlation. OFT [7] optimizes the model or reference state for exact exchange and correlation within some many-body formalism. The optimized reference state [7] defines a "vacuum state" for formal perturbation theory [9]. A shell model is appropriate for atomic electrons because the nuclear attraction dominates the self-consistent radial potential. Occupied orbitals for a particular state in the Hartree–Fock approximation are labeled by quantum numbers n, l, m_l, m_s (or n, j, m) appropriate to the one-electron Schrödinger equation for a central potential. The conventional open-shell Hartree–Fock equations [1] are spherically averaged so that for each value of orbital angular momentum l there is a set of orthonormal radial functions $R_{nl}(r)$, not dependent on the axial quantum numbers m_l and m_s. A "configuration" is defined by a set of occupation numbers $d_{nl} \leq 2(l+1)$ which assign occupied orbital functions to subshells of given n, l but arbitrary m_l, m_s, subject to

$$m_l = -l,\ -l+1,\ \ldots,\ l; \qquad m_s = -\frac{1}{2},\ \frac{1}{2}\ . \tag{1}$$

Conventional Hartree–Fock theory is formulated in terms of eigenfunctions of total orbital angular momentum and spin constructed from the Slater determinants of a specified configuration.

The set of radial functions R_{nl} for occupied orbitals in a reference configuration can be extended to a complete orthonormal set. This generates a complete basis for the atomic N-electron wave function as a hierarchy of virtual excitations, defined by substitution of "unoccupied" orbitals (from the extended set) for orbitals occupied in a particular configuration. An n-electron virtual excitation is defined by n such substitutions [7].

Since the electronic Hamiltonian contains only one- and two-electron operators, only one- and two-electron virtual excitations contribute to the first-order wave function of perturbation theory. The closed-shell Hartree–Fock approximation causes one-electron matrix elements with the reference state to vanish [2], but one-electron virtual excitations cannot be neglected for open-shell states. The total energy or correlation energy can be expressed exactly in terms of the coefficients of all one- and two-electron virtual excitations of the reference state [7]. Valid estimates of these coefficients can be made either by formal perturbation theory [9], or by approximate solution of the matrix eigenvalue problem defined in the basis of all virtual excitations of all orders $n \leq N$, for N electrons. The latter method, superposition of configurations or "configuration interaction," has been widely applied and highly developed in its computational and data-handling aspects [10]. The direct use of relative coordinates for pairs of electrons is not computationally feasible for more than two or three electrons, although it has given very accurate results for two-electron atoms and ions [11].

Because of the great complexity of such calculations for virtual excitations with $n > 2$, several methods have been introduced for approximate incorporation of terms of higher order [10]. These methods either treat electron pairs as uncoupled from each other [12] or modify the variational equations for such separated pairs to allow for the higher-order virtual excitations implied by a cluster expansion of the N-electron wave function [13]. This level of approximation appears to be the most natural extension of theory beyond Hartree–Fock to include electronic pair correlation. In the multiconfiguration Hartree–Fock method,

configuration interaction is incorporated into an iterative variational calculation. A numerical version of this method has been used for accurate calculations of atomic oscillator strengths and photoionization cross sections [14].

These methods have been tested by numerous calculations of total atomic energies or of excitation energies [5], primarily for atoms in the first third of the periodic table. Their most important application, however, has been to the calculation of physical atomic properties that are difficult or impossible to measure experimentally. Theoretical calculations have helped to resolve discrepancies between conflicting experimental data on oscillator strengths [5], and have provided values of polarizabilities, particularly for atomic excited states, that have not been measured.

Hyperfine-structure calculations of high accuracy have helped to establish the fact that electronic correlation affects the three tensorially distinct magnetic hyperfine-structure interaction operators differently, so that three independent parameters must be used to fit experimental data. Theoretical calculations of the electric field gradient at a nucleus, which cannot be measured directly, have made it possible to obtain accurate nuclear quadrupole moments from measured quadrupole hyperfine coupling constants [4, 15].

Low-energy electron scattering by neutral atoms is dominated by the electric dipole polarization potential, essentially a correlation effect. Theoretical calculations by methods that can describe this effect quantitatively have been carried out for electron scattering by hydrogen, helium, alkali metals, and several other atoms [16]. These theoretical results have helped to elucidate observed structural features (resonance and threshold structures) and to establish absolute values of cross sections.

See also: Atomic Structure Calculations, One-Electron Models; Atomic Structure Calculations, Relativistic Atoms; Fine and Hyperfine Spectra and Interactions.

References

[1] D. R. Hartree, *The Calculation of Atomic Structures*. Wiley, New York, 1957. C. Froese Fischer, *The Hartree–Fock Method for Atoms*. Wiley-Interscience, New York, 1977.

[2] C. Møller and M. S. Plesset, *Phys. Rev.* **46**, 618 (1934); L. Brillouin, *Les Champs "Self-Consistent" de Hartree et de Fock*. Hermann et Cie, Paris, 1934.

[3] R. K. Nesbet, *Proc. Roy. Soc. (London)* **A230**, 312 (1955).

[4] N. C. Dutta, C. Matsubara, R. T. Pu, and T. P. Das, *Phys. Rev. Lett.* **21**, 1139 (1968); *Phys. Rev.* **177**, 33 (1969).

[5] A. Hibbert, *Rept. Prog. Phys.* **38**, 1217 (1975).

[6] F. London, *Z. Phys.* **63**, 245 (1930).

[7] R. K. Nesbet, *Adv. Chem. Phys.* **9**, 321 (1965); R. K. Nesbet, *Variational Principles and Methods in Theoretical Physics and Chemistry*. Cambridge Univ. Press, New York, 2003.

[8] W. Kohn and L. J. Sham, *Phys. Rev.* **140**, A1133 (1965); R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*. Oxford Univ. Press, New York, 1989.

[9] H. P. Kelly, *Adv. Chem. Phys.* **14**, 129 (1969); I. Lindgren and J. Morrison, *Atomic Many-Body Theory*. 2nd ed., Springer-Verlag, Berlin, 1986.

[10] I. Shavitt, in *Methods of Electronic Structure Theory*, H. F. Schaefer III, ed., p. 189. Plenum, New York, 1977; A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry*. McGraw-Hill, New York, 1989.

[11] C. L. Pekeris, *Phys. Rev.* **115**, 1216 (1959); **126**, 1470 (1962); K. Frankowski and C. L. Pekeris, *Phys. Rev.* **146**, 46 (1966).
[12] O. Sinanoglu, *Adv. Chem. Phys.* **6**, 315 (1964); **14**, 237 (1969); R. K. Nesbet, *Adv. Chem. Phys.* **14**, 1 (1969).
[13] J. Cizek, *Adv. Chem. Phys.* **14**, 35 (1969); J. Cizek and J. Paldus, *Int. J. Quantum Chem.* **5**, 359 (1971); W. Meyer, *J. Chem. Phys.* **58**, 1017 (1973); W. Kutzelnigg, in *Methods of Electronic Structure Theory*, H. F. Schaefer III, ed., p. 127. Plenum, New York, 1977; W. Meyer, in *Methods of Electronic Structure Theory*, H. F. Schaefer III, ed., p. 413. Plenum, New York, 1977; R. J. Bartlett, *Ann. Rev. Phys. Chem.* **32**, 359 (1981).
[14] C. Froese Fischer, *Comput. Phys. Commun.* **14**, 145 (1978); H. P. Saha and C. Froese Fischer, *Phys. Rev.* **A35**, 5240 (1987); H. P. Saha, C. Froese Fischer, and P. W. Langhoff, *Phys. Rev.* **A38**, 1279 (1988).
[15] J. D. Lyons, R. T. Pu, and T. P. Das, *Phys. Rev.* **178**, 103 (1969); R. K. Nesbet, *Phys. Rev.* **A2**, 661 (1970); J. D. Lyons and T. P. Das, *Phys. Rev.* **A2**, 2250 (1970); R. K. Nesbet, *Phys. Rev. Lett.* **24**, 1155 (1970).
[16] B. L. Moiseiwitsch, *Rept. Prog. Phys.* **40**, 843 (1977); R. K. Nesbet, *Variational Methods in Electron-Atom Scattering Theory*. Plenum, New York, 1980.

Atomic Structure Calculations, One-Electron Models

F. Herman

The one-electron theory of atoms, molecules, and solids [1, 2] has enjoyed wide success in many branches of physics and chemistry. This theory postulates that the exact wave function for a many-electron system can be represented accurately by an approximate many-electron wave function constructed from one-electron wave functions or spin orbitals. Emphasis shifts from a consideration of complex many-electron wave functions depending on the coordinates of all the electrons, to one-electron wave functions. These are easier to treat because they depend only on the spatial and spin coordinates of single electrons. The ground state of the many-electron system can then be described in terms of the occupied spin orbitals, and excitations in terms of transitions between occupied and unoccupied spin orbitals.

The one-electron theory describes the electronic structure and related physical and chemical properties of many-electron systems. The theory is widely used in atomic spectroscopy [3] and many other applications as a conceptual tool. In addition, the theory offers a convenient and systematic framework for carrying out detailed numerical calculations for atoms [4–6]. In such calculations, it is usually necessary to introduce many simplifying assumptions, both physical and mathematical, to make progress.

For some applications, where gross electronic properties are of primary interest, very crude phenomenological models can be adopted. In other instances, where the effects are subtle, it is necessary to employ highly sophisticated theoretical models [7, 8], even though their use leads to extensive numerical computation. Fortunately, large-capacity high-speed electronic digital computers are widely available, as are atomic structure computer codes [5, 9, 10]. Extensive tabulations of the results of atomic structure calculations are also readily at hand [5, 10, 11].

Considerable effort has been devoted to perfecting atomic structure calculations, not only with a view to studying a wide variety of atomic properties, but also as a starting point for computer modeling and simulation in physics, chemistry, biology, and materials science. Such investigations have become increasingly important in industrial and government settings as well as in academe.

A notable characteristic of atomic structure calculations is the wide spectrum of available methods and points of view: different approximations are advantageous for different applications. In this introductory sketch we can hardly do justice to the many ingenious techniques that are currently in use for treating many-electron systems starting with atoms. We will provide a general perspective and some useful references.

Hartree and Hartree–Fock Methods

In orbital theories of many-electron systems, one derives a set of one-electron wave equations by using the variational method. The form of these equations is determined by the manner in which the exact many-electron wave function Φ is represented by one-electron wave functions or spin orbitals ϕ_q where q denotes all the relevant quantum numbers. By using the variational method, one guarantees that the solutions of the one-electron wave equations, the ϕ_q, are the best possible consistent with the assumed form of Φ.

In the Hartree (H) approximation, Φ is represented by a simple product of spin orbitals, ϕ_q, each factor corresponding to a different occupied state q. In the Hartree–Fock (HF) approximation, Φ is represented by an antisymmetrized product of the ϕ_q also known as a Slater determinant and a determinantal wave function. Both treatments lead to an independent-particle model in which each of the electrons moves independently of all the others in a time-averaged potential field produced by all the other electrons and the nucleus. As a result of this time-averaging, spatial correlations in the motions of pairs of electrons produced by their instantaneous Coulomb repulsion are neglected.

In the HF approximation, Φ is represented by a determinantal wave function to take into account the fact that a many-electron wave function must be an antisymmetric function of the electron coordinates. This feature is closely related to the requirement that electrons satisfy the Fermi–Dirac statistics and the Pauli exclusion principle. In contrast to the H approach, the use of determinantal wave functions in the HF method leads to additional terms in the one-electron wave equations, the exchange terms. These tend to keep electrons of like spin out of each other's way, as required by the Pauli exclusion principle. Thus, the HF approximation includes a certain type of spatial correlation between like-spin electrons. This is called exchange and also statistical correlation because it arises from the Fermi–Dirac statistics.

Spatial correlation over and above statistical correlation is usually described simply as correlation. Thus, it can be said that the H approximation neglects both exchange and correlation, while the HF approximation includes exchange but neglects correlation. HF calculations were originally favored over H calculations because they included exchange effects. However, we now realize that exchange effects are to some degree offset by correlation effects. Accordingly, it is best to include exchange and correlation effects together, especially for determining delicate electronic properties. [7, 8].

In the following, we first discuss the essentials of atomic structure calculations within the context of the traditional H and HF methods and the simplified HFS method. Next, we turn to

the density functional method that allows simultaneous incorporation of exchange and correlation effects and greatly improves the physical and chemical accuracy. Finally, we examine the pseudopotential method that provides further simplifications leading to significant reductions in computational effort and the possibility of treating very large systems.

Self-Consistent Iteration and Central-Field Approximation

Since the potential terms appearing in the H and HF one-electron wave equations depend on the wave functions solving these equations, it is necessary to treat these equations by iterative techniques. One chooses an initial set of wave functions and inserts these into the potential terms. Solving the wave equations, one obtains a final set of wave functions for this cycle. The next cycle begins when we use a new set of starting wave functions constructed from a suitable average of the initial and final wave functions of the preceding cycle. This averaging represents a tradeoff between rapid convergence and numerical stability. The process continues until the initial and final wave functions in a given cycle agree with one another to some specified accuracy. In this way we obtain a self-consistent solution in the sense that the wave functions are generated by wave equations whose potential terms are determined by these very wave functions.

One can simplify the self-consistent field equations for an atomic system by taking advantage of the spherical symmetry of the atomic field. One introduces the central field approximation by considering only the spherically averaged component of the atomic field. The complete three-dimensional wave equation can then be separated into an angular wave equation and a radial wave equation by writing the spin-orbital as the product of a radial function $R(r)$, an angular function $Y(\theta,\varphi)$, and a spin function. The angular functions are the well-known spherical harmonics $Y_{lm}(\theta,\varphi)$, where l and m are the azimuthal and magnetic quantum numbers characteristic of a central field. The radial functions, $R_{nls}(r)$, in general depend on n, l, and s, where n and s are the principal and spin quantum numbers. The $R_{nls}(r)$ can be determined by solving the radial wave equations numerically [4, 5] or by the expansion method [6]. In the expansion method, the $R_{nls}(r)$ and all the potential terms are expanded in terms of a suitably chosen set of analytic functions, while the wave equations are solved by matrix methods.

In closed-shell atoms, it is possible to represent Φ by a single determinantal wave function. For the more general case of open shell atoms, it is usually necessary to represent Φ by a linear combination of determinantal wave functions. Each corresponds to a different assignment of the one-electron quantum numbers $(q = n,l,m,s)$ compatible with the assumed overall symmetry of the many-electron atom. This leads to interactions over and above the central field approximation, and to multiplet structure that is of great importance in atomic spectroscopy [3].

Correlation effects can be included by expanding the many-electron wave function Φ as a linear combination of determinantal wave functions, each representing a different electronic configuration. In this approach, known as configuration interaction (CI), it is usually necessary to include large numbers of configurations to insure an accurate description of correlation effects. Apart from dealing with the intricacies of multi-configuration calculations, much of the underlying theoretical effort in quantum chemistry has been devoted to the development of efficient basis sets for representing atomic and molecular orbitals [12, 13]. The same is true in solid-state physics.

CI with optimized basis sets is very popular with quantum chemists. It is widely used and has had considerable success for atoms and moderately small molecules [12, 13]. However, CI is not practical for large molecules and crystals because of the prohibitive amount of computational effort required.

Hartree–Fock–Slater Equations

The HF equations can be simplified by averaging the HF exchange potentials over all occupied states and then using a free-electron model to determine the averaged exchange potential [14]. The exchange potential for any many-electron system can then be approximated at any point $\mathbf{r}$ by the value of the exchange potential $V_{\text{exch}}(\mathbf{r})$ of a free-electron gas having an electronic charge density ρ equal to $\rho(\mathbf{r})$. This is known as the free-electron exchange approximation. This procedure avoids the non-local potentials $V_{\text{exch}}(\mathbf{r},\mathbf{r}')$ inherent in the HF method, at the same time eliminating the need to solve the HF equations for each orbital separately. Attention now focuses on a one-electron wave equation, the Hartree–Fock–Slater (HFS) equation, that contains the averaged exchange potential [1, 5, 14] and is the same for all orbitals.

For atomic systems containing equal numbers of electrons with spin up and spin down (balanced spins), the radial HFS equations take the form

$$\left(-\frac{\mathrm{d}^2}{\mathrm{d}r^2}-\frac{2Z}{r}+V_{\text{coul}}(r)+V_{\text{exch}}(r)+\frac{l(l+1)}{r^2}\right)P_{nl}(r)=E_{nl}P_{nl}(r)\ , \tag{1}$$

where the energy eigenvalues are denoted by E_{nl} and the radial eigenfunctions by $P_{nl}(r) = rR_{nl}(r)$. We will measure distances in Bohr units ($1\,\text{Bohr} = 0.529\,\text{Å}$) and energies in Rydberg units ($1\,\text{Ry} = 13.6\,\text{eV}$). The first term on the LHS is the kinetic energy operator. The second is the nuclear Coulomb potential, with Z denoting the nuclear charge. Next, $V_{\text{coul}}(r)$ is the spherically averaged electronic Coulomb potential, $V_{\text{exch}}(r)$ the free-electron exchange potential, and the last term the centrifugal potential.

The radial wave functions $P_{nl}(r)$ must vanish at the origin and at infinity. They have $n-l-1$ nodes between the origin and infinity. The azimuthal quantum number l ranges from 0 to $n-1$, and the magnetic quantum number m from $-l$ to $+l$. With the radial wave functions normalized, $\int[(P_{nl}(r)]^2\,\mathrm{d}r = 1$, the spherically averaged electronic charge density $\rho(r)$ and Coulomb potential $V_{\text{coul}}(r)$ are

$$\rho(r) = -(4\pi r^2)^{-1}\sum_{nl}\omega_{nl}[P_{nl}(r)]^2 \tag{2}$$

and

$$V_{\text{coul}}(r) = -(8\pi/r)\int_0^r \rho(t)\,t^2\,\mathrm{d}t - 8\pi\int_r^\infty \rho(t)\,t\,\mathrm{d}t\ , \tag{3}$$

where $\omega_{nl}(r)$ is the occupation number for the orbital nl (both spins). In the special case of a closed shell, $\omega_{nl} = 2(2l+1)$. The total number of electrons in the atom is $N = \sum_{nl}\omega_{nl}$, and the ionicity is $Z-N$.

Slater's exchange potential based on his free-electron model can be written as

$$V^{\alpha}_{\text{exch}}(r) = -6\alpha[(3/8\pi)|\rho(r)|]^{1/3}\ , \tag{4}$$

where α is a parameter whose value is 1 according to Slater's intuitive variational derivation [14].

In HF theory, the energy eigenvalues E_{nl} represent one-electron ionization energies (Koopmans' theorem). This is not the case for HFS theory, where ionization and excitation energies are determined by the transition state method [15].

The self-interaction correction, describing the Coulomb interaction of an electron with itself, is taken into account properly in the HF, but not in the HFS method. Consequently, HFS potentials are flawed at large distances from the nucleus. In the interior of non-magnetic solids, where atoms lie close together, this feature is less important than in atoms and molecules. Magnetic solids require special treatment. In spite of its simplified exchange model and the neglect of self-interaction corrections, the HFS method gained favor over the HF method because it usually led to improved ionization and excitation energies and reduced computational effort.

In an early study, the HFS equations were solved with $\alpha = 1$ for all normal neutral atoms in the periodic table [5]. Calculated energy eigenvalues E_{nl} and radial eigenfunctions $P_{nl}(r)$ were tabulated and the computer codes listed. Because of their simplicity and the favorable agreement with experiment obtained, the Herman–Skillman atomic structure codes were and still are widely used for atomic, molecular, and solid state problems [10, 16], and for pedagogical purposes as well [17]. Over time the codes have remained substantially the same except for the incorporation of improved exchange-correlation potentials [10, 21].

Kohn and Sham [18] took issue with Slater's use of $\alpha = 1$ because their variational calculations led to a value of $\alpha = 2/3$. (Note that Eq. 4 with $\alpha = 2/3$ is known as the Kohn–Sham equation.) However, atomic results for $\alpha = 2/3$ often departed more from experiment than those for $\alpha = 1$. This led to the Xα method, where X stands for exchange, and the value of α is optimized for each atom [19]. This pragmatic approach for atoms led to conceptual difficulties in applications to molecules and solids, where it is more reasonable to use the same value of α in all regions of space than different values in different atomic regions.

The significance of the Xα approximation was clarified by HFS atomic calculations that included a charge-density-gradient correction to the exchange potential [20]. The gradient-corrected exchange potential was represented by the expression

$$V_{\text{exch}}^{\alpha\beta}(r) = -6\,[\alpha + \beta G(\rho)]\,[(3/8\pi)|\rho(r)|]^{1/3}, \tag{5}$$

where the dimensionless function $G(\rho)$ is defined as

$$G(\rho) = \frac{1}{\rho^{2/3}}\left[\frac{4}{3}\left(\frac{\nabla\rho}{\rho}\right)^2 - \frac{2\nabla^2\rho}{\rho}\right]. \tag{6}$$

This form was determined by dimensional analysis. Here α and β are variational parameters that were determined by minimizing the total energy for each of a large set of representative atoms using Slater's average-over-configurations method [1, 2]. In this so-called X$\alpha\beta$ approximation, the optimum value of α was found to be equal to 2/3 for all atoms investigated, demonstrating that the Z-dependence of α in the Xα method is due to the neglect of gradient corrections. (The optimal value of β was nearly independent of Z.)

Since this study was based on minimizing the total energy, and most of the total energy resides in the inner shells and very little in the outer shells, the treatment of inner shell electrons

was improved, but that for outer shell electrons was not. These early calculations left much to be desired, but they stimulated many subsequent efforts to improve the treatment of gradient corrections in atoms.

The good agreement between theory and experiment found earlier using the HFS method with $\alpha = 1$ [5] is due in part to the fortuitous cancelation of neglected correlation effects and enhanced free-electron exchange arising from the use of $\alpha = 1$ instead 2/3 (cf. Ref. [21], p. 165).

Density Functional Theory and the Kohn–Sham Equations

The underlying idea of density functional theory [18, 22–25] is that the total energy of a many-electron system is determined by a functional of the charge density $\rho(r)$, $E[\rho(r)]$. From this functional it is possible to derive a potential function $V(r)$ that determines the one-electron wave functions and energies of the system. The theory does not provide an algorithm for determining this functional, but we are assured that the use of improved functionals will result in more accurate descriptions of electronic structure.

The density functional approach to many-electron systems leads to one-electron wave functions and methods of solution analogous to those for the HFS method, with the important exception that the Kohn–Sham exchange potential $V_{\text{exch}}^{\alpha=2/3}(\text{r})$ is replaced by a local potential $V_{\text{exch}}^{\text{corr}}(r)$ representing exchange and correlation effects. This potential can be derived from many-electron theories of the free-electron gas [26, 27]. A sophisticated version of the local density approximation (LDA), the screened-exchange plus Coulomb hole approximation [26], forms the basis of the GW method that is used in solid-state physics to determine energy band structure features with great accuracy [28].

The search for improved density-functional exchange-correlation potentials has been an active field of research for many years and continues to thrive. Some notable advances include generalized gradient approximations (GGA) [29]; semi-empirical schemes where parameters are adjusted to experiment [30]; and orbital-dependent potentials (OEP) [31]. Treatments of self-interaction corrections [32], relativistic effects [33], and spin polarization [34] have also been implemented. These and many other developments have improved the accuracy of density functional calculations considerably [21] (cf. Part II). However, the goal of achieving "chemical accuracy" of about 1 kcal/mole ($= 0.0434$ eV) for ionization, excitation, and relative energies remains elusive.

Pseudopotential Theory

Although the idea of pseudopotentials is very old, dating back to Fermi, recent developments began with a seminal paper by Herring [35] dealing with crystals, though his ideas also apply to atoms and molecules. Herring showed that orthogonalizing the wave functions of the outer shell electrons to those of the inner shells introduces a repulsive term in the wave equation that partially cancels the remaining electronic terms. This process leaves a residual effective potential or pseudopotential that is considerably smoother and weaker than the original "all-electron" potential. Morever, orthogonalization removes most of the nodal structure in outer shell electron wave functions, so that they can be expanded in relatively compact basis sets (orthogonalized plane waves). The use of pseudopotentials greatly increases the number of atoms that can be included in a molecular cluster and the range of problems that can be

treated. Combined with density functional theory, the pseudopotential method paves the way for realistic computer simulation and modeling in many fields of science and technology. [36]

The trail from Herring's work [35] to present-day pseudopotential theory runs as follows: In 1952, the first realistic calculation of the band structure of the diamond crystal [37] was carried out using Herring's OPW method. Seeing the detailed numerical results, Phillips noted the remarkable cancelation brought about by orthogonalization in semiconductors. This led Phillips and Kleinman [38] to devise a practical method for replacing the complicated orthogonality terms by simpler, more easily managed analytical forms. Independently, Harrison [39] pioneered the use of pseudopotentials in the study of the Fermi surface of metals. Further progress included understanding pseudopotential theory at a deeper level [25, 40], and developing efficient semi-empirical [41] and first-principles pseudopotentials [42]. For a comprehensive account of the current state of the theory, see Ref. [21], Chap. 11.

Concluding Remarks

Although we have emphasized atomic structure calculations here, appreciable cross-fertilization has taken place among atomic, molecular, and solid-state calculational studies. Methods for dealing with central field problems in atoms have naturally found their way into studies of molecules and crystals. Extensive investigations of efficient basis sets in molecular calculations provided valuable experience for the subsequent development of sophisticated basis sets for solids. Concepts and methods such as pseudopotentials and density functionals that originated in solid state physics now become increasingly important in atomic and molecular investigations.

Purely theoretical calculations continue to shed insight on complex problems. At the same time, numerical studies have expanded almost exponentially owing to the availability of more powerful computers and the pressing need for detailed results for practical problems, as opposed to analytical results for model problems. Perhaps this trend reflects John C. Slater's point of view (private communication), namely, that you don't really understand physical or chemical properties unless you can actually calculate them. Others like Charles Kittel and Sir Nevill Mott (private communications) took the opposite view, arguing that "back-of-the-envelope" calculations giving order-of-magnitude estimates and essential insights are preferable. Apart from this dichotomy, computational physicists and chemists should bear in mind that their calculations will in all likelihood be improved by future investigators using more powerful algorithms and computational tools, while mathematical theorems proved today will remain true forever [43].

See also: Atomic Structure Calculations, Electronic Correlation; Fine and Hyperfine Spectra and Interactions, Solid-State Physics.

References

[1] J. C. Slater, *Quantum Theory of Atomic Structure*, Vols. 1 and 2. McGraw–Hill, New York, 1960.

[2] J. C. Slater, *Quantum Theory of Molecules and Solids*, Vols. 1 to 4. McGraw–Hill, New York, 1963, 1965, 1967, 1974.

[3] I. I. Sobelman, *Introduction to the Theory of Atomic Spectra*. Pergamon Press, Oxford, 1972; R. D. Cowan, *The Theory of Atomic Structure and Spectra*. University of California Press, Berkeley, 1981; B. R. Judd, *Rept. Prog. Phys.* **48**, 907 (1985).

[4] D. R. Hartree, *The Calculation of Atomic Structures*. Wiley, New York, 1957. For a personal account of Hartree's work in atomic physics, see B. S. Jeffreys, *Comments Atom. Molec. Phys.* **20**, 189 (1987); C. Froese Fischer, *Douglas Rayner Hartree: His Life in Science and Computing*. World Scientific, Singapore, 2004.
[5] F. Herman and S. Skillman, *Atomic Structure Calculations*. Prentice-Hall, Englewood Cliffs, NJ, 1963.
[6] C. Froese-Fischer, *The Hartree-Fock Method for Atoms*. Wiley-Interscience, New York, 1977.
[7] A. Hibbert, Rept. *Prog. Phys.* **38**, 1217 (1975); I. Lindgren and J. Morrison, *Atomic Many-Body Theory*, 2nd ed. Springer-Verlag, Berlin, 1986.
[8] R. K. Nesbet, "Atomic Structure Calculations, Electronic Correlation" (this volume).
[9] Many atomic structure codes are available through *Computer Physics Communications Program Library*, `www.cpc.cs.qub.ac.uk/cpc` and *Quantum Chemistry Program Exchange*, `www.qcpe.indiana.edu`.
[10] S. Kotochigova, Z. Levine, E. Shirley, M. Stiles, and C. Clark, *Phys. Rev. A* **55**, 191 (1997); *ibid*, 5191E. Computer codes are available from these authors: `physics.nist.gov/PhysRefData/DFTdata/contents/html`.
[11] For atomic tabulations, see the internet URL in the previous reference. A wide variety of specialized tabulations appear in *Atomic Data and Nuclear Data Tables*.
[12] J. A. Pople, Nobel Prize lecture, *Rev. Mod. Phys.* **71**, 1267 (1999).
[13] P.-O. Löwdin, "Molecular Structure Calculations" (this volume).
[14] J. C. Slater, *Phys. Rev.* **81**, 385 (1951).
[15] J. C. Slater, *Adv. Quantum Chem.* **6**, 1 (1972); see also Ref. [2], Vol. IV, Sec. 2–5.
[16] K. H. Johnson, *Adv. Quantum Chem.* **7**, 143 (1973); *Ann. Rev. Phys. Chem.* **26**, 39 (1975); F. Herman, A. R. Williams, and K. H. Johnson, *J. Chem. Phys.* **61**, 3508 (1974); see also Ref. [2], Vol. IV.
[17] C. M. Quinn, *Computational Quantum Chemistry – An Interactive Guide to Basis Set Theory*. Academic Press, San Diego, 2002.
[18] W. Kohn and L. J. Sham, *Phys. Rev.* **140**, A1133 (1964).
[19] K. Schwarz, *Phys. Rev.* **B5**, 2466 (1972); see also Ref. [2], Vol. IV.
[20] F. Herman, I. B. Ortenberger, and J. P. Van Dyke, *Phys. Rev. Lett.* **22**, 807 (1969); *Intern. J. Quant. Chem.* **3S**, 827 (1970).
[21] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, Cambridge, 2004. For supplementary material, see http://mcc.uiuc.edu/estructure.
[22] P. Hohenberg and W. Kohn, *Phys. Rev.* **136**, B864 (1963); N. D. Mermin, *Phys. Rev.* **137**, A1441 (1965).
[23] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, Oxford, 1989; R. O. Jones and O. Gunnarsson, *Rev. Mod. Phys.* **61**, 689 (1989); W. Koch and M. C. Holthausen, *Chemist's Guide to Density Functional Theory*, 2nd ed. John Wiley, New York, 2004.
[24] W. Kohn, Nobel Prize lecture, *Rev. Mod. Phys.* **71**, 1253 (1999).
[25] C. Herring, "Solid State Physics" (this volume).
[26] L. Hedin, *Phys. Rev.* **39**, A796 (1965); L. Hedin and S. Lundqvist, *Solid State Phys.* **23**, 1 (1969).
[27] L. Hedin and B. I. Lundqvist, *J. Phys. C* **4**, 2064 (1971); U. von Barth and L. Hedin, *J. Phys. C* **5**, 1629 (1972).
[28] M. S. Hybertsen and S. G. Louie, *Phys. Rev.* **34**, 5390 (1986); **35**, 5585, 5602 (1987).
[29] J. P. Perdew and Y. Wang, *Phys. Rev. B* **45**, 13244 (1992); J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
[30] A. D. Becke, *Phys. Rev. A* **38**, 3098 (1988); C. Lee, W. Yang, and R. G. Parr, *Phys. Rev. B* **37**, 785 (1988).

[31] D. M. Bylander and L. Kleinman, *Intern. J. Mod. Phys.* **10**, 399 (1996).
[32] J. P. Perdew and A. Zunger, *Phys. Rev. B* **23**, 5048 (1981); A. Svane and O. Gunnarsson, *Phys. Rev. B* **37**, 9919 (1988).
[33] J. T. Waber, "Atomic Structure Calculations, Relativistic Atoms" (this volume).
[34] O. Gunnarsson and B. I. Lundqvist, *Phys. Rev. B* **13**, 4274 (1976).
[35] C. Herring, *Phys. Rev.* **57**, 1169 (1940).
[36] M. L. Cohen, *Annu. Rev. Mater. Sci.* **30**, 1 (2000).
[37] F. Herman, *Phys. Rev.* **88**, 1210 (1952).
[38] J. C. Phillips and L. Kleinman, *Phys. Rev.* **116**, 287 (1959).
[39] W. A. Harrison, *Pseudopotentials in the Theory of Metals*. W. A. Benjamin, New York, 1966.
[40] M. L. Cohen and V. Heine, *Solid State Physics* **24**, 1 (1970).
[41] M. L. Cohen and T. K. Bergstresser, *Phys. Rev.* **141**, 1979 (1966).
[42] D. R. Hamann, M. Schlüter, and C. Chiang, *Phys. Rev. Lett.* **43**, 1494 (1979); G. B. Bachelet, D. R. Hamann, and M. Schlüter, *Phys. Rev. B* **26**, 4199 (1982).
[43] See comments by George E. Kimball to the author in F. Herman, *Phys. Today* **37**, 56 (June, 1984).

Atomic Structure Calculations, Relativistic Atoms

J. T. Waber

Relativistic atomic calculations have become available for all of the atoms of the periodic table in the last decade and have extended our understanding in many areas of physics and chemistry. Some of these will be reviewed briefly below.

Apparently, the first such calculation [1] was made for the $_{29}Cu^{+1}$ ion in 1940. It passed relatively unnoticed until Mayers [2] did $_{80}$Hg in 1957. These were formidable tasks done without the aid of modern computers. Only a limited number of atomic calculations of any kind existed before 1963 when Herman and Skillman published their book on *Atomic Structure Calculations* [3]. Within a few years [4–7] a large number of nonrelativistic calculations started to appear with increasing array of complications. With these in hand, the importance of the various improvements could be assessed. Excellent agreement with experiment for energy values was demonstrated in 1975.

Several differences between the nonrelativistic Hartree–Fock–Slater (discussed elsewhere by Herman [8]) and the relativistic Dirac–Slater wavefunctions [9] are important. Instead of a single orbital $P(\mathbf{r})$ for each electron with the quantum numbers (nlm), one writes

$$\Psi_k(\mathbf{r}) = \frac{1}{r}\begin{pmatrix} i^l F(r)\Omega(jlm(\theta,\varphi) \\ i^{l'} G(r)\Omega(jl'm(\theta,\varphi) \end{pmatrix} \tag{1}$$

as the orbital for the kth electron, where $F(r)$ stands for the major radial component and $G(r)$ for the minor. Instead of one spherical harmonic or associated Legendre polynomial $Y_l^m(\theta,\varphi)$ to represent the angular dependence, there are several. The angular dependence for a given component is defined as

$$\Omega_{lj\mu} = A\{B(lj)Y_l^{\mu-\sigma}(\theta,\varphi)\begin{bmatrix}0\\1\end{bmatrix} + C(lj)Y_l^{\mu+\sigma}(\theta,\varphi)\begin{bmatrix}1\\0\end{bmatrix}\}\,, \tag{2}$$

Table 1: Quantum numbers for relativistic atoms.

		Symbol		s	p	d	f	g
	a	j	l	0	1	2	3	4
Number of	-1	$l-\frac{1}{2}$		0	2	4	6	8
electrons	$+1$	$l+\frac{1}{2}$		2	4	6	8	10
Sum				2	6	10	14	18
Kappa	-1			0	1	2	3	4
value	$+1$			-1	-2	-3	-4	-5

where A is a normalizing constant, $B(lj)$ and $C(lj)$ are coefficients dependent on the angular momenta, and the two functions $Y_l^{\mu\pm\sigma}$ are the same type of spherical harmonic as occurs in the nonrelativistic case. Finally $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ represent the Pauli spin matrices which represent up and down spin, respectively. Because the angular momentum l' differs from l by the quantity a (which may be either plus or minus 1), the angular dependence of the major and minor components depend either on l or l' or vice versa. Hence one involves an odd and the other an even power of $\cos\theta$.

The angular momentum j has the definition $j = l + \frac{1}{2}a$. Another aspect of a will be taken up below, and the μ values differ by 1 since $|\sigma| = \frac{1}{2}$. Concerning the two $Y_l^{\mu\pm\sigma}$ functions in the formula for $\Omega_{il\mu}$, the quantity μ is the resolved component of j. For these reasons $\Psi(\mathbf{r})$ is a four-component wave function. Details of the formulation are given by Grant [10] but his notation differs slightly from these formulas. The reader is cautioned that phase factors may differ in the various representations which have been published.

A simple table (Table 1) will indicate these quantum numbers as well as Dirac quantum number κ, and how many electrons can occupy a complete subshell. Thus, for example, the p shell is divided into two subshells; one is called $2p_{1/2}$ and the other $2p_{3/2}$. Each contains $2j+1$ electrons. This is one additional way in which the relativistic and nonrelativistic wave functions and atomic calculations differ.

The phenomenon just mentioned is called spin–orbit splitting and the energy separation between the two subshells varies roughly as Z^4, where Z is the atomic number. This fact will indicate why relativistic effects have only limited importance when dealing with the electrons of atoms and ions with low Z values but become very important for heavy elements.

Some of the details of making a self-consistent-field calculation have been indicated in this volume [8]. For greater detail, the reader is referred to recent papers [11–13]. The probability of finding the ith electron in a spherical shell of radius r reduces to the simple formula

$$\rho_i(r) = [F_i(r)]^2 + [G_i(r)]^2 \tag{3}$$

after integration over the angles θ and φ. The total charge density $\rho(r)$ is obtained by summing this over all the occupied orbitals. Slater's approximation [14] to exchange potential is related to the diameter of a Fermi hole, and hence is given by

$$V_{\text{ex}}(r) = \alpha_s \left[\frac{3}{8\pi}\sum \rho_i(r)\right]^{1/3} \tag{4}$$

where α_s is an adjustable constant in the range of $\frac{2}{3}$ to 1 and he summation runs over all of the

occupied orbitals. Mann [15] has discussed the slightly more complicated equations which arise when exchange is handled in the Hartree–Fock manner.

A convenient way [7] to write the radial equations is as a matrix for a given set of quantum numbers:

$$\frac{\mathrm{d}}{\mathrm{d}r}\begin{pmatrix} F(r) \\ G(r) \end{pmatrix} = \begin{pmatrix} \dfrac{-\kappa}{r} & \dfrac{(V-E_0-W)}{r} \\ \dfrac{-(V++E_0-W)}{r} & \dfrac{\kappa}{r} \end{pmatrix} \tag{5}$$

so that the derivative of F is equal to the sum of radial factors times both F and G – a similar situation applies for the minor component G. That is, the two radial functions are coupled together by first-order differential equations.

In Eq. (5), V is the Coulomb potential Z/r, E_0 is the rest mass of the electron, and W is the energy eigenvalue $E(nlj)$. The relation of the quantum number κ to the more familiar ones, l and j, is listed above.

By differentiating G with respect to r, and substituting, one can get from (5) a second-order differential equation in the major component. This is similar to the Schrödinger equation but contains additional components. One of these additional terms gives rise to the spin–orbit splitting. By assuming that such terms are only a small perturbation, one can estimate the relativistic effects with single-component nonrelativistic wave functions. This was the traditional approach found in textbooks before the relativistic calculations became available.

Relativistic atomic calculations are important for treating phenomena which involve the interaction between the nucleus and electrons. References to typical areas are: Mössbauer spectroscopy [12], hyperfine interaction [13], and beta capture [17].

Another area where relativistic effects are important is in connection with the valence electrons of heavy elements. This can be illustrated by the two diagrams which contrast the "order" of filling electrons in the lowest-energy state of a given shell nl. In a nonrelativistic treatment of the electrons in a rare earth (lanthanide) or actinide element, one first arranges the spin of the f electrons parallel according to LS or Russell–Saunders coupling scheme (Fig. 1). Progressively one spin is occupied after another and the energy increases due to electron–electron interaction. With seven electrons, the resultant is $J = \frac{7}{2}$, corresponding to the seven unpaired spins. Completion of the shell by adding the other $2l+1$ electrons brings the total spin to $J = 0$.

In Fig. 2, there are the two subshells with six and eight electrons separated by a large spin–orbit splitting. According to the modified Hund's rule [18], the first three electrons go in a parallel arrangement hence $J = -\frac{5}{2} - \frac{3}{2} - \frac{1}{2} = -\frac{9}{2}$. However, the next three are opposite in sign so that the sum of the six individual J values, namely J, becomes zero. The seventh electron would lead to a J value of $-\frac{7}{2}$ just as was found for LS coupling. In fact, in many cases, the result found by jj coupling (which is characteristic of a relativistic treatment) is the same as one would obtain by LS coupling; in others, the two results differ. For example, for six electrons, LS coupling would give -3 for J, where as the better answer is 0.

Another area which involves the outer electrons is the anomalous behavior of the angular distribution of photoelectrons. The jj coupling between the bound and the emergent electron in the continuum state leads to observed dependence and explains the occurrence of the Cooper minimum [19] in the asymmetry coefficient β as well as its energy dependence. In contrast,

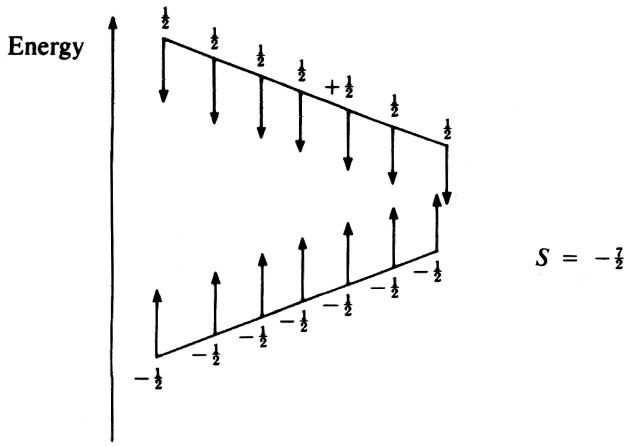

Fig. 1: Schematic of energy involved in adding electrons in *LS* coupling.

LS coupling gives a constant value [20, 21] for β.

The accuracy of the relativistic atomic calculations is indicated by the calculated theoretical binding energy of the $1s_{1/2}$ electron [22]:

Energy eigenvalue	−142.929 keV
Magnetic contribution	+0.715
Retardation effect	−0.041
Vacuum fluctuation	+0.457
Vacuum polarization	−0.155
Total binding energy (theoretical)	141.953 (±0.053) keV
Experimental value [23]	141.963 (±0.013) keV

The agreement is very good with the experimental value of Dittner *et al.* [21] and with the independent theoretical value (also based on a Dirac–Fock calculation) of Freedman *et al.* [23].

This listing also serves to indicate the magnitude of other relativistic effects which cannot be discussed fully here. Both the magnetic and the retardation corrections are involved in the Breit interaction – a coupling of two electrons by means of a virtual photon. The vacuum fluctuation is also called Zitterbewegung. Polarization of the vacuum results from the strong Coulomb field of the nucleus. Its estimation is discussed by Pyykkö [24].

Most of these calculations have assumed that the field experienced is spherically symmetric. While this is guaranteed by Unsold's theorem for a closed shell, it is not true for open shells and a more complex treatment is required for dealing with even relatively simple atoms. The reader is referred to the article in this volume by Nesbet [25].

The other important use of such relativistic wave functions is in the construction of a reasonable molecular or crystal potential when heavy elements are concerned. Two types of relativistic molecular calculations have been done. The "discrete variational method" was used in connection with a linear combination of numerical (relativistic) orbitals by Rosen and Ellis [26–28]. At nearly the same time, Yang and his collaborators developed a relativistic

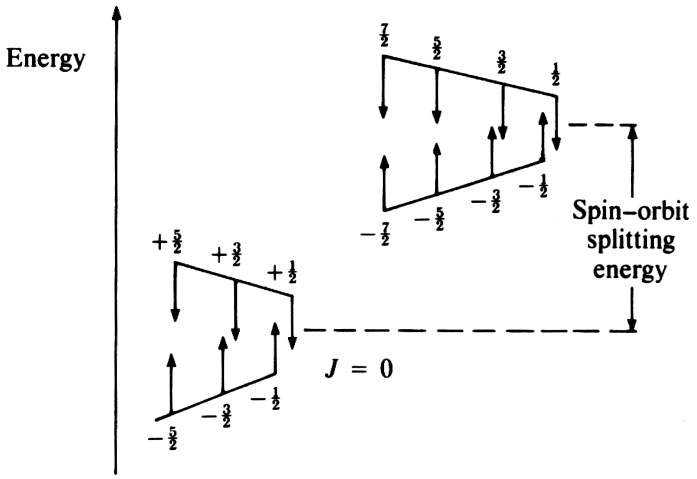

Fig. 2: Energy sketch of adding electrons with various spins in jj coupling.

version of the scattered wave formalism [29–32]. A review article by Pyykkö [24] covers a number of aspects of relativistic quantum chemistry. The earliest relativistic band calculations were made by Loucks [33, 34] who employed the augmented plane wave (APW) method. Sommers [35] developed a relativistic Kohn–Korringa–Rostoker method and Soven [37] a relativistic orthogonalized plane wave. More recently, Koelling and Freeman have published a series of relativistic APW calculations [38–41].

With the ready availability of even multiconfigurational atomic Dirac–Fock MCDF programs of Desclaux *et al.* [42] and Grant *et al.* [43], progress has been made in the last 10 years on the fronts of molecular and atomic structure calculations. Despite this fact, experimentalists continue to use a collection of approximation and *Ansatz* corrections to NR calculations to bring about improved comparisons with their data. These corrections are of questionable validity and may pertain to only portions of the periodic table. The MCDF calculations are "exact" and now with current computers are easy to perform.

While the word "exact" has been used, some difficulties remain; for example, the question of how to treat the interaction between two electrons is not resolved. In the comprehensive review in 1970 by Grant [44] of relativistic atomic calculations, two problems had not emerged, even though Brown and Ravenhall [45] had pointed out that the simple Dirac–Coulomb equation did not yield normalizable solutions. The success of obtaining good numerical eigenfunctions had diverted attention from such fundamental problems.

When finite basis sets were employed, unrealistically negative solutions were obtained and the "Brown–Ravenhall disease" and "variational collapse" became of significance. One cause [46] of the latter was that the same basis set was used to expand the major and minor components. The reader is referred to the studies of Kutzelnigg [47] and Goldman [48], Grant has made a number of comments [49] on these types of problem in recent conference volumes. These difficulties were attributable to the fact that the single-electron relativistic Hamiltonian is not bounded from below, i. e., there exists a negative energy continuum which has features in common with positive energy Rydberg states, and spurious roots with $E \sim -2c^2$ were

obtained [50]. The recommended procedure is to project out properly behaved solutions and constrain the solutions to pertain only to a fixed number of electrons, as was effectively done with numerical orbitals [51].

The use of basis sets was motivated by interest in following the NR procedures of Roothan [52] for molecules. Goldman and Dalgarno implemented a variational procedure for two-electron atoms which avoided spurious roots, variational collapse, and continuum dissolution. Results with $Z \leq 10$ were published [53]. It should be mentioned that the proper nodal behavior is complex. That is, in representing the orbital by such basis functions, the minor component has one more node than the major [54], i. e., the significance increases rapidly with Z.

Concerning the importance of utilizing relativistic orbitals, Torboem, Fricke, and Rosen [55] discuss isomer shifts for low-lying states of IIa and IIb elements and compare nonrelativistic NF, DF, and MCDF results. The inadequacy using the contact term $|\psi_{ns}|^2$ rather than the wave functions integrated over the range of the nuclear charge distribution is illustrated. They note that $np_{1/2}$ orbitals contribute very little to the overall electronics charge density in this region. However, they note that the $ns(n-1)d$ states make an important contribution in the multiconfigurational treatment.

Kim, Huang, Cheng and Desclaux [56] emphasize that care should be taken when comparing DF calculations and experimental spin–orbit splittings $^2P_{1/2}$–$^2P_{3/2}$. They presented results for two negative ions, B^- and F^-. Problems arose because $^2P_{1/2}$ and $^2P_{3/2}$ orbitals do not converge to the same nonrelativistic results and can yield spurious nonrelativistic contributions when one attempts to obtain good agreement with experimental data.

Pitzer and colleagues [57] have investigated the use of an effective potential obtained from DF calculations for molecular problems.

While retardation interaction between two electrons is not very large when treating an atom [54], it will become significant in treating molecules because of larger interelectronic distances. The necessity of treating atomic cases with nonorthogonal orbitals also has not been evaluated for molecular problems.

Zangwill and Liberman [58] have made some improvements of the Liberman–Waber–Cromer program to take in account time-dependent optical response.

Relativistic continuum orbitals have been discussed in the texts by Berestetskii, Lifschitz, and Pitaerskii [59] and Rose [60]. A computer code for such orbitals has been accepted by Computer Physics Communications [61] by Perger and a study involving both discrete and positive energy continuum orbitals expressed in an *ab initio* basis set obtained from the Dirac equation is near completion.

One way of summarizing this article on atomic calculations is to indicate the typical references to various kinds of relativistic atomic calculations which are available as well as some differences (Table 2). Most of these are available as numerical tabulations. The results of Fraga, Saxena, etc. are as coefficients of algebraic basis functions. The maximum atomic number studied for each type is indicated.

Pyykkö [62] organized an exhaustive bibliography of several hundred references covering the period 1916 to 1985. The reader is directed to it for locating the literature on various relativistic studies. The recent paper by Grant and Quiney [63] is an excellent review of the theoretical and computational situation.

Table 2: Some typical relativistic atomic calculations.

Name of type	Author	Type of exchange	max. Z calculated	Year	Footnote
Hartree–Fock	Cohen	None	92	1951	*a*
Dirac–Slater	Carlson *et al.*	Statistical	126	1966	*b*
numerical	Liberman *et al.*	Statistical	126	1966	*c*, *q*, *p*
	Schofield	Statistical	104	1974	*d*
	Rosin, Lindgren	Modified Slater	95	1968	*e*
	Band *et al.*	Statistical	95	1977	*f*
	Huang *et al.*	Statistical	106	1965	*g*
Analytic expansion	Fraga *et al.*	Fock integrals	102	1965	*h*
	Kim		40	1967	*i*
	Kagawa		50	1975	*j*
Dirac–Fock	Mann, Waber	Fock integrals plus	131	1970	*k*
		Breit and correlation corrections	126	1970	*l*
Multiconfiguration	Desclaux	Fock integrals plus	120	1973	*m*
Dirac-Fock	Desclaux	Breit and	126	1976	*n*
	Fricke, Soff	correlation	173	1977	*o*
		corrections			

[a] S. Cohen, *Phys. Rev.* **118**, 489 (1960).
[b] C. C. Lu, T. A. Carlson, F. B. Malik, T. C. Tucker, and C. W. Nestor, Jr., *At. Data* **3**, 1 (1971); (erratum) **14**, 89 (1974); **2**, 63 (1970).
[c] D. Liberman, J. T. Waber, and D. T. Cromer, *Phys. Rev.* **137**, A27 (1965).
[d] J. H. Schofield, *At. Nucl. Data Tab.* **14**, 121 (1974); *Phys. Rev.* **9**, 1041 (1974).
[e] A. Rosin and I. Lindgren, *Phys. Rev.* **176**, 114 (1968); see also I. Lindgren and A. Rosin, *Atom. Phys.* **4**, 93 (1974).
[f] I. M. Band, M. A. Listengarten, M. B. Trzhaskovskaya, and V. I. Fomichev, *Leningrad. Inst. Idernov. Fiz. Report* **298** (April 1977).
[g] K. H. Huang, M. Aoyagi, M. H. Chen, B. Craseman, and H. Mark, *At. Data* **18**, 243 (1976).
[h] S. Fraga and K. M. S. Saxena. *At. Data* **3**, 323 (1971); **4**, 255 (1972); **4**, 269 (1972); **5**, 467 (1973).
[i] Y. K. Kim, *Phys. Rev.* **154**, 17 (1967); (erratum) **159**, 190 (1967).
[j] T. Kagawa, *Phys. Rev. A* **12**, 2245 (1975).
[k] J. B. Mann and J. T. Waber, *J. Chem. Phys.* **53**, 2397 (1970).
[l] J. B. Mann, *J. Chem. Phys.* **51**, 841 (1969).
[m] J. P. Desclaux, *At. Data* **2**, 311 (1973).
[n] J. P. Desclaux, *At. Data* **18**, 243 (1976).
[o] B. Fricke and G. Soff, *At. Nucl. Data Tab.* **19**, 83 (1977).
[p] D. A. Liberman, *Phys. Rev. B* **2**, 244 (1970).
[q] W. Kohn and L. T. Sham. *Phys. Rev.* **140**, A1133 (1965).

See also: Atomic Structure Calculations, One-Electron Models.

References

[1] A. O. Williams, *Phys. Rev.* **58**, 723 (l970).
[2] D. F. Mayers, *Proc. Roy. Soc. (London)* **A241**, 93 (1957). J. P. Desclaux, D. F. Mayers, and F. O'Brien, *Phys. Rev. B* **4**, 631 (1971).
[3] F. Herman and S. Skillman, *Atomic Structure Calculations* (Prentice-Hall, Englewood Cliffs, N.J., 1963).
[4] M. A. Coulthard, *Proc. Roy. Soc. (London)* **91**, 44 (1967); **91**, 421 (1967).
[5] F. C. Smith and W. R. Johnson, *Phys. Rev.* **160**, f36 (1967).
[6] C. C. Lu, T. A. Carlson, F. B. Malik, T. C. Tucker, and C. W. Nestor, Jr., *At. Data* **3**, 1 (1971); (erratum) **14**, 89 (1974); **2**, 63 (1970).
[7] D. Liberman, J. T. Waber, and D. T. Cromer, *Phys. Rev.* **137**, A27 (1965).
[8] F. Herman, this Encyclopedia.
[9] V. M. Burke and I. P. Grant, *Proc. Phys. Soc. (London)* **90**, 297 (1967).
[10] I. P. Grant, *Adv. Phys.* **19**, 747 (1970). A slightly different notation was used in earlier papers [*Proc. Roy. Soc. A* **262**, 555 (1961); *Proc. Phys. Soc. (London)* **86**, 523 (1965)].
[11] D. Liberman, J. T. Waber, and D. T. Cromer, *Comp. Phys. Commun.* **2**, 107 (1971).
[12] B. Fricke and J. T. Waber, *Actinide Rev.* **1**, 433 (1971); *Theo. Chim. Acta* **21**, 235 (1971); *Phys. Rev. B* **5**, 3445 (1972).
[13] J.-P. Desclaux, *Int. J. Quantum Chem.* **6**, 25 (1972); J.-P. Desclaux and N. Bessis, *Phys. Rev. A* **2**, 1623 (1970).
[14] J. C. Slater, *Phys. Rev.* **81**, 385 (1951).
[15] J. B. Mann, J. *Chem. Phys.* **51**, 841 (1969).
[16] J. B. Mann and J. T. Waber. *At. Data* **5**, 201 (1973); *J. Chem. Phys.* **53**, 2397 (1970).
[17] T. A. Carlson, C. W. Nestor, Jr., F. B. Malik, and T. C. Tucker, *Nucl. Phys. A* **135**, 57 (1969).
[18] T. E. H. Walker and J. T. Waber, *Phys. Rev. A* **7**, 1218 (l973).
[19] T. E. H. Walker and J. T. Waber, *Phys. Rev. Lett.* **30**, 307 (1973).
[20] T. E. H. Walker and J. T. Waber, *J. Phys. B* **6**, l 165 (1973).
[21] T. E. H. Walker and J. T. Waber, *J. Phys. B* **7**, 674 (1974).
[22] B. Fricke, J.-P. Desclaux, and J. T. Waber, *Phys. Rev. Lett.* **28**, 714 (1972).
[23] M. S. Freedman, F. T. Porter, and J. B. Mann, *Phys. Rev. Lett.* **28**, 711 (1972). P. F. Dittmer, C. E. Bemis, D. C. Hansley, R. J. Silva, and D. C. Goodmen *Phys. Rev. Lett.* **26**, 1037 (1971).
[24] P. Pyykko, *Adv. Quantum Chem.* (to be published).
[25] R. Nesbet, this Encyclopedia.
[26] A. Rosen and D. E. Ellis, *J. Chem. Phys.* **62**, 3039 (1975).
[27] D. E. Ellis, A. Rosin, and P. F. Walch, *Int. J. Quantum Chem.* **S9**, 351 (1975).
[28] P. F. Walch and D. E. Ellis, J. *Chem. Phys.* **65**, 2387 (1976).
[29] C. Y. Yang and S. Rabii, *Phys. Rev. A* **12**, 362 (1975).
[30] C. Y. Yang, K. H. Johnson, and J. A. Horsley, *Bull. Am. Phys. Soc.* **21**, 382 (1976).
[31] C. Y. Yang, *Chem. Phys. Lett.* **41**, 588 (1976).
[32] C. Y. Yang and S. Rabii, *J. Chem. Phys.* **78**, 6S (1978).
[33] T. Loucks, *Phys. Rev.* **139**, A f333 (1965).
[34] T. Loucks, *Augmented Plane Wave Method* (Benjamin, New York, 1966).
[35] C. Sommers and H. Amar, *Phys. Rev.* **188**, 1117 (1969).
[36] C. Sommers, *J. Phys. C* **3**, 39 (1972).
[37] P. Soven, *Phys. Rev.* **137**, A1706 (1965).
[38] D. Koelling, *Phys. Rev.* **188**, 1049 (1969).
[39] D. Koelling and A. J. Freeman, *Phys. Rev. B* **7**, 4454 (1973).

[40] D. Koelling and A. J. Freeman, *Phys. Rev. B* **12**, 5622 (1975).
[41] D. Koelling and A. J. Freeman, *Plutonium and Other Actinides*, p. 2911 (H. Blank and R. Lindner, eds.). (North Holland, Amsterdam, 1976).
[42] J.-P. Desclaux, *Comp. Phys. Commun.* **9**, 31 (1975).
[43] I. P. Grant, *Comp. Phys. Commun.* **21**, 207 (1980).
[44] Lan P. Grant, *Adv. Phys.* **19**, 747–811 (1970).
[45] G. E. Brown and D. G. Ravenhall, *Proc. R. Soc. (London) A* **208**, 552 (1951).
[46] Y. Ishikawa, R. C. Binning, and K. M. Sand, *Chem. Phys. Lett.* **105**, 189 (1984); **101**, 111 (1983). J. Mark and P. Rozickey, *Chem. Phys. Lett.* **74**, 562 (1980).
[47] W. Kutzelnigg, *Int. J. Quantum Chem.* **25**, 107 (1984).
[48] S. P. Goldman, *Phys. Rev. A* **30**, 1219 (1984); **31**, 354 (1985); **37**, 16–30 (1988).
[49] I. P. Grant, in *Atom Theory Workshop on Relativistic and QED Effect in Heavy Atoms* (H. Kelly and Y.-Ki Kim, eds., pp. 17–19, 200–203, 299–301, American Institute of Physics, New York, 1985). See also *Phys. Rev. A* **25**, 1230 (1982).
[50] See J. Sucher, in *Proceedings of the NATO Advance Study Institute on Relativistic Effects in Atoms, Molecules and Solids* (G. Malli, ed.), Plenum, New York, 1982; also *Proceedings of Argonne Workshop on the Relativistic Theory of Atomic Structure* (H. G. Berry, K. T. Chem, W. K. Johnson, and Y.-Ki Kim, eds.), ANL-80-116, Argonne National Laboratory, Argonne, IL, 1980. See also M. Mittleman, *Phys. Rev. A* **4**, 893 (1971); *A* **15**, 2395 (1972).
[51] W. D. Sepp and B. Fricke, in *Atomic Theory Workshop on Relativistic and QED Effects in Heavy Atoms*, AIP Conf. 136 (H. Kelly and Y.-Ki Kim, eds.), pp. 20–25, American Institute of Physics, New York, 1985.
[52] C. C. J. Roothan, *Rev. Mod. Phys.* **32**, 179 (1960).
[53] S. P. Goldman and A. Dalgarno, *Phys. Rev. Lett.* **57**, 408 (1988).
[54] P. J. C. Airts and W. C. Nieuwport, *Chem. Phys. Lett.* **113**, 165 (1985).
[55] G. Torboem, B. Fricke, and A. Rosin, *Phys. Rev. A* **31**, 2038–2053 (1985).
[56] K. N. Huang, Y.-Ki Kim, K. T. Cheng, and J.-P. Desclaux, *Phys. Rev. Lett.* **48**, 1245 (1982).
[57] Y. S. Lee, W. C. Krumler, and K. S. Pitzer, *J. Chem. Phys.* **67**, 5861 (1977); **69**, 976 (1978); **70**, 288–293 (1979); **73**, 360 (1980); **74**, 1162 (1981).
[58] A. Zangwill and D. Liberman, *Comp. Phys. Commun.* **37**,75–82 (1984).
[59] V. D. Berestetskii, E. M. Lifschitz, and L. P. Pitaerskii, *Relativistic Quantum Theory*, Vol. 1, pp. 113–115. Pergamon Press, New York, 1971.
[60] M. E. Rose, *Relativistic Electron Theory*, p. 82ff, Wiley, New York, 1961.
[61] W. Perger, private communication, Michigan Technological University. Accepted by *J. Comp. Phys.* (1990).
[62] Pekka Pyykko, in *Relativistic Theory of Atoms and Molecules*, Lecture Notes in Chemistry **41**, Springer-Verlag, New York, 1986.
[63] Ian Grant and H. M. Quiney, *Adv. At. Mol. Phys.* **23**, 37–86 (1988).

Atomic Trapping and Cooling

P. van der Straten and H. Metcalf

Optical Forces

The notion of optical forces goes back to Maxwell, but their modern implementation for laser cooling is most commonly described in terms of the momentum of light when it is absorbed by an atom making a discrete transition between states whose energy difference is ΔE. The

magnitude of this momentum exchange is related to the energy through the relativistic formula $p = \Delta E/c = h\nu/c = \hbar k$ where $k \equiv 2\pi/\lambda$ and λ is the wavelength of the light. In order for the momentum exchange between the atom and the light field to be efficient, the light must drive a resonant transition between atomic states, and so must have the right frequency. This leads immediately to a model description that involves only two atomic states, one ground and one excited.

Both absorption and emission exchange nearly the same magnitude of momentum between the atoms and the light, so any net momentum exchange must arise from directionality. At low light intensity I the dominant return to the ground state is through spontaneous emission, and directional exchange is thus implemented because it occurs in random directions. Atoms can only undergo spontaneous emission from their excited states whose lifetime is $\tau \equiv 1/\gamma$, and even under the strongest excitation, they spend no more than 50% of the time in the excited state, so the maximum average rate of spontaneous emission from optically excited atoms is $\gamma/2$ and the maximum force is $\hbar k\gamma/2$. At high intensities the momentum exchange is limited by stimulated emission because the absorption and stimulated emission are both parallel to the laser beam. High-intensity forces are usually produced in the presence of multiple beams so that absorbtion from one can be followed by stimulated emission into the other. The momentum difference between these is imparted to the atoms.

In the low intensity domain the spontaneous emission rate is $\gamma_p = (\gamma s/2)/(1 + s + \Delta^2)$. Here $s \equiv (3\lambda^3\tau)/(\pi hc)I$, and Δ is related to the detuning of the light ω_ℓ from exact atomic resonance $\omega_{\rm atom}$ by $\Delta = 2(\omega_\ell - \omega_{\rm atom})/\gamma \equiv 2\delta/\gamma$. Because spontaneous emission is in random directions its average vanishes so the direction of the force is the same as the direction of the light. This "radiative" force is usually written as $\mathbf{F} = \hbar\mathbf{k}\gamma_{\rm p}$. The optical frequency is measured in the reference frame of the lab which is different from that of the atoms moving at velocity $\mathbf{v}$, and so the Doppler shift $\omega_{\rm D} \equiv -\mathbf{k}\cdot\mathbf{v}$ must be included in the detuning δ. The negative sign arises because the frequency is increased when $\mathbf{k}$ and $\mathbf{v}$ are in opposite directions. In the atomic rest frame the optical frequency is $\omega_\ell - \mathbf{k}\cdot\mathbf{v}$.

The most common form of laser cooling uses counter-propagating beams of light tuned just below atomic resonance. The formula $\mathbf{F} = \hbar\mathbf{k}\gamma_p$ is applied for each of the two beams, but the force has opposite directions and the Doppler shifts are different. The result is a total force $\mathbf{F}_{\rm tot} = -\beta\mathbf{v}$ which damps atomic motion in either direction. Here β is a constant that depends on the atomic and laser parameters. The notion of this damping from the radiative force is readily extended to three dimensions, and such a cooling configuration is called "Optical Molasses".

A pure damping force would bring the atoms to rest and thus to the impossible temperature of $T = 0$. Therefore it is necessary to consider the discreteness of each momentum exchange near the cooling limit. The result is an ultimate low temperature of $T_{\rm D} = \hbar\gamma/2k_{\rm B} \sim 10^{-4}\,{\rm K}$, where $k_{\rm B}$ is Boltzmann's constant and the subscript 'D' refers to the Doppler-shift dependence of the mechanism of cooling. Experiments have shown the inadequacy of the two-level atom model that leads to $T_{\rm D}$, and in fact, both theory and experiment involving real atoms that have multiple energy levels show the ultimate low temperature obtainable by optical cooling in the presence of spontaneous emission is a few times $T_{\rm r} = \hbar k^2/2Mk_{\rm B}$ where M is the atomic mass and the subscript 'r' refers to recoil. At $T_{\rm r} \sim$ few μK the atomic de Broglie wavelength is comparable to λ, and this has important consequences for further cooling.

When atoms interact with nearly-resonant light, they not only absorb its energy and momentum, but they experience also shifts of their energy levels given by $\Delta E_{\rm LS} = \frac{\hbar}{2}\{\sqrt{\Omega^2+\delta^2} - \delta\}$ where the Rabi frequency Ω characterizes the strength of interaction between the atoms and the light such that $\Omega^2 = s\gamma^2/2$. In the limit of $\delta \gg \Omega$ that characterizes many low intensity experiments, $\Delta E_{\rm LS} \approx \Omega^2/4\delta$ is proportional to the light intensity and is therefore called the light shift. This light shift can result in forces on atoms when the light intensity varies in space, such as between the nodes and anti-nodes of a standing wave, because the spatial energy dependence can be viewed as a potential. In some sense, it derives from multiple sequences of absorption and stimulated emission, and is therefore conservative and cannot be used for cooling, just as is the case for any force derived from a potential. Such interactions are labeled the "dipole force" in analogy to the static case, because the light induces an atomic dipole moment, and the atom is in an inhomogeneous field.

Thus optical forces on atoms arise from absorption followed by either spontaneous or stimulated emission, and in general, both processes take place. Atoms can be confined or steered by the dipole force, and cooled by the radiative force, thereby providing physicists with enormously powerful and flexible tools for controlling atomic motion.

Confinement – Light Beams and Magnetic Fields

The sign of the the light shift depends on the detuning. For $\delta < 0$, called "red" detuning, the energy of ground state atoms is lower in more intense light and thus atoms in an inhomogeneous light field are attracted to regions of high intensity. The region near the focus of a single laser beam therefore provides a radial attraction for atoms via the dipole force. If the focus is sufficiently sharp, meaning that the light intensity decreases strongly in either longitudinal direction moving away from the focus, this dipole force can exceed the radiative force that tends to expel the atoms from the focal region longitudinally, especially if the detuning is large enough. Thus a single focussed laser beam having sufficiently large Ω and δ can confine atoms to a quite small region of space. Changing the position of the focus by steering the beam allows atoms to be manipulated as if they were held by tweezers. Such "optical tweezers" have been used on single atoms and on Bose–Einstein condensates, as well as on macroscopic objects.

By contrast, ground state atoms are repelled from the region of high light intensity if the detuning is "blue", $\delta > 0$. Atomic mirrors have been made by focussing a blue detuned laser beam into a sheet of light with cylindrical optics, and a trap has been made with an array of such sheets of light. Atomic mirrors have also been made using the very thin film of light produced by a blue detuned evanescent wave near the surface of a flat piece of glass illuminated from the inside with a beam of light near the critical angle. Finally, blue detuned Laguerre–Gaussian beams with a hollow center have been used to confine atoms in two dimensions and thus guide their motion along the path of the light beam. Two orthogonal, separated sheets of light can cap such an elongated region making a quasi one dimensional trap.

Many new phenomena appear in multiple laser beams because there can be absorption from one beam and stimulated emission into the other. Optical molasses described above is achieved in counter-propagating beams having $\delta \sim \gamma$ and $s \sim 1$ so that spontaneous emission is more likely than stimulated emission. Under these conditions, an atomic sample is cooled. By contrast, in the parameter range $\delta \gg \gamma$ and $s \gg 1$ satisfying $\delta \gg \Omega$, the dominant interaction

is the dipole force. In a standing wave, this dipole force oscillates in space on the wavelength scale thereby forming a periodic potential, and atoms are said to be subject to an optical lattice. (Optical crystal is an inappropriate name except under conditions where all the sites can be filled with exactly one atom.) Atomic motion in such a periodic potential is subject to the well-known conditions described by the Bloch theorem and Bloch wave functions. By careful choices of the mean kinetic energy (temperature) and well depth of the lattice, very many phenomena of condensed matter can studied.

Most atoms have a non-zero magnetic moment in their ground state, and hence can be confined by an inhomogeneous magnetic field. The simplest imaginable case uses a pair of coaxial coils carrying opposite currents to form a quadrupole field. (A single coil cannot work because a local field maximum cannot exist.) The B field at their geometric center is zero, but is non-zero everywhere else in space. Thus atoms whose magnetic moments $\boldsymbol{\mu}$ are properly oriented are attracted to the field zero, and can be confined there if their kinetic energy is less than $\boldsymbol{\mu} \cdot \mathbf{B}_{\max}$ where $\mathbf{B}_{\max}$ is the smallest magnetic field in the vicinity of the coils that constitutes a local maximum. Such magnetic trapping of atoms is free from the disturbing effects of light beams, and can therefore confine atoms with far less disturbance than optical traps. Magnetic traps are widely used for the containment of ultra-cold samples of gas, and particularly Bose–Einstein condensates.

Perhaps the most widely used atomic confinement method is the magneto-optical trap (MOT). It exploits the selection rules associated with the polarization of light and the magnetic sublevels of atoms by making the absorption of light from multiple beams depends upon atomic position in an inhomogeneous magnetic field. The field and light beams are carefully arranged so that atoms not in the center of the trap preferentially absorb light from beams that tend to push them back toward the trap center through the radiative force. The MOT has an extremely large velocity capture range and depth, much larger than either pure optical or magnetic traps.

A much more important property, however, is that the force in a MOT is velocity dependent through the Doppler shift, similar to the force in optical molasses. Both magnetic and optical traps are conservative, that is, atoms entering from one side will readily pass through them and escape from the other side. They can be loaded only by applying them to an already cooled sample of atoms, or by applying a cooling force to atoms traveling in them. By contrast, a MOT tends to slow atoms traversing it so that they cannot escape, and therefore it can capture atoms without auxiliary cooling. Moreover, their capture range can be chosen by the laser and magnetic parameters, and can be changed after a sample is captured to enable further cooling and/or compression. Finally, the magnetic field configuration is the same as that of the quadrupole magnetic trap, so that a MOT can be converted to a magnetic trap simply by shutting off the laser beams, thereby transferring a captured sample from the MOT into a dark and cold purely magnetic trap.

Applications

Although laser cooling has first been discussed in relation with high-resolution atomic spectroscopy, today it is used in many areas of atomic physics. One of the most prominent applications is the research of atomic collisions, where the researchers have been able to study atomic interactions at very low energies. This has led to photoassociation spectroscopy, where

during the interaction between two ultracold atoms a photon is absorbed and the total system is bound in a transient molecular state. Since the translational energy in the initial state is very low, this offers a novel way of studying molecular states with very high resolution. One of the results of these studies is the measurement of the scattering length in the ground state, which can be measured for with unprecedented resolution. This information is of crucial importance for the achievement of Bose–Einstein condensation in such systems.

For sufficient low energy and high densities atomic gases can make a transition to a new state, the so-called Bose–Einstein condensation. Although this has been predicted already in 1925 by S. N. Bose and A. Einstein, its experimental realization in dilute atomic gases was first realized in 1995 using laser cooling and trapping in combination with evaporative cooling. If the atomic gas makes a transition to a condensate, all atoms are in an identical state and have to be described by one macroscopic wavefunction. The coherence properties of such a condensate can be exploited for many novel experiments, like superfluidity, quantized vortices, parametric down-conversion, collapsing condensates and matter–wave interferometry.

The study of Bose–Einstein condensation forms a bridge between the research in atomic physics on laser cooled atoms and many other fields in physics, like low-temperature physics, condensed-matter physics and statistical physics. For example, loading a Bose–Einstein condensate in an optical lattice make it possible to study in this atomic system the quantum phase transition between a superfluid and Mott insulator state. In analogy with electron conduction in condensed matter, the conduction of atoms in an optical lattice can be tuned by increasing or decreasing the optical potential for the atoms, leading to an "conduction" state, where the atoms are distributed over the lattice in a random way, or an "insulator" state, where the atoms are evenly distributed over the lattice sites, thus creating an integer filling of all sites. It is shown, that this transition can be crossed many times without loosing the coherence of the atoms in the superfluid state.

See also: Bose–Einstein Condensation; Cold Atoms and Molecules; Quantum Optics; Ultracold Quantum Gases.

Bibliography

H. J. Metcalf and P. van der Straten, *Laser Cooling and Trapping*. Springer, New York, 1999.

C. J. Pethick and H. Smith, *Bose–Einstein Condensation in Dilute Gases*. Cambridge University Press, Cambridge, 2002.

Atoms

A. P. French

Introduction

Despite all the discoveries that have been made since the beginning of the twentieth century in the fields of nuclear and subnuclear physics, the atom remains the most important type of unitary system in our picture of the physical world. This is in part because the electrically neutral, stable atom is the basic building block in the structure of condensed matter as we are most familiar with it. Under more severe conditions, e. g., at the enormously high temperatures

and densities characteristic of stellar interiors, the individual atom ceases to be an identifiable unit of the structure. Nevertheless, whenever conditions permit, the atom will establish its existence because of its property of being the system of least energy and greatest stability that can be formed from its constituent particles. This property guarantees the atom a permanent place in our description of nature.

Early History

The birth of the theory that the basic structure of matter is discrete, not continuous, is usually attributed to the ancient Greeks, in particular Democritus (ca. 420 B.C.). (The name "atom" comes directly from the Greek atomos – indivisible.) However, it was not until about 1800 that the quantitative study of chemical reactions provided evidence that the behavior of bulk matter might indeed be governed by individual processes on a submicroscopic level. In the latter part of the eighteenth century, A. L. Lavoisier found that the total mass was conserved in chemical reactions, whatever changes in form and appearance took place among the reactants, and J. L. Proust showed that every pure chemical compound contains fixed and constant proportions (by weight) of its constituent elements. (The modern concept of element was propounded by Robert Boyle in 1661.) Building on these results, John Dalton in 1808 enunciated a detailed theory of chemical combination, based on the picture that each element is made up of a host of identical atoms and that the formation of a chemical compound from its elements takes place through the formation of "compound atoms" containing a definite (and small) number of atoms of each element.

From the known mass ratios for many different reactions, Dalton was able to suggest values of mass ratios of individual kinds of atoms. However, his scheme led to certain ambiguities and inconsistencies, which were not resolved until it came to be realized that the basic units of a pure element were not necessarily single atoms, but were often compound atoms in the form of diatomic molecules – a hypothesis first put forward by A. Avogadro in 1811, but which met resistance (despite its success in removing internal contradictions from Dalton's theory) because there was no obvious reason why such well-defined compounds of identical atoms should exist.

Relative Atomic Masses

In 1860 an international conference on atomic weights officially adopted the Dalton–Avogadro scheme, and during the succeeding decades a highly accurate tabulation of the relative atomic masses of the elements was built up from the analysis of thousands of different compounds. A natural unit for this scheme of relative atomic masses was the mass of hydrogen, the lightest atom. This was Dalton's choice, but in 1902 the basis was changed to a slightly different value, namely, one sixteenth of the atomic weight of oxygen. (A prime reason for this change was the practical one that oxygen, in contrast to hydrogen, forms stable and tractable compounds with most elements.) However, the discovery of isotopy (that a given element may have atoms of several different characteristic masses) led to a decision, in 1961, to redefine the atomic mass unit as one twelfth of the mass of the particular isotope carbon-12, and to express chemical atomic weights and the masses of individual isotopes as multiples of this unit.

Avogadro's Number and the Atomic Mass Unit

The atomic-molecular theory was developed in the absence of any direct evidence for the granular structure of matter (although the small irregular movements of microscopic particles in liquid suspension, discovered by the botanist Robert Brown in 1827, were at least suggestive). However, if the theory is assumed to be correct, a definite value must exist for the number of atoms in a given mass of material of known composition. In particular, the *Avogadro number*, N_A, is defined as the number of atoms in a mass of elementary substance equal in grams to its (relative) atomic weight: The mass of an individual atom is thus equal to its atomic weight (in grams) divided by N_A, and the atomic mass unit (1 amu) is numerically equal to $1/N_A$.

One early source of information leading to quantitative estimates of N_A was the study of the properties of gases. The atomic-molecular kinetic theory of gases led to a picture of a gas as made up of small particles traveling at high speeds (hundreds of meters per second). The slowness of the processes of diffusion and mixing in gases implied, however, that individual molecules travel only very short distances before suffering changes of direction through collisions. If there are n molecules of diameter d per unit volume, the mean free path (which can be related directly to the observed diffusion rate) is of the order of $1/nd^2$. To obtain separately the values of n and d, we can use the fact that the total volume of n molecules is about nd^3, and will represent the volume occupied by these molecules if they are condensed to the almost incompressible liquid phase. Analysis along these lines indicated that N_A must be of the order of 10^{24}.

Much more precise values of N_A were obtained later by quite different methods. In 1900, Max Planck inferred a value within 3% of the currently accepted figure as a result of his quantum analysis of the continuous radiation spectrum of hot bodies. In 1916, R. A. Millikan's experimental proof that electric charge exists only in integral multiples of a unit, e, led to a value of N_A as given by the ratio F/e, where F, the faraday, is the amount of electric charge associated with the transport of 1 gram-equivalent of a substance in electrolysis. Later, in about 1930, J. A. Bearden and others made determinations of interatomic distances in crystals, through the process of diffraction of x-rays of known wavelength (measured with a ruled grating); from this the number of atoms in a known mass of crystal could be inferred directly.

The currently accepted value of the Avogadro number is 6.02214×10^{23}, from which follows the result

$$1\,\text{amu} = 1.66054 \times 10^{-24}\,\text{g}$$

Masses and Sizes of Individual Atoms

The relative atomic weights of the known atoms range from about 1 (hydrogen) to 250 (highly unstable transuranic elements). The corresponding absolute masses range from 1.67×10^{-24} g to about 4.1×10^{-22} g. In this range are about 100 different elements (as characterized by the atomic numbers Z in the periodic table) comprising about 300 naturally occurring distinct atomic species – a given atomic species being characterized by (besides its Z value) its *mass number*, A, the integer closest to its relative atomic weight.

The large range of atomic masses is not accompanied by a correspondingly large or systematic variation in size. Atomic radii all lie between about 0.5 and 2.5 Å (10^{-10} m), with no marked increase from the lightest to the heaviest. One of the simplest sources of this informa-

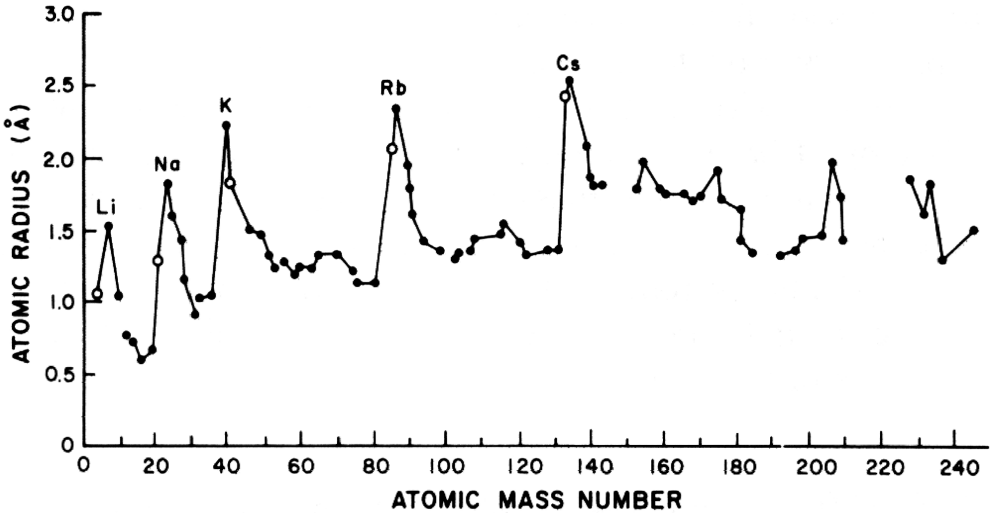

Fig. 1: Atomic radii. Black dots: radius based on experimental analyses of crystals and/or molecules; open circles: radius based on mean-free-path experiments.

tion is a knowledge of the densities of the elements in their solid state; another source is the measurement of cross sections for interatomic collisions. Figure 1 shows how the radii vary with Z; particularly noteworthy are the relatively small radii for the atoms of the noble gases and the especially large radii for the alkali metal atoms that immediately follow them in the periodic table. The general features of Fig. 1 can be understood from the standpoint of the internal structure of atoms (see later).

The Nuclear Atom

The discovery of the electron by J. J. Thomson in 1897 led quickly to the conclusion that all atoms contain some number of electrons. It followed that atoms, being electrically neutral, must be composed of some combination of electrons and positively charged material. Then, during the period 1911–1913, Ernest Rutherford, with H. Geiger and E. Marsden, carried out the alpha-particle scattering experiments that proved that the positive charge and most of the mass of an atom are concentrated in a minute volume, with a radius of less than 10^{-14} m (i. e., less than 10^{-4} of the outer radius of the atom).

Hard on the heels of Rutherford's discovery came the quantum model of the hydrogen atom, published by Niels Bohr in 1913. Using only the known constants of nature (including Planck's constant h), and without the use of any adjustable parameters, Bohr developed an accurate and quantitative model of the hydrogen atom as a nucleus of charge $+e$ (associated with 99.95% of the mass of the whole atom) with a single electron in orbit around it. For the atom in its ground (lowest-energy) state, Bohr's theory gave a radius of 0.53 Å and an ionization energy of 13.6 eV, in good agreement with observation. Furthermore, and perhaps even more striking, his theory accounted naturally for the systematic series of lines in the visible spectrum (Balmer spectrum) of atomic hydrogen. The basis was, according to the theory, that

the possible (quantized) energies of an electron bound to a central charge of magnitude Qe are given by

$$E(n) = -\frac{2\pi^2 me^4 Q^2}{h^2} \times \frac{1}{n^2} \qquad (n = 1, 2, 3, \ldots) \tag{1}$$

and that the wavelengths λ of possible spectral lines are defined by the amounts of energy carried away by photons in quantum jumps between levels, using the relations

$$E_{\text{photon}} = E(n_1) - E(N_2)$$

and

$$E_{\text{photon}} = h\nu = hc/\lambda \, .$$

X-Rays, Atomic Number, and Nuclear Charge

Bohr's theory had little success in accounting for optical spectra other than those of hydrogen or equally simple systems (e. g., singly ionized helium). However, it was possible to apply the theory to the so-called characteristic x-rays emitted by many elements when bombarded with energetic electrons. By 1913 it was established that these x-rays are electromagnetic radiations similar to visible light but of much shorter wavelength (typically of the order of 1 Å). In 1913 H. G.-J. Moseley made a systematic study of the characteristic x-rays and found that their frequencies ν were just what would be expected, according to the Bohr energy-level formula, for a single electron falling to the lowest ($n = 1$) level from a level with $n = 2$ or $n = 3$. A graph of $\sqrt{\nu}$ against atomic number Z was a pair of straight lines (Fig. 2). Moseley concluded that the characteristic x-rays were produced when an atom returned to its ground state after an electron in its lowest possible orbit, close to the nucleus, had been knocked out. He concluded further that the chemical atomic number could be identified with the number of units of positive charge on the nucleus.

Quantum Mechanics and Atomic Structure

With the development of wave mechanics by E. Schrödinger in 1926, the picture of an electron in the field of a nucleus was drastically changed. In place of well-defined quantized orbits it became necessary to think in terms of smoothly varying probability distributions that permitted the electron to be found at any distance from the nucleus, or even inside it. The quantized energies of the possible bound states duplicated the results of the Bohr theory, but in every other way the description was quite different.

The spatial state of an individual electron could be characterized with the help of three quantum numbers, n, l, and m, all integers. The value of n, called the *principal quantum number*, defined the electron energy, just as in the Bohr theory, but the identification of a state also required the two quantum numbers l and m that defined the orbital angular momentum {of magnitude $[l(l+1)]^{1/2}h/2\pi$} of the electron and the projection, of magnitude $mh/2\pi$, of this orbital angular momentum along a specified axis. The quantum analysis required $0 \leq l \leq < n-1$, and $-l \leq m \leq +l$.

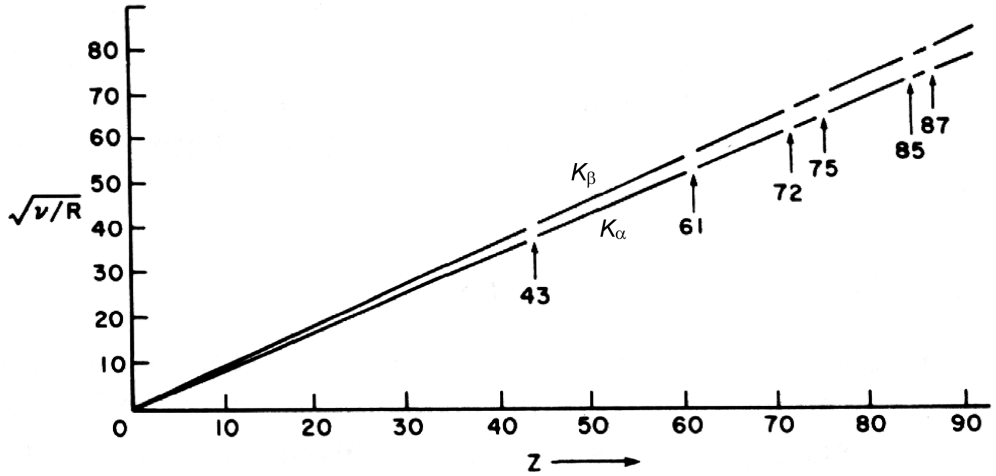

Fig. 2: Linear dependence of $\sqrt{\nu}$ on atomic number for characteristic x-rays. The line marked K_α corresponds to the transitions $n = 2$ to $n = 1$; the line marked K_β corresponds to $n = 3$ to $n = 1$. The graphs show gaps at values of Z corresponding to elements not discovered until later. (The unit of measurement for ν is the rydberg, R, equal to $2\pi^2 me^4/h^3$.)

Added to this was a property first recognized by W. Pauli in 1924: states defined by particular values of n, l, and m are in general split into two components of slightly different energy (as manifested, e. g., in the close doublet structure of many spectral lines). This property, subsequently interpreted as the consequence of an intrinsic spin and associated magnetic moment of the electron (G. Uhlenbeck and S. A. Goudsmit, 1925), meant that the full characterization of the quantum state of an electron in an atom required a total of *four* quantum numbers, the last of which simply took on the values $\pm\frac{1}{2}$ to correspond to the two possible quantized projections ($\pm h/4\pi$) of the spin angular momentum. This allows a total of $2(2l+1)$ different quantum states for given values of n and l.

An essential further ingredient had to be supplied, however. This was the *exclusion principle* of W. Pauli (1924), according to which no two electrons can have the same set of quantum numbers. This principle is the fundamental key to the internal structure of atoms, because it means that the electrons in a many-electron atom cannot all congregate in the lowest energy state ($n = 1$), but are forced to occupy "shells" of progressively increasing energy. We might expect that a given shell would be defined by a particular value of n, since for a *single* electron in the field of a central charge the energy depends only on n. In a many-electron atom, however, the situation is drastically modified by the partial screening of the nuclear charge by the electrons close to it. The innermost electrons, in states with $n = 1$, "see" almost the full nuclear charge Ze, but electrons in states of higher n are exposed to an effective central charge that is less than Ze and depends on l as well as n. Small l, in the quantum-mechanical picture, corresponds to a greater probability for the electron to be near the nucleus and therefore to be exposed to the full nuclear charge. The result of these considerations is to give rise to a fairly well-defined sequence of energy shells, each able to accommodate a certain maximum number of electrons, and corresponding to increasing energy and increasing mean distance from

Table 1: Electron shell structure.

Values of (n,l)	Shell capacity	Cumulative total
(1,0)	2	2
(2,0)+ (2,1)	8	10
(3,0) + (3,1)	8	18
(3,2) + (4,0) + (4,1)	18	36
(4,2) + (5,0) + (5,1)	18	54
(4,3) + (5,2) + (6,0) + (6,1)	32	86

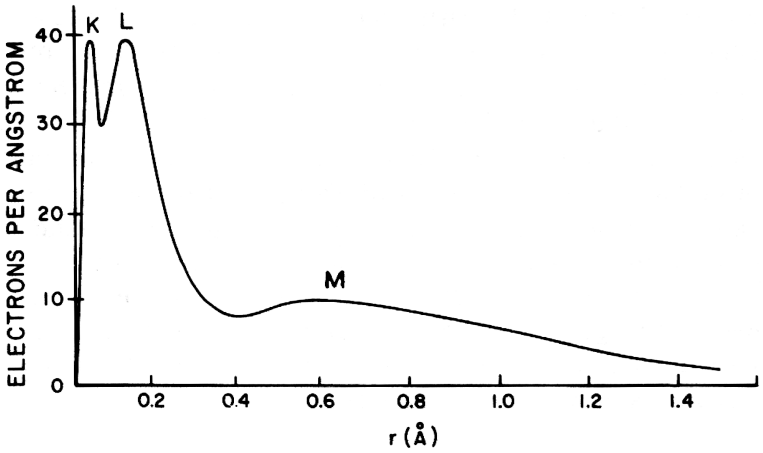

Fig. 3: The radial density distribution for the electron charge cloud in argon atoms, exhibiting the spatial shell structure as inferred from electron scattering experiments. The peaks marked K, L, M correspond to the approximate shell radii for $n = 1$, 2, and 3, respectively. [After L. S. Bartell and L. O. Brockway, *Phys. Rev.* **90**, 833 (1953).]

the nucleus. A partial tabulation showing the shell structure up to $Z = 86$ is shown in Table 1. Completion of a shell leads to a particularly stable and compact structure. The next electron added is relatively weakly bound and can stray quite far from the nucleus. This theoretical model is in good accord with the empirical features of the periodic table of the elements. The atomic numbers $Z = 2$, 10, 18, 36, 54, and 86 are those of the noble gases, He, Ne, A, Kr, Xe, and Rn. The atoms immediately following these ($Z = 3$, 11, 19, etc.) are the alkali metals, which have large atomic radii (cf. Fig. 1) and single, weakly bound valence electrons. Many other less prominent features of the atomic sequence can be understood within the framework of this theoretical model.

The reality of the shell structure has been directly demonstrated in a few cases by probing the atomic structure with x-rays or electrons. Figure 3 shows some experimental results on the radial variation of electron charge density in argon.

General Structure of a Massive Atom

To bring together the results discussed thus far, it may be helpful to consider the complete picture of a particular massive atom with many electrons. Let us take the most abundant isotope of tin, with $Z = 50$, $A = 120$.

The nucleus of this atom has a charge of $+50e$; it contains 50 protons and a number of neutrons equal to $A - Z$, i. e., 70 (but the nuclear structure is not the concern of this article). Closest to the nucleus (on the average) are two electrons in states with $n = 1$; these electrons are exposed to almost the full force of the nuclear electric field. According to Eq. (1), with $n = 1$, $Q = 50$, each electron would be bound with a negative energy equal to almost 2500 times the binding energy of the electron in the hydrogen atom (13.6 eV); this would be of the order of 30 keV. Direct measurements show that it takes x-rays with a quantum energy of 29.2 keV to dislodge an electron from this innermost shell. The average distance of these electrons from the nucleus is comparable to the orbit radius calculated from the Bohr theory for $n = 1$, $Q = 50$; this is about 0.01 Å, or 10^{-12} m – about 150 times the nuclear radius of tin.

Going to the next electron shell (cf. Table 1), we have eight electrons for which $n = 2$ and for which the nuclear charge is significantly shielded, first by the two innermost electrons; second, by the other electrons in this same shell; and third, by penetrating electrons from shells still farther out. The effective central charge is not easy to estimate in this case, but experiment shows that x-rays of about 4 keV can eject electrons from this shell. (Actually, this and subsequent shells have a substructure, based on the involvement of two or more values of the quantum number l, which we shall not go into.) Putting $n = 2$ in Eq. (1), and using the observed electron binding energy of about 4 keV, we would infer an effective central charge of about $35e$ and a mean shell radius of about 0.06 Å.

Proceeding in the same way, we find that the third shell (eight electrons, $n = 3$) has a critical x-ray absorption energy of about 0.9 keV, corresponding to $Q \approx 24$, $r \approx 0.2$ Å; and the fourth shell (18 electrons) has a critical x-ray absorption energy of about 120 eV, corresponding to $Q \approx 12$, $r \approx 0.7$ Å.

This accounts for 36 out of the 50 electrons in the atom. The remaining 14 belong to the fifth electron shell, which can accommodate up to 18 electrons. The situation in this region of the atom is complicated, but we know that these electrons are the main contributors to the outer parts of the atomic charge cloud, which extends to a radius of about 1.5 Å (see Fig. 1). To remove one of these electrons requires an energy of 7.3 eV (the first ionization potential of tin). Thus we see that the energy-level structure within the Sn atom ranges all the way from weakly bound electrons (less than 10 eV) to very tightly bound electrons (tens of keV).

The chemical and spectroscopic characteristics of an element depend on the details of the quantum states of its outermost electrons. In particular, the chemical valence depends on the extent to which the total number of electrons in an atom represents an excess or a deficit with respect to the more stable configuration of a completed shell. In the case of tin, for example, the total of 50 electrons is four short of completing the fifth shell of Table 1. In these terms we can understand why tin is quadrivalent (although it takes a study of finer details to understand in physical terms why it also exhibits divalency). In such matters as these, however, although the properties certainly have their complete basis in electric forces and quantum theory, the theoretical analysis is at best semiempirical.

See also: Atomic Spectroscopy; Atomic Structure Calculations, Electronic Correlation; Atomic Structure Calculations, One-Electron Models; Bohr Theory of Atomic Structure; Elements; Isotopes.

Bibliography

H. A. Boorse and Lloyd Motz, *The World of the Atom*. Basic Books, New York, 1966.
Max Born (tr. J. Dougal), *Atomic Physics*, 8th. rev. ed. Dover Publications, New York, 1989.
A. P. French and Edwin F. Taylor, *Introduction to Quantum Physics*. Norton, New York, 1978.
H. Haken and H. C. Wolf (tr. W. D. Brewer), *The Physics of Atoms and Quanta*, 6th. ed. Springer, New York, 2000.
G. P. Harnwell and W. E. Stephens, *Atomic Physics*. McGraw–Hill, New York, 1955.
G. Herzberg, *Atomic Spectra and Atomic Structure*. Dover, New York, 1945.
Alan Holden, *The Nature of Atoms*. Oxford (Clarendon Press), 1971.
H. G. Kuhn, *Atomic Spectra*. Academic, New York, 1962.
F. K. Richtmeyer, E. H. Kennard, and T. Lauritsen, *Introduction to Modern Physics*, 6th ed. McGraw–Hill, New York, 1969.

Auger Effect

B. Crasemann

An atom that contains an inner-shell vacancy becomes deexcited through a cascade of transitions that are due to two kinds of competing processes: x-ray emission and radiationless , or Auger, transitions. In either process, the original vacancy is filled by an electron from a higher-energy level. In radiative transitions, the released energy is carried off by a photon; in radiationless transitions, this energy is instead transferred through the Coulomb interaction to another atomic electron, which is ejected. The emitted electron is called a K–LL Auger electron, for example, if a K-shell vacancy is filled by an L-shell electron, and another L electron is ejected.

Direct experimental evidence for the existence of radiationless transitions was gained by Pierre Auger through cloud-chamber experiments reported in 1923. X-rays traversing the chamber produced photoelectrons; the tracks of these photoelectrons increased if more energetic x-rays were used. In addition to the photoelectron tracks, Auger observed numerous short tracks, each of which originated at the same point as a photoelectron track (Fig. 1). The length of the short tracks did not change with x-ray energy, but depended only on the kind of gas that was placed in the cloud chamber. Auger was able to show that the length of many of the short tracks corresponded to the energy that would be released in K–LL radiationless transitions. In some cases, Auger found additional tracks, caused by electrons emitted in a second (L–MM or L–MN) step of the radiationless deexcitation cascade of an atom.

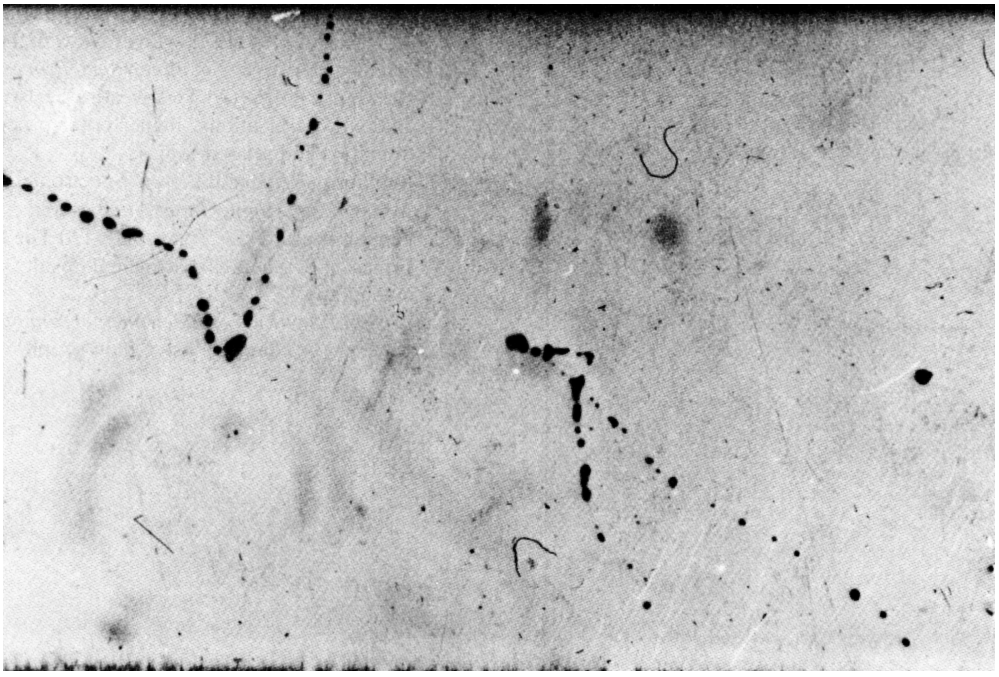

Fig. 1: Photoelectrons and Auger electrons from krypton ionized with 60-keV x rays, photographed in a cloud chamber by Pierre Auger *ca.* 1923. (Courtesy P. Auger).

The quantum-mechanical theory of radiationless transitions was formulated by G. Wentzel in 1927. From perturbation theory, the nonrelativistic matrix element for a direct Auger transition is

$$D = \in \int \psi^*_{n''l''j''}(1)\psi^*_{\infty l_A j_A}(2)\left|\frac{e^2}{\mathbf{r}_{12}^2}\right|\psi_{nlj}(1)\psi_{n'l'j'}(2)\,\mathrm{d}\tau_1\mathrm{d}\tau_2 \ ,$$

where the quantum numbers n, l, j characterize electrons that are identified schematically in Fig. 2. The state of the continuum (Auger) electron is labeled by $\infty l_A j_A$. In the physically indistinguishable exchange process, described by a matrix element E, the roles of electrons nlj and $n'l'j'$ are interchanged (Fig. 2). The total radiationless transition probability per unit time is

$$w_{fi} = \hbar^{-2}|D - E|^2 \ ,$$

if the continuum-electron wave function is normalized so as to correspond to one electron emitted per unit time. The matrix elements D and E can be separated into radial and angular factors. Evaluation of the angular factors depends on a choice of the appropriate angular-momentum coupling scheme. If spin–orbit interaction is neglected, the initial and final two-electron (or two-hole) states can be expressed for different values of the total angular momentum J in the ($LSJM$) representation of Russell–Saunders coupling. For the heavier atoms, inner-shell states are expressed more realistically in j–j coupling.

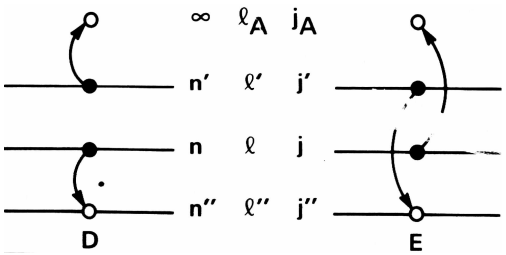

Fig. 2: Energy levels involved in the direct (D) and exchange (E) Auger processes, and notation for the principal, orbital-angular-momentum, and total-angular-momentum quantum numbers that characterize the pertinent electron states.

The selection rules governing radiationless transitions require that the total angular momentum and the parity of the final-state system (ion plus emitted electron) must be the same as of the initial-state ion. A large number of different transitions is generally possible from any given initial state; for example, 2784 matrix elements are required to describe the radiationless decay of an L_3 vacancy in a high-Z atom.

The relative probability that a K-shell hole is filled by an Auger process, rather than by x-ray emission, ranges from 0.999 for the lightest elements ($Z \leq 5$) to 0.02 for uranium. For vacancies in other shells, the Auger transition rate is generally several orders of magnitude greater than the radiative rate, and thus, essentially determines the lifetime of the vacancy. Auger rates can be very fast. Particularly intense are *Coster–Kronig transitions* that shift a hole to a higher subshell within one major shell. Thus, N_1–$N_{4.5}N_{4.5}$ "super-Coster–Kronig" transitions produce a 4s level width $\Gamma > 60\,\mathrm{eV}$ at $Z = 50$, which corresponds to an N_1 vacancy mean life of $< 10^{-17}\,\mathrm{s}$. The limits of validity of perturbation theory are strained when transition rates of such magnitude are calculated.

The Auger-electron energy is

$$E_{\infty l_\mathrm{A}} = E_{n''l''} - E_{nl} - E_{n'l'} \ .$$

The subscripts pertain to the states indicated in Fig. 2; $E_{n''l''}$ and E_{nl} are (absolute values of) neutral-atom binding energies, while $E_{n'l'}$ is the binding energy of $n'l'$ electron *in the presence of an nl vacancy*. The energy of electrons from solid samples can additionally include a contribution from extra-atomic relaxation, typically of the order of 10 eV.

Measurements of electron spectra from radiationless transitions (Fig. 3) and the theory of Auger processes constitute an active field of research, with relevance to fundamental atomic and solid-state theory as well as to surface physics, chemistry, and materials science.

New impetus has been given to Auger spectrometry in recent years with the advent of tunable synchrotron radiation. Selective photoexcitation of atomic subshells near threshold leads to Auger spectra that reveal details of electron rearrangement process, including correlation effects which produce a multifaceted many-electron response to inner-shell ionization. Important new insights into atomic structure and dynamics are being generated through such studies.

See also: Electron Energy States in Solids and Liquids.

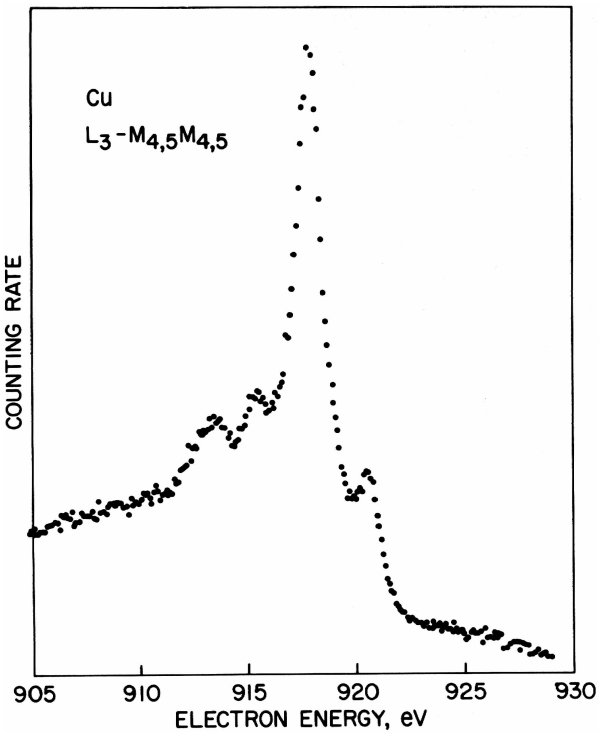

Fig. 3: Energy spectrum of L_3–$M_{4.5}M_{4.5}$ Auger electrons from metallic Cu. The separate peaks result from multiplet splitting due to various couplings of the two final-state holes. (Courtesy Dr. Lo I Yin, NASA Goddard Space Flight Center.)

Bibliography

P. Auger, "The Auger Effect," *Surf. Sci.* **48**, 1 (1975). (E) A first-hand account of the discovery of radiationless transitions.

For up-to-date atomic data see *Photon Interaction Data* (EPDL97); *Electron Interaction Data* (EEDL); *Atomic Relaxation Data* (EADL). Contact for users within the USA: National Nuclear Data Center (NNDC), Brookhaven National Laboratory, Vicki McLane (`services@bnlnd2.dne.bnl.gov`). Contact for users outside the USA: Nuclear Data Section, (NDS), International Atomic Energy Agency (IAEA), Vienna, Austria, Vladimir Pronyaev (`v.pronyaev@iaea.org`).

V. Schmidt, *Electron Spectrometry of Atoms using Synchrotron Radiation*. Cambridge University Press, Cambridge, 1997. A comprehensive treatise.

B. Crasemann, *Can. J. Phys.* **76**, 251 (1998). An overview.

R. H. Pratt, "Some Frontiers of X-Ray/Atom Interactions", in *X-Ray and Inner-Shell Processes*, R.W. Dunford *et al.* (eds.). American Institute of Physics, Melville, 2000; W. Mehlhorn, "Atomic Auger Spectroscopy: Historical Perspective and Recent Highlights", *ibid*.

G. Materlik, C. J. Sparks and K. Fischer (eds.), *Resonant Anmalous X-Ray Scattering – Theory and Applications*. North-Holland, Amsterdam, 1994. Summaries of recent research.

Aurora

E. C. Zipf

An aurora is an often spectacular optical phenomenon that occurs at altitudes above 95 km in the polar regions of the north and south hemispheres. Auroral displays occur most frequently in relatively narrow doughnut-shaped regions which encircle the geomagnetic poles. The Earth rotates underneath these patterns which are fixed with respect to the sun, resulting in a characteristic diurnal vibration in the morphology and temporal behavior of aurora at a particular geographic location. In addition to these geometrical effects, the physical thickness and diameter of the auroral ovals varies considerably with the degree of geomagnetic activity. Enhancements in the magnitude of the solar wind due to solar flares or to a general increase in particle flow from the sun during the normal solar cycle result in the equatorward expansion of the auroral zone which is also accompanied by an increase in the frequency of auroral displays and in their average intensity. The instantaneous position of the auroral zone is determined primarily by the precipitation of electrons with energies in the range 1–20 keV which are guided into the polar regions from the plasma sheet by the Earth's magnetic field lines. As these charged particles descend into the denser regions of the atmosphere, they lose their energy in inelastic collisions with atmospheric atoms and molecules. Most of this energy is consumed in ionizing, dissociating, and heating the atmosphere in the altitude range 95–300 km. Less than 5% of the energy of the primary particles is used to produce the visible radiation for which the aurora is so noteworthy. The plasma densities created by the precipitating electrons are comparable in magnitude to those produced by solar radiation in the normal ionosphere. When auroral activity is unusually intense, still larger ionospheric plasmas are created that cause interruptions in global radio communications (polar blackout).

The most common auroral form is the arc or band which is striking because of its extreme length (hundreds or thousands of kilometers) compared with its width (typically 1–10 km). Auroras are frequently classified in terms of their internal structure: homogeneous or rayed; their apparent motion: active or quiet; their brightness: on a scale from I through IV corresponding to an intensity variation from the brightness of the Milky Way to that of the full moon, respectively; and their visual color: the ordinary green or whitish aurora is designated Type C, while the dramatic veil auroras that fill the entire sky with red light are classified as Type-A aurora.

Three other types of auroral forms deserve special mention. The first is the evening hydrogen are which is produced as the result of proton and electron precipitation and appears in the form of a broad diffuse are equatorward of the main portion of the auroral oval during the evening hours, The second is the Polar Cap Absorption event (PCA) which is produced by very energetic solar cosmic rays (1–100 MeV) that enter the atmosphere over the entire polar cap down to a geomagnetic latitude of 60° where the cosmic ray cutoff limits further penetration. PCA events are associated with major flare activity on the sun and give rise to a bright uniform glow over the poles that is often accompanied by a complete radio blackout in the polar region. The third unusual auroral form is the midlatitude red arc (M-arc) which is a subvisual, broad arc elongated in the geomagnetic east–west direction and located generally between geomagnetic latitudes 41° and 60°. M-arcs are approximately 600 km wide in north–south extent and are found at altitudes above 300 km. Their east–west extent is for thousands of kilometers, possibly circling the entire globe. The light emitted from these arcs is essen-

tially monochromatic consisting of two atomic lines with wavelengths of 630.0 and 636.6 nm emitted by metastable oxygen atoms in the ^{2}D state. This unusual spectrum can be contrasted with the variety of features found in normal auroral radiation. These include many molecular bands emitted by N_2, N_2^+, O_2, and O_2^+ as well as more than 100 spectral lines radiated by atomic O and N and their ions.

Our knowledge of auroral morphology and spectroscopy has been enhanced significantly by recent sounding rocket and satellite experiments. The auroral spectrum has been measured quantitatively from the x-ray region (0.1 nm) to the deep infrared (100 μm), and excitation models for the principal emission features have been developed. Satellites with vacuum ultraviolet imagers have also discovered a new type of auroral form: the theta aurora. Theta arcs are aligned in north–south direction and bisect the auroral oval giving it the appearance of the Greek letter "θ" when viewed from afar.

Perhaps the most dramatic auroral phenomenon is the substorm or breakup event. This highly dynamic display develops from a quiet arc and is characterized by intense swirls, surges, and eddies of multicolored light that expand poleward from the original position of the are and, in a matter of minutes, cover most of the sky. These substorms will frequently last 20 min and are accompanied by disturbances in the local magnetic field that indicate the presence of large currents flowing in the auroral ionosphere. The field-aligned electric currents, which flow into the auroral zone from the magnetosphere and are the source of the magnetic effects associated with aurora, are called Birkeland currents. When these generally vertical currents enter the auroral ionosphere, they change direction and begin flowing horizontally in an east-west sense at altitudes near 100 km until they ultimately find their way along other field lines back to the plasma sheet. The horizontal portion of the current, which is often as much as 1000 km long, is called the auroral electrojet, and it carries currents as high as 10^7 A on some occasions. There is some evidence that the auroral substorms are triggered by plasma instabilities in the electrojet that develop and grow dramatically when the current density in the electrojet exceeds a threshold value. A variety of plasma waves are also generated in auroras. These include Alfvén waves as well as electrostatic drift waves. The excitation of these waves provides a mechanism for heating electrons locally to very high temperatures. As these energetic electrons cool, they produce vibrationally excited molecules that can modify the ion and neutral chemistry of the aurora.

Large electric fields ($\sim 50\,\mathrm{mV/m}$) are occasionally observed along the borders of auroral arcs. These fields produce kinetically energetic ions and contribute to the thermal economy of the upper atmosphere through Joule heating. Some auroras also generate substantial x-ray fluxes with energies of 20 keV or above. These x-rays penetrate down into the stratosphere where they contribute to the formation of nitric oxide. This mechanism is one example of how energy deposited high in the atmosphere (> 90 km) can be coupled into the stratosphere where it can affect the polar ozone budget and may influence long-term climate patterns.

See also: Arcs and Sparks; Atmospheric Physics: Ionosphere; Magnetosphere; Solar Wind.

Bibliography

Joseph W. Chamberlain, *Physics of the Auroras and Airglow*. American Geophysical Union, 1995. A still useful comprehensive text.

A. Valiance Jones, *Auroras*. D. Reidel, Dordrecht, 1974. A comprehensive contemporary text on the physics of auroras.

B. M. McCormac (ed.), *The Radiating Atmosphere*. D. Reidel, Dordrecht, 1971. Emphasis on auroral morphology, particle precipitation, and the aurora as a visual manifestation of large-scale magnetospheric processes.

A. Omholt, *The Optical Aurora*. Springer, New York, 1971. Text emphasizes the emission spectroscopy of aurora.