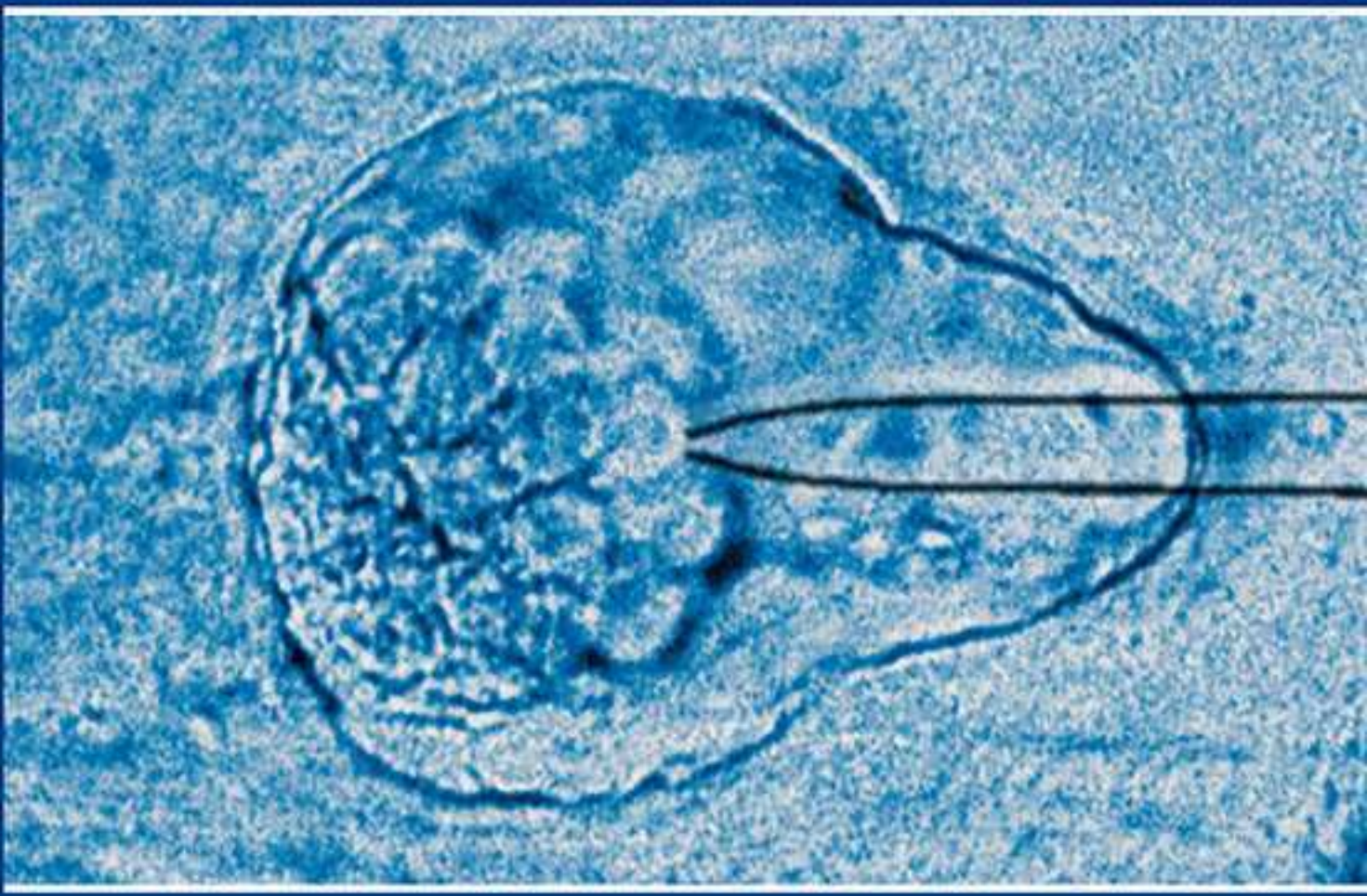
INTRODUCTION TO
Electrophysiological Methods and Instrumentation



FRANKLIN BRETSCHNEIDER
JAN R. DE WEILLE

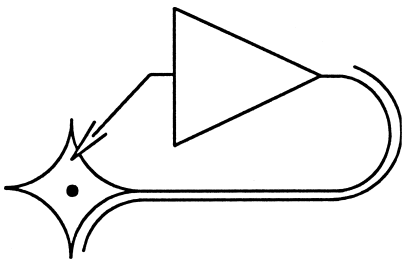
Introduction to Electrophysiological Methods and Instrumentation

This Page is Intentionally Left Blank

Introduction to Electrophysiological Methods and Instrumentation

Franklin Bretschneider
Functional Neurobiology Group
Department of Biology, Utrecht University
Utrecht, The Netherlands

Jan R. de Weille
Institut des Neurosciences de Montpellier
Hôpital Saint Eloi
Montpellier, France



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an Imprint of Elsevier

ELSEVIER B.V.
Radarweg 29
P.O. Box 211, 1000 AE Amsterdam
The Netherlands

ELSEVIER Inc.
525 B Street, Suite 1900
San Diego, CA 92101-4495
USA

ELSEVIER Ltd.
The Boulevard, Langford Lane
Kidlington, Oxford OX5 1GB
UK

ELSEVIER Ltd.
84 Theobalds Road
London WC1X 8RR
UK

First edition 2006

Library of Congress Cataloging in Publication Data
A catalog record is available from the Library of Congress.

British Library Cataloguing in Publication Data
A catalogue record is available from the British Library.

ISBN-10: 0-12-370588-6
ISBN-13: 978-0-12-370588-4

♾ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).
Printed in The Netherlands.

To our teacher and colleague, Dr Robert C. Peters

This Page is Intentionally Left Blank

Contents

Preface

Broadly defined, electrophysiology is the science and technique of studying the electrical phenomena that play a role in the life of plants and animals. These phenomena include the membrane potential, being ubiquitous among living cells, and its changes, which constitute signals playing an important part in the physiology of any organism. These signals may be slow changes caused by the changing concentration of some chemical substance, or the fast transient peaks called "action potentials" or "spikes", which arise by the fast opening of molecular "gates" in the membrane of neurons and similar types of electrically active cells.

Electrophysiological phenomena are the fastest signals in living nature: it has been found that, in directional hearing for example, time differences of less than 20 μs (between the arrival of sound in the left and the right ear) play a part. In addition to fast signalling, electrical processes have been proved to be useful for the sensitive detection of weak signals from the environment. Fish that are bestowed with electroreceptors—sense organs for electricity—have been shown to react to voltages as small as 1 μV across their body wall. The neural code, on the other extreme, is a strong, "sturdy" coding signal that will not be lost in long cables (nerves). By this virtue, the giraffe's brain can feel a mosquito crawling on its toes as precisely as one on its head. Equally powerful are the so-called motoric processes. Best known are the actions of our muscles, which are controlled and amplified largely by electrochemical processes. In addition, some electric fish emit strong pulses of electricity, over 500 V of tension, enough to stun their prey animals and to discourage most enemies. And, although plant life is usually more serene, electrical processes play an important part in metabolism; photosynthesis in particular. Some plants even generate "spikes", such as the well-known *Mimosa pudica* (sensitive plant). The fast withdrawal reflexes are mediated by electrically spreading action. Even though the chemical processes that support life as a whole may be important, studying the fast electrical processes is a fascinating branch of the life sciences, both for the basic satisfaction of our curiosity and for medical purposes.

To the general public, electrophysiological methods are best known from the latter category: ECG (electrocardiogram) and EEG (electroencephalogram) are terms well known from newspapers and television programmes. Apart from these famous clinical applications, a host of methods is in use, such as the recording from muscles (EMG or electromyogram) and from the eye (ERG, or electroretinogram), the electrical measurement of eye movements (electronystagmogram), the recording of nerve activity, the recording of the activity of single cells, whether nerve cells or not, and, through the "patch-clamp" technique introduced mainly by the Nobel laureates Neher and Sakmann, the recording of the activity of single ion channels, those tiny electrical gates in the membranes of all kinds of cells.

Working in electrophysiology implies, apart from a thorough knowledge of these phenomena, an equally well-founded control of one's equipment. In the early days of electrophysiological recording, amplifiers and other tools were often built by the physiologists themselves. Nowadays, many types of instruments for recording, processing and stimulation, versatile and

almost perfect, can be delivered off the shelf from a host of reliable companies. *This does not, however, absolve the user from the obligation to maintain and use the apparatus properly, especially since a lack of knowledge about one's tools may lead to the publishing of erroneous results, which is a waste of time, money and intellectual effort*.

Therefore, despite the streamlined technology, the many electronic devices and computer algorithms available for filtering or post-processing of the signals and for the presentation of the data, all students of electrophysiology must gain proper insight into the working principles of their principal tools, and more specifically of vital components like preamplifiers and electrodes, which are connected to the preparation, the part of living nature that is to be studied. In planning experiments, with the concomitant purchase of instruments, one has to know the possibilities to choose from, and the consequences for the validity of the measurements.

Since most of these instruments depend heavily on electronic circuitry, introductory electronics takes the major part of this book (Chapter 2). In addition, however, we will spend some time on electrochemical processes, such as the ones that are inherent in the use of electrodes, salt bridges and the like, and on the electric—in fact also electrochemical—processes of life itself (Chapter 3). Although this is not meant as a book on electricity theory, we will inevitably spend a few pages on the most basic electrical quantities and processes such as charge, voltage and flow of current through so-called "passive" parts like resistors and capacitors, circuits like voltage sources and filters, and so on (Chapter 1). The complexity of electrophysiological signals and the knowledge to be derived from them lead to many forms of signal and data processing. The spectrum of methods (Chapter 4) may range from the simple recording of an ECG, judged by the eye, to the statistical processing of single-channel opening times by dedicated software and from a dot display of spikes on a computer screen to the analysis of stochastical point processes. In all cases, a proper introduction to the mathematical and electronic procedures involved leads to a better understanding of what is going on, and so reduces the risk of failure.

This book is intended for all students of electrophysiology, especially for readers without a formal training in electronics, signal analysis or electrochemistry, and hopes to serve as a thorough, yet easy-to-digest introduction that should lead all the way up from a first recognition of principles, to both understanding, and the routine application of, the various methods. To this end, this book uses informal language with qualitative explanations, yet using sufficient math to enable the reader to grasp the processes at a sufficiently quantitative level. Most of the jargon, essential if one is to discuss fluently in the area, will be introduced properly, while an index to the key words permits cross-references. This book is as concise as to be useful as a direct study guide, yet may also be used as a work of reference.

Because the principles of electronics are described in an elementary, yet detailed way, and because the discussion is extended to deal with digital instruments, including computer algorithms and mathematics, we hope this book will be useful as a general introduction in instruments and methods, also to people outside the field of electrophysiology. Additional matter is treated in appendices, such as the issue of safety in electrophysiological set-ups and the use of CRT monitors. Thematic literature references and an extensive keyword index complete the book.

Although written down by two authors, this book is the product of years of experience and cooperation with many colleagues and students. In the first place, we would like to thank our teacher and colleague Dr R.C. Peters, who laid the foundation for this book, and encouraged one of us (FB) to extend, improve and publish it over the many years of our cooperation. We owe him many contributions and suggestions. Appendix D, on CRT screen technology, was

suggested by Dr R.J.A. van Wezel and commented on by him and by Mr J. Duijnhouwer. Dr K. Britten kindly provided Fig. D-5. Dr P.F.M. Teunis gave valuable comment and kindly provided the statistical data pertaining to gamma distributions. Many more people provided valuable comment on the first draft, among them Dr A.C. Laan, Mr W.J.G. Loos, Mr R.J. Loots, Mr A.A.C. Schönhage, Mr R. van Weerden, Dr T. Sanderson and several anonymous referees. We also acknowledge the encouragement by Prof. Dr A.V. van den Berg and Prof. Dr W.A. van de Grind. We also acknowledge the smooth cooperation of Dr J. Menzel, Ms M. Twaig and other people at Elsevier. Finally, we would like to thank all our students for explicit or implicit contributions, and for their patience with the earlier versions of this book.

F. Bretschneider
J.R. de Weille
Utrecht, 15 September 2005

1

Electricity

ELECTRICAL QUANTITIES

Although most of the vast field of electricity theory is outside the scope of this book, we will certainly deal with the handful of quantities that play an important part in electrophysiology—for the understanding of instruments as well as electrochemical and neurophysiological processes.

Electric Charge, Current and Potential

The basic quantity is the electric charge, buried in the atomic nucleus as what we call a positive electric charge, and in the electrons surrounding it, which we call negative charge. The unit of electric charge (abbreviated Q) is the coulomb (abbreviated C), defined in (macroscopic) electric circuits in the eighteenth century.

The underlying fundamental constant, found much later (in 1909, by R. Millikan), is the elementary charge, the charge of one electron, which amounts to 1.6021×10^{-19} C. Since this "quantum of electricity" is so small, most electric phenomena we will describe may be considered as continuous rather than discrete quantities.

By the nature of atoms, most substances, and indeed most materials in daily life, are neutral. Obviously, this does not mean that they have no charges at all, but that (i) the number of positive charges equals the number of negative charges and (ii) the opposite charges are so close together that they are not noticeable on a macroscopic scale. This means that a number of substances can be "teased" to release electricity, e.g. by rubbing together. This was indeed the way electricity was discovered in antiquity, and it was examined more systematically from the eighteenth century on. Many science museums are the proud owners of the large static electricity generators invented by, among others van Marum and Wimshurst. These machines generated rather high voltages (around 50 000 V), but at very low current strengths (1 μA), and so were not of much practical use.

Nowadays, most sources of electric energy are electrodynamic, such as the generators in our power plants, as well as in cars and on bicycles. In addition, the electrochemical processes, found originally by Galvani and Volta, are employed in the arrays of galvanic cells we call batteries and accumulators. Both forms of sources deliver the electrical energy at lower voltages

(say, 12 V), but allow far larger currents to be drawn (hundreds of amperes in the case of a car battery).

This brings us to the most important quantities to describe electrical phenomena: the unit of tension, the volt (V), and the unit of current, the ampere (A). Note that the correct spelling of the units is in lower case letters when spelled out, but abbreviated as a single capital. In Anglo-Saxon countries, tension is often called "voltage". Both units have practical values, i.e. it is perfectly normal to have circuits under a tension of one volt or carrying one ampere in the lab or even at home. The definitions are derived from other fundamental physical quantities:

Charge (Q): one coulomb is defined as the charge of $6.241\,460 \times 10^{18}$ electrons.

Current (I): one ampere is a current that transports one coulomb of charge per second.

An overview of electrical quantities, their units and symbols is given in Appendix A.

The origin of the definition of tension, or potential difference, is a bit more intricate. The electrical forces that act on charges (or charged objects) depend not only on the field strength, but also on the distance travelled. Thus, electrical potential (abbreviated as U) is defined in terms of the amount of energy, or work (abbreviation W), involved in the movement of the charge from a certain point in the electric field to infinity (where the electrical forces are zero). If one does not move to infinity but from one point in the field to another point, less energy is involved. This is called the potential difference between the two points. Where to choose the two points will be a matter of practical, quantitative discussion.

Electrophysiologists measure what they call membrane potential by sticking one electrode into a cell. Theoretically, then, the reference electrode should be placed at infinity, where the potential is defined to be zero. In practice, however, the potential difference between inside and just outside the cell is measured. For this purpose, the potential just outside the cell can be considered to be sufficiently close to zero. This is caused by the fact that the membrane has a resistance that is many orders of magnitude higher than that of the fluids inside and outside the cell.

Other circumstances, however, change this view radically: many electrophysiological quantities are recorded entirely outside the cells, such as electrocardiogram, electroencephalogram and a host of signals from nerves and muscles. In this case the potential outside the cell cannot be considered zero! Instead, the potential difference between two extracellular points constitutes the whole signal. Nevertheless, the potential difference across the cell membrane is called potential in the long tradition of electrophysiology. The unit of tension, or potential difference, is the volt.

Tension (U): one volt is the tension between two points that causes one joule (J) of work (W) to be involved in carrying one coulomb of charge from one point to the other.

We use "involved" because the energy is either necessary for or liberated by the movement, depending on the direction.

Resistance

The concept of resistance stems directly from these fundamental quantities: if a certain current flows through an object as a consequence of a tension applied to this object, it exhibits the phenomenon of resistance, which is defined as the ratio of voltage to current.

Resistance (R): one ohm (Ω) is one volt per ampere.

These relations are remembered better in the form of formulas:

$$\begin{array}{rrrr} I = Q/t & \text{or} & \text{ampere} = \text{coulomb/second}: & 1\,\text{A} = 1\,\text{C/s} \\ U = W/Q & \text{or} & \text{volt} = \text{joule/coulomb}: & 1\,\text{V} = 1\,\text{J/C} \\ R = U/I & \text{or} & \text{ohm} = \text{volt/ampere}: & 1\,\Omega = 1\,\text{V/A} \end{array}$$

The latter law is known as Ohm's law, and is very familiar to all people that handle electrical processes. It is often seen in two other forms, depending on which is the unknown quantity:

$$U = IR \quad \text{and} \quad I = U/R$$

This means that, knowing any two quantities, Ohm's law yields the third one. This is used very frequently. In electrophysiology, for instance, one needs to calculate electrode resistances from the voltage that develops when feeding a constant current through the electrode, and membrane resistances from measured current values together with the clamping voltage, and so on.

Resistance is the property of an object, such as a micropipette or a cell membrane. Solids, such as copper, and fluids, such as water, also have resistance, but the value depends on the dimensions of the body or water column. The resistance per unit of matter is called "specific resistance" or "resistivity". The dimension is Ωm ("ohm metre"). In electrochemistry, where the small unit system (cgs system) is still used frequently, the unit of resistivity is Ωcm. As a guideline, freshwater has a resistivity of about $1\,\text{k}\Omega\,\text{cm}$ ($10\,\Omega\,\text{m}$), seawater about $25\,\Omega\text{cm}$ ($0.25\,\Omega\text{m}$). Obviously, metals are better conductors, i.e. they have far lower resistivity values: in the order of $10^{-5}\,\Omega\text{cm}$.

The dimension "ohm metre" may seem odd at first, but is easily explained since the resistance is proportional to the length of a water column, and inversely proportional to the cross-section, which is width $\times$ height, or the square of the diameter. So, it is actually a simplification of $\Omega\,\text{cm}^2/\text{cm}$.

Other, related quantities we have mentioned already are power and energy, or work. The quantity energy (symbol W) has a unit called joule (J). The related, often more interesting quantity of energy per unit of time is called "power" (symbol P), and has the unit watt (W). So, the performance of loudspeakers, car motors and stoves is expressed in W. The longer they are used, the more energy is spent (which must be paid), but power is the best characteristic. Electrical power depends on voltage, current and, through Ohm's law, on resistance:

$$P = UI; \quad \text{or} \quad P = I^2R; \quad \text{or} \quad P = U^2/R \quad (1\,\text{W} = 1\,\text{VA} \quad \text{or} \quad 1\,\text{W} = 1\,\text{V}^2/\Omega)$$

and so on. Work is simply power times time:

$$W = Pt \quad \text{or} \quad W = I^2Rt$$

and so on. The latter formula is known as Joule's law.

Capacitance

The quantity to be discussed next is capacitance. This is the ability to store electric charge associated with a voltage. Now what is meant by "store"? The phenomenon shows up, either wanted or not, when two conducting wires, or bodies in general, are brought close together. If one of the conductors carries a positive charge, and the other one a negative charge, a (relatively high) voltage exists between the two. When brought closely together, however, the electric fields influence each other, thereby partially neutralizing the effect (If there are equal positive and negative charges or if the charges have the same centre of gravity, the net result would be zero charge, or neutrality. This is why atoms in general are neutral). In other words, by bringing two conductors together, the voltage decreases. Therefore, the charge is partially "hidden" or stored. The shorter the distance, and the larger the surface area, the more charge can be stored.

Note that this works only if the two conductors are separated by a very good insulator, such as a vacuum or dry air. Otherwise, a current would neutralize the charges. Other good insulators are glass, most ceramics and plastics. Note also that charge storage is different from what we saw with resistors: a voltage exists across a resistor only as long as a current is flowing through it; the moment the current stops, the voltage will be zero. A capacitance behaves differently. This can be seen by comparing electric quantities with hydraulic ones. Capacitance is an analogue of a vat or water butt. The amount of water is the analogue of an electric charge, the flow of water is the analogue of an electric current, and the water level corresponds to an electric voltage. When water flows in a vat, the water level builds up slowly, depending on the total amount of water poured in. In a small vat, a certain water level is reached with a smaller amount of water than in a large vat. The larger vat is said to have a larger storage capacity.

In the same way, a capacitor is a vat for electric charge, and the word "capacitance" is derived directly from this analogy. The unit of capacitance (symbol C) is farad (symbol F, after Faraday).

Capacitance (C): one farad is the storage capacity that causes a tension of 1 V to arise by transferring one coulomb of charge.

$$\text{or} \quad C = Q/V$$

Check the following derived formulas:

$$C = It/V; \quad C = t/R; \quad \text{and} \quad Q = CV$$

The capacitance exhibited by two conductors depends on distance, and hence on the form of the objects. Wires, spheres and irregular shapes have part of the surface area closer, and part farther from the other conductor. For two parallel plates, the capacitance can be calculated easily:

$$C = \varepsilon_0 \varepsilon_r A/d$$

Here, A is the surface area ($l \times w$ for a rectangle, πr^2 for a circle), d is the distance between the plates, ε_0 is a constant called the "absolute permittivity" of free space, also absolute dielectric constant, and has a value of 8.854×10^{-12} F/m. Finally, ε_r is called the "relative permeability" or "relative dielectric constant", often dielectric constant for short, and is determined by the material between the plates. By definition, vacuum has a dielectric constant of unity, air has

a value only slightly higher (1.00058), but other insulators have more marked effects on the capacitance. The list below gives approximate values (the dielectric constant is temperature-dependent, and often also frequency-dependent):

glass	about 3.8–6.7
mica	6.0
paper	3.7
polyethylene	2.35
PVC	about 4.5
china	about 6–8
water	about 81

Apparently, the dielectric medium influences the electric field in the gap between the plates. The relatively high value of water is due to the dipole form of the water molecules, together with their mobility in liquid water. Thus, a large part of an applied electric field "disappears" in the re-orientation of water molecules.

Magnetism

Magnetic processes, like electrical ones, are known since antiquity. The intimate relations between the two were found much later, however: only about two centuries ago.

The main relation is the following: an electric current induces a magnetic field, but the converse is not true—a magnetic field does not induce an electric current. Only *a changing magnetic field* generates an electric current. In fact, this is the main cause for the popularity of alternating current in society, such as the electrical power lines.

Because of this relation, magnetic quantities are often expressed in electrical units. A useful measure of magnetic processes is the magnetic flux density, also called magnetic inductance, or inductance for short, which has a unit called "tesla".

Inductance: one tesla (T) is the inductance that exerts a force (F) of 1 newton (N) at one metre (m) distance from a wire carrying a current of 1 A.

$$\text{or} \quad 1\,\text{T} = 1\,\text{N}/(\text{Am})(\text{"newton per ampere metre"})$$

The term "flux density" can be seen by the following relation:

$$1\,\text{T} = 1\,\text{J}/(\text{Am}^2)(\text{"joule per ampere metre squared"})$$

which means that, integrated over one square metre, a magnetic induction corresponds to an amount of energy.

Self-Inductance

The interaction of electric current and magnetic field leads to the notion of self-inductance in the following way: a changing electric current induces a changing magnetic field, but a changing magnetic field induces a current again. Since this induced current is in the opposite direction,

the net effect is that the current that flows is lower than it would be without self-inductance. The unit of self-inductance (symbol L) is called henry (symbol H). Since self-inductance arises by changing current strength, the description of the process involves time, and follows from the relation:

$$U = -L \, \mathrm{d}I/\mathrm{d}t$$

In words, the induced voltage is proportional to the change of current strength $\mathrm{d}I/\mathrm{d}t$ and the magnitude of the self-inductance L (and has the opposite polarity). Hence:

Self-inductance: one henry is the self-inductance that causes a voltage of one volt to develop when the current strength changes (minus) one ampere per second.

Apparently, self-inductance is a fundamental property of any conductor carrying current. Nevertheless, self-inductance is strongest when a long wire is wound into a coil, also called an inductor or solenoid. By joining many turns of wire together, the magnetic fields caused by one and the same current are added. Adding a core of iron, ferrite or any other ferromagnetic material enhances the magnetic induction still further. So, the magnetic properties of iron can be expressed in a quantity similar to the dielectric constant in storing electric charge. It is called "relative magnetic permeability", and has the symbol μ_r. Vacuum and many materials have a permeability of virtually unity. Materials that have a (very slightly) smaller value are called diamagnetic (μ_r about 0.99999); materials with a slightly higher permeability (about 1.01) are known as paramagnetic. In fact, all materials show diamagnetism to some degree, but in paramagnetic materials, the latter property dominates the first. Far more conspicuous are the so-called ferromagnetic materials, which have permeability values far higher than unity. The best known is iron, of course, but some iron oxides (ferrites), nickel and chrome are also ferromagnetic. The table below shows approximate values for a few well-known materials:

cast iron	600
ferrite	1000
pure iron	5000
alnico	8000
mu metal	20 000
permalloy	10^5
supermalloy	10^6

The highest values are for metal alloys specifically designed to have extremely high μ_r. Since μ_r depends on the magnetic field strength, the values in the table are approximate maximum values. Because of this behaviour, ferromagnetic materials are used as core materials in electromagnets, transformers, etc. and in audio and video tapes, computer discs, etc. Iron is suited only for low frequencies (50 Hz–50 kHz), and ferrites for both low and higher frequencies (say up to 30 MHz). At still higher frequencies, self-inductance gets so dominant that no core is needed. This is exploited in radio and television sets, but does not play a role in electrophysiology. Note, however, that coils made to block (h.f.) radio waves from an electrophysiological set-up must use ferrites as core materials to be effective.

Direct and Alternating Current; Frequency

An electric current that flows in a certain direction, and does never change direction, is called a direct current, abbreviated DC (or dc). In a more strict sense, a DC is a current that is constant over time. The voltage associated with a direct current is called a "DC voltage". Static electricity, such as that caused by the charge built into "sticky" photo albums, and the voltages of galvanic cells (batteries and accumulators) are examples of DC voltages. Again, the term is used frequently in the more strict sense of a constant DC voltage. A battery or other power supply is said to deliver a DC voltage. For the batteries, this is only approximately true, since in most types the voltage decreases slowly by exhaustion during use. Nevertheless, the notion of a DC is useful in contrast with an alternating current, or AC, which is a current changing direction all the time, usually on a regular basis. The mains voltage, for example, changes direction 50 times per second (in America 60). Strictly speaking, the mains voltage changes direction 100 (or 120) times per second: first from plus to minus, then from minus to plus again. The convention, however, is to count the number of repetitions, or cycles per second, called the "frequency". The unit of "per second", or "s^{-1}", is hertz (Hz).

In electrophysiology, one distinguishes frequency from rate: a sinusoidal or other well-known waveform, such as a sound wave, is said to have a frequency (in Hz), whereas a pulse train, such as a train of spikes (nerve impulses) is said to have a certain rate. This distinction is made for an important reason: if a frequency is changed, all parts of the process are accelerated or slowed down. A receiving apparatus, such as an amplifier, must be adjusted to allow the changed frequency to pass. Spikes, however, have the same shape, and hence the same speed the voltage is changing with, irrespective of their rate. This has consequences for the recording apparatus (amplifiers): see Chapter 2. To underline the difference, a separate unit called the "adrian" (after the pioneer electrophysiologist Lord Edgar D. Adrian) had been proposed, but did not catch on in the community of electrophysiologists. For rates, the unit is usually notated simply as "/s", "s^{-1}" or "sp/s". The difference is illustrated in Fig. 1-1. The sinusoidal shape is considered to be the "basic" alternating current. There are several reasons to do this. In the first place, a sine wave arises in electric generators, such the bicycle dynamo, by rotating a magnet in a coil, or alternatively rotating a coil in a magnet.

Thus, most electric power distributed in society is made up of a sine wave at 50 Hz (in America 60 Hz). Secondly, it can be shown mathematically that, upon transforming a signal waveform to the frequency domain, a sine is the "building element" having only a single

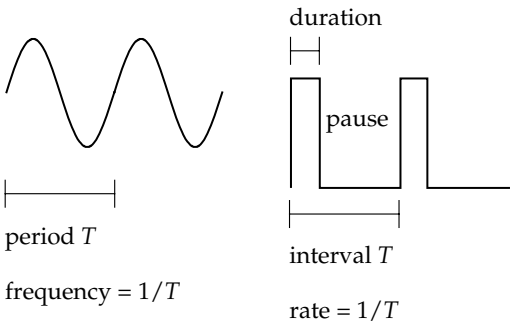


Fig. 1-1 Parameters of (sine) frequency and (pulse) rate.

frequency. Other waveforms, such as square wave, sawtooth or triangle, can be considered as combinations of sine waves at different frequencies. A square wave with a frequency of 100 Hz, for instance, has components at 300, 500, 700 Hz and so on. These higher frequencies are always a multiple of the basic, or fundamental, frequency, and are called "harmonics", "overtones" or "partials". Note that the fundamental frequency, or fundamental for short, is also called the "first harmonic".

Actually a sine wave should be called a "sinusoidally changing voltage" rather than a "wave", since the word "wave" means the spatial spreading of disturbances in a medium, whereas our sine "wave" is only present at the output terminals of the instrument. Nevertheless, it is a common habit to speak of sine and square waves.

A sine wave is a function of time, and so the voltage at any moment can be described by the following function:

$$U_t = U_{max} \sin(\omega t)$$

Here, U_t is the voltage at any moment t, U_{max} is the maximum voltage, usually called the "amplitude", and ω is the angular frequency, i.e. the number of radians per second. The angular frequency is related to the frequency (f, the number of cycles per second) by

$$\omega = 2\pi f$$

When connected to a resistor, an alternating voltage will give rise to an alternating current

$$I_t = I_{max} \sin(\omega t)$$

where I_t and I_{max} are the momentary current and the current amplitude respectively. The relation of voltage and current are again determined by Ohm's Law:

$$U_t = RI_t \quad \text{and} \quad U_{max} = RI_{max}$$

Reactance

Things get more complicated when we apply alternating voltages and currents to parts that are not simply resistors. For example, we have seen earlier that applying a DC voltage to a capacitor will lead to charging, but once the capacitor is charged to the full input voltage, no current will flow: the two conductors are separated by an insulator, allowing no current to flow between them. When one applies an AC voltage to a capacitor, however, a current (an alternating current) will flow, and will keep flowing as long as the AC voltage is applied. Because the input voltage changes polarity many times per second, the capacitance will be charged positively, then negatively, then positively again, and so on. Thus, despite the insulator, a current seems to flow through it.

Note that although no charge can flow through the insulating layer, the AC current flow through a capacitance is very real. It is not difficult to understand that the magnitude of the current that flows will depend on the capacitance value: a large capacitor stores a larger charge at a given voltage, and will sustain a larger current when reverse-charged many times per second. Thus, a 1 pF capacitor will support only a minute current, even when connected to the 230 V mains. In fact, this is the situation that arises when one picks a mains cable with the

hand: the so-called stray capacitance formed by the copper leads and the hand, separated by the plastic insulation, amounts to one or a few picofarad. The resulting current, about 300 nA, is harmless, and even imperceptible. To the contrary, a capacitor of, say, 100 µF would support a rather large current when connected to a high-voltage AC circuit, so that it could be fatal to touch.

Although the mentioned currents resulting from stray capacitances are minute, they can be measured, and are often picked up inadvertently by sensitive electronic devices, causing a 50 or 60 Hz interference called "hum". This is frequently a nuisance for electrophysiologists trying to measure microvolts in laboratory rooms, fitted with 230 V (or 120 V) mains cables, outlets, lighting and the like. The best strategy is to keep mains cables far from sensitive instruments, and especially as far as possible from the specimen in an electrophysiological set-up. Alternatively, the specimen, or the entire set-up, can be kept in a Faraday cage, which will screen any static electromagnetic fields effectively.

Apart from the capacitance value, the current through a capacitance will depend on the frequency, i.e. on how often the charge is reversed. Thus, we need a notion, similar to resistance, to describe the ability of a capacitance to sustain a current. This is called the "reactance", or capacitive reactance, and can be computed using the following formula:

$$I_{\max} = C\ \mathrm{d}U/\mathrm{d}t = C\ \mathrm{d}(U_{\max}\sin(\omega t))/\mathrm{d}t$$

which can be converted to:

$$I_{\max} = \omega C U_{\max}\cos(\omega t)$$

The capacitive reactance (symbol X_C) follows from the ratio of voltage to current:

$$X_C = 1/(\omega C)$$

Like resistance, it is expressed in ohm. Simple calculations show that at 50 Hz, a 100 pF capacitance has a reactance of about 32 MΩ, one of 100 µF of 32 Ω. Estimating the capacitance of a person standing next to a mains outlet at 10 fF, the current flow would be 690 pA. Although this is small (and harmless), it is nevertheless far more than the current through a single ion channel. Note that at 10 MHz, a 100 pF capacitance would have a reactance of only 160 Ω. Thus, a capacitor that blocks the mains frequency may pass radio-frequency interference (RFI).

By a similar argument, the way in which an inductance (solenoid) interacts with an AC can be described. Because the counter-voltage that develops in a coil depends on the change of current strength, the reactance of a self-inductance, or inductive reactance (X_L) is also frequency-dependent:

$$X_L = \omega L$$

Here, the reactance increases with frequency. Since an inductance does not block a direct current, it can be used to smooth the current from mains-operated power supplies (see Chapter 2).

Current and Voltage Sources

The theoretical entities (ideal) voltage source and (ideal) current source play important roles in electricity theory. A voltage source is defined as a component that maintains a constant

voltage across its two terminals, irrespective of the current drawn from it. We will discuss later on to what extent practical voltage sources can approximate this ideal. In an analogous way, a current source is defined as a component that drives a certain current strength through any connected load, irrespective of the voltage that is needed to do so (and hence irrespective of the resistance of that load). Again, practical current sources will only approximate this ideal.

COMPONENTS, UNWANTED PROPERTIES

Most of the above-mentioned electrical quantities can be put in the form of components, where each property has some well-defined value. A component sold specifically for the property of resistance is called a "resistor". The usual construction is a small rod with two connecting wires. The rod is either made of a composite material having the desired resistance (composition resistors) or of an insulating ceramic material covered with a film of metal or carbon (metal film and carbon film resistors). It is important to note that the properties of carbon composition resistors are rather bad: they have a high temperature coefficient, a high drift and aging, and are by far the noisiest type. Carbon film resistors are slightly better, but metal film is the best buy. Metal film resistors have a far lower temperature coefficient and are more precise and stable than the other two types of resistors. A fourth type is the wire-wound resistor, in which a wire of a metal alloy, again with a low temperature coefficient, is wound on a ceramic body. Metal film resistors are used most frequently today, but one is not always free to choose: wire-wound resistors are only practical from arbitrarily low values (0.1 Ω or less) up to about 10 kΩ. In addition, they may have too much inductance for high-frequency applications. Metal film resistors are made from 100 Ω up to about 5 MΩ. For still higher values, as often used in electrophysiology, one has to rely on carbon film resistors, and compensate for the greater errors involved, e.g. by selection of components and by calibration of the instruments.

The value of the resistor is printed on the rod body, usually as coloured rings (see Fig. 1-2). This is done because printed numbers wear off very quickly, notably the decimal

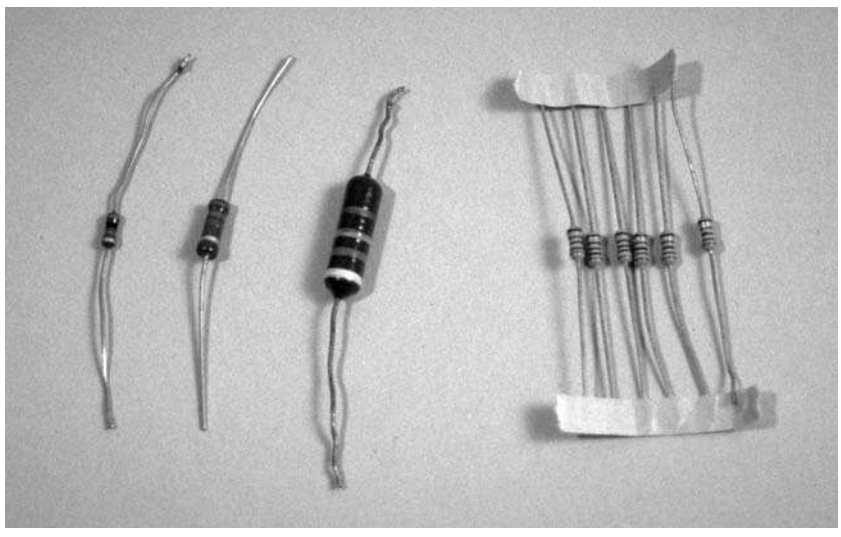


Fig. 1-2 Practical colour-coded resistors.

point. Therefore, in cases where the value is printed in numbers on the component's body, the decimal point is replaced by the multiplier, e.g. "5k6" is printed instead of "5.6 kΩ", "1n2" instead of "1.2 nF".

The colour code is as follows: carbon resistors have three rings. The first two stand for digits of the resistance value, the third one codes for a multiplication factor, i.e. the number of zeros following the two digits. The meaning of the colours is shown below; it is rather easy to remember: the ten digits are formed by the colours of the spectrum, preceded by earth colours (black and brown), and followed by sky (i.e. cloud) colours (grey and white):

black	0	or no zeros
brown	1	or 0
red	2	or 00
orange	3	or 000
yellow	4	or 0 000
green	5	or 00 000
blue	6	or 000 000
purple	7	or 0 000 000
grey	8	or 00 000 000
white	9	
gold	–	1/10
silver	–	1/100

Thus, a resistor with red, purple, yellow has a value of 270 000, or 270 kΩ, and a resistor with orange, white, gold has a value of 3.9 Ω (39×0.1).

Things can be more complicated, however. The above-mentioned resistors have usually a fourth ring that codes for the tolerance, or accepted, one-sided error. Here, gold signifies 5%, and silver 10% tolerance. Since metal film resistors are far more stable, they usually have better tolerances: usually 2% or 1%. Therefore, these resistors have five rings: three digit rings, a multiplier and a tolerance ring, the latter mostly being red or brown.

A component that behaves as an almost pure capacitance is called a "capacitor". In principle, capacitors consist of two conducting plates, or electrodes, separated by a thin insulating layer. However, practical designs may differ widely, depending on the desired magnitude range, the necessary precision and the voltages the insulator must be able to withstand (see Fig. 1-3). Capacitors can be obtained in the range from about 1 pF (10^{-12} F) up to about 10 mF (the abbreviation MFD, still printed occasionally on capacitors, means μF). The first capacitors, used to collect static electricity by eighteenth-century pioneers, consisted of a glass jar, covered with aluminium on both sides. These "Leyden jars" could withstand thousands of volts. The smallest modern capacitors (1 pF–10 nF) consist of a mica sheet, silvered or aluminized on both sides. These capacitors are very stable and may be manufactured to narrow tolerances. Larger values (about 1 nF–5 μF) are obtained with a sandwich of two aluminium foils with a plastic foil in between. These are handy, cheap and fairly stable. The values are less precise, however. Capacitors with still larger values would either grow to unwieldy sizes or need impractically thin plastic films. Therefore, a metal-oxide coating on an aluminium foil is used as insulator instead. These capacitors consist of a metal can, filled with a conducting (yes, conducting) salt

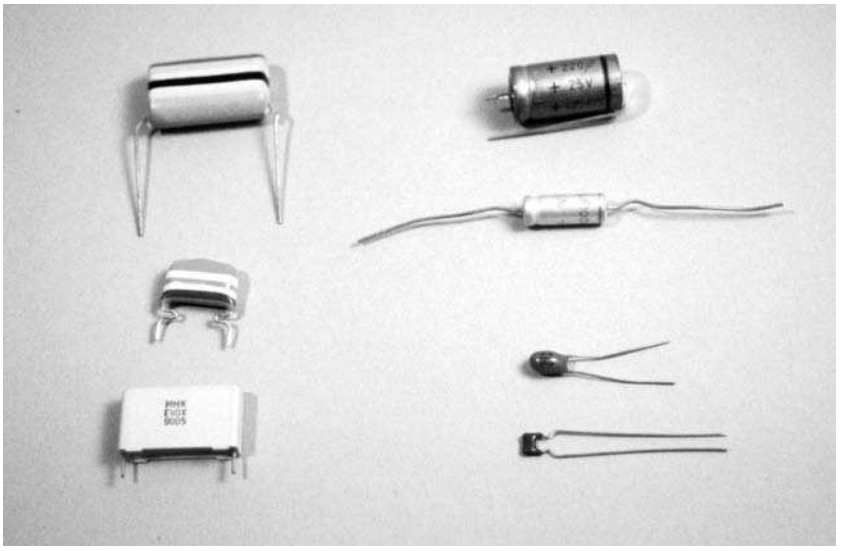


Fig. 1-3 Practical capacitors. Different types and values. The two upper right ones are electrolytic capacitors.

solution in which a coiled aluminium foil is suspended. Can and solution (electrolyte) together form the first electrode, the aluminium foil being the second one. The insulator is merely a thin coating of aluminium oxide (Al_2O_3), made by forcing a current through the finished assembly. Because the insulator can be made very thin, these so-called electrolytic capacitors can be made up to very large values: over 10 mF (10 000 μF).

Unfortunately, this ingenious form of capacitor has a number of disadvantages. First, the thickness of the oxide layer cannot be made very precise, so rated values are only approximate. Factory tolerances might be stated as "$-10+50\%$". This means that a capacitor of nominal value 100 μF might have a true capacitance somewhere between 90 and 150 μF. To make it worse yet, the value may change during use. Secondly, the thin insulators cannot withstand very high voltages, so most high-capacity "electrolytes" are intended for low-voltage circuits. A lucky circumstance is that the composition of the salt solution can be chosen to oxidize the aluminium and so to "repair" small holes automatically. The electrolyte (salt solution) serves further to warrant the contact between the metal electrode and the oxide layer, and so is part of one of the conductors. In view of the dielectric constant, water would make an ideal dielectric to make capacitors with, if only it would not conduct. A further peculiarity of electrolytic capacitors is that they are polarized: the oxide-carrying foil must be the anode, i.e. must always be kept positive with respect to the other one.

A special type of electrolytic capacitor is the "tantalum capacitor". Here, the anode is made of tantalum, and the dielectric is a thin layer of tantalum oxide (Ta_2O_5). The electrolyte is absorbed in a thin paper foil. These capacitors can be made relatively small, are more stable than conventional "electrolytes" and have slightly better tolerances. They are the preferred choice for most electronic devices, especially in the signal chain. Conventional electrolytes are used mainly in power supplies, as well as in the cheaper forms of consumer electronics.

Unwanted Properties, Impedance

The different types of components described above can all be obtained commercially, and are usually good-quality products. Nevertheless, it is important to note that resistors and capacitors do have unwanted properties, which are in part due to fundamental, hence unavoidable, laws. In short, resistors do have capacitance, and capacitors do have resistance.

Fundamentally, the two ends of a resistor can be considered as two conductors that are fairly close together (mostly less than a centimetre). Therefore, any resistor will have a few pF of capacitance. This is hardly a problem with low resistances (less than about 10 kΩ), but becomes very prominent at higher values, especially in the GΩ resistors used in intracellular and patch-clamp amplifiers. In addition to resistors, all parts of an electronic circuit have unwanted or "stray capacitance". Adjacent wires in a cable, or even adjacent lanes on a printed-circuit board, have capacitance with one another, and any wire has a stray capacitance with its environment, such as the instrument case, which is usually grounded.

Conversely, capacitors have at least two unwanted forms of resistance. The first one is the resistance of the dielectric, which should be infinitely high, but is often noticeable as a "leak". The best insulators, such as mica or glass (the latter one only at room temperature), have nearly ideal insulating properties, but capacitors with plastic insulators may show a noticeable "leakage resistance" that shows up in parallel with the capacitance. Although this is usually high, hundreds of megohms, its existence must be kept in mind, and may play a role in electrophysiological measurements, often where gigaohms are involved. Capacitors will also show a fundamental series resistance, since the conductors—metal coatings or foils and connecting wires—have a small but non-zero resistance. Usually, the series resistance may be neglected, but will become prominent with large foil-capacitors carrying relatively high currents. In addition to resistance, any coiled-up capacitor foil will exhibit self-inductance.

Practical self-inductances are made by winding long, thin copper wires into a coil, or solenoid, often around an iron or ferrite core. This may yield values from less than 1 μH, useful in radio and TV receivers to tune to a specific broadcast programme, up to tens of henries, used to reduce ripple in power supply circuits (see Chapter 2). Because of the copper wire, often long and thin to cram a large inductance value into a small volume, inductors have a relatively high series resistance, which can be neglected rarely, if at all. Current through a coil thus heats it, spending (or "dissipating") energy according to Joule's law. Thus, coils are complex components, and must in fact always be considered as a (perfect) self-inductance in series with a resistor. This means that the effective "resistance" of a coil is composed of a "true" resistance caused by the wire and the reactance due to the inductive processes. The notion describing this "total AC resistance" is called "impedance" (symbol Z, unit Ω). Thus, for a practical inductor, the impedance is the sum of the two described properties:

$$Z = R + XL \quad \text{or} \quad Z = R + \omega L$$

Since impedance is frequency-dependent, any statement about impedances must include an explicit or implicit statement about the frequency pertaining to it. Often, this leads to conventions, such as in the case of loudspeakers, where the nominal impedance, such as the familiar 8 Ω, is to be taken at a frequency of 1 kHz. In practice, a loudspeaker coil has a resistance of about 3 Ω, and a reactance of 5 Ω at 1 kHz. Note that at lower frequencies, the reactance will be lower and at higher frequencies higher than the stated value.

The impedance of other components, such as a leaky capacitor, can be derived by a similar argument. Here, capacitance and leakage resistance are parallel rather than in series. Parallel and series circuits will be dealt within the next chapter.

Note that perfect or "ideal" capacitances and inductors do not dissipate energy; only resistances dissipate energy. This is comparable to the mechanical equivalents of self-inductance, capacitance and resistance in mechanical systems: mass, spring stiffness and friction, respectively. A car that would have only mass (inevitable) and springs (added, although a certain elasticity is inherent in all materials) would bounce restlessly on an uneven road, since the bouncing energy would not be dissipated (in reality, a little energy is dissipated by air friction, but this is far from sufficient). Therefore, real cars have friction added, in the form of oil-filled dashpots, to dissipate most of the bouncing energy.

Inductors (coils, solenoids, see symbols in Fig. 1-7) are not used much explicitly in electrophysiology. An exception is the trick to wind a mains or signal cable onto a ferrite rod to reduce radio-frequency interference. To be effective, this makeshift coil must be situated close to the wall of the Faraday cage that harbours one's set-up. Practical forms of solenoids are shown in Fig. 1-4.

Much used in all sorts of technical set-ups is the notion of a "transformer", which is basically composed of two or more coils wound onto a single core (Fig. 1-5). The first coil, where an input AC voltage is applied, is called the "primary winding", or primary for short. The second and any further coils, where other voltages can be tapped, are all called "secondary winding". In the transformer, the mutual relations of electric currents and magnetic fields are exploited for the benefit and flexibility of electric circuits. That this device is able to transform voltages needs no argument, but it is nevertheless useful to discuss the virtues as well as the limitations.

The most basic characteristics of a transformer are the "turns ratio" and the "impedance ratio". Since the two solenoids share a common magnetic field, the turns ratio determines the ratio of input and output voltages. But a certain turns ratio can be reached in different ways: a primary of 10 turns, together with a secondary of 1 turn has the same ratio as the combination

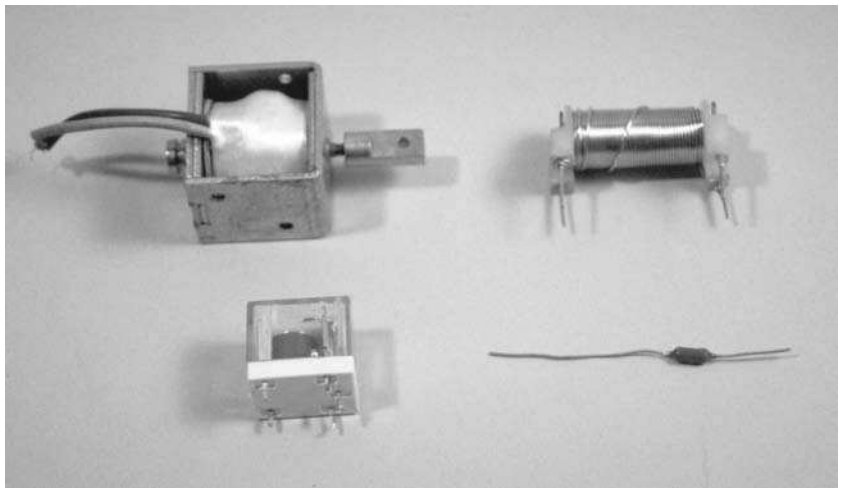


Fig. 1-4 Practical coils (solenoids, inductors).

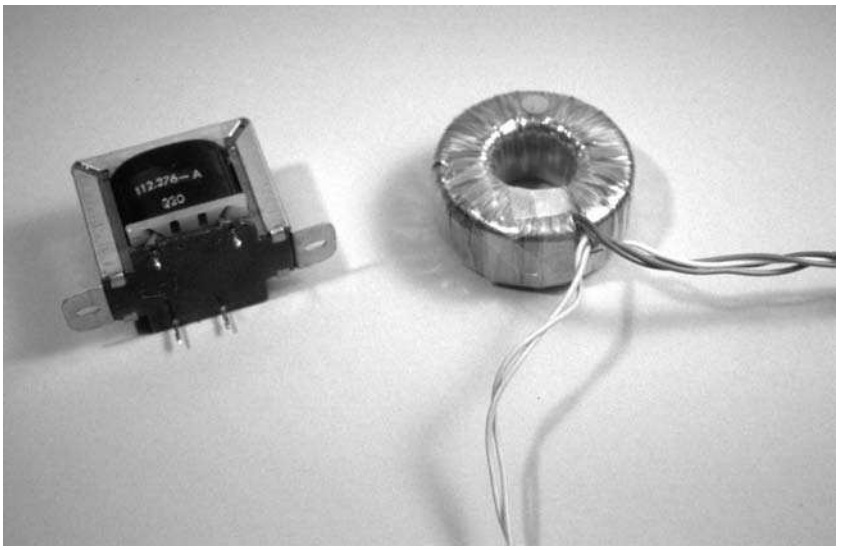


Fig. 1-5 Practical transformers. A conventional, E-core type at left and a more modern ring-core type at right. The thin wires are connected to the primary winding (high voltage, low current), the thick ones to the secondary (lower voltage, stronger current). Ring-core transformers are preferred because they produce less stray magnetic fields.

of 10 000 turns and 1000 turns. The difference lies in the impedances of the two windings. The impedance of a transformer winding is related to the square of the number of turns, and so the *impedance ratio is the square of the turns ratio*. Usually, the impedance of the primary must be matched to the intended input voltage, and the output impedance must be matched to the expected load.

In addition, the dimensions and the construction of a transformer determine how much energy, or rather power, may be converted without overheating it. Therefore, most transformers are specified to these three quantities, e.g. "230 V to 12 V, 25 W". Note that the latter designation states the amount of power that can be transferred, not the power dissipated by the transformer. Since power is voltage times current, the rating mentioned above is occasionally stated as "25 VA". In addition to passing useful power on, a transformer does dissipate a bit of energy itself. This is caused partly by resistance of the copper wires, partly by currents generated in the core. This may amount to about 20% of the power transmitted. As an example, a transformer that delivers 24 W to a load (12 V × 2 A) might spend about 29 W from the source (230 V × 0.125 A).

It will be obvious that the mutual influence of electric current and magnetic field is limited to alternating current, i.e. any DC applied to a transformer is not transmitted. Note that it does, however, cause dissipation, i.e. heating the device up. A second factor limiting the amount of DC that may be forced through a transformer is that a direct current magnetizes the core. This may cause the core to be saturated, preventing proper functioning. Transformers are used in most situations where a high voltage must be converted into a lower voltage, but may also be used in the reverse situation: pushing a low voltage up (a step-up transformer). Note, however, that a transformer does not add energy to a signal, and so cannot be used instead of an amplifier.

In addition to the power loss, transformers have other bad habits. Even with AC, the impedances of the windings are frequency-dependent, and so are the magnetic properties of the core. Therefore, transformers can be used only in a limited frequency range. Even then, a transformer will almost always distort a signal. This is why transformers are hardly ever used to transform signals. They are indispensable, however, in power supply circuits, transforming the mains voltage into either higher voltages, needed to drive oscilloscope and TV picture tubes, or lower voltages, needed for most transistor circuits.

Apart from transforming voltages, transformers are also useful to insulate two current circuits electrically from each other. A well-known example is the "shaver outlet" found in bathrooms. This is a 1:1 transformer (230 V in, 230 V out) that may save lives because it insulates the circuit connected to the shaver from the mains voltage, one of the conductors of which is grounded. The shaver circuit is said to be "floating", since none of the terminals is grounded. The effect is that unintentional contact with a "live" wire does not cause a current to flow through the human body to the ground. By the same token, specially designed transformers are used to insulate medical instruments from the patient circuit (ECG electrodes, etc.).

In Chapter 2, we will discuss the virtues of and methods for letting an electrophysiological recording circuit "float" (differential recording), which is used to reduce interference.

Cables

Special attention must be focussed on the properties of connection wires, usually called "cables". The form of the wires used to connect lamps, vacuum cleaners and other household appliances is hardly important. Obviously, the copper conductors must be thick enough to carry the necessary currents, and the insulation must be thick enough to prevent shock hazard. Often, a third wire is added to provide for a ground connection to the metal parts of the appliance. The same demands hold for cables used to power scientific instruments, but in addition, the handling and measuring of weak signals puts new and special demands on the connections used. Anyone who tried to wire up a record player with the type of wire used to install, say, an electric door chime will no doubt have found that a penetrating hum pervades the signal, making it totally useless. Therefore, all signal links must employ "shielded cables". The simplest form, used occasionally, is to put one conducting "shield", usually a copper braid, around both signal wires. Although this resolves the hum problem by keeping out the electric fields emanating from the mains cables (discussed before), the stray capacitance between the two signal wires remains. If a number of wires, carrying different signals, are bundled in this way, the stray capacitances cause one signal to "leak" to other wires in the bundle. It is not prevented by a shield around the whole bundle.

This signal leakage, called "cross-talk", can be prevented by covering each wire with a separate shield. Since a cylindrical shield around a single wire is concentric, or coaxial with it, these cables are called "coaxial cables" (coax for short). By using only coaxial cables to transport weak electric signals, interference by electric fields, both from the mains ("hum") and from the broadcast radio waves, as well as cross-talk, are prevented to a large extent. However, the stray capacitance between the wire and its shield remains, and this will prove to be a nuisance in several electrophysiological recording situations. The commercially available coax cables have a capacitance of about 100 pF/m of length. In many situations, this is no problem, but the stray capacitance gets dominant in two situations: high-frequency (e.g. in radio engineering) and high-impedance (e.g. in many electrophysiological recording chains).

In fact, cable capacitance, together with the resistance of a microelectrode, forms a filter that will block higher-frequency components when not well designed. These filters will be dealt with later on in this part of the book. The most obvious solution is to keep cables short in critical parts of the set-up.

A second solution is to use a preamplifier that does not need an input cable at all, because it is situated directly behind the electrode holder. This construction is called a "probe amplifier". A third solution consists of the use of special low-capacitance cables. This form of coax uses plastic foam instead of a massive insulation between wire and shield. Because foam is half plastic, half air, the relative dielectric constant of foam is between that of plastics (about 2–3) and that of air (1). These cables may have capacitances of about 30 pF/m. Often, however, such as in intracellular and patch-clamp recording, the use of the "probe" type configuration is mandatory.

Users of electronic components must always be aware of the unwanted properties of a component, know the approximate magnitude (colloquially called the "order of magnitude"), and so estimate the effects before application of that component in a specific research set-up.

CIRCUITS, SCHEMATICS, KIRCHOFF'S LAWS

Electrical components are joined into "electric circuits". The most primitive way to do so is simply to tie the connecting wires firmly together. Because the surfaces of wires tend to oxidize and so lose contact, other methods were invented to make connections that last permanently. Soldering, i.e. the connection of (copper) wires with an easily melting lead/tin alloy, is the method used most often. Mass production leads to the development of the "printed-circuit board" (PCB), where all wires necessary for a particular circuit are manufactured in one stroke, being etched in the form of lanes on a metallized, glass-fibre-reinforced plastic sheet (Fig. 1-6).

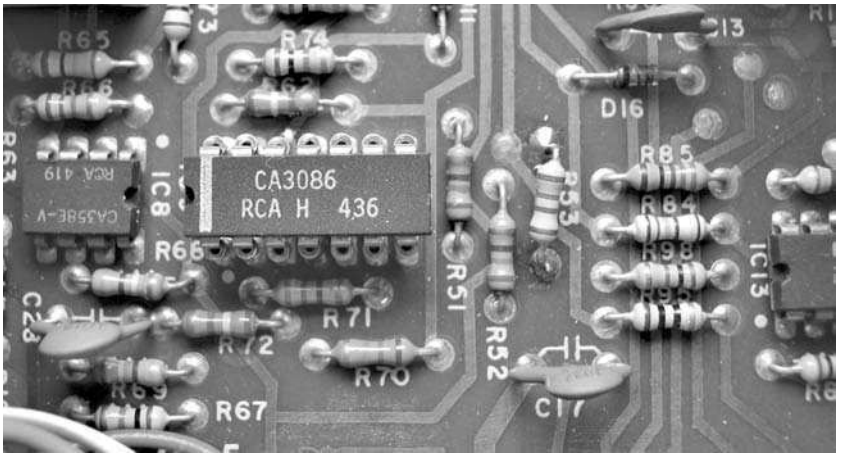


Fig. 1-6 Part of a printed-circuit board, showing several components in addition to the printed copper lanes.

The course of the connections between components on a PCB is often very long and sinuous, running around numerous other components. Therefore, it is almost impossible to analyse the function of a circuit by looking at the PCB. To explain the action of an electronic circuit, a schematic, graphical representation called a "circuit diagram" is drawn, in which specific symbols that indicate the different components are connected by clear, straight lines.

Throughout the scientific and engineering practice, different forms of symbols are used, all more or less abstract representations of the form or function of the components they represent, and laid down in, unfortunately, a number of industrial standards in numerous countries. The symbols used in this book do not conform fully to one of these standards (although they are close to the main European use), but are chosen mainly for clarity. The symbols used for the components dealt with in this chapter are depicted in Fig. 1-7.

In addition, we will need symbols for switches, voltmeters, outlets and plugs of both "genders", electrodes used in electrophysiology, and for such things as earth and ground connections. A number of accessory symbols are shown in Fig. 1-8. Note the distinction between case ground (case, frame or all connected metal parts of an instrument or set-up, which may have contact with the ground, but may also float) and ground (the same as before, but definitely tied to ground). Switches exist in a variety of forms, which differ in the number of poles and the number of positions, hence SPST (single pole, single throw) for a single-pole on/off switch and so on.

Note also that the (case) ground symbols suggest an "open end", where in reality all ground connections form a closed circuit. Finally, note the distinction between a wire crossing and

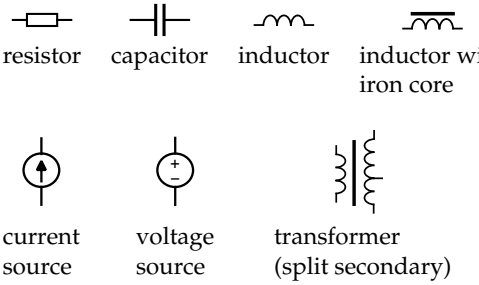


Fig. 1-7 Symbols for circuit diagrams.

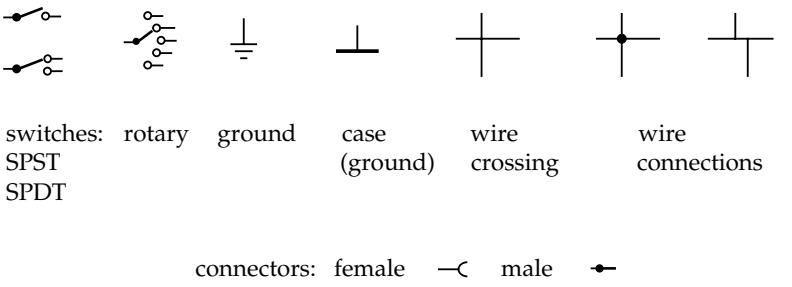


Fig. 1-8 Accessory symbols.

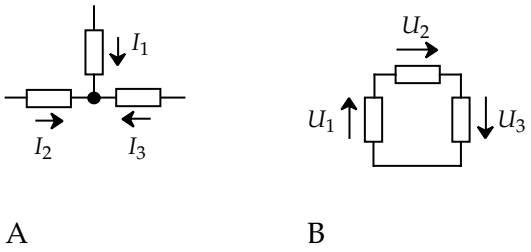


Fig. 1-9 Node (A) and Loop (B).

a wire connection. The latter form reduces the risk of ambiguity. Further symbols will be introduced where necessary.

An overview of schematic symbols is given in Appendix B.

Most symbols, such as resistors and capacitors, can be used in both the horizontal as well as the vertical position. For some components, a specific orientation is recommended. A second convention is to let the main signal path point in the reading direction, i.e. from left to right. Thus, the primary winding of a transformer is drawn preferably as the one left of the core. A further useful standard is to draw the power supply line at the top of the diagram, the reference (ground) line at the bottom. In the case of dual power voltages, the positive one is at the top, the negative one is usually drawn below the ground (reference) level. Note that most symbols represent a single component, such as a resistor or capacitor, but some are "abbreviations" for more complex entities, such as the notions of a current source or voltage source. Symbols representing entire instruments, used frequently in so-called block diagrams, will be dealt with in Chapter 2.

Before discussing the properties of circuits, we need two fundamental laws pertaining to the behaviour of electric quantities in circuits. Joining components together creates "nodes", where several components meet, and "loops", through which electrical currents may flow. The two notions are illustrated in Fig. 1-9.

The two important laws are formulated first by Kirchoff (around 1850 AD). The first one, Kirchoff's node law, states that the sum of all currents entering a node must be zero. Thus, in the example of Fig. 1-9A, $I_1 + I_2 + I_3 = 0$. Taking the directions of the currents into account, if $I_1 = 1\,\text{A}$ and $I_2 = 3\,\text{A}$ (towards the node), I_3 must be $-4\,\text{A}$ or $4\,\text{A}$ away from the node.

The second law is the Loop Law, which states that the sum of all voltages in a loop must be zero. Thus, in Fig. 1-4B, if U_1 is $1.5\,\text{V}$, U_2 and U_3 must add up to $-1.5\,\text{V}$. Kirchoff's laws follow directly from fundamental laws like the conservation of energy. Stated simply, neither voltages nor currents can get "lost" in a circuit. Kirchoff's laws form a simple and powerful tool to analyse the processes in any electric circuit quantitatively.

COMPOSITION OF SIMILAR COMPONENTS: ATTENUATORS

The most simple and basic circuits are compositions of components of the same type, such as resistors, connected in parallel, in series, or in a combination thereof. In any case, a composition of two or more resistors connecting two points (nodes) can be replaced by a single resistor, the value of which can be computed using Ohm's and Kirchoff's laws. Examples are given below.

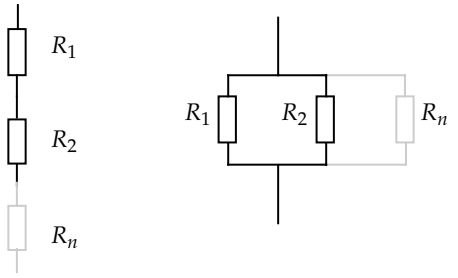


Fig. 1-10 Resistances in series and in parallel.

The simplest form is the connection of two resistors in series. The total resistance between the two end points is simply the sum of the two component resistances. This principle holds for any number of resistances in series: see Fig. 1-10.

$$R_{\text{tot}} = R_1 + R_2 + \cdots + R_n \quad \text{or} \quad R_{\text{tot}} = \sum_i R_i$$

When connecting resistors in parallel, the total resistance between the two points is lower than any of the component resistors, and is often formulated as follows.

$$1/R_{\text{tot}} = 1/R_1 + 1/R_2 + \cdots + 1/R_n$$

In fact, here it is easier to use a quantity called "conductance" (symbol g), which is simply the inverse of resistance.

$$g = 1/R$$

The notion of a conductance is used very often, especially in electrophysiology, such as in dealing with the amount of current through a membrane, or passed by an ion channel.

The unit of conductance is siemens (S). In the USA, this unit, in fact the "inverse ohm", or Ω^{-1}, is often called the "mho" ("ohm" spelled backwards). So, a resistance of 1 MΩ is the same as a conductance of 1 μS, etc. Thus, the formula for parallel resistances reduces to the simple addition of the component conductances:

$$g_{\text{tot}} = g_1 + g_2 + \cdots + g_n \quad \text{or} \quad g_{\text{tot}} = \sum_i g_i$$

Similar arguments lead to the effect of combinations of other parts, such as capacitances. It is easy to see that connecting capacitors in parallel increases the total capacitance, just like a number of wine vats on the same floor increase the storage capacity at the same level of the liquid.

$$C_{\text{tot}} = \sum C_i$$

To the contrary, a chain of capacitors in series has a lower capacitance. This can be understood by noting that it amounts to increasing the thickness of the dielectric. The formula is analogous to the case of the parallel resistances:

$$1/C_{\text{tot}} = \sum_i 1/C_i$$

The inverse of capacitance has been used formerly, but has never become popular. Guess how Americans used to call this unit … yes, the daraf; no kidding!

With self-inductances, the case is similar to the connections of resistances: in series circuits, the self-inductances sum, in parallel circuits, the inverses of the self-inductances sum.

So far we dealt with multiple components connecting two nodes, but as we argued, these circuits may be reduced to a single component. Parallel and series circuits are used occasionally, for instance to compose values that are not on the market. An other application is to spread the power dissipated over more than one component: in this way a 1 Ω resistor that can dissipate 10 W safely may be composed by connecting 10 resistors of 10 Ω, each rating 1 W, in parallel.

More often, however, resistors are combined to create different voltages, for example to attenuate a signal that is too strong. A well-known example is the so-called oscilloscope probe, intended to extend the range of input voltages by using resistors to divide the input voltage by a factor of, say, ten. Two resistors in series, used as a "voltage divider" or "attenuator" circuit, is shown in Fig. 1-11.

Here, the circuit has three nodes: ground, input and output. The output voltage as a function of the input voltage can be derived with the following formula.

$$U_{\text{out}} = U_{\text{in}} \times \frac{R_2}{R_1 + R_2}$$

The factor $R_2/(R_1 + R_2)$ is called the attenuation factor.

For alternating current, two capacitors or two self-inductances connected in the same way can be used as attenuators. Often, a voltage needs to be attenuated in a variable way. The most familiar example is the volume control on a radio, TV or audio set. Thus, we need two resistors with a variable attenuation factor. This is done by providing a (carbon, wire-wound or metal film) resistor with a moveable tap. Such a component is called a "potentiometer" and is shown in Fig. 1-12. Since it is purely resistive, a potentiometer may be used to attenuate

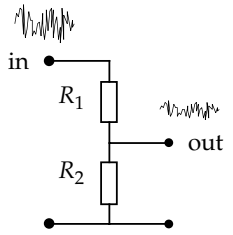


Fig. 1-11 Voltage divider circuit.

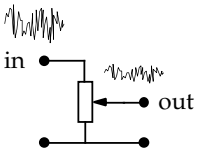


Fig. 1-12 Potentiometer as a volume control.

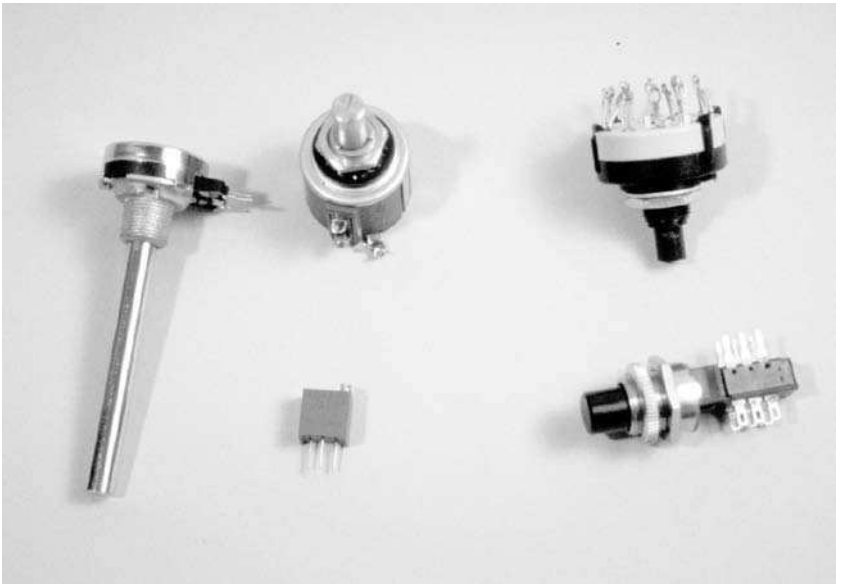


Fig. 1-13 Practical potentiometers (left and centre) and switches.

both AC and DC voltages. Note, however, that stray capacitances may lurk around the corner. Practical forms of potmeters and switches are shown in Fig. 1-13.

Practical Voltage Sources and Current Sources

Apart from the useful attenuators described above, voltage division often occurs unintentionally. This is shown best when we consider practical voltage and current sources. The ideal of a voltage source, maintaining a constant voltage across a load, is approached to a fair extent by the well-known batteries and accumulators. However, by short-circuiting a battery, a reduced voltage is the effect, rather than an infinite current.

In fact, the stated (nominal) voltage of 1.5 V of the familiar "dry cell" exists only when no current is drawn from it. The more current is drawn, the more the voltage decreases. In fact, this can be explained as a resistance in the voltage source.

Thus, a real voltage source can be considered as an ideal voltage source in series with a (small but noticeable) "source resistance" (R_{src}). This is also called "internal resistance" or "output resistance", since it is manifest at the output of the voltage source. Thus, a practical voltage source can be represented in the way shown in Fig. 1-14A. The source resistance is not an added component, but rather an unavoidable property of any voltage source. The "load resistance" (R_{load}) is the representation of the lamp, motor, instrument or whatever is connected to this voltage source. It is easy to see that source resistance and load resistance together form a voltage divider, each getting their share of the total voltage, also called the "electromotive force".

The symbol used for an electromotive force is E, to distinguish it from the "practical", "actual" or useable voltage U. Apparently, an ideal voltage source has a zero source resistance, and the lower the source resistance, the better a voltage source behaves. As a rule of thumb,

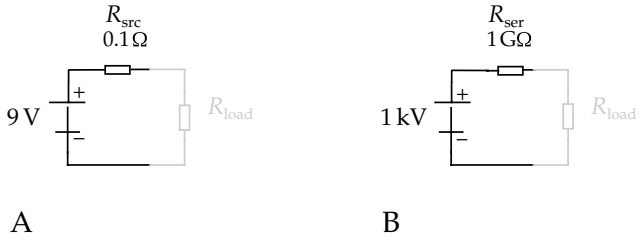


Fig. 1-14 Practical voltage (A) and current (B) sources.

the voltage can be considered to be constant as long as the load resistance is high relative to the internal resistance. How much higher will depend on the precision wanted. Usually, a ratio of 1:100 is acceptable, since this will introduce an error of only 1%. Although batteries and accumulators are not bad, the design of any instrument must take the inherent loss of voltage into account. Readers who wondered why a 3 V flash light needs a 2.2 V bulb have found the answer.

Practical current sources are encountered rarely in daily life. A possible exception is the solar cell, which at constant illumination yields a constant current, proportional to the amount of light falling on it (at an approximately constant voltage of 0.5 V). Contrary to a voltage source, an ideal current source has *infinite source resistance*. Therefore, a practical current source can be built using a high-voltage source in conjunction with a high series resistance, such as shown in Fig. 1-14B. In this example, the 1 kV source combined with the 1 GΩ resistor delivers approximately 1 μA to the load, irrespective of the load resistance, as long as the load resistance is small with respect to the applied source resistance. How much lower will again depend on the wanted precision. Note that in this case, the low source resistance of the 1 kV power supply is increased artificially by the 1 GΩ resistor. Resistors of this order of magnitude are used among others in microelectrode preamplifiers to provide a test current for measuring the resistance of the glass capillary. Since capillary resistance in an experiment may exceed 100 MΩ, the errors are often not negligible, and must be corrected by means of a table or calculation.

VOLTAGE AND CURRENT MEASUREMENT

The reliable measurement of voltage and current faces the same problems related to internal resistances. The classical way to measure a voltage is by a galvanometer, or moving-coil meter, in which a thin-wired coil fitted with a pointer is suspended in a strong magnetic field. A current through this coil causes it to rotate around its pivot, so that the pointer moves over a graduated scale. By design, the deflection may be made to be very nearly linear with the current strength. Moving-coil meters may be made fairly sensitive, such as 100 μA for the maximum, or full-scale deflection (fsd). Since the coils have resistances of about 1 kΩ, full-scale deflection is reached at a voltage of about 0.1 V. Today, the cheapest "voltmeter" for use in domestic electrical circuitry is still based on this moving coil, or microammeter (see Fig. 1-15 left).

Not surprisingly, however, this basic device is superseded by electronical ones, called "DVM"s (digital voltmeters, Fig. 1-15 right). The voltage to be measured is amplified electronically, and usually converted into digital form, presented on an LCD screen.

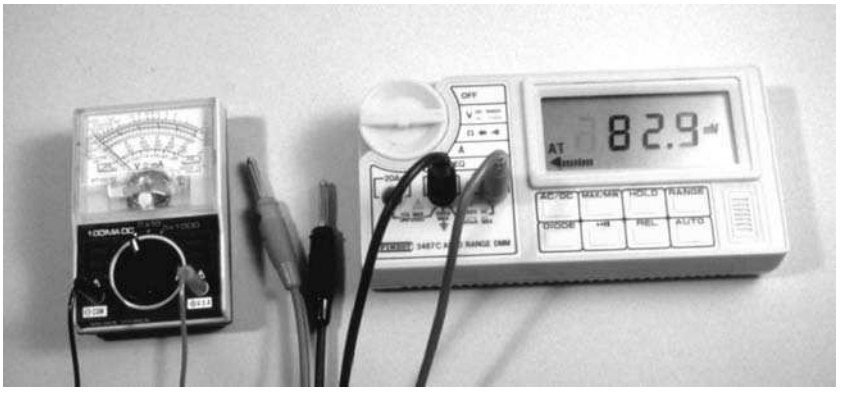


Fig. 1-15 Moving-coil meter (left) and digital voltmeter (DVM, right).

Unfortunately, the mode of operation and the limitations of such devices are hard to see through. Therefore, we will digress briefly on voltage and current measurement with a moving-coil meter. DVMs may have better performances, but the principles to observe remain the same.

The functioning of a moving-coil voltmeter is not hard to understand. A voltage across its terminals causes a current to flow through the wire. The current causes a magnetic field to be developed, which in turn moves the pointer. This means that a voltmeter *draws a bit of current out of the source* that it intends to measure the voltage of. The current drawn will reduce the voltage more or less, depending on the ratio of the internal resistance of the voltage source and the resistance of the meter coil. This is illustrated in Fig. 1-16. The circuit is shown at the top left. However, we are more interested in the electrical properties of the components. Any real-life voltage source has an internal resistance, and so has the voltmeter. Therefore, we need to derive the so-called equivalent circuit. This is drawn at the right. Here, R_{int} is the internal resistance of the voltage source, and

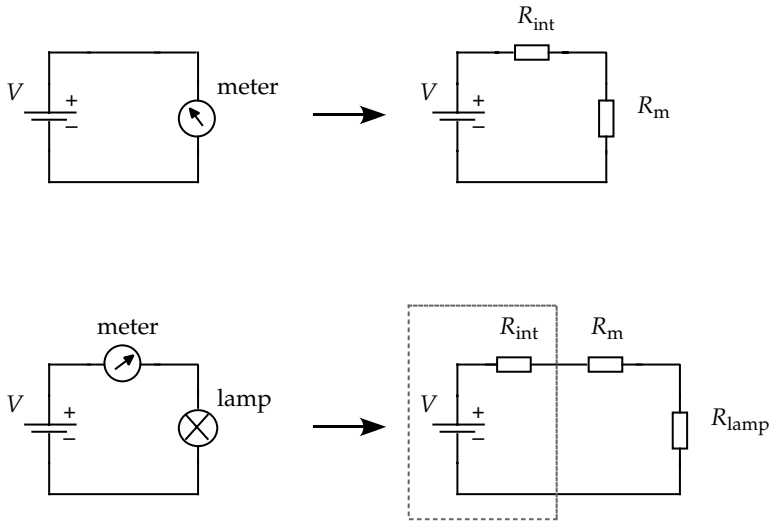


Fig. 1-16 Errors in the measurement of voltage (top) and current (bottom). The real circuits are at left, the equivalent circuits at right. The dashed rectangle shows which properties belong to the battery.

R_m the resistance of the meter coil. As a practical example, let us take a small battery such as the ones used in hearing aids and watches. The voltage will be about 1.5 V, the internal resistance about 1 Ω. Suppose the voltmeter has a resistance of 1 kΩ. Then, the meter draws about 1.5 mA from the battery (1.5 V over 1001 Ω). In this case, the voltage drop across the battery's internal resistance is 1 mV. Thus, the voltage is underestimated by 1 mV. In all but the most demanding cases, this can be neglected. Nevertheless, a voltage is principally underestimated by the influence of the measuring device. The higher the meter's resistances, the better the approximation.

For current measurement, the meter must be inserted into the electrical circuit, in series with the other component. In a flashlight, for instance, to measure the lamp current, we must break the circuit somewhere and insert the current meter.

This is shown in Fig. 1-16, bottom row. Here, the demands are the other way round: the current that flows through the circuit must not be hindered by the current meter, and so a very low meter resistance is needed. Suppose, for example, that the lamp in our example behaves as a resistance of 3 Ω. When connected to a 1.5 V battery, the lamp current would be expected to be 0.5 A. However, even without a meter in the circuit, the lamp current also flows through the battery's internal resistance. If we take this again at 1 Ω, the total resistance in the circuit amounts to 4 Ω. Therefore, the real lamp current would be 0.375 rather than 0.5 A. If our current meter would also have a 1 Ω resistance, the current would be reduced further to 0.3 A (1.5 V over $3+1+1=5\,\Omega$). Again, the measurement is an underestimation of the true value. A lower meter resistance improves the approximation.

In the examples above, the errors involved are not too serious: the battery voltage is estimated slightly too low, and the current through the lamp is reduced a bit. For measurements in electronic and electrophysiological circuits, however, the errors would be far too high, and might amount to almost 100%. Therefore, electronic voltmeters having a far higher input impedance must be used.

For any voltage measurement, the input impedance needed can be computed from the values of source impedance (i.e. the circuit under test) and meter impedance, which together form a voltage divider circuit. To get a fair precision, the impedance of the meter should be about 50 to 100 times as high as the circuit's impedance. In current measurement, the meter resistance must be lower than the other resistances in the circuit, by approximately the same factor.

When monitoring the mains current drawn by a lamp or a similar appliance, losing a few hundred millivolt from the 230 V is no problem, but in low-voltage circuits, a similar voltage drop may be significant. Finally, electrophysiological signals are obviously too small to be measured in this way, being mostly less than 100 mV themselves. Therefore, we need electronic voltmeters and preamplifiers, treated in Chapter 2. Next to the specialist equipment, however, a simple DVM that can measure voltage, current and resistance (as most types do) comes in handy, e.g. for the checking of resistances, testing of cables, batteries, and so on.

COMPOSITION OF UNEQUAL COMPONENTS: FILTERS

The next circuit to be treated is composed of a resistor (R) and a capacitor (C). This important circuit behaves as a filter, and is therefore called an "RC filter". Other combinations of the basic components are the RL filter and the LC filter. The latter is very popular, although

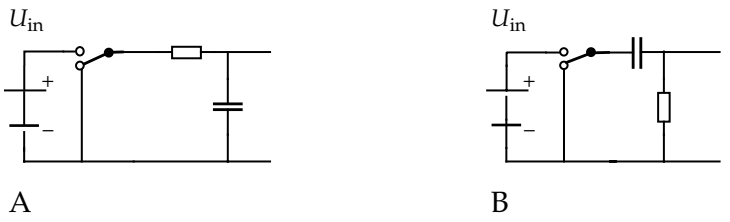


Fig. 1-17 The two forms of RC filter: low-pass (A), high-pass (B).

few people will realize that their houses are full of LC filters. LC filters are the universal, resonant filters that can single out one broadcast frequency from the multitude of signals that fills the ether. Therefore, tuning a radio or TV set is done with LC filters. RC filters are employed most for electrical measurements and for the manipulation of electrical signals in general.

Since a filter has an input and an output, the RC filter exists in two configurations, shown in Fig. 1-17.

The two forms are called low-pass and high-pass filter, names that will become apparent through the analysis of what happens when we feed signals into them.

Let us analyse the low-pass configuration first. We have seen earlier that a capacitor can be charged by feeding a current into it. This we will do by adding a voltage source (of, say, 1 V) and a switch (Fig. 1-17). To be sure that we start with an uncharged capacitor, it is shorted until the start of the experiment. For the sake of simplicity, we take a capacitor of 1 F and a resistor of 1 Ω. At the moment we have flicked the switch, the voltage appears at the input, and so across the resistor (since the capacitor is uncharged, the output voltage is still zero). The charging current can be computed easily: $I = U/R$, that is, $1/1 = 1$ A. From the properties of electric current, and the units chosen, this would charge the capacitor in one second, that is, if the current would remain constant. But it is easy to see that the current decreases continuously. After a certain time, the capacitor is charged to 0.5 V. At that time, the voltage across the resistor is reduced to 0.5 V, which in turn reduces the charging current by the same factor. In other words, charging gets progressively slower, and the input voltage is approached asymptotically, in a way called an "exponential function". This is shown in Fig. 1-18.

The mathematical description of this process is fairly simple, and follows from the above discussion. We saw that the charging rate, that is the change of the output voltage ($\mathrm{d}V/\mathrm{d}t$), is proportional to the voltage across the resistor, and this is equal to the difference of output and input voltage ($U_{out} - U_{in}$):

$$\mathrm{d}V/\mathrm{d}t = K(U_{out} - U_{in})$$

where K is a constant that determines the time scale of the charging process. The solution of this simple differential equation is the exponential function given below:

$$U_{out} = U_{in}(1 - \mathrm{e}^{-Kt})$$

It will be clear that the constant K must depend on both R and C. With a higher resistance, charging would be slower because less charge is transferred per time unit, whereas with a higher capacitance, the charging would also be slower, this time because more charge is needed

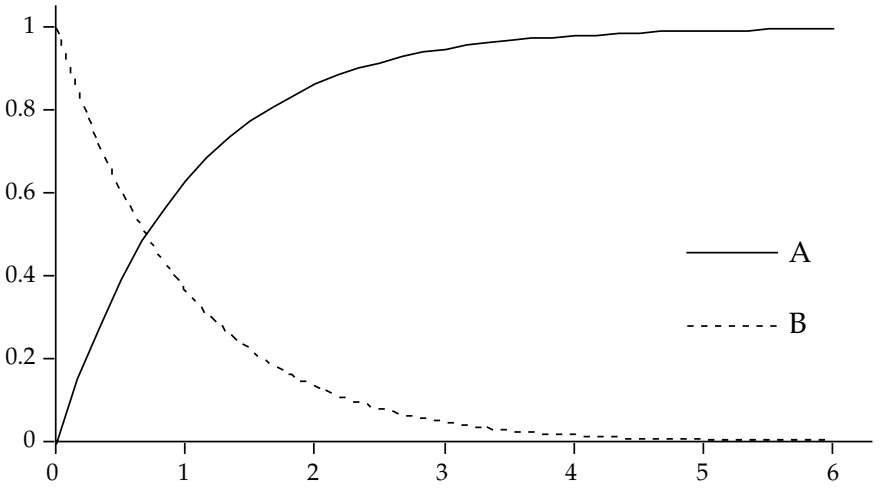


Fig. 1-18 Step responses of RC filters.

to develop a certain voltage. The quantity determining the time scale is the product RC, usually called the "time constant" (symbol τ). The process is analogous to the charging of a water butt through a (thin) tube. Thus:

$$K = 1/RC \quad \text{or} \quad K = 1/\tau, \quad \text{hence} \quad U_{\text{out}}/U_{\text{in}} = 1 - \mathrm{e}^{-t/\tau}$$

By dividing through the input voltage, a normalized, or dimensionless, quantity is obtained that runs from zero to one in an asymptotic way. This is shown in Fig. 1-18, curve A.

Mathematically, this is the same process as radioactive decay: when half the original amount is left, decay is also half as fast. Therefore, the characteristic used to describe the speed of the decay is the half-life, or the time span in which half of the substance falls apart. With capacitor charging, we could also compute the half-voltage time, but in practice, the base (e) of the exponential function is used (dotted line in Fig. 1-18). Thus, the time constant (τ) represents the time in which the output voltage is changed to $(1 - 1/\mathrm{e})$ times the voltage step at the input.

The exponential curve shown is called the "step response" of the RC filter. Other filters will in general show other step responses, so that a number of filter circuits can be characterized by their step responses. If the switch is flicked again, the capacitor will be discharged. Since the current flows through the same resistor, the discharge curve is similar to the charging curve, except that it is inverted, going from +1 V to 0 V (Fig. 1-18, curve B).

What about the response to sinusoidal signals? Most often, a filter is fed with a complex signal, which may consist of a sine wave of a certain frequency or a combination of sinusoids with different frequencies. It is not difficult to understand what happens when a sine signal is fed into the low-pass filter described above. With the step response, we have seen that, because of the time it takes to charge the capacitor, the output voltage lags behind the input voltage. The same happens when we feed a sine signal into a low-pass filter: the output signal lags behind, in other words the phase of the sine wave is changed. In addition, the sluggishness of the response causes the amplitude to be lower than the amplitude of the input signal.

This amplitude decrease, or attenuation, depends on the frequency of the signal, and a frequency-dependent attenuation is just the definition of a filter. At very low frequencies, the capacitance is charged so fast that it has hardly any effects on the output signal: the output signal is virtually identical to the input signal. At progressively higher frequencies, however, the effects of the capacitor become more prominent, causing both a phase delay and a lower output amplitude. At very high frequencies, hardly any output signal is left; these frequencies are said to be "filtered out". The degree of filtering, plotted as a function of the frequency (more precisely the logarithm of frequency), is shown in Fig. 1-19. This is called the "frequency response", or "frequency characteristics". A is called the "amplitude characteristic", φ (Greek lower case "phi") the "phase characteristic".

The figure shows that the transition from passing the signal unaltered to filtering is a gradual one, sloping down with increasing frequency. The phase delay runs from 0 to 90° ($\pi/2$ radians), also in a very smooth way with increasing frequency. Despite this wide area in which the filter affects the input signal, we can choose an arbitrary but practical point as the boundary, called "corner frequency", also called "roll-off" or "cut-off frequency". The frequency that is best taken as a standard, shown as a dotted line, is the point at which the signal is attenuated by 3 dB (stated otherwise, the gain is −3 dB). The phase is delayed by 45° ($\pi/4$).

Since all RC low-pass filters show the same *form* of filtering behaviour, the cut-off frequency is a complete and sufficient description to characterize it.

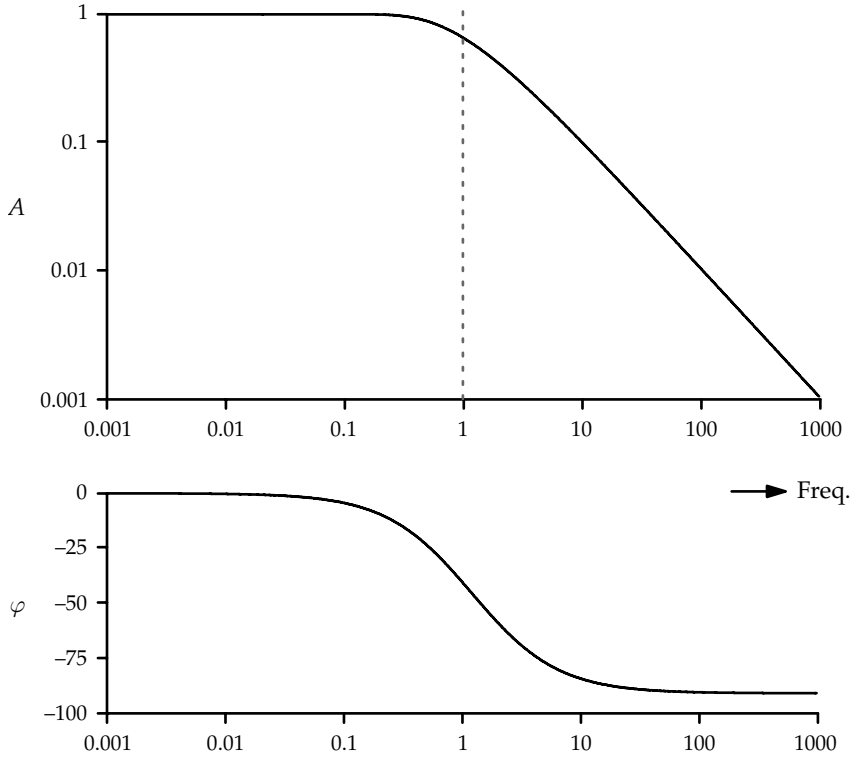


Fig. 1-19 Frequency characteristics of low-pass filter.

The amplitude is expressed in the well-known logarithmic measure called the "decibel". The decibel is a relative measure, and originated in telephone technology, where power, rather than voltage, is the interesting quantity. Hence, the unit of sound power, the "bel" (B, after Alexander G. Bell, inventor of the telephone), was defined as a factor of ten (one order of magnitude) relative to some convenient standard. Because most measurements span a few bels, and because about 0.1 bel is a just perceptible change, the decibel (dB) became most frequently used. The decibel is often used to describe the gain (G) of amplifiers. In that case, the reference is simply the input voltage. For filters and attenuators, where the output voltage can only be equal to or less than the input voltage, the gain values are zero or negative, respectively.

For the description of filter behaviour, we are discussing voltage rather than power ratios. Therefore, the formula for the conversion of the "gain" of a circuit into the decibel form is:

$$G = 20\log(U_{\text{out}}/U_{\text{in}})\,\text{dB}$$

Here, U_{out} and U_{in} are the voltages to be converted, and the factor of 20 arises because (i) a bel is 10 dB, and (ii) power is the square of voltage, contributing a square under the log sign or the equivalent, a factor of two in front of it.

For our low-pass filter, the gain at very low frequencies is virtually 0 dB, and becomes more and more negative with increasing frequency. At the cut-off frequency, the gain proves to be -3 dB; hence the cut-off is also called the -3 dB point. Check that here the output voltage is $1/\sqrt{2}$ times the input voltage, or about 70.7%. The ratio of output to input powers is 1/2, or 50%. Thus, as a simple rule of thumb, the roll-off frequency is where both power transmission and phase shift are "one half".

Having described the low-pass filters by their time behaviour as well as their frequency characteristic, it is easy to extend the discussion to the high-pass version: the characteristics of the high-pass filter are the complement of the functions described above. Thus, the step responses of the high-pass filter can be taken from Fig. 1-18 again, but in this case curve B depicts the charging behaviour (starting at the moment the input voltage is switched on), whereas curve A is the discharging curve (from switching off). Thus, contrary to the response of the low-pass filter, the voltage step at the input is transmitted unattenuated, after which the voltage returns to zero. The voltage step is, as it were, "forgotten". This is also called "adaptation", and is seen very frequently in responses of sense organs and other neurophysiological structures. Note that after the input step back to zero, a negative peak develops. This is also seen in neural signals, for instance in the response of photoreceptors, where the spontaneous activity is lowered or even suppressed immediately after switching a light source off.

The frequency characteristics of the high-pass circuit are also complementary: the highest frequencies are passed unattenuated, whereas amplitude reduction and phase shift occur at increasingly lower frequencies (Fig. 1-20). The cut-off frequency, here taken at unity again, is again the -3 dB point.

The phase shift is also in the opposite direction, running from 0° (at infinite frequency) to 90° ($\pi/2$) lead, rather than lag, with 45° as the cut-off frequency.

Thus, at very low frequencies, the output leads 90°, that is, the peak of the output signal occurs earlier than the peak of the input signal. At a first glance, having an output signal before the input may seem a violation of causality. This is not so, however, because we are describing a steady-state situation, where the input signal already existed for a number of cycles. If we examine the onset of a sine signal, we will see that causality is preserved, and that the phase lead builds up within one or two cycles. This is illustrated in Fig. 1-21 for both the low-pass and high-pass filters.

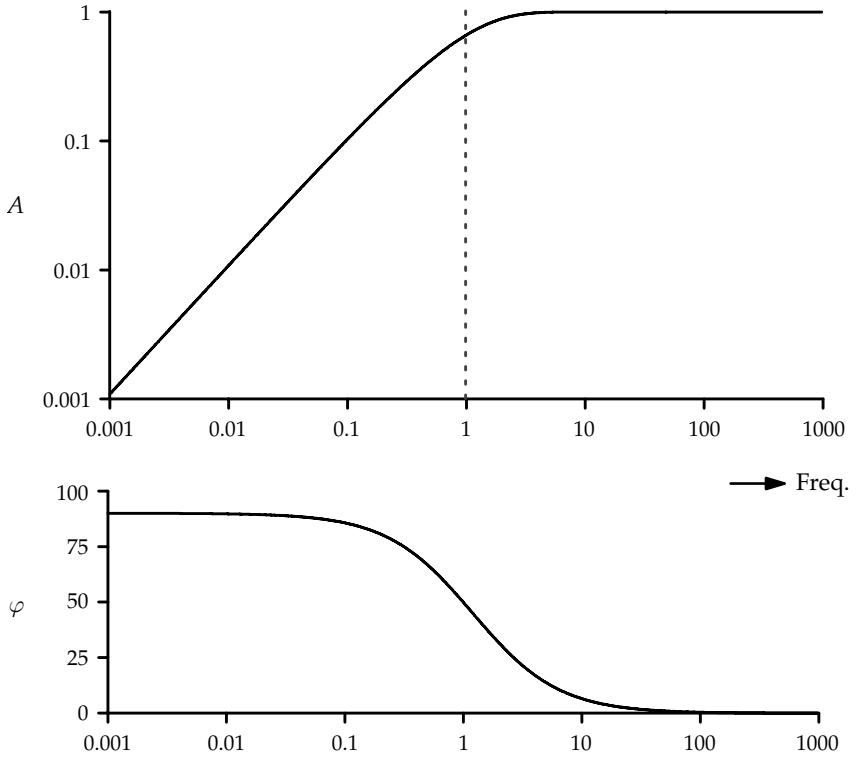


Fig. 1-20 Frequency characteristics of high-pass filter.

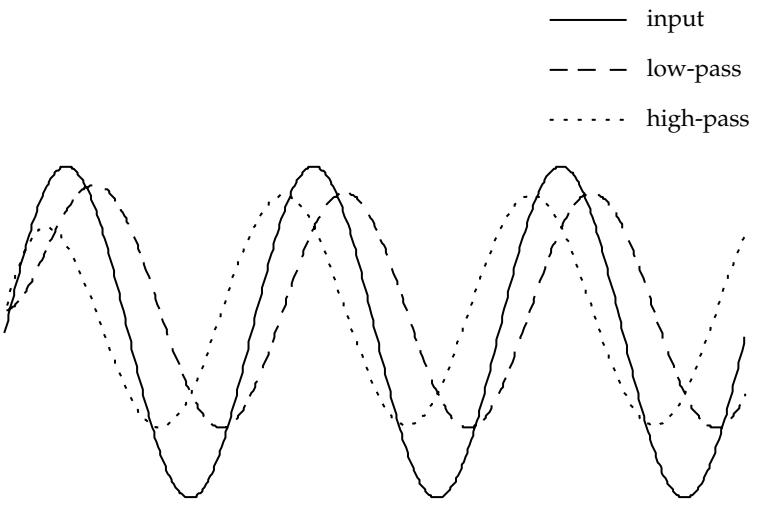


Fig. 1-21 Response of filters to the onset of a sine signal.

Note that at 0 Hz, i.e. at a pure DC, the response of a high-pass is essentially zero, which agrees with the expected insulation by the dielectric layer of the capacitance. At DC, the notion of phase ceases to be useful, although mathematically it is still defined.

Integration and Differentiation

Apart from filtering, the action of low-pass and high-pass filters may be thought of as mathematical integration and differentiation respectively. This can be illustrated with the responses of filters to sine signals just described. Consider the frequency characteristics of the high-pass filter again. At very low frequencies (say, a factor of ten below the cut-off frequency), the response to $U_{in} = \sin(t)$ is $U_{out} = \cos(t)$, which is the derivative, or differential quotient, of the input function.

This behaviour is not limited to sinusoidal signals: a high-pass filter is said to "differentiate" any low-frequency signal. Conversely, a low-pass filter, in this case driven at frequencies far above cut-off, is said to "integrate" the input signal. Indeed, the integral of $\sin(t)$ with respect to time is $-\cos(t)$, and this is indeed the function describing the attenuation *cum* 90° phase lag we saw earlier. Again, this property is not limited to sine signals. In fact, RC filters can be used to convert time signals into their time derivatives or integrals, but in a very inconvenient way: in order to get the desired property, the signal frequency must be far from the cut-off frequency, and so the output amplitude will be low, typically less than 10% of the input signal. Thus, an amplifier will be needed to push up the signal level again.

With the use of amplifiers, however, better differentiators and integrators can be built. This is treated in Chapter 2.

A second consequence of the principle of operation is that an RC differentiator or integrator must be fed with signals in a limited frequency range. Outside the mentioned range, a filter either passes the signal unaltered, or distorts the signal into a strange hybrid of the input signal and its time derivative or integral. This can be derived from the step responses shown in Fig. 1-18. The input step signal can be extended to a square wave signal, which is a series of steps from negative to positive and back. The response of different RC filters to such a square signal is shown in Fig. 1-22 (low-pass) and Fig. 1-23 (high-pass).

Fig. 1-22 Distortion of a square signal by low-pass filters.

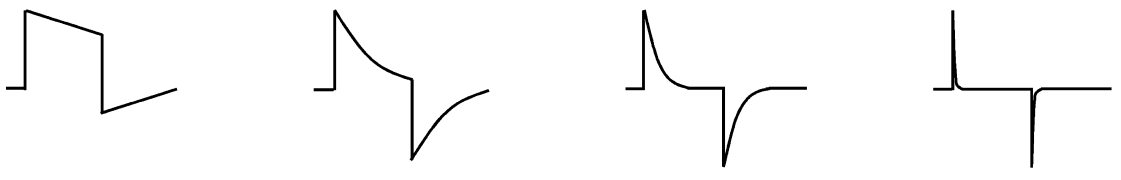


Fig. 1-23 Distortion of a square signal by high-pass filters.

Note that these graphs may also be read to show the response of one low-pass filter to square signals of increasing frequency, and of one high-pass filter to square signals of decreasing frequency. In that case, the subsequent graphs have different time scales.

From Fig. 1-22 it can be seen that a square wave is flattened more and more into a signal resembling the true time-integral, which is a triangle (check that in the interval between two steps, a square wave is simply a constant, and that the time-integral of a constant is a linear function). A linearly ascending or descending voltage is called a "ramp", whereas a linearly rising and falling signal is called a "triangle".

The situation for a high-pass is similar, although it might not be obvious at first why the peaks of the "differentiated" square signal resemble a time-derivative of that square. To clarify this issue, we must examine the rising edge of the input signal on a much faster time scale. Since the rising edge of any signal is not infinitely fast, we will find the sigmoidal shape shown in Fig. 1-24A. The concomitant output signal is shown in B, and can be seen, on the same stretched time scale, to be approximately equal to the time derivative of the input signal.

LC Filters

For completeness, we will digress briefly on LC filters, without going deeply into the theory. The two possible forms of combining a capacitance with a self-inductance are illustrated in Fig. 1-25.

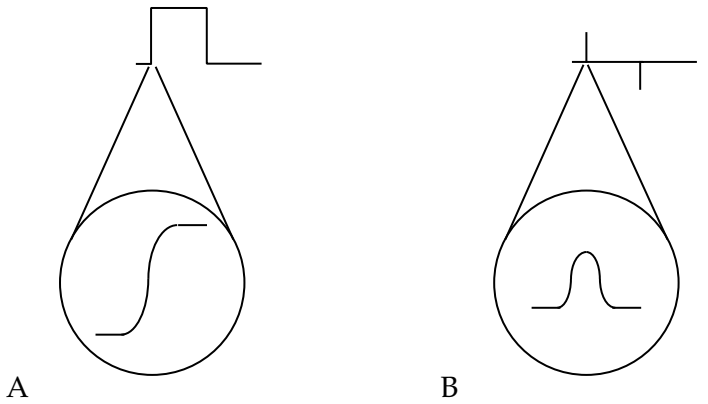


Fig. 1-24 Rising edge (A) and derivative (B).

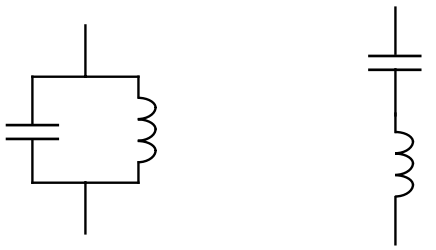


Fig. 1-25 Parallel and series forms of LC circuits.

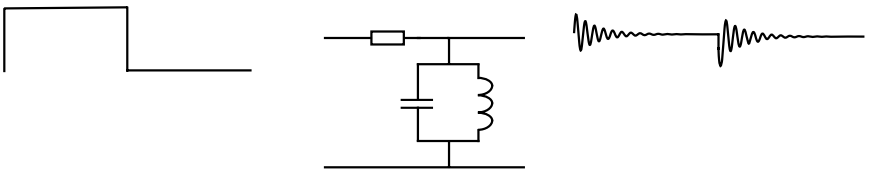


Fig. 1-26 Ringing by an LC circuit.

An LC circuit comprises of two components, both of which show reactance, and in an opposite way: a capacitor passes only changes, as we have seen, whereas the magnetic properties of inductances oppose changes in an electric current. The combination shows resonant behaviour, just as its mechanical equivalent: the combination of a mass and a spring.

The resonant frequency of an LC filter follows from:

$$f_{\text{res}} = \frac{1}{2\pi\sqrt{LC}}$$

where f_{res} is the resonant frequency in Hz, L the self-inductance in H, and C the capacitance in F. An LC filter is said to be tuned to this frequency. This means several things, depending on the way it is used. When fed with a step, impulse or pulse signal, the resonance shows up prominently, and is known as "ringing". An example is shown in Fig. 1-26.

Sinusoidal signals are treated as follows. At the resonant frequency, the impedance of the parallel circuit is maximal, i.e. higher frequencies are shunted by the capacitor, lower frequencies by the self-inductance. In combination with resistors, an LC circuit can be used to single out a narrow frequency band around its resonant frequency, and this is why it is employed as tuning circuit in radio receivers and the like. The series configuration has minimum impedance at the resonant frequency, so this configuration can be used to eliminate a narrow frequency band from any input signal. Therefore, the series LC circuit is called a "suction circuit".

In electrophysiology, it may be used to get rid of interference signals with known and constant frequency, such as the 50 (or 60) Hz mains frequency, or the carrier frequency of a nearby broadcast station.

This ends the chapter on electric processes. The manipulation of electrical signals is enhanced enormously by the use of what is called electronics. This is the topic of Chapter 2 of this book.

2

Electronics

ACTIVE ELEMENTS

So far, we dealt exclusively with the so-called passive components, which do not add energy to the signal (for the moment, consider any wanted quantity or message as a signal; later we will define the notion of a signal in a more precise way). What is "amplification"? Broadly speaking, mankind has found several ways to amplify its powers, literally and figuratively, and these may be compared to find both the possibilities and the limitations. The first step was to harness animal power and to exploit it for human purposes. In this way, a man (say 60 kg) controls the power of a bull (say 1000 kg) to plough the land. This example from antiquity illustrates the principle of amplification: the farmer uses his muscle power as a signal that controls a much stronger force, the bull's muscles. It also shows the expenses: the bull needs a large amount of food that must be delivered on a regular basis. More recently, the same principle was employed in the steam engine, where the energy from burning coal is converted into mechanical energy, that could drive a revolving shaft used for milling, pumping and the like. Note that in both cases the energy is delivered in a bulky form that, by itself, does not perform the wanted operations.

Active elements in electronics perform according to the same principle. The weak signal from a microphone or an electrode must be amplified to be able to feed a loudspeaker, or to show it on an oscilloscope, or to get it sent through a long cable. The necessary energy comes from a so-called power supply. In few cases, this may be an alternating current tapped directly from the mains outlet (lamp dimmers and variable-speed drills, for example), but most electronic apparatus are powered by a direct current, either from a battery or derived indirectly from the mains AC through a so-called rectifier, which we will analyse later on.

Amplification must not be confused with transformation. In Chapter 1 we have seen that a transformer may output a higher voltage than was fed into it, but with a proportionally lower current, so that power is conserved, not increased. In fact, there are losses of power into heat, making the power conversion always less than 100% effective.

Since the advent of vacuum tubes in the beginning of the twentieth century, the principle of controlling a large bulk force with a smaller, "structured" one became feasible for electrical quantities. Later, after the Second World War, semiconductors such as the transistor took over; so the largest part of this book is devoted to them. However, most of us spend many days staring at computer and TV screens, which are still working according to the vacuum tube

principle. Because of this continued use and because the working principle of vacuum tubes is easier to grasp than the physics of semiconductors, we will start explaining this venerable device. Later on, the different types of semiconductors, each with their specific virtues for electrophysiology, will be dealt with appropriately.

VACUUM TUBES AND SEMICONDUCTORS

In modern electronics, the active elements are usually transistors and similar "semiconductor" components. The basic structure of a semiconductor is a silicon crystal, or rather a sandwich of crystals of silicon to which slight amounts of other elements are added. Unfortunately, the processes in such crystals cannot be understood without a thorough knowledge of quantum mechanics. Hence, we will treat the mode of operation of transistors in a rather superficial way. In the old days of electronics, vacuum tubes were the active elements. Although nowadays these large, hot bulbs are seldom used, their mode of operation is much easier to understand.

Therefore, a brief digression on the principle of electric current in vacuum tubes is necessary here.

The first fundamental property is one-way conductance, or "rectification". In the late nineteenth century, pioneers like Braun and Fleming had found that in an evacuated glass bulb an electron current may flow as a kind of "rays" (cathode rays) if very high voltages are applied. Among others this led to the development of the picture tube, which is still in use today. Fleming found also that, at fairly low voltages, electrons will flow only from a heated electrode (metal wire) to a cold one, but not in the reverse direction. Such a device, having two electrodes, is called a "diode", and is depicted in Fig. 2-1.

This effect can be demonstrated by connecting a battery to the tube, anode positive. If an AC is applied, current will flow only during one half of the period: the AC is said to be rectified, since only one polarity is passed. By this principle, diodes are used to convert the mains AC into a DC suited to use as power supply for all kinds of apparatus.

Lee De Forest found (about 1905) that addition of a third electrode, called the grid (see Fig. 2-2), made the flow of electrons controllable. Moreover, it turned out (later) that the effect could be stronger than the cause, in other words that amplification takes place—electronics was born!

The influence of an electric field on current flow in vacuum is fairly easy to understand, and can be compared to the operation of a garden hose using your hand: a small amount of energy (a twist of the wrist) may turn the peaceful watering of your lawn into getting much energy from an angry neighbour. By a similar principle, a tiny current through the grid can control

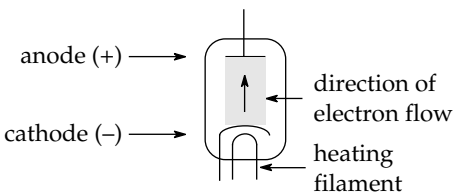


Fig. 2-1 Vacuum tube diode schematic.

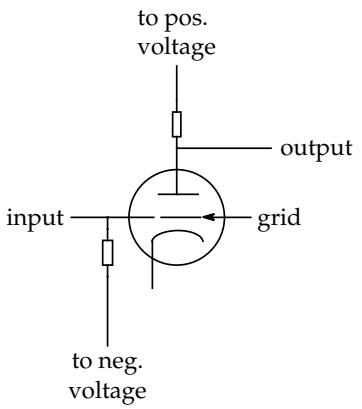


Fig. 2-2 Triode tube: the primeval amplifier.

a much larger current through the tube's other two electrodes, viz. by influencing the electric field, and thus either attracting or repelling the electrons.

The triode and similar vacuum devices have been the basic active elements of electronics from the very beginning up to the 1960s, when they were largely superseded by transistors. Therefore, our main attention will be attracted, in the next chapter, by these versatile, cheap and easily miniaturizable building blocks. Yet to date, the principle of the vacuum tube is still indispensable for special purposes, e.g. high-voltage and high-power circuits (trains, radio stations), and also still serves in most of the TV sets and computer displays using CRT picture tubes.

SEMICONDUCTOR DEVICES

Because electric currents in semiconductors flow in crystals, these devices are also called "solid-state" devices. The materials suited for the wanted phenomena are neither conductors (such as metals) nor insulators (such as glass, ceramics, etc.) and are therefore called semiconductors. In an ideal insulator, all electrons are bound to the atomic nucleus. Metals, to the contrary, have so-called free electrons that may move from atom to atom, thereby transporting electric charge. In other words, they may sustain an electric current. Semiconductors hold an intermediate position: at absolute zero temperature, they behave as insulators, whereas at room temperature they have a few free electrons. The materials with these properties are four-valued: germanium and silicon. These elements have four electrons in their outermost shell. Germanium was used first, but to date, silicon is used most often, and will be used in subsequent examples. In the pure material, however, the current that can flow is far too low to be of practical value. This type of silicon is called "intrinsic".

The interesting properties only arise by "doping" (the addition of very small amounts) with other elements, called "impurities". The addition to a silicon crystal of a five-valued element, such as arsenic or antimony, causes more free electrons, since only four do fit in the silicon's crystal lattice. The fifth electron is bound so weakly that it may be considered as free. Therefore, pentavalent elements are called "electron donors". The excess electrons can move freely through the crystal, and so turn it into an almost-conductor. Silicon prepared in this

way is called n-type silicon, or "n silicon" for short. In a similar way, trivalent elements, such as indium or gallium, cause a shortage of electrons, and are called "electron acceptors". By the movement of the existing electrons, the deficit of an electron is also mobile, and called a "free hole". It *acts like a positive particle*. Current flowing in such a semiconductor is called a "hole current". Silicon treated in this way is called "p silicon". The ability to use charge carriers of both signs constitutes a major difference with conduction in vacuum, which is by electrons exclusively. In semiconductor technology, interesting devices arise from the combination of p-type and n-type silicon. Occasionally, the intrinsic form ("i silicon") is used.

DIODES AND TRANSISTORS

In summary, a p semiconductor contains many free holes, but hardly any free electrons, whereas an n semiconductor contains many free electrons, but hardly any free holes. If these two opposite types of semiconductors are joined (or rather grown together) to form a so-called p–n junction, a device results that conducts an electric current only in one direction. This is shown schematically in Fig. 2-3.

If such a junction is connected so that the p-side is positive with respect to the n-side, the following happens. In the n-side, electrons are rushing towards the p–n junction, whereas in the p-side, holes are rushing equally in the direction of the junction. At the junction, the two types of charges, electrons and holes, combine. If the junction is connected in the reverse direction, hardly any current or no current at all will flow, because in the p-zone there are insufficient free electrons, whereas in the n-zone there are insufficient free holes. Only at very high reverse voltages (hundreds or thousands), the junction will break down (Fig. 2-3, left side of I/V curve).

Apparently, this device behaves as a diode, a "semiconductor diode". Practical diodes are illustrated in Fig. 2-4.

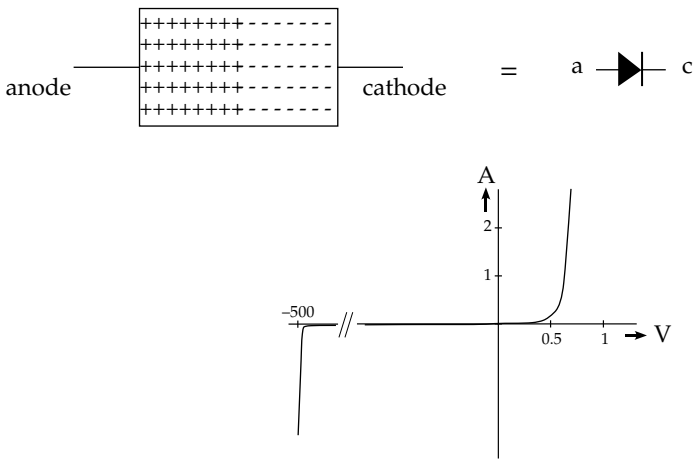


Fig. 2-3 p–n junction, its symbol (diode) and conduction characteristic.

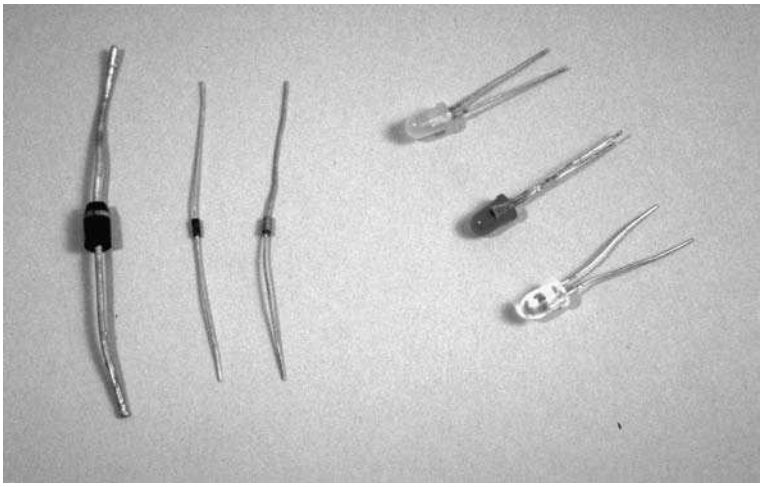


Fig. 2-4 Practical diodes. The three at the right are LEDs, or light-emitting diodes.

A disadvantage of silicon diodes is that conduction in the forward direction will take place only if the driving voltage exceeds the (approx.) 0.7 V threshold value (see I/V characteristic in Fig. 2-3). This means that lower voltages cannot be rectified.

This device, however simple, was truly revolutionary, since its inherently asymmetrical conductance performs the same task—rectification—as the vacuum tube diode, whereas it can be made far smaller, does not need a filament to heat the cathode, and works with lower voltages. If the polarity of a connected voltage is such that current flows, the diode is said to be connected in the forward direction, if the polarity is reversed, the diode is connected in the reverse direction. To build a better rectifier, silicon diodes can be joined into an array of four, which performs "full wave rectification": see Fig. 2-5. This so-called "diode bridge" circuit is the universal solution for rectification in power supplies.

A special type of diode emits light when connected in the forward direction. These light-emitting diodes (LEDs), made out of gallium arsenide or other exotic semiconducting materials, are used frequently to replace coloured signal lamps. The newest designs can even emit a bright, bluish-white light suited for illumination purposes. Note that LEDs have higher threshold voltage—1.6–3.6 V—depending on the colour (i.e. on the materials used).

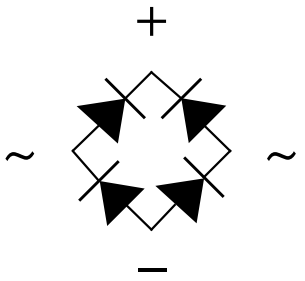


Fig. 2-5 Diode bridge.

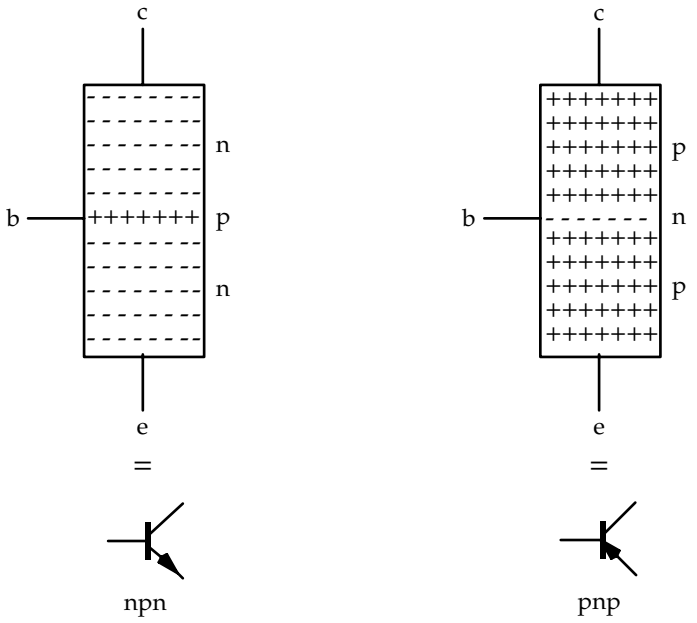


Fig. 2-6 Transistors.

The most important invention in semiconductor devices, however, is the semiconductor triode, usually called "transistor" (a contraction of transduction and resistor). As might be expected, a transistor has three zones of p and n silicon.

This implies there are two possible configurations: pnp and npn. The two types of devices are called "pnp-transistor" and "npn-transistor". They are said to be "complementary". The terminology of the three connections differs from the vacuum-tube jargon, since the middle zone (p or n) is called the "base", whereas the other two (n or p) are called "emitter" and "collector". Schematics are shown in Fig. 2-6.

The names can be explained from the original way of building a pnp-transistor: a thin slice of n silicon, the base, is joined with two drops of p-material (e.g. indium). By diffusion of the indium into the base crystal, two p-zones arise. Charge carriers are injected, or emitted, by one of the outer electrodes, and are collected by the other one, which explains their names. Again, our explanation of the mode of operation of transistors has to be superficial and qualitative, since a true explanation of the underlying principles would need a thorough knowledge of solid-state physics and hence of quantum mechanics. Fortunately, such knowledge is not necessary to use semiconductors in a proper way, as long as one has a good notion of voltages and currents in the device, and the relationships between them.

Superficially, the transistor configuration resembles two diodes connected back to back. If this were the case, no current would ever flow from emitter to collector. By quantum-mechanical phenomena, however, more interesting things happen: charge carriers, in the pnp example holes, may traverse the base, and so constitute a current through the transistor. The effect is that the middle electrode may control the current through a transistor, analogous to the way the grid controls the current through a vacuum tube. Practical forms of transistors are shown in Fig. 2-7.

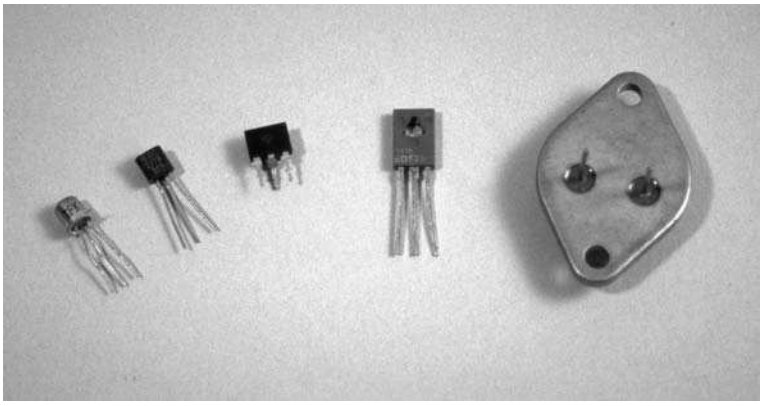


Fig. 2-7 Practical transistors. The size reflects the maximum current that can be handled.

OTHER SEMICONDUCTOR TYPES

The properties of semiconductor devices are governed by solid-state physics, which is a complicated science drawing heavily on quantum mechanics. Therefore, the processes leading to the myriad different devices and components used in all branches of science, including electrophysiology, are hard to grasp for outsiders. Therefore, we will mention only those components, and their properties, that are often most useful or even vital to electrophysiological measurements.

To start with, a "Zener diode" is a kind of diode in which applying a voltage in the reverse direction leads to an abrupt breakdown at a very sharply defined voltage. Not only is this breakdown reversible, the precise value depends on the composition of the semiconductor layers, and may be chosen between about 2.5 and 25 V. Therefore, Zener diodes are used as voltage stabilizers by forcing a current in the reverse direction, and using the maintained, well-defined voltage across the terminals as reference value for power supplies, voltmeters, AD converters, etc. Note that in this case, the cathode is the positive terminal. By the way, Zener diodes designed for 5.6 V Zener voltage proved to have the best properties as to stability, temperature (in)dependence and the like, and are often used as a basis for the derivation of other voltages. In the world of electronic tubes, a similar device is the neon lamp, since gas discharges usually work at a relatively constant voltage depending on gas composition and pressure, electrode distance, etc.

Many semiconductors show photoelectric phenomena, so diodes may be used either to emit or to receive light. One of the oldest and cheapest designs is a thin film of cadmium sulphide (CdS), between two conducting terminals. This device, called "light-dependent resistor" (LDR), acts as a resistor that has a very high value in the dark (over 10 MΩ), which may decrease to less than 100 Ω in bright light. LDRs are, however, rather slow: they can only follow constant light levels or relatively long flashes, but are widely used in photographic exposure meters, automatic cameras and so on. "Photodiodes" and "phototransistors" resemble normal diodes and transistors, but are specifically designed for the reception of light. In principle, normal transistors are intrinsically light-sensitive, which may be shown by opening the metal case of, e.g., a BC107 carefully. In photodiodes, the light-dependent current flows in the reverse direction, whereas the collector–emitter current in a phototransistor flows as if it were caused

by a base current proportional to the light flux. Like normal transistors and diodes, these components need a power supply to utilize the light sensitivity.

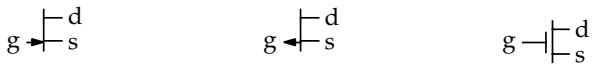
To the contrary, so-called photovoltaic cells and PIN photodiodes generate a current, which is approximately linearly dependent on the light flux when the cell is short-circuited (PIN stands for positive, intrinsic, negative; this diode has a thin, intrinsic silicon layer between the two sides of the junction, where charge carriers kicked by photons may "leak through"). The photo current may be measured with a sensitive current meter or amplifier, where the meter itself acts as the short-circuit. When left open, the light-induced voltage of these components may reach about 0.5 V in bright light. Solar panels are simply stacks of photovoltaic cells that generate 6, 9 or more volt useful to power small apparatus or charge accumulators. The photovoltage generated by a PIN photodiode, when left open, is approximately proportional to the log of the light intensity. Because of this property, they may be used to read density units or photographic "stops" directly, when connected to an amplifier with a high input impedance.

The semiconductor type that is no doubt the most important one in electrophysiology is the field-effect transistor or FET. The basic form, which is in fact older than the transistors described before, consists of a silicon strip or channel between two conducting terminals. In the middle is a p–n junction which, contrary to the normal transistor, is used connected in the reverse direction, that is, if the third terminal forms an anode, it is made positive with respect to the channel, so that no "base current" will flow. The important process is that the electric field in this reverse-biased junction causes the resistance of the channel to change.

In the most familiar form, the so-called n-channel depletion FET, the channel is made of n silicon, and the electric field in the junction is able to reduce the current in the channel, hence depletion. When a voltage is applied to the middle terminal, the current through the channel is modulated. It is therefore appropriately called the gate. The connections to the channel are called source and drain; see Fig. 2-8. In many respects, the FET resembles the vacuum tube, where it is also an electric signal field that modulates the current flowing from cathode to anode, but with a number of important differences. In the first place, an FET is small and has all mentioned advantages of other semiconductors. Second, semiconductor processes are more versatile than the electron flow in vacuum. Therefore, FETs may be made in different "flavours" to suit various needs. Like pnp and npn transistors, FETs come in complementary types, here called n-channel and p-channel, in which the polarity of voltages is reversed.

In addition to the mentioned depletion type, FETs may be made with a thin intrinsic region, so that the channel conducts no current unless the gate electrode bears a certain voltage. These are called enhancement-type FETs. As to most (profitable) FET properties, depletion and enhancement types perform equally well. Like normal transistors, the FET is symmetrical in principle, but asymmetrical in practice, that is, source and drain cannot be exchanged.

The property of FETs which is most important to electrophysiology is the high input impedance. The only junction of this transistor is reverse biased, so that virtually no current flows: the gate *voltage* only controls the current through the channel. Now, this must be



n-channel p-channel MOSFET

Fig. 2-8 Field-effect transistors.

corrected before you think we have reached Utopia. Any junction, operated in reverse bias, conducts a very small, but nevertheless important, so-called leakage current, and this current has to be provided by the input circuit. The great advantage of the FET over the normal so-called junction transistor is the amount of current needed at the input (base and gate, respectively). As we have seen, transistors need about 1 μA of current to function properly, whereas the leakage current of an FET is orders of magnitude lower: about 1 to 10 pA (10^{-12} A)! These current values may be translated into the so-called input resistance, or input impedance, by applying Ohm's law. If, in a normal transistor, 1 μA of base current flows at 0.1 V of base voltage, the equivalent input resistance would be $0.1/10^{-6}\,\Omega$, or 100 kΩ. As input stage of an electrophysiological amplifier, these transistors would be useful only if the electrodes used had a far lower resistance. This is the case for the large metal plates used in electrocardiography and similar techniques, but it would be utterly useless to monitor the electrical life of a single cell, where currents in the order of nanoamps are all there is.

Here, the FET scores far better: 1 pA per 0.1 V, or 100 GΩ of input resistance, is no exception. Therefore, FETs are the main building blocks of electrophysiological pre-amplifiers, and are used in a host of other situations where only small currents are available. And this is not even the maximum attainable: a variant called the MOSFET, short for metal-oxide semiconductor field-effect transistor, has a still higher input resistance, i.e. a still lower leakage current. This is made possible by doing away with the junction altogether, and just "gluing" the gate electrode to the channel via a thin metal-oxide layer. Metal oxides are very good insulators, and so the gate is effectively insulated from the channel. Yet, the voltage on the gate controls the channel current again, making it a valuable component for high-impedance preamplifiers and so-called electrometers used in many types of physical, astronomical (e.g. star light), chemical (pH meter) and biological (intracellular and patch-clamp) measurements. The MOSFET has a few hidden flaws, however, such as a higher noise level, in comparison with the above-mentioned "normal" FET, somewhat confusingly called junction FET, or j-FET.

Warning: Because of the high gate resistances, all FETs are extremely vulnerable to electrostatic voltages, and can be destroyed by handling, or merely by touching with a finger or plastic ballpoints. This also holds for devices that use FETs as input transistors, such as electrophysiological preamplifiers.

AMPLIFIERS, GAIN, DECIBELS AND SATURATION

The fulcrum of every electrophysiological set-up consists of an amplifier, called the preamplifier because it is the first apparatus connected directly to the measuring electrodes. In the early days of electrophysiological recording, amplifiers were made with vacuum tubes, and were often built by the physiologists themselves. Nowadays, semiconductors are used almost exclusively in pre-amps, and many types of pre-amp, versatile and almost perfect, can be delivered off the shelf from a host of reliable companies. This does not, however, absolve the user from the obligation to maintain and use the apparatus properly, especially since a lack of knowledge about one's tools may lead to the publishing of erroneous results, which is a waste of time, money and intellectual energy.

So, what are these important properties of preamplifiers?

Gain

This is the most conspicuous property, not necessarily the most important one. Usually, the voltage gain factor is defined as the ratio of output voltage to input voltage. But how much gain is necessary for an experiment depends on the ratio of the voltages to be measured over the voltage that can be "seen", or measured, by, e.g., voltmeters or by the most universal indicator, the oscilloscope. Usually, one takes 10–100 mV as measurable, although many oscilloscopes and electronic voltmeters may be useful down to 1 mV (indeed, they have amplifiers built in, which makes the notion of "necessary" gain a bit vague). The consequence is that for intracellular recording of membrane potentials, being tens of millivolts, no extra gain is needed. Thus, microelectrode preamps often have a gain switchable between 1 and 10. Why we cannot do without preamps will be clear later on.

The opposite case arises if the electrical life of one or more cells is measured from the outside or from a larger distance, such as in the case of extracellular recording electrodes, including electrocardiograms, electroencephalograms and the like, where a rather high gain is necessary, since the signal voltages in these cases may be reduced to a few μV or even less. In these cases, the gain factor is of crucial importance, and might be 10^4 to even 10^6. Because of these large and hard-to-read figures, the gain factor of amplifiers is usually expressed in decibels. A gain factor expressed in decibels is simply called gain. By the log transform, a large range of values may be expressed as fairly small numbers. In addition, when connecting two amplifiers in series, the gains just add, but gain factors must be multiplied.

The voltage gain is computed using the following formula, taken from Chapter 1:

$$G = 20\log(U_{\text{out}}/U_{\text{in}})\,\text{dB}$$

Thus, an electrophysiological preamp with a gain factor of 10 000 has a gain of 80 dB. The same relation holds for current gains.

The power gain follows from

$$\mathrm{G_P} = 10\log(P_o/P_i),$$

where G_P is the power gain, P_o and P_i the output and input powers, respectively. In audio circuits, one is still interested in powers. For instance, the power emitted by the head of a tape player, or by the venerable long-play record pick-up, is about 10^{-9} W (1 nW). In order to be audible, the loudspeakers emit, say, 10 W. Thus, a record player's amplifier has a power gain of $10\log(10/10^{-9})$, or 100 dB.

Note that in computing power gain, one has to take the input and output impedances (resistances) into account. Take a microelectrode preamplifier with a *voltage* gain factor of unity (i.e. 0 dB). Since the input impedance will be in the order of 1 TΩ (teraohm, or 10^{12} Ω), and the output impedance might be 10 Ω, the power gain factor is 10^{11}, or 110 dB. No wonder we cannot do without a preamp!

Bandwidth

The bandwidth tells what frequencies are amplified. The well-known audio amplifier has to amplify sound frequencies, i.e. about 20 Hz to 20 kHz. Both lower and higher frequencies are

not perceived by the human ear, and so need not be amplified or passed at all. Bats and crickets, by the way, would need audio amps with higher frequencies.

In other words, amplifiers have to filter the incoming signal, and pass only the wanted frequencies. Fundamentally, all amplifiers are low-pass filters, in that an infinite frequency is not physically possible. At high frequencies, the capacitance of any wire with its environment (the instrument case, the shield of the cable) acts as a parallel capacitor that attenuates the signal. At the highest frequencies, the wires themselves behave as self-inductances, enhancing the low-pass effect. This should come as no surprise since "in no time, nothing can change". Design and purpose determine which frequencies will be admitted, but there is always an upper limit.

But is there also a lower limit? In a strict sense, a frequency of 0 Hz does not exist either: this would mean a current or voltage that had been present for an infinite time, and will remain forever. For practical purposes, however, a direct current, or DC voltage, may be considered to have zero frequency: within the period of interest it will not change (polarity). The voltages of batteries, power supplies and the like fulfil this criterion, and there are no physical objections at all to measure them. A voltmeter is in principle capable of measuring the 1.5 V battery voltage as long as it exists. Early electrophysiological amplifiers had a lower limit to the pass band. In this case, the filtering was caused by capacitors used as coupling elements between the amplifier stages. A coupling capacitor between the collector of one stage and the base of the next stage simplifies the design because usually the collector bears higher voltages (e.g. about 6 to 30 V) than the base (about 0.7 V). This is still used in all cases where the DC value of a signal is not necessary, or not even wanted. Such amplifiers are called AC amplifiers, as opposite to DC amplifiers, the bandwidths of which reach down to zero.

The upper limit of the bandwidth of an amplifier, indeed of any circuit, is limited by so-called parasitic capacitance, also called stray capacitance. Parasitic capacitances are everywhere: any conductor has capacitance with its environment, even the shortest wire between two components. Especially coaxial cables, where the shield is close to the conductor all the way, have a capacitance of about 100 pF/m. Special foam insulation cables may have 30 pF/m, but using air as insulation keeps stray capacitances to the minimum. Keeping wires as short as possible and as far as possible from other conductors are measures used to reduce parasitic capacitances, both in the design of instruments and in wiring up an electrophysiological set-up. In addition, parasitic capacitances arise internally in resistors, transistors, switches and so on, so that full knowledge of the used components may help to keep stray capacitances in control, and thus to prevent the bandwidth from getting lowered unintentionally. Especially a preamp input circuit with a connected microelectrode is prone to unintended loss of bandwidth.

Often, however, the bandwidth of amplifiers is limited intentionally, to allow only the wanted signal to pass unchanged, whereas disturbing signal components such as hum and noise must be eliminated as far as possible. The most simple way of filtering, and the one most frequently used, is to incorporate the familiar RC filters, treated in Chapter 1. Thus, an amplifier fitted with a high-pass and a low-pass filter can be considered as a "bandpass filter with gain". As with these filters, the bandwidth (BW) of an amplifier is defined as the frequency band between the -3 dB points. It must be kept in mind that, despite this convention, signal frequencies outside this pass band are not fully suppressed, and one must always be aware of the frequency characteristics of the mentioned filters. In many designs, higher-order filters may be used that filter more sharply outside the wanted pass-band. Adjustable, electronic, higher-order filters are sold as separate instruments. The principles are discussed later in this

text. Even with the most sophisticated filter design, however, passing signals inside the pass-band completely and at the same time rejecting signals outside it completely is a physical impossibility. It is also possible to build band-reject filters in an amplifier, e.g. to reduce hum, but the same limitations hold here: it is impossible to reject the 50 Hz hum fully while keeping signal frequencies of, say, 49.99 and 50.01 Hz fully.

Most electrophysiological amps have rather broad pass-bands, whose positions depend on the function. For example:

for nerve membrane potentials	about 0 to 3000 Hz
for electrocardiograms	about 0.1 to 30 Hz
for electroencephalograms	about 1 to 50 Hz
for nerve or muscle spikes	about 300 to 3000 Hz
for plant action potentials	about 0 to 1 Hz
for the analysis of spike shape	about 0.1 to 100 kHz
for single-channel recording	about 0 to 50 kHz

Amplifiers with very narrow pass-bands are used mainly in other branches of electronics, such as radio communication, where one wants to tune in on a single broadcast frequency. There are, however, electrophysiological experiments where one needs to amplify a very narrow frequency band, being almost "a single frequency". This is usually performed with a so-called lock-in amplifier, which will be discussed later.

It is important to note that the passband of an amplifier must be chosen on the basis of the frequency contents of the signal, not merely the repetition frequency of a signal. To be sure: in order to allow a 1 Hz sine wave, one needs a passband centred on 1 Hz, but to pass a spike train, one needs to pass the fast-changing process we call a spike, or action potential, irrespective of the repetition frequency. Thus, a nerve spike amplifier needs a passband centred on about 1 kHz, whether a neuron fires 100 spikes per second or 1 spike per hour!

A second consequence of the use of filters in an amplifier is that outside the passband, signals are (linearly) distorted, in the sense that frequencies below the passband are differentiated, and frequencies above the passband are integrated.

Input and Output Impedances

These are crucial quantities that determine the usefulness of an amplifier in a particular situation. With regular electronics, however, the output impedance of any device is between 1 and 50 Ω, and so will cause no problems. The only demand is that the output impedance of a preamp should be about 30–50 times lower than the resultant input impedance of all devices that are connected to it. Since oscilloscopes, recorders, computer inputs and other electronic circuits have relatively high input impedances (100 kΩ to 1 MΩ), this demand is easy to fulfil. Passive voltmeters (moving-coil meters), some recording devices and audio amps may have lower input impedances, however, so one has to keep track of impedances when wiring any set-up. And never connect a loudspeaker (4–8 Ω) directly to any device that is not an audio amplifier designed to drive them.

People accustomed to wiring up their own audio chain will contend that input and output impedances must be matched. Indeed, in power-transferring circuits, impedance matching is often best: an 8 Ω amplifier must be used to drive 8 Ω loudspeakers. This is true, but with voltage measurement circuits, things are different. This can be explained as follows.

If an amplifier with an output impedance of 8 Ω is not loaded at all (the load impedance is infinity), the voltage at its output is maximal, but since no current flows, no power is developed. If the same amp is short-circuited (load impedance is zero), the maximum current flows, and the maximum power is developed, but unfortunately the output voltage is zero, and the power is spent entirely within the amplifier itself. The optimum energy transfer is attained by impedance matching, where exactly half of the power is transferred to the load impedance (in this case the loudspeaker). This may seem a poor result, but is nevertheless the best attainable. At the same time, the output voltage is exactly half the maximum voltage mentioned earlier, and this is the main reason to do it differently in circuits that have the purpose to measure voltages. If the membrane potential of a certain neuron is −60 mV, we want to measure this value entirely, not one-half of it.

In addition, impedance matching would be far too critical in this situation: if the cell impedance would change, the input impedance of our preamp would have to be changed accordingly. Therefore, one adopts the mentioned high-impedance rule: if each instrument has an input impedance that is about 30–50 times as high as its predecessor output resistance, the signal is attenuated only a few per cent and this is usually considered acceptable. For precision measurements, the ratio has to be raised accordingly.

A notable exception to this rule is when one wants to measure current rather than voltage, such as in the well-known patch-clamp recording. In this case, the ideal current amplifier (usually a current-to-voltage converter) has zero impedance, so the flow of current is not hampered. Here one has to obey the rule that the input impedance must be 30–50 times as *low* as the impedance of the preparation.

Maximum Signal Strength, Distortion

The maximum signal strength (in our case voltage) an amplifier can generate depends on a number of constructive details, the power supply voltages being the most important limitation. Transistor circuits operate with low voltages, although a few special (and expensive) types may handle a few hundred volts. Typically, modern amplifiers for measuring purposes run on 5–15 V, and that limits the output voltage to these values, or even a few V lower, as will be explained below.

Remembering the properties of conventional transistors, about 0.5 to 0.7 V is lost inherently in the way transistors function, so that in circuits with many cascaded transistor stages, some 2 to 5 V may be lost. Although FETs do not have any threshold, and so may be operated from zero up to the power supply voltage, most so-called FET amplifiers also use conventional transistors. As a guideline, then, it is safe to assume that an amplifier powered with 15 V may deliver about 10 V of signal voltage. Since most professional electronics operate on dual power supplies, i.e. plus and minus 15 V, the voltage swing at the output of an amp will range from about −10 to +10 V. Some amplifiers, especially battery-operated ones, may be limited even to a few volts (plus and/or minus). In any case, the amplifier's specifications sheet must be consulted to learn the maximally sustained output voltages.

What about the input voltage? This quantity seems to be derived simply from the output voltage and the gain. In principle, an amp that sustains ±10 V and amplifies one thousand times (60 dB) will be saturated if the input voltage exceeds ±10 mV. Things get more complicated, however, if filter circuits are involved. In this case, the input stages of an amplifier may be saturated by a DC voltage (electrode polarization!) which goes unnoticed because later on in the circuit it is filtered

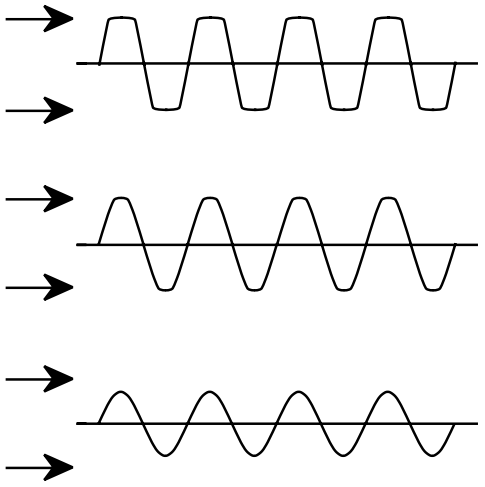


Fig. 2-9 Clipping.

out. Note that this means that the user may be kept unaware of the malfunctioning because the output voltage is nicely zero. Therefore, electrophysiologists have to have sufficient knowledge of the internal structure of their amps, and of the recording conditions, to prevent this situation from arising in an experiment, for instance when electrode polarization voltages rise slowly, as they often do, under aging, changing ion concentrations and the like.

What else may happen when an amplifier is operated with too strong signal voltages? In most cases, one is warned because the shape of the signal changes: a sinusoid of increasing amplitude "bangs its head" against the maximum output voltage, and is said to be distorted. This so-called clipping is shown in Fig. 2-9: the lower trace fits between the two saturation limits and so is undistorted. The middle and upper traces show increasing grades of distortion. It must be kept in mind that the transition to saturation is not always as sharp as is shown here, so the distortion "sneaks in" at slowly increasing amplitudes, leaving the user unaware again.

If the signal is a small AC, such as a spike train from a neuron, saturation at one side, caused by the above-mentioned increase in polarization voltage, may result in an apparent decrease in spike amplitude. Mistrust any such signals unless you are absolutely sure that amplifier malfunctioning is not the real cause.

In some cases, amplifier saturation is used to convert a sinusoidal signal into a square wave signal, but apart from sound effects in pop music, distortion should be recognized and avoided.

NOISE, HUM INTERFERENCE AND GROUNDING

By the virtues of electronic circuitry, weak signals arising in biological preparations may be amplified to drive oscilloscopes, audio amps, and a host of instruments for recording and analysis. Obviously, there will be a limit to the gain, or amplification factor, that may be employed before one reaches a fundamental or practical limit, where some kind of "interference", or unwanted signal, dominates over the wanted one. In a broad sense, any unwanted signal component is occasionally called "noise", but that term is best reserved for the random fluctuations

of electrical quantities that make the familiar, hissing sound if fed to a loudspeaker. Noise, hum and other forms of interference can be coped with successfully, but stem from different sources, and must be treated accordingly. Noise, then, is most fundamental because it is omnipresent in all physical, chemical and biological systems at temperatures higher than absolute zero, and so cannot be eliminated entirely. However, by carefully designing a recording set-up, noise can be—and must be, in most cases—minimized for the given situation.

Amplifier noise is illustrated in Fig. 2-10.

Note that not the absolute noise amplitude but rather the noise level relative to the signal strength is what determines the success of an electrophysiological recording. This quantity is called signal-to-noise ratio or S/N ratio and is usually expressed in dB. By a few, not-too-complicated calculations, one may assess the possible S/N ratio of a certain situation even before attempting an experiment. Usually, the signal strength depends on the way of recording, and can be estimated in advance using the following rules of thumb (in decreasing order):

Intracellular potential and spike recording	about 10–100 mV
Electrocardiogram (ECG)	about 100 μV–1 mV
Extracellular spike recording	about 10–100 μV
Electroencephalogram (EEG)	about 1–10 μV
Extracellular current measurements (with the so-called vibrating probe)	about 100 nV–1 μV

The famous patch-clamp recording techniques measure current rather than voltage:

On-cell or excised patch recording	about 1–10 pA
Whole-cell patch-clamp recording	about 100 pA–10 nA

Since the exact signal voltages depend much on the detailed properties of the preparation and the recording geometry, these are only indicative values, which nevertheless give a good idea about the order of magnitude involved, and will be used in the subsequent discussion.

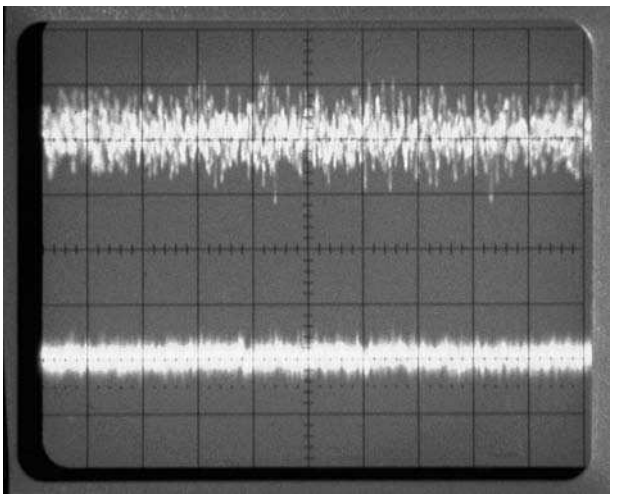


Fig. 2-10 Amplifier noise as it may show on an oscilloscope screen.

What about noise amplitudes? The most basic form of noise is called thermal noise or Johnson noise, and appears in any electrical circuit or component that has resistance. It arises by the random movement ("Brownian motion") of charge carriers in any conducting body. It can be shown, from statistical mechanics, that the power is dependent on temperature and bandwidth:

$$P_n = 4kTW$$

Here, P_n is the noise power, k is Boltzmann's constant (about 1.38×10^{-23} J/K), T is the absolute temperature (room temperature being about 293 K), and W is the used frequency bandwidth (i.e. $f_{max} - f_{min}$). As an example, the noise power over a 10 kHz bandwidth at room temperature amounts to 1.62×10^{-16} W.

With the help of the relations $P = U^2/R$ and $P = I^2R$, we can compute the voltage across or the current through a resistor:

This yields formulas for noise voltage and current:

$$U_{eff} = 2(kTRW)^{1/2} \quad I_{eff} = (4\,kTgW)^{1/2}$$

Here, U_{eff} and I_{eff} mean the effective value of the voltage and current respectively, which will be explained below, R is the resistance (in Ω), and g the conductance ($1/R$; in S). It is important to keep these formulas in mind, especially the fact that noise amplitude is dependent on resistance and on bandwidth: both can be controlled to a certain extent in an electrophysiological experiment. The noise voltage formula is used most, because most electrical measurements are voltage measurements. However, in configurations for current measurement, such as in patch-clamp recording, the noise current must be computed from the conductances involved.

Next, let us explain what is meant by effective voltage. In Chapter 1 we explained the notion of effective value of an alternating current, and found it to be $1/\sqrt{2}$ times the peak value. Now, most types of noise have a Gaussian amplitude distribution, which means that values around zero occur most, and values farther from zero (both positive and negative) are increasingly rare. Thus, a noise signal has no formal "peak amplitude" like any deterministic signal, but has an effective value, that would again be the value of a DC that gives the equivalent amount of heat. For Gaussian noise, this is the well-known root mean square (RMS) value, or standard deviation, which is the value from zero to one of the points of inflection of the Gauss curve. In this connection, it is useful to wonder what a noise voltage looks like on an oscilloscope. Low-frequency noise may look like irregular bouncing of the trace, but high-frequency (white) noise simply broadens the trace into a fuzzy horizontal band. In practice, the width, or amplitude, of this visible noise band is about 4–6 times the effective value, depending somewhat on the brightness of the oscilloscope beam relative to the ambient light level.

What about the bandwidth of thermal noise? A noise signal has a so-called continuous spectrum, i.e. it can be considered as a mixture of an infinite number of frequencies, each with infinitesimally small amplitude. Therefore, the greater the admitted bandwidth, the greater the noise amplitude. From the formula, it can be seen that the noise voltage is dependent on the square root of the bandwidth. Since power is proportional to the square of the voltage, the power spectrum is flat: equal frequency bands hold the same noise power, whether from 0 to 1 kHz or from 1.000 to 1.001 MHz. Therefore, thermal noise is also known as white noise, a term derived from the fact that white light contains all wavelengths at (approximately) the same intensity. Although the parallel is somewhat sloppy (in what we perceive as white light,

all wavelengths in fact do not carry the same power density), but the notion of a "white" spectrum is nevertheless well defined.

The mentioned formula for noise voltage leads to a fundamental rule for electrophysiological, and indeed all other, measurements: don't use a wider bandwidth than necessary to preserve signal shape. A ten times higher bandwidth results in a $\sqrt{10}$ times, or about three times, higher noise level. Note that the upper limit to the bandwidth is usually the most important one in this respect. Since we used mostly a rather wide passband, reducing the bandwidth from 1–1000 Hz by a high-pass filter to, say, 10–1000 Hz, one has still about 1000 Hz of noise-admitting bandwidth. Note also that, since frequencies just outside the formal bandwidth do contribute a bit of noise, the noise bandwidth of an amplifier is slightly larger than the -3 dB limit suggests, usually (i.e. if first-order filters are applied) $\pi/2$ (approx. 1.57) times as high. Since noise calculations are usually intended to monitor orders of magnitude, the effect of this correction is of minor practical importance.

Armed with the knowledge just acquired, we may compute the thermal noise of a glass microelectrode before even having tried one. Let us assume that the required bandwidth is 5 kHz. Since the resistance is the other important quantity, the tip resistance of the pipette is the main cause of thermal noise. Micropipettes intended for intracellular recording usually have a resistance of about 100 MΩ. Filling these values in the Johnson formula yields $V_{\text{eff}} \approx 100\,\mu\text{V}$. This means that signals to be measured must be substantially larger than this value. Fortunately, intracellular signals are usually in the millivolt region, and extracellular (semi/micro) electrodes have larger tips, and hence lower resistances, than intracellular ones. Unfortunately, there are more types of noise that play a part in electric measurements.

The second type to deal with is called shot noise. It arises by the random movement of charge carriers through a barrier, such as a vacuum tube cathode, a p–n junction or a cell membrane. Therefore, shot noise occurs not only in all kinds of devices such as amplifiers, but also in living matter. Shot noise arises because even in a steady (DC) current, the number of charge carriers passing the border at any moment is a statistical quantity. Thus, it can be compared to the noise hail stones make on a metal roof: each hail stone contributes a minute tap, imperceptible in the whole hissing sound that is nevertheless made up of myriads of these tiny sound impulses. In the same way, the electrons leaving the cathode of a tube, cross the p–n junction of a semiconductor or the membrane of a cell, form a current of which the random fluctuations are a form of shot noise. The magnitude of shot noise can be derived by statistical arguments, in which the elementary charge, the total number of charges (hence the total current) and the time, or its inverse, the bandwidth, play a part. The current noise amplitude follows from

$$I_{\text{eff}} = (2\,eIW)^{1/2}$$

where e is the elementary (electron) charge (about 1.6×10^{-19} C), I is the current, and W is the bandwidth. Note that the fluctuations become more prominent at lower current strengths. Thus, a membrane current of 1 μA using a bandwidth of 10 kHz has a shot noise component of 57×10^{-12} A, or 57 pA, which is only a minute fraction of the total current. At 1 pA, the order of magnitude of ion-channel currents, the noise amounts to about 6×10^{-14}, or 6%, and at 1 fA (10^{-15} A) the current would exist mostly as shot noise: 1.8×10^{-15} or 180%!

Like Johnson noise, shot noise has a "white" spectrum and a Gaussian amplitude distribution. It must be mentioned that a certain type of electrophysiological signal resembles shot noise. This is the case when one records with relatively large electrodes from a whole nerve bundle.

With this so-called gross-activity recording, each spike is a minor click, barely detectable if at all, whereas the amplitude of this noise-like signal correlates well on the average spike activity of the fibres in that bundle. The spectrum, however, is not entirely white, since a spike is not an (infinitely short) impulse. Hence, the spectrum of a gross-activity signal will be the same as a single-unit or population spike recorded under comparable circumstances.

Note that, if an electrophysiological signal has the characteristics of noise, extra care must be taken to discriminate it from instrumental noise.

A third type of noise is the so-called excess noise, pink noise, or 1/f noise. This is often the largest noise component, but unfortunately the hardest to grasp, in a signal. The name is derived again from the power spectrum, which in this case is inversely proportional to frequency. The term "pink" is again a loose parallel with light, and alludes to light-red light, where long wavelengths (and hence low frequencies) are more abundant than shorter ones. One-over-f noise originates again in tubes and semiconductors, and presumably also in the cell membranes we study, but on the other hand is not as fundamental as the other two noise types. Therefore, the amount of 1/f noise in transistors, and hence in amplifiers, may be reduced by certain aspects of the design, and thus may be influenced by the designer, not by the user, of scientific instruments. Because of the mentioned power spectrum, excess noise is most conspicuous at low frequencies. At higher frequencies, there is always a point beyond which the white-noise types are dominant. The transition point is called the $1/f$ corner (Fig. 2-11), and is a test characteristic of commercially available preamplifiers.

The other forms of interference to an electrophysiological signal are less fundamental, yet often hard to combat. This holds especially for hum, the often persistent periodical disturbance (see Fig. 2-12) that stems from the mains lines; abundant in any house or laboratory. The frequency will be 50 or 60 Hz depending on the continent: 50 Hz in (among others) Europe and 60 Hz in (among others) the Americas.

Careful design of sensitive electronic apparatus and a careful and "clean" way to build a set-up help to minimize the detrimental influence of hum on electrophysiological signals. The main therapy against hum trouble consists of grounding and shielding. This illustrates that there are in fact two forms of hum.

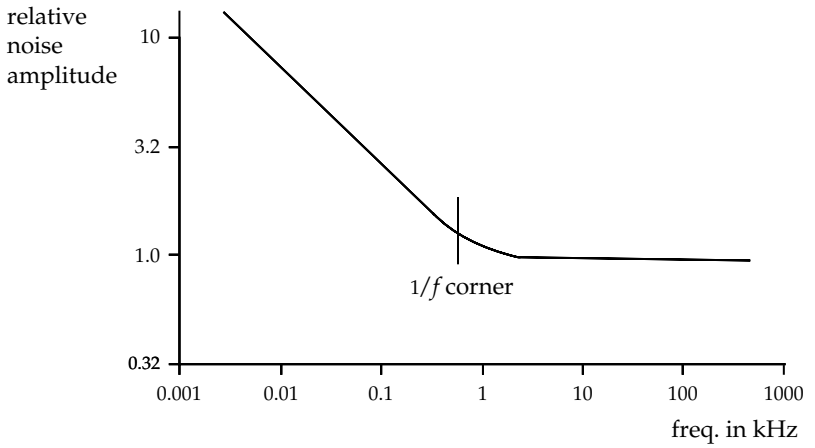


Fig. 2-11 One-over-f noise and corner.

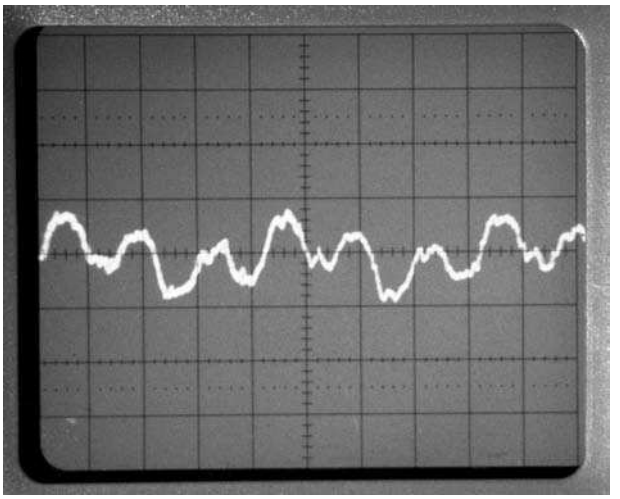


Fig. 2-12 Hum as it may show on an oscilloscope screen. Since the shape is not a pure sinusoid, we conclude that it is at least partially caused by magnetic fields from the mains.

The first one may be called electrostatic, since it is caused by the electric field emanating from the mains lines. These have to be kept as far as possible from an electrophysiological preparation, since they carry relatively high AC voltages. The electric field emanating from mains lines is sinusoidal, and so causes a hum signal that has the same shape. A simple and effective cure is shielding, i.e. covering instruments, cables or the whole set-up by a grounded, conducting layer of metal. Most scientific instruments are built in a metal case, and signals are transported via shielded, or "coaxial", cables (see Chapter 1) between the instruments. The preparation itself, and any other "open" parts of the set-up, may be enclosed in a so-called Faraday cage. In hospitals, complete rooms for EEG recording may be shielded this way.

The other type of hum is caused by the *current* flowing through the mains lines, instrument cables, etc., causing a magnetic field that may reach our set-up. Contrary to the "static hum" described above, a "magnetic hum" signal is often peaked, caused by the bumpy current flowing through gas discharge tubes (such as fluorescent lights), power supplies and motors (refrigerators, centrifuges, etc.), and from switching apparatus on and off. Here, the only cure is keeping distance, together with a proper grounding technique. For the weak signals encountered in electrophysiological research, a proper way of grounding is absolutely vital, and consists of several measures, explained below.

The above-mentioned grounding of all instrument cases, cables and shields is suffice to keep static voltages from building up.

If one grounds too eagerly, however, ground loops may emerge, and these are the causes of much trouble. This can be explained by recalling the principles of electromagnetism. By mutual induction (the same principle that makes transformers work), an AC current through a mains or ground cable induces a voltage in any nearby wire. Whether this (AC) voltage will be accompanied by a current in this wire depends on the impedance of the loop: an open circuit carries no current, and the voltage appears simply at the ends of the wire. If this is, say, 1 mV, it will be relatively harmless in most electrophysiological set-ups. If, on the other hand, the loop is closed, making the loop impedance, say, 10 mΩ, a current of 100 mA would be the result. This

current may induce yet another hum voltage in other nearby cables, and so on. Unfortunately, this situation arises if a chain of measuring instruments is wired up in the normal way, i.e. by plugging all instruments in an AC wall outlet and interconnecting the instruments by means of coaxial cables. This is illustrated in Fig. 2-13A. The ground loops, indicated by circular arrows, may span more than a square metre, and so can pick up numerous stray fields from mains cables and transformers, especially from the instruments themselves. Several remedies are in use. The first solution, shown in Fig. 2-13B, is interrupting the shields of the coaxial cables connecting the individual instruments, so that the safety ground connections of the instruments provide the sole ground connection. A complication is that the safety ground, supplied with the mains outlets, is often too "dirty" to be used with delicate measuring instruments. This is because these wires carry ground leak currents from any apparatus connected to them, such as refrigerators, centrifuges, heating baths and other laboratory instruments. These large and often "peaky" currents may cause many millivolts to develop across the ground circuit, notwithstanding its very low internal resistance. For electrophysiological measurements, a better instrument ground is often mandatory, and should be provided in every laboratory. In this case, the best ground "mecca" is at the input circuit of the pre-amplifier, i.e. at the specimen ground connection (Fig. 2-13C, left part).

This introduces another problem, however, since the solution sought by many electrophysiologists—removing the mains grounding wires from the instruments—is on bad terms with security. This is especially hazardous because in building a set-up, most people connect the power cords first, then wire up ground and signal connections. The problem is aggravated by the fact that in computers and other instruments employing MOSFETs, omitting the ground connection may blow all chips at once, and so must be totally avoided.

The best solution to this grounding dilemma is either to use only the security ground (in case it is sufficiently clean; Fig. 2-13B), or to separate the ground circuits of digital (computers, counters, etc.) and analogue (preamps, etc.) apparatus. The two ground circuits may be connected by a resistance of, say, 10 Ω because in that case no significant loop current can flow, whereas from a static viewpoint all components are connected to "ground" potential. This is shown in Fig. 2-13C (right part).

Here, introducing a 10 Ω resistor into the ground loop reduces the loop current by a factor of up to 1000 (because the loop resistance jumps from about 10 mΩ to 10 Ω). Some measuring instruments, such as preamps, even have the ground circuits of their input and output separated by such a 10 Ω resistor. This is particularly useful, since an electrophysiological preparation may have "false" ground connections through thermostats, saline flow circuits and the like. In many practical situations, one has to fiddle around with grounding wires to find the best configuration, i.e. the way that minimizes hum and interference.

The ultimate solution to the multiple ground syndrome is the use of isolation amplifiers. Although not standard in every lab yet, this solution must be considered seriously, since it will not be unduly expensive, and may prevent a lot of trouble. Electronic musical instruments are already coupled through optically isolated (MIDI) interfaces.

Occasionally, one may find a form of hum that has a frequency of 100 (120) Hz rather than 50 (60) Hz. Usually, this stems from insufficiently filtered power supplies, and so has an obvious remedy: increase the ripple filter capacitance or buy a better power supply.

The last form of interference that may vex electrophysiologists is radio-frequency interference, or RFI, mostly from the strongest broadcast stations, occasionally from a nearby radio amateur, or even from the campus beeper network.

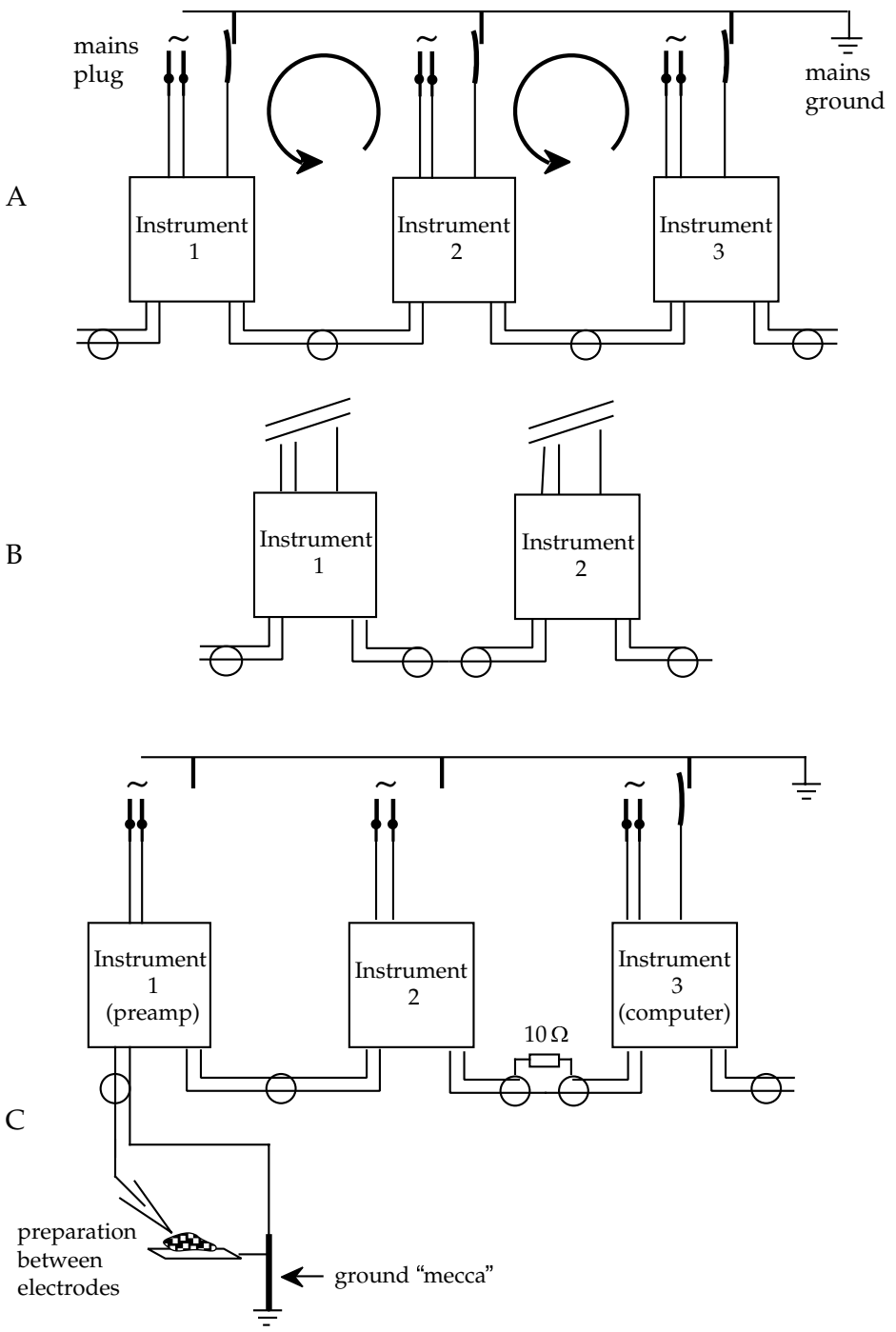


Fig. 2-13 Grounding a chain of measuring instruments.

In principle, the remedies for these disturbances are the same as described for static hum, i.e. shielding and Faraday cages. In reinforced concrete buildings, where the steel rods provide some shielding, it often helps to keep electrophysiological set-ups far from the windows, i.e. closest to the centre of the building.

DIFFERENTIAL AMPLIFIERS, BLOCK DIAGRAMS

The basic form of amplifier we discussed so far has only one input. Thus, the output voltage U_{out} can be described as the input voltage with respect to ground, U_{in}, multiplied by a gain factor, here called M:

$$U_{out} = M \times U_{in}$$

However, it is often necessary to measure the voltage between two points where neither of them is connected to ground. In electrophysiological recording, one has, for instance, often one electrode in or near a nerve fibre, the measuring electrode, and a second one nearby, the reference electrode. This configuration is called differential recording and needs a differential amplifier.

Such a preamp has two input terminals, usually called "+" and "−", or "A" and "B", respectively, and the circuit is such that the output is dependent on the difference between the two input signals, again multiplied by a gain factor M:

$$U_{out} = M(U_A - U_B)$$

Ideally, this would mean that a voltage V_C, *present on both inputs*, would not appear at the output, because it would be cancelled by the subtraction:

$$U_{out} = M[(U_A + V_C) - (U_B + V_C)]$$

In electrophysiological recording, the relevant signals are often as follows:

V_A—or signal, voltage at the recording electrode;
V_B—or reference, voltage at the reference electrode; and
V_C—voltage caused by some kind of interference (hum, RFI).

The latter is called a common-mode voltage, whereas the difference $(V_A - V_B)$ is called the "differential voltage" or signal. Ideally, then, a differential amplifier would just amplify the differential signal without being influenced by the interference. In a practical situation, the signal might be a 1 mV ECG (left arm minus right arm), whereas radio waves or mains cables cause an interfering common-mode signal of, say, 200 mV. Without a differential preamp, then, the ECG would be totally obscured by the interference.

In practice, however, differential amplifiers are not perfect, i.e. mathematically exact, so traces of a common-mode signal can be found at the output if it is relatively large, as in the previous example. The degree to which a preamplifier suppresses a common-mode signal is called the common-mode rejection (CMR), and is usually expressed in dB. The CMR is defined

simply as the gain of the amp for differential signals (the "normal" gain) minus the gain for a common-mode signal. As a practical test for preamps, one can (and must!) determine the CMR by connecting a signal source such as a sine-wave generator to both A and B inputs at the same time. If, for example, the gain of an amp is 60 dB (i.e. 1000×), and a common-mode signal is attenuated by 20 dB (i.e. the common-mode gain is −20 dB), the CMR is $60-(-20)=80$ dB. Well-designed electrophysiological preamps should have a CMR of at least 100–120 dB.

To indicate the difference, an amplifier having only one input is called a single-ended amplifier. In fact, a differential amplifier can be considered as two amps in one: if a single signal voltage is applied to input A, while grounding input B, the amp behaves as a normal, single-ended amplifier. If input B is connected to the signal source, and A grounded, we have again a single-ended amp—however, one that reverses the signal polarity:

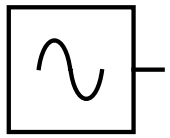
$$U_{\text{out}} = -MU_{\text{in}}$$

This configuration is called an inverting amplifier. This is often useful by itself, and is the reason why the A and B inputs are also called + and − inputs respectively. In words, the inputs are called inverting input (B or −) and non-inverting input (A or +). Note that the + input does not need to be fed with a positive voltage, and that the − input need not necessarily be kept negative: the symbols stand only for the polarity of the output with respect to the corresponding input.

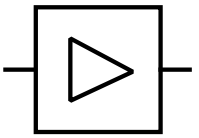
To discuss differential amplifiers any further, we need a convenient shorthand to distinguish the different types of amplifiers. Up to now, we have seen the schematic symbols used to represent transistors and other semiconductors, which together may make up an amplifier or other electronic instruments. These symbols are useful to elucidate the functioning of simple circuits, but it will be obvious that most instruments are so complicated that it is neither necessary nor feasible to draw all components separately. Thus, new symbols were defined to represent complete instruments rather than components. These include both single-ended and differential amplifiers, voltmeters, oscilloscopes, signal generators, filters, loudspeakers or audio amps, and so on. Schematics composed of these meta symbols are called block diagrams, and are indispensable in describing one's set-up. The most frequently used, and best standardized, block symbols are depicted in Fig. 2-14. As you see, most symbols consist of a rectangle with input and output connections, as well as a schematic indication of the function or contents of the box. New ones can be invented easily to describe new or special functions. In the most general form, a block symbol consists of a box containing the name or type number of an electronic apparatus. The boxes must be connected with straight lines, as short and neatly arranged as possible. By adding (arbitrary) symbols to reflect measuring and ground electrodes (Fig. 2-15), experimental subject, tissue preparation or cells and the like, the fundamentals of any electrophysiological recording situation can be made clear.

Every experimenter should make a habit of drawing a complete block diagram of the set-up used before doing experiments.

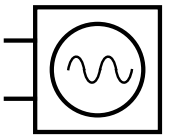
Differential amplifiers are often used in electrophysiology in situations where interference voltages are high—or signal voltages are low—as in the case of most laboratories doing extracellular nerve recording, EEG, or patch-clamp experiments. A proper explanation of the advantages of differential recording needs a discussion of the geometry of the electric fields involved, but the simplification used in Fig. 2-16, although not entirely correct, may help to grasp the principle. Here, the cell or nerve to record from is indicated as the "object", whereas the rest of the animal, tissue and/or saline is called the "animal".



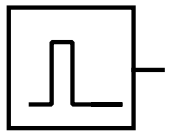
sine generator



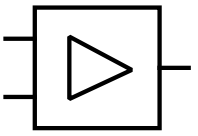
amplifier single-ended



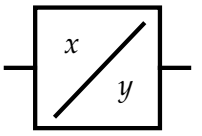
oscilloscope



pulse generator

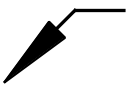


amplifier differential



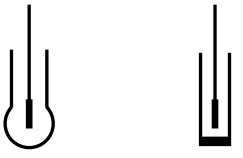
x to y converter

Fig. 2-14 Symbols for block diagrams.



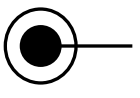
metal microelectrode

glass microelectrode



pH and reference electrodes

ECG (Einthoven) electrode



precordial or EEG electrode

ground electrodes

Fig. 2-15 Suggested symbols for electrodes.

The main point is that grounding electrodes need to have a low resistance (or impedance, since most forms of interference are alternating currents), say 1 kΩ or less, whereas measuring electrodes have to be small, and hence have a much higher impedance: megohms or more. If one would use a 10−100 MΩ microelectrode as a ground electrode, even the relatively weak currents stemming from nearby mains cables or the local broadcast station would give rise to high interference voltages, masking the signal to be measured.

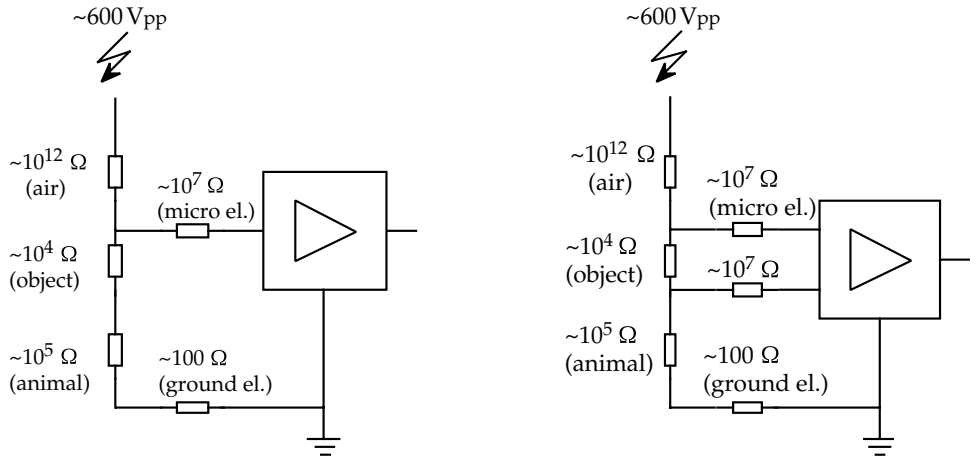


Fig. 2-16 Single-ended versus differential recording.

OPERATIONAL AMPLIFIERS, FEEDBACK

The so-called operational amplifier, or op-amp for short, is the basic building block of the analogue computer, the large signal processing machine that was used in laboratories before the digital computer made its soaring rise. Like the digital computer, the op-amp is used so widely nowadays, and has become so cheap (and so small), that the many devices incorporating op-amps are no longer considered to be "computers". The name still refers to the mathematical "operations" that can be performed with these devices.

The first operational amplifiers, built in the 1940s with vacuum tubes, were inverting amplifiers, and, although nowadays most op-amps are differential by design, many functions are performed with the non-inverting input grounded, thus reducing it to the inverting amp referred to in the last paragraph.

The clue to understanding the functioning of op-amp circuits is to understand the principles of feedback, since this is what makes them work in a so well-defined way. Feedback comes in two flavours, positive feedback and negative feedback, and the classical example of the latter form is how cyclists and car drivers can keep their vehicles on a straight path despite all kinds of diverting forces: wind, irregular pavement, sloping road surfaces and so on. A gust of wind from the left, for example, would push the vehicle to the right, and the driver counteracts this force with an equally strong force in the opposite (or "negative") direction. As a result, the driver continues in the wanted straight path. In other words, by compensating all diversions continuously, the wanted situation can be maintained indefinitely. Positive feedback, on the other hand, is like steering with crossed hands: any disturbance will be amplified instead of corrected, and will usually end in disaster. In electronics, feedback is used in the same way: negative feedback to stabilize any wanted situation against unwanted influences or fluctuations, positive feedback to generate electrical "disasters" such as pulses or other types of oscillations.

Let us return to negative feedback and see how this is applied with operational amplifiers. A simple, one-transistor amplifier might have a gain of about 50 to 300 depending on several properties of the used transistor, such as temperature, aging and so on. By cascading transistors, gain may be pushed easily to over one million times (120 dB), but the uncertainties grow

proportionally. Moreover, if such a component would have to be replaced, the properties of the whole device might get altered in an intolerable way. So, most transistor types are convenient but unreliable components, having wide ranges of gain, temperature coefficients and so on. As we will show, negative feedback may be used to stabilize gain, so that an amplifier may be built that has distinctive properties independent of temperature, aging and the properties of unreliable components.

An op-amp is an amplifier with a high gain and a very high input impedance. A feedback circuit is provided by a few added components, often resistors. A simple circuit suitable to analyse the mode of operation is depicted in Fig. 2-17. The op-amp is used as an inverting amplifier (+ input grounded). The feedback is effected through R_2 (10 kΩ), whereas R_1 (1 kΩ) may be called input resistor. Let us analyse this circuit for the case in which the amplifier would have a gain factor of 1000, and the input impedance is so high that virtually no current flows through the input terminals. Let us assume further that there are no offsets, which means that an input signal of zero yields an output signal of zero. If we connect an input signal, say +1 V, a current will tend to flow through R_1 that makes the inverting input positive. Obviously, the gain of the amplifier will tend to drive the output to very high negative values, but by the action of R_2, the voltage at the inverting input is reduced again. This is the principle of negative feedback, and will yield an equilibrium state at approximately the following values. The output signal will be about −9.9 V, the voltage at the inverting input will thus be about +9.9 mV. The end result is an amplifier with a gain of almost (minus) 10 times.

Now what will happen if we increase the gain of the op-amp? If the gain factor is increased to one million, the output voltage will be closer to 10 V, whereas the voltage at the inverting input will be a mere 10 μV. Thus, the circuit is now a nearly perfect times-10 amplifier. This leads to the simple rule of thumb to understand the op-amp functioning: if the gain approaches infinity, and the input impedance too (no current through the inputs), the output voltage of the op-amp will be such that the inverting input is held at (very nearly) the same potential as the non-inverting input. If the non-inverting input is grounded, such as in our example, the inverting input is kept *actively* at ground potential. Therefore, it is called the virtual ground. In this case, the output voltage follows simply from the current involved: an input voltage of 1 V drives a current of 1 mA through R_1. Because of the infinite input impedance, the same current must flow through R_2, yielding an output voltage of $-R_2/R_1$ V.

Our amplifier has a modest gain factor, but, it is virtually independent on the gain of the used op-amp (the gain without feedback, called the open-loop gain). The closed-loop gain, although only 10, is much more reliable, and depends mainly on the values of the added components. In real op-amps, the open-loop gain may drift and be temperature-dependent between, say, 50 000 and 250 000, but this has only negligible consequences for the properties of the entire circuit. Needless to say that the used resistors must be precise enough, and have a good

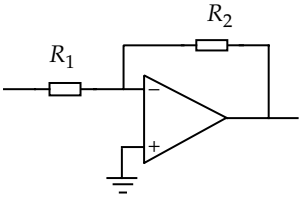


Fig. 2-17 Basic op-amp circuit.

stability, but this is much easier to accomplish: metal film resistors have usually a tolerance of 1% or better (and may be selected to match within 0.1%), and have a very low temperature coefficient: about a thousand times better than most semiconductors. Practical op-amps are cheap (a few dollars to about 100 dollar), have open-loop gains of at least 100 dB (100 000×), and have input impedances of at least 1 GΩ. The bandwidths reach into the MHz range. So, by designing op-amp feedback circuits, we can harness these near-perfect devices to perform all kinds of tasks, or operations, necessary to collect or process electrophysiological signals.

A few examples are given below. Figure 2-18 shows an op-amp as adder. Since the inverting input behaves as the virtual ground, currents into this point simply add in a linear way: the output current is the sum of the input voltages. By choosing different values of R_1 and R_1', the inputs can be given different weights. Subtraction is done with the circuit of Fig. 2-19. This is a way to build a differential amplifier. Note, however, that the input resistors R_1 and R_1' determine the input impedance of the whole circuit, so it is not suited as differential input stage for electrophysiological amplifiers. Finally, the principles of addition and subtraction may be extended to any number of input signals by adding extra input branches. The use of op-amps is not limited to amplification or summation, as is shown in the following circuit (Fig. 2-20).

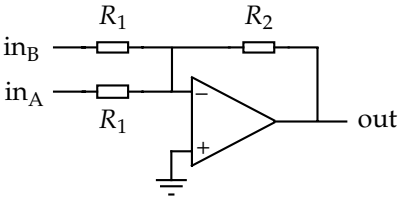


Fig. 2-18 Op-amp adder.

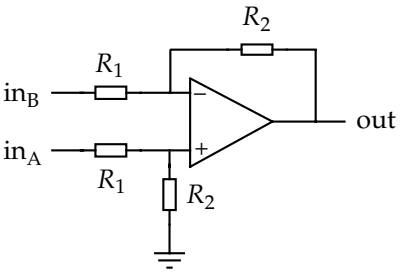


Fig. 2-19 Op-amp subtractor.

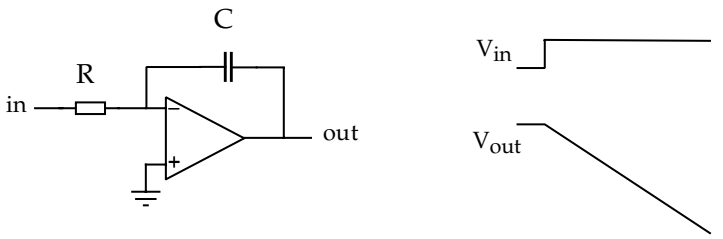


Fig. 2-20 Op-amp integrator.

A close look at the circuit of Fig. 2-20 shows that by adding a capacitor, we have built an electronic integrator: the output is the integral of the input voltage, with RC as the familiar time constant:

$$U_{out} = -1/RC \int U_{in}\, dt$$

We will analyse the integrator by letting $R = 1\,M\Omega$, and $C = 1\,\mu F$, and feeding the input again with 1 V DC This yields an input current of $1\,\mu A$, which charges the capacitor with 1 V/second. Thus, the output voltage runs away with (minus) one volt per second. Apparently, then, measures have to be taken not to get stuck at the maximum output voltage. This means that practical integrators will need extra components to prevent build-up. The most simple one is a reset button across the capacitor to discharge it at will, but in most applications, the signal voltages are limited electronically, for instance by reversing the input signal. This is the basic principle of the so-called function generator, which will be dealt with later on.

By exchanging the resistor and the capacitor of the integrator, the circuit is turned into a differentiator, which renders the derivative of the input current. We will leave the analysis of this circuit to the reader. As with the case of the integrator, practical differentiators need additional components to keep some properties within bounds. As a hint, bear in mind that the input impedance is a pure capacitance, and that the input current will increase with increasing frequency. Integrators and differentiators can be considered as RC filters, made ideal by the action of the op-amp.

Whereas passive RC filters have a roll-off frequency, above or below which the 6 dB/oct slope flattens off, the op-amp circuits described above maintain their +6 dB (differentiator) or −6 dB (integrator) slope over their whole functional bandwidths. Likewise, the output of a differentiator leads 90° in phase to the input, whereas the output of an integrator lags 90°. Finally, the output voltage may be higher than the input voltage.

We will meet these circuits again, since they play an important part in constructing electrode test and compensation circuits in intracellular and patch-clamp amplifiers.

All former circuits, having the non-inverting input grounded, inverted the input voltage, and so were an inverting amplifier, an inverting adder and so forth. By using the inverting input. The circuit of Fig. 2-21A is surely the most simple op-amp circuit, since it has no external components other than a piece of wire: the output is connected to the inverting input directly. Using the same rule of thumb, the output voltage turns out to be the same as the input voltage. Therefore, it is called a follower or voltage follower. What would be the advantage of an amplifier that does not amplify (left for the reader—remember the high input impedance of op-amps)? Even better still, the intrinsic input impedance of the device (which depends on the type of transistor used) is multiplied by the open-loop gain. Because of this useful feature, the basis of any intracellular microelectrode amplifier is a follower circuit. Figure 2-21B shows

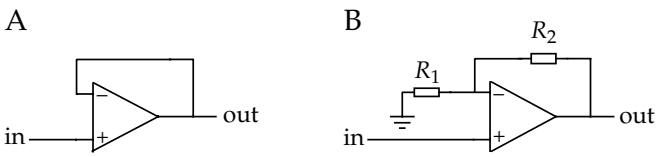


Fig. 2-21 Simple voltage follower (A) and follower with gain (B).

how to add gain to such a non-inverting amplifier. Installing a number of resistors, together with a rotary switch, turns this circuit into a versatile amplifier with a selectable gain factor of, e.g. 1, 2, 5, 10, 20, 50 and 100. Practical forms of the op-amp are shown in Fig. 2-22.

The last op-amp circuit we will discuss is the current-to-voltage converter (CVC): a simple but important tool in electrophysiology. Figure 2-23 shows that the op-amp is fitted with only a feedback resistor, and that the input is directly connected to the virtual ground. This implies that the circuit connected to it is effectively grounded: the input impedance is near-zero, like a true ground connection. Therefore, it is also called a virtual ground circuit. In addition to providing a ground connection, this circuit measures the current to (virtual) ground, and so can be used to ground an electrophysiological preparation and to measure the stimulus current flowing through the preparation. Therefore, a CVC is also part of the famous voltage-clamp amplifier. If a good, high-impedance op-amp is used, feedback resistors as high as 10^{10} or even $10^{11}\ \Omega$ can be used. Under certain precautions, currents as low as 0.1–1 pA can be measured. Therefore, a CVC with a feedback resistor in the above-mentioned range is the heart of any patch-clamp amplifier, which is discussed later on.

This ends our review of op-amp circuits. More complex applications, such as microelectrode preamplifiers, are discussed in the next chapter. The realm of applications of op-amps is virtually unlimited. Using the same electronic building block, the feedback circuit, usually consisting of a mere two to three parts, determines a host of functions that can be performed.

But what would happen if the feedback circuit is left out, so if an op-amp is used open-loop? In this case, the (almost) infinite gain drives the output to the maximum voltage (about the

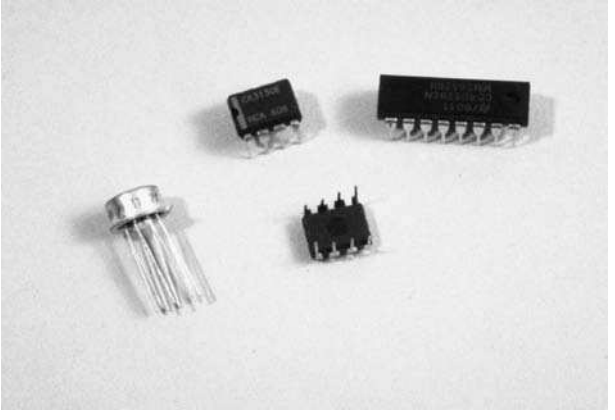


Fig. 2-22 Integrated circuit (IC) op-amps.

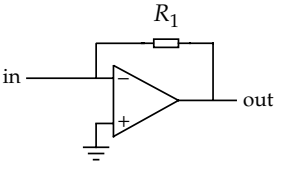


Fig. 2-23 Current-to-voltage converter.

power supply voltage), either the positive or the negative one, depending on minute differences between the input voltages.

Although such a configuration is likely to be unstable, and might cause a lot of problems for the designer, it is nevertheless useful, and called a comparator. The output voltage signals a difference between two voltages with a high precision. If the open-loop gain of such a comparator op-amp is 100 dB, the difference between the voltages at the inverting and the non-inverting input needs only to exceed 100 μV to yield a 10 V output voltage. Thus, if the voltage at the inverting input is at least 100 μV higher than that at the non-inverting input, the output is fully negative; if it is 100 μV lower, the output is fully positive. In fact, special op-amps that sustain stable, open-loop performance are sold explicitly as "comparators".

ELECTRONIC FILTERS

Operational amplifiers are also the ideal building blocks for electronic filters. In many recording situations, one has more filtering needs than the low-pass and high-pass RC filters, dealt with in Chapter 1, can provide. These basic, so-called first-order filters have a filtering slope of 6 dB (high-pass) or −6 dB (low-pass) per octave outside their pass-bands. They can be cascaded to get higher orders. The most simple design would be to combine two low-pass sections into a second-order low-pass (Fig. 2-24). This combinations yields a −12 dB/oct roll-off slope, accompanied by a 180° phase lag asymptote.

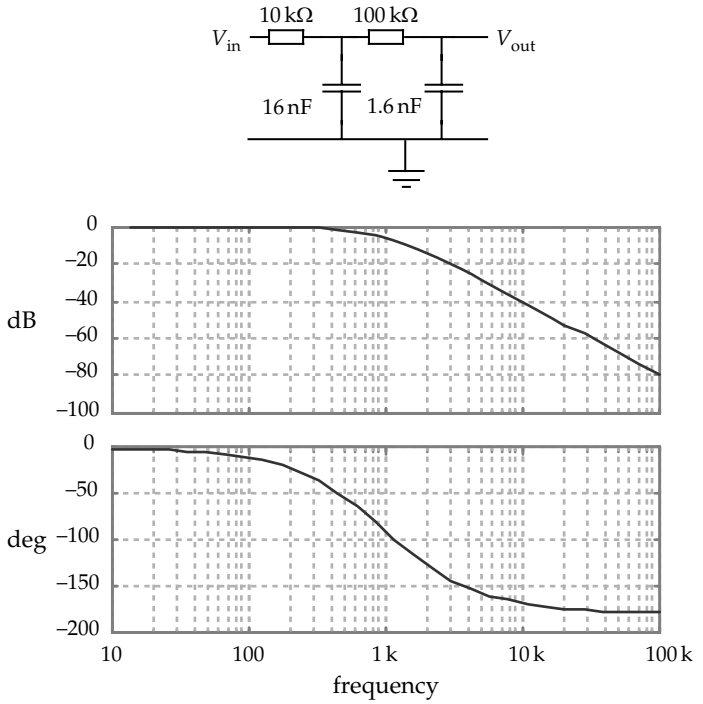


Fig. 2-24 Two low-pass sections cascaded to obtain a second-order low-pass filter. Top: circuit diagram, bottom: frequency characteristics.

This passive solution to a higher-order filter suffers from two serious flaws. In the first place, the two sections are coupled directly, and so their actions are mutually influenced. To minimize this influence, different *R* and *C* values have been chosen for the two sections in the example, but in general, one would need an amplifier between the stages (and preferably in front of the first section, as well as after the last one). However, op-amps have the possibility of using filters as feedback paths. This provides for the necessary separation, and makes the filter characteristics far more flexible. A second-order low-pass with a single op-amp is shown in Fig. 2-25. Frequency and damping (the latter determining the peak in the frequency–response curve) can be varied by adjusting the components marked *f* and *d* respectively.

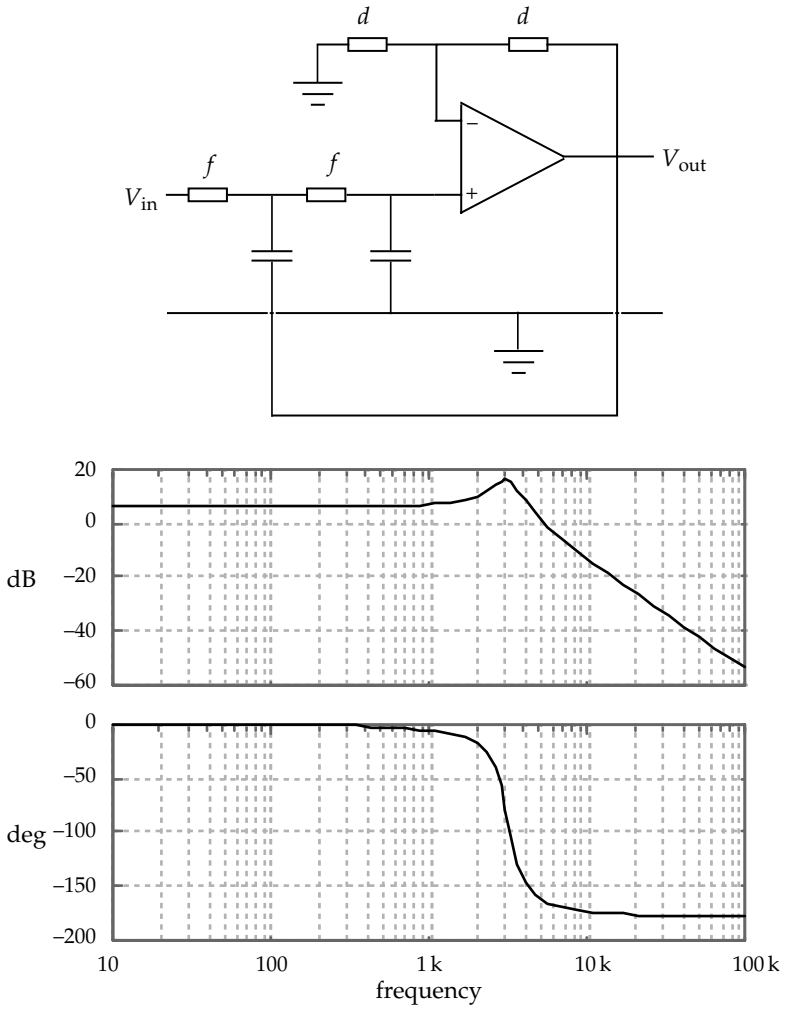


Fig. 2-25 A second-order active low-pass filter with a single op-amp. The frequency is determined by the value C of the capacitors and by the resistors marked *f*; damping by the resistors marked *d*. Top: circuit diagram, bottom: frequency characteristics.

Higher-order filters (4th, 6th order, etc.) are usually built out of second-order sections like the one shown here. For odd orders (3rd, 5th, etc.), one first-order section is added.

By altering the feedback circuits and the relative values of the different components, a number of filter characteristics may be obtained, which differ widely in their frequency behaviour. Some filters have a beautifully flat passband, but a non-linear phase shift, others have ripples in the passband, but cut far steeper beyond the cut-off frequency. In addition, high-pass, band-pass and band-reject configurations can be built. The different versions are often named after their inventors, such as Butterworth and Chebyshev. The table below shows the properties of the most popular ones. The properties printed in bold type are often the reasons to use a particular design (provided the disadvantages can be tolerated).

Type	Pass-band performance	Transition	Stop-band performance	Phase or delay performance
Butterworth	**flattest**	slow	fair	fair, non-linear
Bessel	decreasing	slow	fair	**best (linear)**
Chebyshev I	rippled	fast	monotonic	non-linear
Chebyshev II	decreasing	fast	rippled	non-linear
Elliptic (Cauer)	rippled	**fastest**	rippled	non-linear

ELECTROPHYSIOLOGICAL PREAMPLIFIERS

Although specialized instruments such as microelectrode amplifiers are occasionally still made with separate, or discrete, transistors and other components, most electrophysiological instruments are composed of, or can be represented by, op-amp circuits. Below are a few examples of circuits for the main forms of electrophysiological recording: extracellular, intracellular and patch-clamp. Note that, for flexibility, precision and stability, practical amplifiers can be considerably more complex than the ones shown here.

Amplifier for Extracellular Recording

An amplifier suited for extracellular recording is shown in Fig. 2-26. Here, we need a high gain (a gain factor of 1000 or more), a fairly high input impedance (at least 100 MΩ), the possibility to bar DC voltages ("AC coupling"), and filters to limit the bandwidth. Op-amps A_1 and A_2 form a differential pair with a high input impedance and a high gain, determined by:

$$G = 1 + \frac{2R_2}{R_1}$$

The signal is converted from differential to single-ended by A_3. This circuit is known as instrumentation amplifier. To provide further gain and/or filtering, amplifiers A_4 and A_5 are added. For simplicity, these amps are shown in the simple follower configuration, preventing mutual influencing of the filter sections. In practical designs, the resistors and capacitors indicated with asterisks are sets, selectable with rotary switches, so as to provide flexible gain and bandwidth. For clarity, offset and calibration circuits are also omitted.

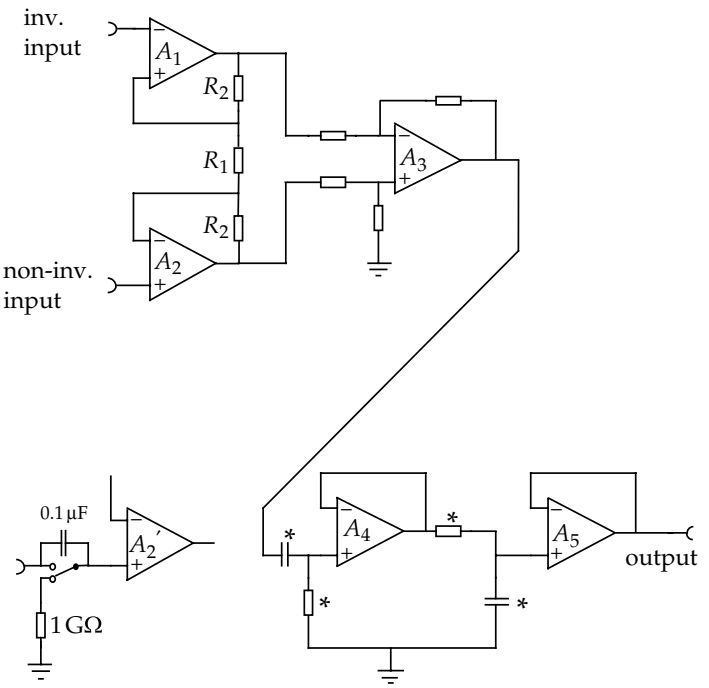


Fig. 2-26 High-gain differential preamplifier.

The alternative input circuit at the lower left (op-amp A_2') is a way to implement AC coupling. It is usually provided at both inputs. The reader is invited to compute the cut-off frequency using the values shown. Unfortunately, such a circuit reduces the common-mode rejection at low frequencies, largely because the components used cannot be matched better than about 1%, and because carbon must be used for resistors in the GΩ range. A very popular type of amplifier from the 1960s on is the EG&G PAR113[1] low-noise differential preamplifier, shown in Fig. 2-27.

Amplifier for Intracellular Recording

A second, important way of recording is by means of an intracellular, glass capillary microelectrode. An amplifier schematic is shown in Fig. 2-28A. Since intracellular voltages are relatively high, we do not need much gain: a factor of 10 suffices. Because intracellular microelectrodes can have impedances of over 100 MΩ, a very high input impedance is mandatory. Therefore, the first op-amp (A_1) is connected as a voltage follower. The second one (A_2) provides some gain (R_2/R_1), together with a DC offset control (potentiometer P_1). Another potentiometer P_2 is used to provide a controllable input capacitance compensation. This is often necessary because at the high impedance level used, the stray capacitances of pipette holder, cables and/or input circuit

[1] Now replaced by model number 5113.

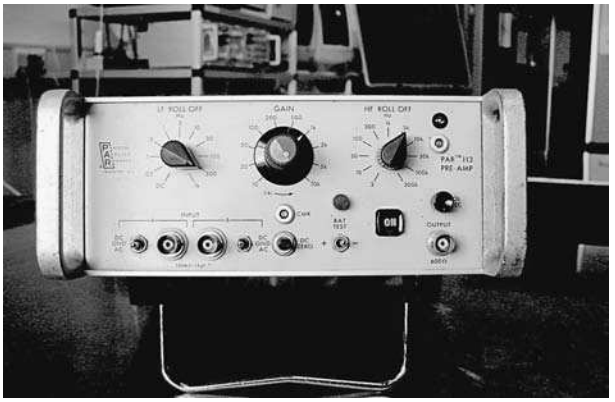


Fig. 2-27 Practical extracellular amplifier: the famous PAR 113 low-noise, differential preamp.

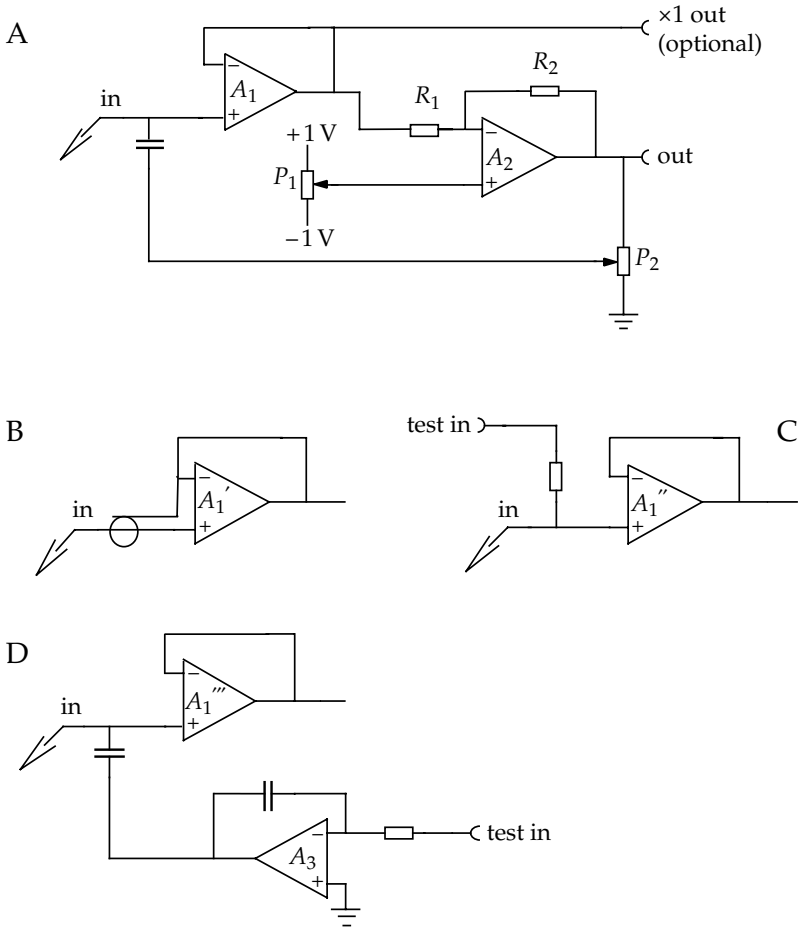


Fig. 2-28 Microelectrode voltage amplifier.

will reduce the bandwidth too much. As an example, a 100 MΩ pipette together with 10 pF stray capacitance reduces the bandwidth to a mere 160 Hz, insufficient to record most nerve and muscle spikes. The capacitance compensation circuit is in fact a form of positive feedback, here for high frequencies, and must be used with care. Overcompensation leads to a very peaked response, ringing, or even to oscillation. An alternative way to reduce the effect of cable capacitance is to drive the shield of the cable with the ×1 output of the follower amp. This technique, known as guarding and shown in Fig. 2-28B, diminishes the bandwidth reduction without the use of positive feedback, and so without the risk of overcompensation. Here, the harmful effect of the cable capacitance is reduced to almost zero because the voltage across that capacitance is reduced. This is because with the shield grounded, the full input voltage charges the cable capacitance, whereas here only the op-amp error voltage, which is far lower, charges the cable capacitance. Guarding is often helpful when the use of relatively long cables is necessary, but does not help to reduce the effects of other stray capacitances such as the capacitance of the pipette shaft and taper. Since glass micropipettes have a variable impedance, which may vary between experiments and also during recording, the amplifier must be fitted with a circuit to measure the electrode impedance in situ. Figure 2-28C and D show two alternative ways to accomplish this. The first circuit (A_1'') uses a high resistor, 1 GΩ or more, to inject a small current into the electrode at the input. When operated with a square wave at the test input, the amplitude of the square signal at the electrode can be converted to read electrode impedance. A disadvantage of this circuit is that the input impedance of the amplifier is reduced to the value of the test resistor. A better way is shown in Fig. 2-28D (op-amp A_1'''), where a small capacitance (about 1 pF or less) is used to feed a test signal in. Usually, because the capacitor differentiates the test signal, and must be fed with a triangle signal, rather than a square signal, an integrator is built in, to convert the test square into a suitable triangle.

Patch-Clamp Amplifier

The last type of amplifier to discuss is the patch-clamp amplifier, illustrated in Fig. 2-29. Because here membrane *currents* rather than voltages are measured, we have a situation different from the voltage amp discussed above.

A patch-clamp amplifier measures the current flow from the electrode (pipette) to the ground. Therefore, the core of a patch amp is a virtual ground, or a CVC. Note that, although a current meter needs to have a low input impedance, the electronics used must have a very high input impedance nevertheless. This is because the currents to be measured are so small: in the range of 1 nA down to less than 1 pA. This also means that stray capacitances tend to spoil the bandwidth again. Therefore, a number of correction and compensation circuits are needed. Fig. 2-29 shows a somewhat simplified circuit diagram of a patch amp suited to measure single-channel as well as whole-cell currents. Op-amp A_1 is the CVC. It needs to have a very high input impedance, yet have excellent low-noise properties.

The feedback resistor R_1 is usually 1−10 GΩ, depending on the range of current strengths to be measured. The non-inverting input is not grounded, but is connected to an input circuit providing the command voltage. Thus, the pipette can be voltage-clamped, usually by step inputs to the command input op-amp (A_4). The difference between the output voltage of A_1 and the command voltage reflects the membrane current. This is accomplished by the difference amplifier A_2. Op-amp A_3 forms a circuit to correct the bandwidth. It has unity gain at low frequencies, but enhances high frequencies. By the choice of the filter components, the high

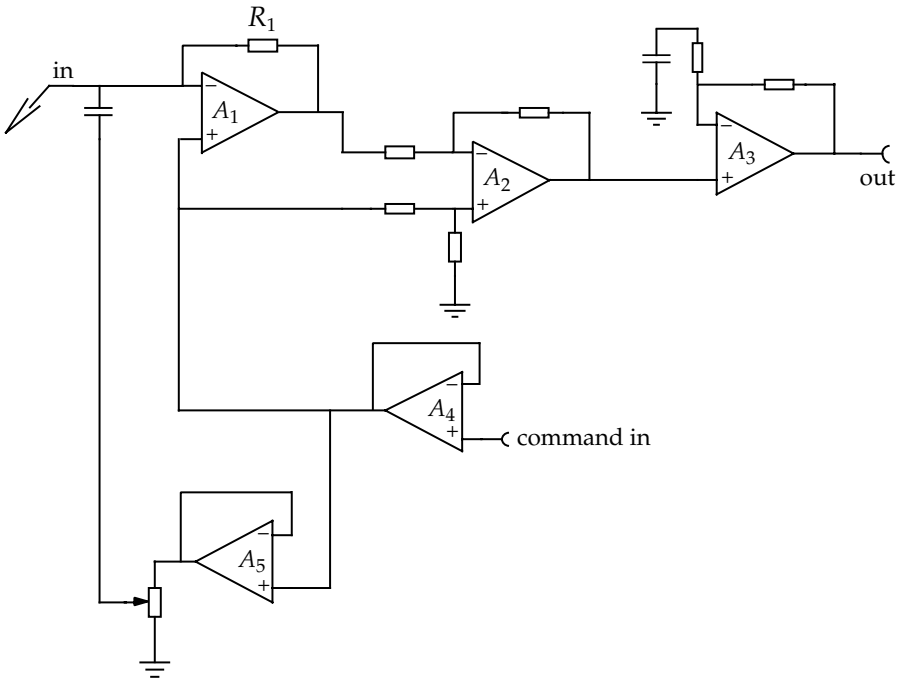


Fig. 2-29 Simplified patch-clamp amplifier.

enhancement can be chosen to compensate for the bandwidth reduction caused by the stray capacitances at the input. A_5 improves the response to command voltage steps by charging the stray capacitances to the new voltage through a small, separate capacitor. In addition, practical patch amps employ compensation circuits for the cell membrane capacitance and for the pipette resistance in the case of whole-cell recording.

A practical patch-clamp amplifier is the Axopatch 200, shown in Fig. 2-30. It has a separate head-stage, which harbours the patch pipette directly (Fig. 2-31).

The advent of integrated circuits (ICs) and the field-effect transistor (FET) technology has greatly helped to reduce instrumental noise in patch-clamp amplifiers by grouping the elements of the first amplification stage (A_1 in Fig. 2-25) including the high-resistance feedback resistor,

Fig. 2-30 Axopatch 200 patch-clamp amplifier.

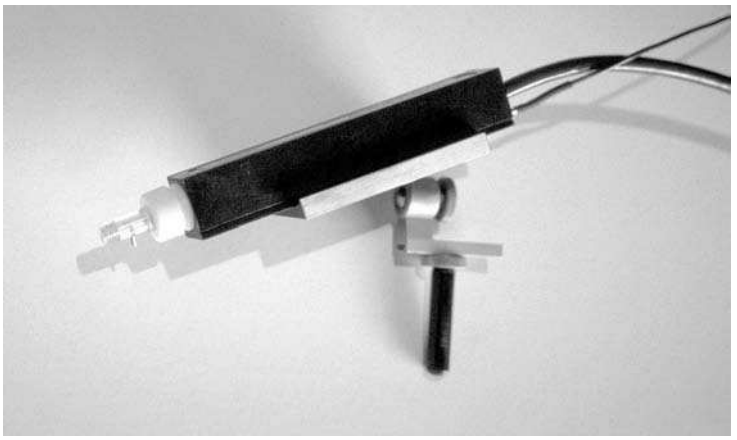


Fig. 2-31 Head stage of the Axopatch 200.

which constitutes the principal source of noise, on a single chip. It has thus been possible to reduce noise power to below $10^{-30}\,A^2/Hz$ at 100 Hz, $10^{-29}\,A^2/Hz$ at 10 kHz, giving a usable bandwidth of approximately 10 KHz under optimal conditions. As was explained in an earlier section, at least part of the noise is due to random thermal movements of charge carriers in the material making up the electronic components (Johnson noise). To avoid the Johnson noise component in the feedback resistor of the first amplification stage, this resistor can be replaced by a capacitor, converting A_1 into an integrator circuit (see Fig. 2-20). A disadvantage of this design is that the feedback capacitor needs to be decharged regularly (usually at a rate of 1 KHz) to prevent saturation of the signal. An additional improvement of the S/N ratio is cooling of the head stage. The combination of these two techniques has led to a 10-fold improvement of S/N ratio and hence of bandwidth.

In practice, an electrophysiology set-up will not contain a separate amplifier dedicated to each of the experimental conditions that were discussed so far. In particular, a set-up for patch-clamping will often have to do with the same patch amplifier to measure intracellular membrane potential. All commercially available voltage-clamp amplifiers have a so-called "current-clamp" option that allows for membrane potential measurements. In current-clamp, the current injected into the cell has to be exactly zero. This goal is achieved if the voltage difference across resistor R_1 in Fig. 2-29 is zero, hence the output of A_1 should be equal to the membrane potential (V_m). By connecting this output (A_1) to the non-inverting input of A_4 (which in voltage-clamp receives the command voltage), this condition is met and the stages A_2 and A_3 will have zero output, indicating zero injected current. The same problems with pipette capacitance, mentioned for microelectrode amplifiers, also apply here. In order to record the membrane potential reliably, the amplifier-input capacitance and the pipette capacitance have to be compensated correctly before breaking the membrane and passing to the whole-cell configuration. Although a single capacitance compensation circuit is shown in Fig. 2-29, usually at least three are provided in order to compensate for input capacitance, pipette capacitance and membrane capacitance. Of course, in current-clamp mode, the membrane capacitance is left uncompensated.

The sequence of events leading to the recording of the membrane potential is thus the following:

1. Fix the pipette in its holder and with the pipette still in the air, compensate for the amplifier input or stray capacitance (typically 1–3 pF) while in voltage-clamp.
2. Go to the "cell-attached" voltage-clamp mode and compensate for the pipette capacitance using a second compensation circuit (typically 3–10 pF).
3. Now break the membrane and switch to current-clamp.

This works fine as long as the compensated capacitances remain constant during an experiment, which is not always the case. If, for example, the level of the bath solution changes by a perfusion system, the pipette capacitance varies, which is especially dangerous if the solution level drops below the starting level. In that case, the pipette capacitance becomes overcompensated, the amplifier ocillates and large currents are injected into the cell.

A second important inconvenience of voltage-clamp amplifiers, if used in current-clamp mode, is related to the high amplification factor of the head stage A_1. As the amplifier is designed primarily to convert very small currents to large potentials, very small errors at the input (e.g. a very small bias current, inherent to all electronic devices) may lead to important voltage-offsets at the inverting input of A_1 (as much as tens of millivolts when recording from small cells). Not all manufacturers of voltage-clamp amplifiers provide adequate nulling of bias currents. Hence, it is often difficult, although not impossible, to measure the membrane potential reliably using a voltage-clamp amplifier.

Two-Electrode Voltage-Clamp Amplifier

The single-electrode voltage-clamp amplifier relies on the fact that the cell or membrane patch has a high resistance (typically 100 MΩ to 10 GΩ) with respect to the pipette resistance (1−10 MΩ) such that the voltage drop across the pipette is relatively small and can be compensated easily. If the cell resistance is low, as for example in the case of *Xenopus* oocytes, this condition is not met. In that case, a two-electrode amplifier is more appropriate. In Fig. 2-32, one electrode is used to measure the membrane potential. The amplified difference (A_2) between command potential and the membrane potential is fed into the cell by the second electrode. Amplifier A_3 monitors the current injected.

Measurement of Membrane Capacitance in Voltage-Clamp

After having recorded responses from a series of cells, the need is sometimes felt to normalize the responses with respect to the size of the cell, especially if the size varies widely fom one cell to another. As the membrane surface area of a cell is proportional to the membrane capacitance, it suffices in general to normalize the data with respect to the membrane capacitance. The latter can be easily measured by applying a voltage step and recording the current in response of this step. This response contains two components: a steady-state component due to the ohmic membrane conductance (I_L in Fig. 2-33) and a capacitive component (I_C). The time-integral of I_C then gives the capacitive charge Q.

$$Q = \Sigma I_c \cdot \Delta t$$

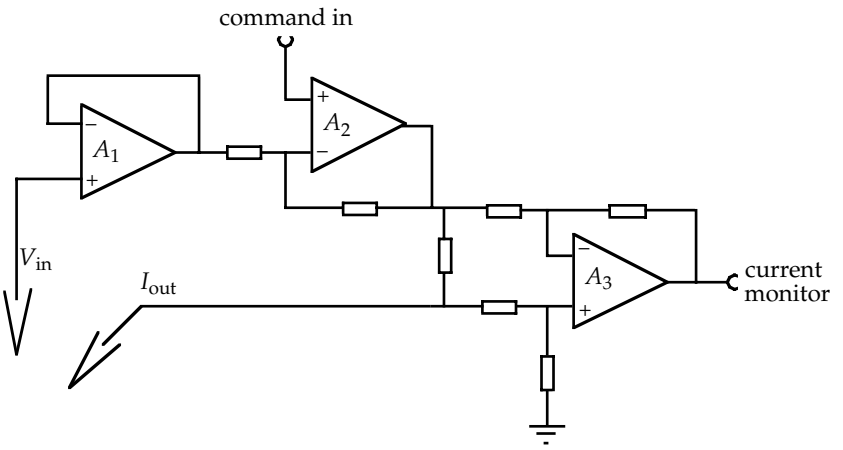


Fig. 2-32 Two-electrode voltage-clamp set-up.

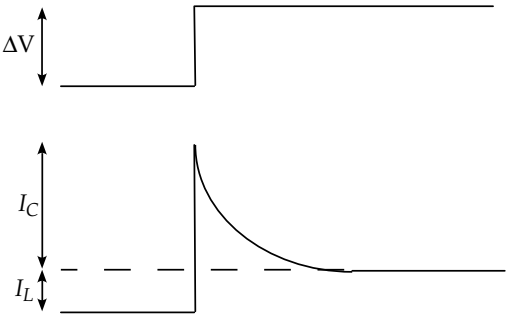


Fig. 2-33 Measurement of membrane capacitance.

By dividing by the voltage step, the membrane capacitance C_m is obtained:

$$C_m = \frac{Q}{\Delta t}$$

Recording of Secretory Events

During secretion, secretory vescicles fuse with the plasma membrane expanding its surface for a while before they are recycled. The changes in membrane surface are accompanied with changes in membrane conductance and membrane capacitance that, in mast cells, are in the order of 10 fF ($1\,\text{fF} = 10^{-15}\,\text{F}$) per event. Although the method described in the previous paragraph can be used to get a rough estimate of the membrane capacitance, it is far too inexact to resolve single secretory events. Instead, methods that use a small amplitude sinewave voltage are often used. These methods rely on the electrical model as shown in Fig. 2-34.

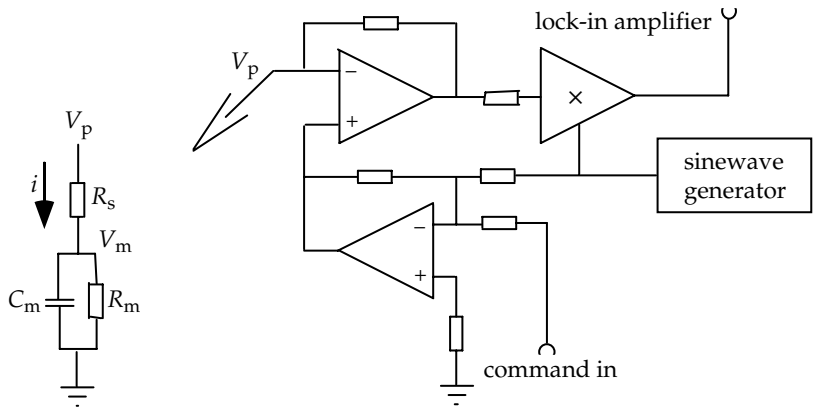


Fig. 2-34 Recording of secretory events.

In this model R_s is the series resistance or the pipette resistance, G_s ($=1/R_s$) the series conductance, R_m the membrane resistance, G_m ($=1/R_m$) the membrane conductance, C_m is the membrane capacitance, V_p is the pipette potential or voltage-clamp potential and V_m the actual membrane potential. The membrane current under voltage-clamp is then:

$$i = V_m \cdot G_m + C_m \frac{dV_m}{dt}$$

This is equal to:

$$i = (V_p - V_m)G_s$$

and hence, $V_m = V_p - \frac{i}{G_s}$

Substitution of V_m in the first equation gives:

$$V_p G_m G_s - i(G_m + G_s) + C_m \frac{d(V_p G_s - i)}{dt} = 0$$

Passing to complex notation with $V_p = e^{-j\omega t}$ and $i = I(\omega)\, e^{-j\omega t}$ gives:

$$G_m G_s e^{j\omega t} - (G_m + G_s)\, I(\omega) e^{-j\omega t} + C_m \frac{d\left(G_s e^{-j\omega t} - I(\omega) e^{-j\omega t}\right)}{dt} = 0$$

And thus:

$$G_m G_s - (G_m + G_s) I(\omega) + j\omega C_m I(\omega) - j\omega G_s G_m = 0$$

Finally splitting $I(\omega)$ into real and imaginary parts: $I(\omega) = A + jB$ gives:

$$A = G_s \frac{G_m^2 + G_m G_s - \omega^2 C_m^2}{(G_m + G_s)^2 + \omega^2 C_m^2}$$

$$B = G_s \frac{\omega G_m G_s + 2\omega C_m}{(G_m + G_s)^2 + \omega^2 C_m^2}$$

The real part, A, vanishes for a certain combination of C_m and ω.

$$A = 0 \quad \text{for:} \quad G_m^2 + G_m G_s = \omega^2 C_m^2$$

Hence, for a certain frequency of stimulation, small variations of the membrane capacitance can be measured without contamination by membrane conductance changes. In practice, the test frequency, ω, is fixed at a value between 1 and 5 kHz and C_m is adjusted by "cheating" with the capacitance compensation of the voltage-clamp amplifier to obtain $A = 0$. During this compensation, the imaginary output must be eliminated. This is done by chopping the voltage-clamp output current with a lock-in amplifier (this is an amplifier that inverts the signal during half the sine period) and averaging over about 10 periods. After compensation, the phase angle between sine wave stimulus and lock-in amplifier input is shifted 90° to measure the imaginary part of the output current that now reflects changes in capacitance (see Neher and Marty, 1982; Lindau and Neher, 1988). The present generation of computer-operated voltage-clamp amplifiers contain membrane capacitance measurement as a standard option.

A second way to measure quantal release from cells electrically is by oxidation of the substance released. Acetylcholine, adrenaline and serotonine are easily oxidized by a carbon electrode held at about 700 mV by a voltage-clamp circuit. To do so, a 5 μM carbon fibre is inserted in a glass or plastic pipette and then sealed with epoxy or by heat (in case of a plastic pipette). Often, after sealing the tip, the carbon fibre is broken off or polished to expose a clean surface. In addition, the tip is soaked in ethanol, butanol or propanol for 15 minute prior to experiment. Then the carbon electrode is brought very close to a cell and clamped at 700 mV. Each quantum released generates a spike of current ranging from 10 to 100 pA, depending on the distance between releasing site and electrode.

Amperometric detection of other substances than catecholamines may be more difficult. For example detection of insulin (by oxidation of sulphur bridges) requires special carbon electrodes on which a film of ruthenium oxide/cyanoruthenate is deposited and is operated at 850 mV. These electrodes degrade rapidly and often other options are sought. One such an option is charging, for example, β-cells (which secrete insulin) with serotonine, which is then co-secreted with insulin. A second option is coculture with a cell type that responds to the secreted substances. Hence, the second cell type serves as the detector.

A third option is to measure secretion optically, using fluorescent dyes such as TMA-DPH (1-(4-trimethylammonium)-6-phenyl-1,3,5-hexatriene). TMA-DPH is a hydrophobic compound that can nevertheless be dissolved in water in micromolar concentrations. When it comes into contact with the cell membrane, it rapidly dissolves in the lipid phase, at the same time becoming fluorescent (excitation 340 nm, emission 430 nm). When the cell is stimulated, secretory granules fuse with the membrane, take up TMA-DPH and are subsequently internalized, making the cell interior fluorescent. Then the extracellular solution is replaced by a solution devoid of TMA-DPH, which also removes the dye from the extracellular leaflet of the plasma membrane, due to partitioning between aqueous and lipid phases. Then the cell may be stimulated again and now intracellular fluorescence decreases upon fusion of the secretory vescicles with the plasma membrane (see Zhou and Misler, 1996; Cox and Kulesza, 1984; Illinger and Kuhry, 1994).

POWER SUPPLIES AND SIGNAL SOURCES

After the foregoing, complicated circuitry, the structure of power supplies is simple and fairly straightforward. In the early days of electronics, huge piles of galvanic cells called batteries were the only sources of DC. Since the invention of diodes, the alternating mains voltage can be converted into direct current, yielding a strong and inexhaustible power source. Today, more elegant forms of battery are still used to power portable instruments, and in addition some electrophysiological instruments use batteries because keeping them disconnected from the mains may help to reduce interference levels. Most apparatus, however, needs mains power to furnish the needs of op-amps and other building blocks: usually +15 and −15 V. Digital circuits such as counters, frequency meters and computers also need a single +5 V power supply. The mains voltage, 230 or 110 V, can be stepped up or down easily to furnish higher or lower voltages as required.

An AC is converted into a DC by a so-called rectifier, which in its simplest form consists of one or more diodes and a storage capacitor. The simplest form is illustrated in Fig. 2-35 (top). The signal passed by a diode is rectified in the literal sense, but the pulsating DC is not yet suitable to power our instruments. Therefore, a large storage capacitor is added, as a water butt that stabilizes the water level of a source which delivers in gusts.

The voltage on the capacitor is a fairly "flat" DC, although it has a ripple caused by the charging of the capacitor once every cycle of the mains voltage, hence at 50 (60) Hz. Therefore, this basic, or half-wave, rectifier is seldom used. A better form is the full-wave rectifier, employing a diode bridge (see Fig. 2-35, bottom). The result is shown in Fig. 2-36:

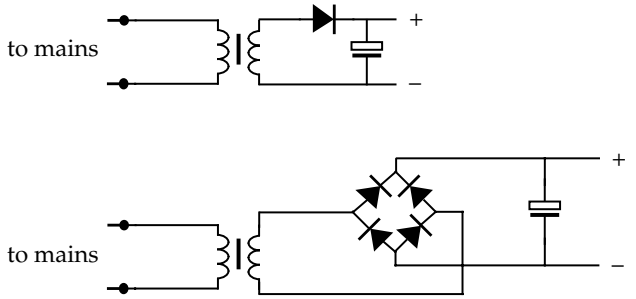


Fig. 2-35 Half-wave and full-wave rectifiers.

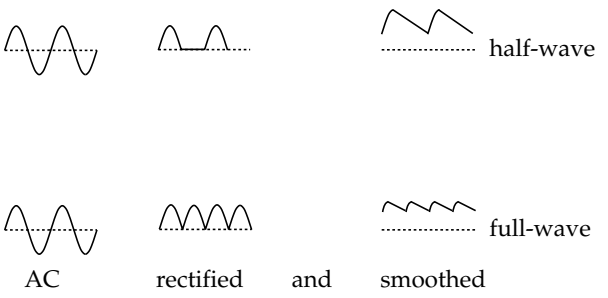


Fig. 2-36 Ripple of rectifier circuits.

by sending both the positive and the (inverted) negative peaks to the capacitor, the ripple is substantially lower, since the capacitor is now charged twice each mains cycle, hence at 100 (120) Hz.

To reduce the ripple further, and to make the output voltage independent on mains voltage fluctuations, most instruments are fitted with a stabilized power supply, in which electronic circuitry (in fact, a kind of powerful op-amp) is added to flatten the voltage further. If a 100 (120) Hz hum signal is found in an electrophysiological set-up, it will most probably stem from an ill-dimensioned power supply.

Apart from built-in power supplies, separate devices to deliver direct current, also called power supply, are sold, usually with two or three adjustable voltages (see Fig. 2-37). In the laboratory, they are used not only to deliver power to operate special or home-made circuits, but also to administer DC stimuli, such as (the DC component of) the command voltage to a voltage-clamp amplifier.

In addition to direct currents, electrophysiologists need several sources to deliver test and measurement signals, such as sine and square waves, pulses and ramp or sawtooth-shaped voltages. The instruments used are therefore called generator, so sine generator, pulse generator and so on. A sine wave can be generated by an oscillator, which is the electronical equivalent of a pendulum clock. Oscillators are also used in radio transmitters and other signal sources, but for frequencies lower than about 1 Hz, the large capacitors needed become unpractical. A more versatile way to generate low-frequency sine and square signals is by a device called a function generator, named so because it generates three "functions": sine, square and triangle, simultaneously. The mode of operation is illustrated in Fig. 2-38.

The core of the circuit is an op-amp integrator, fed from a so-called comparator circuit. If the comparator output is high (say 10 V), the integrating capacitor is charged, yielding a negatively sloping voltage. At a certain threshold, the comparator output flips to negative

Fig. 2-37 A dual, adjustable laboratory power supply.

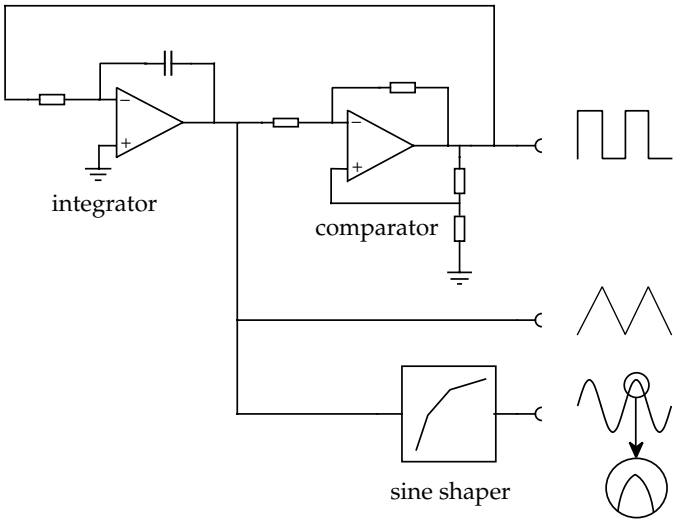


Fig. 2-38 Schematics of a function generator.

(−10 V), so that the capacitor is charged in the opposite direction. Note the positive feedback used in this circuit to get the wanted slack, or hysteresis, between the two "flipping" levels. In this way, a square and a triangle (linearly up- and down-going voltage) are generated simultaneously.

However, the signal we need most is a sine wave. To this end, the triangle is fed through an ingenious circuit consisting of a ladder of diodes and resistors that "bends" the triangle until it approximates a sine wave. Unfortunately, this synthesized function sine is less precise than the intrinsic sine generated by an oscillator circuit, and this can be a nuisance, especially when the sine is differentiated somewhere in an electrophysiological measurement circuit. In that case, small peaks may show up at the positive and negative peaks of the sine signal. Nevertheless, the advantages of the function generator prevail: by using large capacitors and high resistances, very low frequencies can be generated—often as low as 0.001 Hz. In addition, the frequency may be controlled by feeding extra current into the integrator part. In this way, one has what is called the voltage-controlled oscillator, or VCO mode (using the word oscillator in a broader sense). Most of these tricks are built into the type shown in Fig. 2-39.

Pulse generators also use capacitors, resistors and op-amps to generate pulses of adjustable polarity, amplitude and duration. The most important characteristics of sine and square signals: period, amplitude, peak-to-peak amplitude and RMS amplitude are illustrated in Fig. 2-40, together with the characteristics of pulses. Here, T indicates the period or the interval time, A the amplitude and R the RMS value. The peak-to-peak value is $2A$. In a pulse series, or pulse train, the ratio of on-time to total time is called the duty cycle, and is usually expressed in percentages. In the example, the duty cycle is about 25%. Needless to say that a train of action potentials is a kind of pulse signal that has variable interval times, called spike interval or, more precise, interspike interval times (the duration of an action potential may be called the intraspike interval). A square signal can be considered a bipolar pulse of equal positive and negative durations.

Fig. 2-39 A practical form of function generator. The left knob and switches control waveform and frequency, the rightmost knobs control the amplitude and DC offset. The middle controls are for creating sweeps, asymmetric waveforms, etc.

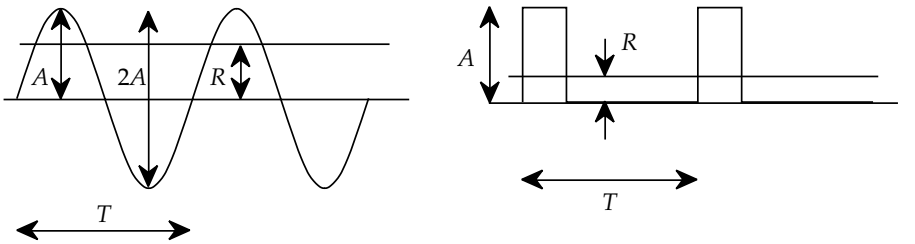


Fig. 2-40 Properties of sine and pulse signals.

A further characteristic of sine signals is the effective value or root-mean-square value (RMS). The mains voltage of 230 V is one such RMS value, since the peak voltage in that case is about 325 V. The reason is that the power (work) delivered by an alternating current is proportional to the square of the mean value. Hence, a 325 V peak sine wave generates an amount of heat (work, etc.) that corresponds with a direct current of 230 V, the root of this mean square. For a sine wave, the peak value is $\sqrt{2}$ times the RMS value. In the case of pulse signals, the RMS value depends on the duty cycle. If a pulse train has a very low duty cycle, the RMS value is far lower than the peak value. The ratio of peak to RMS value is called the crest factor. This is an important specification of instruments that have the read-out calibrated in RMS value, since to measure the proper RMS voltage, the much higher peak voltage of short pulses must be passed undistorted.

The output impedance of power supplies is usually very low, 0.1 Ω or less, as this is necessary to deliver power at a virtually constant output voltage. A power supply can be used as a DC source, provided it is not short-circuited unintentionally during wiring-up. Most other signal generators have a higher output impedance: usually 50 Ω (to comply with the so-called 50 Ω telecommunication impedance standard). Some older apparatus may still use the older standard of 600 Ω. Here one needs to be careful, because if many instruments are connected (audio amps, computers, pen recorders, long cables, etc.), the output voltage might decrease slightly. At the other end, some function generators may drive an 8 Ω loudspeaker directly.

ELECTRONIC VOLTMETERS

The passive voltmeter, based on the "microammeter", or moving-coil meter, described in Chapter 1, has a rather low input impedance, and a rather low sensitivity. Therefore, electronic voltmeters and AVO meters are built that consist, essentially, of a microammeter with a preamplifier. In this case, the input impedance for voltage measurements can be sufficiently high, for sufficiently low current measurements. A modern, electronic voltmeter has usually a digital read-out, turning it into a digital voltmeter, or DVM, which is convenient and reduces the chance of misreading values. The principal components of a DVM are shown in Fig. 2-41.

The input "conditioners" include input attenuators and shunt resistors, rectifier and current source, necessary to measure resistance, DC and AC voltage and current. For more explanation see digital techniques, below. For the moment, it is important to note that DVMs may be made more precise than the ones employing a moving-coil meter, which has a precision hardly better than 2.5%. Digital meters can be designed to have a precision of 0.05% or even better, and may show up to 5 or 6 significant digits. Needless to say that the most precise types are the most expensive ones, and that the most common types of meter are definitely not better than 0.5–1% error tolerance. In addition, precision instruments must be recalibrated regularly, say on a one-year basis, to warrant their high precision and, since this is often neglected in laboratory practice, the practical accuracy may be lower than the nominal one. In addition, the stated precision only holds for the DC ranges. Thus, the ohms and AC ranges are less precise than the stated overall precision.

Measuring AC signals has another caveat: voltmeters that measure RMS value must pass the peak amplitude without getting saturated, or else the pulse amplitude will be underestimated. Thus, so-called true RMS voltmeters intended to measure pulse trains must have a high crest factor (see "properties of pulses" above). However, the most common voltmeters do not measure true RMS value, but attenuate the measured AC voltage by a factor of $\sqrt{2}$, so that the indicated value *is calibrated only for sine signals*.

Electrometers

A special type of voltmeter called electrometer is essentially an AVO meter with an extremely high input impedance (or resistance). This is accomplished by a MOSFET input stage. By the

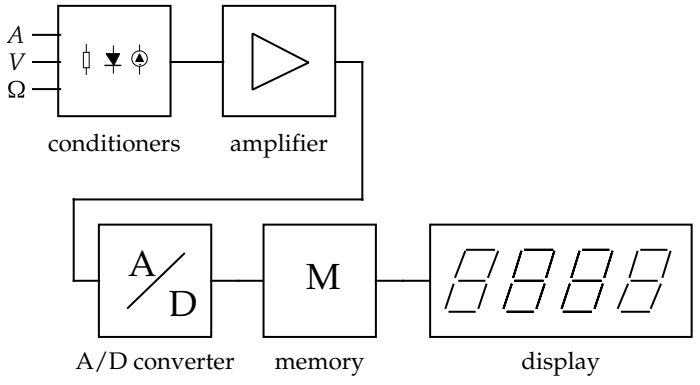


Fig. 2-41 Block diagram of a digital voltmeter.

use of built-in shunt resistors and current sources, an electrometer can be used to measure voltages from about 100 μV to 10 V, currents from about 10^{-14} to 10^{-5} A, and resistances of 10^5 up to 10^{14} Ω. Electrometers are indispensable for the testing of amplifiers, electrodes, etc. In fact, an electrometer can be used as a high-impedance preamplifier for electrophysiological recording. In fact, a very high input impedance is always beneficial.

The reason MOSFETs are not used everywhere is that these transistors produce more noise than j-FETs. For certain measurements, however, MOSFETs are the only way. Apart from the use as a separate test instrument, electrometers are used in conjunction with pH electrodes, or ion-selective electrodes in general. This is necessary because these electrodes use a membrane of glass or a liquid ion exchanger as ion-sensitive part, and have impedances of 10^9 up to 10^{12} Ω. At such high electrode impedances, the use of guarding is mandatory. This holds especially for the intracellular versions.

THE CATHODE RAY OSCILLOSCOPE

No doubt, the cathode ray oscilloscope (CRO), or oscilloscope (or even scope for short), is the instrument used most in all branches of physical science and technology. Obeying the simple principle of displaying one or more voltages against time, this is the most versatile tool for the monitoring and measurement of electrophysiological signals. In fact, the oscilloscope has turned into a kind of "sixth sense", or extension of our own senses. All kinds of signals, sinusoids as well as square or more complex waveforms, can be analysed easily if the frequency is at least about 10 to 20 Hz. Pulses such as nerve spikes, and to a certain extent irregular spike trains may also be observed on the scope. General purpose oscilloscopes have a bandwidth of at least 20 MHz, sufficient for most electrical signs of life. Very low frequencies, on the other hand, are better recorded on a chart recorder.

The core of an oscilloscope is a kind of picture tube, resembling the familiar one used in TV sets and computer monitors. In the neck of the tube, a fine, pencil-shaped beam of electrons, a cathode ray, is generated by an electrode system resembling that of a triode. Indeed, again electrons are emitted by a heated cathode. The grid, here called the Wehnelt cylinder, is used to modulate the amount of current in the beam, and hence the beam intensity. Instead of one anode, an array of electrodes carrying different voltages acts as an electron lens, focussing the beam onto the screen. The complete electrode array is known as the electron gun. If the beam would be merely projecting onto the screen, a fine spot would appear in the middle of the screen. To draw the kind of trace we want, the beam is deflected by two pairs of extra electrodes, the deflection plates. This is depicted in Fig. 2-42. One pair deflects the beam in

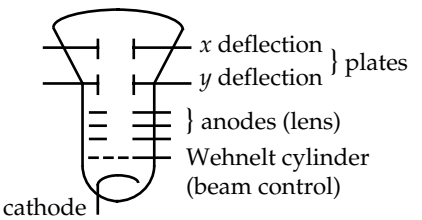


Fig. 2-42 Schematic symbol of cathode ray tube.

the vertical direction, and is fed with the voltage to be measured (the signal). The second pair deflects the beam in the horizontal direction.

Feeding the horizontal pair by a sawtooth signal takes care of writing the signal as a function of time. Therefore, the horizontal deflection circuit is called the time base. This differs from the mode of operation of a TV screen, where the beam is deflected from outside the tube by strong electromagnets, which write a continuous raster of parallel lines, where the TV signal consists of the intensity modulations to generate an image on the screen. In an oscilloscope, the beam intensity is usually constant throughout the writing of the trace, and is suppressed during the time the beam returns to its starting position. Synchronization is also different: in a TV set, the drawn lines are "broken off" at the point where the TV camera sends its scanning spot back. This is called a line synchronization signal, line sync for short. A second sync signal (field or frame sync) sends the spot back to the top of the screen after it has written the last, or the lowest, line. When no TV station is received, both horizontal and vertical deflection signals are running in their own pace. They are said to be free-running (see Appendix D).

In an oscilloscope, a stable image is obtained in a different way, called triggering: after having written a signal trace, the spot is sent back and will wait at the far left of the trace until a starting signal, or trigger, is generated. The trigger signal may be derived from the vertical signal (e.g. a certain voltage level) or from a separate, or external, trigger signal. For instance, when we stimulate the eye of an animal with light flashes, we want to record an electrical signal from the start of the flash. When showing sinusoidal signals, the positive zero-crossing is often converted into a trigger signal.

The complete block diagram of a simple oscilloscope is shown in Fig. 2-43. A practical oscilloscope is shown in Fig. 2-44.

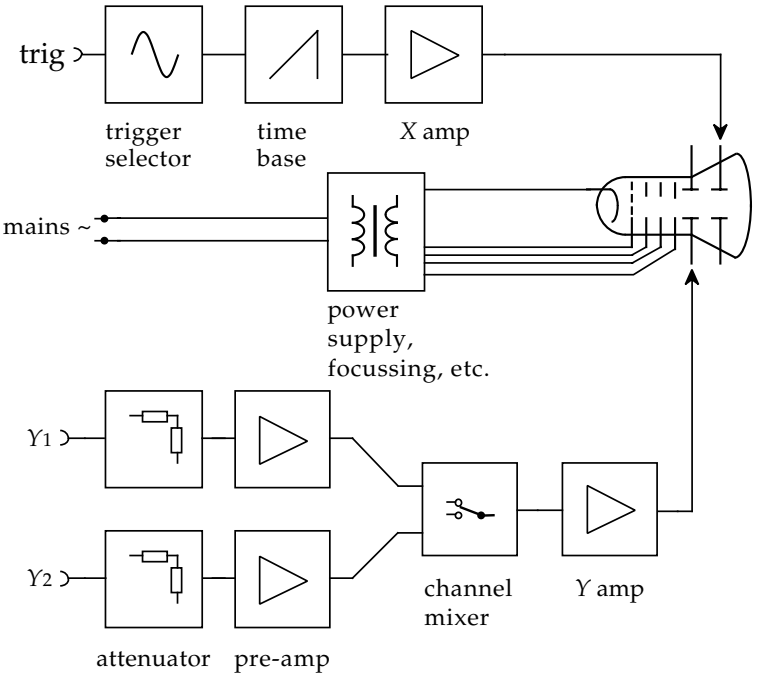


Fig. 2-43 Block diagram of cathode ray oscilloscope.

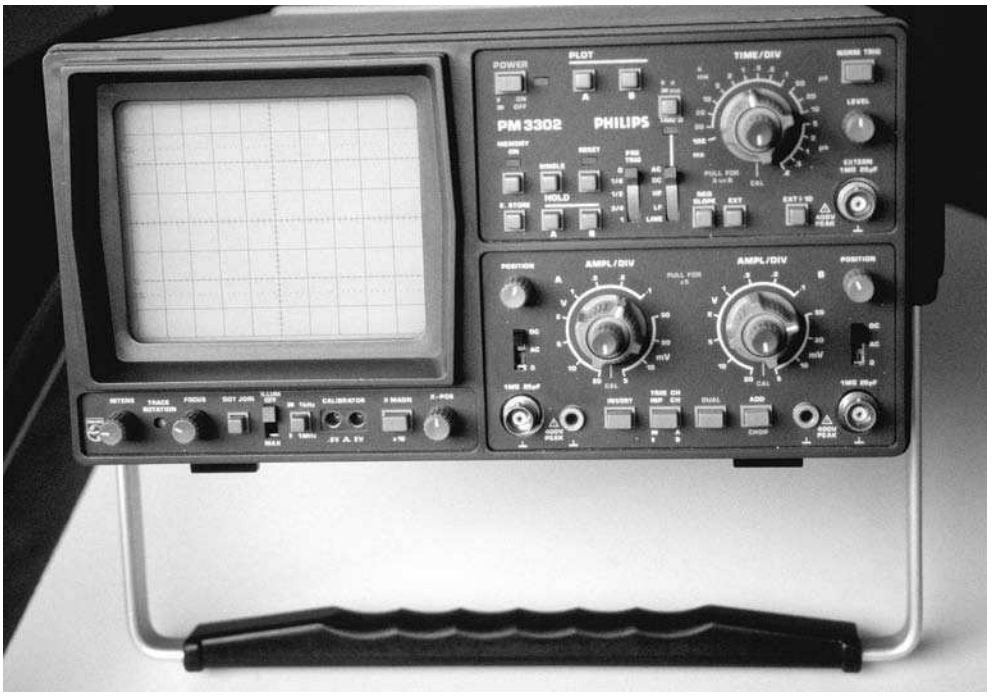


Fig. 2-44 A conventional two-channel oscilloscope (CRT display).

LCD Screen Oscilloscopes

Liquid crystal display (LCDs) screens are increasingly popular today, so the advent of small, portable "pocket oscilloscopes" with an LCD screen may come as no surprise. These little helpers are usually small, handy and battery-powered, and so may be carried easily to do any job (see Fig. 2-45). In addition, they are available at very affordable prices. However, most types have a single Y/T trace, as well as a rather limited resolution, both in time (about 25 samples per division) and in voltage (8 or even 7 bits vertical resolution). Their ease of use has earned them a place in most labs, albeit in addition to the full-blown, higher-resolution, dual-channel CRT scopes.

Important Properties of Oscilloscopes

Since the oscilloscope is a kind of "sixth sense", or sensory "extension piece" to people studying electrical phenomena, it is of paramount importance to the validity of measurements and judgements of signals to have a thorough knowledge of its mode of operation and performance.

The most fundamental properties follow directly from the principles discussed above. For the vertical deflection, scopes are fitted with amplifiers, and so obey all laws we discussed there. For many branches of physics, the bandwidth is of crucial importance to the success of the measurements, and with ingenious techniques, bandwidth of more than 1 GHz has been

Fig. 2-45 A portable oscilloscope with LCD screen.

obtained. Thus, even people working on television and radar circuits can monitor their signals on an oscilloscope.

It will be obvious that for electrophysiological work, the bandwidth must also be sufficient to allow the fastest acts of life, action potentials, channel openings and so on, to be monitored. With modern, even modestly priced, oscilloscopes, this condition will, however, be always fulfilled. Most commercial scopes are designed for radio and television repair technicians, and have usually a bandwidth of 10–20 MHz, amply sufficient for electrophysiological signals, at least for the signals that succeed to pass our preamps. A second "vertical" characteristic is the sensitivity, usually expressed in volt per scale division (one division being usually 1 cm, or one-half inch). By virtue of our use of preamplifiers, we do not need extreme demands on sensitivity: usually 5 or 10 mV/cm suffices.

The usual horizontal signal is the time base, the speed of which is expressed in (m)s/division. Most oscilloscopes support about 0.2 s/division up to less than 1 μs/division. Note that if a low-frequency signal, such as a spike train, is stretched by a fast time base setting, for instance to analyse the rising edge of the spike, the display may get very dim, or barely visibly if at all.

Note that most oscilloscopes support uncalibrated intermediate settings by providing potentiometers for all major adjustments. This can cause grossly erroneous readings, and so when using an oscilloscope to make measurements of amplitudes, delay times and so on, it is important to keep the relevant setting in the calibrated position. This is usually either the fully clockwise or the fully anticlockwise position, where the button "snaps in".

Often, one is interested in synchrony, shape or any other comparison of two rather than one (vertical) signal. Therefore, many oscilloscopes support dual-beam or dual-trace operation. The former, rather expensive, principle employs two electron guns and two sets of

deflection plates in the cathode ray tube, so that two signals can be observed truly synchronous. Most present-day scopes employ the dual trace mode, in which the two traces are generated by a "soft" principle in a normal picture tube. The trick is to switch, or alternate, the single trace between two different input signals, thus fooling the eye to see two synchronous traces.

This can be done in two modes, designated as alternate and chop. In alternating traces, first a whole trace is written using the first input channel. At the end of the trace, or during flyback of the spot, the deflection plates are connected to the second input signal, so that the next trace depicts the other signal. When this happens fast enough, our eyes see two simultaneous traces, so that we can make the necessary comparisons. At too-long time base settings, however, the display "breaks down" into the component traces, so that we can no longer fuse the images. It must be noted that this mode is only valid for perfectly periodical signals, because one period of the first waveform is compared with the next period of the second waveform. This means that two spike trains cannot usually be compared in this way. In all cases where slow or irregular signals are studied, the traces must be chopped. With this mode, the deflection plates are alternating between the two input signals with a very high frequency, usually over 100 kHz. In this mode, the two traces consist of dots that are alternately derived from first and second input signal, again giving a strong dual-trace sensation. In this case, however, the alternation will become visible at very short time-base settings, but by continuously fluctuating the chopping frequency, this can be kept at an unobtrusive level.

Some dual-trace oscilloscopes permit the connection of one of the input signals to the horizontal deflection circuit, thus creating an X/Y display. For distinction, the normal mode of operation is called Y/T display. The X/Y mode is useful for the comparison of sinusoidal signals, and gives rise, then, to the well-known Lissajous figures. These are used for phase measurements, mostly in communication and colour television circuits, where the colour information is coded in the phase of two carrier signals.

With the restrictions mentioned, the dual-trace oscilloscope is the standard working horse of electrophysiology, where almost always two concurrent signals are studied: stepped voltages and the response of a membrane patch, stimulus and the response of a sense organ or synapse, two signals from adjacent sense organs or of different pairs of EEG electrodes on the skull, a spike train and a signal derived from it and so on, so forth. From this list it will be clear that, occasionally, more than two signals will have to be compared. To this end, the dual-trace principle has been extended to combine, or multiplex, three or more input signals, yielding four-trace oscilloscopes, and so on. Since our ability to analyse several things at once is limited, the task of comparing multiple traces must often be performed on recorded, rather than displayed, signals, and is nowadays mostly left to computers. A simple way to capture transient signals is with a storage oscilloscope, in which the trace written on the CRT screen, causing a positive "trough" on the inner side of the fluorescent layer, is made more long-lasting by an accessory electron gun near the screen. Although the storage scope is still in use in some labs, it is superseded mostly by the digital oscilloscope, in which the sampled waveform is stored in a digital memory and displayed repetitively on the screen. In addition, digital scopes are built that can perform all sorts of processing of the signals, such as averaging, amplitude histogram construction or Fourier transformation. In fact, such a digital scope is a form of special-purpose computer designed to store and analyse electrical signals.

Obviously, computers are far more powerful and versatile to store, measure and analyse signals, and so the virtues of the oscilloscope lie more in instantaneous monitoring and control of the experiment, excellent to detect the presence and shape of the wanted signal, and so

to assess the validity of an experiment. Note that measurements with the oscilloscope are of rather limited accuracy. Partly inherent in the working principle—a flying spot on a rather small screen—partly by practical design parameters, the accuracy is seldom better than, say, 5%. Especially the read-out parallax of scopes fitted with external graticules are bound to cause erroneous readings, which might exceed 10% of inaccuracy. Thus, more precise AC or DC voltage measurements must be performed with a digital voltmeter, whereas precision analysis of more complicated waveforms must be left to computers having ADCs of the desired precision.

DIGITAL ELECTRONICS, LOGIC

Contrary to what many people think, digital techniques did not originate in electronics. In fact, we perform "digital" operations, which means "counting on your fingers", from ancient times on. Today, the word "digital" is used to indicate any quantity, operation or measurement that involves a finite number of states. It is used as antonym of analogue (US analog), which by definition means any quantity that can vary continuously, possibly between certain limits.

These names might be a bit confusing, but can be explained historically. Electrical signals are often used to represent some other physical or chemical quantity. For example, the petrol gauge in a car consists of a potentiometer connected to a float. If the petrol level rises, so does the voltage, and if it falls, the voltage falls. Since the electrical signal from the gauge mimics the petrol level, it is called an "analogue" of the petrol level. As a contrast, all things that can be counted are called digital. For instance, the car's odometer indicates the mileage in digits, where values intermediate between two numbers are either absent or unintelligible. The odometer may be called a digital distance meter, in contrast to such things as rulers and tape measures, where in principle any position can be read off.

In fact the distinction is merely a practical one, since with sufficient magnification, most quantities we consider to be continuous prove to be quantized, or "grainy". As an example, the length and weight of things are considered analogue quantities, although there is a formal quantization in the length or weight of single atoms.

By the same principle, electrical quantities such as charge and current (charge per time) are quantized by the elementary charge, being about 1.6×10^{-19} C. Indeed, there are cases in which one has to take quanta in account, such as with the reception of light by photoreceptors, but in everyday life, this graininess may be neglected, and voltage, current and so on are treated as analogue quantities. Digital signals, such as the symbols language is made of, date also from prehistoric times, but in dealing with digital electrical signals, we will use the early, electromechanical telephone exchange as a starting point. This venerable circuit, still in use today, employs so-called relays, having ten positions that could be reached by feeding it with one to ten pulses, generated by the familiar telephone dial. By dialling more than one digit (decimal digit), we can theoretically reach a finite, but arbitrarily large number of people. In the mean time, back in the 1940s, researchers working on prototypes of the digital computer found that binary relays, being either on or off, were more versatile as switching element. With the advance of electronics, vacuum tubes and transistors have replaced the large, slow and noisy relays.

To analyse the functioning of digital circuits, we may observe that making the signal binary is the alternative solution to the problem we encountered with transistor amplifiers, namely

that individual components have different, and fluctuating properties. In Chapter 2, we learned that by increasing the gain of an amplifier and by using feedback, we could surmount this problem. Making the signal digital is another way to make reliable circuits with, so to speak, unreliable components. The simplest digital system is binary, i.e. it uses only two alternative states, usually called on and off (electrical), or true and false (logical), or zero and one (number system). These circuits use two voltages, often zero and +5 V, to reflect these two conditions. Contrary to the voltages in amplifiers and comparable transistor circuits, these two voltages are the only states that are allowed.

Fig. 2-46 shows a simple, so-called digital logic circuit. It is easy to analyse the functioning: if the input voltage is zero, the transistor is shut off, so that the output voltage is virtually equal to the power supply voltage, which in this case is +5 V.

If, on the other hand, the input signal is +5 V (a logical "one"), the transistor conducts, pulling the output voltage to (almost) zero. Indeed, by the intrinsic properties of transistors, a small voltage is left (about 0.5 V); therefore the operation of binary circuits is defined more precisely. All voltages of +0.8 V or less are taken to signify the logical "zero", whereas voltages in excess of +2.4 V are taken as logical one. The range in between is the "forbidden zone". Properly designed digital circuits will only pass this range as quickly as possible (i.e. in a few nanoseconds!), and will never "linger" longer than a fraction of a microsecond in it. Within these specifications, our little circuit of Fig. 2-46 functions as a so-called inverter: the logical number is inverted, which means turned into its opponent state (true becomes false, and vice versa).

For such logic circuits, it is customary to describe the function in a so-called truth table, expressed either in logical terms or in the form of numbers:

hence:

input	output
true	false
false	true

or:

input	output
1	0
0	1

The functioning may be summarized still shorter by a formula:

output = NOT input

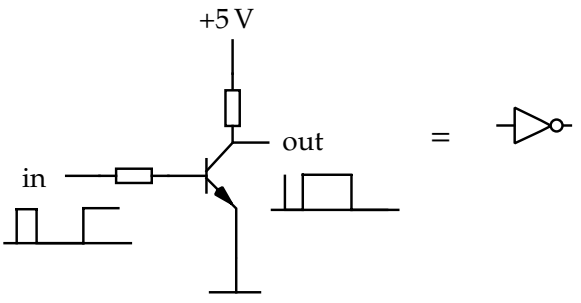


Fig. 2-46 Simple logic circuit (inverter).

Note that both formula and truth table are exhausting (i.e. sufficient) descriptions of the inverter. Note further that the circuit functions independent of the precise gain factor of the used transistor, provided the gain is higher than a certain minimum value.

Although the inverter is the easiest to understand, and is very often used in digital circuitry, its function as such is not very interesting, so let us examine other members of the logic "family". Many logical operations are performed with so-called gates, which can be described as circuits in which the flow of information in one branch is influenced by other branches. The simplest gates have two inputs and one output (the usefulness of more outputs will be dealt with later).

A gate can be made by the addition of one resistor to the inverter of Fig. 2-46: see Fig. 2-47.

The truth table is as follows:

in1	in2	out
0	0	1
1	0	0
0	1	0
1	1	0

In words, the output is "0" if either input, or both inputs, are "1". In the jargon of mathematical logic, the corresponding formula is:

$$\text{out} = \text{NOT} \quad (\text{in1 OR in2})$$

Therefore, this circuit is called a NOT OR gate, or NOR gate for short. By adding the inverter, or NOT circuit, of Fig. 2-46, the NOR is converted into an OR gate. The plus sign in the symbol indicates the OR operation (which is somewhat similar to the algebraic addition). In the same way, AND (and NAND gates can be constructed, where the output is one (zero) only if both

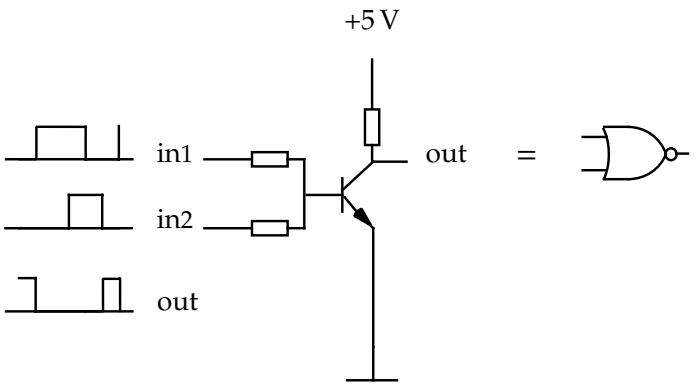


Fig. 2-47 Two-input NOR gate.

inputs are one. A further, important type is called "exclusive or", abbreviated XOR. The truth table is given below left:

XOR			XNOR		
in1	in2	out	in1	in2	out
0	0	0	0	0	1
1	0	1	1	0	0
0	1	1	0	1	0
1	1	0	1	1	1

In words, the output is one only if the two inputs are different. The negated form, exclusive nor or XNOR, is shown at right. Here, the output is one only if the inputs are equal. This circuit is therefore called binary comparator and plays an important role in digital pattern recognition.

In drawing logic circuits, one is interested in the function, rather than constructive details of the circuit. Therefore, new symbols were defined to signify the function of gates, inverters and so on. Examples are given in Fig. 2-48.

In addition to gates, a circuit called flip-flop plays an important part in digital electronics. A flip-flop, officially called a *bistable multivibrator*, is a simple though powerful circuit, built basically with two transistors. This is shown in Fig. 2-49A. In this example, the first transistor is shown in the conducting, or saturated, state, which causes the second transistor to be shut off. This in turn tends to keep the first transistor saturating, so that this condition will last

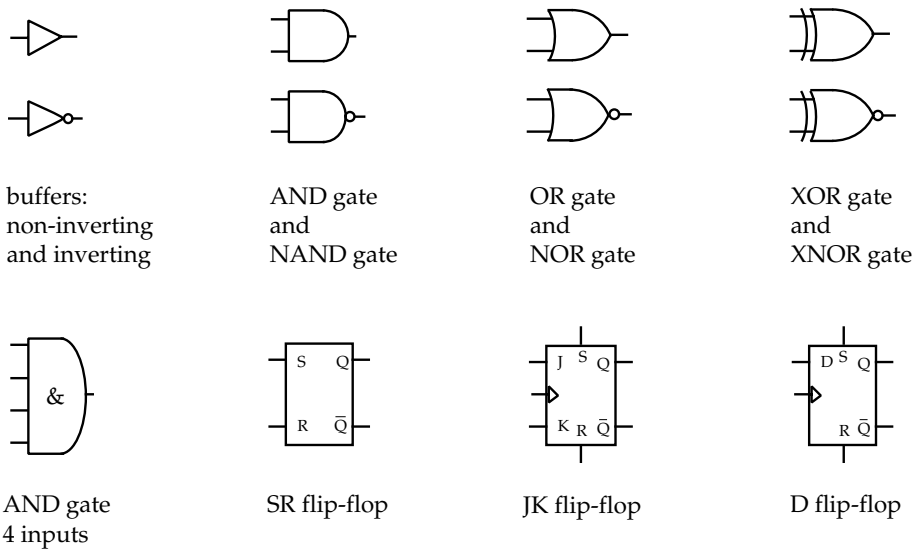


Fig. 2-48 Symbols for gates and other logic components.

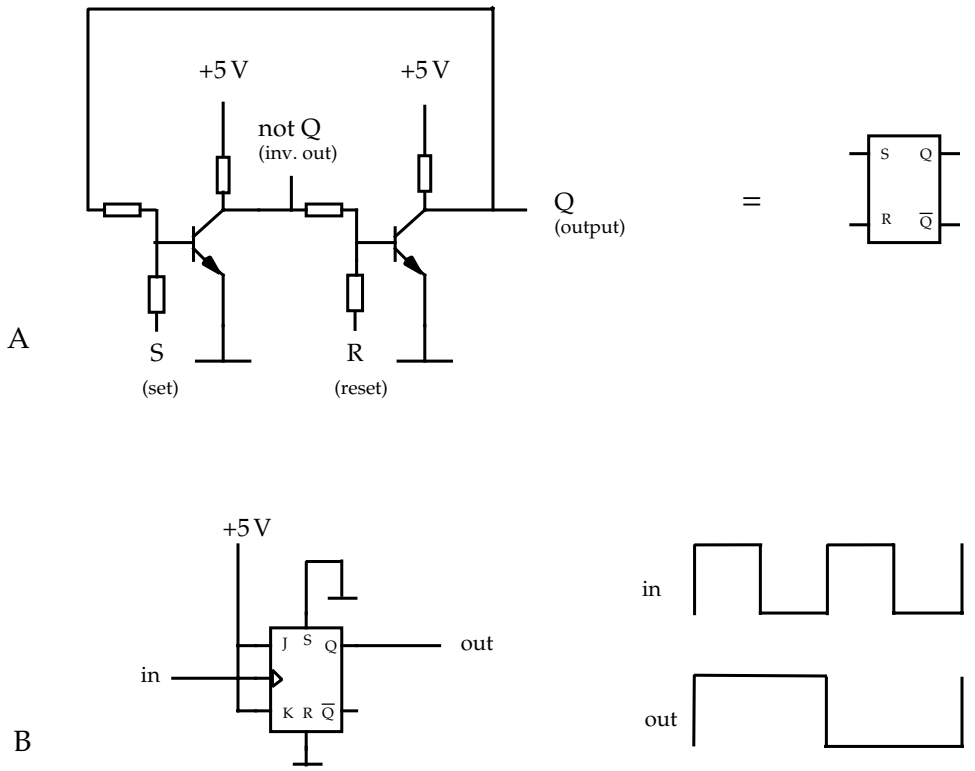


Fig. 2-49 Principle of basic set/reset (A) and divide-by-two (B) flip-flops.

forever. If we could somehow force the second transistor to start conducting ("flip" it), the first one would be shut off, and this would again last forever. If we could trickle it to flip again ("flop"), the first condition will be restored. This circuit therefore has a "memory": it simply retains the condition it was forced into, and it will hold on forever unless it is flipped again or until the power is shut off. In the latter case, by the way, turning power on results in an unpredictable state: this is why many digital instruments need to be reset, or be put into a certain starting condition intentionally, after switching on the power. Most modern instruments, such as counters, frequency meters, computers and so on, perform a reset automatically.

Note that we may consider each of the two transistors as the output, and so that a flip-flop has two outputs, in schematics usually designated Q and ~Q (NOT Q), and that both outputs are often used simultaneously to drive several other circuits. A flip-flop has a built-in inverter, as it were. In large arrays, however, such as in memory circuits, there is only space to connect one output, if at all.

Indeed, the flip-flop is the basic element of the first forms of digital memory (random-access memory, or RAM). It is still used in memory chips called static RAM (the larger and faster dynamic RAM, or DRAM, used in personal computers use a different principle, where a capacitor is used to retain the ones and zeros).

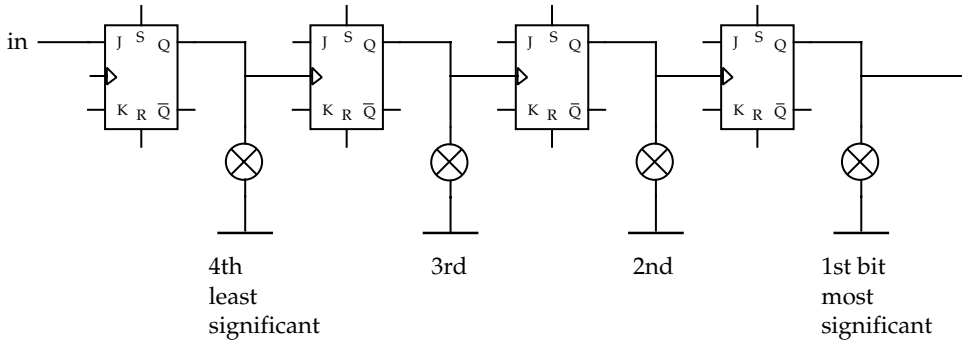


Fig. 2-50 A chain of flip-flops.

To control the state of a flip-flop, several kinds of input circuit are added, which we will not describe in detail. The most important input signals are set and reset (see RS flip-flop in Fig. 2-48. A type of flip-flop called the J-K flip-flop can be wired so that the output flips after each full cycle of the input pulse. Since flipping twice yields one output pulse, such a flip-flop divides the input frequency by a factor of two, and is hence called divide-by-two flip-flop: see Fig. 2-46B. A chain of flip-flops may divide the input frequency by 4, 8, 16, 32 and so on. This is called a frequency divider circuit.

The circuit of Fig. 2-50 shows a chain of four divide-by-two flip-flops, each output driving a (small) lamp. Therefore, aside from dividing an input frequency by a factor of 16, this circuit may be used to count to 16, since the 16 possible states can be monitored, or read out, by looking at the lamps.

These 16 possible states, running from 0 up to 15, are shown below:

0000	0100	1000	1100
0001	0101	1001	1101
0010	0110	1010	1110
0011	0111	1011	1111

where zeros and ones stand for lamps in the off or on condition respectively.

Indeed, we are counting from 0 to 15 in the binary system (often abbreviated as bin): each flip-flop output represents one binary unit of information, or bit. The leftmost column reads, from top to bottom, zero, one, two and three, the next column from four to seven, and so on. These four bits together, reflecting 16 possible states, are occasionally abbreviated as a nibble, and are often written as one digit in the hexadecimal system (hex for short). And since we have only enough numerals for the decimal system (dec for short), the capital letters A through F are used to fill in the missing symbols. Thus, counting from 0 to 15 in the hexadecimal system goes as follows:

0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F

After F comes . . . indeed, 10, in this case not to be pronounced as "ten", but as either "one, zero" or as "sixteen". More familiar than the nibble is the composition of two nibbles, or eight bits, into one byte. Obviously, one byte reflects $16 \times 16 = 256$ possible states, and is often written as two hex digits, the largest one being hex FF, or decimal 255.

The preference of digital circuits for powers of two is still found in many other properties of computers, such as the "kilo", actually meaning $1024 = 2^{10}$ rather than the familiar 1000, "Mega", meaning $1024^2 = 1\,048\,576$ (and neither 1 000 000, nor 1000 kilo), and so on.

Despite this bias to powers of two, we can convert, with a few extra gates, a four bit counter to jump from 1001 (nine) directly to zero. In this way, we can build divide-by-ten (decade divider) circuits and decimal counters. Because the factor of ten is achieved with binary circuits, this is called binary-coded decimal, or BCD. A chain of decimal counters is called a digital counter, or counter for short (since counting is digital by definition), and is used often in electrophysiology to count action potentials or other events.

Usually, however, one is not interested in total counts, but in frequencies, or counts per time interval. Therefore, a counter is usually fitted with a gate, which is switched on for a predetermined time. Most often, one second is chosen, since this yields a readout of frequency directly in hertz. The one-second pulses are usually derived from a so-called clock circuit, based on a quartz oscillator. The basis of such a square wave generator is an astable multivibrator, a circuit that resembles a flip-flop, but has no stable states at all: it changes states at a high rate. The frequency is stabilized by using the mechanical resonance of a quartz crystal as a "tuning fork". Starting with a 32 768 Hz crystal, one gets one-second gate pulses (followed by one-second pauses) using a chain of 16 divider flip-flops.

The full schematic of a digital frequency meter is given in Fig. 2-51. Here, by the way, the indication "digital" is not superfluous, since traditionally frequency meters were analogue, accumulating voltage pulses in a capacitor. The input signal, which might be a sine wave such as shown in trace A, is turned into a square wave by a circuit called a Schmitt trigger. This is a kind of op-amp circuit resembling the comparator (open-loop circuit, see earlier section), but with a little positive feedback added to assure that states intermediate between "0" and "1" (the "forbidden" states) are traversed very quickly. Thus, it helps to make the edges steeper. The "straightened" input signal is shown in trace B. Trace C shows the one-second gate pulse, and trace D that part of the input signal that passes the gate and hence is counted. At gate closure, the counter shows the number of pulses counted, and hence the frequency of the original signal. The memory section is added to be able to display a value while the next measurement is being made.

Digital measuring instruments have become increasingly popular. In part, this is driven by the technical possibilities to integrate a large number of transistors and accessory components onto a single "chip", but is based also on the superior properties of digital over analogue meters. The most important virtues are reliability and precision. The functioning of digital instruments is still more independent on the whims of individual components than the op-amp-like circuits we described earlier. If we feed exactly 13 pulses to the simple counter described above, the displayed state will each time be a faithful "1101" (hex "D"). We may proudly announce: digital operations are *exact*. This is not to say that digital measurements have infinite precision! The precision of the above-mentioned frequency meter depends mainly on the precision of the crystal used as the time base, and a bit on the oscillator circuit built around it. Without special precautions, most quartz crystals, such as the ones used in electronic wrist watches, have a basic precision up to one part in 10^5, which is usually sufficient for general laboratory practice.

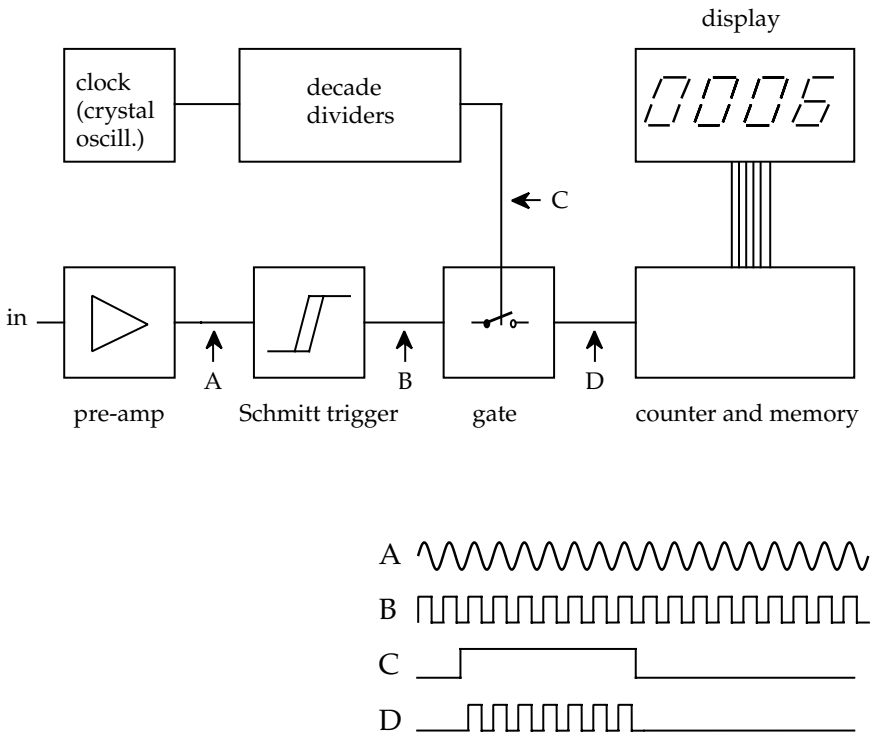


Fig. 2-51 Schematics of a frequency meter.

Precision is raised in some instruments by keeping the crystal at a constant temperature (in an "oven").

An other feature that contributes to the reliability of digital instruments lies in the nature of the display. In addition to the above-mentioned row of lamps, practical instruments may employ luminous (LED) or black (liquid-crystal) digits, which are seldom misread. To the contrary, the needles of conventional voltmeters can be misread by observing from an oblique angle, or by making mistakes in reading the needle position between the different types of scale division lines.

Note that the above statement about the input signal must be true to get the precise answer we expect. This might seem obvious, but it is important to realize that the precision of any digital instrument will collapse if it is fed with a "dirty" signal. This is illustrated in Fig. 2-52, where noise on a low-frequency sine signal causes multiple false triggers, which will yield fluctuating and completely erroneous frequency readings. In electrophysiology, one often has signals that are noisy and/or contain spurious interference "spikes" (not to be confused with the wanted action potentials), stemming from refrigerators and hosts of other machines connected to the electrical mains.

Apart from these possible pitfalls, digital instruments are the most reliable and versatile tools available to process electrophysiological, or indeed any, signals. To prevent the above-mentioned pitfalls, the user has always the responsibility to check whether the input signals

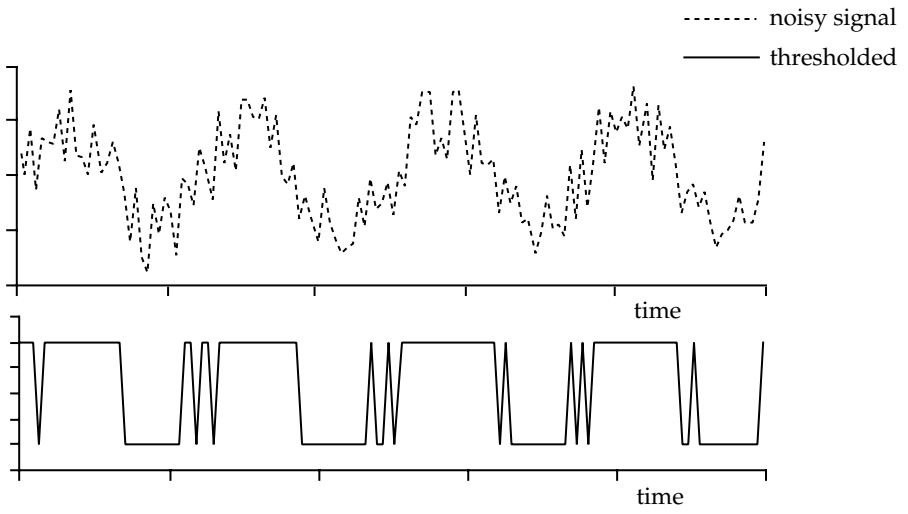


Fig. 2-52 Erroneous input signal to a frequency meter.

are suited to be processed digitally. Alas, the often cited phrase "garbage in–garbage out" applies to all our signal processing!

A/D AND D/A CONVERSIONS

The segregation of measuring instruments in analogue and digital is not of a categorical kind. Indeed, most real-world quantities, such as weight, length, voltage an so on are on an analogue scale, and must be converted into digital form to be processed and displayed. This conversion step is called analogue-to-digital conversion, and the circuit doing it is called an analogue-to-digital converter, or ADC. The inverse circuit, a digital-to-analogue converter (DAC), is also often used, especially in combination with computers. A DAC is necessary, for example, if the output of a digital meter must be written onto a paper chart, if a digital instrument is used to control the gain of an amplifier, or the intensity of a light used as a stimulus and so on.

The frequency meter described above may serve as a vantage point to illustrate the principles of AD and DA conversion. The simplest form of a DAC is shown in Fig. 2-53. Here, we added resistors to the outputs of the four flip-flops from Fig. 2-50, and add the resultant currents with an op-amp. A current through these resistors flows only when the output of the flip-flop is "1".

By choosing the resistance value ratios to correspond with the values ("weights") of the bits, the output voltage consists of 16 steps, from 0 to 15 units, and thus forms an "analogue" representation of the states of the counter.

However, practical DACs are a bit more complex, the main reason being that the voltage pertaining to a logical one may in reality fluctuate between about 2.5 and 5 V, and so would not form a reliable output signal component.

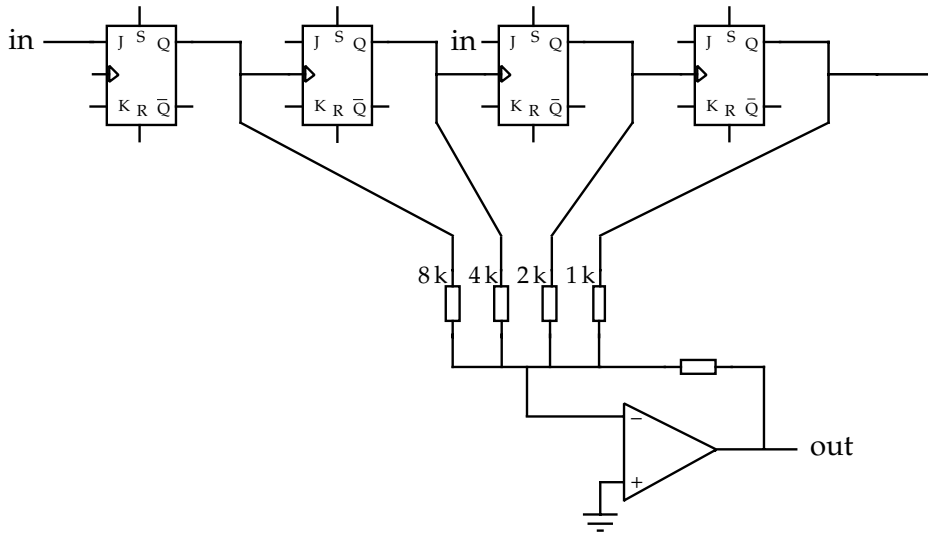


Fig. 2-53 Simple digital-to-analogue converter.

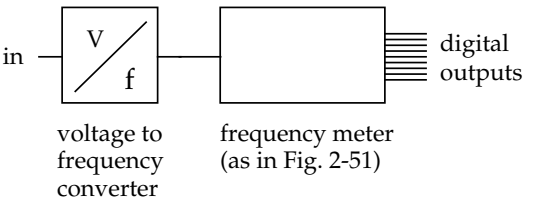


Fig. 2-54 Simple analogue-to-digital converter.

How to convert an analogue signal into a digital one? One of the most simple ways to perform an AD conversion is illustrated in Fig. 2-54. First, the voltage to be measured is converted into a proportional frequency, and then fed to the digital frequency meter described before. Although this form of ADC is simple to understand, its resulting measurements are not very precise in practice, and there are far more, and better, ways to perform the conversion from analogue to digital. These are, however, outside the scope of this book.

Nevertheless, the discussion on ADCs and DACs puts the finger on the weak spots in digital techniques: any digital instrument is *only as good as its AD (and/or DA) converter!*

Therefore, digital voltmeters are not inherently precise, and the specifications of many DVMs used in the lab, especially the cheaper ones, are not much better than analogue ones could be. Fortunately, other advantages, such as the reliable read-out, the compact and sturdy construction, remain valid, as well as the ease of use: many DVMs are of auto-polarity and auto-range, which means they sense the polarity and the approximate value of the input signal respectively, and adjust themselves to give the proper display range.

A series of digital measurements, or numbers, derived from an analogue signal constitute a digital signal. The fact that the continuous real-world quantity is chopped up, or sampled, into separate, or discrete parts has consequences that any scientist should know. The

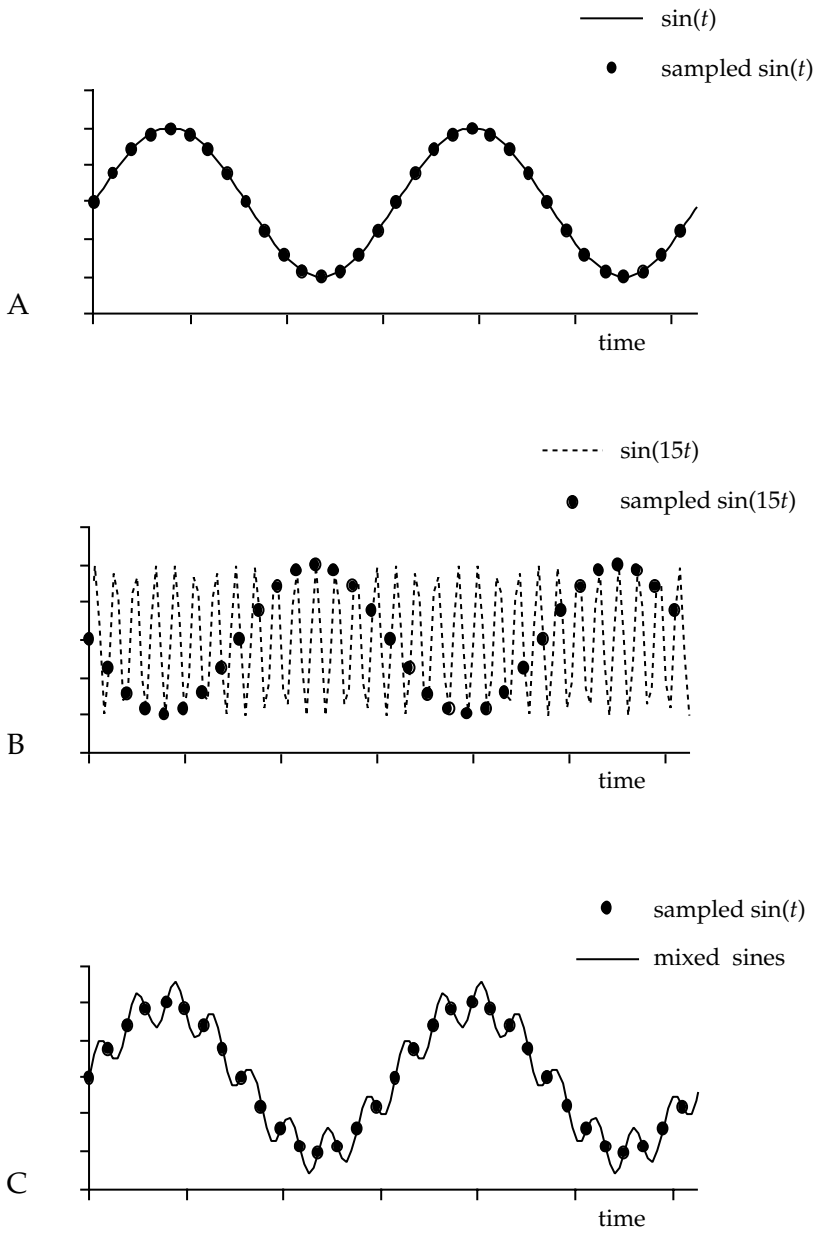


Fig. 2-55 Problems of waveform digitization.

process is called sampling or digitizing. One of the pitfalls is shown in Fig. 2-55. The upper traces (A) show graphic representations of a sine wave and its digitized counterpart. For the human eye at least, the digital signal is a fair representation of the original, analogue signal.

Figure 2-55B, however, shows a case in which something went wrong: the high frequency sine is sampled too infrequently, which results in a digital representation that reflects a non-existent, low-frequency sine wave. This phenomenon is called aliasing (from "alias"—a false name or identity).

It can, and must, be prevented by taking care that the sampling frequency fulfils the Nyquist criterion, which states that any waveform of interest must be sampled with a frequency of more than two times the (highest) frequency of interest (hence called the Nyquist frequency). For electrophysiological measurements, this implies that we must either choose a sampling frequency so high that it is, without question, higher than twice the highest signal frequency, or we have to limit the bandwidth of the preamp in order to fulfil the Nyquist criterion. To this end, some digital instruments have a special anti-alias filter built in. This is often a higher-order filter that allows signal frequencies from zero to slightly under the Nyquist frequency to pass unattenuated, yet prevents aliasing by cutting off higher frequencies very sharply, up to −100 dB/octave. Nevertheless, one must always keep in mind that a digital signal is undetermined between two samples, i.e. that one is not informed about the real world during the time between two samples. Thus, the digital signal shown in Fig. 2-55A might stem from the (improbable) signal shown in C.

COMPUTERS

The most complex and versatile, and certainly the best known type of digital apparatus is the computer, the operation of which can be derived from the simpler digital circuits we dealt with above. Modern personal and office computers, including the ones in electrophysiological labs, are very complex machines, which are hard to fathom. For the users, much depends on the operating system (OS), such as Microsoft's Windows, Apple's OSX of the many versions of Unix. Many good books to learn to control these OS'es exist. In physiology, however, computers must be extended to handle electrical signals. Some turn-key systems do exist, but most researchers have to bolt their electrical apparatus to the computers themselves. In these cases, a bit of insight into the architecture might help to understand the possibilities and problems of data acquisition and analysis. Therefore, although this is not a book on computer techniques, a few basic operating principles will be treated below.

Apart from counting and measuring, digital circuits are very well suited to do mathematical computations, and indeed any kind of manipulation of numbers in general (the numbers may of course represent letters, words or other symbols). Adding two numbers, for instance, can be performed with a combination of gate circuits. Multiplying and dividing a number by a power of two is particularly simple. This can be illustrated with our 4-bit counter: take the decimal number 6 (bin 0110), and shift the bits one position right, or one position left.

decimal:	12	6	3
	<− left	right −>	
binary:	1100	0110	0011

Obviously, shifting right once more, i.e. dividing an odd number by two, will yield fractional numbers, but these can be handled easily by digital circuits if one makes the necessary conventions (0001.1 would have to mean decimal 1.5).

Conversely, shifting left farther will yield numbers higher than (decimal) 16 (11 000 = decimal 24). When bounded to the original four bits, we get rounding and overflow respectively:

$12 \times 2 =$ overflow	$3/2 = 1$ (rounded)
*1000	0001

where the asterisk stands for the signalling of the overflow condition. In larger binary chains, this is simply the signal to set the next, more significant bit, but any digital instrument should signal overflow of the most significant bit to the user (e.g. an "overflow" LED on a counter, or an error condition in a computer program).

In the early days of digital electronics, the problem with digital circuits was that different operations demand different forms of wiring of the circuits performing the calculations, and that it had become unwieldy to change the wiring by patch panels, or by flipping switches all the time. This was solved by the principle of the computer, where the circuits are made flexible, and where "changing the wiring" is performed in a highly mechanized or electronic way, and made sensitive to electronic command signals called instructions.

Therefore, the single, most essential property of a computer is that it is a digital circuit, the function of which is determined in rapid succession by a set of input signals, rather than by the wiring. The sequence of instructions is called the program. In principle, these instructions could be fed in on demand, as an input signal, but in practice, a number of instructions is stored at once, in a circuit called digital memory. In principle, a memory is a large array of flip-flops, each holding one bit of information. One can write into memory by setting the appropriate flip-flops to zero or to one at will, and read from memory by connecting the outputs of the relevant flip-flops to the desired output wires. This stored-program feature is the second important characteristic of the computer.

Thus, the main functional parts of a computer are the processor, which performs most operations, and the memory, which holds the program as well as the data, i.e. the numbers that have to be processed. The processor may be a single chip, but is often a group of related, partly specialized chips, such as a math coprocessor and a memory management unit (MMU). The speed of computer functioning is largely dependent of the clock frequency, and hence speed, of the processor. This is the speed mentioned in advertisements, tests, etc. However, the overall speed, available to the user, depends to a large extent on other architectural details, and on the speed of other components like the hard disks and other storage media, and even on the network connections, if available.

A third, indispensable computer characteristic is the bus structure of the connections between these different parts. A bus is a set of wires that run across all component circuits. Together, the bus wires code for as many bits of information as the number of wires, often 16, 32 or more. In general, three busses can be distinguished that provide for the flow of all signals in a (simple) computer.

In the first place comes the data bus, which carries the instructions as well as the numbers that are processed, and the address bus, which specifies which memory cell is to be read from or written into. A third set of common wires, occasionally called the control bus, switches the right parts on and off, so that the screen gets the information to be displayed, and the printer the information that must be rendered on paper. For most types of personal computer, the bus speed, rather than the processor speed, is the bottleneck in the maximum overall speed that

can be attained. Therefore, a further component called a cache memory is added to speed up most operations. This is a small RAM chip, connected to the processor with a special, faster type of bus. How this speeds up the work can be understood by the fact that, if the computer fetches the contents (byte) of a memory location, most often the next series of addresses will be needed a short time later. Therefore, if a certain address is asked for, the computer's OS puts the contents of a whole block of main memory into this cache memory, so that subsequent memory calls are handled faster. Many computers even have more levels of cache memory: one on the chip, the others as close as possible to it (backside and motherboard caches).

The busses together interconnect not only all internal parts, such as processor, memory, hard and floppy disk drives and so forth, but also external parts (peripherals) such as keyboard, mouse, printers and so on. A *simplified* block diagram is shown in Fig. 2-56.

To work with a computer, it is obvious that it must have a user interface, i.e. a set of devices to control it. This usually comprises of a screen, also called a monitor, together with a keyboard and a mouse. Most often, the screen is a television-like CRT (cathode ray tube), although LCDs (liquid crystal displays) are increasingly popular.

In addition to controlling a computer, CRT screens are used frequently to generate complex patterns used in visual science. In this case, it is of utmost importance to know the spatial resolution, colour space, scan raster (i.e. pixel rate and frame rate), geometric precision and the like. More than once, artefacts stemming from the image generation have been showing up in electrophysiological experiments. The weal and woe of CRT screens when used in experiments is treated in Appendix D.

In addition to the mentioned peripherals, most computers have extra, free connectors, or slots, to connect a host of other instruments, such as the AD and DA converters necessary to handle analogue signals from the real world, modems to send data through a telephone line, extra printers or monitors, etc.

The above-mentioned parts are collectively called the hardware of the computer, in contrast with the software, which is the familiar description of the set of all possible or available

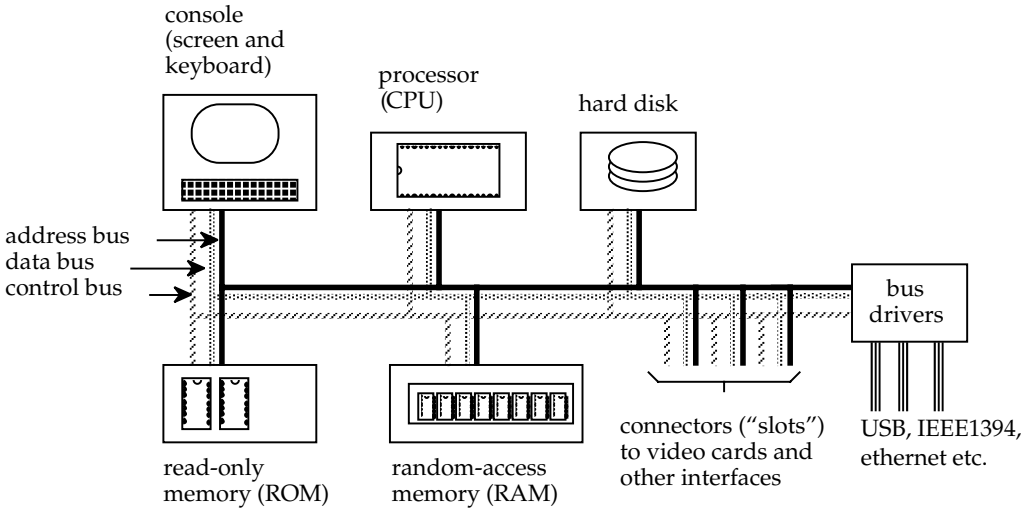


Fig. 2-56 Simplified schematic plan of a computer.

programs, routines or instructions. The software is "soft" in the sense that it is fundamentally a collection of (mathematical) ideas about how to process data, and thus exists only in the brains of people. To feed software into a computer, some form of carrier of these ideas is necessary. Formerly, the carrier was a punched paper tape or a deck of punched cards, nowadays it is mostly a hard disk, compact disk or other magnetic and/or optic medium. Flash memory chips (RAM chips that retain the data without the need for a power supply) are increasingly popular for daily data exchange, whereas recordable (or rewritable) DVDs are a good choice for long-term data storage. All kinds of data carriers together are called the media.

The principle of the computer and the first working prototypes were designed before and independent of the development of electronics. The first mechanical devices that may be called computers date back from the 1820s: when Charles Babbage worked on his (unfinished) "analytical engines". In the 1940s, several of such mechanical calculators were built using electromechanical relays, comparable to the ones in telephone exchanges. Not much later, however, the usefulness of vacuum tubes as electronic "switches" was recognized: they needed less current, could switch very much faster, and could be built smaller than relays. This led to the construction of the first electronic computers, such as the "electronic numerical integrator and computer", ENIAC, in 1943–1946. The vacuum-tube computer grew up to the very limits of construction: the more components, the more frequently an occasional failure would hamper functioning of the whole apparatus. Remember the chain of flip-flops we discussed as a counter: counting would stop or become erroneous if any of the participating flip-flops would fail. So, a computer with more than about 50 000 tubes would hardly function, if at all. The advent of transistors, being not only much smaller, but also far more reliable components got the principle of the computer where it is now: millions (or even billions) of transistors, joined into large-scale ICs, performing billions of instructions per second.

The first generations of (large) computers, and the fastest ones in use today, employed small magnetic rings, or cores, as one-bit memory cell. Zeros and ones were stored as the direction of the magnetic field: clockwise or anti-clockwise. The term core memory, still heard today, stems from these machines. With the advent of large-scale metal-oxide semiconductor (MOS) memory chips, the prices came down by a factor of more than 1000. These chips, nowadays called random-access memory (RAM—because each cell can be accessed at will, contrary to the tapes used earlier). This, together with the construction of the single-chip microprocessor, led in the late 1970s to a stripped-down version of the large computer, called "microcomputer", which was affordable for small companies and even family households. This microcomputer evolved into today's personal computer, a device that does not need any introduction at all.

However, for a better understanding of the functioning of these everyday machines, and to see what is involved in the use of computers in electrophysiological research, we will discuss the main principles and the gross construction of the average PC, together with examples of data manipulation.

The PC is evolving very rapidly, and so giving details about specifications of PCs is a hopeless task, since this kind of information is often obsolete the very moment it appears in print. Therefore, we will stick to a simple, hypothetical type, not unlike the first microcomputers, the working principle of which is most easily grasped.

A freshly built, or "virgin", computer would consist of all the necessary parts, yet, having no program, would not be able to perform any tasks at all. Therefore, the most basic pieces of program, or routines, that are necessary to make a computer able to process commands from the keyboard, or to read something from a disk, is built into a fixed-contents memory, or read-only memory (ROM). The collection of programs to access the disks, listen to the

keyboard, put text or graphics on the screen, and so on is called the operating system (OS), or, since the disk plays the central part, a disk operating system (DOS). Although a few brands of computers have their operating system in ROM, most have only the most basic routines, the basic input and output system (BIOS) in ROM. This BIOS is factory-programmed to read the OS from disk into RAM at start-up, or, rather, to read that part of the OS needed at start-up. Because ROM chips are not easily interchanged or reprogrammed, they are collectively called the firmware of a computer, i.e. the relatively fixed part of the software.

All other instructions to a computer, be it parts of the OS or user applications, are called software, and are fed to the computer on demand, usually either from the built-in hard disk or from any kind of removable media. The instructions must be put in a form the computer understands, in a programming language (or computer language), such as Basic, Pascal or Fortran. Higher-lever languages such as Mathematica®, Matlab® and Sysquake® are increasingly popular, because more complicated tasks can be programmed with a few, relatively simple instructions. LabView is a graphical programming language specifically designed for the direct control of AD cards and other laboratory instruments.

Strictly speaking, however, computers do not understand these languages directly. The statements from these so-called higher languages must be translated first, often by the same computer, into machine language. The type of machine language used depends on the processor, and differs widely among different brands of chips. Taken still more literally, the only "language" a computer understands is a certain succession of voltages, or patterns of voltages, applied to the set of bus wires. Obviously, these electrical instructions are abbreviated, or made more abstract, to the benefit of the human programmer. Thus, the voltages of 0 and +5 V are rendered as the numbers 0 and 1; and these are grouped further into hexadecimal digits, and so on. In the latter form, a piece of machine language for some processor would look like this:

2C 07 02 46 00 0F E8 4F 06 46 00 30 … etc.

Some of these numbers code for instructions, others for addresses where to get or store a number, and some are numbers that are to be used, such as constants. Obviously, this is still hard to read, and led to the development of the assembly language, by giving instructions, processor registers, etc. short names. These abbreviations, called mnemonics, help to understand what the code performs. In assembly language, the above program fragment would be notated as:

```
move.l    d7,d6
andi.w    #$f,d6
lsr.w     #4,d7
addi.w    #$30,d6
```

Here, a programmer will be able to recognize the instructions easily as "move a long-word", "immediate AND", "left shift" and so on. By the way, in the "daily computer life", assembly language is often called "machine language", in contrast with the higher programming languages mentioned earlier.

Writing complex computing or data manipulation programs entirely in assembly is a tedious task, since even the most simple calculation would cost several thousand lines of code.

Note that the assembly language example given above is only a *minute* part of a decimal-to-hex conversion routine. Therefore, most programmers use a higher language to write their

programs in, such as Pascal, Basic or C. These languages look more like "algebra, stated in English", and are not only easier to read and remember, but also far more compact. The simple statement "c = a + b" would need many assembly-level instructions to fetch numbers, put them in registers, convert them to binary form, add them byte by byte (or send them to a floating-point processor), convert them back to decimal, and so on.

As an illustration, the listings below show similarities and differences between three popular programming languages: Fortran, Basic and C. The output of all three mini-programs is the same: a table of the first 6 natural numbers and their squares.

Program text	Added remarks
C IN FORTRAN	C = comment
INTEGER I, IMAX	declare loop counter and max.
PROGRAM	start of program
IMAX = 6;	set loop maximum
DO 100 I = 1,IMAX	start of loop
WRITE(6,200) I, I**2	main instructions, yield the output
100 CONTINUE	end of loop
200 FORMAT (I8, 1X, I8)	specification of printing format
END	end of program
5 REM IN BASIC	REM means remark (start of program is implicit)
10 IMAX = 6	setting loop maximum (declaration is implicit)
20 FOR I = 1 TO IMAX	start of loop
30 PRINT I, I^2	main instructions
40 NEXT I	end of loop
50 END	end of program
/* IN C: */	remark
#include <stdio.h>	invoke library of standard input and output (stdio)
main()	start of program
{	
int imax = 6;	declaring and setting the loop max
int i;	declaring the loop counter
for (i = 1; i <= 6; i++)	start of loop
printf ("%8d %8d", i, i*i);	main instructions
}	end of program

Fortran ("formula translator") was the first higher language that became widely used in the 1950s. It was extended and modernized over the decades, and therefore is still in use today, mainly in physics and technology. BASIC is said to mean "beginner's all-purpose symbolic instruction code", but it is almost certain that the acronym preceded its explanation. It was invented in the 1960s, mainly for didactic purposes, just like Pascal in the 1970s. Apart from serving beginners' needs, both proved useful, however, for full-blown scientific programming tasks, and are used very often, especially in personal computers.

Any program written in a higher-level language must be translated into the mentioned machine code by either of two methods: compiling and interpreting. This is performed by

computer programs (!), known as compiler and an interpreter respectively. An interpreter translates one instruction (or program line) at a time, and then executes it. Thus, a Basic interpreter would take line 10 of our Basic example, and do the following: reserve a few bytes of memory to hold the variable IMAX, attach this name to the address of this memory register; convert the character "6" into binary representation (00000000 00000110) and put it in the reserved memory locations. Then it would read the next line.

A compiler, to the contrary, would take the complete program text, in this respect called source code, and convert all instructions into machine-language form. Thus, the output of a compiler (called object code or runtime code), is a complete, independently working program (stand-alone program, which may be given (or sold) to people that do not need to own the compiler used. Large, stand-alone, programs are also called executable (EXE), application (Program) or simply program. This is in contrast with an interpreter text, which needs the interpreter each time it is executed (i.e. at runtime).

However, there is a price to the added convenience of higher-level languages: most of the code generated by these compilers or interpreters does not run so fast as when the same task would be put directly in machine language form. In this respect, the C language generates the code running fastest, which is said to be the most efficient code. Therefore, most user applications, such as word processors, spreadsheets and image analysis programs, are written in "C". Even then, assembly language is often still used for the most time-critical parts of the programs.

3

Electrochemistry

INTRODUCTION, PROPERTIES OF ELECTROLYTES

Electrical processes in living organisms take place in watery solutions containing salts, proteins, carbohydrates and a host of other organic and inorganic substances. These processes are dominated to a large extent by various salts. Therefore, we will need a good understanding of the properties of electrolyte solutions and of the processes associated with them. In addition, most methods to get measurements from the wet medium are carried out with electronic instruments, which must be connected somehow to the process studied. Therefore, we are interested also in the processes at the electrodes used for measurement and stimulation. More precisely, the point of focus is the boundary between the electrode and the solution, or metal/electrolyte interface (conducting solutions are called "electrolyte"). Insulators do not support an electric current, of course, but may show electrical influences when brought into contact with electrolyte solutions; this is because the surfaces of many insulators consist of charged particles. Finally, electrical processes take place at the boundaries between different solutions. Thus, we will have to deal with the following boundaries:

1. metal/electrolyte
2. insulator/electrolyte
3. electrolyte 1/electrolyte 2.

We will treat the metal/electrolyte, or metal/solution, interface in greater detail below. Boundaries at insulators such as glass and plastics give rise to so-called electrokinetic effects, which will be described briefly, later. The third case is called a liquid junction or diffusion boundary. It arises where two different solutions meet in situations where convection or mixing is prevented, such as in gels, porous substances and capillaries. This causes the dreaded diffusion potential, or liquid junction potential (LJP), which forms a threat to the validity of many (DC) electrophysiological measurements.

Knowledge of these processes can be extended easily to include the electrical phenomena at the cell membrane, where two solutions are separated by a selectively permeable membrane. Thus, we will add the following system:

electrolyte/membrane/electrolyte

The processes involving membrane potentials and conductances are important in neurophysiology. In addition, potentials across artificial membranes are used to measure ion activities, such as pH.

Electrolytes

Electrolytic solutions, as their name suggests, arise because a number of substances that are electrically neutral in dry condition can dissociate, or split, into charged particles, or ions, when they are dissolved in a suitable solute (usually water). In principle, all dissolved salts are dissociated fully into their constituent ions, but not all salts dissolve readily.

The degree to which a salt dissolves (and splits) depends on the solubility product, which is the product of the concentrations of the constituent ions. For silver chloride (AgCl), an important substance in electrophysiology, the solubility product, i.e. $[Ag^+] \times [Cl^-]$, is about 10^{-10}. Dissolved in pure water, the solid AgCl dissolves until both ion concentrations are 10^{-5}. When other salts are present, things may change. In a physiological saline environment, for example, the chloride concentration is in the order of 0.1 M. Therefore, solid AgCl will dissolve only up to a silver ion concentration of 10^{-9}.

Some salts, such as rock minerals, are so insoluble that their dissolution may take lots of water during millions of years (or takes place by wearing down mechanically).

Acids and bases are a bit different. In principle, acids split into hydrogen ions (H^+) and an anion that characterizes the particular acid. In the same way, bases split into a characteristic cation and the negative hydroxyl ion (OH^-). The strength of acids and bases is characterized by their respective dissociation constants. Some acids and bases are almost fully dissociated, and hence called "strong", whereas many others dissociate only partially and hence called "weak". Examples of strong acids are hydrochloric acid (HCl) and nitric acid (HNO_3); of weak acids, carbonic acid (H_2CO_3) and hydrogen sulphide (H_2S). Note that, in solution, weak acids and bases are for the greater part undissociated, and so do not take part in reactions. Therefore, weak acids and bases are very useful to make buffer solutions: if some H^+ (or OH^-) is "used up" by the substance the buffer is added to, a bit of acid (or base) extra is split, so that the H^+ (or OH^-) concentration is kept approximately constant.

In electrophysiology, several salts play important parts. A simple electrolyte, such as sodium chloride (NaCl), splits into one cation (Na^+) and one anion (Cl^-). In general, hydrogen and metals are positively charged, whereas the other half of the original salt carries a negative charge. This was found by early electrochemists by forcing an electric current through a salt solution. The metals, being positive, accumulate at the cathode (negative electrode), and hence are called cations, whereas the negative ions are called anions, because they migrate towards the anode (positive electrode).

The ions can move through the water, and so they are able to transport charge, just like electrons in a vacuum or like the electrons and holes in a semiconductor.

We will summarize the properties of electrolytes with a simple example, using NaCl. But simple electrolytes do not show simple behaviour, as will become clear later on. In addition, the solute itself, water (H_2O), shows remarkable features. First, the hydrogen atoms in the water molecule are distributed asymmetrically, causing the water molecule to have a positive side and a negative side: it is said to be polar. This is the cause of the very high relative dielectric constant of pure water: about 84 (most conventional dielectrics have a dielectric constant of about 2 to 6). This feature is of direct consequence to the behaviour of metal electrodes put in

an electrolyte solution, explained further on. Water would be the ideal dielectric for capacitors if it would be a better insulator. In fact, however, water conducts slightly, because of its partial ionization into H^+ and OH^-. Under normal conditions (room temperature and atmospheric pressure), a fraction of about 10^{-7} is split. Therefore, the pH of neutral solutions at room temperature is 7. At higher temperatures, more water is split, until at the boiling point (100 °C), the pH is about 6.

The presence of ions, being mobile charges, makes salt solutions fairly good conductors of electricity. How well they conduct is usually expressed as resistivity, abbreviated ρ (Greek "rho"), and often expressed in Ωcm (cgs system) or Ωm (SI). The inverse of resistivity, called the conductivity and abbreviated g, is also used frequently. The unit of conductivity is siemens (S, see Chapter 1). Note that resistivity is a characteristic of a substance rather than of an object. The resistivity, expressed in Ωcm, is the resistance of a column, or tray, with a length of 1 cm and a contact surface area of 1 cm^2, in other words, the resistance of a cube of 1 cm edge. In the case of other forms of column, the resistance is directly proportional to the length, and inversely proportional to the contact area. As a consequence, the resistivity is expressed in Ω cm^2/cm, which simplifies to Ωcm. This is easy to understand, since a longer column can be thought of as composed of resistors connected in series, whereas a wider column can be considered as a circuit of resistors in parallel.

Conduction of electrolytes is not as good as in metals, wherein the free electrons may cause resistivity values as low as 10^{-6} Ωcm. Yet, strong electrolyte solutions, such as 3 M KCl, have a resistivity of no more than 5–10 Ωcm. Seawater, having ion concentrations of about 0.5 M, has also a rather low resistivity. Other values can be taken from the following table, which gives approximate resistivities of common electrolyte solutions at 18 °C:

copper (for comparison)	1.7	μΩcm
3 M KCl	5	Ωcm
seawater (mid-ocean)	22	Ωcm
seawater (coastal areas)	25	Ωcm
physiological saline (about)	70	Ωcm
0.1 M KCl	90.9	Ωcm
0.01 M KCl	820	Ωcm
freshwater	0.2–10	kΩcm
distilled water	1–5	MΩcm
pure water (theoretically)	23	MΩcm

The value given for pure water, at a pH of 7, can be approached closely if boiled (i.e. degassed) water is deionized in a mixture of ion-exchange resins. Distilled water and deionized water, which are allowed to equilibrate with air, absorb carbon dioxide, which is partially dissociated into H^+ and HCO_3^-, and so causes the lower resistivities given. The two lower concentrations of KCl given are used to calibrate resistivity meters.

Many of the stated resistivity values have direct consequences for electrophysiological measurements. For instance, the resistance of glass capillary electrodes depends on the resistivity of the filling fluid. To keep the electrode resistance to a minimum, strong salt solutions such as 3–4 M KCl or KAc (acetate) are used. When dealing with the skin and skin sense organs of aquatic animals, the resistivity of the environment, being either freshwater or seawater, may play a role. Measurement of small currents in seawater is difficult, because they yield far lower

voltages than in freshwater. The composition of physiological salines, or "Ringer's fluids", is chosen to be similar to the natural body fluid of the used animal species in both chemical composition and resistivity.

In simple solutions, the resistivity depends on the salt concentration, but not in a completely linear way. In very dilute solutions, the ions are relatively independent, and behave, despite the presence of water molecules, more or less like atoms in a gas. In this case, a doubling of the concentration causes the conductivity to be doubled, and so the resistivity to be halved. Hence, resistivity is inversely proportional to concentration. In stronger solutions, however, the ions interact with each other in a rather complicated way, described among others in the famous works of Debije, Hückel and Onsager. Roughly speaking, the effect of interactions can be described as if ions are screening each other, reducing their electric "visibility" in the solution.

The main effect of these interactions is that the electrically active concentration, or activity for short, of ions in concentrated solutions is less than that expected from the true concentration. Thus, the ion activity can be considered to be the *electrically effective concentration*. To be able to distinguish this quantity from the true concentration, i.e. the total amount of substance present, this latter quantity is called, somewhat superfluously, the stoichiometric concentration. This means "by weight", i.e. the amount found by chemical analysis, by drying and weighing, etc., or simply the amount "put into it". Because the interactions of ions take place mainly at high concentrations, at concentrations of 0.1 mM or less, the activity is virtually the same as the concentration. At concentrations of over 1 mM, however, the screening becomes apparent. As an example, a 100 mM KCl solution behaves electrically as if the concentration were 76 mM. Therefore, the ion activity is said to be 76 mM. The conversion factor is called the activity coefficient f:

$$f = a/c, \quad \text{or} \quad a = cf$$

where c is the concentration and a the activity of an ion.

At higher concentrations, the activity coefficient can become as low as 0.55, so that a 3 M KCl solution behaves electrically as if it were 1.65 M. There is a complication, however. Activity coefficients of individual ion species cannot be determined directly. The activity of ions is reflected most prominently in the conductivity of salt solutions, but these consist necessarily of cations and anions, so that a resultant, or mean activity, coefficient is obtained. Figure 3-1 shows the mean activity coefficients of KCl, NaCl and $CaCl_2$ as a function of concentration, derived from several compilations of measurements (Landolt and Börnstein 1923–1936; Parsons 1959).

Under certain assumptions, the activity coefficients of salts having a common anion can be used to estimate the activity coefficients for individual ion species.

The result of such computations is shown in Fig. 3-2.

Note that, especially for divalent ions such as calcium, the reduction of the activity coefficient is very prominent: at 1 mM, the activity is already reduced to 0.9, and at 0.5 M the activity coefficient is reduced to a mere 0.2. At concentrations of 1 M and higher, the activity coefficients are hard to determine, and are usually estimated by compiling the results of different methods. Some of these measurements suggest that at very high concentrations, the activity coefficient rises again to reach values far higher than 1 (dotted lines in Fig. 3-2). This would mean that the negative interactions found at lower concentrations are changed into positive, or reinforcing, interactions. Although the mechanism hereof is not understood fully yet, the effects are large and must be taken into account whenever strong salt solutions are used, e.g. as salt bridges.

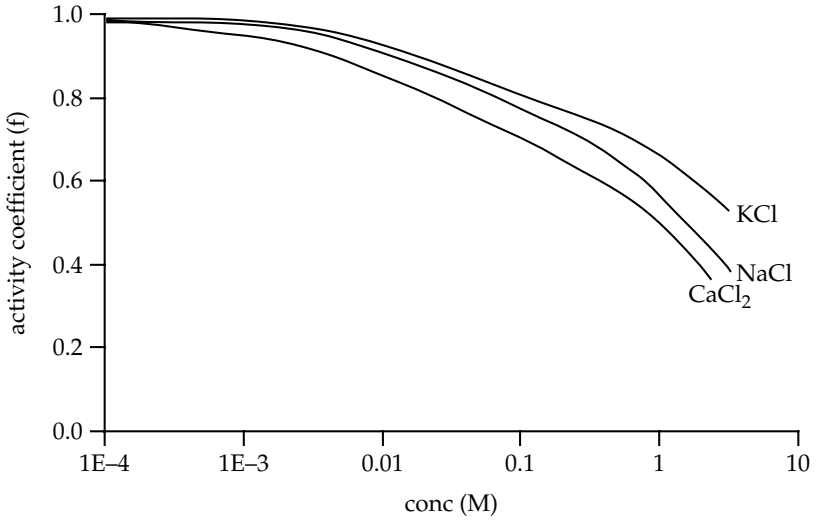


Fig. 3-1 Mean activity coefficients of salts versus concentration.

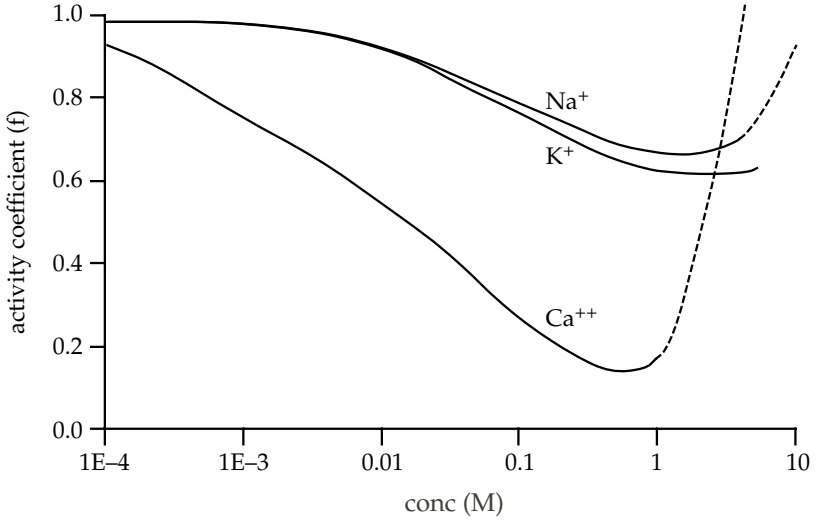


Fig. 3-2 Estimated activity coefficients of cations as a function of concentration (or ionic strength).

It is important to note that in all electrochemical calculations, ion activities rather than concentrations must be used.

Because of the interactions mentioned, the electrical resistivity of salt solutions depends on the activities of all the constituent ions. In addition, a quantity called the ion mobility plays an important part. For an ion to transport electricity, it must be mobile, i.e. it must be able to move with respect to the solute. The mobility of the different, familiar ions varies widely. It seems to depend on the radius of the ions, but many ions are slower than expected. Therefore,

one has to assume that some cation species drag water molecules with them, and this so-called hydration mantle is strongest in small cations. Mobility is abbreviated u, and expressed in $10^{-8}\,\mathrm{cm^2\,s^{-1}\,V^{-1}}$. The table below shows the mobility of a number of common ion species at room temperature.

Cations	u	Anions	u
H^+	36.3	OH^-	20.5
Li^+	4.0	Cl^-	7.9
K^+	7.6	NO_3^-	7.4
Na^+	5.2	HCO_3^-	4.6
NH_4^+	7.6	Ac^- (acetate)	4.2
Ag^+	6.4	$1/2SO_4^-$	8.3
$1/2Ca^{++}$	6.2	$1/2H_2PO_4^-$	3.7
$1/2Mg^{++}$	5.5	$1/2\ HPO_4^-$	5.9

From the table, hydrogen seems to be an exception, having a mobility far greater than all other metal, metalloid and composite ions. Therefore, it is suggested that hydrogen ions move as free protons. Note that the mobility of a single ion species cannot be determined directly, so that the values in the table are computed from many experiments involving different salts with either a common cation or a common anion. The mobility of an ion contributes directly to the conductivity of the solution. The conversion factor from mobility u to conductivity g follows from:

$$g_\mathrm{i} = Fu_\mathrm{i}$$

where g_i is the *contribution* of one equivalent (one mole of a univalent ion, half a mole of a divalent ion, etc.) of an ion species *to the conductance* of a solution, and u_i is the mobility of that ion. The proportionality constant F is called Faraday's constant. This is the amount of charge carried by one equivalent, and has the value of 96 500 (coulomb per equiv.). Faraday's constant is found very often in electrochemical relations, such as the equations of Nernst dealt with later on.

When dealing with composite electrolytic solutions, being mixtures of several salts at different concentrations, the notion of concentration can be replaced by the total ionic strength. The ionic strength S is defined by:

$$S = \frac{1}{2}\sum c_\mathrm{i} z_\mathrm{i}^2$$

where c_i is the concentration, z_i the valence of each ion species. The square arises because the electrostatic force between two point charges is proportional to the product of the two charges. Thus, the ionic strength of a 10 mM KCl solution is 10 mM, and that of a 10 mM $CaCl_2$ solution is 30 mM.

Although the ionic strength is a useful indication of the approximate electrical properties that can be expected, the detailed behaviour of ions may still depend on the precise composition of the solution, i.e. on the individual activities and mobilities. Because of interactions

between ions of different species, the activity coefficients shown above depend on ionic strength rather than on the concentration of only one ion species. To be more precise, therefore, ionic strength, rather than concentration, is the quantity shown along the horizontal axis in Figs. 3-1 and 3-2.

It is important to note that the exact composition of many solutions, such as the natural body fluids, is not known. Therefore, many electrochemical quantities *can only be approximated*.

In subsequent sections, we will need the following constants frequently:

molar gas constant	R	8.314	$\mathrm{J\,mol^{-1}\,K^{-1}}$
Avogadro's constant	N	6.03×10^{23}	$\mathrm{mol^{-1}}$
Faraday's constant	F	96 485	C/equiv.
elementary charge	e	1.6022×10^{-19}	C

THE METAL/ELECTROLYTE INTERFACE

At a metal/solution interface, three things can happen:

1. "Nothing"—the boundary behaves as capacitance.
2. Faradaic processes—the well-known redox reactions.
3. Non-Faradaic processes—adsorption and other complications.

Which of the above will happen depends on the compositions of the metal and of the solution, and on the voltage applied.

A major complication, in fact for all electrochemistry, is that the processes at one metal/solution boundary, called an electrochemical half-cell, cannot be studied in isolation. They can be studied only by joining two half-cells into a full electrochemical cell. This implies that any electrochemical measurement is in fact the difference of two components that can be separated only by the careful design of several experiments and by choosing an appropriate standard.

The standard adopted universally in electrochemical research is called the normal hydrogen electrode. It has been chosen carefully in such a way that it can be reproduced within very narrow tolerances in the laboratory, and consists of a porous platinum electrode saturated with hydrogen gas at one atmosphere pressure, in a solution containing one mole of hydrogen per litre (1 normal, or N). Although it may seem strange that hydrogen is used as a metal, this is found to be the best way to build a standard half-cell, and has been used for more than a century. In electrophysiology, however, potential measurements are almost always relative, and so may be taken against any stable but arbitrary standard. In the following discussion, all stated potential *differences* are meant to be the potential difference across a whole electrochemical cell, whereas all stated *potentials* are half-cell potentials taken with respect to the above-mentioned standard.

The best-known examples of electrode reactions pertain to the combination of a metal in a solution of a salt of that metal, such as copper in copper sulphate solution or zinc in zinc chloride solution. Since most metal salts are poisonous, this type of electrode cannot be used very often. An example is a pair of copper electrodes in porous pots filled with copper sulphate,

which is used in earth sciences to measure geoelectric fields. In electrophysiology, we need non-polarized electrodes in combination with physiological saline and, in general, solutions that are not poisonous to living tissues. In these cases, the solution does not contain the electrode metal in significant concentrations, and this has severe consequences for the processes that take place to move charge across the boundary.

As an example, we will take the half-cell composed of a mercury electrode and a 1 M KCl solution (Bard & Faulkner, 2001). At potentials greater than +268 mV, mercury is oxidized by combining with chloride ions:

$$2\mathrm{Hg} + 2\mathrm{Cl}^- \rightarrow \mathrm{HgCl}_2 + 2\mathrm{e}^-$$

At strongly negative potentials, at least 2.9 V, potassium ions are reduced (and form an alloy with mercury):

$$\mathrm{K}^+ + \mathrm{e}^- \rightarrow \mathrm{K(Hg)}$$

Capacitance of Polarized Electrodes

In the range of potentials between the two voltages mentioned above, no electrode reactions occur, and so no steady current (DC) flows whatsoever. This can be seen from a graph of the current flow (I) as a function of potential (E), the so-called I/V characteristic (or better I/E). This is illustrated in Fig. 3-3 (left). Note that the I/V characteristic of a pure resistance would be a straight line, the slope of which depending on the value of the resistance. In the potential range mentioned, the I/V characteristic of the metal electrode is flat, i.e. no current flows. In other words, the resistance is infinite. In this range, the electrode is called a polarized electrode. Although no DC is supported, a polarized electrode behaves as a rather large capacitance. This can be understood by the following argument.

An electrode can be considered as a border between two conducting substances (the metal and the solution). Since no charge can cross that border, it behaves as an insulator, and thus the whole acts as a capacitance. It can also be seen that the capacitances formed have very high values. If the boundary would be sharp, the "insulator" would be infinitely thin, and so the capacitance would be infinite. Since in reality the capacitance is finite, there must be a finite gap between the charges on opposite sides. In the nineteenth century, Helmholtz formulated a theory about this, the so-called electrical double layer.

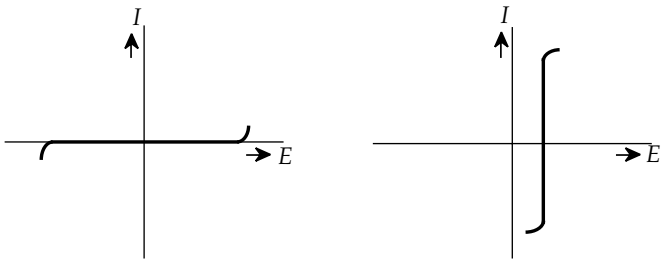


Fig. 3-3 *I/V* characteristics of ideal polarized and non-polarized electrodes.

One side of the double layer consists of the electrons in the metal. Because of the high conductivity of metals, this layer is very thin—far less than a nanometre. The other side, the charge in the solution, is more complex. Several processes determine the effective thickness of this layer. In the first place, ions have finite diameters, and can only approach the boundary down to the radius of the ion. In the second place, some ions adhere to the metal surface, thus keeping others from getting closer. These ions have hardly any mobility. This inner layer is called the "Helmholtz layer", and determines the properties of an electrode to a great extent. Farther out in the solution, the ions are more and more mobile, approaching the mobility in the bulk solution.

The profile resembles that of an "atmosphere", i.e. it assumes the same shape as the air pressure at different altitudes in the earth's atmosphere. In this way, the potential changes gradually from the value in the metal to the value in the solution. This is shown in Fig. 3-4.

The size of the double layer depends on the ion species involved, their concentrations, as well as on the potential, and may vary between less than 1 nm in concentrated solutions to tens of nanometre in very diluted solutions. Since the capacitance is dependent on the thickness of the double layer, the factors that affect the double layer will affect the capacitance, too. Therefore, metal electrodes, very popular in electrophysiology, have complicated properties that must be assessed or calibrated in any practical situation.

Faradaic Processes

Also called oxidoreduction reactions, redox reactions or electrolysis. Faradaic processes are the best-known electrode reactions, taught in any school chemistry programme. As an example, we will take the case of a zinc electrode in a zinc sulphate solution. By making the electrode sufficiently positive, zinc atoms can be oxidized into zinc ions:

$$Zn \rightarrow Zn^{++} + e^{-}$$

At sufficiently negative electrode potentials, zinc ions are reduced:

$$Zn^{++} + e^{-} \rightarrow Zn \downarrow$$

The metallic zinc precipitates onto the surface of the electrode.

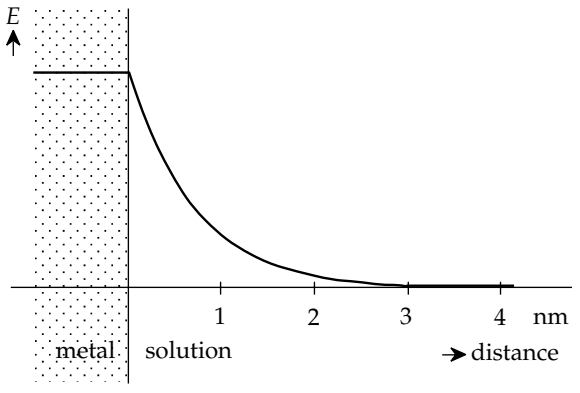


Fig. 3-4 Potential distribution across a metal/electrolyte interface.

In such a half-cell, the electrode assumes a stable equilibrium potential, virtually independent of current strength. The potential has been shown to depend on the concentration of the salt and on a theoretical property of the metal called the metal's solution pressure, or the tendency to dissolve. Nernst established the relation between these two as:

$$E = \frac{RT}{zF} \ln\left(\frac{c}{K}\right)$$

where E is the electrode potential, T is the absolute temperature, z is the valence, c is the concentration of the metal ion and K is the mentioned solution pressure. Although the latter is a hypothetical quantity, and is even criticized as to its physical meaning, the differences between metals are very real. Some metals, such as silver, gold and platinum, do not dissolve readily (i.e. oxidize), and are called "precious metals". Others, such as aluminium, iron and zinc, are oxidized very easily. Thus, all metals can be ordered according to their relative solution pressures. Hydrogen, although not a metal in the strict sense, fits in this electrochemical series perfectly. Its use as a standard was mentioned before.

The list below shows electrode (i.e. half-cell) potentials for a number of reactions, all taken at "normal" (one equivalent per litre) concentrations of the salt. These are called "standard electrode potentials", and are abbreviated E_0.

System	E_0 (in volt, at 25 °C)
Zn^{++}/Zn	−0.761
Fe^{++}/Fe	−0.440
Pb^{++}/Pb	−0.126
H^+/H_2	*0 by definition*
$AgCl/Ag^+$	+0.222
calomel	+0.281
Cu^{++}/Cu	+0.337
Hg^{++}/Hg	+0.789
Ag^+/Ag	+0.799
Au^{+++}/Au	+1.50

Note that the standard hydrogen electrode half-cell has a pH of zero. Since in biochemical systems, pH values are about 7 rather that at zero, biochemists often use electrode potentials referred to a hydrogen electrode at pH = 7.0. These are indicated as E_0' and can be found by subtracting 406 mV from the respective values in the table (why?). Standard potentials of other redox systems, such as metabolic systems, can be expressed in the same way, and are called standard redox potentials. These are almost always expressed as E_0' values.

As a result of Faradaic processes, these electrode half-cells conduct electric current with very little change of the potential. Therefore, a metal in a solution of one of its salts is known as a "non-polarized electrode". The I/V characteristic of an ideal non-polarized electrode is shown in Fig. 3-3, right. The electrode voltage is constant over a large range of current strengths. In that respect, the I/V characteristic of non-polarized electrodes is the opposite of the characteristics of polarized electrodes, where the current is zero over a large voltage range.

Practical Electrodes

Practical electrodes do not exactly follow the I/V characteristics of the ideal polarized or non-polarized electrodes. This is shown in Fig. 3-5.

In the case of a polarized electrode, which as we saw shows mainly capacitance, a small current can nevertheless traverse the electrode surface because a tiny amount of water can be split. Thus, the I/V characteristic is not exactly horizontal. In fact, a polarized electrode can be represented by a leaky capacitance, i.e. a capacitance shunted by a (relatively high) resistance. In the same way, a non-polarized electrode cannot conduct current without any deviations of the potential. Hence, a non-polarized electrode may be represented by a voltage source (*the* electrode potential) in series with a (relatively low) resistance. Such voltage is often called an "electromotive force" (emf, although it is not a force in the usual physical sense), and abbreviated as E. Both the so-called equivalent circuits are shown in Fig. 3-6.

Practical forms of electrodes are shown in Figs. 3-7 and 3-8.

Electrochemical Cells, Measuring Electrodes

The aforementioned processes pertain to one electrode, dipped into a salt solution. Any practical case, however, will consist of a full electrochemical cell, made up of two electrodes in one solution, or alternatively of two joined half-cells. The latter case comprises a liquid junction, which will be dealt with later. In general, two forms of electrochemical cell are used. For the first one, one picks metals as far apart in the list of electrode potentials as possible. In this case, the resulting cell voltage is as high as possible. This is used for batteries and accumulators, the electrochemical cells that are used as energy sources. The other extreme is used in electrophysiology: two electrodes are made from precisely the same metal, and dipped into a

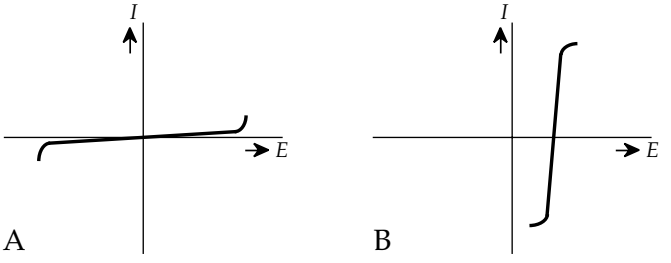


Fig. 3-5 *I/V* behaviour of practical polarized and non-polarized electrodes.

Fig. 3-6 Equivalent circuits of polarized (A) and non-polarized (B) electrodes.

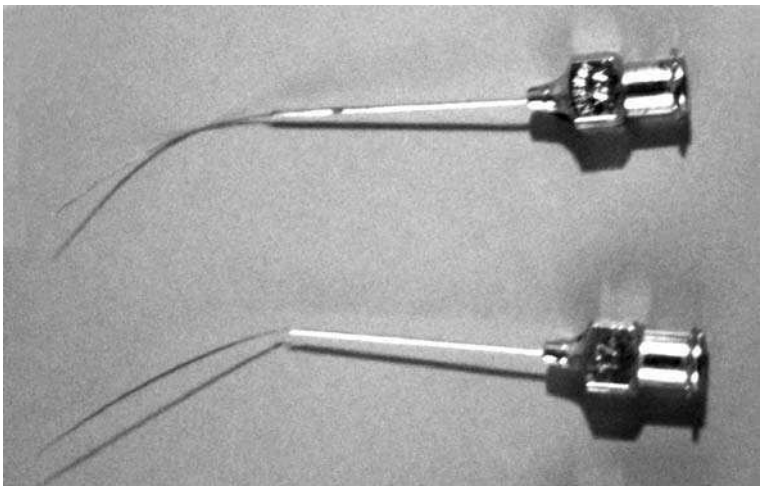


Fig. 3-7 Practical electrodes: pointed and coated tungsten wire electrodes, mounted in all-metal hypodermic needles.

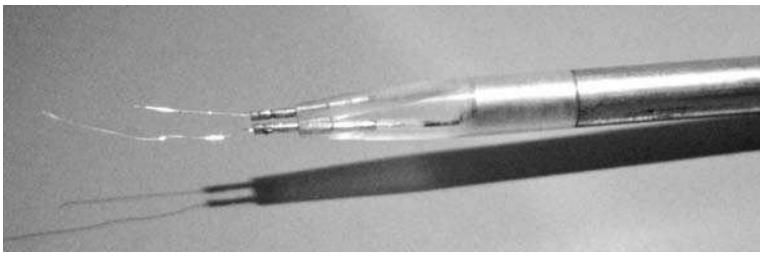


Fig. 3-8 Practical electrodes: thin silver wire electrodes in a holder.

homogeneous solution. In this case, the potential difference should be zero. Since electrophysiological signals are often very small, however, any small deviations of the ideal, or "expected" value of zero constitute a voltage called the "electrode polarization". If no substantial current is fed through the electrodes, it is called the "spontaneous polarization". Since the polarization in this case is a random variable with zero mean, both amplitude and polarity are hard to predict, and can interfere with electrophysiological measurements unless one takes appropriate precautions.

The Silver/Silver Chloride Electrode

In electrochemistry, the reliable, non-polarized standard electrode used for over a century is the Hg/HgCl, or calomel, electrode. In can be made stable and is reproducible to very narrow tolerances necessary for the precision of fundamental, electrochemical research. However, since calomel electrodes contain liquid mercury, they must be handled with care, and kept in the upright position. In addition, electrophysiological measurements do seldom need the precision mentioned above. Therefore, the similar, non-polarized Ag/AgCl electrode is used far more often. Ag/AgCl electrodes can be made from thin silver wires, or bought as sintered pellets

with an attached connection wire. Silver wires may be cut and bent on demand, and may be held in any position. Therefore, in electrophysiology, Ag/AgCl is used almost exclusively for the recording of DC and low-frequency potentials.

Having discussed the Faradaic processes at a metal electrode dipped in one of its own salts, it may seem strange that a metal covered with a solid, insoluble salt is used as a non-polarized electrode. So let us first see:

- why Ag/AgCl is an electrode at all;
- why it is a non-polarized one.

The mode of operation relies on two facts, which answer these questions:

1. In the AgCl crust, silver atoms have a low but useful mobility.
2. The Cl atoms, in contact with a solution containing chloride, can be exchanged just like metal atoms in contact with a salt of that metal.

Thus, the Ag/AgCl electrode can be considered as "a chorine electrode in a salt of chlorine". Here, the anion rather than the cation is the reactive substance that supports an electric current at a virtually constant potential. The potential is determined by the concentration of Cl^- in the solution, and can be assessed with Nernst's equation, yielding the familiar 58 mV per decade. The stability of Ag/AgCl electrodes is served further by the fact that physiological salt solutions have a high and stable chloride content, which is about 150 mM. By the low solubility of AgCl, the concentration of silver ions near an Ag/AgCl electrode is very low, usually about 1 nM. However, other anions in living tissue that can form insoluble salts of silver, such as sulphide, can "poison" an Ag/AgCl electrode. The poisoning effect of sulphide, for instance, can be understood by recalling that the silver sulphide that is formed would try to establish a Nernst equilibrium with sulphide ions (more precisely HS^- or S^-). Since sulphide concentrations are highly variable, the potential of an involuntary Ag/Ag_2S electrode would be unstable. To avoid contamination, the electrode proper is usually screened from body fluids, water, etc. by an electrode holder filled with a clean, filtered KCl solution (called a "salt bridge", and dealt with below).

Non-Faradaic Processes

These processes include adsorption and desorption of ions at the electrode surface as well as structural changes of the substances involved in the electrode reactions. Because of these processes, electrochemical measurements, including electrophysiological ones, depend on *approximations* to the theories presented. Therefore, measured values of electrode potentials, impedances, etc. will often deviate from the predicted values. This situation aggravates when intracellular or extracellular body fluids are involved, especially since their complex ion composition may include proteins and other molecules with intricate chemical properties.

ELECTROKINETIC PROCESSES

The electrical double layer at the metal/electrolyte interface, described above, has a counterpart at the boundary between an insulator, such as glass, and an electrolyte solution. Although it may seem less obvious, the separation of charges at these boundaries leads to several electrical

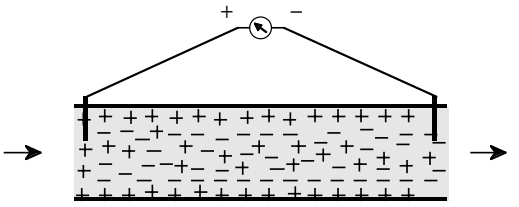


Fig. 3-9 Streaming potential in a salt-solution-filled glass capillary tube.

phenomena, found and described in the nineteenth century. They are collectively called "electrokinetic processes". These phenomena can show up unintentionally in electrophysiological measurements. The existence of an electrical double layer at an insulating surface was concluded from the voltage that develops across a porous plug where an electrolyte solution is forced through it—a process known as a "streaming potential". It is best illustrated with a glass capillary filled with a salt solution and having measuring electrodes at the ends (Fig. 3-9).

A sustained flow of liquid is indicated by the arrows. It was found that, by the nature of the wall and the ions in the solution, one of the ion species is adsorbed to the surface, leaving an "atmosphere" of counter-ions in a narrow zone along the wall. Together, the ions and counter-ions form an electrical double layer, which is analogous to the Helmholtz double layer described for the metal/electrolyte boundary. The movement of the liquid drags the ion "atmosphere" in the direction of the flow, leaving a charge of opposite polarity (slightly) behind. This shows up as a streaming potential, which can be measured with a pair of electrodes at the ends. Strength and polarity depend on which ion species is absorbed, and how strong, and hence on both the wall material and the composition of the fluid. Given these circumstances, the potential difference depends entirely on the speed of flow. Often positive ions adhere to the surface, whereas the negative charge is dragged towards the end of the tube. This is the case shown in the figure. Porous substances, such as cotton wool, paper and some ceramics, can be thought to consist of numerous capillary spaces, and so show the same phenomena. In natural waters, the same process, here called "filtration potential", causes voltage gradients to develop across the sand or clay bottom. The currents associated with these voltage gradients can be perceived by some species of fish that have special, so-called ampullary electric sense organs.

LIQUID JUNCTION POTENTIALS

Often, electrophysiological measurements involve one or more transitions from one solution to another. The most familiar one, known from the chemistry lab, is the transition from the filling fluid of a pH reference electrode to the external medium. In electrophysiology, one has to deal with the transitions from the microelectrode filling fluid to either the cytosol of an impaled cell or the extracellular body fluid or saline. In all these cases, the transition between two electrolyte solutions with different compositions is known as a "liquid junction". Here, a liquid junction potential (l.j.p. also called diffusion potential)* develops, despite the fact that the different ion

* Note that some authors call a membrane potential that arises by passive ion permeabilities rather than by ion-pump activity a diffusion potential. Therefore, the former name is less ambiguous.

species are allowed to diffuse freely across the boundary. The potentials arise because different ion species move at different speeds through the solute, thereby creating a slight separation of charges. The ion mobility, stated in Section 1, is the quantity that determines the magnitude of the junction potential. From the table of ion mobilities of various ions given in Section 1, it can be noticed that K^+ and Cl^- have closely matched ion mobilities. This is one of the main reasons why KCl is so important in electrochemical measurements.

The simplest liquid junction consists of a single salt at two concentrations. In this case, the liquid junction potential E_l follows from the following formula, derived by Nernst:

$$E_l = \frac{RT}{F}\left[\frac{u^+ - u^-}{u^+ + u^-}\right]\ln\frac{c_1}{c_2}$$

where c_1 and c_2 are the two concentrations of the salt, u^+ and u^- are the ion mobilities, and other symbols as defined earlier (p. 109).

As an example, the l.j.p. between solutions of 0.1 M KCl and 0.01 M KCl is 1.12 mV. NaCl solutions cause larger l.j.p.'s, because the mobilities of Na^+ and Cl^- ions are more different from one another: taking 0.1 M versus 0.01 M again, $E_l = 12\,\text{mV}$.

For more complex junctions, the l.j.p. can be computed with the aid of Henderson's formula:

$$E_l = \frac{RT}{F}\left[\frac{\sum_i \left\{\frac{u_i}{z_i}(c_{2i} - c_{1i})\right\}}{\sum_i \{u_i (c_{2i} - c_{1i})\}}\right]\ln\frac{\sum_i u_i c_{1i}}{\sum_i u_i c_{2i}}$$

To approximate the exact l.j.p.'s as closely as possible, the dissociation of water in H^+ and OH^- must be taken into account. Note that the pH of demineralized water and of unbuffered, neutral solutions is about 5 rather than 7, because CO_2 from air is dissociated partially into hydrogen (H^+) and bicarbonate (HCO_3^-) ions.

The table below states l.j.p.'s of a number of junctions that can show up in the electrophysiological practice. The above-mentioned factors have been taken into account.

It must be kept in mind that the computed values are approximations again, mainly because the mobilities are known for very dilute solutions only, whereas in electrophysiology one uses concentrations of 0.1 M and higher.

Solution 1	Solution 2	E_l ($E_1 - E_2$)
3 M KCl	0.1 M KCl	1.66 mV
3 M KCl	0.15 M NaCl	0.89 mV
0.1 M KCl	0.15 M NaCl	−5.4 mV
3 M KCl	1 mM $CaCl_2$	3.68 mV
0.1 M KCl	1 mM $CaCl_2$	1.21 mV
3 M KCl	demineralized water	5.69 mV
0.1 M KCl	demineralized water	40.3 mV
0.15 M NaCl	demineralized water	44.1 mV

From the table, it is clear that, *in general*, the l.j.p.'s involving 3 M KCl are relatively low, whereas lower concentrations of KCl cause relatively higher potentials (the situation given for 1 mM $CaCl_2$ is only an occasional exception: for other concentrations of $CaCl_2$, the situation might be reversed again).

The beneficial influence of high KCl concentrations can be explained by analysing Henderson's formula. The middle factor, involving the sum of concentration differences multiplied by the mobility, causes the salt with the highest concentration to dominate the l.j.p. In other words, choosing a high concentration of a salt where cation and anion have almost equal mobilities helps to keep l.j.p.'s low. This is the main reason that KCl solutions of 3 M or higher are used so often in electrochemical chains containing liquid junctions. By the same argument, KCl is also used as the main filling fluid of glass microelectrode pipettes. Here, a second reason to use concentrations of at least 3 M is that it helps keeping the tip resistance as low as possible. Occasionally, however, the leakage of potassium or chloride from the tip would have an unwanted influence on the cell one is recording from. In these cases, other salts can be used, such as LiCl in cases where potassium must be avoided or KAc (acetate) if chloride must be avoided. Unfortunately, these filling fluids cause higher l.j.p.'s. The mobilities of NH^{4+} and NO_3^- are matched better, but are too poisonous for many tissues.

MEMBRANE POTENTIALS

In general, the object of electrophysiological experiments consists of studying membrane potentials and changes thereof. Stated more precisely, a membrane potential is a potential difference between the inside and the outside of a biological cell.

The magnitude of membrane potentials can be derived from the case of the liquid junction. For the simple case of one permeant ion, the formula of Nernst stated above can be used, taking either u^+ or u^- as zero. The formula then reduces to the familiar one:

$$E_{eq} = -\frac{RT}{zF} \ln \frac{c_{in}}{c_{out}}$$

Here, the potential E_{eq} may be called an "equilibrium potential" because the situation reaches a thermodynamic equilibrium. In addition to the familiar constants, c_{in} and c_{out} are the concentrations, more precisely the activities of the permeant ion at the inside and outside of the membrane, respectively. For single-valued ions ($z = 1$) such as K^+ and Na^+, the Nernst formula amounts to the well-known 58 mV per decade of concentration ratio at room temperature. At the body temperature of warm-blooded animals, the value is 61.5 mV/decade).

Derivation of the Equilibrium Potential

Simple diffusion of an (uncharged) molecular species in one dimension is given by Fick's law:

$$J = -D\frac{\partial C}{\partial x}$$

where J is flux, C concentration, x distance and D the diffusion constant. Electrodiffusion, which describes the diffusion of charged particles such as ions in an electrical potential gradient, takes

an extra voltage-dependent term. With V the electrical potential, u the mobility of the particle and z its charge:

$$J = -D\frac{\partial C}{\partial x} - uzC \cdot \frac{\partial V}{\partial x} \qquad \text{(Eq. 3-1)}$$

According to the Stokes–Einstein equation, D and u are related by:

$$D = u\frac{RT}{F}$$

where R, T and F are familiar constants. Combining the latter two equations gives the Nernst–Planck equation of electrodiffusion:

$$J = -u\left(\frac{RT}{F} \cdot \frac{\partial C}{\partial x} + zC \cdot \frac{\partial V}{\partial x}\right) \qquad \text{(Eq. 3-2)}$$

In equilibrium, the net flux J equals 0 and therefore:

$$\frac{\partial V}{\partial x} = -\frac{RT}{zF} \cdot \frac{\partial C}{C\partial x} = -\frac{RT}{zF} \cdot \frac{\partial \ln C}{\partial x}$$

Integration yields the Nernst equilibrium equation:

$$V_1 - V_2 = \frac{RT}{zF} \cdot \ln\left(\frac{C_2}{C_1}\right)$$

It gives the transmembrane equilibrium potential ($V_1 - V_2$), given the intra- and extra-cellular concentrations of an ion species, C_1 and C_2 respectively.

The Reversal Potential

Equation 3-2 is difficult to solve, if at all, without simplifying it to some extent. Goldman, Hodgkin and Katz have done so by assuming a constant electrical field, i.e. V changes linearly across a slab of width d that is voltage-clamped at the borders.

Hence, $\frac{\partial V}{\partial x} = \frac{V_1 - V_2}{d} = \frac{V_m}{d}$ for all x within the slab, where V_m is the membrane potential. Then Eq. 3-2 becomes:

$$J = -u\frac{RT}{F} \cdot \frac{\partial C}{\partial x} - u\frac{zCV_m}{d}$$

rewriting gives:

$$\frac{\partial C}{\frac{-JF}{uRT} - \frac{zFCV_m}{dRT}} = \partial x$$

which, after integration, yields:

$$\frac{-dRT}{zFV_m} \cdot \ln\left(\frac{\frac{zFV_mC_1}{dRT} + \frac{JF}{uRT}}{\frac{zFV_mC_2}{dRT} + \frac{JF}{uRT}}\right) = d$$

with C_1 and C_2 the concentrations at the borders of the slab. Exponentiation gives:

$$\frac{\frac{zFV_m C_1}{dRT} + \frac{JF}{uRT}}{\frac{zFV_m C_2}{dRT} + \frac{JF}{uRT}} = \exp\left(-\frac{zFV_m}{uRT}\right)$$

The result is rewritten, representing the flux of a single ion species across a voltage- and concentration-clamped slab, as follows:

$$J = -\frac{uzFV_m}{d} \cdot \frac{C_1 - C_2 \exp\left(-\frac{zF}{RT}V_m\right)}{1 - \exp\left(-\frac{zF}{RT}V_m\right)}$$

The current density, I, is obtained from the ion flux, J, by multiplication with $z \times F$:

$$I = -\frac{uz^2F^2V_m}{d} \cdot \frac{C_1 - C_2 \exp\left(-\frac{zF}{RT}V_m\right)}{1 - \exp\left(-\frac{zF}{RT}V_m\right)}$$

The (ion) permeability P is defined as D/d, with D being the diffusion coefficient. Combining the definition with the Stokes–Einstein equation above gives the relation between permeability, P, and mobility u:

$$P = \frac{uRT}{d}$$

Replacing u/d in the flux equation by P_i/RT, where P_i is the permeability of ion species i, and indexing the flux, charge and concentrations give:

$$I = -\frac{P_i z_i^2 F^2 V_m}{RT} \cdot \frac{C_{i1} - C_{i2} \exp\left(-\frac{z_i F}{RT}V_m\right)}{1 - \exp\left(-\frac{z_i F}{RT}V_m\right)}$$

The net current of all ion species combined equals 0 at equilibrium (i.e. at the current reversal potential), or:

$$\sum_i I_i = 0$$

and hence:

$$\sum_i \frac{P_i z_i^2 F^2 V_m}{RT} \cdot \frac{C_{i1} - C_{i2} \exp\left(-\frac{-z_i F}{RT}V_m\right)}{1 - \exp\left(-\frac{z_i F}{RT}V_m\right)} = 0$$

If only the monovalent ions K^+, Na^+ and Cl^- are considered, z_i^2 is always 1. Noting that for anions it suffices to interchange C_1 and C_2 to keep the equation homogeneous with respect to the sign of z in the exponential
because $\frac{C_1 - C_2 e^{a}}{1 - e^{a}} = \frac{C_2 - C_1 e^{-a}}{1 - e^{-a}}$:

$$P_K\left\{C_{K1} - C_{K2}\exp\left(-\frac{FV_m}{RT}\right)\right\} + P_{Na}\left\{C_{Na1} - C_{Na2}\exp\left(-\frac{FV_m}{RT}\right)\right\} + P_{Cl}\left\{C_{Cl2} - C_{Cl1}\exp\left(-\frac{FV_m}{RT}\right)\right\} = 0$$

$$P_K C_{K1} + P_{Na} C_{Na1} + P_{Cl} C_{Cl2} = \{P_K C_{K2} + P_{Na} C_{Na2} + P_{Cl} C_{Cl1}\}\exp\left(-\frac{FV_m}{RT}\right) \text{ or :}$$

$$\frac{P_K C_{K1} + P_{Na} C_{Na1} + P_{Cl} C_{Cl2}}{P_K C_{K2} + P_{Na} C_{Na2} + P_{Cl} C_{Cl1}} = \exp\left(-\frac{FV_m}{RT}\right)$$

After taking the natural logarithm on both sides and inversion of the quotient to eliminate the negation in the exponent, the Goldman equation for monovalent ions, relating the intra- and extra-cellular ion concentrations to the reversal potential, $V_m(I = 0)$, is obtained:

$$V_m = V_1 - V_2 = \frac{RT}{F}\ln\left(\frac{P_K C_{K2} + P_{Na} C_{Na2} + P_{Cl} C_{Cl1}}{P_K C_{K1} + P_{Na} C_{Na1} + P_{Cl} C_{Cl2}}\right)$$

Ion Selectivity

The Goldman equation can be used to determine ion permeability ratios, or ion selectivity, of a given channel. If we wish to determine the permeability ratio of Na over K of a cation channel, the cell containing the channels of interest is perfused both intracellularly and extracellularly with known concentrations of K^+ and Na^+ ions and the reversal potential, V_{rev}, is measured using a voltage-clamp amplifier. Then according to the Goldman equation:

$$V_{rev} = \frac{RT}{F}\ln\left(\frac{P_K [K]_o + P_{Na} [Na]_o}{P_K [K]_i + P_{Na} [Na]_i}\right)$$

If the permeability ratio P_{Na}/P_K is r, then the above equation becomes:

$$V_{rev} = \frac{RT}{F}\ln\left(\frac{[K]_o + r[Na]_o}{[K]_i + r[Na]_i}\right)$$

from which, after exponentiation, r can be easily obtained. A special case is the situation in which only one cation species (say K^+) is perfused intracellularly, while the other (Na^+) is perfused extracellularly at the same concentration, in the absence of other cations.

$$V_{\text{rev}} = \frac{RT}{F}\ln(r)$$

This gives:

$$r = \exp\left(\frac{FV_{\text{rev}}}{RT}\right)$$

Electrodes Sensitive to pH and Other Ions

Artificial membranes, from glass or plastic, may also develop membrane potentials. This has proved useful for ion activity measurement. The so-called glass electrode used for the measurement of pH consists of a thin-walled glass sphere, fitted with an Ag/AgCl electrode and filled with 0.1 N HCl. It was found early in the twentieth century that a glass membrane, basically an insulator, is slightly permeable to H^+ ions, and thus can serve as a pH-sensitive electrode. When the electrode is immersed in a solution of unknown pH, a Nernst potential of 58 mV per pH unit develops, which can be used to measure the pH of the unknown solution.

The pH meter proper is in fact an electrometer (a voltmeter with a very high input impedance), with tick marks reading one pH unit per 58 mV. At pH = 1, the potential across the glass membrane is zero, since the hydrogen ion concentrations on both sides are equal. In this case, the potential of the reference electrode determines whether the potential measured will also be zero. Usually, the reference electrode consists of an Ag/AgCl electrode in a saturated KCl solution. Modern pH meters employ a combination electrode, in which the glass electrode and an AgCl reference electrode are combined into one probe (Fig. 3-10).

This principle of ion-selective-membrane electrodes has been extended to other ions. First, types of glass that have a higher permeability for Na^+ than for K^+, or the reverse, were developed. These can be used as Na-electrodes and K-electrodes respectively, provided that the pH is buffered. This is necessary because the H^+ permeability is still higher (note that in absolute terms, these conductances are very low). Later, several oil-like substances were developed that each show a selective permeability for one ion species. These are known as

Fig. 3-10 A pH electrode with built-in reference (AgCl) electrode.

liquid ion exchangers (LIX) and were developed for a number of ion species, such as hydrogen, potassium, chloride, calcium and magnesium. To be used as ion-selective membranes, such substances are soaked up in thin plastic fabric membranes, or put directly into the tip of a glass micropipette. The latter form allows intracellular ion activity measurements. Since the ion permeability, and hence the conductance, of an LIX is very low with respect to the concentrated salt solutions used to fill micropipettes, the tip resistance of an ion-selective electrode is very high: $10^9\,\Omega$ or more. Ion-selective microelectrodes may have a resistance of even $10^{12}\,\Omega$ or higher! Therefore, only electrometer-grade amplifiers employing MOSFETs can be used to measure the Nernst potentials of these electrodes.

ELECTRODES: PRACTICAL ASPECTS

The Glass Micropipette

Strictly speaking, a glass "microelectrode" is not an electrode at all: it is a salt bridge with a (sub-) microscopic tip diameter. Of course, the pipette fluid is in contact with an Ag or Ag/AgCl electrode, mounted together in a fluid-filled holder fitted with a cable. The whole assembly behaves as an electrode.

Glass micropipettes are manufactured—starting from thin-walled capillary tubes of 1–2 mm diameter—by means of a pipette puller, a device made especially to pull pipettes with different properties to fulfil the needs for different types of recording. Intracellular recording and patch-clamp recording, for instance, demand different shapes and tip diameters.

In general, the structure of a micropipette can be described in terms of shaft, taper and tip (Fig. 3-11, top). The angles of taper and tip vary between designs, but the tip angle is usually only a few degrees of arc. Because of the ultrafine tip, the most conspicuous property of a glass micropipette is its high resistance. Because of the conical shape, virtually all resistance resides

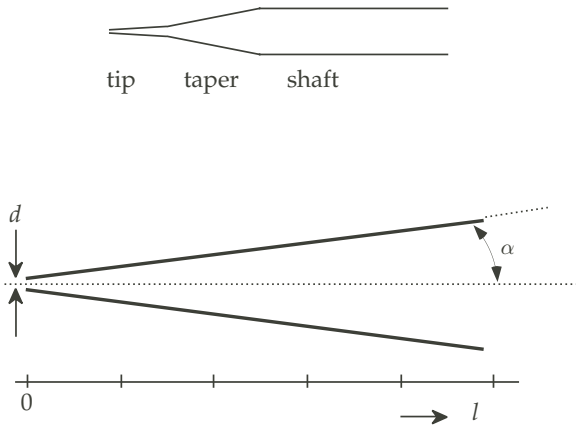


Fig. 3-11 General outline of micropipette and derivation of tip resistance.

in the tip. This can be computed with the following formula, which expresses the (partial) resistance R_{tip} of the pipette tip up to a distance of l from the tip.

$$R_{tip} = \frac{\rho}{\pi \cdot \tan^2(\alpha)} \left(\frac{2\tan(\alpha)}{d} - \frac{1}{l + \frac{d}{2\tan(\alpha)}} \right)$$

The resistivity of the filling fluid is designated as ρ (rho), the tip angle as α and the tip diameter as d. The situation pertaining to the formula is illustrated in Fig. 3-11, bottom.

A graph of this formula is shown in Fig. 3-12. Figure 3-13 shows a practical form of micropipette and holder.

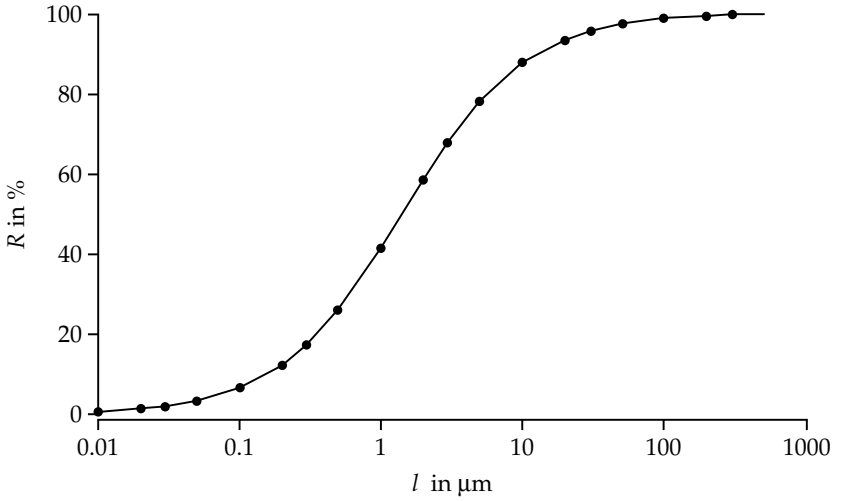


Fig. 3-12 Graph of (relative) resistance contribution versus length.

Fig. 3-13 A glass micropipette in a holder. The thin side tube allows to apply pressure (tracer injection) or suction (membrane patch attachment).

Apart from the potential of the Ag/AgCl electrode and the tip resistance, both of which can (and must) be measured in each case, the wall of a glass microelectrode shows capacitance that behaves as a stray capacitance shunting the signal to a certain extent. The capacitance value depends on the wall thickness and the perimeter, and since both thickness and perimeter are reduced towards the tip, the capacitance per unit length is approximately constant.

Finally, glass micropipettes, in general, annoyingly develop the so-called tip potential. This is an added potential jump at the tip that occurs often after impalement of a cell, and that is thought to arise by the partial clogging of the tip with large molecules such as proteins or cell fragments. Tip potentials are very capricious, and can amount to 30 mV. They develop only at tip diameters below about 0.5 μm, and so can be assessed after an experiment by breaking off the tip to a diameter well over this size.

Patch Electrodes

The procedure to make patch electrodes is very similar to pulling microelectrodes. Often a multi-step electrode puller is used such that the taper of the pipette is short, which makes it easier than long-tapered pipettes to back-fill and to remove air bubbles. The pipettes are made on the day of the experiment. If they are kept too long it is difficult, if not impossible, to make good seals with the cell membrane, possibly due to the slow hydration of the glass. After pulling, the wall of the tip may be enlarged for improved sealing using a microforge. This step is not essential and is often omitted.

As the tip of the pipette is relatively large (up to a few microns), ions and small molecules in the the pipette solution interchange freely with the cell contents. For small fluorescent molecules like fluoresceine or Fura, it takes only a few seconds after rupture of the cell membrane to spread from the soma of a cerebellar purkinje cell to far into the dendritic arborisation. Therefore, the patch pipette is filled with an iso-osmotic solution (typically 150 mM KCL, 1 mM Mg^{2+}, a pH buffer such as HEPES and a Ca^{2+} buffer such as EGTA). If filled with this solution or a solution of equal ionic strength, patch pipettes should have a resistance between 1 and 10 MΩ, depending on the size of cell to patch. The pipette resistance can be estimated by applying air pressure to the interior while keeping the pipette tip in 95% methanol. The pressure required to expel air bubbles from the tip increases with electrical resistance. It takes a pressure of 2 Bar to expel air bubbles from a 5 MΩ pipette. The presence of divalent ions (usually Mg^{2+}) in the pipette solution is important to obtain and maintain a seal.

When preparing pipettes for excised patches, it is in general necessary to reduce capacitive currents (which are in the order of nA) in order to be able to measure the unitary currents of interest (which are in the order of pA). This is done by coating the glass pipettes, excluding the last 100 μM close to the tip, with a resin such as Sylguard. The syrupy liquid is easily applied to the electrode using a fire-polished "Pasteur" pipette and then hardened by heating with a microforge. The same procedure also reduces noise in the recording due to random movements of charges in the glass (see also Chapter 2).

Subsequently, the patch electrode is filled in two steps. First, the tip is filled by suction, usually by fixing the pipette on a syringe and pulling the piston. Second, the pipette is back-filled using either a syringe needle or a plastic pipette tip that is pulled over a Bunsen burner to a diameter that can enter the glass patch pipette and reach down to the taper. Air bubbles are then removed by tapping lightly on the pipette with a fingernail. A more sophisticated

way to remove air bubbles is to pass a bolt along the side of the pipette in a kind of "sawing" movement, the pitch of the bolt causing the pipette to vibrate.

The Semi-Permeable Patch

One of the advantages of the whole-cell suction pipette technique is that the ion composition of the medium of the cell interior is well defined, as ions and molecules in the cell equilibrate rapidly with the pipette contents. Although it is a benefit to be able to control ion gradients over the plasma membrane, the washout of the enzymatic machinery of the cell and its second messengers is not always a blessing. It may be especially a nuisance if one intends to study the regulation of ion channels by intracellular factors. Washout of larger molecules, such as proteins, may be delayed by using pipettes with small tips, at the same time degrading the voltage-control. Even then, smaller molecules, such as ATP and cAMP, are still rather effectively removed from the cytosol.

To overcome this difficulty, pore-forming proteins may be included in the pipette solution (Fig. 3-14, left). Having made a seal with the plasma membrane (the "cell-attached" configuration), these proteins gradually insert themselves into the membrane allowing electrical contact and exchange of ions with the cell interior, while keeping larger molecules inside the cell. The molecular cut-off size depends on the choice of the pore-forming protein. The α-toxin from *Staphylococcus aureus* forms pores that let pass ions and molecules up to 1000 Da. Digitonin permits the exchange of structures, which are of the size of a mitochondrion. The most popular pore-forming substance amongst electrophysiologists is nystatin. It conducts monovalent ions, but bars (at least to a large extent, if not fully) the way to Ca^{2+}. Nucleotides cannot pass the nystatin pore either. Unfortunately, nystatin is not easy in its use. It is sensitive to light and almost insoluble in water. Moreover, it prevents formation of a seal. Therefore, the tip of the pipette should be filled with a solution devoid of nystatin and back-filled with a freshly sonicated solution containing (typically 100 μg/ml) nystatin. Then the seal is made (and microscope illumination switched off), after which nystatin slowly diffuses down the

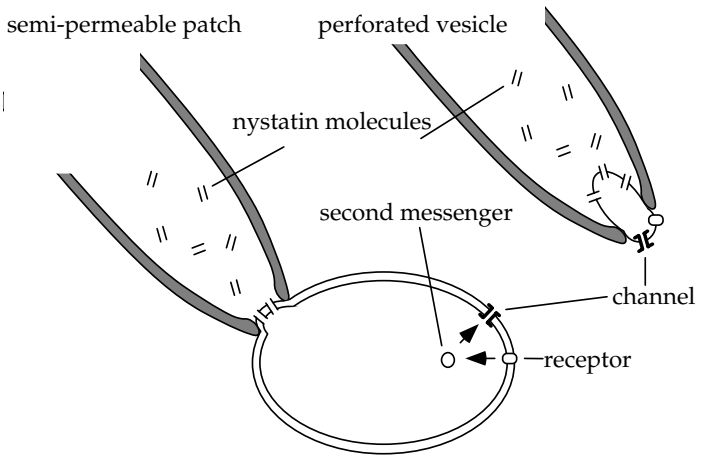


Fig. 3-14 Two methods for making semi-permeable patches.

tip and integrates itself with the patch membrane. Its integration may be followed by monitoring the membrane capacitance that should increase during a period of about 20–30 min. and then remain stable. An interesting variation on the semi-permeable patch technique is the perforated-vesicle (Fig. 3-14, right). Rather than making an inside-out patch, a vesicle is excised from the cell using a nystatin back-filled pipette. This results in a configuration similar to the inside-out patch, but with cytosolic constituents concentrated close to the plasma membrane (Levitan and Kramer, 1990). This technique is not for the impatient or the weak of nerve.

Ground Electrodes

Any ground electrode needs to have a low impedance, and so they are usually made relatively large. Sometimes, a large Ag/AgCl pellet is used to ground the preparation, but more often, a metal wire, plate or wire mesh serves as the ground connection (Fig. 3-15). The use of naked metals may seem odd at first sight, since we saw earlier that metal surfaces are polarized electrodes that generate relatively large and variable polarization voltages. However, sufficiently large metal electrodes have fair DC properties, and this can be understood by the following argument. As we discussed earlier, small metal surfaces have random potentials, centred around a certain mean, or expected value. Since the deviations from the mean are random variables, the average value of a number of such potential values will be closer to the mean than the individual values.

Since a large electrode surface area can be considered to be the sum of a large number of smaller areas, a large metal plate has a lower and more stable polarization potential than a smaller one. In addition, a silver surface will be chlorinated spontaneously during the usage with chloride-containing solutions, causing the DC properties to improve over time. This effect is also used in the "ground" electrode of a building, which consists usually of a long copper tube or pole, dug deeply into the ground (below the subsoil water level). For very critical DC recordings, a non-polarized ground electrode is still recommended, however.

Fig. 3-15 A silver strip ground, electrode in an insect preparation tray.

VOLUME CONDUCTION: ELECTRIC FIELDS IN ELECTROLYTE SOLUTIONS

Sources of electric current in water or salt solutions give rise to so-called stationary electric fields. The theory of electricity was developed in the eighteenth century with static forms of electricity. Charges in space, maintaining electrical potential differences, were kept away from leaking by dry air and by insulators like glass or ceramics. In a conducting medium, however, an electric current would flow and discharge any static source within a fraction of a second. Therefore, to generate an electric field in a fluid, we have to supply current continually (DC and/or AC). The mathematical descriptions of electric fields in water and in an insulating space are similar, but in a conducting medium, current sources replace the charges in a static field, and current densities replace field strengths.

Three main forms of stationary electric field can be distinguished, and can be used or encountered in the electrophysiological practice:

1. homogeneous field,
2. monopole field, and
3. dipole field.

These are illustrated in Fig. 3-16, and will be described briefly below.

Homogeneous Electric Field

A "homogeneous field" arises when an electric current is fed through two parallel electrodes that cover the short sides of an elongated tray. This configuration, sometimes called an electrolytical trough, is shown in Fig. 3-16A. It can be used to illustrate the quantities used to describe electric fields in conducting fluids. Let us assume that we feed a direct current (of strength I) through the trough from right to left, by making the right electrode positive with respect to the left one. Because of the constant cross-section of the tank and because the electrodes fit the short sides entirely, current flows in straight lines, parallel to the long sides of the trough. Since the properties of the field are the same throughout the tank, it is called a homogeneous field.

The quantities that describe the electric field in the liquid are analogous to the quantities that describe current through an object, viz. an electrical resistance, but expressed per spatial unit. In analogy to resistance R, we have the resistivity ρ ("rho"), which is usually expressed in Ωcm. In analogy with current, we have the current density J, expressed in A/cm^2. Finally, the voltage gradient, grad(V), expressed in V/cm, replaces the voltage. Resistivity, current density and voltage gradient, thus, describe the properties of a unit cube, which has $R = \rho$, $I = J$ and $U = \text{grad}(V)$. Obviously, these quantities are related via an analogue of Ohm's Law:

$$U = IR \quad \leftrightarrow \quad \text{grad}(V) = J\rho$$

To describe the potential, we need a reference value. In a practical circuit, one of the electrodes would be grounded, but theoretically, it is more attractive to take zero at the middle of the trough (note that otherwise the electrode polarization would introduce an error). Thus, the

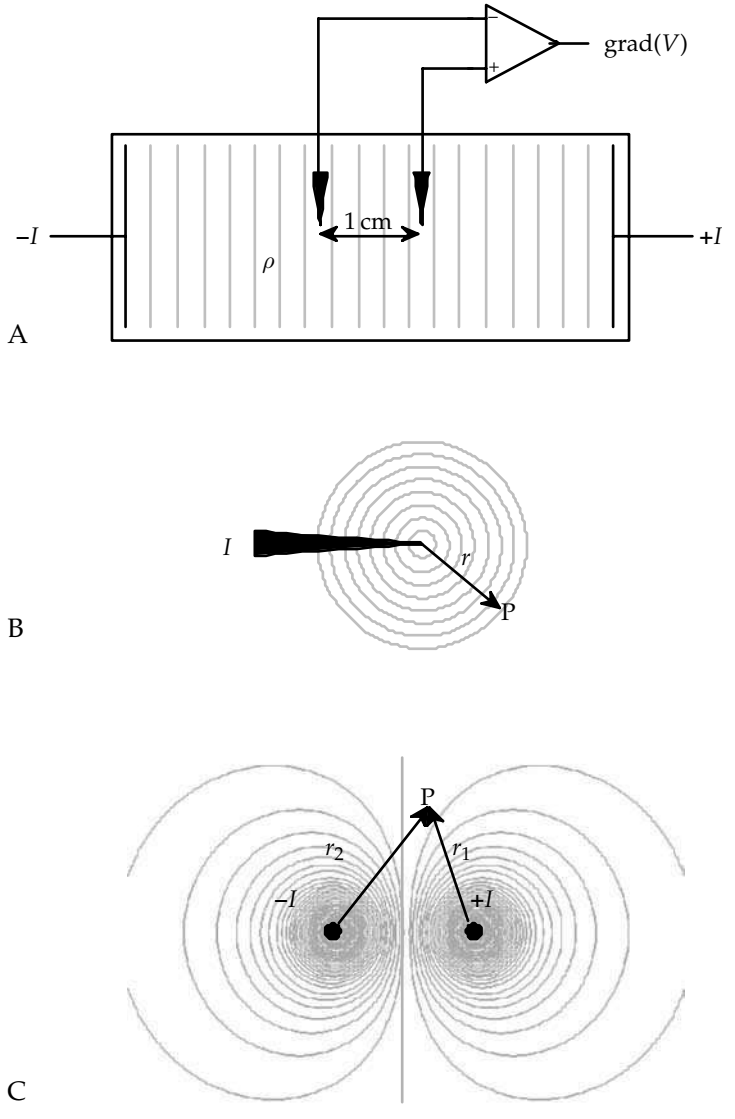


Fig. 3-16 Homogeneous, monopole and dipole fields.

potential (E) can be determined as the voltage between a measuring electrode and a central reference electrode, l cm apart, and follows from:

$$E = l\,\mathrm{grad}(V) \quad \text{or} \quad E = J\rho l$$

Lines connecting points of equal potential, or "equipotential lines", are parallel to the electrodes that generate the field. The homogeneous electric field can be used to test and calibrate one's measurement set-up.

Monopole Field

The situation at an isolated electrode, or monopole, is completely different. Theoretically, a monopole field arises, when the reference or ground electrode is situated at infinity (and is infinitely large). In practice, to approximate a monopole field sufficiently, the ground electrode has to be far away only with respect to the distances involved in the measurements. For example, the field around a point-shaped electrode in a 10 cm × 10 cm × 2 cm tray is sufficiently monopole-like over a radius of at least a few millimetres. The electrode itself must, theoretically again, be infinitely small, i.e. point-like. If one uses the tip of a glass micropipette, the monopole situation is sufficiently approximated, even close to the tip.

The shape of a monopole field is spherical, with the current I emanating radially from the tip, and grad(V) and J decreasing with increasing area, hence with the square of the radius r. For any point P, current density and voltage gradient follow from:

$$J_P = \frac{I}{4\pi r^2} \quad \text{grad}(V)_P = \rho J = \frac{\rho I}{4\pi r^2}$$

Taking the potential of the ground electrode (at infinite distance and infinitely large) as zero, the potential in P follows from:

$$E_P = \frac{\rho I}{4\pi r}$$

Equipotential lines are shown in Fig. 3-16B. Like the homogeneous field, a monopole field can be used for test purposes, provided one is able to estimate the distance r between the pole and a measuring electrode.

Dipole Field

A "dipole" can be considered as two monopoles, the first of which carries a current $+I$. Since the other one "absorbs" the same current, it carries a current $-I$. The electrodes are also known as "source" and "sink" respectively.

To compute the voltage gradient, current density or potential at any point, the contributions of both poles simply sum, taking the respective distances r_i ($i = 1, 2$) into account.

Therefore, current density, voltage gradient and potential at any point P follow from:

$$J_P = \frac{I}{4\pi r_1^2} + \frac{-I}{4\pi r_2^2}$$

$$\text{grad}(V)_P = \rho J = \frac{\rho}{4\pi}\left(\frac{I}{r_1^2} + \frac{-I}{r_2^2}\right)$$

$$E_P = \frac{\rho}{4\pi}\left(\frac{I}{|r_1|} + \frac{-I}{|r_2|}\right)$$

This leads to the pattern shown in Fig. 3-16C.

Since in the middle of the poles the contributions of the two poles cancel, the line through the centre of the dipole, perpendicular to the line connecting the poles, has zero potential. This mid-line is drawn in Fig. 3-16C. Close to the electrodes, the equipotential lines are almost

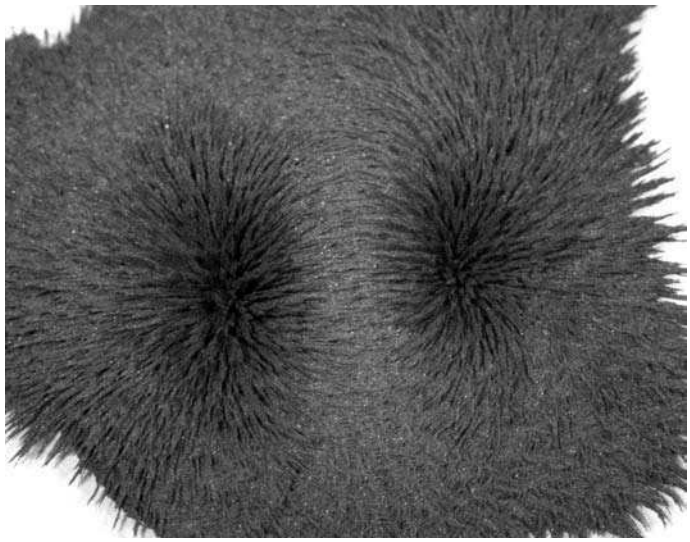


Fig. 3-17 Iron filings showing magnetic fieldlines of a (hidden) bar magnet.

circular (monopole-like) and too closely spaced to be distinguished in the figure. The formulas can be extended to any number of poles distributed through space.

Dipole fields arise in a number of situations, of both natural and technical origin. However, because of their complex form, dipole fields are less suited for testing and calibration purposes. Note that, properly speaking, current density and voltage gradient are vectorial quantities, having both a magnitude and a direction. For simplicity, the values given here are the maximum values that pertain to the mentioned directions. In directions perpendicular to the stated ones, both current density and voltage gradient are zero. To be more precise, they follow a cosine law: as an example, the current density at an angle of 45° to the direction of the field lines, in the abovementioned homogeneous field, would be $J\cos(45°)$, or about 0.71J.

The shape of the dipole field is best known from that of a bar magnet. In fact, the electric and magnetic dipole fields share a common mathematical form. In Fig. 3-17, the magnetic field lines of a hidden bar magnet are made visible by the layer of iron filings.

This description of the spatial distribution of electric currents in water ends our discussion of electrochemical processes. Further study remains necessary, however. The literature listed in Appendix I treats much of the material presented here in greater depth or detail, and is meant as a guide to this end.

4

Signal Analysis

INTRODUCTION

In general, electrophysiological signals have to be processed and analysed. However, there are many types of electrophysiological signals and recording techniques, so the methods to process and analyse them will be equally diverse.

For analysis, signals can be subdivided into the following broad categories: intracellular voltages, extracellular voltages and transmembrane currents. One of the signals most frequently used and processed is the action potential, or spike.

Analysis might deal with the shape of the voltage or current, i.e. the amplitude and time course of a sensory or postsynaptic potential, an action potential, or the opening and closing of an ion channel. Such signals are called analogue signals, in contrast with digital (finite-precision numeric) signals. Alternatively, for many purposes, each action potential may be considered as a point process, of which size and shape are taken for granted. Here, all aspects of the distribution in time are the all-important quantities. One might be interested in the average spike frequency, in stimulus-related occurrences of spikes, in simultaneous occurrences of spikes and so on.

As an example, a photoreceptor cell stimulated with light flashes responds by a depolarization, which may be picked up by an electrode in (or near) the cell. The nerve fibre conducts a varying spike rhythm, which may be recorded too. Both signals need specific processing methods. This is illustrated in Fig. 4-1, where the receptor potential is averaged, and the spike series is processed into a so-called instantaneous frequency plot.

We will start with the most straightforward analysis techniques, applicable when both input and output of the structure under study can be considered as continuous functions of time, like the light intensity and the receptor potential in the above example. Specialized methods to analyse an action potential series (colloquially called a "spike train") and methods to analyse the opening and closing statistics used to characterize ion channel types will be treated later on.

ANALYSIS OF ANALOGUE POTENTIALS

Systems Analysis

We may define what we call a "system" as some part of the real world that is defined by the input(s) and output(s) studied. Systems in general may have multiple inputs and outputs. For simplicity, however, we will deal here only with systems having one input and one output.

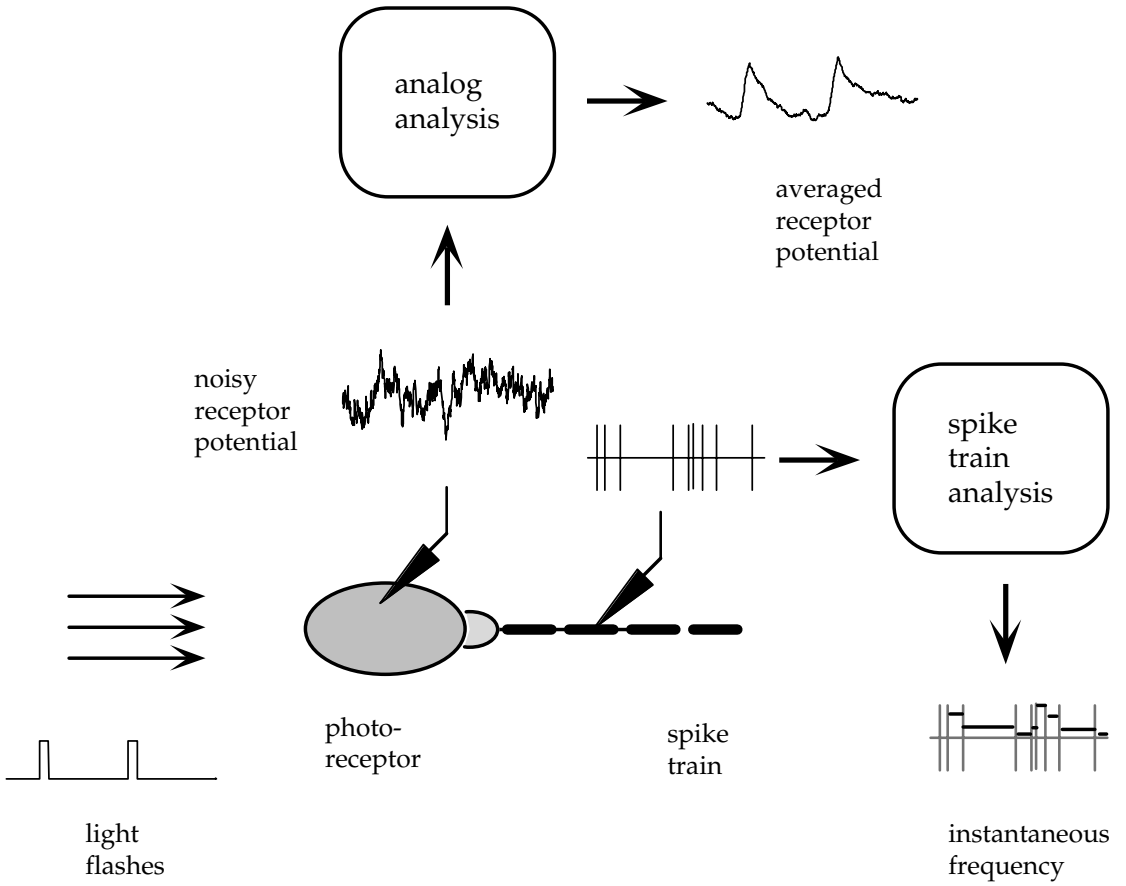


Fig. 4-1 Signal analysis summarized for a photoreceptor preparation.

The standard procedure of what is called a "systems analysis" is to feed an input signal into the system, record the output signal, and try to find a relationship between the two. The found relation is called a "model" of that system. In everyday language, hearing the notion of a model, we think of a scale model: reduced in size, such as a ship model; or sized up, like the models of atoms and molecules. In systems analysis, the model may be a physical one, such as an electronic filter that mimics some property of a neuron, or, more generally applicable, a mathematical model. The corresponding mathematical theory is called "systems theory"; we will discuss some of its basics later on.

The methods of systems analysis are by no means limited to electrophysiology, or even to physical science: the principles are equally useful for technical systems, in ecology or economics. To imagine the latter case, an input might be a change in interest rate, and the observed output the development of investments.

The principle of systems analysis is illustrated in Fig. 4-2, left side.

The arrow from system to model reflects that the model is based on the found properties of the system. Once a model has been formulated, it can be treated like the real system: feed it with an input signal, record the output. In most, if not all, cases, however, the model will be

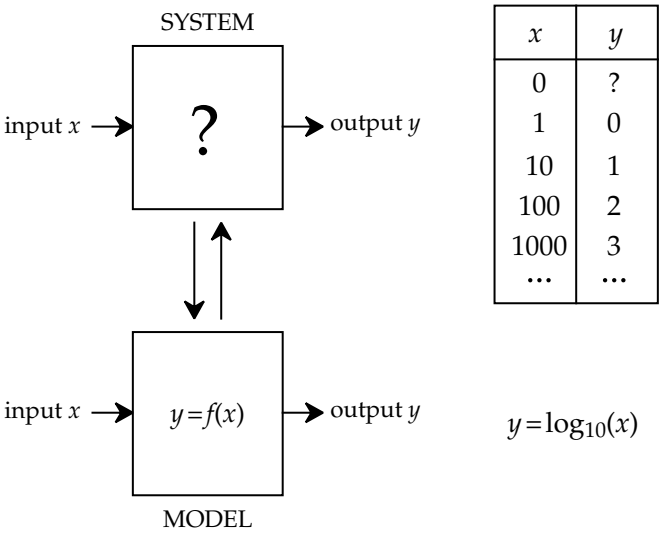


Fig. 4-2 Systems analysis principle (left) and an example result (right).

neither sufficiently precise nor complete, which necessitates going back to study the real-world system. Therefore, a second arrow points back from model to system.

As an example of the systems-analytical practice, consider the case of the photoreceptor cell again. The cell has been stimulated with light flashes of different intensities, while measuring the amplitude of the cell's response. The outcome could be as depicted in the short table of Fig. 4-2: at a stimulus intensity of 10 lux, the potential is 1 mV; at 100 lux it is 2 mV, and so on. This simple example is chosen to illustrate two main points:

1. The relation between the input and output quantities is derived tentatively from the results of an experiment, unless there is a fundamental reason to pick a specific mathematical function. Here, a logarithmic relation is an obvious choice, but its success is never warranted.
2. Once a model has been formulated (here $y = \log_{10}(x)$), it leads to predictions of results that would be obtained under circumstances different from the ones used in the experiment. If the model equation leads to unlikely results, the model must be replaced or refined by new measurements. In our example, the log relationship would lead to an impossible situation at zero illuminance, where the model predicts log(0). Returning to the real system, the actual response in the dark (stimulus zero) might give us the necessary correction to the model equation.

This ends our simple example. Because we compare the amplitudes of the input and output signals, this form of analysis is said to be performed in the "amplitude domain". This is useful by itself, but systems theory is centred on time functions. The reason for this will become clear soon.

A crucial notion is that, in all but a few trivial cases, the input and output signals change in time, and the details of the time behaviour reflect the most important aspects of functioning. Therefore, systems theory is centred around the analysis of time functions, which is said to occur in the "time domain". As we will see later on, the inverse of time, i.e. frequency, is equally important, and hence much of the theory will deal with the "frequency domain".

If we analyse a system, we may make some change, such as a raising the temperature, a light flash, an electrical pulse, and so on. All of them are functions of time, and so the input signal is given the symbol $f(t)$. The output is also a function of time, and is called $g(t)$ (the g being simply the letter following f in the alphabet). This is illustrated in Fig. 4-3.

Systems analysis is the technique to get information about the action of a system, expressed as a relation between input and output. This system property will also be a function of time, and is called the "transfer function" $h(t)$. The transfer function reflects an important property of any system, viz. the way it responds to changes of the input, hence of some circumstance or quantity in the environment. Therefore, systems analysis is sometimes called "dynamic systems analysis". If we would choose a simple, static component like a resistor as a system, the response to any input would follow from Ohm's law, and hence would be rather uninteresting. However, most real-world systems, be it physical, chemical or biological, show rather complex dynamical behaviour, and so will need some effort to analyse and describe.

Thus, $g(t)$ is some relationship between $f(t)$ and the system behaviour $h(t)$. This relation will prove to be a so-called convolution, a mathematical term that needs further explanation, the symbol © stands for it.

Convolution

To explain the process of convolution, we will use the well-known low-pass RC filter.

If a very short pulse is fed into this circuit at $t = 0$, the output jumps to some value, taken here to be unity for simplicity. Immediately after the pulse, the capacitance starts to discharge, causing the output to decrease in an exponential way: $g(t) = 1 - e^{-t/\tau}$.

The pulse is "forgotten" slowly, and the longer you wait, the less the output reflects the original pulse strength (Fig. 4-4). Principally, the pulse never dies out entirely, but in practice, after some time, we consider the output to be zero again.

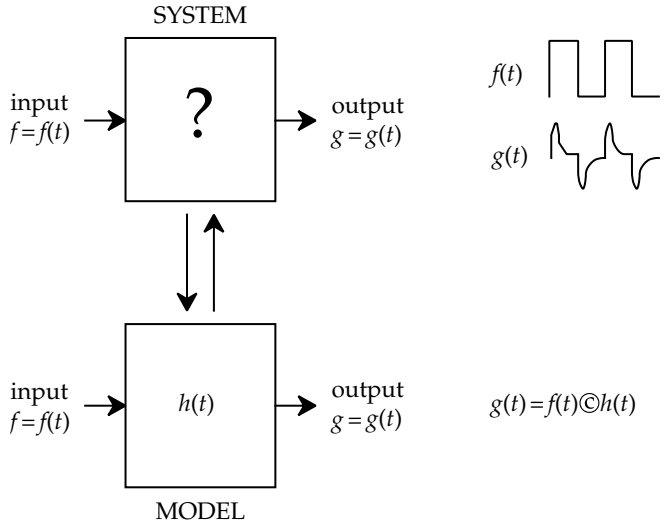


Fig. 4-3 Systems analysis principle for time functions (left); the system function $h(t)$ (lower left) and example input and output functions (upper right).

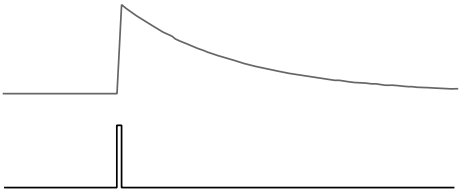


Fig. 4-4 An input pulse and the response of an RC filter.

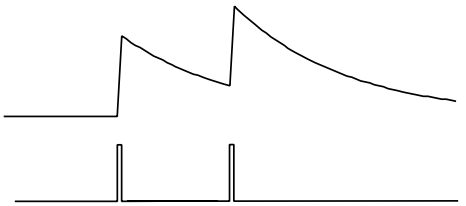


Fig. 4-5 The response to two successive input pulses.

If we administer a second pulse, the same process repeats, but in this case the output voltage is the sum of the remnants of both pulses, taking their respective times of occurrence into account (see Fig. 4-5). The summation of input amplitudes can be generalized to apply to any input signal, continuous in time. The output is the sum of all input signals, taking their times of arrival into account. For continuous time functions, the sum changes into a time integral.

Apparently, we have two kinds of time functions interwoven, or "convolved": the input signal, and the "memory" function of the RC filter.

To compute the output, we need to reckon with both time functions at once. This leads to the so-called convolution integral:

$$g(t) = \int_{-\infty}^{t} f(t) \cdot h(t-\tau)\mathrm{d}\tau$$

This is a mathematical way to express the way in which two time functions are "sliding along one another". Since there are two time functions, the integral comprises two time variables, called t and τ. The regular, or "real time", variable t is the instant at which we want to know the output $g(t)$, but this single value depends on the whole history, i.e. the time integral time from $-\infty$ to t. This is the meaning of the integration variable τ.

To get the flavour of convolutions, imagine two pieces of cardboard, each with a rectangular hole in it, stacked on top of each other, and positioned between the observer and a lamp (Fig. 4-6, left). In this situation, the total amount of light that will pass the cardboard barrier is determined by the amount of overlap between the two holes. If we slide the first piece of cardboard along the second one horizontally (x direction), the amount of light grows slowly until the overlap is maximal, and then diminishes again.

The amount of light $\boldsymbol{L}$ is determined by the convolution of the two hole functions. If the two rectangles are identical, the convolution is a triangle (top right figure). If one of the holes is wider than the other, there is a region of "slack", in which the light output remains constant. In that case, the convolution is trapezium shaped (bottom right).

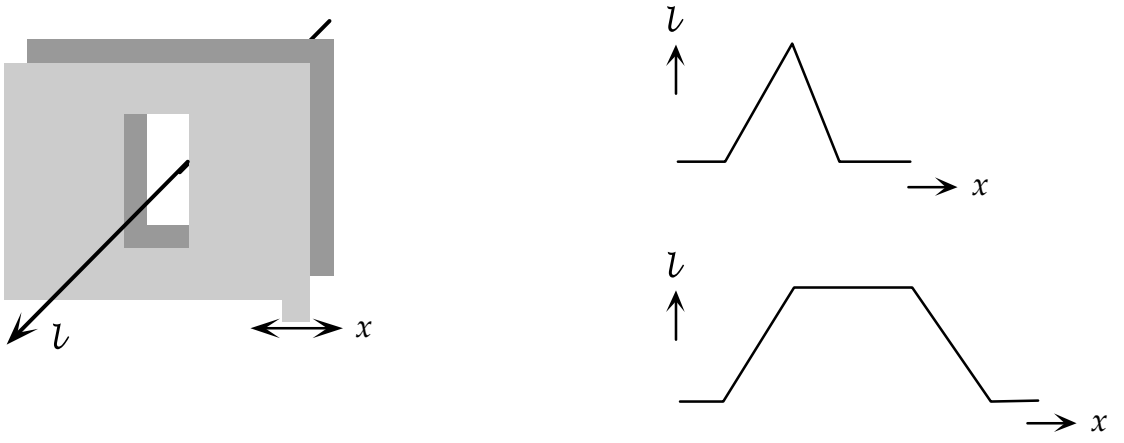


Fig. 4-6 Convolution explained with a light ray and sliding cardboard masks.

The process of convolution is used so much in systems analysis that it is given a separate symbol. Originally, this was the asterisk (the symbol *), but since the advent of the computer, the asterisk is often used as multiplication sign (to prevent confusion with the letter x). Therefore, we have chosen the © symbol here.

In the case of low-pass filter, the convolution integral that has to be determined is between the input signal $f(t)$ and the filter property $h(t)$. It will be notated as follows:

$$g(t) = f(t) © h(t)$$

Unfortunately, in all but the most simple cases, derivation of convolution integrals is tedious, even for people with sufficient mathematical skills. And things get still worse: to derive the transfer function of a system, the measured output signal and the input signal must be deconvolved.

$$h(t) = g(t) ©^{-1} f(t)$$

Here, the symbol $©^{-1}$ stands for a deconvolution integral, which is the inverse operation of the convolution.

An attractive solution is to transform the functions into the frequency domain. In that case, a convolution is replaced by a simple multiplication, a deconvolution by a division.

How does one transform an entire time function into the frequency domain? Transforming a single number from time domain to frequency domain is simple, of course: inversion. If an alternating current has a period of 20 ms, its frequency is 1/0.02, or 50 Hz. The inverse transform, from frequency domain to time domain, is identical: inversion!

Here, however, we have to transform complete functions or signals between two domains, and so we need other mathematical tools.

These tools, or mathematical procedures, are called integral transforms. The two most frequently used transforms are due both to eighteenth-century mathematicians: Laplace and Fourier. These will be explained below. Note that the convention used to distinguish time domain functions from their corresponding frequency domain counterparts is to notate time functions with lower case letters (f, g and h), and frequency functions with capitals (F, G and H respectively).

For both transform types, an inverse form exists that serves to transform functions backwards, i.e. from frequency domain to time domain. These inverse transforms are almost identical to their respective forward counterparts. Although the Laplace and Fourier transforms resemble one another, their respective uses differ in practice. "Laplace transforms" are used mostly in analytical (i.e. algebraic) work, whereas the "Fourier transform" is used mostly in numerical analysis (i.e. calculations). This is shown in the table below.

	Laplace Transform	Fourier Transform
Mathematical form	$F(\mathrm{s}) = \int_0^\infty f(t)\,\mathrm{e}^{-st}\,\mathrm{d}t$ $s = \sigma + \mathrm{j}\omega$	$F(\mathrm{j}\omega) = \int_{-\infty}^{\infty} f(t)\,\mathrm{e}^{-\mathrm{j}\omega t}\,\mathrm{d}t$
Signals	transient and periodical	periodical
Time axis	$0 < t < \infty$	$-\infty < t < \infty$
Practical application	analytically (determining transfer functions)	numerically (computation of frequency characteristics)
Resources	table of Laplace transforms	computer with FFT (Fast Fourier Transform) algorithm

We will discuss both transforms in more detail below.

The Laplace Transform

The procedure for the application of ***L***aplace transforms to derive the output of a system from the input and transfer functions is as follows:

$$\begin{array}{ccc} f(t) & h(t) & g(t) \\ \Downarrow & \Downarrow & \Uparrow \\ \boldsymbol{L} & \boldsymbol{L} & \boldsymbol{L}^{-1} \\ \Downarrow & \Downarrow & \Uparrow \\ F(s) & H(s) & G(s) \end{array}$$

Of course, usually the system transfer function is the unknown, and must be derived from the input and output functions:

$$\begin{array}{ccc} f(t) & h(t) & g(t) \\ \Downarrow & \Uparrow & \Downarrow \\ \boldsymbol{L} & \boldsymbol{L}^{-1} & \boldsymbol{L} \\ \Downarrow & \Uparrow & \Downarrow \\ F(s) & H(s) & G(s) \end{array}$$

The inverse Laplace transform, $\boldsymbol{L}^{-1}$, is often omitted, because we are satisfied to have the frequency domain transfer function, $H(s)$. This is called the "frequency characteristics" and renders the system properties in an intelligible way.

At this stage, one may wonder whether the situation is improved at all: to replace the problematic convolutions and deconvolutions by more simple operations, we have to perform a number of forward and/or inverse Laplace transforms. These transforms are likewise complex mathematical procedures. At a first glance, this action does not seem to help at all. The improvement, however, stems from two facts:

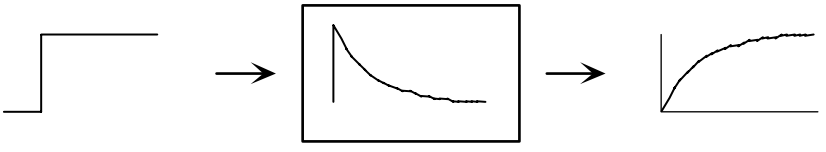
1. Many Laplace transforms have been derived already, and are tabulated in books. This means that the task of deriving a specific Laplace transform reduces to finding it in the tables.
2. Since, in the frequency domain, functions can be combined using multiplication and division rather than convolutions and deconvolutions, complex functions not in the tables can be split up often into component functions that are tabulated. In a way, this compares to an artist's palette, on which the variety of colours needed can be derived from a small number of primary colours.

Below is a short list of Laplace transforms, necessary for the example illustrated subsequently.

Note that the variable of a Laplace-transformed time function is denoted as s. This variable is called "complex frequency". The meaning of "complex numbers" in general and complex frequency is treated in Appendix E. For the moment it suffices to consider s as the equivalent of frequency (in Hz). The variable t stands for time (in seconds), whereas a and c are constants.

Description	Time domain function	Frequency domain function
Ramp	$f(t) = ct$	$F(s) = \frac{1}{s^2}$
Step	$f(t) = c$	$F(s) = \frac{1}{s}$
Impulse	$f(t) = \delta$	$F(s) = 1$
Low-pass RC filter	$h(t) = \mathrm{e}^{-\mathrm{at}}$	$H(s) = \frac{1}{s+a}$
Output, see text	$g(t) = 1 - \mathrm{e}^{-\mathrm{at}}$	$G(s) = \frac{1}{s(s+a)}$

Example: a step input to a low-pass RC filter.



$$
\begin{array}{ccccc}
f(t) = 1 & (\copyright) & h(t) = \mathrm{e}^{-\mathrm{at}} & (=) & g(t) = 1 - \mathrm{e}^{-\mathrm{at}} \\
\Downarrow & & \Downarrow & & \Uparrow \\
\boldsymbol{L} & & \boldsymbol{L} & & \boldsymbol{L}^{-1} \\
\Downarrow & & \Downarrow & & \Uparrow \\
F(s) = \dfrac{1}{s} & (\times) & H(s) = \dfrac{1}{s+a} & (=) & G(s) = \dfrac{1}{s(s+a)}
\end{array}
$$

Thus, the well-known output of a low-pass filter to an input step is derived here by means of a multiplication in the frequency domain of the step input and the filter transfer function.

Another interesting property follows from the transform table above, namely that multiplying by s is equivalent to differentiation, and division by s is equivalent to integration of a function. This is illustrated by the frequency domain function of a ramp (a linearly rising voltage). The step function, $f(t) = \text{constant}$, is just the derivative of the ramp, and $1/s$ is $1/s^2$ multiplied by s. The next stage, comparing the step function with the impulse (or delta) function, may need a further explanation. A step is a sudden change of voltage (or any other quantity). At a certain moment, the quantity is supposed to jump (instantaneous, or infinitely fast) to a different value. In other words, the derivative is an infinitely high change in an infinitely short time. This function, called the "impulse", or "delta function", is a very useful mathematical notion, notwithstanding the fact that is not realizable physically. In practice, however, the impulse can be approximated by a finite (but very short and very high) rectangular "pulse".

The beauty of the impulse as an input signal lies in the frequency domain—here it is unity: $F(s) = 1$. A spectrum that has the same strength at all frequencies is called a "white" spectrum.

Sending an impulse into any system is equivalent to a stimulation with all frequencies at once. The result is that the output signal $G(s)$ is identical to the system response $H(s)$. In the time domain, if the input signal is an impulse, $h(t) = g(t)$. In this case, the response $g(t)$ is called the impulse response. The attractiveness of an (approximated) impulse lies not only in the identity of output and system transfer function, but also in the "whiteness" of the spectrum: instead of stimulating patiently with a number of sine stimuli of different frequencies in succession and collecting the response amplitudes and phases, stimulating a system with an impulse yields a full-blown frequency characteristic in one fell swoop.

The impulse also has disadvantages, however. The main problem is that an infinitely short, infinitely strong signal is best approached by a very short and hence very strong signal. The large amplitude may cause several problems. These will be dealt with later, together with an alternative (white noise). For the moment, it suffices to remember that the mathematical impulse can be approximated by a physical pulse that is shorter than the time scale one wants to investigate. The spectrum of such a pulse is white up to the highest frequency of interest. As a rule of thumb, the time scales should differ by about one order of magnitude: if we want a system response to be valid up to 100 Hz (i.e. $(10\,\text{ms})^{-1}$), it is OK to use a pulse of about 1 to 2 ms duration.

The Fourier Transform

In recording electrical signals, one has the input and output signals not as explicit functions but as changing voltages in time. Digitized with a computer, these signals form arrays of numbers in memory. This is where the Fourier transform (FT) comes in.

Basically, the Fourier transform is an integral transform much like the one by Laplace, and so it can be used for essentially the same kind of operation: transforming time data into the frequency domain and vice versa. However, as indicated in the table given earlier, the practice is different. The frequency variable of the Fourier transform is $j\omega$, and so it is applicable only to periodical signals. In Appendix E, this consists of all signals lying on the vertical axis ($j\omega$ axis) in the figure.

Performing an FT can be compared to matching the input signal to a number of sinusoids with different frequencies, and determining the amplitude and phase of the signal content at each of these component frequencies. This will be explained later.

Usually, the Fourier transform is performed numerically, using a computer. Since computers work with a finite number of limited-resolution numbers, the integral transformation describing the FT is converted into its discrete version, the DFT (discrete Fourier transform). The DFT will be treated later on.

The FT has its inverse function, or inverse Fourier transform (IFT), that converts a frequency domain signal into its time domain counterpart.

Since performing an FT neither adds nor reduces the information content of the signal, the operation is fully reversible, and the inverse transform of the inverse transform is again a forward transform: I(IFT) = FT. Thus, one can collect FT pairs: pairs of signals—one in the time domain, one in the frequency domain—that are related to one another by an FT/IFT.

Figure 4-7 depicts the most frequently used pairs. Note that the figure may be read left-to-right as well as right-to-left. This means that, if the left column represents time functions, the right column yields the corresponding Fourier-transformed frequency functions. The converse is also true: if the left column is interpreted as frequency functions, the right column gives the corresponding, inverse-Fourier-transformed time functions.

The reader will have noticed that any representation in the frequency domain should consist of both an amplitude and a phase graph. Here, in the right column, we attempt to depict both amplitude and phase in a single graph. This is possible because in the examples given here, the phase is either 0 or π (counter-phase; 180°) radians. So in the right column spectra, energy at zero phase is drawn above the horizontal (frequency) axis, whereas counter-phase spectral energy is below the axis. Note also that it is a custom to draw spectral energy in narrow bands as vertical lines starting at the f-axis (hence spectral lines, whereas continuous spectra are drawn as curves.

Another important fact is that the left and right graphs are not drawn to scale, and have arbitrary time and frequency calibrations.

Speaking about the right column, we have to explain yet that the graphs are drawn with the origin (0 Hz) in the middle. Mathematically, the Fourier transform yields negative frequencies as well as positive ones. In discussing real-time electrical signals, this may seem odd, because it is unclear what could be meant by "a frequency of minus 100 cycles per second". In this situation, negative and positive frequencies can be interpreted as rotating clockwise and anticlockwise respectively. Using an arbitrary reference, we can speak of a shaft rotating at 100 rpm when it rotates anticlockwise, and −100 rpm when it rotates clockwise. Similarly, a car driving in reverse gear could be assigned a speed of −5 km/h. In optical applications of the Fourier transform (not treated here), positive and negative frequencies represent simply light rays to the right and left of the optical axis, respectively, and so need to be distinguished. In the analysis of signals in time, however, the energy content at negative frequencies may be "folded back" onto the positive-frequency axis, which means adding their energies up. The Fourier transform of a 1 W electrical signal at 50 Hz consists of 0.5 W at 50 Hz and 0.5 W at −50 Hz.

We will inspect each row of Fig. 4-7 (Fourier pair) separately.

a. depicts an impulse at $t = 0$. The spectrum is white, as we have seen from the Laplace transform. Since the function value of the Laplace-transformed impulse is unity at all frequencies, an impulse is the neutral element in the frequency domain, irrespective of the

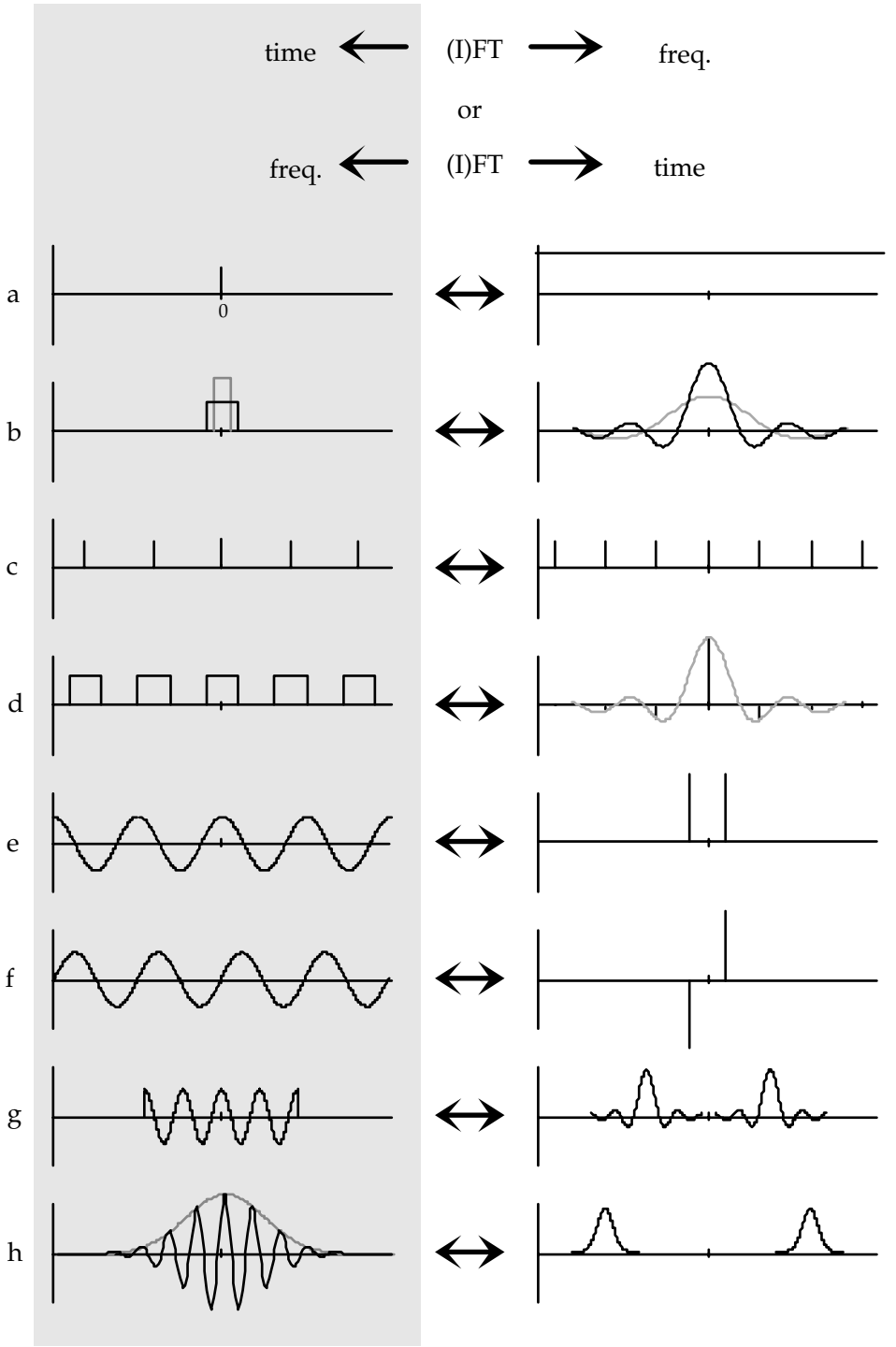


Fig. 4-7 Fourier transform pairs. Left column: time function, right column: frequency function, or the other way round.

choice of transform; it has a "white" spectrum. Reading (a) backwards reveals that a signal that is always the same, i.e. a direct current, has a spectrum that shows only one line, situated at 0. In other words, all energy is at a frequency of zero, which agrees with our notion of a direct current. As stated earlier, an impulse cannot be made physically, and must be approached by a finite pulse.

b. depicts a finite pulse (symmetrically around $t = 0$) and its frequency spectrum. A long pulse has a relatively narrow spectrum (black lines), a shorter pulse has a wider spectrum (grey lines). In the limiting case, an infinitely short pulse, this reduces to the situation at (a), an infinitely wide (i.e. white) spectrum. The rule of thumb given earlier for the use of a finite pulse to represent an impulse can be derived from the relations given here. The form of the spectrum (amplitude as a function of frequency) has a sinc shape ("sinc" being an abbreviation of the $\sin(x)/x$ function).

c. shows the situation in which an impulse is repeated periodically, which is often called an "impulse train". As can be seen, the frequency domain representative is also an impulse train. All spectral lines are equally high, so it can be considered to be a "white" spectrum that exists only at certain frequencies.

d. is a combination of (b) and (c): a repeated, finite pulse. The spectrum is a combination of the line spectrum caused by impulses, now with the sinc function as an envelope (grey curve). Most energy is at zero, so, such a pulse train has a strong DC content.

e. shows a cosine function, i.e. an example of the well-known sinusoid, or "pure tone" signal function. As is expected, the spectrum consists of a single line at the frequency f that corresponds to the inverse of the period (2π radians). This is in keeping with the idea of a sinusoidal (harmonic) motion at a single frequency. However, mathematically, the energy is divided evenly between the frequencies $+f$ and $-f$ Hz.

f. shows the sine function. Intuitively, one would expect the spectrum to be the same as the one at (e), since the sine and cosine functions differ only in their relative phase. Indeed, the spectrum consists again of two lines at $+f$ and $-f$, now with opposite phases. This reflects the difference between so-called even and odd functions, which will be explained below.

g. depicts a tone burst, i.e. a sinusoidal signal that exists for a limited time. Note that, mathematically speaking, the sine and cosine signals shown at (e) and (f) are supposed to exist indefinitely. The tone burst can be considered as a combination of (b) and (e): a cosine limited to a finite, rectangular pulse-shape, or "window", as it is called. Indeed, the frequency domain shapes of both time functions can be seen: a sinc-shaped spectrum, centred around the sine frequencies $+f$ and $-f$ Hz. Note that the left and right graphs are not drawn to scale. This is necessary to show the essential shapes in both time and frequency domain. Note also the implication that a short tone pulse strictly does not possess a single frequency, but has a frequency band (or range of frequencies) instead. In hearing research, for instance, one tries to measure the performance of our hearing system as to tone frequency discrimination. It must be taken into account that a tone burst has a less sharply defined frequency *physically*, i.e. irrespective of the skills of our hearing system. In addition, the spectrum is not monotonous: the sinc function that forms the envelope of the spectrum has so-called side lobes. The spectrum has several peaks, some of them with inverted phase. This led to the development of other shapes of tone bursts. Switching a tone smoothly on and off, by increasing and decreasing the amplitude gradually (rather than by a switch), improves the shape of the spectrum in certain ways.

h. shows a Gaussian-shaped tone pulse. Mathematically, an elegant way of smoothing in the time domain is the Gauss function, the well-known bell-shaped curve that is encountered often in statistics. The FT of a Gauss curve is again a Gauss curve. For the spectrum of a Gaussian-shaped, or "windowed", tone pulse, this means that there are no side lobes (at the expense of a somewhat wider main lobe). In signal analysis, more window shapes are used. The art of "windowing", i.e. shaping the time course of a pulse to improve the spectrum, is dealt with later on.

Odd and Even Functions

The difference between the sine and the cosine functions ((e) and (f) above) illustrates the difference between what is called odd and even functions. As can be seen in (e), the cosine function is symmetrical with respect to the origin (time zero). This can be formulated as $\cos(t) = \cos(-t)$. The cosine is called an "even function" because of this property. The sine function belongs to the other category called "odd functions" that show anti-symmetry around time zero. Thus, $\sin(t) = -\sin(-t)$.

Many more examples could be given. However, most other cases can be derived from the ones given here, using the following, simple rules.

— Time and frequency are each other's inverse. Shorter times correspond to wider spectra, etc. The Fourier pairs may be interpreted backward. Examples: a sinc-shaped pulse yields a rectangular spectrum ((b) read backward); two impulses at opposite sides of time zero yield a sinusoidal spectrum ((e) or (f) read backward).
— Addition is linear: the sum of two signals yields a spectrum that is the sum of the constituents (taking the phases into account).
— Time differences leave the amplitude spectrum untouched, but change the phases. An impulse at any moment other than zero has the same, white, amplitude spectrum, but phases that increase with frequency. For example, an impulse at $t = 1\,\text{ms}$ has a phase of 360° at 1 kHz, etc.

These rules can be combined. As an example, take two successive pulses, identical in amplitude and duration, but one positive and one negative. In this case, the DC components cancel, whereas other parts of the spectrum do not, in general.

Linearity

The foregoing principles and methods of systems analysis rely on a special property of the system in question: the linearity of the system. It means that the response to a sum of input signals must be the sum of the individual responses. If we call the input signals A and B, and the outputs α and β respectively, the linearity principle can be depicted as in Fig. 4-8.

In this example, A is a square-wave signal, and B a low-frequency sine. From the output signals depicted, one can deduce what kind of system we are dealing with: a high-pass filter.

The demand for linearity seems a rather strong limitation of the applicability of system analytical methods. However, if the linearity principle does not hold, mathematical manipulations such as the Fourier transform and convolution to predict the system response would not work, since they implicitly assume that this response can be described by the summation of sine

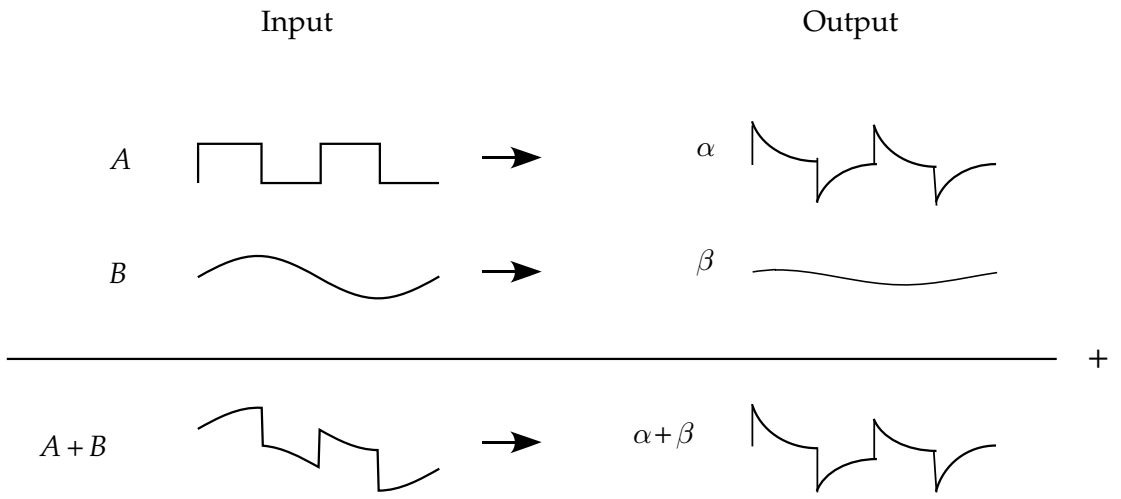


Fig. 4-8 The response of a linear system to the sum of two input signals.

waves or past inputs respectively. One may wonder whether the systems we want to analyse are linear and what will be the consequences if they are not. Fortunately, many systems are linear, at least over a fairly large range of amplitudes. Passive electronic components such as resistors and capacitors and circuits made thereof behave in a linear fashion. To the contrary, many active circuit elements like diodes and transistors, as well as instruments such as amplifiers do not behave linearly, or do so only in a limited amplitude range. The same holds for ion channels, synapses and other neurobiological processes.

In Chapter 2, we have met the saturation of an amplifier, leading to the distortion of the (sinusoidal) signal. Note that "saturation", or "distortion", is a useful measure of (and a warning for) non-linearity. Distortion implies the generation of harmonics, i.e. of frequencies not present in the input signal. If a sinusoid is distorted by a saturating amplifier, part of the signal power will appear in the harmonics. The stronger the distortion, the more power is in the harmonics.

Especially, biological systems may show very complex, non-linear forms of behaviour. The electrical responses of nerve cells, in particular the time course of action potential series, are often strongly non-linear. How, then, do we know whether the methods of linear systems analysis can be applied?

There are several ways to cope with non-linearity:

1. *Approximate linearity*. Fortunately, the linearity demands are not very strict, and many systems are *approximately* linear. If one feeds a sinusoidal signal in, say, an amplifier, the output signal may look fairly sinusoidal when viewed on an oscilloscope screen. If one compares output and input signals simultaneously on the screen, however, any small distortion will show up. The question to be answered is: how much distortion can be accepted by the methods of systems analysis? A pragmatic answer used as a guideline in systems analysis is: 10% distortion. Since distortion implies the generation of other frequencies, one can express the degree of distortion by the relative contribution of the harmonics to the total signal power. If a pure sinusoid is distorted 10%, that fraction of the signal power is converted into higher frequencies

(harmonics). It means that still 90% of the signal power is at the fundamental, i.e. at the input, frequency. Note that, even at less than 10% distortion, a signal may look quite distorted when viewed on an oscilloscope screen, so if a signal looks fairly sinusoidal on-screen, one can be sure that linearity is warranted to a fair degree.

2. *Small-signal analysis*. Even in explicitly non-linear systems, approximately linear responses can be obtained if signal amplitudes are kept low. As an example, the voltage change of an animal photoreceptor cell is approximately logarithmically dependent on the light intensity. However, to determine the frequency response of such a sense organ, one simply stimulates with a light intensity that is modulated only partially.

3. *Linearization*. If one knows the mathematical form of the non-linearity involved, the output signal may be transformed with the inverse of that function before performing the analysis. If the output depends, say, on the square of the input signal, one simply takes the square root of the output signal before applying systems analytical methods. In the example discussed in Fig. 4-2, one could linearize the sensory cell's response by taking the antilog of the output. If a signal is linearized mathematically, the signal amplitude does not need to be kept low.

4. *Non-linear systems analysis*. There are methods developed explicitly to treat (strongly) non-linear systems. These will be treated later on.

Analogue-to-Digital and Digital-to-Analogue Conversions

So far, we have treated signals that are truly continuous waveforms in time. Analogue apparatus, such as amplifiers, filters, pen recorders and the like, keep the continuity in time. Of course, we cannot measure these signals with infinite precision, but the thresholds of detection and precision are gradual, depending on noise levels, interference, etc. In contrast, most signal processing nowadays takes place within computers, instruments that can treat only finite series of finite-precision numbers.

This means that the first step in signal processing (after amplification and, usually, display on an oscilloscope screen) consists of converting the continuous signal into a discrete and digital one (i.e. exisiting only at a finite number of moments, and expressed as numbers). This is done by a circuit called an analogue-to-digital converter (ADC). The result is a series of numbers, kept in computer memory (RAM) and eventually written to disk or to any suitable storage medium. A "sample" is thus defined as a single numerical measurement from the continuous world. Thanks to the generous amounts of RAM in present-day computers and the fast processing speeds, a signal can be digitized in a sufficiently large number of sufficiently precise numbers. Nevertheless, it must be kept in mind that digitization limits the resolution both in time and in amplitude, and the choice of digitization parameters is crucial to the validity of the digital data.

The process of sampling a continuous, real-world quantity can be considered as a mathematical operation on that quantity. To find the essential properties of sampled signals, a theory called sampling theory was developed in the period 1930–1950, by celebrities like Shannon and Nyquist. Fortunately, from these theories we need to remember only a few basics. A first, basic rule is that we cannot make inference on processes we did not sample. A sampled signal has two important quantities: the sampling interval (notated with a lower case t) and the total sampling time, notated with a capital T. See Fig. 4-9.

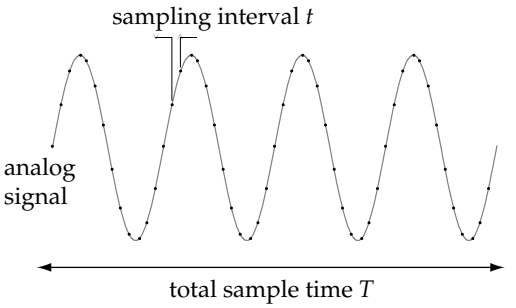


Fig. 4-9 Digitizing an analogue waveform. The analogue signal, here a sine, is depicted in grey, together with the digitization times (black dots). The total time T might be 1 s, the sampling interval t 1 ms (note: not drawn to scale for clarity).

For a series of k samples, spaced t seconds apart, the total sampling time is $T = kt$ seconds. A little confusion might arise because, sometimes, the entire series of samples is also called "a sample" (which it is, of course; a small piece of a real-world signal). To avoid the confusion, a series of consecutive samples (sampling points) is better called a "sweep".

The very fact of "sampling" a real-world quantity poses limits on the information content of the resulting digital signal. First, processes that take more than T seconds have been sampled only partially, so insufficient information about them is available. In the frequency domain, this implies that we have no information about frequencies lower than $1/T$ Hz. Second, the sampling interval limits the time resolution: between two samples, essentially nothing of the original signal is known. These limitations are both obvious and important, and can be explained best with a numerical example.

Imagine a 1 s sound having 1000 samples, spaced 1 ms apart. Since we have only one second, we do not have information about frequencies lower than 1 Hz. In addition, the sampling interval of 1 ms limits the time resolution of our sample: the "spaces" between our time samples are empty, i.e. unknown. This limits the upper boundary of the frequency spectrum fundamentally to 500 Hz. Remember that one needs at least two samples per period to characterize a frequency. This is the important "Nyquist criterion". If a signal is sampled too infrequently, i.e. with two samples per period (of the highest frequency present) or less, insufficient information on the sampled waveform is obtained. Worse still, a phenomenon called "aliasing" occurs. Both the "Nyquist frequency" and the pitfalls of aliasing were treated in Chapter 2.

A voice signal sampled in the described way would be barely useful, since it would contain only frequencies up to 500 Hz. This implies that the higher frequencies (harmonics) that help to recognize and understand the voice signal are absent.

Note that the Nyquist frequency is the absolute, fundamental limit of digitization.

It is better to choose a sampling frequency somewhat higher than the Nyquist value. Digital audio, for example, uses sampling frequencies of 44.1 or 48 kilosamples per second to cover the audio spectrum, i.e. frequencies below 20 kHz.

For the electrophysiological practice, the sampling frequency has to be chosen carefully, depending on the type of signal one observes. Since fast AD converters are affordable today, sampling frequency is usually chosen to be higher than the required minimum value.

Thus, for the examples mentioned in Chapter 2 (amplifiers), useful sampling rates would be as in the table below.

for nerve membrane potentials	about 10 kilosamples/s
for electrocardiograms	about 100 samples/s
for electroencephalograms	about 100 samples/s
for nerve or muscle spikes	about 10 kilosamples/s
for plant action potentials	about 10 samples/s
for the analysis of spike shape	about 200 kilosamples/s
for single-channel recording	about 100 kilosamples/s

Notes:

1. Oversampling only yields larger data files without giving more information. Despite the availability of cheap storage media, collecting useless amounts of digital data should be avoided.
2. To avoid aliasing, one must filter the analogue signal in order to avoid higher-than-Nyquist signal frequencies. Remember that some AD converter cards have so-called anti-alias filters built in; but most of them have not.

Signal Windowing

In digital signal processing, the data consist of a finite sample of discrete values, taken from a practically continuous physical world. As discussed earlier, this limits the amount of information that can be deduced from the sampled signal both in time and in amplitude. Essentially nothing can be said about the physical reality before the first and beyond the last sample points. Worse still, the abrupt beginning and end of a sample introduce artefacts when performing transformations such as the FT. This is illustrated in Fig. 4-7, where Fourier pair **e** shows an infinitely long sinusoid and its Fourier transformed counterpart, which consists of two spectral "lines" at + and − the frequency of that signal. A short sample of such a signal is shown in trace **g**, and, indeed, it has a spectrum very different from the former one: instead of lines at certain precise frequencies, the spectrum consists of two continuous functions, spanning a range of frequencies (the sidebands) and having side lobes where the phase is reversed. Apparently, the properties of a short sample of a continuous function may deviate substantially from the properties of the original. A first conclusion of this observation is that a long sample is better than a short one. Indeed, if the sample length is about 10 s, the width of the sidebands is reduced to about 0.1 Hz, so that the spectrum resembles the Fig. 4-7e better. No matter how long the sample, however, it will always be bounded by a beginning and an end. So-called windowing functions are invented to alleviate this problem. The trick consists of reducing the amplitude of the signal sample values in the vicinity of the bounds. This is performed by multiplying the sample values with a function that tends to zero at the boundaries. A useful function is a shifted cosine, called the "Hanning window" after its inventor, the German J. von Hann. This is depicted in Fig. 4-10A.

The other functions shown are called "Hamming window" (B, after R.W. Hamming) and 'Gaussian window' (C). Note that the Gaussian envelope is necessarily truncated, so that the amplitude at the ends of the sample is not exactly zero. Still more windowing functions exist, all

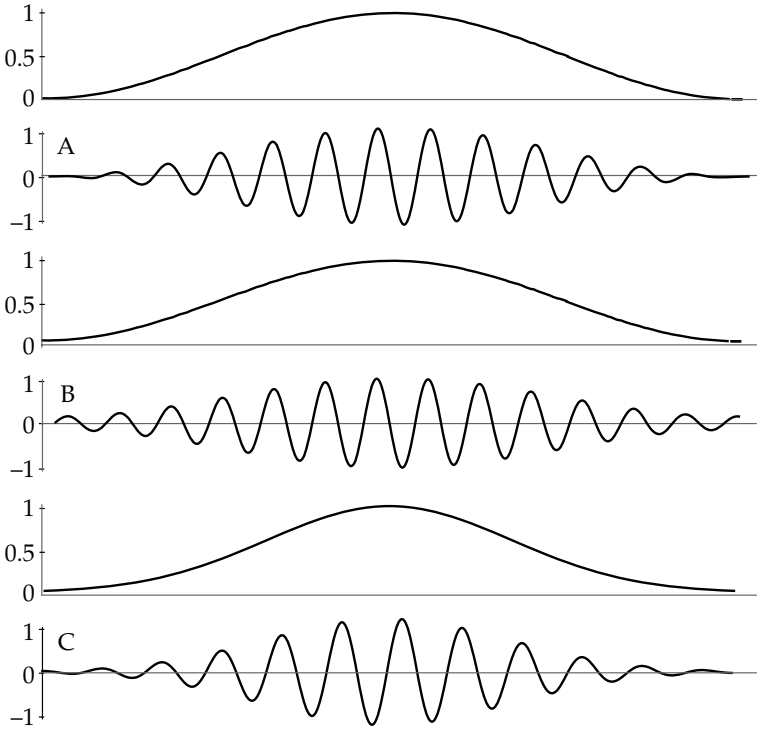


Fig. 4-10 A window function (upper graph) and its resultant signal sample shape (bottom graph) for three popular window functions: A—Hanning; B—Hamming; and C—Gaussian.

having the same purpose: to reduce the signal amplitude at both ends of the sweep (window) to zero or to a small value. All these window functions yield better representations in the frequency domain, as will be illustrated below. Finally, applying no such window function at all is in fact a window by itself, called the "rectangular window" (because the signal values are kept unattenuated everywhere).

The effect of using different window functions show up dramatically in the frequency domain. This is shown in Fig. 4-11.

In principle, a pure sine signal, such as in A, should show up in the frequency domain as a single spectral component, or "spectral line". In other words, all the energy exists at a single frequency, provided the sine wave exists at all times. However, since any sample spans a finite time, the spectral lines have a non-zero width, equal to the inverse of the sweep duration.

Things get worse when the sweep contains a signal truncated in an arbitrary way, and, indeed, this is almost always the case when sampling real-world signals. An example is illustrated in trace B: both ends of the sample do not connect, since the sweep starts at a large positive value and ends at a large negative one. The effects in the frequency domain are dramatic. The spectral line (or lines at frequencies of both +16 and −16 units) are widened substantially, thus reducing the precision with which the signal spectrum can be analysed. In addition, the wider lines may occlude any nearby frequency components. The Hanning window (C) reduces the problem to a large extent. Note, however, that the amplitude spectrum

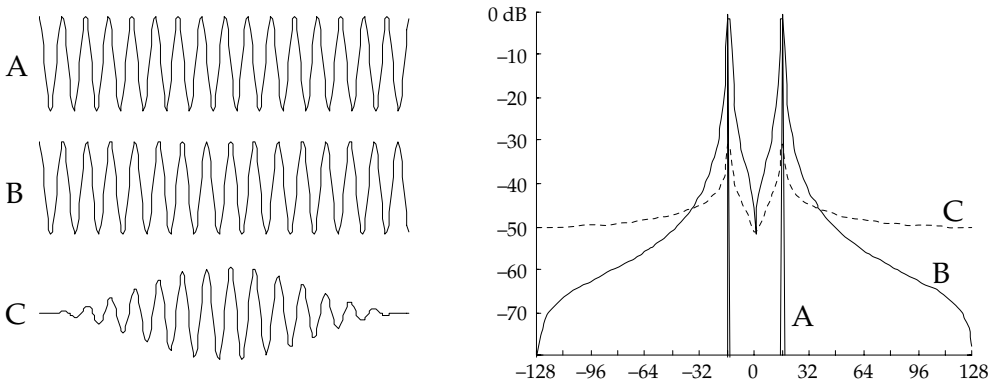


Fig. 4-11 Signal samples (256 points) with rectangular versus Hanning window. Left: shape of signal sample; right: frequency spectrum. A: a sinusoid sample with exactly 16 cycles in the rectangular window. B: a sinusoid sample with about 15.5 cycles in the rectangular window. C: the same as B, but multiplied with the Hanning window.

is rendered with a logarithmic *y*-axis, reading decibels. At very low amplitude values (more than 40 dB below the peak), the Hanning-windowed signal spectrum is wider rather than narrower than the rectangularly windowed one. Although this will not usually play a part, at times one of the other window shapes mentioned above may yield better results. In the analysis of loud sounds, for instance, −50 dB might very well be within the audible range. Investigators of sound and hearing must choose their window functions carefully.

Digital Signal Processing

The simplest form of digital signal improvement is averaging a number of consecutive sweeps.

Signal Averaging

"Signal averaging", also called "signal recovery", is the most straightforward method of digital signal processing. It can be applied in all those cases in which a repetitive signal is administered to a human, an animal or a preparation. Brief electrical pulses, tone bursts or light flashes are usually given, say, once every few seconds. If the response of a neuron, a sensory cell or a brain nucleus is "buried" in noise, signal averaging may improve the signal (more precisely the signal-to-noise (S/N) ratio) to a large extent. In many cases, a recognizable response emerges only after substantial averaging, such as in the case of "event-related potentials" (ERP) in the brain. A good example is the weak response of a photoreceptor cell to a dim light flash, such as shown in the introduction to this chapter. The response of the cell (say, a slight depolarization) is obscured by the noise in the recording. This may include both membrane noise stemming from the cell and noise from the electrodes and preamplifier.

Figure 4-12 shows a weak pulse buried in noise and the improvement in S/N ratio obtained by the averaging of different numbers of sweeps.

This form of signal averaging is used widely for the analysis of physiological and other biological (or physical, for that matter) signals. However, its use depends entirely on the availability of a clean "trigger signal" that signals the time when the averaging should start.

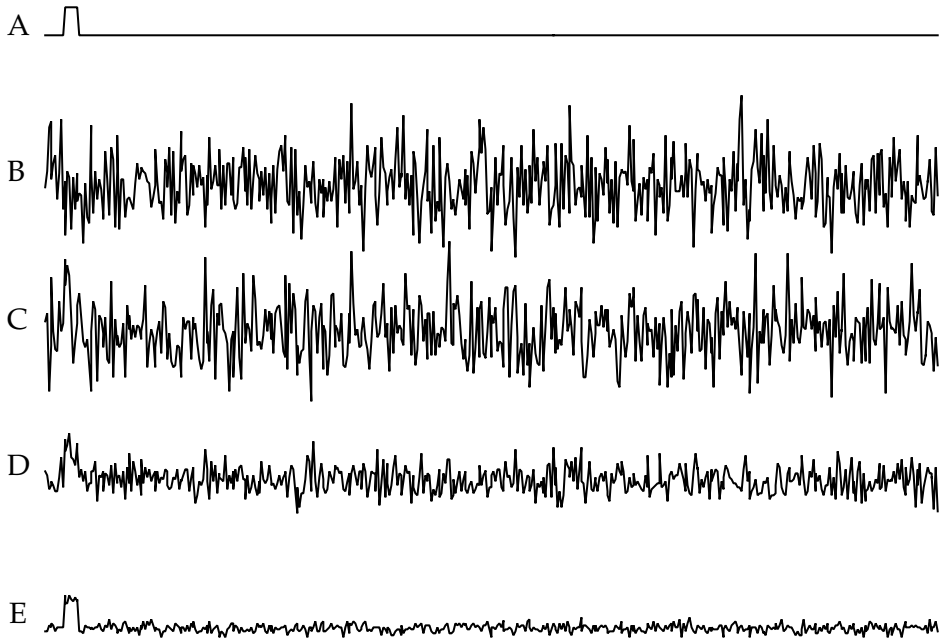


Fig. 4-12 A weak pulse buried in noise and the improvement in S/N ratio obtained by the averaging of different numbers of sweeps. A signal, B noise, C signal + noise, D average of 5 sweeps, E average of 50 sweeps.

In neurophysiology, this is often the case, since we have a stimulus signal or some other event, time-related to the signal investigated. However, in situations where there is no distinct starting command, other techniques are needed. One such possibility lies in correlating one or more signals. This is the next subject.

Autocorrelation

Consider a noisy signal (such as in Fig. 4-12) that has been sampled fast enough to be able to see its details. If the signal were a purely random signal, then the value of each sample would be completely different from the preceding or the following sample. What we see instead is that each sample is followed by one of a similar value: adjacent samples are correlated. A measure of the degree of correlation between samples in a signal $x(t)$ that are separated by a distance t is given by multiplying the signal with a time-shifted version of itself:

$$A(\tau) = M[x(t) \cdot x(t-\tau)] \qquad \text{(Eq. 4-1)}$$

where τ is the time shift and $M[\ ]$ reads: "the mean of".

Because $x(t)$ is being correlated with itself, $A(t)$ is called the "autocorrelation function" (ACF). It is clear that $A(t)$ takes on positive values if $x(t)$ and $x(t-\tau)$ are on the average very similar, whereas if $x(t)$ and $x(t-\tau)$ vary independently of one another, it will be zero. $A(t)$ for the signal in Fig. 4-13 is shown in the same figure. As might be expected, it is a function that is symmetrical around $t = 0$.

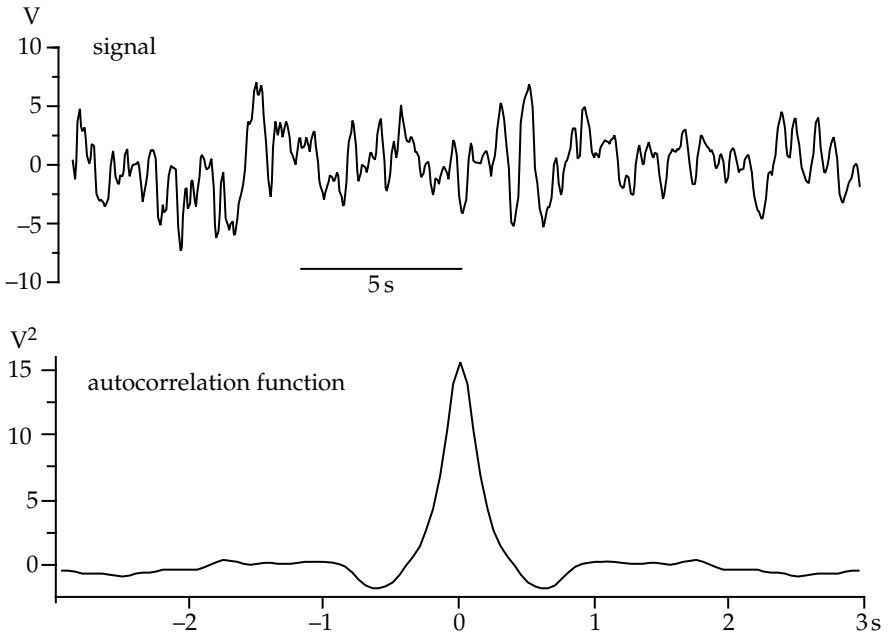


Fig. 4-13 A noisy signal and its autocorrelation function.

Going from a sampled signal to a continuous signal, 4-1 becomes:

$$A(\tau) = \int x(t) \cdot x(t-\tau)\,\mathrm{d}t$$

Things become different when the noisy signal contains a periodic component, such as a sine, square or pulse. For periodic signals, the ACF does not die out at larger time shifts. Since a sine wave repeats itself every period, the ACF shows peaks at those values. In addition, it will be clear intuitively that the autocorrelation of a sine wave yields negative values after half a period, because here one multiplies a positive with a negative value. The result seems simple: the ACF of a sine is again a sine. Note, however, that the horizontal axis of the ACF reads time shift rather than time. Other waveforms are not preserved in their ACFs: for example, the ACF of a square has a triangular shape. Apparently, some information as to the signal shape is lost.

Nevertheless, a lot of information on the signal can be inferred from the shape and scaling of ACFs, which is illustrated in Fig. 4-14.

In the case of pure, band-limited noise signals, the ACF shows the bandwidth (A), and the steepness of the bandwidth limitation (B). The ACF of periodic signals does not die out at higher time shifts, and retains some information on the waveform (C). Finally, most real-world signals will contain both a periodic component and noise, which can be segregated by looking at the ACF (D). From the latter example, one can determine amplitude and bandwidth of the noise component from the peak at low time shifts, as well as amplitude and frequency of the sinusoidal component from the ACF at higher time shifts.

The autocorrelation procedure implies averaging, and is a powerful way of detecting *any* periodic signal buried in noise. Note that, contrary to the signal averaging method described

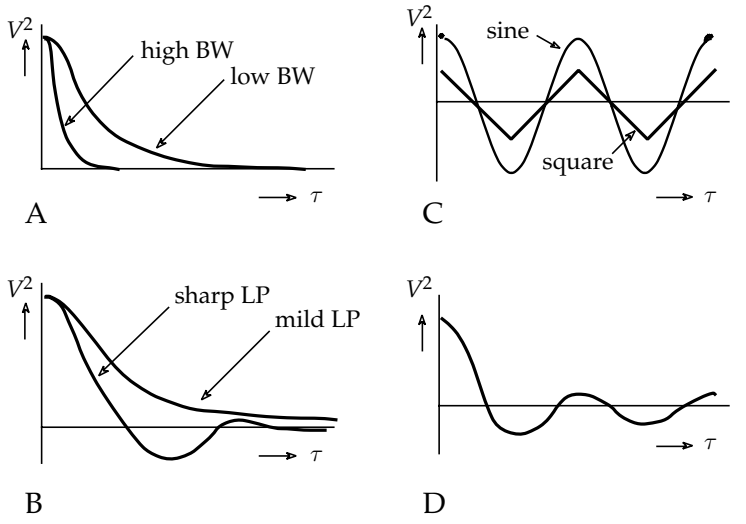


Fig. 4-14 Autocorrelation functions of noisy and periodic signals. A: noise signals with different bandwidths, B: mild versus steeper BW limitation (1st order and higher order low-pass respectively), C: periodic signals and D: a combined sine + noise signal.

above, no clean stimulus or trigger signal is necessary. Therefore, autocorrelation may also be used to find unknown spontaneous periodic signals in signals from neurons, sense organs or brain. In astronomy, the existence of pulsars (pulsating stars) was found by the saving grace of autocorrelation: the pulsations are far too weak to be detected directly.

Crosscorrelation

In addition to time-shifting a single signal, we may also wish to correlate $x(t)$ with a different signal $y(t)$. We then obtain the "crosscorrelation function" (CCF):

$$C(\tau) = \int x(t) \cdot y(t-\tau)\,\mathrm{d}t \qquad \text{(Eq. 4-2)}$$

Crosscorrelation functions can be used, for instance, to establish whether two neurons are functionally coupled and if so, to determine the time lag between activity in one neuron and that in the other. Eq. 4-2 is also called "convolution integral". As we will see later, it is a powerful tool in data analysis, signal detection and filter design.

A special case of crosscorrelation is the Fourier transform. If we replace y by a sine wave function of frequency f, we get

$$A(f) = \int x(t) \cdot \sin(2\pi f t)\,\mathrm{d}t$$

$A(f)$ represents the amplitude of the sine wave having a frequency f as present in the signal x, i.e. the Fourier coefficient of $\sin(f)$. The cosine series is found similarly:

$$B(f) = \int x(t) \cdot \cos(2\pi f t)\,\mathrm{d}t$$

Using complex notation, the last two equations can be combined into a single, compact equation:

$$X(\omega) = \int x(t) \cdot e^{-j\omega t}\, dt$$

with $\omega = 2\pi f$ and $j = \sqrt{(-1)}$. An important property of the Fourier integral is that the components $X(\omega)$ are orthogonal (independent or uncorrelated) with respect to each other, or:

$$\int e^{-jst} \cdot e^{-j\omega t}\, dt = 0$$

This property indicates that its inverse function exists. In fact, reconstitution of the signal x from its spectral components $X(\omega)$ is simply a summation of all sines and cosines in the spectrum:

$$x(t) = \int X(\omega) \cdot e^{j\omega t}\, d\omega$$

The "convolution theorem" gives a straightforward relation between convolution of two functions in the time domain and (complex) multiplication in the frequency domain. It states that if $g(t)$ is the convolution (©) of $x(t)$ and $y(t)$, then the Fourier transform of $g(t)$, written $G(\omega)$, is the product of the Fourier transforms of $x(t)$ and $y(t)$, or:

$$\text{if } g(t) = x(t) \copyright y(t) \quad \text{then} \quad G(\omega) = X(\omega) \cdot Y(\omega) \qquad \text{(Eq. 4-3)}$$

Thanks to this theorem, it is relatively easy to carry out the inverse of convolution. This is called "deconvolution" and it is often used to compensate for signal deterioration due to known sources, such as, for example, in improving the image of a star seen through an imperfect lens. If in Eq. 4-3 we know $g(t)$ and $y(t)$ and wish to know $x(t)$ then:

$$x(t) = F^{-1}\{G(\omega)/Y(\omega)\} \qquad \text{(Eq. 4-4)}$$

where F^{-1} stands for the inverse Fourier transform

If the convolution integral of Eq. 4-3 is rewritten with slightly different symbols, then it can be regarded in a general way as the response y of a (linear) system to a stimulus x. The transfer function $h(t)$ embodies the properties of the system under study:

$$y(t) = \int h(t - \tau) \cdot x(t)\, dt$$

The advantages and disadvantages of the impulse were described earlier. Since the theory holds only if one may assume linearity, the impulse should remain relatively small. In that case, it represents little energy and the output might be difficult to extract from the ever-present background noise. A way to circumvent this problem is to use "white noise" as the input signal. Although it may seem strange, in the frequency domain, white noise is very much like an impulse: it has a white spectrum , i.e. all frequencies have equal amplitudes. Only the phases differ: in an impulse, all phases are zero at $t = 0$, whereas the phases of the components of white noise are distributed randomly.

As with the ACF, the shape and scaling of a CCF yields information on the process to be analysed. This is illustrated further in Fig. 4-15, where the crosscorrelation of noise with deterministic signals is shown. With crosscorrelation, the common component in two signals can be revealed, even in the presence of much noise. Obvious applications in neurophysiology

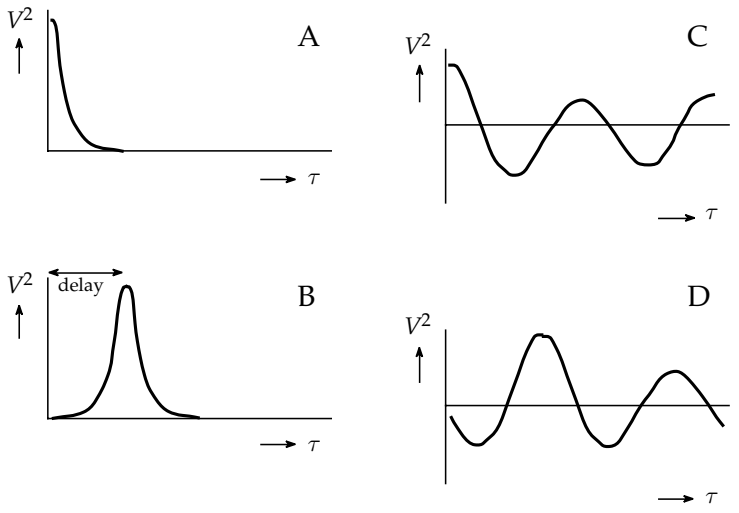


Fig. 4-15 Crosscorrelation functions of different signal situations. A: two uncorrelated, band-limited noises. B: a noise and a delayed version of that noise. C: a sine with a noisy sine. D: a sine with a delayed noisy sine.

are the detection of ERPs, which are usually both buried in brain activity that is unrelated to the stimulus.

The Discrete Fourier Transform

Having dealt with the do's and don'ts of signal digitization, we return to the digital, or discrete, form of the Fourier transform (DFT). Algorithms to implement the FT exist for the most popular programming languages such as Basic, C and Pascal. In addition, math packages for personal computers, such as Mathematica®, Matlab® and Sysquake®, have built-in routines to perform the FT and IFT.

Let us take our earlier example again: a 1 s record having 1000 samples at 1 ms intervals. The DFT would contain 500 spectral lines, 1 Hz apart. This is in keeping with the limitations stated by the Nyquist criterion. If we do not have any information on the signal where this record was taken from, we have no information on frequencies lower than 1 Hz, so the "space" between the spectral lines is empty (unknown). Only if we would extend our sample to span 2 s, we would have spectral lines every 0.5 Hz. And at the other end, we have spectral information up to 500 Hz. Taking samples every 0.5 ms would extend our spectrum to 1 kHz, and so on.

In principle, performing a DFT of sample size n involves n^2 computations (involving sine and cosine functions, so many time-consuming floating point operations). For the early computers, in the 1950s and 1960s, this was a tedious task. In 1965, however, Cooley and Tukey published a far more efficient algorithm known since as the fast Fourier transform, or FFT for short. The FFT avoids any superfluous computations, but to obtain the maximum efficiency, the sample size must be a power of two. Thus, an FFT might comprise of 256, 512 or 1024 samples. The last value is called "one k". Our modern computers allow FFTs of 8 k or more to be performed routinely. The size chosen will depend usually on the nature of the input signal, on the time it takes to perform the FFT, and the desired precision of the result (the frequency domain version of the signal).

The Detection of Signals of Known Shape

Equation 4-4 can help the detection of signals of known shape but unknown size. An example of such a detection problem presents itself if we want to analyse a record containing postsynaptic currents. If only one type of neurotransmitter is released, then the decay times are often similar between individual currents. They are not identical however, due to differences in the distances of the sites of transmitter release and the soma. The voltage-clamp record shown in Fig. 4-16 has been recorded from cerebellar granule cells in culture.

To detect an event we do the following:

1. Take the Fourier transform of both the stretch of data and the template. The template is an idealized prototype of a postsynaptic current without noise and of unit amplitude.
2. Divide the two spectra.
3. Take the inverse Fourier transform of the resulting spectrum. The lower trace in Fig. 4-16 shows the deconvoluted data. Sharp peaks indicate where the onset of synaptic currents occurred.

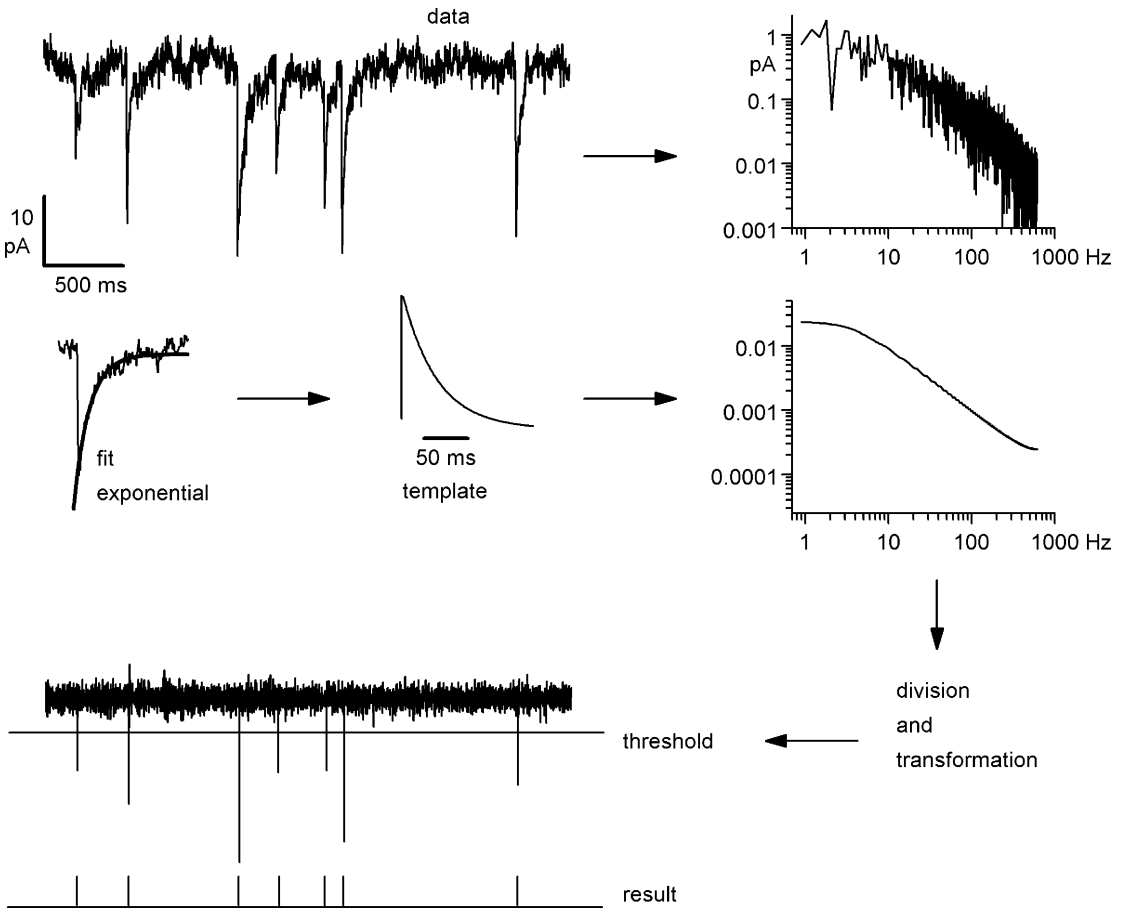


Fig. 4-16 Detection of signals of known shape (explanation in text).

4. Finally, using a threshold to separate the peaks from the background noise, we get the event times we were looking for. Note that the amplitudes of the deconvoluted spikes are not always proportional to the original synaptic currents. This is due to an imperfect match between the template and the synaptic current in question that has for instance a less typical decay time.

The template may be created in two ways. A few of the larger synaptic currents may be fitted with an exponential function, and the average decay time may then be used to create the template (the method used in this example). The second method consists of fitting the Fourier spectrum of the data with the appropriate function. Note that in Fig. 4-16, the data spectrum and the template spectrum are similar. This is because the relaxation time constant of the synaptic currents dominates the data spectrum. On top of that, the spectrum contains "white" and "pink" ($1/f$) noise (instrument and thermal noise). To get the Fourier transform of the exponential we integrate from 0 to ∞:

$$E(\omega) = \int_0^\infty \mathrm{e}^{-kt} \cdot \mathrm{e}^{-\mathrm{j}\omega t} \mathrm{d}t$$

yielding:

$$E(\omega) = 1/(k + \mathrm{j}\omega)$$

which after separation of real and imaginary parts yields:

$$E(\omega) = k/(k^2 + \omega^2) - \mathrm{j}\omega/(k^2 + \omega^2) \qquad \text{(Eq. 4-5)}$$

If we now create the power spectrum, $E^2(\omega)$, we have a function that is easy to fit:

$$E^2(\omega) = 1/(k^2 + \omega^2) \qquad \text{(Eq. 4-6)}$$

This function, a spectrum resembling the frequency characteristic of a low-pass filter, is called a Lorenzian. By adding white (w) and pink (p) noise the final function to fit to the power spectrum will be:

$$F(\omega) = a/(k^2 + \omega^2) + w + p/\omega^2$$

with $1/k$ being the relaxation time constant of the "average synaptic current".

Digital Filters

The above procedure is an example of a digital filter. Since data acquisition and analysis is done almost always with computers, it is worthwhile to examine the principles of digital filtering on a more general basis.*

* Although digital filtering is more powerful and flexible than the old analogue filters, one should not underestimate the importance of a simple RC filter in the pre-amplifier. If an electrophysiological pre-amplifier is saturated by a strong polarization voltage, or by a 50 (60) Hz hum or any other form of unwanted signal, the signal may get distorted beyond recognition, and no digital operation afterwards can help.

The most elementary form of a digital filter that is useful to reduce wide-band noise is to average a small number of consecutive samples, and use the series of averages as a filtered signal. This is illustrated in Fig. 4-17. Of the noisy signal, every five samples are averaged. This reduces HF noise in the sample somewhat. However, we end up with time resolution (samples per second) far less than that of the original signal. Samples 1–5 yield the first data point, samples 6–10 the second, and so on. There is a better way, however, that keeps the time resolution almost unchanged. In averaging consecutive samples, there may be some overlap. Thus, one may average samples 1–5, samples 2–6, samples 3–7 and so on. This procedure is known as a "running average" or "moving average". In this case, the output signal has almost as many samples as the input signal (at the start and/or at the end of the sweep one loses a

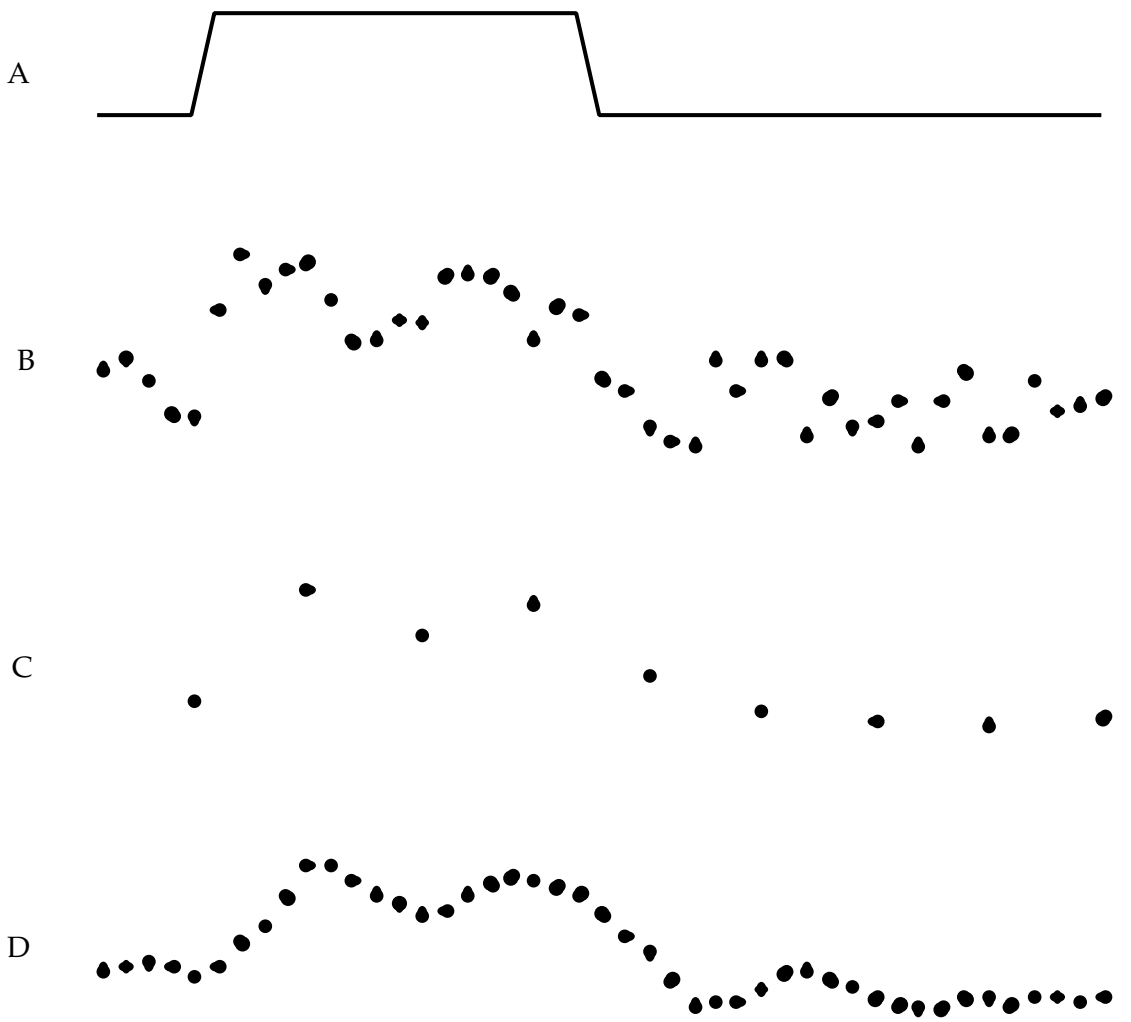


Fig. 4-17 Effect of simple digital filters. A: a real-world pulse, B: a noisy, sampled version thereof, C: a series of averages of 5 samples each, D: a running average over 5 samples.

few points). As you can imagine, this behaves as a low-pass filter, since any abrupt change of the signal will be "smeared out" over a small number of output samples. The well-known "blurring" of digital photos is just a two-dimensional version of a running average. This serves to reduce image noise (or grain), at the expense of sharpness.

This type of filter is known more generally as a "finite impulse response", or FIR filter. As you can see in Fig. 4-18, the response to an impulse dies out after as many samples as the averaging interval includes.

In this example, all five samples are averaged with equal contributions. This is by no means the only possibility: One might let the centre samples prevail, by assigning the different samples different weights. For instance, sticking to the same example, we can average $0.25y_1 + 0.5y_2 + y_3 + 0.5y_4 + 0.25y_5$ (and divide by 2.5). The series of weights (0.25 0.5 1 0.5 0.25) again is called a window (see Signal windowing above). The equal-values window again is called a "rectangular window". Different window shapes may be used to get specific results.

Analogue filters, such as an RC filter, have an unlimited (infinite) impulse response. For example, a capacitor, charged at $t = 0$, will lose its charge asymptotically, its charge theoretically never reaching exactly zero. Although digital computations are limited to discrete steps, it is nonetheless possible to simulate such asymptotical, real-world behaviour with digital filtering algorithms. These are called (you might have guessed) infinite impulse responses (IIR) filters. They are also known as recursive filters because the filter output is used over and over again. The simplest version is a simulation of an RC low-pass filter. The procedure is simple indeed: At each sample point, take 20% of the new sample value and add it to 80% of the existing value. This is the new, filtered, value. Repeat the process over the whole digitized signal (i.e. the sweep).

In a formula, each output value $Y(t)$ at time t depends on the previous output value $Y(t-1)$ and the input value $X(t)$:

$$Y(t) = 0.2X(t) + 0.8Y(t-1), \quad \text{or, more general:} \quad Y(t) = pX(t) + (1-p)Y(t-1)$$

The factors may be varied over a wide range, provided the two factors add up to 100%, to keep the signal amplitude unchanged (hence p and $1-p$). Using larger values would introduce a gain, which should not pose a problem, by the way.

Taking $p = 20\%$ again, the impulse response would consist of the numbers 1, 0.8, 0.64 (i.e. 0.8×0.8), 0.512, 0.4096 and so on. As you can see, this series approaches zero in an exponential way. However, despite the "infinite" in the name, the response will reach zero in

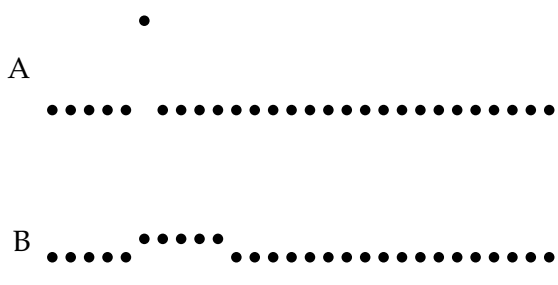


Fig. 4-18 A digital impulse (A) and the response of a 5-sample running average filter to it (B).

a finite time, by rounding errors, i.e. when the smallest quantization level has been reached. The time constant is approximately equal to the inverse of the factor p.

This simple example may serve to show the principle of digital filtering. However, the art of digital filtering has grown into a specialized branch of digital technique and the concomitant math developed with it, so far more sophisticated and specialized filtering methods can be applied with success to electrophysiological data. The next section is an example of such a method.

Fourier Filters and Non-Causal Filters

Electrophysiological data is almost always recorded on hard disk by a computer for off-line analysis. This gives the possibility to acquire the data with maximal bandwidth and filter it later depending on the type of analysis chosen. We start with a set of data in digitized form. In general, there are two approaches to the problem of digital filtering: (1) the frequency domain approach and (2) the time domain approach. However, the distinction between the two is not sharp, as one approach can often be restated in terms of the other.

After the discovery of the FFT, filtering in the frequency domain has become the most popular one. According to the theory of Fourier, a signal can be thought of as a sum of sine waves of different frequencies, amplitudes and phase shifts. It is then easy to remove unwanted frequencies from the spectrum.

This process can then be resumed following three steps:

1. calculation of the sine wave spectrum (Fourier transform),
2. removing or otherwise manipulating of the sine wave parameters, such as their amplitudes, and
3. reconstitution of the filtered signal by an inverse Fourier transform.

A simple example shown in Fig. 4-19 may illustrate the procedure. A pulse-shaped signal was recorded against a background of 50 Hz mains interference (upper left). The sine wave amplitude spectrum was then taken (step 1, upper right). The large peak at 50 Hz and its harmonics at 100, 150, 200 and 250 Hz were then removed (step 2, lower left) and the signal was reconstituted (step 3, lower right).

Filtering in the time domain is based on the convolution integral of Eq. 4-2:

$$y(t) = \int_0^\infty x(t) \cdot h(t-\tau)\, \mathrm{d}\tau$$

or in shorthand:

$$y(t) = x(t) \copyright h(t)$$

where x is the input data, h is the impulse response of the filter and y is the filtered signal. In the case of a simple RC filter, the function h is:

$$h(t) = k\mathrm{e}^{-kt}$$

where the RC time of the filter equals $1/k$.

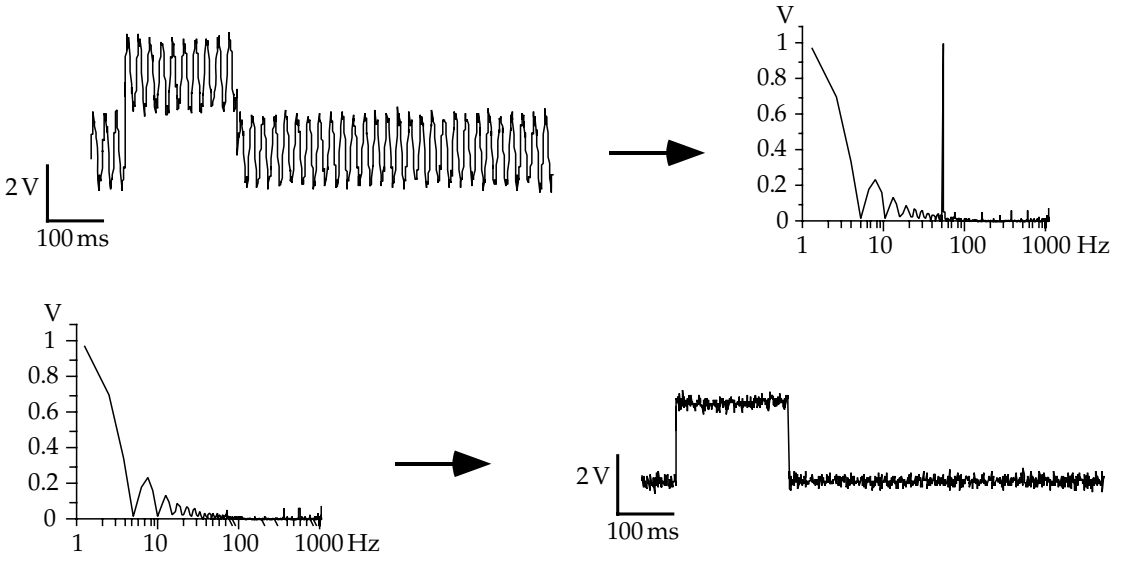


Fig. 4-19 Frequency-domain filtering.

According to the convolution theorem (13), the Fourier transform of our RC filter, $h(t) = ke^{-kt}$, should give us the exponential time characteristics as shown in Part I.

$$H(\omega) = k \int e^{-kt} \cdot e^{-j\omega t} dt$$

When integrating from 0 to ∞ and after separation of real and imaginary parts this yields the same result as in Eq. 4-5, scaled by a factor of k:

$$H(\omega) = \frac{k^2}{k^2 + \omega^2} - \frac{jk\omega}{k^2 + \omega^2}$$

As we have seen in Eq. 4-16, the real part , or $k^2/(k^2 + \omega^2)$, is the so-called Lorenzian function. The amplitude spectrum of the RC filter can now be found by multiplication of H(ω) with its complex conjugate and taking the square root (for explanation of complex conjugate, see Appendix E):

$$|H(\omega)| = \sqrt{H(\omega) \cdot H^*(\omega)} = \frac{k}{\sqrt{k^2 + \omega^2}}$$

We see that for low frequencies ($\omega \ll k$) $H(\omega) = 1$, and, for high frequencies ($\omega \gg k$), the formula reduces to k/ω. Hence the slope of the filter after the cut-off point is 6 dB/octave (see Chapter 1). This little exercise shows that filter algorithms that work in the frequency domain can be designed starting from impulse responses in the time domain and vice-versa. Now let us take a closer look at the time-domain RC filter. When we rewrite $h(t)$ in discrete form it becomes $h[i]$:

$$h[i] = k\, e^{-ki} \Delta t$$

and the convolution becomes:

$$y[j] = \frac{k}{T} \sum_i x[j-i] \cdot \mathrm{e}^{-ki} \Delta t$$

where x is an array that contains the data to be filtered and y the filtered data. The array h contains the filter coefficients. T is the time span of the summation over n and Δt is the sample interval between the data points. The process of filtering is shown in Fig. 4-20. The contents of array h and the input array x are shown on the top. The input (impulse) signal is zero everywhere except at index 5. The filtering starts (step 1) with the x and h arrays aligned at index 0. The sum of products, $x[i] \cdot h[i]$, which is 0 in this example, is stored in $y[4]$. Then the array h is shifted one position to the right and the sum is stored in $y[5]$, etc. until the end of the x array is reached. Upon exit, y contains the filtered impulse signal, which, in our case, is a scaled copy of the filter function itself. This illustrates why the function h is called the "impulse response".

Note that the filter in Fig. 4-20 uses only points that are located in the past to calculate the present output. Such filters are called causal filters. These filters can also be constructed with electronic elements in real life. However, since digital filters are applied on data that is stored on disk, it is possible to use future data to calculate the present filter output. This is called non-causal. It is especially useful to resolve problems concerning the beginning of the output array. In the last example, the elements 0, 1, 2 and 3 of the output array were ignored. Now that it is known what will come, it is easy to predict that the contents of the first four elements should be 0. An example of a non-causal filter is a simple extension of the causal RC filter discussed above. Suppose that the impulse response h is again an exponential, but now symmetrical around the $t = 0$ axis:

$$h(t) = k\,\mathrm{e}^{-k|t|}$$

where $|t|$ is the absolute value of t.

As this is a symmetrical function, the Fourier transform consists only of a cosine series (there is no imaginary part):

$$H(\omega) = k^2/(k^2 + \omega^2)$$

This is the Lorenzian function that we have seen before. Because there is no imaginary part, there are no phase shifts introduced by the filtering process. It is a phase-less filter. For high frequencies ($\omega \gg k$), the slope of the filter is proportional to $1/\omega^2$, which corresponds to 12 dB/octave. Note that the same output would have been obtained if the data were passed through the causal RC filter twice, once in forward direction and once backwards.

A special case of non-causal filters is the recursive filter. Recursive filters, which we met already as IIR filters, use both input and output data. They are difficult to design and imply solving a system of linear algebraic equations. However, once their characteristics are established they are very compact. The example given below is, again, a low-pass RC filter:

$$y[n] = (1-a) \cdot x[n] + a \cdot y[n-1]$$

Note that in contrast to the causal filter, this filter uses only two array cells at n and $n-1$. Its transfer function is (Hamming, 1983; Marmarelis and Marmarelis, 1978):

$$H(\omega) = \frac{1-a}{1-a \cdot \mathrm{e}^{-\mathrm{j}\omega\Delta t}}$$

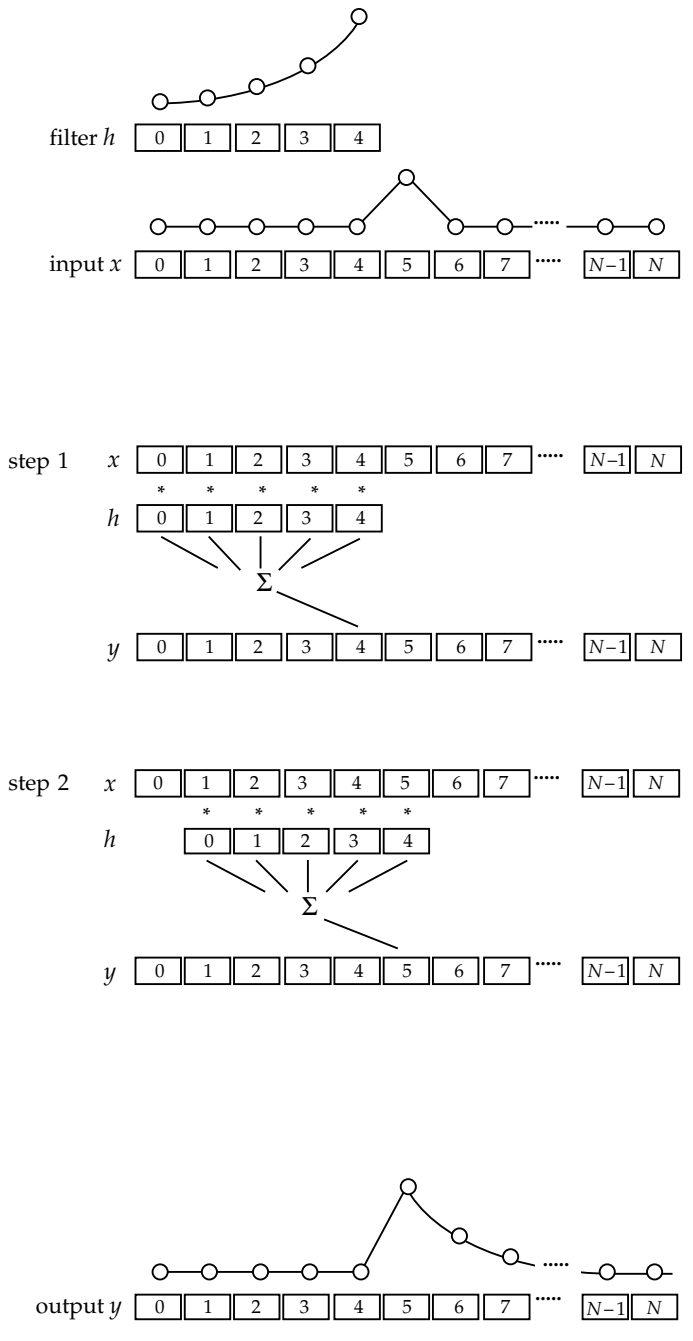


Fig. 4-20 A digital filtering process.

from which it is not too difficult to find the relation between cut-off frequency (3 dB point) and the coefficient *a*. Realizations of this filter in C code and in Matlab® are shown in Appendix G.

Non-Linear Systems Analysis

The analysis methods described above all depend on the approximate linearity of the systems involved. The methods for assessing whether a system can be considered to be linear have been discussed earlier. For all those cases where the system under study is strongly non-linear, we need different methods to analyse their function. In electrophysiology, non-linear systems abound. Action potential generation of course is a very non-linear process. The fact that some ion channels open or close when the membrane potential is changed means that membrane resistances are very variable and voltage-dependent. This is quite different from the simple, constant resistors we assumed when treating filters, voltage dividers and so on.

To analyse neuronal function, both analogue potentials and action potential series can be evaluated. In the latter case, some measure of spike activity, such as frequency or instantaneous frequency, must be derived to apply systems analysis. Spike trains behave in a non-linear way very often. Even if a peripheral sense organ shows a spike frequency that is approximately linearly dependent on the stimulus, later stages of processing in the brain will generally react in a non-linear way. Many brain cells for instance are silent until some condition is fulfilled. Therefore, we need a general method to analyse and describe non-linear systems. These methods exist, but are inevitably more complicated than their linear counterparts.

The Formal Method: Wiener Kernel Analysis

Earlier, we saw that the transfer characteristics of a linear system can be derived from the impulse response. If a system is essentially linear, the impulse response can predict the response to any arbitrary signal. For instance, if we would stimulate with two pulses in succession, the response will be simply the sum of two impulse responses, taking the time between them into account. In a substantially non-linear system, however, this rule is not warranted at all. As an example, we will analyse a simple non-linear system that behaves linearly up to a certain hard boundary. In fact, *any* physical or physiological system will show this behaviour when stimulated with strong inputs. Amplifiers, for instance, are bounded by their power supplies, often + and −12–15 V. Neuronal potentials are limited usually by the equilibrium potentials of sodium (about +60 mV) and potassium (−90 mV).

The output of such a system when confronted with one of its boundaries is illustrated in Fig. 4-21.

In the linear version, the response to a second impulse is added to the remains of the earlier impulse. It is easy to understand that the shape of the response is changed as a function of both the size and the relative timing of the input impulses. The latter is shown in (B).

It can be understood intuitively that, to describe the non-linearity, one has to take both the amplitudes and the relative timing of the two impulses into account. In principle, however, an input signal is not limited to two impulses, and we should analyse the responses to three impulses with their amplitudes and relative timing, then four impulses and so on.

This seemingly tedious task can be described formally by a branch of higher mathematics called Wiener Kernels, named after Norbert Wiener and developed by Wiener after Volterra. With this method, the full response of any system may be described by a series of so-called kernels k_0, k_1, k_2, k_3 and so on. "Kernel" is mathematical jargon for a functional (a function of

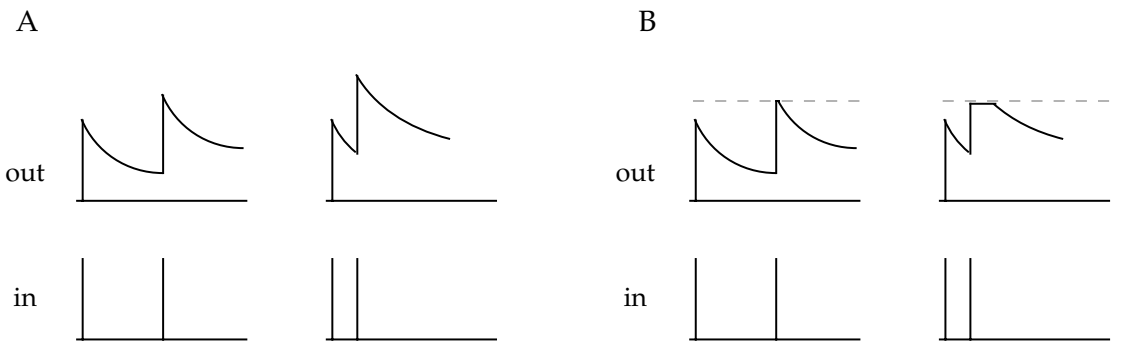


Fig. 4-21 Responses of a linear (A) and a clipped system (B).

a function). For a complete description of this method, the reader is referred to the book by Marmarelis and Marmarelis (1978). A simplified explanation is given below.

The zero-order kernel k_0 is time independent, constituting simply the DC value of the output in the absence of an input signal. The next two kernels depend on one (τ_1) and two (τ_1, τ_2) respectively time variables that determine the times of occurrence of the input impulses shown above, and can be expressed as follows:

$$G_1 = \int_0^\infty h_1(\tau)x(t-\tau)\mathrm{d}\tau$$

$$G_2 = \int_0^\infty \int_0^\infty h_2(\tau_1\tau_2)x(t-\tau_1)x(t-\tau_2)\mathrm{d}\tau_1\tau_2 - P\int_0^\infty h_2(\tau_1, \tau_1)\mathrm{d}\tau_1$$

Here, h_1 is the impulse response, hence the linear part of the transfer function, whereas h_2 reflects the non-linear interactions between two input impulses. P is the power density of the input signal. Without going into further detail, the reader can guess what the next term will look like: $G_3 = \int\int\int h_3$(form with 3τ's) . . . etc.

At first one will be disappointed to learn that any non-linearity can be described by . . . an infinite series. Fortunately, the practice proved to be hopeful. Early investigators computed the first three to four kernels, and found out that most non-linear systems, at least in physiology, can be described by the first two kernels: next to h_1, which describes the linear part of the transfer function, the non-linearities show up mostly in h_2. Higher-order kernels are said to be almost "empty" i.e. they contain a very small fraction of the energy of the output signal, hence can be neglected for all but the most demanding situations.

Next, we will discuss what Wiener kernels look like. The zero-order kernel, h_0, is time independent, and so constitutes a single number (viz. the DC value of the output). The first-order kernel, h_1, is a function of time (i.e. one time variable, τ), and so can be depicted as a line in a 2D graph of $h_1(\tau)$ versus τ. The second-order kernel h_2 is a function of two time variables τ_1 and τ_2, and so can be depicted as a surface in a 3D graph of $h_2(\tau_1, \tau_2)$ against τ_1 and τ_2. This surface depicts the non-linear interactions between two impulses. An example from neurophysiology, taken from a paper by one of the authors (de Weille), may elucidate the significance of the second-order Wiener kernel. Figure 4-22 shows the response of secondary neurons in the brain of a catfish, processing electrosensory information. These neurons are

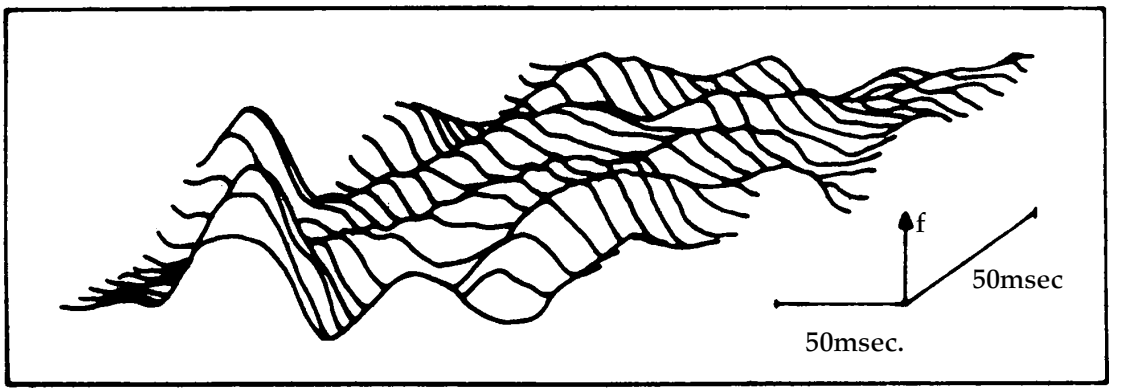


Fig. 4-22 Second-order Wiener kernel of a secondary electroreceptive neuron. Horizontal axes: time in ms, vertical axis: $h_2(\tau_1, \tau_2)$ in (arbitrary) units of spike frequency. From de Weille, 1983.

often silent, responding only to some features of the sensory input (alternating electric fields in the water surrounding the fish).

The peaks in this "non-linear landscape" show that the strongest non-linearity occurs at certain combinations of τ_1 and τ_2. At first sight, this way of reporting the non-linear behaviour of a neuron may seem strange and hard to interpret. A convincing application of this technique lies in modelling, i.e. in predicting the response of the neuron in question from the found transfer function, in this case with the non-linearity included. This is shown in Fig. 4-23.

The measured response of the neuron is simulated better if both h_1 and h_2 are used, in other words, when the approximated non-linearity is taken into account.

The Informal Method: Output Shape Analysis

Independent of this formal mathematical approach, non-linearity can be described qualitatively, viz. by a sort of taxonomic determination of the type of non-linearity involved. This means that first the type of distortion is determined by graphical inspection of the data, preferably followed by assigning a quantitative value to the degree of non-linearity.

This can be applied to the electronic instruments used in recording, but also to neurobiological subjects. Even then, examples are often taken from electronics. In an interesting, but often poorly understood paper, MacKay (1963) tried to explain the so-called power law (aka Stevens' law) found in perception research with the hypothetical non-linear behaviour of neurons.

The informal identification method of non-linear systems (processes) works by comparing the type of output non-linearity of the system in question with a palette of shapes of known non-linear processes. Often, the shape of the output signal contains strong cues to the non-linearity involved. Figure 4-24 shows graphics pairs of an I/O curve (upper graph) and the concomitant signal shape (lower graph) in time.

In the upper left corner, the figure shows the linear I/O curve, together with a sinusoidal (i.e. undistorted) signal shape. The various forms of non-linear processes lead to different forms of distortion, which can be recognized from a measured, physiological signal.

Note that a truly logarithmic curve cannot exist: it has a singularity (infinite output) when the argument (input) is zero. In physiology, however, approximate logarithmic processes do

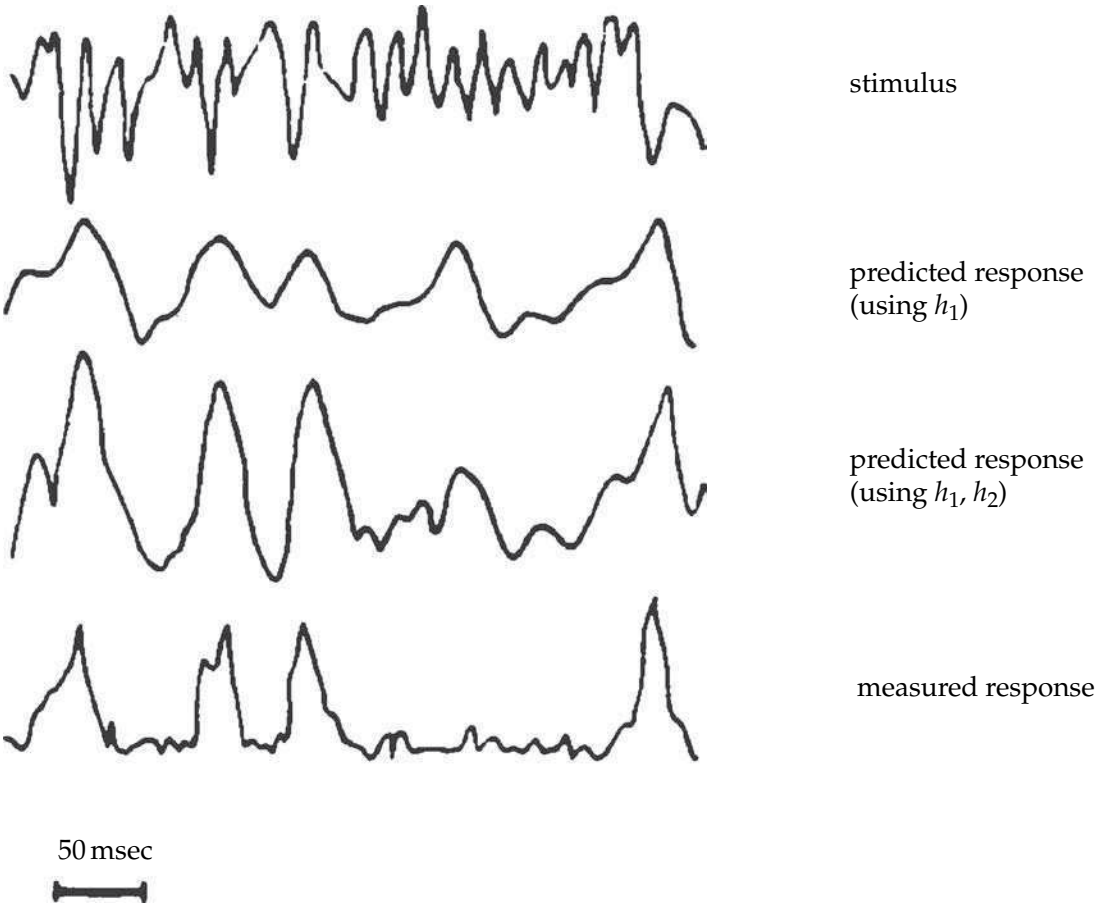


Fig. 4-23 Prediction of responses of a secondary electroreceptive neuron to a pseudo-random electric stimulus (upper trace) using either h_1 or both h_1 and h_2. Bottom trace; actual response (averaged). Response peaks about 50 spikes/s.

occur often, albeit in a limited amplitude domain. The other non-linearities shown have no singularity, but may nevertheless be limited to certain amplitude domains. The square root curve for instance is limited mathematically to non-negative values. Apart from being more complex, non-linear systems analysis is also more exciting, which we hope you will agree upon after reading the next paragraph.

The Importance of Non-Linearity

In view of the various complications when dealing with non-linear systems, it will come as no surprise that systems analysis is not very popular, at least in biology. In many cases, keeping stimuli small and exploiting the other tricks mentioned before will simplify the analysis by keeping the responses approximately linear. However, limiting ourselves to linear systems

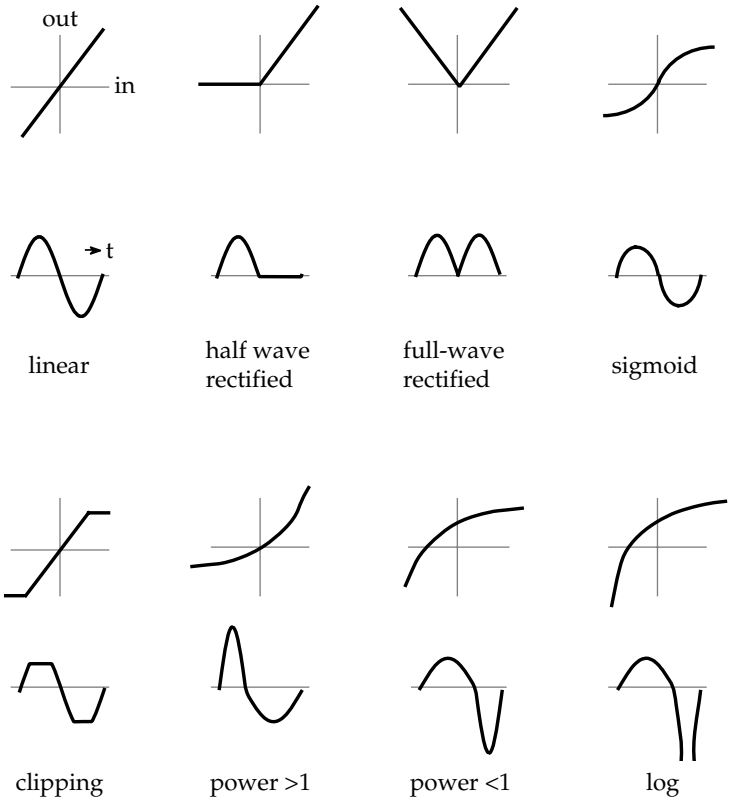


Fig. 4-24 Palette of non-linear operations. For each case, the upper graph shows the I/O characteristic, and the bottom graph the resultant distortion of a sinusoidal signal. The linear case, top left, is given for comparison.

would sell our analytical capabilities short. In fact, linear systems can be summarized largely as high-pass filters, low-pass filters and gain, or any combination thereof.

Non-linear systems can be analysed to a greater extent than linear ones. The principle is illustrated in Fig. 4-25. The upper row, A, shows how a sinusoidal signal is altered (i.e. filtered) by a cascade of a low-pass and a high-pass filter. The (qualitative) graph shows the output being both attenuated and phase-shifted. In row (B), the order of the two filters is reversed, which yields the same output. Therefore, a physiologist investigating a cascade of linear processes cannot determine, at least by input/output analysis, what the order of the components is. This is regrettable, since we want to "make the black box transparent".

The non-linear cascades shown in (C) and (D) show a high-pass filter and a clipping circuit. This yields completely different results depending on the order. If the clipping is performed first (C), the sine wave is distorted (flat tops). This distorted signal is then high-pass filtered, which leads to the odd signal shape shown. If the filter is the first segment, however, the sine may be attenuated so much that no clipping occurs afterwards (D). For physiologists, this enhances systems analysis into a tool that may reveal the order of subsystems without the need for an electrode or other physical contact in the middle, i.e. between the subsystems.

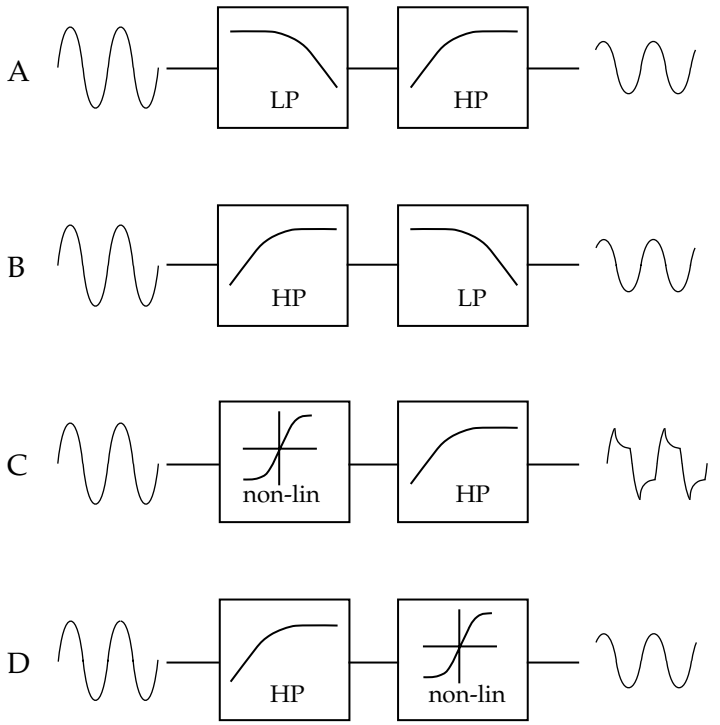


Fig. 4-25 The order of subsystems: linear vs. non-linear. The order of linear subsystems (A and B), in this case a low-pass (LP) and a high-pass (HP) filter, cannot be determined by I/O analysis, since they yield identical outputs. If one or more non-linear processes are involved, the output can depend on the order of the subsystems.

ANALYSIS OF ACTION POTENTIAL SIGNALS

Except for unicellular species such as amoebae and paramecia, all animals need to convert neural signals into trains of action potentials to be able to send them through long nerve fibres. It is tempting to call the spike code a digital code, to segregate it from the clearly analogue potential variations that arise in sense organs, neuronal cell bodies and so on. Indeed, given the regenerative mechanism that restores the amplitude of an action potential everywhere along the nerve fibre, the amplitude may be considered a binary signal: at any moment, there is either a spike or no spike.

However, most series of action potentials are continuous in time. The information being transmitted remains "analogue" in the sense that it may be encoded in spike density (spike frequency modulation). The physiologically important parameter may also reside in the temporal delay or coherence with respect to the electrical activity of other neurons converging onto the same target neuron. In any case continuously varying spike intervals contain the information of interest.

Before tackling the analysis of spike signals, the distinction must be made between single-unit and gross activity. In most cases, one records, with a suitably small electrode, the activity of a single nerve fibre. The fibre may be connected to a sensory cell or a neuron, and is usually called a unit. Hence the name single-unit recording. In other cases, especially with extracellular

recording, the signal contains more than one spike train. A multi-unit signal, which contains a small number of spike trains, is very confusing and often of little use. In some cases, the spike trains can be segregated by their respective amplitudes. If the fibres have different distances to the recording electrodes, their spike trains will have different amplitudes (Fig. 4-26) and so can be segregated by a computer algorithm.

If one records from a whole nerve, the number of units contributing to the signal is so large that they cannot be segregated. In fact, the resultant, so-called gross activity signal behaves as a form of noise.

Population Spike and Gross Activity

"Gross activity" is the term for the joint activity of tens, hundreds or even thousands of fibres present in a single nerve or muscle. In some special cases, the activity (firing) of all fibres is synchronous, for instance if the whole nerve or muscle is stimulated. This is often the case in clinical applications, where the functioning of nerves and muscles after a trauma can be assessed by recording the activity elicited by pulses from external (i.e. transcutaneous) stimulation. In this case, the summed spike signals may be recognizable as a relatively large spike, as if it were a single unit. This is called a population spike, or compound action potential (CAP): see Fig. 4-27. The sharpness of the resultant spike is determined by the degree of synchrony. If the different fibres differ in conduction velocity, the spikes in parallel fibres tend to diverge, so that the best signal is obtained close to the source (stimulus electrode).

Under more natural circumstances, however, spike trains in parallel fibres tend to be independent. In this case, a random, noise-like signal called gross activity arises. More precise, it has the properties of the so-called shot noise, each spike being a "shot" (see the Chapter 2). The amplitude of this noise is a good measure of the average spike activity in the nerve or muscle measured. Therefore, although the spike frequencies cannot be recorded directly, the spike code can be analysed to a fair extent. The desired amplitude can be found by a very simple circuit called a "diode pump". It consists of a diode and a low-pass filter. The output of such a circuit shows the average spike activity in a whole nerve (Fig. 4-28), provided the input signal is strong enough to overcome the diode threshold voltage of about 0.7 V (see Chapter 2, diodes). Alternatively, computer algorithms performing essentially the same operations may be used with signals of arbitrary amplitude.

The functions of some hitherto unknown sense organs have been found by gross activity recording (Fig. 4-29).

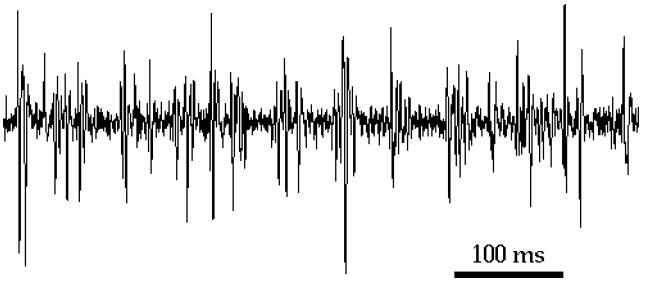


Fig. 4-26 A multi-unit recording from a small muscle (ball of the thumb).

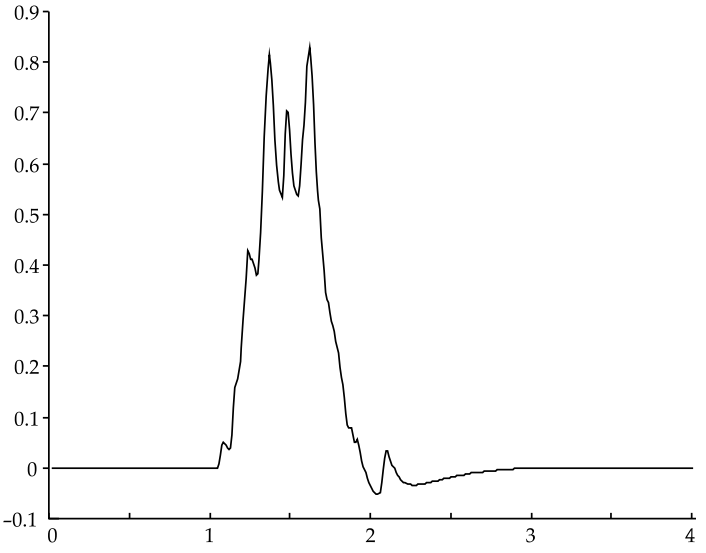


Fig. 4-27 A population spike (simulated).

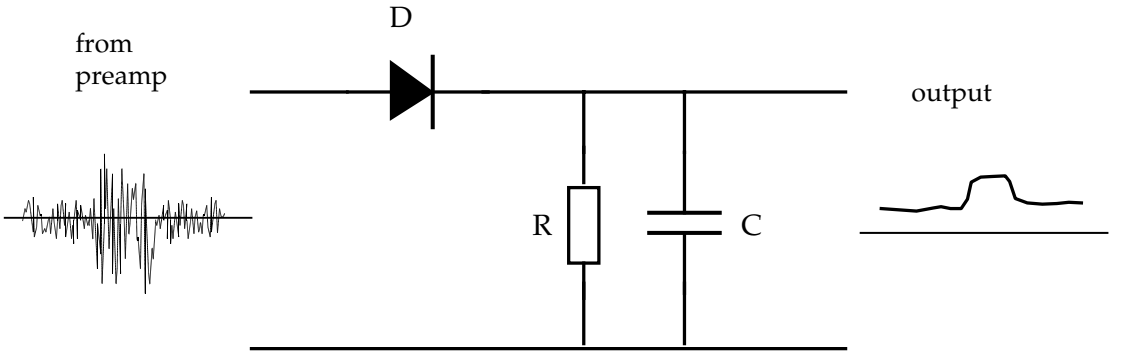


Fig. 4-28 Gross-activity analysis by a "diode pump".

Recording from the Skin Surface

The Electrocardiogram

The recording method best known to the public is surface recording from the human skin. This started with the pioneering work of Einthoven, among others, who recorded the human electrocardiogram (ECG) using large metal plate electrodes connected to the arms and legs of the patient. The only recording device at that time was the string galvanometer, the time movement of the wire being recorded on photographic paper with a built-in microscope.

Today the ECG, recorded with a variety of recording methods, is an indispensable diagnostic method in the heart clinic. The basic set-up for a routine ECG is still the so-called Einthoven triangle, using recording electrodes on the two arms and one leg of the patient (Fig. 4-30 left). This type of recording yields the familiar ECG shape with its sharp QRS peak and the more

Fig. 4-29 Gross-activity recording of a nerve that proved to innervate a novel type of chemoreceptor in the skin of a fish (from Peters & van Steenderen., 1987; courtesy of Dr R.C. Peters). Bar 1 second.

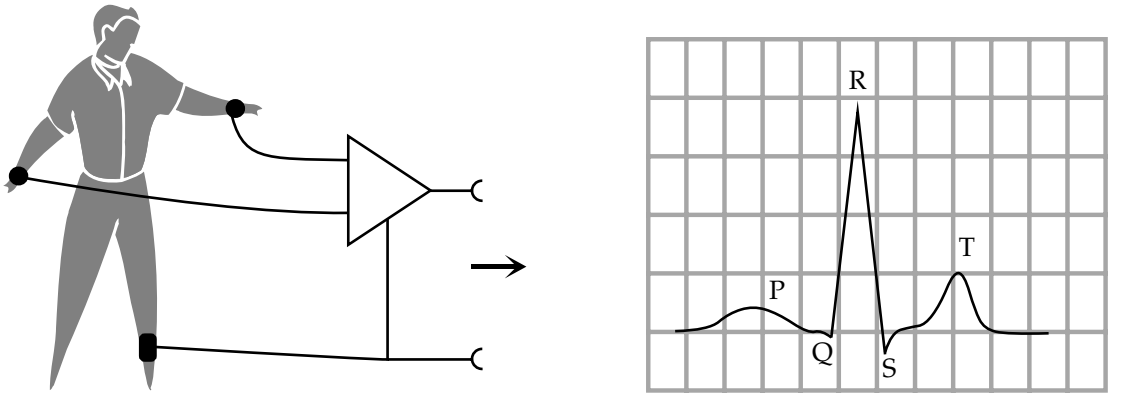


Fig. 4-30 ECG recording with the Einthoven triangle (left) and the resulting ECG waveform (right). The amplitude of such an ECG is about 1 mV.

gradual P and T components (Fig. 4-30 right). Despite the large distance between the heart and the surface electrodes, the ECG is a form of single-unit recording. However, the shape of the ECG is largely determined, not by the shape of the cardiac action potential, but by the way the action potential, initiated in the sinoatrial node and propagated through all heart chambers, moves through space. Each phase of the ECG is an indicator of one of the electrical processes that together cause the heart beat, and so, signal processing is usually limited to visual observation by an expert.

For a more detailed picture of heart functioning, so-called precordial electrodes are used, i.e. a row of electrodes on the chest, and so in front of the heart.

The electrical demands for ECG recording can be derived from the electrodes and the amplitude of the signal obtained. ECG electrodes are usually fairly large metal electrodes made of stainless steel, nickel or silver, and connected to the skin with a salty, well conducting electrode gel. The relatively large surface area, together with the salt, warrants a low resistance

(more correctly impedance) and so does not put severe demands on the input impedance of the ECG amplifier: about 1 MΩ suffices. The voltage gain must be fairly high (say 1000×), and filters must be built in to block hum and reduce noise. The necessary bandwidth is about 1 to 30 Hz.

However, recording from human beings implies an electrical contact between the patient and the electrical apparatus. Since the latter is usually powered by the local mains, special medical instruments are needed, i.e. extra insulated and protected amplifiers, recorders, etc. The problems of medical recording and the specifications of the suitable equipment is discussed in Appendix C.

The Electroencephalogram

Like heart activity, electrical processes from the brain can be recorded from the outside. In this case, however, one does not record from a "single unit". To the contrary, the electroencephalogram (EEG) is the ultimate of gross-activity recording. Electrodes on the scalp, usually an array of a dozen or more, will record the average activity of millions of neurons, from large parts of the brain. Nevertheless, a functional segregation of brain functions can be made by choosing electrode positions carefully. The waveforms recorded now do not reflect the spike frequencies of some brain cells, but rather the degree (and timescale) of synchronization of large populations of brain cells. If all brain cells would fire uncorrelated, the result would again be a form of shot noise treated earlier, and so would show an almost "white" spectrum (the bandwidth limited mainly by the spectral contents of spikes). This is not the case with the EEG, where often a single frequency (or a narrow band of frequencies) is most prominent. Even in the earliest EEG records, made by Berger in the 1920s, a few prominent frequencies can be seen (see Fig. 4-31). These rhythms were designated Greek letters, which in part are still

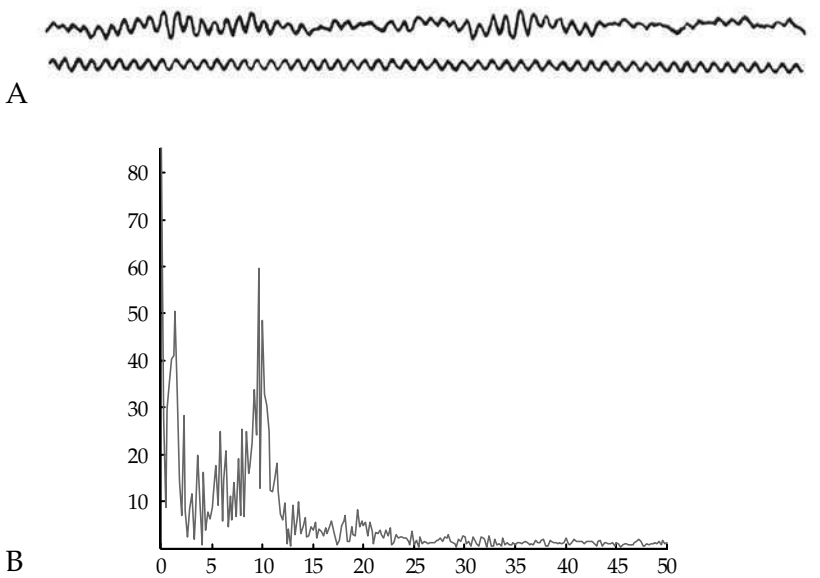


Fig. 4-31 (A) An EEG by Berger of his 15-year-old son, together with a 10 Hz reference signal (lower trace). From Berger, H. (1929). (B) Amplitude spectrum (0–50 Hz) computed from the graph, showing peaks at about 1 and 10 Hz.

in use today: the alpha rhythm, around 10 Hz, during resting when awake; the beta rhythm (wideband, irregular) when performing mental tasks; delta waves (about 4 Hz) during deep sleep. All in all, EEG signals span a frequency band of about 1 to 50 Hz.

EEG electrodes are usually somewhat smaller than their ECG counterparts (often small metal discs with or without silver chloride coating), but are still large enough to have a fairly low impedance (a few kΩ at 50 Hz). Gain demands for an EEG amplifier are however very high, because the signal is usually only a few tens of μV in amplitude. Limiting the bandwidth with higher-order filters is also important.

Both time domain and frequency domain are important for the processing of EEG signals. Like cardiograms, EEGs are usually analysed visually by an expert. Some types of EEG show characteristic patterns called "spindles" in the time domain, where the amplitude is waxing and waning gradually. In addition, the amplitude spectrum, obtained by Fourier-transforming the encephalic signal, may give additional cues.

The EEG described so far is a spontaneous activity of the brain, depending in a global way on the mental state of the subject. Responses to sensory stimuli, which may give more specific answers as to sensory processing, can also be recorded. However, since the EEG stems from a huge number of neurons, responses to specific stimuli are usually too weak to be recognized from a single record. This is why such evoked potentials, more generally dubbed event-related potentials (ERPs), only become visible after substantial signal averaging (the event is not necessarily a sensory stimulus: it can also be a motor action of the subject). Of course, the averaging can be triggered on a fixed point, such as the start of the stimulus (an acoustical beep, a light flash, a brief touching of the skin, etc.) or the command for an action. An example is shown in Fig. 4-12. Averaging some 64–256 sweeps usually reveals a significant electrical response. Size, latency and other quantities may then be related to the processing of the stimulus in question by the brain.

Other Surface Recording Techniques

A number of other internal electrical activities of humans and other land animals can be recorded from the skin, most of them used again for clinical diagnosis. Although some of these signals stem from DC processes in the body rather than from action potentials, it is useful to treat them here together.

Examples are the electromyogram (EMG) from one or more muscles and the electronystagmogram, the recording of eye movements by electrodes at the temples. The multi-unit recording shown in Fig. 4-26 is an example of an electromyogram.

The technique indicated with the term "lie detector" also falls in this category. It is actually a measurement of the resistance of the skin of the hand palm, which changes under the influence of the autonomic nervous system (i.e. the control of sweat glands). Although the term "lie detector" is a misnomer, the palmar skin resistance is a useful measurement, since it reflects the balance of (ortho-) sympathic and parasympatic activities. It is thus a recording of covert behaviour, and as such useful in medical, psychological and psychophysical research. Usually, skin resistance recording is combined with the recording of other physiological signals, such as the ECG, breathing rhythm, an EMG of postural muscles and so on. The DC skin potential may also be used. The combination is called polygraphy and may give a better representation of the physiological state of the subject than each single quantity.

Note that polygraphy does not necessarily mean poly-electrode: the scheme in Fig. 4-32 describes a device we built for recording the ECG, EMG, skin resistance and skin potential

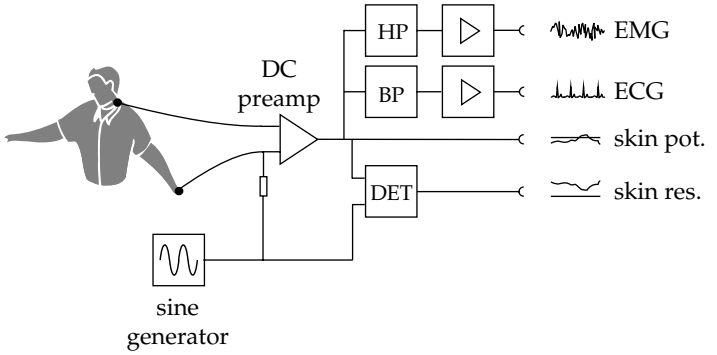


Fig. 4-32 Simplified schematic of a polygraph using two electrodes. Ground wires etc. are omitted. HP high-pass filter, BP band-pass filter, DET detector (converst amplitude of AC into a DC).

simultaneously with only two electrodes connected to the patient. The skin impedance is measured by supplying a sinusoidal current to an electrode on the hand palm.

The trick is done further by positioning the other electrode on a neck muscle, and segregating the signal components by electronic means: a DC pre-amplifier for the skin potential, a detector/amplifier (DET) to derive the skin impedance from the AC signal, a low-frequency band-pass filter (BP) and post-amplifier for the ECG and a high-pass filter (HP) with post-amplifier to detect the EMG.

The quantities skin impedance and electromyogram are processed with circuits like the one showed in Fig. 4-28. The resulting "polygrams" are usually interpreted directly from the chart records, i.e. in the time domain.

Two final remarks to skin surface recording:

1. All types of skin surface recording benefit from an electrically screened set-up. This is why the walls of ECG and EEG recording rooms are often covered with metal to form a large Faraday cage.
2. Often, human skin surface recordings are made with dedicated, medical-grade apparatus that fulfil the electrophysiological demands, including the safety aspects dealt with in Appendix C. Portable versions of ECG and EEG recorders are used by general practitioners to diagnose at the patient's bed. These apparatus have still better filters built in to yield clear results despite the lack of a screened room.

Single-Unit Activity

The simplest situation to illustrate the problem of analysing action potential series is the example shown in Fig. 4-1: a sense organ picking up some quantity from the environment and encoding it in the form of an action potential series, or spike train for short. The encoding of light flashes by a photoreceptor and its subsequent encoding as a spike train is only the beginning of a long journey of that signal trough several parts of the peripheral nervous system and the brain. During that journey, the form of the signal will alternate many times between an analogue potential (receptor potential, postsynaptic potential) at the points where the signal must be processed, and a spike train in the long-distance transport sections.

The intrinsic properties of the spike code determine the fate of any neural signal but, in addition, they will determine which analysis methods are suitable for physiologists to hunt down the functioning and the function of the studied neurons.

Uncertainty and Ambiguity in Spike Series

Formally, a series of action potentials (spike train) is a signal continuous in time. However, for most analyses of neural signals, the duration of each spike can be neglected, and only the timing of the pulses carries information. Moreover, most spike trains are irregular, making statistical analysis necessary. Therefore, a time series of action potentials is considered to be a stochastic point process. A point process because it exists only at certain moments in time, and stochastic because fluctuations are conspicuous. As we will see later on, the fluctuations may even form an essential part of the code.

A typical spike train might look like Fig. 4-33.

This signal can be considered as a regular firing frequency modulated by noise. Both components may be characterized by well-known statistical measures such as mean and standard deviation. In the nervous systems, such a signal might represent the activity of a sense organ, the contraction force of a muscle fibre or the activity of a neuron in the brain. Apart from the fluctuations, the spike train will be modified and varied in time. In a sense organ, this might stem from a changing stimulus (any important quantity in the environment); in a motor circuit controlling a muscle, it stems from instructions from the brain to change the activity of that particular muscle fibre (or motor unit).

Figure 4-34 shows a spike train modified by a change in stimulus/activity level.

Note that the spike code *per se* introduces an uncertainty, since a sudden change in activity will only be reflected in the spike train at the moment of the next spike. Between two spikes, the signal is uninformative. The ubiquitous fluctuations cause a second type of uncertainty: the central nervous system must perform some kind of statistical analysis to find out whether the change has a relevant cause (such as a change in ambient light intensity) or is a mere fluctuation.

In addition to uncertainty, a spike signal may be ambiguous. An example of this ambiguity occurs in directional hearing, which depends partially on timing differences between the sound-evoked spikes from the left and right ears. This is illustrated in Fig. 4-35 for a sound source situated to the right of the listener. In most situations, our brain is capable of deriving the direction of the sound source from the subtle differences in arrival times of spikes from our

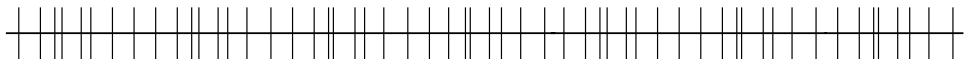


Fig. 4-33 A spike train.

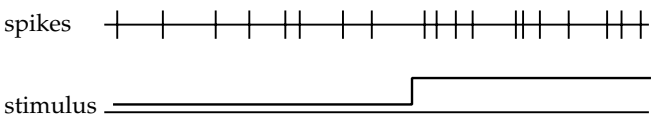


Fig. 4-34 Uncertainty at a change in activity level.

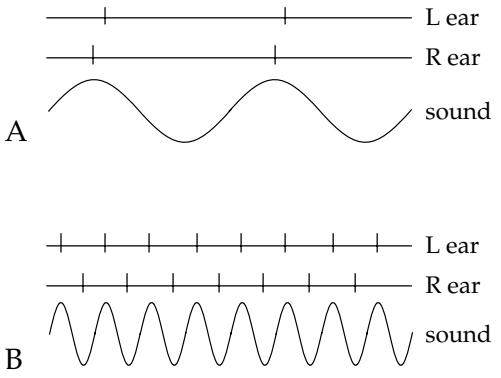


Fig. 4-35 Ambiguity in directional hearing. (A) Timing difference allowing perception of the direction of the sound. (B) Ambiguity, leading to a diffuse sound sensation.

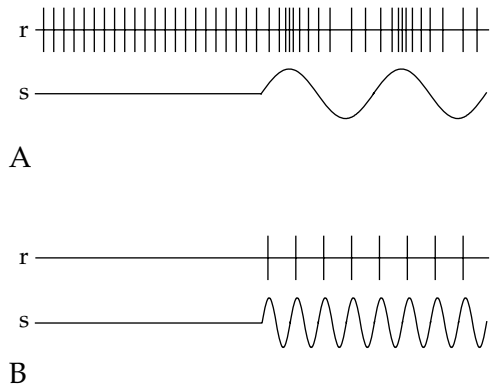


Fig. 4-36 Spontaneously active versus silent nerve cells.

left and right ears. This is shown in A. With neuronal circuits, the short delay between spikes from the right and left ears can be translated into a sensation of direction, and be distinguished from the much longer delay left–right. At a certain frequency the two intervals become equal (shown in B) and the perceived sound direction is getting indeterminate: the sound seems to stem from a broad front.

In sense organs, two main types of activity can be distinguished: silent and spontaneously active. Some sensory cells and fibres show spontaneous activity, i.e. in the absence of a stimulus, a more or less regular spike train is visible. The spontaneous rate is modulated by the sensory input. Other sense organs are silent at rest, and only fire when a certain threshold is crossed. These two cases are illustrated in Fig. 4-36.

The fluctuations in spike timings can be very large, and occur in virtually every nerve fibre. It may seem purely detrimental to the value of spike signals in encoding real-world quantities. However, it is not always true: the fluctuations may serve to overcome certain disadvantages of the point process code.

Suppose, for instance, that a neuron from the ear fires in response to a sound stimulus at a certain (not too high) frequency. If the spike code were entirely deterministic (i.e. without any fluctuation), the following would happen. At very low intensities, no spikes fired at all. Above threshold, a single spike per period would appear. A further increase would not influence the spike signal, until the point where the sound amplitude reached a second threshold, above which two spikes per period would appear. Hence, a strictly regular spike train would code amplitude in a digital (discrete) way. More subtle amplitude changes would go unnoticed. Enter the noise: because of the ubiquitous fluctuations, the situation is very different. At extremely low sound intensities, again no spikes will appear. However, even at relatively low intensities, a spike will appear occasionally. The stronger the sound, the higher the *probability* that a spike will be fired, until it is about one spike per period. At that point, an occasional second spike per period will occur, and so on. This is shown in Fig. 4-37. In this case, the amplitude is sensed in a continuous, threshold-less way, *provided the c.n.s. is allowed to average the spike signal*; either in time or over a number of parallel fibres. In reality, both ways of averaging occur.

Whatever the part played by the fluctuations, electrophysiologists investigating a spike signal face the task of separating the deterministic changes from the stochastic ones. Fortunately, there exists a plethora of methods to extract relevant information on the functioning of neurons and neuronal networks.

The basic ones will be treated below.

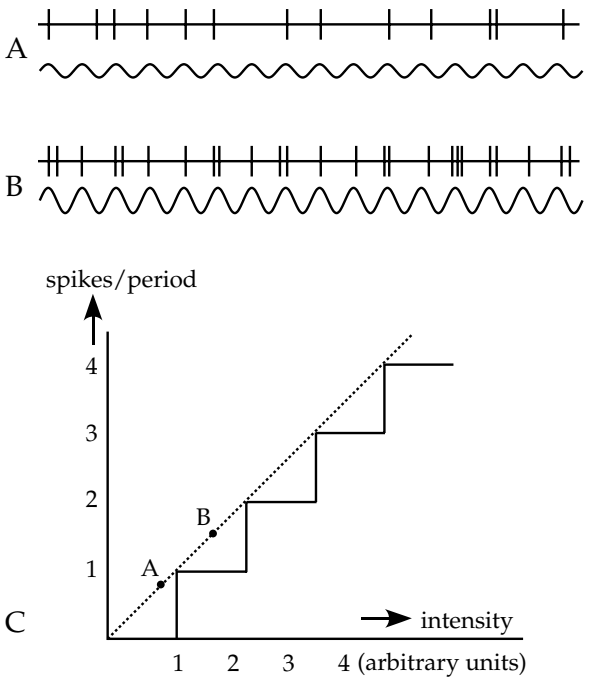


Fig. 4-37 The advantage of fluctuations in a spike signal. (A) Sound intensity where about 0.8 sp/period arise. (B) A stronger sound yields about 1.6 sp/period. (C) The input/output curves with (straight line) and without (staircase) fluctuations.

Interval Histogram

A number of spike pattern analysis methods is based on the interval time between spikes, the "spike interval" or, more precisely, the "interspike interval" (formally, the duration of the action potential is also a time interval; the intra-spike interval). The most compact description of a (large) number of interval values is, of course, the classical statistical set of moments. The "mean", also known as the "first moment", is the single most important characteristic. This describes the deterministic content in the signal, but gives no clue as to the scatter, or degree of (un)reliability of the spike train. Therefore, we need at least one other measure such as the "standard deviation". This is the most used (and misused) measure of scatter. It is called the "second central moment" formally because it reflects the weighted (root mean square) deviation of the data from the mean. Of course, this holds only for data series that follow approximately a normal distribution. If a distribution of interspike intervals is markedly skewed or peaked (i.e. a distribution that differs significantly from the normal, or Gaussian, distribution), one might need further moments, such as the "third central moment", called "skew" for short, and the "fourth central moment" or "kurtosis" (peakedness). Details and formulas can be found in most books on statistical methods.

Unfortunately, spike interval distributions are often skewed, which has several reasons and depends on the spike-generating mechanism. The two extremes may illustrate this: on one end a perfectly regular spike train entirely determined by a pacemaker mechanism, and at the other end a highly irregular, entirely random one. The first case is approximated by, for example, the cardiac pacemaker, and by some sensory receptor cells and fibres. A perfect pacemaker would have a very unusual, spike-shaped interval histogram: all values in one bin. In real-world pacemakers, however, there is always some noise, which stems from the random opening of ion channels, a fluctuating number of synaptical vesicles and other molecular processes. This causes the interspike intervals to fluctuate around a certain mean value, often giving rise to an approximately Gaussian distribution. This situation, best described as "largely deterministic plus slightly random", is shown in Fig. 4-38, trace 1. Finally, one can imagine an entirely random interval series, which is again approximately the case with some real-world neurons.

The reader may wonder why action potentials, arising successively in one and the same physical structure (a neuron), can be found to be largely independent from one another. In general, any physical structure will have some degree of inertia, or memory. Indeed, in most neurons, the degree of depolarization determines the overall firing rate or firing probability, and it will always take some time to change the membrane potential and the firing rate. Depending on the duration of this dependence, successive spikes may show either more or less interdependency, the degree of which can be assessed by statistical inference.

However, the high degree of independence of successive spikes in some neurons can be explained by two facts. First, far more ion channels are open during an action potential than in between the spikes. In other words, each spike causes a brief shunting of the local membrane potential; a kind of "reset" pulse erasing the "memory" of past situations to a certain degree. The degree of shunting is determined by the shape of the neuron soma and spike-initiating zone and by the number and distribution of voltage-sensitive ion channels. This partial shunting, or partial reset, explains most irregular spike firing. Second, interspike interval times are often longer than the membrane time constants involved, so the electrical "memory" is very short.

Most neurons show a compromise between the extremes of a pacemaker, having a regular spike rate, and a noisy, virtually independent spike generator. Figure 4-38, trace 2, shows a spike train recorded from a sensory neuron and its interval histogram in the absence of

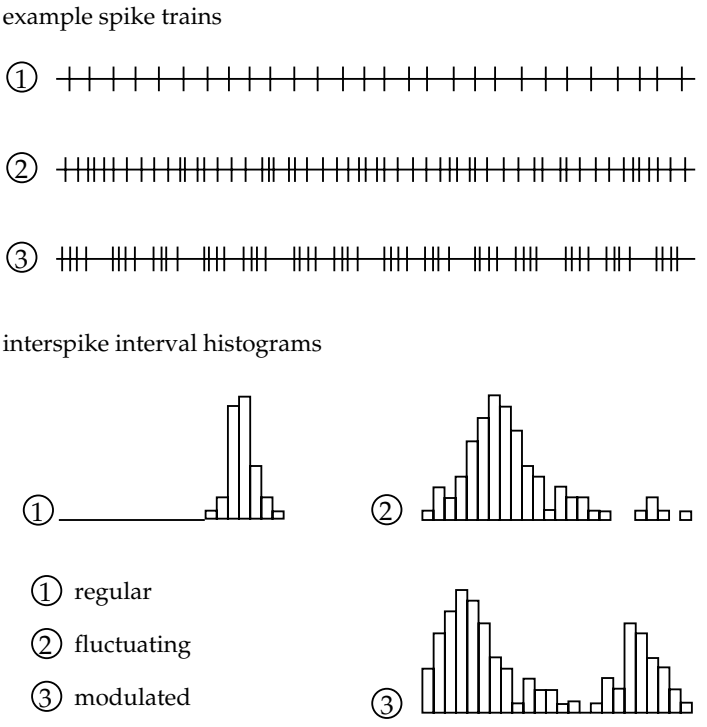


Fig. 4-38 Schematic representation of sample spike trains and their respective interval histograms: (1) a fairly regular series with a narrow, approximately normally distributed interval distribution; (2) a more irregular spike train with a skewed histogram; (3) A spike train modulated by a deterministic signal (sinusoid) and its histogram.

stimulation. Contrary to trace 1, this represents the situation "largely random plus slightly deterministic". The histogram is skewed with a tail at the long-interval side. Trace 3 shows the effect of a periodical, i.e. deterministic modulation of the spike rate.

Poisson Processes

To explain why spike interval distributions are often skewed, we need to delve a bit deeper into the statistical principles underlying spike generation. At relatively long spike interval times and/or a high degree of shunting, subsequent spike occurrences are virtually independent. In statistical theory, a series of mutually independent random events forms a relatively simple and well-known case known as a "Poisson process". The number of events per time unit (i.e. the rate) of such a process follows the Poisson distribution. This is a skewed distribution, well known from most books on fundamental statistics. A histogram of a Poisson-distributed spike series is shown in Fig. 4-39 left.

However, spike analysis is performed usually at the single interval level (see below). In a Poisson process, the histogram of the interspike interval lengths has the shape of a negative exponential. This is shown in Fig. 4-39 right.

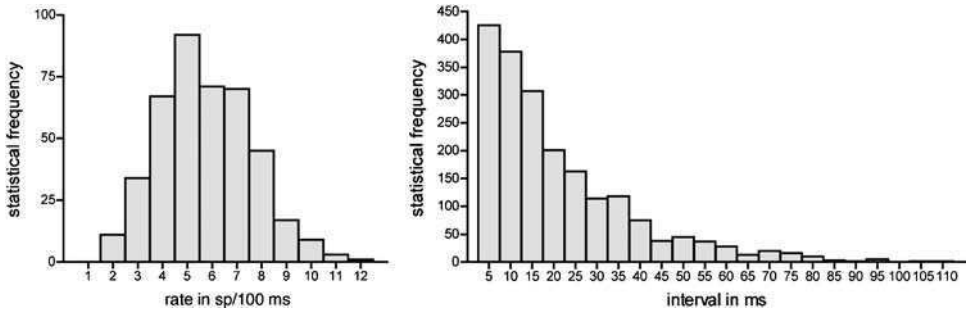


Fig. 4-39 Left: Rate histogram of a Poisson-statistics spike train. Right: Interspike interval histogram of a simulated Poisson (i.e. independent) spike interval series.

The exponential distribution has only one parameter: the mean frequency, usually called λ. The amount of scatter reflected in the standard deviation (the second central moment) is simply $\sqrt{\lambda}$.

Thus, in principle, one could expect the interval histogram of some spike series to follow this exponential distribution. Most interval histograms made from real spike trains, however, have different shapes. This is shown in Fig. 4-40. Graph A shows the exponential distribution discussed above. A first modification would be caused by the refractory period, which is inherent in the neuronal spike-generating mechanism. This would cause the distribution to shift to the right (graph B).

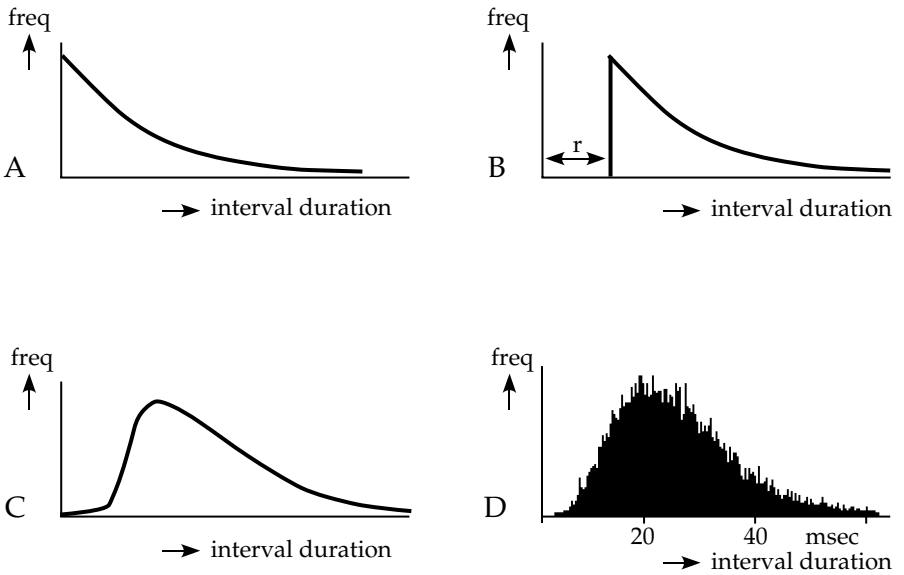


Fig. 4-40 Theoretical interval distributions: (A) exponential distribution; (B) the same, but taking the refractory period into account; (C) a gamma distribution; (D) a spike train from a sensory receptor fibre (ampullary electroreceptor of a catfish; see Teunis *et al.*, 1991).

A more fundamental consideration is that the spikes may indeed be mutually independent, and that, in addition, each spike is the result of a process in which a number of independent events is necessary to trigger one spike. One might think of small postsynaptic potentials, where a spike is only elicited after, say, three to four psp's in a short time span. In this case, the distribution followed is a so-called gamma distribution, which derives its name from the gamma (Γ) function, which is part of its mathematical description.

A gamma distribution shifted by the aforementioned refractory period is shown in Fig. 4-40C. Graph D, finally, is an interval histogram made from neuronal data (single-unit recording of spontaneous activity of an ampullary electroreceptor organ).

The Gamma Distribution

A "gamma distribution" has two parameters, called λ and r. The λ is a scale parameter, reflecting the time scale (mean spike rate). Gamma distributions with different λ's have essentially the same shape, only stretched more or less on the time (interval duration) axis. The parameter r is called the "shape parameter". The example in Fig. 4-40C has $r = 4$, which means that four independent events are needed to cause each spike. Thus, the overall shape of a gamma distribution is determined by r. For $r = 1$, the gamma distribution reduces to the exponential distribution treated earlier, and so is skewed strongly. As $r \to \infty$, the gamma distribution approaches the Gaussian ("normal") distribution, and so gets more and more symmetrical. Most interval histograms can be fitted with a gamma distribution. By the way, this does not necessarily mean that the underlying cause is clear.

The Mathematics of Random Point Processes

Finally, the shape of interval histograms can be treated in a more general way.

Consider a process that generates random events at an average rate of μ events per second. Further, suppose that the probability of an event occurring is independent of time. Divide an epoch of duration t into m bins of equal lengths, with m being sufficiently large to prevent the occurrence of more than 1 event per bin. Then the process is a homogeneous Poisson point process. If μ is a function of time, the process is an inhomogeneous Poisson process that we will not discuss here. The bin size is:

$$\delta t = t/m$$

The probability of having one event per bin ($P(1)$) is $\mu\delta t$ and the probability of having no event in a bin ($P(0)$) is $1 - \mu\delta t$. It follows that the probability of having no event in the epoch of duration t is:

$$P(0) = (1 - \mu\delta t)^m = (1 - \mu t/m)^m$$

Taking the limit of $P(0)$ for increasing m gives:

$$P(0) = \lim_{m\to\infty} (1 - \mu t/m)^m = \mathrm{e}^{-\mu t}$$

The probability to get 1 event in any of the bins is:

$$P(1) = m\mu\delta t(1 - \mu\delta t)^{m-1} = \mu t(1 - \mu t/m)^{m-1}$$

Taking the limit of $P(1)$ for increasing m:

$$P(1) = \lim_{m\to\infty} \mu t(1-\mu t/m)^{m-1} = \mu t\ \mathrm{e}^{-\mu t}$$

The time it takes before the first event occurs is called the "waiting time" and its associated probability density function is called the waiting time distribution, W. Suppose the event occurs in the last, mth bin, the other bins 1 through $m-1$ being empty, then the probability,

$$\begin{aligned} \delta P(1) &= \mu\delta t(1-\mu t/m)^{m-1} \quad \text{or :} \\ \delta P(1)/\delta t &= \mu(1-\mu t/m)^{m-1} \end{aligned}$$

Taking the limit of $\delta P(1)/\delta t$ for increasing m gives the probability density:

$$W(1) = \mu \mathrm{e}^{-\mu t} \qquad \text{(Eq. 4-7)}$$

Hence, the event intervals (e.g. spike intervals) have an exponential probability density distribution; the intervals are said to be Poisson distributed. From this equation, the mean duration between two successive events can be obtained:

$$t_{\mathrm{mean}} = \int_0^\infty t\ \mu \mathrm{e}^{-\mu t}\mathrm{d}t = 1/\mu$$

which, as expected, is 1 over the event rate and the variance is:

$$\sigma^2 = \int_0^\infty t^2 \mu \mathrm{e}^{-\mu t}\mathrm{d}t = \frac{1}{\mu^2}$$

The ratio of the standard deviation to the mean interval is called the "coefficient of variation", denoted by c_{v}:

$$c_{\mathrm{v}} = \sigma/t_{\mathrm{mean}} = 1$$

The c_{v} characterizes the variability in the event intervals. A distinguishing feature of a homogenous Poisson process is that $c_{\mathrm{v}} = 1$.

The probability of the occurrence of n events, with the nth event in the last, mth, bin and with $n \ll m$ is:

$$\delta P(n) = \mu\delta t(1-\mu t/m)^{m-n}(m\mu\delta t)^{n-1}/(n-1)!$$

and hence,

$$\delta P(n)/\delta t = \mu(1-\mu t/m)^{m-n}(m\mu\delta t)^{n-1}/(n-1)!$$

Here the factor $m^{n-1}/(n-1)!$ relates to the distribution of $n-1$ events in m bins. Now, after taking the limit for increasing m, the probability density $W(n)$ is obtained:

$$W(n) = \lim_{m\to\infty} \mu \frac{(1-\mu t/m)^{m-n}\cdot(m\mu\delta t)^{n-1}}{(n-1)\,!} = \frac{\mu^n t^{n-1}\mathrm{e}^{-\mu t}}{(n-1)\,!} \qquad \text{(Eq. 4-8)}$$

This waiting time distribution, which describes the waiting time until the nth event, is also known as the standard gamma distribution. It is easy to verify that $W(n)$ reduces to Eq. 4-7 for $n = 1$.

When fitting a function to experimental data, it is often desirable to have floating point parameters to fit, rather than integers. This can be achieved for Eq. 4-8, by replacing the $(n-1)!$ faculty with the gamma function $\Gamma(n)$, which is a kind of faculty for floating point numbers:

$$W(n) = \frac{\mu^n t^{n-1} e^{-\mu t}}{\Gamma(n)} \qquad \text{(Eq. 4-9)}$$

where n may now be a floating point.

The interpretation of Eqs. 4-8 and 4-9 is that each spike in a train may be caused by a process in which a number of independent events are necessary to trigger one spike. One might think of small postsynaptic potentials (psp's) where a spike is only elicited after, say, three or four psp's in a short time span.

Markov Chains

The assumption in Eq. 4-8 is that all events leading to the final observable nth event have, in a probabilistic sense, identical intervals, i.e. the rate constants all equal μ. If we wish to drop that restriction we could think of the problem in the following way:

$$S_1 \underset{\mu 1}{\rightarrow} S_2 \underset{\mu 2}{\rightarrow} \cdots \rightarrow S_{n-1} \underset{\mu(n-1)}{\rightarrow} S_n \qquad \text{(Eq. 4-10)}$$

Starting from state 1, S_1, the system evolves through different states until it reaches the observable state S_n. The transitions between states are Poisson distributed, but now having different rate constants μ_i. Such a series is called a "Markov chain".

More mathematics of Markov chains can be found in Appendix F.

Time Series Analysis: Spike Rate, Interval Series and Instantaneous Frequency

Apparently, histograms show several interesting aspects of neural functioning, but they fail to show one fundamental quantity: time. Therefore, we will focus now on the analysis of spike trains in time.

Spike Frequency or Rate

The most intuitive way of analysing a time series of events is to measure the frequency, i.e. the number of events in a predetermined time interval. This is shown in Fig. 4-41. An irregular

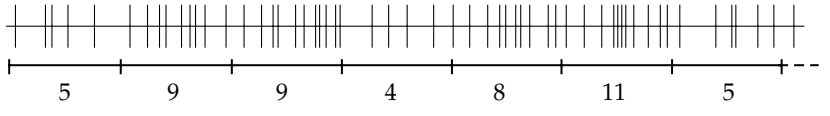


Fig. 4-41 Spike rate or frequency.

spike train is divided into a number of fixed time intervals. The number of spikes counted in each interval is a measure of the spike frequency, also called the "spike rate". The distinction between rate and frequency may need an explanation. The well-known notion of a frequency stems from the harmonic (sinusoidal) motion, in which the motion (or any other physical quantity) is fluent. If the frequency is doubled, so are all aspects of the signal, such as speed and acceleration. In the frequency domain, the whole spectrum is shifted by a factor of two. In a spike train, however, the duration of the individual events is always the same, and so is independent of the number of events per time unit. In this case, it is better to call the number of events per time unit a "rate" rather than a frequency.

To amplify neural spikes, an amplifier must have its *frequency* bandwidth centred at about 1 kHz, irrespective of the spike *rate*. Even one spike per hour would still be a short pulse. Obviously, there must be differences in the frequency domain too, but these are far smaller than in the above-mentioned case of a sinusoid. Since both frequency and rate are expressed in the quantity s^{-1} or Hz, it has been proposed to use the name adrian for the unit of spike rate, in honour of the famous pioneer electrophysiologist Lord Edgar Adrian. The name did never catch on and, since the frequency content of spike trains is known well enough by the people recording them, using the term "spike frequency" poses no problem in practice. The most common expression is "spikes per second", abbreviated "sp/s".

The series of values shown in Fig. 4-41 demonstrates the disadvantages of spike rate as a measure of neuronal activity. If the measuring interval is taken rather short, as in the figure, the number of spikes per measurement is very small, yielding a rather crude measure of the signal. If the measuring window is taken longer, say one or a few seconds, the measured rates are more precise (say between zero and a few hundred spikes per second), but now the time resolution is unacceptably low. Even a snail could not survive when informed about its environment only once every couple of seconds.

A number of problems in neural coding arise because the frequency of the "carrier" (the average spike rate) is not much higher than the frequency with which an organism needs to be informed about its environment. The neural code resembles the principle of frequency modulation (FM) used in radio transmissions, but the orders of magnitude are very different. An FM radio station transmitting at, say, 100 MHz codes for speech and music which contains frequencies up to about 20 kHz. So, from the carrier wave's viewpoint, the frequency needs to change only very slowly. This is called a rate code, and many nerve cells do approximately the same—however, at very different time/frequency scales. In human hearing, for example, the spike rate from cochlear nerve fibres may be identical to the frequency of the sound wave, say 400 sp/s at 400 Hz. Comparing this with the duration of the shortest syllables (or musical notes), about 20 ms, shows that the spike frequency needs to change every few spikes. Since there is not enough time to determine *the* spike rate, it is better to call such a code a time code, or interval code, rather than a rate code. Each individual spike (or spike interval) may contain relevant information about the signal.

This has consequences for the way in which spike trains must be analysed to infer the relevant neural information from them.

Interval Series and Instantaneous Frequency

To obtain a more detailed record of the processes coded by spike signals, we will evaluate each individually occurring spike. In the early days of electrophysiology, the spike interval times

had to be measured by hand from chart recordings or photographs of spike trains. This was a tedious task, since getting interesting results implies the processing of thousands of spikes. Nowadays, we perform the same task with the aid of a computer. Mark the formulation of the previous sentence: man is still in control, or has to be. This means that, although we are glad that we do not need to process the bulk data, we still need to check the computer algorithms that recognize spikes from an input signal, and count the proper interval. To this end, it is a good idea to analyse a small sample by hand and compare it with the computer analysis of the same spike train.

As an illustration, a very short sample is shown in Fig. 4-42. The spike signal is shown in the bottom trace, the interspike interval series (t_i) and the instantaneous frequency the inverse of t_i, are shown in the middle and upper trace respectively. If this spike train is analysed in real-time, each spike interval time is known only at the occurrence of the next spike. Therefore, both measures of spike activity are undetermined during the first interval, i.e. in the time between spike #1 and spike #2 (hatched areas). In the next interval, the value of the first interval (or its inverse) is plotted, and so on. In summary, both interval series and instantaneous

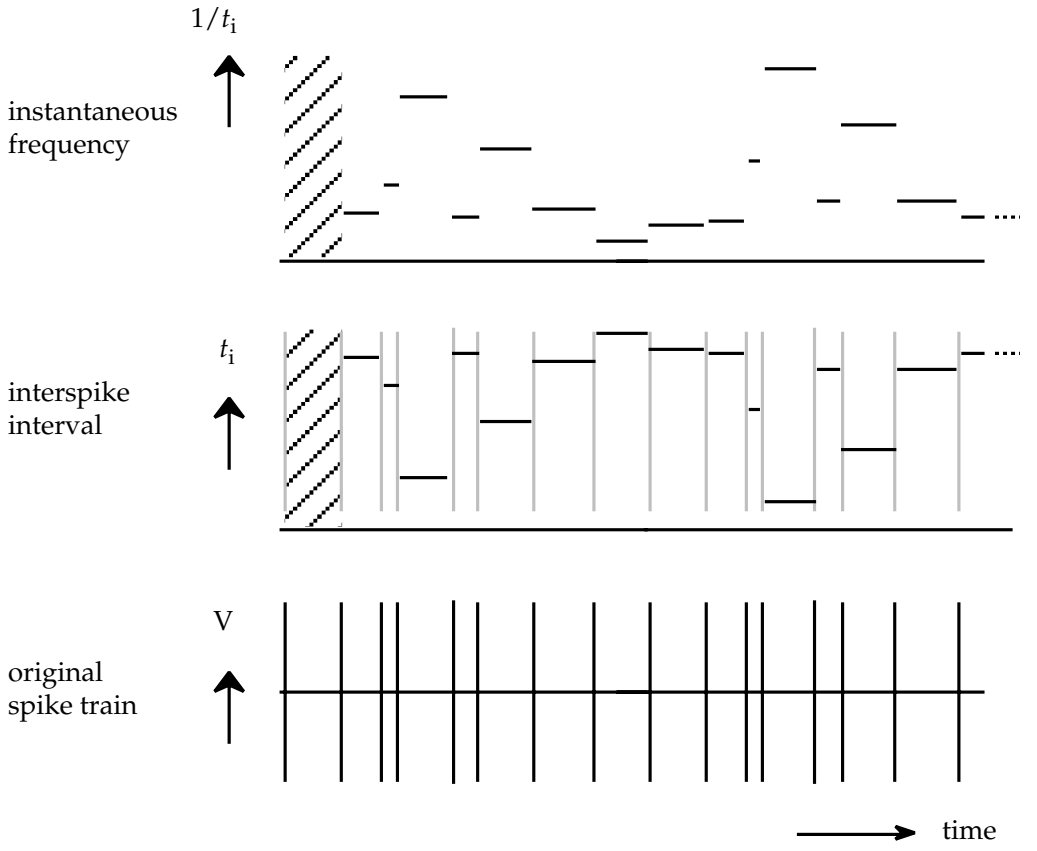


Fig. 4-42 From bottom to top: a small sample of a spike train, a plot of the interspike interval times, and its inverse, the instantaneous frequency.

rate are lagging the duration of one spike interval. Obviously, if the spike train is analysed later, i.e. after its complete acquisition, this delay does not arise. The figure shows that the interval series and the instantaneous frequency series are approximate mirror images of one another. Small intervals yield large rate values, long intervals yield small rate values. This is the general picture for fairly small changes in spike rate or interval, but it does not hold for large excursions of the spike rate. The inverse of a sine function is also a periodical signal, but a very distorted one.

This distortion can be illustrated with an analogue (continuous) example. Figure 4-43 shows a sine wave $\sin(t)+1.2$ and its inverse $1/(\sin(t)+1.2)$. Although, in principle, interval time and instantaneous frequency carry the same information, the choice of one over the other is not arbitrary. In the nervous system, the spike rate is often approximately proportional to a physiological quantity such as the strength of a signal from the environment (eye, ear, etc.) or the position of our own head, limbs, etc. (via proprioceptors). Therefore, spike rate is arguably a better measure than interval time. If there is hardly any activity, the intervals are getting very long. If the activity ceases altogether, the interval time tends to infinity. This is not in keeping with our intuition, in which a dead neuron is not infinitely active. If, in the example of Fig. 4-43, the amplitude of the sinusoid (upper trace) would be increased only slightly, the spike frequency would touch the zero line, and hence the inverse would grow to infinity. The bottom line is that although spike data are collected as a series of interspike interval times, the conversion of this series into an instantaneous frequency series yields a better record of the neurophysiological activity studied.

Dot Display

A simple method to display the activity of a neuron, called a "dot display" or "raster display", consists of plotting each spike as a dot on a picture tube or on chart paper, and using our own visual system as a pattern analyser. Figure 4-44 shows a dot display of the response of some neuron to a repeated stimulus. Each horizontal trace plots all spikes that occurred shortly before and some time after a stimulus (such as an electric pulse, a light flash, a sound click and so on). Each spike train shows relatively large random fluctuations, but by repeating the

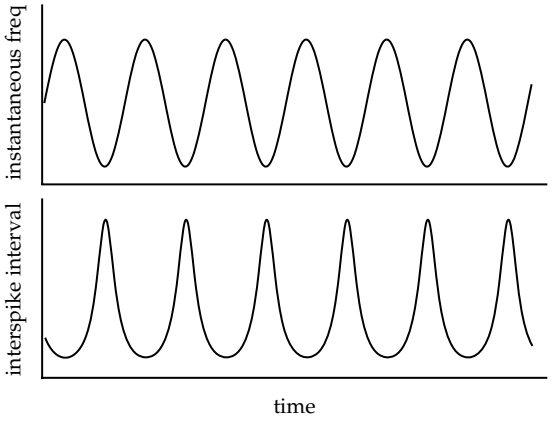


Fig. 4-43 A sinusoidal curve and its inverse.

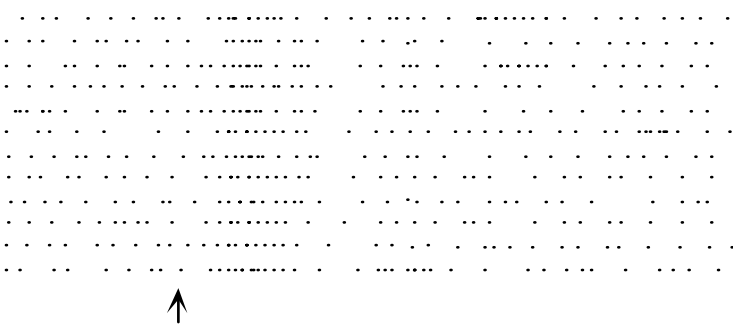


Fig. 4-44 A dot display of a hypothetical neuron that responds to a repeated stimulus, in addition to strong fluctuations in spontaneous activity. Arrow shows the moment of stimulation.

stimulus a large number of times and lining up all traces vertically, a pattern is seen to arise. In the example given, the fluctuations cause bursts of relatively closely spaced spikes as well as gaps to occur spontaneously in each trace. In addition, however, the higher activity directly after the stimulus, and a gap somewhat later, is seen to occur in all traces.

This type of display can give a researcher a good impression of the degree in which a neuron responds to a certain stimulus. Before personal computers were common, it was often the only way to analyse spike data. Nowadays, however, spike train analysis is mostly done with computers. An automated version of the dot display is the so-called post-stimulus time histogram, or PSTH for short.

Stimulus–Response Characteristics: The PSTH

Traditionally, the PSTH plots a frequency histogram (here "frequency" means the *statistical* frequency) of spike occurrences in a number of bins at regular times, starting from the beginning of the stimulus. However, in sensory physiology, the stimulus is often a continuous waveform, such as a sound wave or a modulated light intensity. In this case, the stimulus is always present, and the PSTH is dimensioned so that one period of the stimulus is rendered. Such a plot is called a peri-stimulus time histogram, having the advantage that the traditional abbreviation PSTH can be kept. The principle is illustrated in Fig. 4-45.

A neuron is stimulated with an appropriate stimulus signal. This may be either a short pulse (such as a light flash to a photoreceptor, a sound click to a hearing organ, an electrical pulse to an interneuron, etc.) or a continuous waveform (such as a sinusoidal or compound sound wave for a hearing organ, a periodical head rotation for a semicircular canal organ, etc.). In Fig. 4-45, a sinusoidal signal is used to get a PSTH. The spike train is recorded repeatedly, starting from a fixed point in the stimulus waveform (sweeps 1 through n in the figure). At relatively short time scales, only one or a few spikes per sweep will occur, yielding a crude histogram at first. After many sweeps, however, a detailed histogram reveals the average response of the organ to the stimulus. Note that in the PSTH, the T stands for the times of occurrence of the spikes with respect to a fixed trigger point in the stimulus signal. Interspike interval times play no role here.

In addition to illustrating the way in which physiologists can probe the response of a single nerve fibre, Fig. 4-45 may serve to illustrate the way in which the nervous system is able

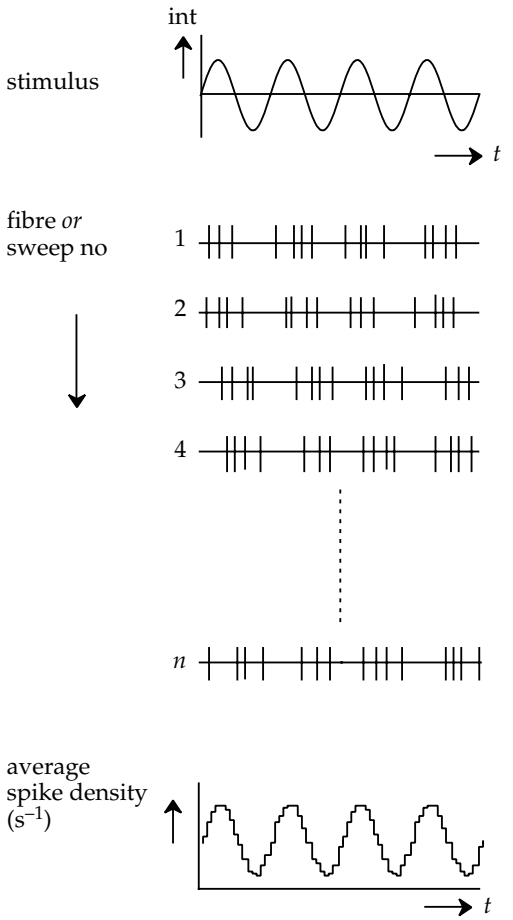


Fig. 4-45 Principle of PSTH determination.

to process data reliably, given the fluctuating spike trains shown. In this case, the traces 1 through *n* represent the simultaneous responses of a number of parallel fibres. Although electrophysiologists record usually from one fibre or a very limited number of fibres at any given time, the nervous system can rely on a massively parallel information stream. In most sense organs, such as the eye, the cochlea (hearing organ), the organs of equilibrium, and many skin sensors for touch, temperature and so on, many parallel fibres represent virtually the same quantity.

In the labyrinth, for example, the nerve from a semicircular canal organ consists of a few hundred fibres, all representing the same signal: a head rotation around one axis. It is plausible that nature performs an operation similar to the PSTH scheme by letting a number of fibres from the above-mentioned nerve converge onto a small number of brain cells. Combining the psp's from these parallel fibres averages the spike occurrence times in a way similar to the one described for PSTH determination, yielding a reliable measure of the original input (a head rotation in the example). The PSTH is a sensitive technique to find a small but nevertheless

important response to a periodical input signal. If the response is strong enough, the modulation of the spike rate influences the interspike interval histogram. This is the case shown in Fig. 4-38, trace 3. However, this is a very insensitive indicator of spike rate modulation. A far better method to detect the properties of a spike train, which is used frequently during experiments, is to make the spikes audible. Since each spike causes a distinct click of about 1 ms duration, any small audio amplifier and loudspeaker will do. Our ears plus brain form a very sensitive time series processor!

ANALYSIS OF NERVE MEMBRANE DATA

Terminology: The Hodgkin and Huxley Channel

In 1952, a paper "A quantitative description of membrane current and its application to conduction and excitation in nerve" was published by Hodgkin and Huxley, in which they presented a mathematical model for the functioning of Na and K channels in the nerve axon. Although ion channels had not been found yet at that time, the terminology that was introduced in this paper to describe membrane permeability is still used today to describe the kinetics of (voltage-sensitive) ion channels.

According to their model (the H&H model), a channel contains a pore that is kept closed at the resting membrane potential by a number of "activation gates". Upon depolarization, the activation gates open up and ions can traverse the pore until it is closed again by (a number of) inactivation gate(s). Figure 4-46 shows a schematic representation of the voltage-sensitive Na^+ channel according to the H&H model. During the brief moment that all gates are open simultaneously, an inward Na^+ current passes through the pore.

The gates are supposed to operate independently of one another and to open and close stochastically with Poisson-distributed life times. Hence the kinetics of each gate may be considered as a mono-molecular chemical reaction. The energy to move the gates is furnished by the transmembrane electric field. In the H&H theory, the opening of a single gate requires the translocation of a single elementary charge across the membrane.

Due to the stochastic nature of the movement of the gates, the macroscopic current measured from a large population of Na^+ channels is not square as shown for a single channel in the figure above, but a function of four exponentials (Fig. 4-47).

Upon return to the resting membrane potential, a tail current may be observed that decays rapidly due to the closing of the activation gates (deactivation).

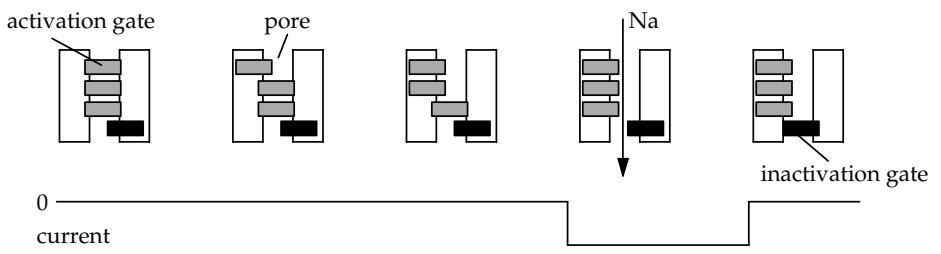


Fig. 4-46 Schematic showing the different states of a voltage-sensitive sodium channel.

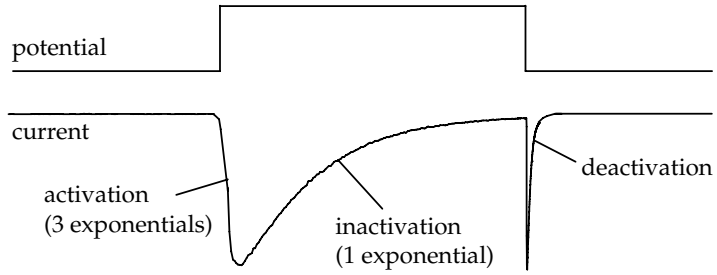


Fig. 4-47 The resultant, macroscopic, current through many ion channels.

Analysis of Macroscopic (Whole-Cell) Currents

The first and most important step in the analysis of patch-clamp experiments is taken before the actual experiment: its design. What do we want to measure and how do we eliminate currents and channels we are not interested in? What data do we have to feed to our analysis program such that it can deal with the question that we ask?

As an example, let us suppose we wish to determine the voltage dependence of activation of the low threshold (L-type) calcium channel.

To eliminate voltage-dependent sodium currents from many preparations, tetrodotoxin (TTX) can be applied extracellularly. Unfortunately, not all voltage-dependent Na^+ currents are sensitive to this toxin, but fortunately, they are not sensitive either to inhibitors of the L-type Ca^{2+} channel such as the dihydropyridines (e.g. nifedipine) or diphenylalkylamines (e.g. verapamil). Potassium currents may be eliminated by replacing extracellular K^+ by tetraethylammomium (TEA) ions and by replacing intracellular KCl by a mixture of CsCl and TEACl (K^+ channels are poorly conductive for CsCl, but since they are usually much more numberous than Ca^{2+} channels, a small cesium current may contaminate the records if 10 mM of TEACl is not added to the pipette solution). N-type Ca^{2+} channels are inhibited by omega-conotoxin, and P-type Ca^{2+} channels are inhibited by omega-agatoxin. There is no inhibitor for the T-type Ca^{2+} channel, but its voltage dependence and kinetics are different from the L-type channel.

The contribution of voltage-insensitive ("leak") currents may be subtracted from the records after measuring their amplitude at potentials at which the Ca^{2+} channel is inactive. Subsequent linear extrapolation should then give an estimation of their contribution at other potentials.

The contribution of currents that are still left may be eliminated by recording once in the absence of L-type Ca^{2+} blockers and recording a second time in the presence of L-type Ca^{2+} blockers. Difference in the records then yields the true L-type current. Often, extracellular calcium is raised from 2 to 10 mM in order to increase the size of calcium currents. As L-type Ca^{2+} channel inactivation is due to binding of Ca^{2+} ions to an intracellular domain of the channel protein, increasing calcium currents increases inactivation, but Ca^{2+} may be replaced by Ba^{2+}, which binds poorly to this site, thus reducing inactivation. However, the substitution of Ba^{2+} for Ca^{2+} also shifts the voltage dependence of calcium channel activation. Relatively large currents carried by Ca^{2+} channels may be obtained by eliminating all extracellular calcium and recording the currents in 140 mM NaCl. In that case, the L-type channel current is carried by Na^+ ions, but again, the voltage dependence of channel activation is changed.

The Current to Voltage (I/V) Curve

An I/V curve is obtained by presenting a succession of square voltage pulses of increasing amplitude to a voltage-clamped cell and recording the current responses. For each trace, the peak (maximum) current is measured and plotted as a function of the voltage of the square pulse. The I/V curve in Fig. 4-48 was obtained in conditions to measure calcium currents in the absence of extracellular Ca^{2+}, hence in extracellular 140 mM NaCl, 0 Ca and TTX 100 nM, intracellular 25 mM NaCl, 100 mM CsCl, 10 mM TEACl and 2.5 mM EGTA.

The I/V curve is the result of the activation of both T- and L-type channels. The same series of pulses given after inhibition of L-type channels by nifedipine gives the I/V curve for the T-type channel (Fig. 4-49, left). Then subtracting this curve from the first curve gives the I/V curve for the L-type channel (right).

The reversal potential, which is the potential where the current changes polarity, is 43.5 mV in this example and close to the equilibrium potential for sodium. As the whole-cell current is proportional to both the fraction of open channels and the driving force (the difference between reversal potential and actual potential), the fractional conductance C/C_{max} can be obtained by dividing the data points in the graphs above by the driving force and normalizing with respect to the maximum conductance. This gives, in Fig. 4-50, the so-called activation curves for T-type channels (open circles) and L-type channels (closed circles).

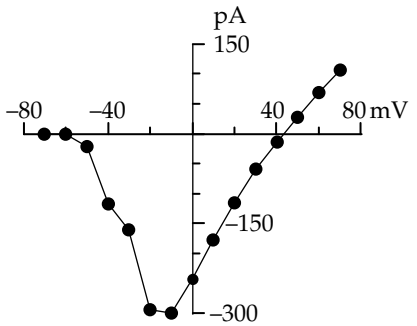


Fig. 4-48 *I/V* curve of a calcium current.

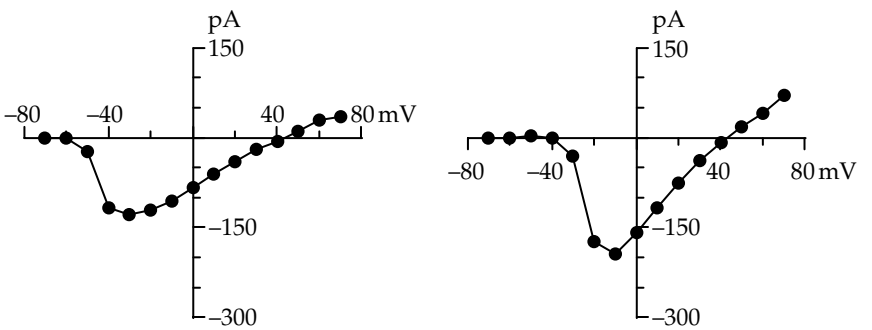


Fig. 4-49 Segregated *I/V* curves of the type T (left) and type L (right) channel.

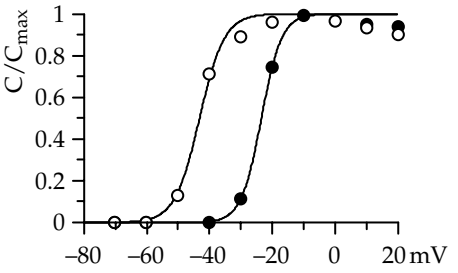


Fig. 4-50 Activation curves of T and L types of calcium channels.

Leak Subtraction by Extrapolation

May an example illustrate how this method works (Figs. 4-51 and 4-52).

Figure 4-51 left represents a family of K^+ currents recorded in response to a series of square voltage pulses of increasing amplitude (hence, a similar *I/V* protocol as in the preceding

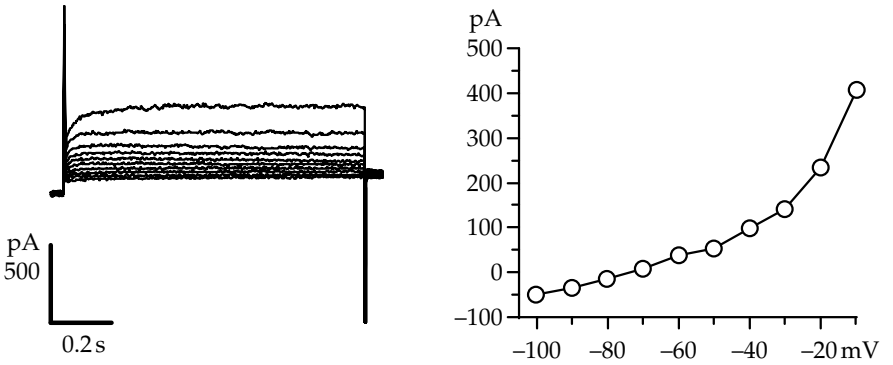


Fig. 4-51 Potassium currents and the *I/V* curve before leak subtraction.

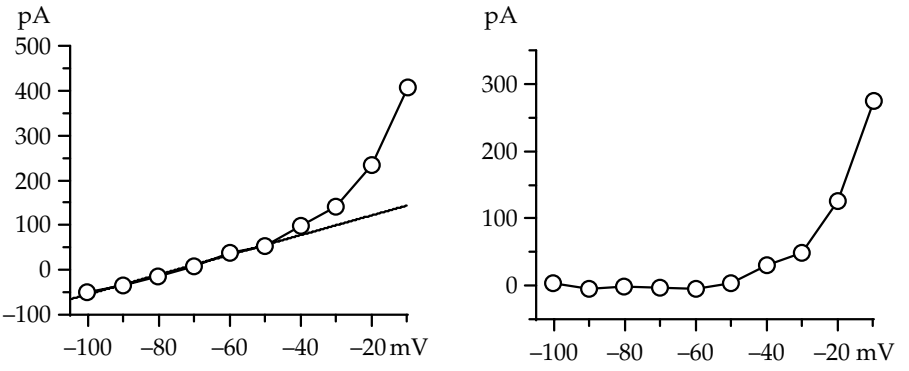


Fig. 4-52 Leak subtraction (left) and the corrected *I/V* curve (right).

paragraph). The figure to the right shows the resulting I/V curve. A priori knowledge of the K^+ channel in question (an outward rectifier) says that it should not open below $-60\,mV$. Hence the first 5 data points represent a linear leakage current. Linear regression of these 5 points gives a leak conductance of 2.2 nS (Fig. 4-52, left), and after the subtraction of the leak, the curve to the right is obtained.

Leak Subtraction by Prepulses: The P/N Method

This method is used to eliminate linear leak and (most of) the capacitive currents.

In the preceding paragraphs, the I/V protocol consisted of a series of square depolarizing voltage pulses (Fig. 4-53).

In the P/N (one positive over N negative pulses) protocol, each depolarizing pulse of amplitude v is preceded by a number N of hyperpolarizing pulses of size $-v/N$. The responses to the $N+1$ voltage jumps are summed with the result that all linear responses cancel out and the non-linear response of the voltage-sensitive current of interest remains. This is shown in Fig. 4-54 for $N = 2$.

Of course, the resting potential must be so chosen that it lies well outside the range of the voltage sensitivity of the current of interest.

Capacitive currents of opposite polarity are elicited by the hyperpolarizing and the depolarizing pulses and tend to cancel. However, they are not quite so because the relaxation time constant of the capacitive current is a function of the membrane resistance, which is high during the hyperpolarizing pulses, but gradually decreases during the depolarizing pulse due to the opening of voltage-sensitive channels. Similarly, upon return to the resting potential, a tail current may change the relaxation time constant of the capacitive current. Hence, the P/N protocol cannot altogether compensate for a poorly set clamp amplifier.

Noise Analysis: Estimating the Single-Channel Conductance from Whole-Cell or Large Patch Recordings

Part of the noise recorded in the whole-cell configuration is due to the random opening and closing of ion channels. This noise contains information from which the single-channel conductance and the total number of channels can be obtained. Consider a cell that contains predominantly a population of non-inactivating voltage-dependent channels (e.g. non-inactivating outwardly rectifying K^+ channels). If the probability p of channel opening is low (e.g. at hyperpolarized

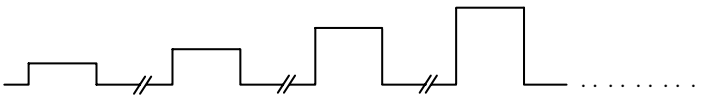


Fig. 4-53 Series of pulses used in I/V curve measurement.

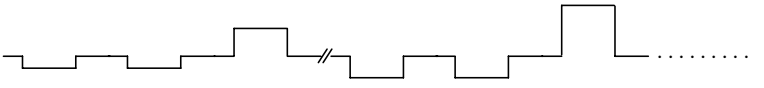


Fig. 4-54 Pulse series used in the so-called P/N protocol, for $N = 2$.

potentials), then the mean current generated by the population of channels is low, and the random fluctuations around the mean current are small. At the other extreme, if all channels are open all the time ($p = 1$), the mean current is maximal and the random fluctuations around the mean are again small. Half-way between these conditions, i.e. at $p = 0.5$, the mean current is half maximal and the random fluctuations are the largest, since now channels open and close all the time. Hence plotting the variance $s(p)$ of the macroscopic current as a function of the mean current $m(p)$ gives a hump-like graph (Fig. 4-55).

According to Neher and Stevens (1977), the relation between variance $s(p)$ and mean $m(p)$ is:

$$s(p) = i \cdot m(p) - \frac{(m(p))^2}{N}$$

with p the probability of channel opening, i the unitary current and N the total number of channels. Differentiation of s with respect to m gives:

$$\mathrm{d}s/\mathrm{d}m = i - \frac{2m}{N}$$

Hence at zero mean current ($p = 0$), $\mathrm{d}s/\mathrm{d}m = i$. Fitting the histogram with a parabola:

$$s(x) = a + b(x - x_0)^2$$

gives two equations: at $x = 0$, $\mathrm{d}s/\mathrm{d}x = -2 \cdot b \cdot x_0 = i$; and at the intersection of the parabola with the $y = s(x) = 0$ axis, $x = N \cdot i$. Therefore, fitting the histogram with a parabola yields both the unitary current and the total number of channels. Often, the variance at zero mean current is not zero, due to amplifier and thermal membrane noise. In such a case, the variance at zero mean current gives an estimate of the background noise that should be subtracted (vectorially) from the data points before fitting the curve with a parabola. Hence, if the background variance is b, then the actual channel noise variance a for each data point is obtained as follows:

$$a = (\sqrt{s} - \sqrt{b})^2$$

where s is the observed variance.

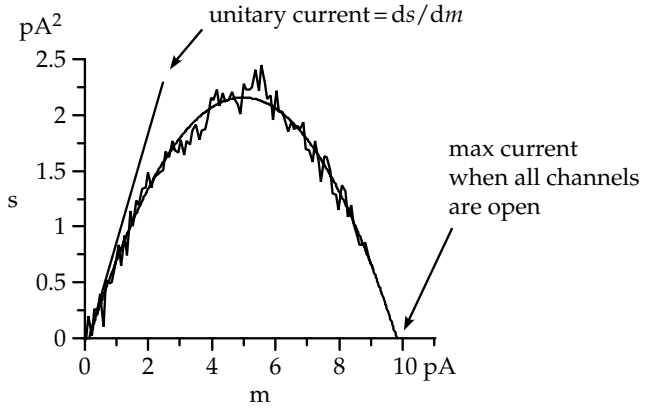


Fig. 4-55 Variance of the macroscopic current as a function of the fraction of open channels.

Above, stationary channel activity was assumed, but the method can also be applied to transiently active channels, like the voltage-sensitive Na^+ channel. The only thing that is important is to obtain variance and mean by varying the open channel probability. During a voltage jump from a hyperpolarized to a depolarized potential, the Na^+ currents start with all channels closed ($m = 0, s = 0$). During the activation phase of the current, the probability of channel opening increases ($m > 0, s > 0$) and then decreases again when inactivation sets in. If one presents many identical voltage jumps to the cell, the macroscopic current of each of the responses will be very similar except for the random fluctuations. By subtracting the mean response from each trace, these fluctuations remain and they can be used to estimate the variance at each phase of the mean current. Plotting these estimates as a function of the mean current gives a figure similar to the one above.

NOISE ANALYSIS: ESTIMATING CHANNEL KINETICS

Before the advent of single-channel voltage-clamp, "channel kinetics" could be estimated either by fitting transient macroscopic currents with (multiple) exponential functions or by spectral analysis of steady-state currents. Whereas the first method works well with voltage-sensitive currents, the second is often required in those cases where transient currents cannot be (easily) obtained, such as for constitutively active or ligand-gated channels. Consider a hypothetical cell containing constitutively active channels that flip between open and closed states with the following kinetics:

$$\mathrm{C} \underset{k_2 = 3}{\overset{k_1 = 1}{\rightleftharpoons}} \mathrm{O}$$

If it were possible to force all channels in the closed state (C) and then have them relax to equilibrium, a macroscopic current with a relaxation rate constant of $k_1 + k_2 = 4\,\mathrm{s}^{-1}$ would have been obtained. Since this is not possible, we have to use the fact that the cell contains only a limited number of channels that produce measurable current noise. The general idea behind the theory that was developed by DeFelice and others in the 1970s (DeFelice, 1977) is that the noise fluctuations represent statistical deviations from the mean that subsequently relax to equilibrium with, in this case, a rate constant of $k_1 + k_2$. As we have seen in Eq. 4-6, exponential decay in the time domain translates into a Lorenzian function in the frequency domain. Hence, by taking the spectrum of the channel noise and fitting the spectrum with a Lorenzian function, the relaxation rate constant may be found. In Fig. 4-56, 20 s sweeps of the activity of 256 channels were simulated, the mean spectrum was obtained and fitted with a single Lorenzian function, yielding an estimated rate constant of $3.67\,\mathrm{s}^{-1}$.

Analysis of Microscopic (Unitary) Currents

The analysis of the "unitary current" (or single-channel current) usually involves four steps:

1. the estimation of the unitary current,
2. the detection of opening and closing events,
3. determination of the number of channels in the patch and
4. measurement of dwell times.

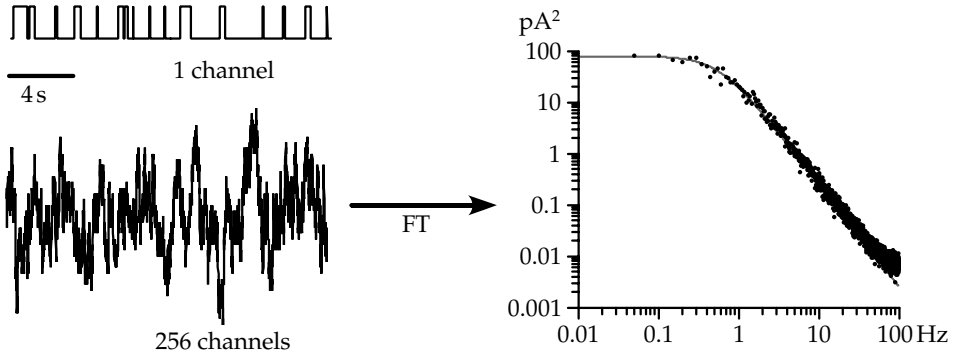


Fig. 4-56 Simulated macroscopic membrane current (lower left) and its power spectrum (right).

1. *Estimation of the unitary current.* The unitary current may be obtained from the current–density histogram. During acquisition, the output voltage of the patch-clamp amplifier has been sampled by an ADC with a certain precision, say 12 bit, such that the span of values at the output of the converter ranges from −2047 to +2048. Suppose the converter has been calibrated such that 2048 corresponds to 20 pA. The current–density histogram reflects the frequency with which each of the 4096 values between −2047 and 2048 (−20 and +20 pA) occurs in the record. In the absence of channel openings, the values will cluster around 0 pA. Due to random charge movements in both amplifier and membrane patch, the peak at 0 pA has a bell shape obeying a Gaussian distribution. If the patch contains a single channel that opens and closes regularly, then the current–density function will show an additional peak. The two peaks are separated by a distance (in pA) that corresponds to the unitary current. If more channels of the same type are present, then a series of equidistant peaks will appear in the current–density histogram (Fig. 4-57).

The distance of the peaks can, of course, be measured by hand, but it is better to fit a number of equidistant Gaussians to the histogram, the difference of the means giving the unitary current (7.0 pA in this example). Fitting gives at the same time the system offset (note that the peak at 0 pA is not quite centred around zero). This offset needs to be subtracted from the records before detection of opening and closing events may be carried out. The example given above concerns a relatively large voltage-insensitive K^+ channel without overlap between the bell-shaped curves and without baseline fluctuations. This condition is not always met and, especially if channel conductance is low (<1 pS), baseline drift and fluctuations need to be removed by subtraction of a polynomial or a cubic spline.

A special case of baseline fluctuation is the one caused by changes in the clamp voltage e.g. when stimulating the opening of voltage-dependent channels by a jump in potential. The voltage jump generates a capacitive current, which, if not carefully compensated with the amplifier settings, tends to contaminate the current record, as shown in Fig. 4-58.

The first two records show channel openings during a depolarizing pulse, while the third stimulation did not elicit openings. The fact that some records do not contain openings can be used to eliminate the stimulus artefacts. By taking the average of the traces devoid of channel openings and subtracting the average from each individual trace, most of the capacitive current is removed (Fig. 4-59).

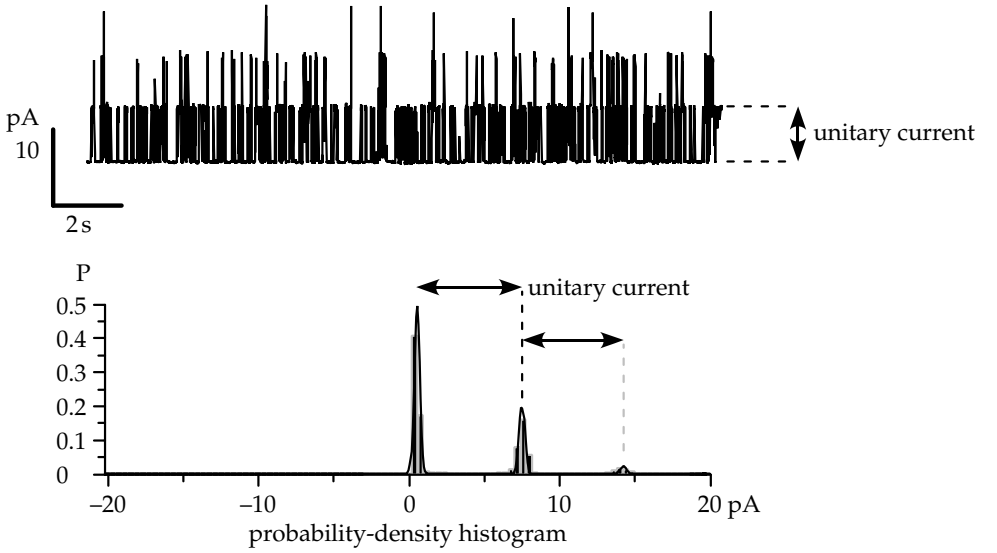


Fig. 4-57 A channel current and its probability density histogram.

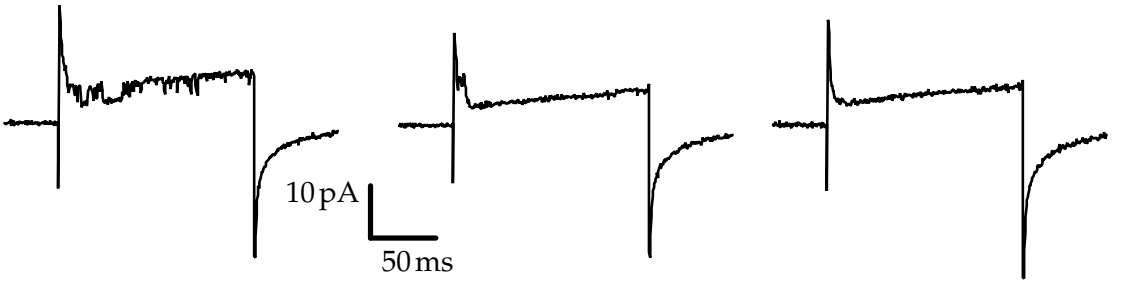


Fig. 4-58 Capacitive current peaks that hamper channel current measurement.

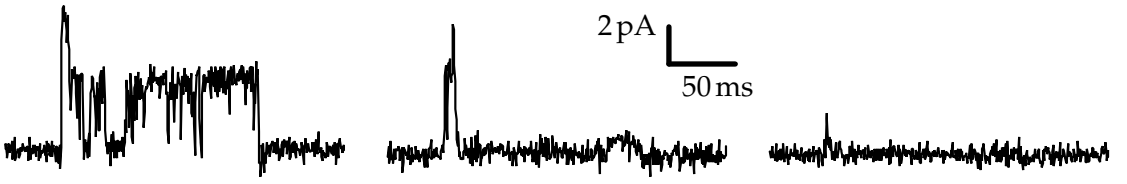


Fig. 4-59 Removal of capacitive current by subtraction.

2. *Detection of opening and closing events*. Once all artefacts are removed, baseline fluctuations subtracted and the value of the unitary current determined, the detection of opening and closing events can take place.

The simplest way to detect channel openings is to set thresholds at $(n + 0.5) \times i$, where i is the unitary current and n an integer greater than or equal to 0 (Fig. 4-60). Then, values above

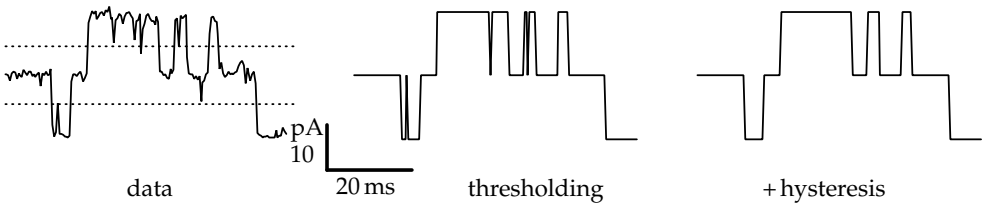


Fig. 4-60 Detection of channel openings by thresholding, using hysteresis to eliminate brief threshold crossings.

threshold n correspond to n openings (the dotted lines in the figure below indicate thresholds for $n = 1$ and $n = 2$, the middle trace shows the result of thresholding).

As can be seen from the first two traces, short spurious excursions from the current state, barely crossing the threshold, may lead to detection of events. Although it is not shown here, this phenomenon is a nuisance when background noise is more important or if the record is less filtered. In those cases, each channel transition from one state to the next may be accompanied by a few rapid crossings of the threshold before the channel settles at the next state. This situation can be remedied by adding some amount of hysteresis to the thresholds (rightmost trace). A threshold with hysteresis is a threshold split into two sublevels with the signal having to cross both levels in order to change state. The larger the distance between the sublevels, the larger the hysteresis. Thus, having "idealized" the patch-clamp records, the times between channel transitions can be measured to create dwell-time histograms.

3. *Estimation of the number of channels in a patch*. When multiple channels are present in a membrane patch, the number of channels needs to be known in order to carry out correctly the kinetic analysis using dwell-time histograms. A conservative estimation is given by counting the maximum number of levels that occur in the patch-clamp records. However, especially if open channel probability is low, the rare occasions when all channels were open simultaneously might have been missed. The binomial analysis discussed below gives a second estimate. Suppose that the patch contains N channels. All channels are identical and have a probability p of being open. The probability of being closed is $1 - p = \mathrm{q}$ (say). For $N = 2$, the probability that both the channels are open is p^2, the probability that one is open and one closed equals $2pq$ and both closed is q^2.

For $N = 3$, this gives $p^3, 3p^2q, 3pq^2$, and q^3 for 3 open, 2 open/1 closed, 1 open/2 closed and 3 closed, respectively. The coefficients of the terms can be found by expanding "Pascal's triangle":

$$
\begin{array}{ccccccccc}
 & & & & 1 & & & & \\
 & & & 1 & & 1 & & & \\
 & & 1 & & 2 & & 1 & & \\
 & 1 & & 3 & & 3 & & 1 & \\
1 & & 4 & & 6 & & 4 & & 1
\end{array}
$$

The numbers in each new row are the sum of the two numbers just above, adding 1s at the borders. The coefficients may also be calculated using the formula:

$$c_k = \frac{N!}{k!(N-k)!}$$

In the following stretch of data, at least three channels are present as judged by the number levels in the trace at the left and the number of peaks in the current–density histogram at the right (Fig. 4-61).

Now, two unknowns need to be solved: N, the total number of channels and p, the probability of opening. We first calculate the product of N and p:

$$Np = (S_1 + 2S_2 + 3S_3)/S_t$$

Here S_1 is the surface of the peak in the current–density histogram corresponding to 1 open channel, S_2 is the surface corresponding to 2 open channels, S_3 is the surface for 3 open channels and S_t is the total surface of the histogram. In the case of this example, the current–density histogram could be fitted by 4 equidistant Gaussians with identical variance, meaning that all peaks are isomorphic. Therefore, it is not necessary to calculate the surface of each peak. It suffices to take the square of the amplitude of each peak as a measure of its surface and hence:

$$Np = (A_1^2 + 2A_2^2 + 3A_3^2)/(A_0^2 + A_1^2 + A_2^2 + A_3^2)$$

In the current example, $A_0 = 0.080072$, $A_1 = 0.081944$, $A_2 = 0.050924$ and $A_3 = 0.021246$ giving $Np = 0.8197086$. If N were 3 then p would be $0.8197086/3 = 0.2732362$.

With the hypothetical N and p, the expected surfaces ($E_0 \ldots E_3$) of the current–density histogram can be calculated using the approach discussed above. The sum of χ^2 deviations between expected and observed surfaces ($O_0 \ldots O_3$) gives a measure of error (dE):

$$dE = S_i(E_i - O_i)^2/O_i \quad \text{which for } N = 3 \text{ equals } 3.943723e-3$$

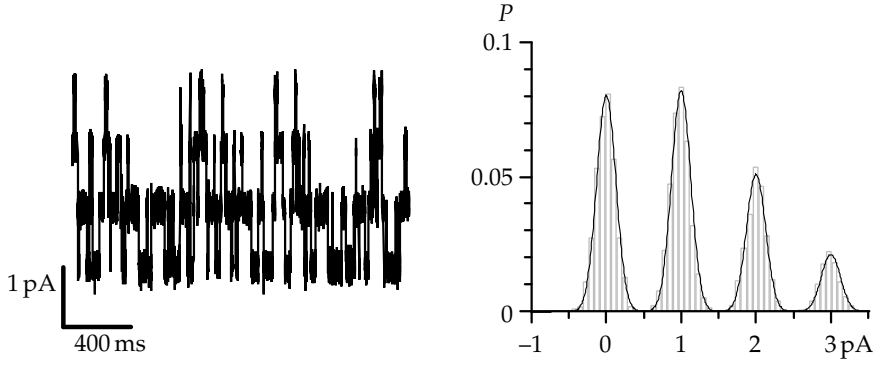


Fig. 4-61 Example current with at least three channels.

The same calculation can be carried out supposing 4 and 5 channels with p 0.2049271 and 0.1639417, respectively. This results in:

$N = 3 \quad dE = 3.943723e-3$
$N = 4 \quad dE = 6.825411e-5$
$N = 5 \quad dE = 1.234697e-3$

Hence it is most likely that the patch contains 4 channels.

4. *Measurement of dwell times*. As its name, at least partially, suggests, a dwell time histogram describes how often and how long a signal is in a certain state n. Figure 4-62 shows an example for a patch containing a single K^+ channel ($n = 0$ (channel closed) and $n = 1$ (channel open)).

From both histograms it is clear that short dwell times are more abundant than longer ones. In general, channel kinetics are considered to behave like radioactive decay or mono-molecular chemical reactions. This implies that transition probabilities are independent of time and that decay follows an exponential time course. The above histograms are fairly well fitted by single exponential distributions which give the rate constants (in s^{-1}) for a simple mono-molecular model of the behaviour of the channel (C = closed state and O = open state):

$$C \underset{k_2 = 399.5}{\overset{k_1 = 64.5}{\rightleftharpoons}} O$$

If the patch contains more than one channel, the analysis is somewhat more complicated. In the following example, the patch contained 3 channels, resulting in 4 dwell-time histograms. Each of the histograms is fitted with a single exponential (Fig. 4-63).

Suppose that the 3 channels are identical and each can be either in the closed state (C) or in the open state (O). If all channels are closed, then the apparent rate constant leading from $n = 0$ to $n = 1$ is $3k_1$. Similarly, if all channels are open, then the apparent rate constant from $n = 3$ to $n = 2$ is $3k_2$. If only one channel is open, then the apparent rate constant is $2k_1 + k_2$, since two channels are available for opening and one for closing. For two open channels it is $k_1 + 2k_2$. In general, with N being the total number of channels in the patch and n the current state:

$$k(n) = (N - n)k_1 + nk_2(s^{-1})$$

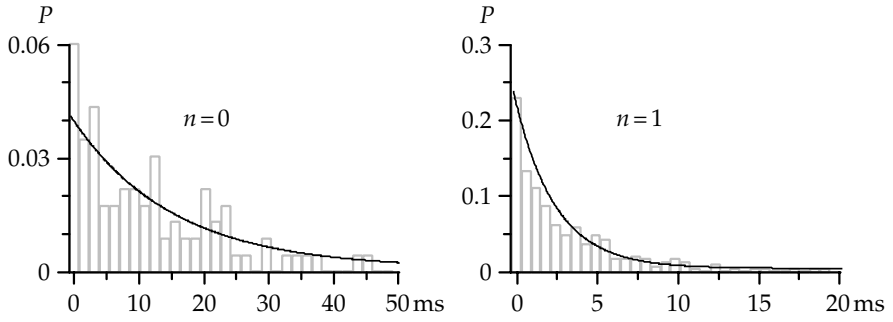


Fig. 4-62 Dwell-time histograms for a single potassium channel. Left: closed, right: open state.

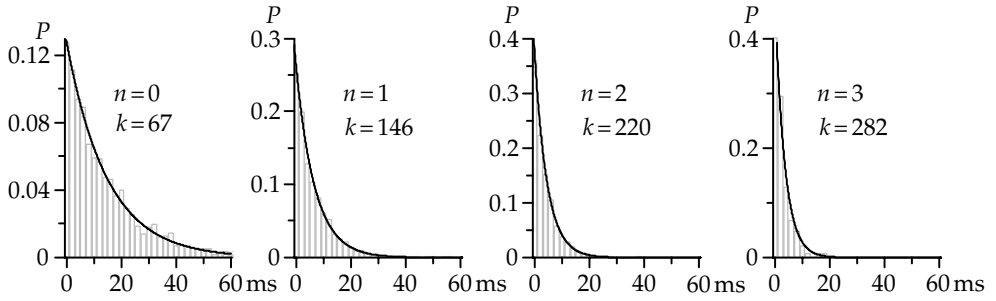
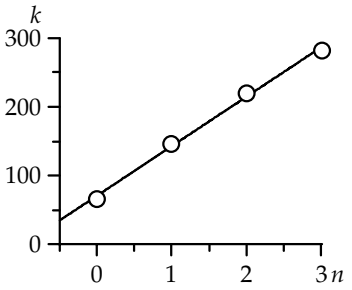


Fig. 4-63 Dwell-time histograms for a patch with three channels.

With the data of the figure above, we get 4 equations:

$$\begin{aligned} 3k_1 &= 67 \\ 2k_1 + k_2 &= 146 \\ k_1 + 2k_2 &= 220 \\ 3k_2 &= 282 \end{aligned}$$

It is of course easy to use the first and the fourth equations to find the rate constants k_1 and k_2, but then not all available data would have been used. Moreover, often the probability of having all channels open simultaneously is low with the consequence that the associated dwell-time histogram (the fourth equation) is poor and has to be ignored. The relation between state n and apparent rate constant k is a linear one and therefore we can find estimates for k ($n = 0$) and k ($n = 3$) using linear regression:



This yields a coefficient of 71.9 and an intercept of 70.9 from which we can get $k_1 = 23.6$ and $k_2 = 95.5$.

Not all dwell-time histograms can be fitted by a single exponential. In Fig. 4-64, activity of a single channel was recorded. The closed time histogram ($n = 0$) is poorly fitted by a single exponential. The plot of the residue-of-fit underneath the dwell time histogram shows a triphasic time course that disappears if the histogram is fitted with two exponentials ($n = 0$, middle graphs). A non-random distribution of the data points around the fit as in the bottom left figure is an indication that the number of degrees of freedom of the fitting function is too low. The open time histogram ($n = 1$, right-most graph) is well fitted with a single exponential.

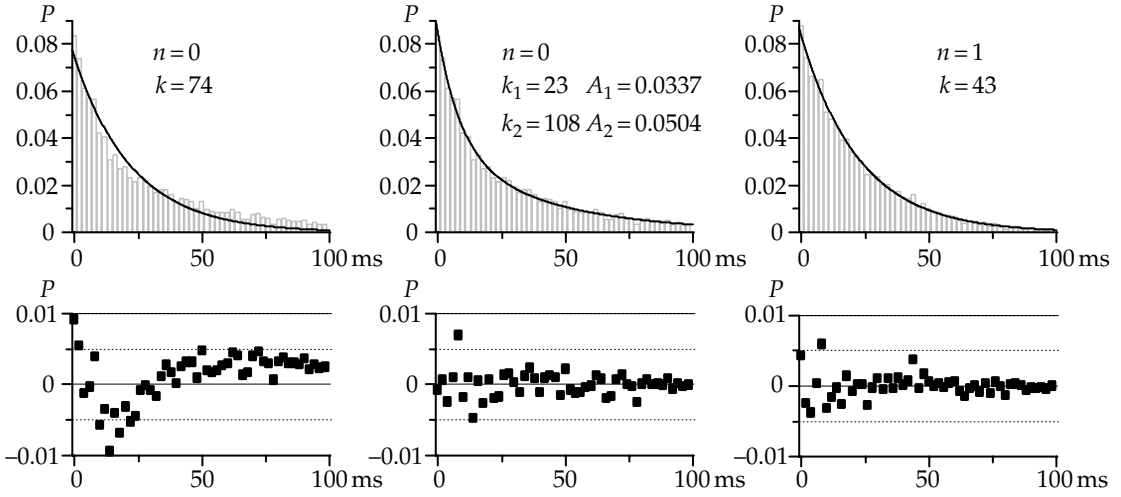


Fig. 4-64 Dwell-time histograms suggesting two different pathways to the open state.

The fact that two exponentials are required to fit the closed time distribution indicates that two different pathways lead to the open state. The dwell time histogram is actually the sum of two distributions: one describing the transition of C_1 to O and the other describing the transition of C_2 to O. This observation allows us to propose the following model to explain the data:

$$C_1 \underset{k_3}{\overset{k_1 = 23}{\rightleftharpoons}} O \underset{k_2 = 108}{\overset{k_4}{\rightleftharpoons}} C_2$$

Of k_3 and k_4 we know that $k_3 + k_4 = 43$. In equilibrium net fluxes are zero:

$$k_1 \cdot C_1 = k_3 \cdot O$$
$$k_2 \cdot C_2 = k_4 \cdot O$$

hence, $\frac{k_3}{k_4} = \frac{k_1 \cdot C_1}{k_2 \cdot C_2}$ where $\frac{k_1 \cdot C_1}{k_2 \cdot C_2}$ corresponds to $\frac{A_1}{A_2}$, the ratio of the amplitudes of the two exponentials in the closed time distribution. With these two equations, two unknowns can be solved, giving $k_3 = 17$ and $k_4 = 26$.

Unfortunately, our model ($C_1 \leftrightarrow O \leftrightarrow C_2$) is not the only one that explains a bi-exponential closed time distribution. Consider the following chain:

$$C_1 \leftrightarrow C_2 \leftrightarrow O$$

Here the open time distribution is mono-exponential, while the open state may be reached directly from C_2 or indirectly from C_1. Without further information it is impossible to decide in favour of either of the two models.

If the channel is a ligand- or voltage-gated we might force all the channels in state C_1 by removing the ligand or by hyperpolarization respectively. Upon return to ligand/polarization we measure the time lapse between the onset of the stimulus and the first channel opening.

Doing so for many (1000 or more) stimulus presentations results in the so-called first-latency distribution. If the first model is correct, the first-latency distribution will be mono-exponential with rate constant k_1, if the second model is correct, the first-latency distribution will be bi-exponential. In that case, the chain $C_1 \leftrightarrow C_2 \leftrightarrow O$ causes a delay in the channel opening and therefore the maximum probability of channel opening does not occur at $t = 0$, but at a later time point as in Fig. 4-65.

The smooth line depicts the bi-exponential fit.

Calculating Dwell Time Histograms from Markov Chains

Above we were able to deduce some of the channel properties from inspection of the dwell time histograms and exponential fits made thereof. The number of exponentials necessary to fit the dwell time distributions and the shape of these distributions gave some clues about possible models describing the channel behaviour. In the current section we will do the inverse: we postulate a model and then calculate the dwell time histograms and the macroscopic current.

As we have seen above, a model of channel activity (e.g. $C_1 \leftrightarrow O \leftrightarrow C_2$) may consist of a number of states (e.g. C_1, O, C_2) interconnected by arrows depicting exponential decay from one state to another. In the section discussing spike intervals and Poisson processes and in Appendix F it is shown that the rate constants leading from one state to another are most conveniently arranged in a square transition matrix and that such a matrix represents a set of linear differential equations. We will take up the arguments of that section, using them to calculate dwell time distributions.

The First Latency Distribution

The first latency describes the time it takes for a channel to open for the first time. Suppose our model is

$$C_1 \underset{\mu 1}{\rightarrow} C_2 \underset{\mu 2}{\rightarrow} \cdots \underset{\mu(n-1)}{\rightarrow} C_n \qquad \text{(Eq. 4-11)}$$

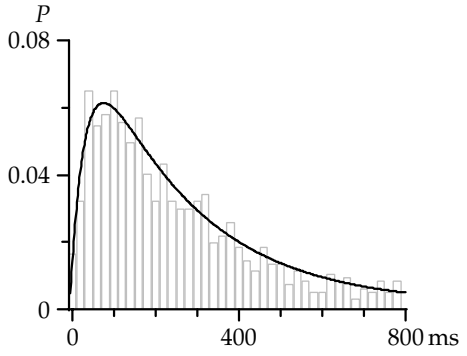


Fig. 4-65 First-latency distribution with bi-exponential fit.

with all channels in state C_1 at $t = 0$ and transition rate constants μ_i. Apart from the symbols (C for Closed and O for Open rather than S for State), this model is exactly the same as the one that we developed to calculate the waiting time distribution between successive spikes in a Poisson-distributed spike train (4-10). Therefore, the math is the same and the result (for three closed states leading to the open state) is the same (see F-7):

$$Fl(t) = \mu_1\mu_2\mu_3\left[\frac{\exp(-\mu_1 t)}{(\mu_2-\mu_1)(\mu_3-\mu_1)} - \frac{\exp(-\mu_2 t)}{(\mu_2-\mu_1)(\mu_3-\mu_2)} + \frac{\exp(-\mu_3 t)}{(\mu_3-\mu_2)(\mu_3-\mu_1)}\right]$$

Of course, this result does not give the general solution to the first latency problem. For one, the channel is not necessarily in C_1 at $t = 0$. Sometimes it might be in C_2 or in C_3 (let us restrict ourselves to $N = 3$ for the time being). In order to get the first latency distribution for arbitrary occupation of states at $t = 0$, the equivalents of the equation above have to be calculated for the cases that the channel is always in state C_2 or always in C_3 at $t = 0$. This will give us three first latency distributions Fl_1, Fl_2 and Fl_3 for the channel being in C_1, C_2 or C_3 at $t = 0$ respectively. Further suppose that the probability to be in state C_1 at $t = 0$ is p_1, to be in state C_2 is p_2 and to be in state C_3 at $t = 0$ is p_3. The weighted sum of these distributions gives the more general result:

$$Fl(t) = p_1 \cdot Fl_1(t) + p_2 \cdot Fl_2(t) + p_3 \cdot Fl_3(t)$$

The quantitatively identical result would have been obtained if starting from:

$$c_1 = \begin{pmatrix}(\mu_3-\mu_1)(\mu_2-\mu_1)\\ \mu_1(\mu_3-\mu_1)\\ \mu_1\mu_2\end{pmatrix} \qquad c_2 = \begin{pmatrix}0\\ \mu_3-\mu_2\\ \mu_2\end{pmatrix} \qquad c_3 = \begin{pmatrix}0\\0\\1\end{pmatrix}$$

and $p = z_1c_1e^{-\mu_1 t} + z_2c_2e^{-\mu_2 t} + z_3c_3e^{-\mu_3 t}$

We would have chosen the vector p_{t0} as follows:

$$p_{t0} = \begin{pmatrix}p_1\\p_2\\p_3\end{pmatrix} = z_1c_1 + z_2c_2 + z_3c_3$$

and would have proceeded similarly from thereon.

It may seem that models containing more than one open state pose a second complication. However, the presence of multiple open states does not change anything fundamental to the math involved, as the only thing we are interested in is the total time spent in the closed states and we do not care to what open state the system exits. For example, the models having transition rate constants μ_1 and μ_2:

$$C_1 \underset{\mu_1}{\rightarrow} O_1 \leftrightarrow O_2 \underset{\mu_2}{\leftarrow} C_2 \quad \text{and} \quad C_1 \underset{\mu_1}{\rightarrow} O \underset{\mu_2}{\leftarrow} C_2$$

Evidently, have identical first latency distributions, and the models:

$$O_1 \underset{\mu_1}{\leftarrow} C \underset{\mu_2}{\rightarrow} O_2 \quad \text{and} \quad C \underset{\mu_1+\mu_2}{\rightarrow} O$$

also have identical first latency distributions.

Now that we know what the ingredients are, we can formulate a general approach to calculate any first latency distribution. First, create the matrix A representing a set of differential equations (as in F-2, $X' = AX$) from the Markov transition matrix.

$$A = \begin{bmatrix} -\Sigma c_1 & k_{21} & k_{31} & k_{41} & k_{51} \\ k_{12} & -\Sigma c_2 & k_{32} & k_{42} & k_{52} \\ k_{13} & k_{23} & -\Sigma c_3 & k_{43} & k_{53} \\ k_{14} & k_{24} & k_{34} & -\Sigma c_4 & k_{54} \\ k_{15} & k_{25} & k_{35} & k_{45} & -\Sigma c_5 \end{bmatrix}$$

where k_{ij} represent the rate constants from state i to state j, and Σc_i the sum of the rate constants in the ith column, hence the sum of all rate constants leading away from state i. Because we are only interested in the time spent in the closed states, part of the matrix is superfluous. It is therefore a good idea to reorganise the matrix such that transitions between closed states appear in the upper left corner. This is most easily done by interchanging rows and columns associated with the open and closed states. Suppose that in the above matrix columns 1, 2 and 5 correspond to closed states and columns 3 and 4 to open states. It suffices to switch columns 3 and 5 and rows 3 and 5 to obtain:

$$A = \begin{bmatrix} -\Sigma c_1 & k_{21} & k_{51} & k_{41} & k_{31} \\ k_{12} & -\Sigma c_2 & k_{52} & k_{42} & k_{32} \\ k_{15} & k_{25} & -\Sigma c_5 & k_{45} & k_{35} \\ k_{14} & k_{24} & k_{54} & -\Sigma c_4 & k_{34} \\ k_{13} & k_{23} & k_{53} & k_{43} & -\Sigma c_3 \end{bmatrix}$$

Now the upper 3×3 block of A represents transitions between closed states, giving the reduced transition matrix R:

$$R = \begin{bmatrix} -\Sigma c_1 & k_{21} & k_{51} \\ k_{12} & -\Sigma c_2 & k_{52} \\ k_{15} & k_{25} & -\Sigma c_5 \end{bmatrix} \qquad \text{(Eq. 4-12)}$$

Then, by the methods we discussed in the section about Markov chains (Appendix F), we have a computer program to calculate the eigenvalues and then solve for each eigenvalue λ:

$$|R - \lambda I| = 0 \quad \text{to obtain the eigenvectors.}$$

The only thing that still misses in order to be able to calculate the first latency distribution is the probability of occupancy of the closed states at $t = 0$:

$$p_{t0} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}$$

If the model is used to compare it with experimental results, then it is a good idea to design the experiment such that the probability to be in a specific closed state, initially, equals 1 (e.g. by removing all ligand or by membrane hyperpolarization).

The case in which the system is in equilibrium will be discussed in the next paragraph.

The Closed Time Distribution

The closed time distribution describes the time it takes for a channel to open given that it is closed. The problem of obtaining the closed time distribution much resembles the problem of obtaining the first latency distribution, except that we are not only interested in the first opening but also in subsequent openings. This suggests that we first take the first latency distribution, wait until the next closing, again take the 'first' latency distribution and sum the two distributions, wait until the third closing etc, until we reach the end of a period of length t for which we wish to calculate the closed time distribution. As we do not really follow individual closing events, our approach is probabilistic.

At every instant, the probability that a channel goes from an open state to a particular closed state is proportional to the probability to be in one of the open states multiplied by the rate constants leading from those open states to the particular closed state. So if we want to know the average probability to enter a particular closed state during a stretch of t seconds following a situation that may be non-equilibrium, we have to calculate the integral of the evolution of each open state during t seconds and multiply these integrals by the rate constants leading to the particular closed state. This will give us (again for the model with three closed states):

$$p = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}$$

where the vector p describes the average probability to enter each of the closed states during a period of t seconds after a (non-)equilibrium condition. Calculating the first latency distribution with the vector p as initial condition gives the closed time distribution.

If the system is in equilibrium, the first latency distribution and the closed time distribution are identical and the calculations to obtain the vector p are more simple. In order to get the vector p, we have to solve three equations in three unknowns (p_1, p_2 and p_3). From Eq. 4-12 we have the three differential equations:

$$R = \left|\begin{array}{ccc|c} -\Sigma c_1 & k_{21} & k_{51} & 0 \\ k_{12} & -\Sigma c_2 & k_{52} & 0 \\ k_{15} & k_{25} & -\Sigma c_5 & 0 \end{array}\right|$$

which should be read as follows (remember that $\Sigma c_1 = k_{12} + k_{15}$):

$$\begin{array}{lll} -p_1\Sigma c_1 + p_2 k_{21} + p_3 k_{51} = 0 & \text{or} & -p_1(k_{12}+k_{15}) + p_2 k_{21} + p_3 k_{51} = 0 \\ -p_2\Sigma c_2 + p_1 k_{12} + p_1 k_{52} = 0 & \text{or} & -p_2(k_{21}+k_{25}) + p_1 k_{12} + p_3 k_{52} = 0 \\ -p_3\Sigma c_3 + p_1 k_{15} + p_2 k_{25} = 0 & \text{or} & -p_3(k_{51}+k_{52}) + p_1 k_{15} + p_2 k_{25} = 0 \end{array}$$

As it should be in equilibrium, the number of forward and backward transitions to a particular state is the same.

Unfortunately, one of the three equations is redundant; the matrix, R, is said to be singular. Therefore, a new one must replace one of the equations. We might impose that the sum of the probabilities to be in any of the closed states is unity:

$$p_1 + p_2 + p_3 = 1$$

The system of equations, with M the modified matrix R, can thus be modified into:

$$M = \left[\begin{array}{ccc|c} -\Sigma c_1 & k_{21} & k_{51} & 0 \\ k_{12} & -\Sigma c_2 & k_{52} & 0 \\ 1 & 1 & 1 & 1 \end{array}\right]$$

This set is easily solved by hand for the case of only three equations, but becomes much more difficult with increasing number of closed states. Fortunately, computer routines carrying out matrix inversion can do the job for us (see Watkins, 2002).

The Open Time Distribution

The open time distribution is calculated in much the same way as the closed time distribution. The only thing that changes is that we swap columns and rows of the transition matrix such that all transitions between open states (rather than closed states) end up in the upper left corner of the matrix.

The Macroscopic Current

With all the knowledge we have now, the calculation of the macroscopic current associated with a given Markov chain is easy. First we set up a transition matrix, A, just like we have done in the first latency paragraph. Next we calculate the eigenvalues and eigenvectors. In the last step we have to calculate the scaling factors for each of the eigenvectors from the initial conditions.

Pseudo code for routines to calculate the macroscopic current and dwell time distributions can be found in Appendix H.

Example: Simulation of the Hodgkin and Huxley Voltage-Dependent Sodium Channel

As we have seen previously, the voltage-dependent sodium channel according to Hodgkin and Huxley can be described by the action of four independent gates. Their mechanical model can be translated into the reaction scheme of Fig. 4-66.

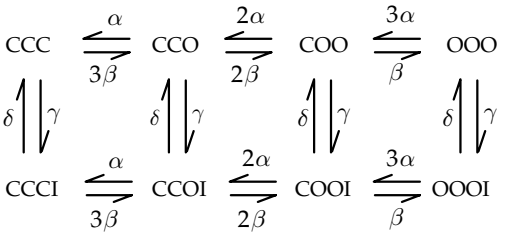


Fig. 4-66 Translation of the Hodgkin and Huxley model of the voltage-dependent sodium channel into a Markov reaction scheme. Each activation gate opens with a rate constant α and closes with a rate constant β. The forward and backward rate constants for the inactivation gate are γ and δ respectively. CCC means all three activation gates are closed and the inactivation gate is open. CCOI means two activation gates closed, one open and the inactivation gate closed. Only the state OOO is conductive.

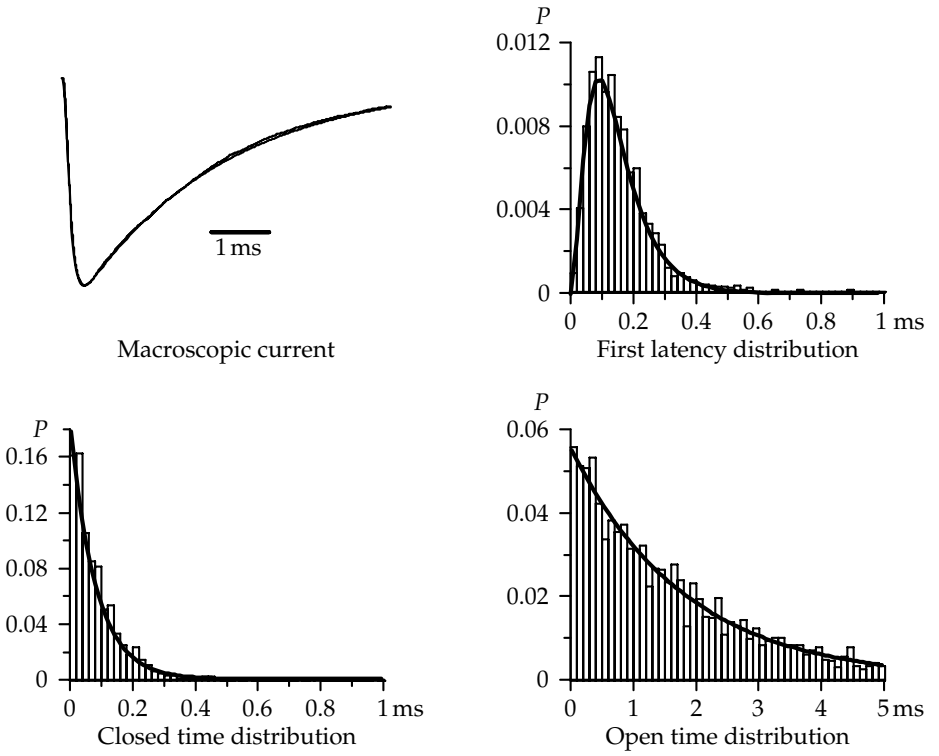


Fig. 4-67 Comparison between dwell time distributions obtained with the Markov chain approach and Monte-Carlo simulation of the Hodgkin and Huxley voltage-dependent sodium channel. Bar dwell time histograms were obtained with Monte-Carlo simulation and smooth lines were calculated by eigen decomposition. Note that for the macroscopic current, the two simulations overlap to such an extent that the curves are indistinguishable.

With $\alpha = 50\,s^{-1}$; $\beta = 12\,000\,s^{-1}$; $\gamma = 400\,s^{-1}$; $\delta = 7\,s^{-1}$, which are the rate constants corresponding approximately to a membrane depolarization to -10 mV, the following transition matrix can be made (zero entries are left blank for clarity):

$$\begin{bmatrix}
 & 50 & & & & & & 7 \\
36\,000 & & 100 & & & & 7 & \\
 & 24\,000 & & 150 & & 7 & & \\
 & & 12\,000 & & 7 & & & \\
 & & & 400 & & 12\,000 & & \\
 & & 400 & & 150 & & 24\,000 & \\
 & 400 & & & & 100 & & 36\,000 \\
400 & & & & & & 50 &
\end{bmatrix}$$

Then applying the routines described above, the results shown in Fig. 4-67 are obtained.

Appendices

A: SYMBOLS, ABBREVIATIONS AND CODES

Symbols

Quantity		Unit	
A	area	m^2	
a	ion activity	M	molar
C	capacitance	F	farad
c	ion concentration	M	molar
d	distance, diameter	m	metre
E	electromotive force	V	volt
F	force	N	newton
f	frequency	Hz	hertz
f	rate	s^{-1} or Ad	adrian
f	ion activity coefficient	(dimensionless)	
G	gain	(dimensionless, often in dB)	
g	conductance	S	siemens
H	magnetic inductance	T	tesla
I	current	A	ampere
J	current density	A/m^2	
L	self-inductance	H	henry
l	length	m	metre
P	power	W	watt
Q	charge	C	coulomb
R	resistance	Ω	ohm
r	radius or distance	m	metre
S	ionic strength	M	molar
t	time	s	second
U	tension, voltage	V	volt
u	ion mobility	cm^2	$s^{-1}\,V^{-1}$
V	voltage	V	volt
W	work, energy	J	joule
X	reactance	Ω	ohm
Z	impedance	Ω	ohm
ε	dielectric constant		
ρ	resistivity	Ωm	
μ	Magnetic permeability		
τ	time constant	s	
ω	angular frequency	rad/s	

Abbreviations

AC	alternating current
AD	analogue to digital
ADC	analogue-to-digital converter
AVO	ampere, volt, ohm (meter)
BCD	binary-coded decimal
BIOS	basic input/output system
BW	bandwidth
CMR(R)	common-mode rejection (ratio)
CPU	central processing unit
CRO	cathode ray oscilloscope
CRT	cathode ray tube
DA	digital to analogue
DAC	digital-to-analogue converter
DC	direct current
DOS	disk operating system
DRAM	dynamic random-access memory
DVM	digital voltmeter
ECG	electrocardiogram
EEG	electroencephalogram
EMF	electromotive force
ENG	electroneurogram
ERG	electroretinogram
FET	field-effect transistor
GND	ground
jFET	junction field-effect transistor
LDR	light-dependent resistor
LIX	liquid ion exchanger
LJP	liquid junction potential
MOSFET	metal-oxide silicon field-effect transistor
OS	operating system
PCB	printed-circuit board
RAM	random-access memory
RFI	radio-frequency interference
ROM	read-only memory
RMS	root mean square
SPST	single-pole, single throw (switch)
SPDT	single-pole, double throw
VCO	voltage-controlled oscillator
XOR	exclusive-or (function, circuit)

Decimal Multipliers

T	tera	10^{12}
G	giga	10^{9} (*)
M	mega	10^{6} (*)
k	kilo	1000 (*)
m	milli	10^{-3}
μ	micro	10^{-6}
n	nano	10^{-9}
p	pico	10^{-12}
f	femto	10^{-15}
a	atto	10^{-18}

(*) Note that in computer jargon, "k" stands for 1024, "M" for 1024^2 (1 048 576) and "G" for 1024^3 (1 073 741 824).

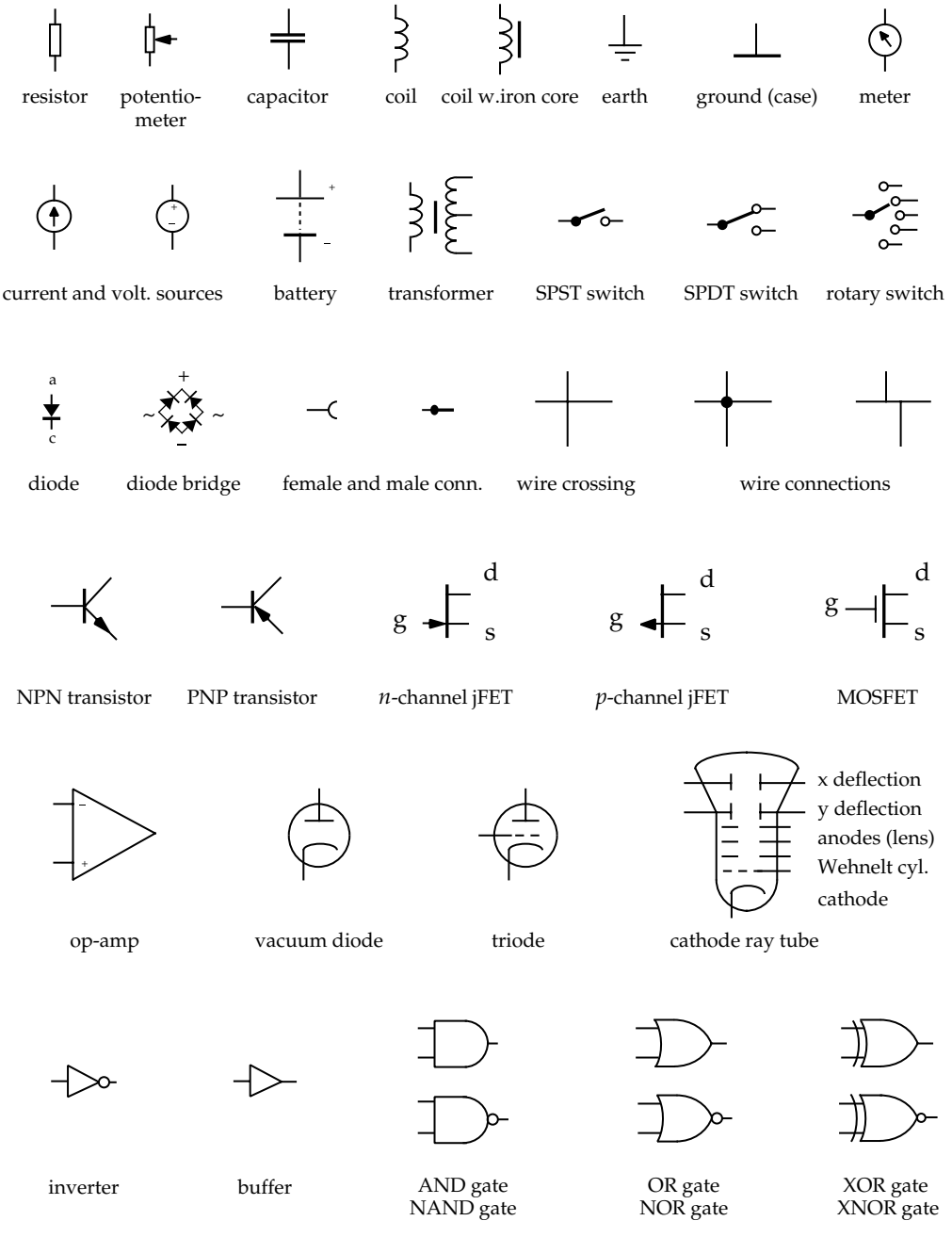
Colour Code for Resistors

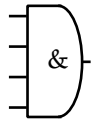
silver		10^{-2} or 10% tolerance
gold		10^{-1} or 5% tolerance
black	0	no zeros
brown	1	0 or 1% tolerance
red	2	00 or 2% tolerance
orange	3	000
yellow	4	0 000
green	5	00 000
blue	6	000 000
violet	7	10^{7}
grey	8	10^{8}
white	9	10^{9}

Use:

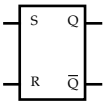
for carbon resistors: 4 rings ddm t (2 digits, multiplier and tolerance).
for metal film resistors: 5 rings dddm t.
occasionally for capacitors: ddm signifying value in pF.

B: SYMBOLS FOR CIRCUIT DIAGRAMS

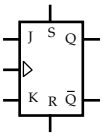




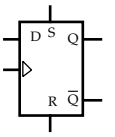
4-input NAND



RS flip-flop



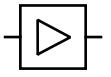
JK flip-flop

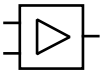


D flip-flop

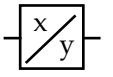

pulse generator


sine generator


amplifier


differential ampl.


oscilloscope



x-to-*y* converter


metal microelectrode


glass microelectrode


pH glass electrode


reference electrode


ECG (einthoven) electrode


EEG or precordial electrode


ground electrodes

C: ELECTRICAL SAFETY IN ELECTROPHYSIOLOGICAL SET-UPS

Regular Instruments

Regular electronic instruments, used for electrophysiological work, are reasonably safe as to the hazard for electric shocks from the mains supply. In principle, any set-up is grounded, which means that all metal parts that can be touched are connected to earth. The earth, or ground, connection runs along the mains wires and is provided by the local electricity company, at least in some countries. The types of connectors differ per country: rim clip, third pin, round or flat pins, etc. All serve the same purpose, viz. connecting the joint metal parts of an instrument to ground. If a fault occurs in which the mains voltage comes into contact with such a part, the very low impedance of the ground cables limits the voltage that will build up to safe values. Usually, the corresponding mains fuse blows, signalling the fault condition.

Electrophysiologists face a dilemma, though. The ground connection provided along with the mains wires suffices for safety, but is often too burdened with interference signals to be used as a ground for the recording of weak electrical signals. Many sources contribute to the spurious voltages on the ground wires. In the first place, gas discharge lamps (fluorescent lighting, high-intensity microscope illuminators, etc.) produce inductive transients, because they ignite and quench twice in each period of the mains frequency (twice because the voltage must be above a certain threshold whereas the direction of the current in the lamp is unimportant). A second notorious source of interference signals are motors switching on and off, such as are found in refrigerators, centrifuges, thermostatic baths and so on. Therefore, many labs have installed their own "clean" earth connection, usually in the form of an iron pipe driven deeply into the ground. Such grounding points can be ordered from companies installing lightning rods.

If one possesses such a clean, extra earth connection, however, one is blessed with two grounds, and interconnecting them is certainly not a good idea! In that case, a ground loop is formed that may encompass many square metres. This may cause substantial currents to flow, again deteriorating the properties of the grounding point as a reference for measurement circuits (A in Fig. C-1). The dilemma, then, lies in choosing which ground to use. If one decides to use the clean measurement ground, all ground cables and connectors to the mains ground circuit must be cut. This violates the safety rules, because any instrument connected to the mains outlet would lack its proper safety ground until the connection to the clean ground is made.

Apart from the aspect of safety for the people that operate laboratory instruments, the safety of the instruments themselves is at stake when one abandons the normal ground circuit. Failing to ground a digital instrument such as a computer may blow all its integrated circuits ("chips") in a split second. Therefore, as a golden rule, NEVER use computers without a ground connection. True, some people use their PC's at home, in rooms not fitted with grounded outlets, seemingly without nasty consequences. However, proper functioning is not warranted (by the manufacturer), and damage to data may show up very irregularly, hiding the true cause. In addition to leakage currents from the power line, static electricity, which may build up simply by walking across a nylon carpet, can play havoc with electronic circuitry.

A compromise is possible, however, by interconnecting the two grounds through a small series resistance (B in Fig. C-1). In this case, oscilloscopes and other instruments can be grounded safely to the mains ground, whereas a pre-amp or preparation grounding point is connected to the clean instrument ground. This may seem a contradiction, but remember that the spurious voltages across a ground circuit amount to no more than a few millivolts.

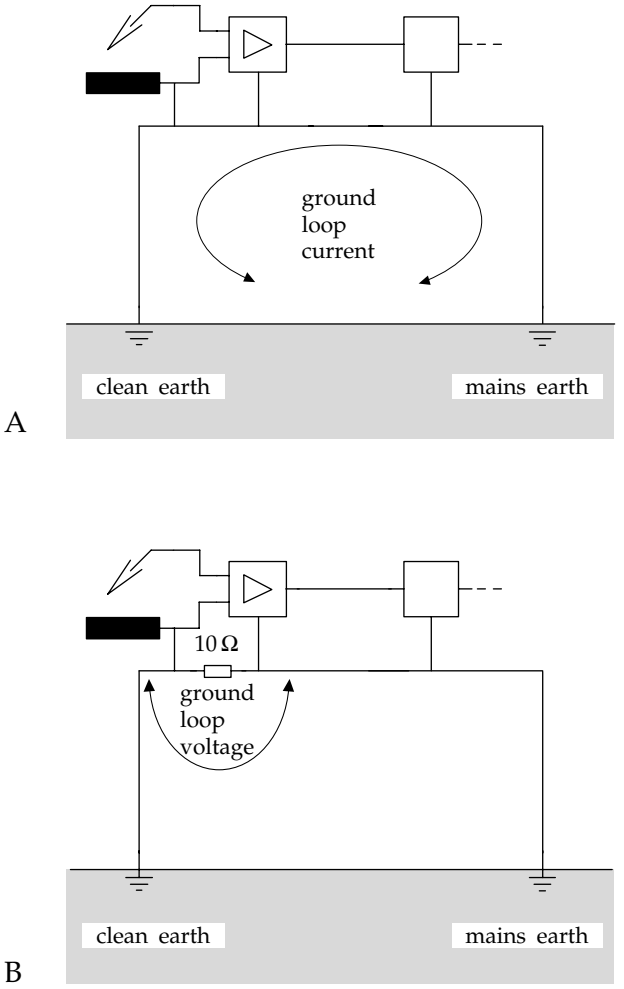


Fig. C-1 Grounding methods: (A) conventional, (B) with ground loop resistor.

The reason these low voltages are harmful to sensitive measurements stems from the low impedance of the ground circuit. Five millivolts of interference signal across 10 mΩ (milli ohm) of cable resistance causes a current of 0.5 A to flow through the entire ground circuit. If the ground loop resistance is increased to 10 Ω, which is a factor of 1000 higher, the current is reduced by the same amount. In this case, almost the entire interference voltage appears across this 10 Ω resistor. In the case depicted in B, that voltage forms only a minor contribution to the (amplified) output voltage. The same principle of segregating ground circuits is employed in some hybrid (digital and analogue) instruments, such as multimedia computers. Here, the ground circuit is split into a "digital ground" and an "analogue ground", because the fast-switching digital circuits produce spurious transient currents to ground.

What about the inherent safety of the instruments involved? Laboratory instruments are (or should be) built according to international standards guarding the safety of mains-operated gear. These standards are set up by the International Electrotechnical Committee (IEC) and

published in a large series of numbered volumes. For lab instruments, the safety standards are described mainly in IEC 1010 ("Safety requirements for electrical equipment for measurement, control and laboratory use") and IEC 664 ("Insulation coordination for equipment within low-voltage systems").

Like household apparatus, lab instruments employ double insulation in cables and in any internal connections that are exposed to wear or stress. Double insulation prevents one from touching a dangerously high voltage under a so-called single-fault condition. This means any situation in which the basic layer of insulation around a live wire is interrupted or damaged, or any single inadvertent contact with a mains-voltage carrying conductor.

IEC 1010 specifies, among other things, the minimum distances required between conductors carrying relatively high voltages. In general, any AC voltage of over 30 V rms (42.4 V peak) and any DC voltage over 60 V is considered as a "hazardous live". Obviously, mains voltages belong to this category.

By specifying the minimum distances ("clearance", meaning the size of the gap in air, and "creeping", which is defined as the shortest distance across an insulating body), the risk of an electric shock is considered to be reduced to acceptable levels. However, even if these standards are met, an electric current may flow between circuits that are connected to the mains and the circuits that may be touched by humans. This current, called the leakage current, arises in principle in all parts that carry alternating current and have mutual capacitance, which amounts to everything. The influence depends of course on the magnitude of the capacitance. In mains-powered instruments, the principal source is the transformer used to step down the mains voltage. The primary winding carries the mains voltage, and is wound tightly around the core. Usually, the secondary coils are wound directly on top of the primary. Because of the relatively high capacitance between parts so close together, both the core and the secondary windings carry a leakage current. Under circumstances of proper grounding, this current will flow to ground. However, in an ungrounded instrument, the leakage current may develop sizeable potentials on the metal parts, which can be felt if one touches them.

Another important source of leakage current at the mains frequency stems from interference filters built into most electronic equipment. Some instruments, such as vacuum cleaners, refrigerators and other frequently switching machines, generate short peak currents and/or voltages that may flow back into the power lines. In other instruments, especially digital ones, these peaks may cause a malfunction. To damp the peaks, most instruments are fitted with an interference filter, usually an LC low-pass filter, at the entry point of the mains/power cable. The most common type is shown schematically in Fig. C-2.

The values of the components differ a bit between brands, but are in the order of magnitude given below:

C_1: about 10–100 nF
C_2: about 2–5 nF
L: about 0.5 mH

Note the capacitive voltage divider formed by the two capacitors C_2. If the ground lead is omitted, the instrument case will float at half the mains voltage (i.e. 55 V AC in the USA, 115 V AC in Europe). If one touches the case of an instrument under such conditions, a current will flow through the circuit: mains—capacitor C_2—body—shoes—ground. The values of the capacitors involved (C_2) limit the current to a harmless value of about 0.36 mA (@ 230 V, 50 Hz,

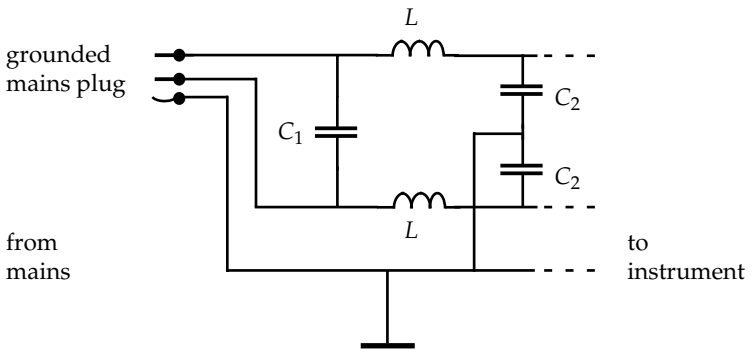


Fig. C-2 Mains interference filter.

$C_2 = 5\,\text{nF}$). However, if the (single!) insulation between the two plates of that capacitor would give way, one would be in direct contact with the full mains voltage and power.

Medical Instruments

For medical apparatus, i.e. instruments that are intended to be connected to a human subject, more stringent safety rules apply. Thus, one is allowed to connect intact animals or preparations to a regular set-up, but no human subjects. For use on human subjects, more strict rules apply as to shock hazard, leakage currents from the mains voltage, etc. These rules are laid down in IEC 601 "Medical electrical equipment" and its later amendments. In short, IEC 601 defines the apparatus and parts used, and states rules as to the specifications to which medical electronic instruments must be built, as well as the methods to test them. Examples of these standards are given below.

To reduce the risk of an electric shock, medical electronic instruments are built with a higher degree of electrical insulation between such components as the primary and secondary windings of transformers. Additional circuits serve to insulate the patient circuit from the rest of the (mains-operated) set-up or instrument. These additional safety measures explain at least part of the higher price one must pay for medical-grade apparatus. Alas, painting an oscilloscope white does not turn it into a medical-grade one!

The most reliable circuits to insulate an input circuit from mains-operated instruments are optical and radio-frequency (RF) couplers, or isolators. These consist of a transmitting part at the patient side and a receiving part at the mains-operated side. By this principle, an opto-coupler consists of a light-emitting diode, the intensity of which is modulated by the signal (patient ECG, etc.), a transparent plastic layer that can withstand high voltages, and a photodiode or similar light detector to pick up the emitted light signal (see A in Fig. C-3). An RF isolator (B in Fig. C-3) employs a tiny transmitter and a receiver. The transmitting and receiving coils are separated by a plastic sheet, yet so close together that no antennas are necessary, and no appreciable radio wave signal escapes to the environment. For these patient protection devices to function, the patient side of the device must be powered either by batteries or by a mains supply that has equally well-insulated components (see SELV p. 220). Even within the category "medical", instruments on the market differ in the degree of protection against electric shocks and excessive leakage currents.

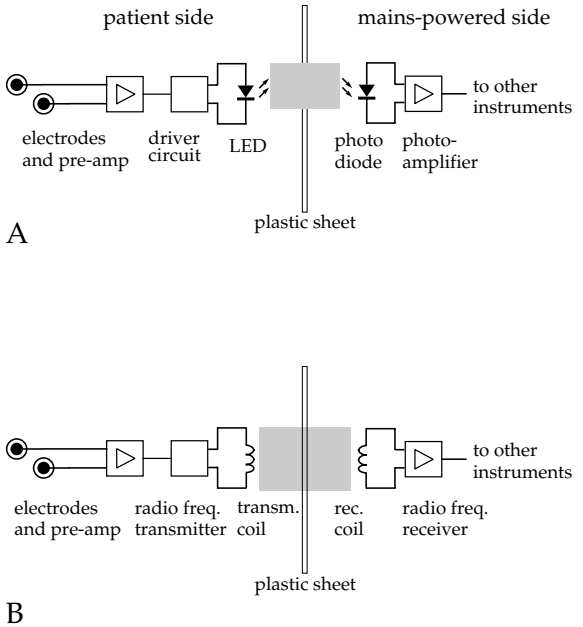


Fig. C-3 Opto-isolator (A) and RF isolator (B).

Therefore, IEC standards segregate medical instruments in classes and types, the most important of which are described briefly (and in a less formal way than in the corresponding IEC standards books) below.

Class I This class has, in addition to basic insulation, its metal parts connected to a protective earth conductor. In short, grounded apparatus. In our homes, washing machines, refrigerators, etc. belong to this category.

Class II This class has an extra layer of insulation, in other words, these are double-insulated instruments. Many household appliances fall into this category: shavers, hair dryers, portable audio, etc.

Type B Medical equipment specified as to maximum leakage current and reliability of earth connection. As a guideline, the patient leakage current in this type of apparatus under normal circumstances will be lower than 100 μA.

Type BF This is the same as B, but having an attached part of type F (floating; see below). The floating part consists usually of the pre-amplifier in contact with the patient electrodes.

Type C This type meets higher demands as to leakage and safety. Usually, it is fitted with a floating part (type CF). As a guideline, the patient leakage current in this type of apparatus under normal circumstances will be lower than 10 μA.

Type F This is a part of the equipment that is floating, i.e. so well insulated from ground that no appreciable leakage current will flow even when that part should come into contact with the mains voltage. Used as an addition to types B and C.

Instruments with the highest degree of protection are designated as “safety extra low-voltage”, or SELV, equipment. This is defined as being lower than 25 V AC or 60 V DC. These instruments are powered by low voltages, usually derived from the mains voltage by a special, so-called SELV transformer. Such transformers meet very high insulation standards to avoid electrical shocks and excessive leakage currents in the patient circuit. Obviously, instruments used in heart surgery, etc. are built to the highest safety standards.

Instruments that do not meet the appropriate standard from the list above are not safe enough to be used for the respective purposes. It is the responsibility of the laboratory staff to verify whether their instruments meet the standards required to work safely.

D: THE USE OF CRT MONITORS IN VISUAL EXPERIMENTS

Image Generation in CRT Monitors

Since the advent of television, people are accustomed to looking at CRT screens (cathode ray tube screens, or TV screens). A CRT is very akin to the oscilloscope dealt with in Chapter 2. In its application as a picture tube, however, the image is generated by having the narrow (focussed) electron beam scan the whole screen to build up a rectangular image. The screen is covered on the inside with the so-called phosphors, minerals that emit light when struck by the electron beam. Ideally, the scanning is performed so fast that, for the human eye, it merges into a stationary image.

The image is built up by scanning the object in a pattern of horizontal lines, left to right, then top to bottom. Each of these images is called a "frame". In a TV or video system, scanning one frame takes about 33 to 40 ms (i.e. a frame rate of 30 or 25 Hz respectively). Computer monitors scan faster: usually 75 to 85 fps (frames per second). After each line, the spot has to be swung back to the left, and after each frame, back to the top. During both of these so-called flybacks, the intensity is suppressed (i.e. black). See Fig. D-1.

A television set has, in addition to the image-generating circuits, a receiver section, sound circuits, and often additions like videotext, picture-in-picture and so on. Newer sets have a digital image memory, which stores each frame to be shown, and so permits higher scanning rates for display, making the image look more quiet and stable.

Today, the CRT screen, usually called a "monitor", is still the most widely used type of display for computers, closed-circuit video, etc. A monitor must be fed a so-called video signal, which consists minimally of a luminance (brightness) signal and two sync (synchronization) signals. In addition, most systems provide a few extra signals, which code image colour. The signals necessary are explained below.

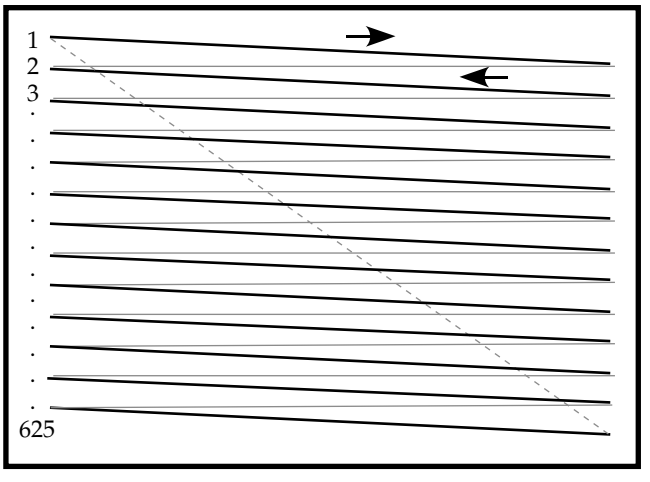


Fig. D-1 Scanning pattern on a picture tube. The image is generated in the same way as we read: left to right and top to bottom. The thin horizontal lines are the (invisible) flyback traces, which take only a few microseconds. The raster flyback, drawn as a diagonal dotted line, actually takes as much time as it takes to write 12 lines.

The 50 or 60 Hz frame rates were chosen for technical and economic reasons. Unfortunately, the frame rates chosen in the early days, and still in use today, are barely high enough for the human eye to see a TV picture as a stationary image: most people see a 50 fps image as unquiet, if not outright flickering. It is most conspicuous (and annoying) when viewing a screen from the corners of the eyes, because in the human retina, the peripheral photoreceptor cells respond faster than the central ones. In the beginning, the intensity of picture tube images was rather low, but nowadays, more efficient and endurable phosphor types allow intensities that are better suited for daytime use. However, at higher intensities, the human visual system works faster, making the flickering more obtrusive than before. Therefore, computer monitors evolved not only to display more pixels on a line and more lines in a frame, but also to employ higher frame rates, rendering far more stable-looking images. Newer TV standards, such as the DVB system used by TV satellites, also allow for better image quality. Nevertheless, the high broadcast frequencies used (10–12 GHz) allow for thousands of channels.

Frame Rates and Interlacing

In the beginning of TV times, the frame rate was chosen to be synchronous with the power mains frequency (namely, to avoid hum being visible as a moving darker or lighter band in the image). Thus, north American TV (NTSC standard) has a frame rate of 60 fps, whereas the European TV (PAL standard) provides for 50 fps. The information rate (the bandwidth, or what we would call "the pixel rate" in a digital system) in both systems is approximately the same. The difference lies in the number of lines per frame. American TV generates 525 lines per frame, the European system 625. The bandwidth of the video (luminance, or intensity) signal is about 5 MHz. Sending 50 or 60 fps would, however, need twice the bandwidth (then) available.

Therefore, in practice, a TV frame is composed of two half-frames, which are interlaced to form the entire frame. Such a coarse or half-frame is called a "field". Thus, the odd lines 1, 3, 5, . . . , 525 are written first (indeed within 1/50 s). Subsequently, the beam jumps back to the top, but with a slight offset, so that the next field (containing the even lines 2, 4, . . . , 524) is interlaced with the odd ones (Fig. D-2). Thus, the image source has to provide only 25 or 30 fps, whereas the eye sees 50 or 60 scans per second. Very close to the screen, one can see this interlacing as a slight hopping or flickering of adjacent lines. At normal viewing distances, however, both fields blend nicely into one more detailed image.

Computer monitors have higher frame rates, today over 100 fps, and do not use interlacing. Here, image generation is called a "progressive scan", where all lines are written in the top-to-bottom order. Nevertheless, the image is still built up by scanning, and this has consequences for some applications that are described below.

The Video Signal

A video image must be written onto the screen, and this is performed by scanning a raster (rectangular area). As you can see in Fig. D-2, the lines are written slightly downward (because of the time it takes to write from left to right), and the flyback is faster, hence more horizontal. Obviously, during the flyback, the electron beam must be pinched off to avoid writing on the screen. This is called "blanking". Blanking must take place during both line flybacks and raster (field) flybacks. The brightness values of the image, black through white, are represented by

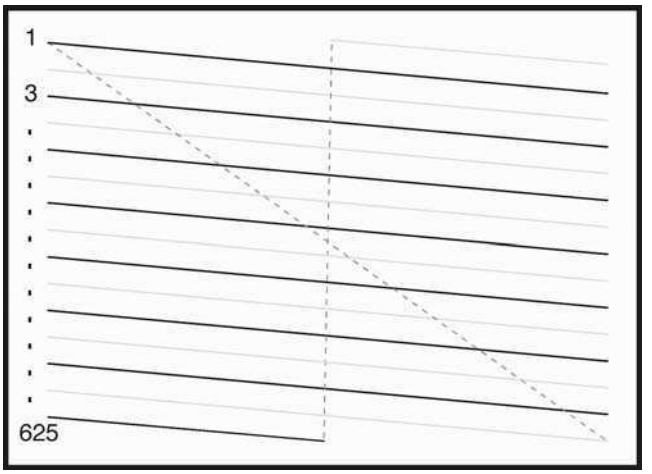


Fig. D-2 Interlaced scanning pattern, with line flybacks omitted. The first field consists of the odd-numbered lines (black). In the middle of the last line, the spot flies back to the point where the second field (grey) must be filled in. In fact, this yields a half "line zero". During flyback, the beam is blanked out.

voltages, as is the blanking. In addition, other voltage levels signal the sync pulses. A video camera must provide all these signals, either separately or combined. All in all, a combined, or composite video signal is a complicated waveform, shown in Fig. D-3. Most TV standards, such as the American NTSC (colour) and EIA (B&W) standards as well as the European PAL and SECAM (colour) and CCIR (B&W), use a 1 V total amplitude, subdivided into sync, blanking and image brightness values. White is the highest voltage, the sync level the lowest. The blanking level is just a bit lower than black, ensuring absolute invisibility of the flyback traces.

At the other end, a TV set or monitor analyses this composite signal into its original components, feeding brightness values to the picture tube, and sync pulses to the line and raster generator electronics. Image brightness and sync signals can be discriminated by voltage: 0 to about 0.3 V is a sync pulse, 0.3 V is the blanking level, and higher values are image content. The simplest technique to separate the horizontal (line) and vertical (field) sync pulses, still used

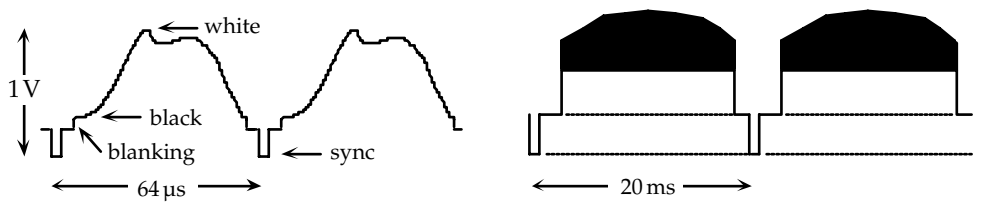


Fig. D-3 Schematic diagram of a composite video signal. Because the signal is detailed to several time scales, the diagram is split into two parts. The left diagram shows two consecutive lines, with the h-sync (line sync) pulses. The white, black, blanking and sync pulse levels are shown. In the right drawing, the signal is staggered to show two entire video fields, separated by v-sync (field sync) pulses. The dotted lines represent the h-sync pulses (not to scale; the distinction between odd and even fields is also not visible on this scale). The times indicated are for the European (CCIR) system.

in TV sets and video monitors, is to feed the sync signal into both a low-pass and a high-pass filter. So, the first one passes the long v-sync pulses while attenuating the very short h-sync pulses, whereas for the second one it is the other way around. However, video recorders, digitizers, etc. use a more precise identification of the sync signals. The recording industry uses this fact to build copy protection schemes into commercial video tapes and DVDs by tampering with the sync signals in a way that does not interfere with viewing the movies on screen, but prevents copying.

In colour television, the signal is still more complicated to allow encoding the colour of each image point. In principle, a colour video signal consists of five signals: red, green, blue, h-sync and v-sync. To transmit such a signal onto a radio wave carrier, all these signals must be combined into one composite colour video signal, which we will not treat in detail here. However, things are different (easier) in computer monitors. Because the computer and the monitor are connected with a short cable, the red, green, blue and sync signals can be sent through separate wires. These monitors are indicated as RGB. Since the combination of signals leads to compromises in quality, RGB video has a better quality than composite video. An intermediate form is S-video (S for separate), where the signals are split between two conductors, rather than five.

The early TV standards, still in use for broadcasting, were chosen as a compromise between image quality on one side and, on the other hand, both the limitations in the technique of that time and the fair distribution of radio channels (then only VHF). Thus, the total bandwidth of a TV channel was (and is) limited to about 7 MHz. The bandwidth of any video signal is the product of the number of separable image points (pixels in digital systems) and scan lines (the so-called definition) and the number of frames per second. Both standards, 525 lines at 30 fps and 625 lines at 25 fps, yield a bandwidth of about 5 MHz. Including the sound channel encoded with the images, a TV "channel" is 7 MHz wide. Increasing the bandwidth would reduce the number of channels, and hence stations that can broadcast simultaneously in one frequency band (such as VHF). Computer monitors are connected directly to their screens, and so the bandwidth is far less limited. Modern, high-resolution monitors provide images of, say, 1280×960 pixels per frame, and 70 to over 100 fps. The bandwidth may amount to 100 MHz or more. Although this poses no problem in principle, it is the reason you cannot extend monitor cables at will without deteriorating the image.

The Use of CRT Monitors in Electrophysiology

Cathode ray tube screens are used frequently in many branches of science, for widely varying purposes. In biology, electron microscopes, confocal microscopes and an increasing number of "normal" light microscopes are fitted with a video camera and monitor. In these cases, the screen image is supposed to be a faithful reproduction of a real-world object. In visual science, computer monitors are used as visual stimulators. Here, the screen image is computer generated, and should reproduce the investigator's intentions as to colour, size, shape and timing of the stimuli.

The image on a computer monitor comprises a rectangle, centred on the screen, surrounded by a small black border. In contrast, a TV image fills the entire screen surface, leaving no black borders. In order to ensure this screen-wide image, a TV or video monitor has a certain degree of overscan, i.e. the scanned raster is somewhat larger than the viewable area. Television programme and movie producers take this into account, and keep important details far from the

image borders. When one intends to use a TV screen or video monitor for scientific purposes, this should be kept in mind: drawing or programming image content should remain within a "safe" area in the middle of the screen.

Computer monitors do not suffer this problem. However, when using CRT devices to present images for scientific purposes, a number of aspects have to be checked or controlled. These deal with colour and brightness rendition, line and raster linearity, and especially with all sorts of timing issues.

Contrast, Gamma and Other Brightness Issues

The notion of a gamma stems from classical photography. It is known for more than a century that the contrast of a photo depends on the film emulsion used and on the way the film is developed. Ideally, the faithful reproduction of the shades of any real-world object would imply that the light intensities in the scene are rendered linearly into image intensities. However, it was recognized in the early days of photography that this is often neither feasible nor desirable. Photographic film can span contrast ratios of at most 1000:1, usually less, so that most real-world scenes would be saturated (shadows pitch black and/or highlights bleached out). Photos printed on paper have a still lower contrast, say 50:1. By controlling gamma (explained below) in addition to exposure, photographers can take care that the most important image details are preserved.

In general, the relation between object brightness (B) and image brightness (b) is a power curve, $b = cB^{\gamma}$. The exponent, called *gamma*, reflects the shape of the brightness imaging curve. The constant c is called *contrast*. Note that gamma is not the same as contrast, although the effects may resemble each other; see Fig. D-4. A third quantity, brightness, simply shifts the entire curve upward (brighter) or downward (dashed line in left graph).

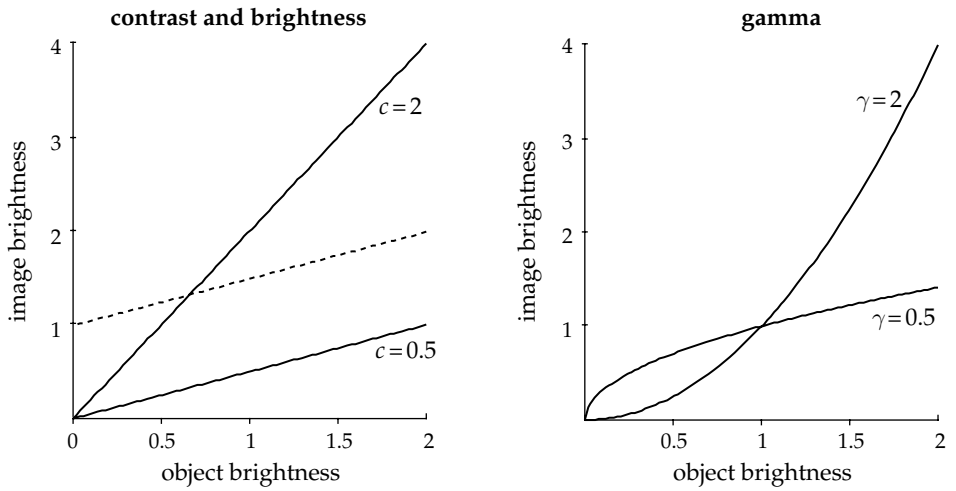


Fig. D-4 Imaging light intensity: contrast, brightness and gamma. Left: two contrast values at $\gamma = 1$; the dashed curve has $c = 0.5$ at a higher brightness level. Right: two gamma values at $c = 1$.

With a low gamma, image brightness rises fastest in the dark parts, slower in the lighter areas. A high gamma results in the opposite.

The gamma problems in the use of monitor screens stem from the fact that the gamma of the CRT device itself (i.e. the analogue imaging part) is far from 1, namely about 2.5. This is mainly due to the electrostatic properties of the so-called Wehnelt cylinder, the electrode in a picture tube responsible for regulating the electron beam's intensity. From the beginnings of television, the high gamma in the domestic TV sets was corrected at the transmitting end, namely by giving all TV cameras a gamma of 0.45. This results in an overall gamma of 0.45×2.5, or 1.125. A gamma slightly higher than unity yields pleasant, brilliant images. In colour television, it is important to apply the same gamma correction to all three colour channels (red, green and blue). Otherwise, severe discolouration can occur as a function of intensity.

In video cameras, the gamma correction of 0.45 is still in use. In addition, many commercially available CCD cameras (for microscopy, etc.) have the possibility to switch to a gamma of one. In this case, the video voltage is linearly dependent on object intensity, and may be evaluated by electrical measurements.

The gamma used in computer-generated screen images is a far more complicated matter. The on-screen gamma depends on the computer hardware, the monitor hardware and the software used. Many computer systems (Mac's, PCs) and programs (Photoshop) have their own gamma calibration programs. For a full appreciation, the reader is referred to the literature references given at the end.

In addition to the problems of non-linear brightness curves, the phosphors used to create the image will get dimmer in time, because they wear out by the incessant electron bombardment. Also, screen images meant to be uniform may be dimmer near the edges. This error can originate in the camera (image tube, optics, far less with CCD chips) and/or in the monitor (beam deflection geometry).

How critical a faithful image intensity is will depend obviously on the purpose of the image. Before getting desperate from the foregoing story, it is good to know that a monitor screen can be calibrated by a screen calibrator, a kind of hand-held photometer. With such an aid, contrast, gamma and colour rendition can be checked and calibrated.

A last complication to mention is the limit in intensity imposed by the high-voltage power supply that feeds the electron beam. The maximum power that this circuit can deliver limits the integral brightness of the whole screen. Thus, white lines and small patches can be far brighter than an entirely white screen.

Colour Coding

Cathode ray tube monitors and comparable systems (LCD screens, projectors, etc.) produce coloured images by the so-called "additive colour mixing" principle, where any colour can be built by tiny spots of red, green and blue light (RGB) in different proportions. This is very different from the daily experience of most of us, since we are accustomed to mixing paint rather than light to produce different colours. Mixing paint works according to the subtractive colour mixing principle. A white surface reflects approximately all light falling on it, and so looks white. Any dot of paint absorbs some of the light. If red is absorbed, the resultant colour is a shade of blue-green called cyan; absorption of green leads to magenta (a kind of purple), and absorption of blue yields yellow. So-called ink-jet printers use this CMY (cyan, magenta, yellow) scheme. Applying all three inks should yield black. However, because the black so

obtained is not as pure as a special black ink (such as soot), printers use black as a fourth "colour". Thus the system is indicated as CMYK (the K for black is to avoid confusion with blue).

In mixing light spots, things work the other way round. An unused monitor screen is black (apart from reflections, which should be avoided indeed). The tiny phosphor dots at the inner surface of the CRT emit either red, green or blue light when excited by an electron beam. The resultant colour is determined by the relative amounts of red, green and blue light emitted. The dots for red, green and blue are so close together that, from a proper distance, the three components fuse into a single colour. Using a magnifying glass, the phosphor pattern can be observed, though.

The computer or video camera must deliver three signals, which together code for red, green and blue. In principle, equal amounts of R, G and B should yield white. However, there is again a snag. In the first place, the human eye is not as sensitive for blue as for red and green light. In addition, the three phosphor chemicals are not equally effective. Therefore, the relative amounts of red, green and blue to mix in order to get "white" form a set of gain constants, built into the TV set or monitor. Third, there are several definitions of "white". In fact, when imitating tungsten light, one needs more red and yellow, and less blue then when rendering sunlight or fluorescent lighting. Most monitors allow adjustment of the white point to suit the needs of the user.

Geometry

The spatial reliability (shape) of images on a CRT screen depends critically on the linearity of the signals that generate the raster (called the *deflection signals*). The signals responsible for the horizontal (line) and vertical (raster) scanning of the screen should be perfect saw-tooth shaped. In practice, this is feasible only up to a certain degree. The relative amplitudes of the vertical and horizontal dimensions of the screen image are relatively easy to test and adjust, e.g. by measuring (or even by observing) how a circle is rendered. The human eye is fairly good at assessing deviations from the circular shape. However, non-linearities in the deflection signals tend to produce harder-to-recognize deformations, which would give a circle an ovoid shape (not merely elliptical but having a blunt and a more pointed side, like an egg). Although modern monitors perform relatively well, the on-screen reproduction of shape will never be as good as, for instance, a laser printer or photocopier.

Timing

Timing of stimuli on a CRT screen may present several difficulties. In the first place, the fact that the image is written with a flying spot implies that different parts of the image are written at different times. The horizontal time shift is very small (at most some 60 μs), but the bottom part of each image is written substantially later than the top part. Depending on the monitor's frame rate, this difference can be as large as 20 ms. Visual neurons in early cortical areas may respond to this timing difference. The timing differences will be less at higher frame rates. Fortunately, most computers in use today can provide for several rasters, having different spatial sizes and several relatively high frame rates.

Finally, the phosphors may play a part in the timing accuracy. A phosphor emits light not only when struck by the electron beam but also a short time afterwards. In most monitors, this

phosphor decay is approximately exponential, having a time constant of about a millisecond. However, this implies that traces of light may be visible for tens, or even hundreds, of milliseconds (when you shine a flashlight or strobe flash on your TV screen at night, you might see the whole screen light up for several seconds).

The most annoying timing artefact has been found by people recording electrophysiologically from the (vertebrate) visual system while using a CRT monitor to generate the visual stimuli. Often, the raster frequency shows up in the electrophysiological recordings, either as a frequency component or, with spike train recording, as phase locking of the spikes to the raster frequency. Figure D-5 is an unpublished graph of Britten and Heuer, showing a spectral peak at the raster frequency.

Spatial and Brightness Resolution

In addition to the foregoing, it should be realized that all computer-controlled images are digital, i.e. both spatial and colour aspects arise in discrete steps. A single picture element is called a "pixel". In the old-fashioned video signal, still much in use today, only the vertical dimension of the screen image is limited in a digital way: it is determined by the number of scan lines. The horizontal dimension is analogue. It is limited of course (by the bandwidth

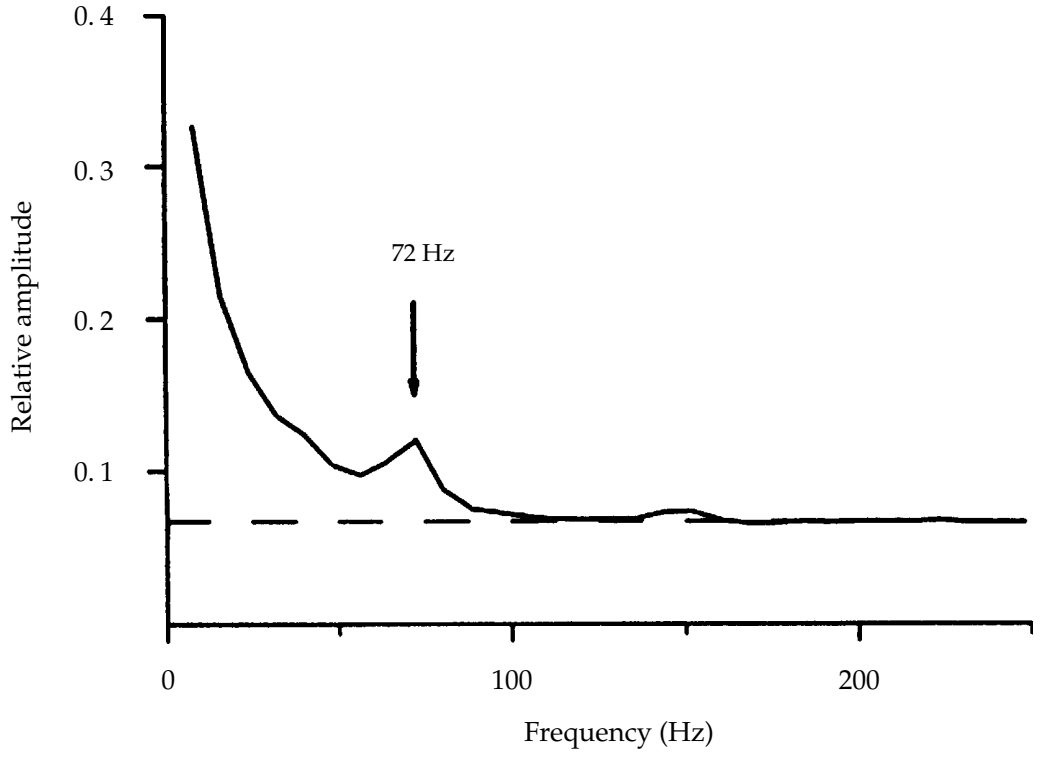


Fig. D-5 Spectrum of the spike rate from a neuron in the medio-temporal cortex of the macaque. The peak is at 72 Hz, the monitor's raster frequency. Courtesy of Dr K. Britten.

of the video circuitry), but not rendered in steps. In other words, a very small white dot on a black background (or the reverse) will be smeared out a little, but the centre of gravity of the spot can be at any place on the screen line. In modern video cameras, having a CCD chip for imaging, the horizontal component is also discrete, because the video chip consists of a 2D array of tiny, light-sensitive cells. The light intensity in these cameras is still an analogue quantity (omitting the limit caused by the quantal nature of light).

Obviously, computer-generated images are also digital in light intensity, and hence in colour. The so-called colour depth is the number of bits used to encode brightness. Most systems use an 8-bit colour depth, which means 256 levels from black up to (and including) white. Often, 0 stands for black, 255 for white. Colour systems have three such 8-bit channels, coding for the red, green and blue components of the colour image. Other colour systems exist, such as "Lab" (lightness and two colour components a and b) and "CMYK" (cyan, magenta, yellow and black, the pigments used in colour printing), but these are outside the scope of this text.

For most daily-life images, 8-bits black-and-white or 24-bits colour coding (8 bits per channel) is enough: the step size is barely visible, if at all. Note that, for some scientific purposes, higher-bit depth coding may be necessary. Also note that the step size is closely related to gamma. With any non-linear brightness relation, the step sizes in dark and in light areas will be different.

The next complication lies in the dot raster of the picture tube (CRT) itself. Colour is created by using three electron beams, which impinge on either red, green or blue phosphor dots. This raster resolution, generated by a video camera or a computer, should be compatible with the sizes of these screen dots, expressed as the dot pitch. Most monitors are multiscan, i.e. they accept a wide range of resolutions and frame rates, and most computers can generate different rasters, say from 640×480 pixels at 60 Hz (frames per second) up to perhaps 1024×768 or even 1600×1200 pixels (depending on monitor size) at 85 or 90 Hz. However, it will be useless to choose a raster resolution that would generate pixel sizes smaller than the screen dots of a given monitor. Even if screen dot size and pixel size are approximately equal, so-called moire patterns may arise, which disturb the image. The dot pitch of most monitors is in the order of 0.25 mm. This implies that one can generate useful rasters of about 1024×768 or 1280×960 pixels on a 17-inch monitor.

As you can see, the CRT display is a versatile device, which will be useful in a variety of applications, provided that one knows the working principles and heeds the warnings.

E: COMPLEX NUMBERS AND COMPLEX FREQUENCY

Complex numbers are denoted in general by $z = a + \mathrm{j}b$, where a and b are the real and imaginary part of the complex number z. The imaginary unit vector, j, has the property: $\mathrm{j}^2 = -1$. Thus, j is identical to the imaginary unit, notated i in mathematical texts. Since in electricity theory the letter i is used to represent electrical current, the letter j is chosen for the imaginary unit (convention in technical texts).

Basic operations of two complex numbers are straightforward:

Additition:

$$(a + \mathrm{j}b) + (c + \mathrm{j}d) = (a + c) + \mathrm{j}(b + d)$$

Multiplication:

$$(a + \mathrm{j}b) \cdot (c + \mathrm{j}d) = (ac - bd) + \mathrm{j}(ad + bc) \quad \text{(here we used } \mathrm{j}^2 = -1\text{)}$$

Division:

$$\frac{a + \mathrm{j}b}{c + \mathrm{j}d} = \frac{(a + \mathrm{j}b)(c - \mathrm{j}d)}{(c + \mathrm{j}d)(c - \mathrm{j}d)}$$

$$= \frac{(ac + bd) + \mathrm{j}(bc - ad)}{c^2 + d^2}$$

The so-called complex conjugate is very useful to simplify complex formulae, e.g. to segregate real and imaginary parts. The conjugate of a complex number z has j replaced by $-\mathrm{j}$ and is denoted here by z^*. Hence, if $z = a + \mathrm{j}b$, then $z^* = a - \mathrm{j}b$. Of course the conjugate of the conjugate of z is z itself: $(z^*)^* = z$. It can be easily verified that if x and y are complex numbers:

$$x + y = y + x \quad \text{and} \quad xy = yx$$

The real part of $x = a + \mathrm{j}b$, $\mathrm{Re}(x)$, can be reformulated using the conjugate:

$$\mathrm{Re}(x) = (x + x^*)/2, \quad \text{since} \quad ((a + \mathrm{j}b) + (a - \mathrm{j}b))/2 = a \tag{E-1}$$

and similarly,

$$\mathrm{Im}(x) = (x - x^*)/2\mathrm{j}, \quad \text{because} \quad ((a + \mathrm{j}b) - (a - \mathrm{j}b))/2\mathrm{j} = b \tag{E-2}$$

If $x = a + \mathrm{j}b$, then $xx^* = (a + \mathrm{j}b) \cdot (a - \mathrm{j}b) = a^2 + b^2$. The value $a^2 + b^2$ is always positive or zero. Its positive square root is called the *modulus* or the *absolute value* of the complex number x and is denoted by $|x|$. Hence $|x|^2 = xx^*$.

Since complex number are actually pairs of numbers, they can be depicted as vectors in a 2D space, where the vector projections on the two orthogonal axes represent the real and imaginary part of the complex number (Fig. E-1).

Suppose a is the angle between a vector ($\mathbf{v}$) of length r_1 and the real axis.
Then:

$\mathbf{v} = x + \mathrm{j}y = r_1 \cdot (\cos a + \mathrm{j} \sin a)$, where x and y are the coordinates of $\mathbf{v}$ in complex space.

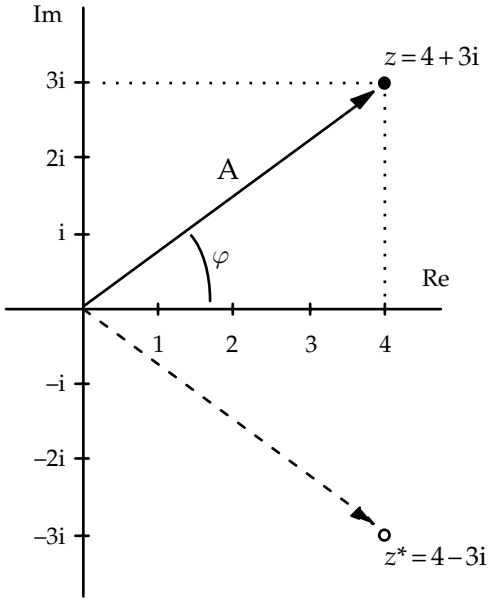


Fig. E-1 A complex number and its conjugate as vectors on the Re/Im plane.

Now, if we have another vector $\mathbf{w} = r_2 \cdot (\cos b + \mathrm{j} \sin b)$, where b is the angle between W and the real axis, then

$$\mathbf{vw} = r_1 r_2 \cdot [(\cos a \cos b - \sin a \sin b) + \mathrm{j}(\sin a \cos b + \sin b \cos a)]$$

which simplifies to:

$$\mathbf{vw} = r_1 r_2 \cdot [\cos(a+b) + \mathrm{j} \sin(a+b)]$$

Hence multiplication of two complex numbers involves the multiplication of their moduli and the summation of their angles.

It has been shown by Euler (using an argument based on Taylor series expansion, which we will not repeat here) that:

$$\cos a + \mathrm{j} \sin a = \mathrm{e}^{\mathrm{j}a}$$

when replacing a by $-a$ it follows that:

$$\cos a - \mathrm{j} \sin a = \mathrm{e}^{-\mathrm{j}a}$$

Combining this result with (E-1) and (E-2) gives the useful equations:

$$\cos a = (\mathrm{e}^{\mathrm{j}a} + \mathrm{e}^{-\mathrm{j}a})/2 \quad \text{and} \quad \sin a = (\mathrm{e}^{\mathrm{j}a} - \mathrm{e}^{-\mathrm{j}a})/2\mathrm{j}$$

$$\text{Rewriting with } \mathbf{v} = r_1 \cdot \mathrm{e}^{\mathrm{j}a} \quad \text{and} \quad \mathbf{w} = r_2 \cdot \mathrm{e}^{\mathrm{j}b} \quad \text{gives:} \quad \mathbf{vw} = r_1 r_2 \cdot \mathrm{e}^{\mathrm{j}(a+b)}$$

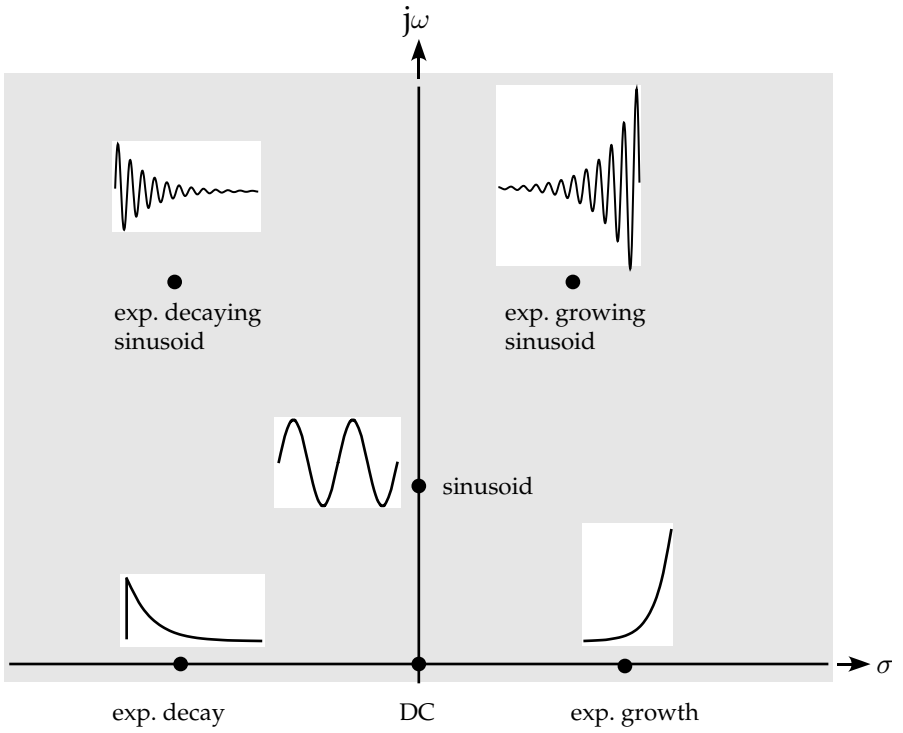


Fig. E-2 Explanation of the meaning of the complex frequency Laplace variable *s*. The real part is rendered along the *x* axis, the imaginary part along the y axis. The fat dots are examples of the different signal categories. Except at DC, a stylized figure shows what the signal at each point would look like.

The Meaning of Complex Frequency

It was stated earlier that the Laplace transform can be used for any signal, being either periodical (sinusoid, square and so on) or transient (pulse, exponential decay, etc.). This follows from the variable *s*, which is called *complex frequency*. This variable is complex, and so is in fact a pair of variables:

$$s = \sigma + j\omega$$

The quantity behind the j, called the imaginary part, ω, represents the conventional frequency ($\omega = 2\pi f$ rad/s). In addition, the real part of *s*, σ, stands for non-periodic (transient) functions, such as growth or decay. This can be illustrated best by plotting in a real (σ) versus imaginary ($j\omega$) graph again. Figure E-2 shows how all sorts of signals fit into this graph: sine waves of increasing frequency are lying on the *y* axis; exponential growth and decay are laid out on the *x* axis, with faster decay farther left, faster growth farther to the right. The origin represents a direct current (frequency zero; no growth and no decay).

F: THE MATHEMATICS OF MARKOV CHAINS

The presentation of the problem in Chapter 4 is a special case of the Markov chain approach, as, in addition, Markov chains may have both forward and backward rate constants and the linear chain as shown above may have side chains. In that general case and supposing we have three states, S_1 through S_3, then the system may be represented by a transition matrix:

$$\mathbf{A} = \begin{bmatrix} \cdot & k_{21} & k_{31} \\ k_{12} & \cdot & k_{32} \\ k_{13} & k_{23} & \cdot \end{bmatrix}$$

where the rate constants k_{ft} indicate the rate from state f to state t. Note that zero matrix elements are left out for clarity.

A Markov chain resumes in fact to a problem of solving a set of differential equations. With probabilities p_1, p_2 and p_3 to be in each of the three states S_1 through S_3, and the rate constants k_{ft} as shown above, the set of differential equations is:

$$\begin{aligned} \mathrm{d}p_1/\mathrm{d}t &= p_2k_{21} + p_3k_{31} - p_1\,(k_{12} + k_{13}) \\ \mathrm{d}p_2/\mathrm{d}t &= p_1k_{12} + p_3k_{32} - p_2\,(k_{21} + k_{23}) \\ \mathrm{d}p_3/\mathrm{d}t &= p_1k_{13} + p_2k_{23} - p_3\,(k_{31} + k_{32}) \end{aligned} \tag{F-1}$$

or in matrix form:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \begin{bmatrix} -k_{12} - k_{13} & k_{21} & k_{31} \\ k_{12} & -k_{21} - k_{23} & k_{32} \\ k_{13} & k_{23} & -k_{31} - k_{32} \end{bmatrix}$$

Hence it suffices to complement the Markov transition matrix at the diagonal entries with the negative of the sum over each column to obtain the corresponding set of differential equations. Equations like Eq. F-1 are written in matrix shorthand as:

$$\mathbf{X}' = \mathbf{A}\mathbf{X} \tag{F-2}$$

where $\mathbf{X}$ is a vector, $\mathbf{X}'$ the derivative of $\mathbf{X}$ and $\mathbf{A}$ is a square matrix.

This equation is suspected to have solutions of the form:

$$\mathbf{X} = \mathbf{C}\,\mathrm{e}^{\lambda t}$$

i.e. an exponential function, with $\mathbf{C}$ a constant vector.

Substitution into Eq. F-2 gives:

$$\begin{aligned} &\mathbf{C}\lambda\,\mathrm{e}^{\lambda t} = \mathbf{A}\mathbf{C}\,\mathrm{e}^{\lambda t} \\ &\text{or} \quad (AC - \lambda C)\,\mathrm{e}^{\lambda t} = 0 \end{aligned} \tag{F-3}$$

This can only be true if:

$$AC - \lambda C = 0$$

Because $C = IC$, where I is the identity matrix (a matrix with ones on the main diagonal and zeros elsewhere), Eq. F-3 becomes:

$$(A - \lambda I)C = 0$$

It has been proven (see Rainville and Bedient, 1981) that Eq. F-3 only has solutions if the determinant of $(A - \lambda I)$ is zero:

$$|A - \lambda I| = 0 \tag{F-4}$$

The determinant of a matrix is rather difficult to calculate by hand, but becomes much easier if the matrix is triangular, since the determinant of a triangular matrix is simply the product of its diagonal elements. Therefore, most computer programs try to reduce a matrix as in Eq. F-1 to triangular form by so-called similarity transforms:

$$\mathbf{A}_2 = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ \cdot & a_{22} & a_{32} \\ \cdot & \cdot & a_{33} \end{bmatrix}$$

See Watkins for a more in-depth treatment of the problem. $\mathbf{A}_2$ indicates the transformed matrix $\mathbf{A}$ and a_{ij} the, a priori, non-zero elements of $\mathbf{A}_2$. Then:

$$A_2 - \lambda I = \begin{bmatrix} a_{11} - \lambda & a_{21} & a_{31} \\ \cdot & a_{22} - \lambda & a_{32} \\ \cdot & \cdot & a_{33} - \lambda \end{bmatrix}$$

and therefore:

$$|A_2 - \lambda I| = (a_{11} - \lambda)(a_{22} - \lambda)(a_{33} - \lambda) = 0$$

This is a polynomial of degree 3 having three solutions:

$$\lambda = a_{11}, \quad \lambda = a_{22} \quad \text{and} \quad \lambda = a_{33}$$

The three lambdas are called the *eigenvalues* of the square matrix $\mathbf{A}$. The next thing to do is to find the (eigen)vectors, $\mathbf{C}$, of Eq. F-4. For $\lambda = a_{11}$ this will be:

$$(A_2 - \lambda I)\, C = \begin{bmatrix} 0 & a_{21} & a_{31} \\ 0 & a_{22} - a_{11} & a_{32} \\ 0 & 0 & a_{33} - a_{11} \end{bmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = 0 \tag{F-5}$$

where the vector $\mathbf{C}$ has three elements c_1, c_2 and c_3.

It follows that:

$$c_2 a_{21} + c_3 a_{31} = 0, \quad c_2(a_{22} - a_{11}) + c_3 a_{32} = 0 \quad \text{and} \quad c_3(a_{33} - a_{11}) = 0$$

This gives $c_2 = c_3 = 0$ and c_1 is free to choose. Hence an eigenvector corresponding to the eigenvalue $\lambda = a_{11}$ is:

$$\mathbf{C}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

The other two eigenvectors are obtained similarly. Now, since the three vectors correspond to the transformed system $\mathbf{A}_2$, the last thing that has to be done is to carry out the inverse similarity transform on the three vectors to obtain the final result, which is of the form:

$$p = z_1 C_1 \exp(a_{11}t) + z_2 C_2 \exp(a_{22}t) + z_3 C_3 \exp(a_{33}t)$$

with p, C_1, C_2 and C_3 vectors. The parameters z_1 through z_3 are scalars that are determined by the boundary conditions.

Now that we know how to proceed in general, let us return to the initial problem concerning the chain of events shown in 4-10 on pg. 198. In that case, with $n = 4$ the transition matrix would have been:

$$\begin{bmatrix} -\mu_1 & 0 & 0 & 0 \\ \mu_1 & -\mu_2 & 0 & 0 \\ 0 & \mu_2 & -\mu_3 & 0 \\ 0 & 0 & \mu_3 & 0 \end{bmatrix}$$

Actually, we are not interested in the final observable 4th state, but only in the time spent in the pathway leading to it, so the system to solve reduces to:

$$A = \begin{bmatrix} -\mu_1 & 0 & 0 \\ \mu_1 & -\mu_2 & 0 \\ 0 & \mu_2 & -\mu_3 \end{bmatrix}$$

Because A is already triangular, no transformations are required to determine the eigenvalues, which obviously are $-\mu_1$, $-\mu_2$ and $-\mu_3$. Solving $(A - \lambda I)\, C$ as in Eq. F-5 results in:

$$C_1 = \begin{pmatrix} (\mu_3 - \mu_1)(\mu_2 - \mu_1) \\ \mu_1(\mu_3 - \mu_1) \\ \mu_1\mu_2 \end{pmatrix}$$

and

$$p = z_1 C_1 \exp(-\mu_1 t) + z_2 C_2 \exp(-\mu_2 t) + z_3 C_3 \exp(-\mu_3 t) \tag{F-6}$$

Initially, at $t = 0$, only the first state is occupied or:

$$p_{t=0} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = z_1 C_1 + z_2 C_2 + z_3 C_3$$

hence

$$\begin{aligned} &z_1(\mu_3 - \mu_1)(\mu_2 - \mu_1) = 1 \\ &z_1\mu_1(\mu_3 - \mu_1) + z_2(\mu_3 - \mu_2) = 0 \\ &z_1\mu_1\mu_2 + z_2\mu_2 + z_3 = 0 \end{aligned}$$

Solving the three equations gives:

$$z_1 = \frac{1}{(\mu_3 - \mu_1)(\mu_2 - \mu_1)}$$
$$z_2 = \frac{-\mu_1}{(\mu_3 - \mu_2)(\mu_2 - \mu_1)}$$
$$z_3 = \frac{\mu_1\mu_2}{(\mu_3 - \mu_2)(\mu_3 - \mu_1)}$$

Back substitution of this result into Eq. F-6 gives the time-dependent occupation for each of the three states:

$$p_1(t) = \exp(-\mu_1 t)$$
$$p_2(t) = \frac{\mu_1}{\mu_2 - \mu_1}\{\exp(-\mu_1 t) - \exp(-\mu_2 t)\}$$
$$p_3(t) = \frac{\mu_1\mu_2\{(\mu_3 - \mu_2)\exp(-\mu_1 t) - (\mu_3 - \mu_1)\exp(-\mu_2 t) + (\mu_2 - \mu_1)\exp(-\mu_3 t)\}}{(\mu_2 - \mu_1)(\mu_3 - \mu_2)(\mu_3 - \mu_1)}$$

Before exiting to state 4, the system remains in one of the three lower states with probability $p(t) = p_1(t) + p_2(t) + p_3(t)$. The waiting time distribution, W, is a probability density function and therefore $\mathrm{d}p(t)/\mathrm{d}t$ needs to be calculated. After summation and differentiation:

$$W(3) = \mu_1\mu_2\mu_3\left\{\frac{\exp(-\mu_1 t)}{(\mu_2 - \mu_1)(\mu_3 - \mu_1)} - \frac{\exp(-\mu_2 t)}{(\mu_2 - \mu_1)(\mu_3 - \mu_2)} + \frac{\exp(-\mu_3 t)}{(\mu_3 - \mu_2)(\mu_3 - \mu_1)}\right\} \qquad \text{(F-7)}$$

Figure F-1 shows an example for $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$.

One might hope that Eq. F-7 would reduce to Eq. 4-9 for $\mu_1 = \mu_2 = \mu_3 = \mu$ and $n = 3$. Clearly this is not the case, as the denominators of the quotients in Eq. F-7 become zero.

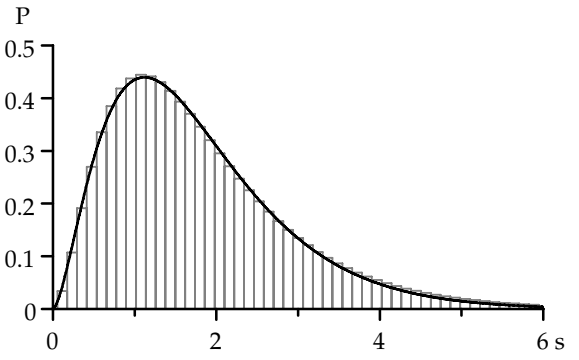


Fig. F-1 Waiting time distribution for four states with transition rate constants $\mu_1 = 1, \mu_2 = 2, \mu_3 = 3\,\mathrm{s}^{-1}$. The distribution was fitted with the Gamma distribution of Eq. 4-9, yielding $\mu = 1.535$ and $n = 2.724$. Despite the fact that the model, assuming $\mu_1 = \mu_2 = \mu_3$, is incorrect, the fit is rather good.

In fact, a problem occurs when trying to solve $(A - \lambda I)C = 0$ to obtain three (independent) eigenvectors:

$$A = \begin{bmatrix} -\mu & 0 & 0 \\ \mu & -\mu & 0 \\ 0 & \mu & -\mu \end{bmatrix}$$

Clearly, A has three repeated eigenvalues $-\mu$.

$$(A - \lambda I)\, C = \begin{bmatrix} 0 & 0 & 0 \\ \mu & 0 & 0 \\ 0 & \mu & 0 \end{bmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = 0$$

Apart from the the obvious solution $0 = 0$, this gives two equations:

$$\mu c_1 = 0$$
$$\mu c_2 = 0$$

and therefore $c_1 = c_2 = 0$ and c_3 may be chosen freely. Hence, one solution is:

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \cdot \mathrm{e}^{-\mu t} \tag{F-8}$$

and because that is the only one we will get by trying $\mathrm{e}^{-\mu t}$ as a solution, other solutions have to be looked for. Suppose that the solution $\mathbf{X}$ has the form:

$$\mathbf{X} = \mathbf{C}(t) \cdot \mathrm{e}^{-\mu t} \text{ with } \mathbf{C}(t) \text{ a vector with time-dependent elements: } \begin{pmatrix} c_1(t) \\ c_2(t) \\ c_3(t) \end{pmatrix}$$

Substitution into Eq. F-2 gives:

$$\begin{pmatrix} c_1(t) \\ c_2(t) \\ c_3(t) \end{pmatrix} \cdot -\mu\, \mathrm{e}^{-\mu t} + \begin{pmatrix} c_1'(t) \\ c_2'(t) \\ c_3'(t) \end{pmatrix} \cdot \mathrm{e}^{-\mu t} = \begin{bmatrix} -\mu & 0 & 0 \\ \mu & -\mu & 0 \\ 0 & \mu & -\mu \end{bmatrix} \cdot \begin{pmatrix} c_1(t) \\ c_2(t) \\ c_3(t) \end{pmatrix} \cdot \mathrm{e}^{-\mu t}$$

After elimination of $\mathrm{e}^{-\mu t}$ and rewriting, the result is:

$$\begin{pmatrix} c_1'(t) \\ c_2'(t) \\ c_3'(t) \end{pmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ \mu & 0 & 0 \\ 0 & \mu & 0 \end{bmatrix} \cdot \begin{pmatrix} c_1(t) \\ c_2(t) \\ c_3(t) \end{pmatrix}$$

and thus:

$$c_1'(t) = 0$$
$$c_2'(t) = \mu c_1(t)$$
$$c_1'(t) = \mu c_2(t)$$

After integration of each of the three equations:

$$c_1(t) = a \quad \text{and so} \quad c_2'(t) = \mu a$$
$$c_2(t) = \mu a t + b \quad \text{and so} \quad c_3'(t) = \mu^2 a t + \mu b$$
$$c_3(t) = 0.5\mu^2 a t^2 + \mu b t + c$$

where a, b and c are arbitrary (integration) constants. Writing out this result shows that the solutions for X are of the form:

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \cdot \mathrm{e}^{-\mu t} + \begin{pmatrix} 0 \\ \mu a \\ \mu b \end{pmatrix} \cdot t\,\mathrm{e}^{-\mu t} + \begin{pmatrix} 0 \\ 0 \\ a\mu^2/2 \end{pmatrix} \cdot t^2\,\mathrm{e}^{-\mu t}$$

Choosing $a = b = 0$ and $c = 1$ yields the eigenvector found in F-8. The two other solutions are found similarly by setting $a = c = 0$, $b = 1$ and by setting $a = 1$, $b = c = 0$. The result is a linear combination of the three solutions:

$$p = z_1 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \cdot \mathrm{e}^{-\mu t} + z_2 \left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \cdot t + \begin{pmatrix} 0 \\ 1/\mu \\ 0 \end{pmatrix} \right\} \cdot \mathrm{e}^{-\mu t} + z_3 \left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \cdot t^2 + \begin{pmatrix} 0 \\ 2/\mu \\ 0 \end{pmatrix} \cdot t + \begin{pmatrix} 2/\mu^2 \\ 0 \\ 0 \end{pmatrix} \right\} \cdot \mathrm{e}^{-\mu t}$$

At $t = 0$ only the first state is occupied, which yields $z_1 = z_2 = 0$ and $z_3 = \mu^2/2$. Before exiting to state 4, the system remains in one of the three lower states with probability $p(t) = p_1(t) + p_2(t) + p_3(t)$. The waiting time distribution, $W(3)$, is obtained as previously by differentiating p(t):

$$W(3) = \mu^3 t^2 \mathrm{e}^{-\mu t}/2$$

which is identical to Eq. 4-8 and 4-9 for $n = 3$.

This little exercise shows first the essential principles for the manipulation of Markov chains and second that Eq. 4-2 constitutes a special case of the Markov chain approach, indicating that the latter is the more general of the two. It may appear that the eigen decomposition as shown above is somewhat complicated and rather cumbersome to apply each time we wish to test or study a kinetic model. Actually, it is quite the contrary as the analysis that was carried out here by hand is always the same, no matter the model at hand. It can therefore be well performed by a computer routine, which makes the analysis almost as easy as a mouse click.

G: RECURSIVE (NON-CAUSAL) FILTERS

The next piece of C program code applies the low-pass filter of Part 4

$$y[n] = (1-a)\cdot x[n] + a\cdot y[n-1]$$

The data are passed twice through the filter; first with increasing index values, the second time from back to front, to obtain a result without phase shifts and to have a slope of 12 dB/octave.

```
void RCFilter(float *in, int np, float cutoff, float timebase)
{ // in[ ] contains the data to be filtered
// np is number of data points (=dimension of in[ ] )
// cutoff is cut-off frequency in Hz
// timebase is sample interval in seconds
short i;
float a,b,y;
  a=2-cos(2*pi*cutoff*timebase);
  if (a!=1)
  { a=a-sqrt(a*a-1);
    b=1-a;
    y=b*in[0];
    for (i=0;i<np;i++) in[i]=y=b*in[i]+a*y;
    y*=b;
    for (i=np-1;i>=0;i--) in[i]=y=b*in[i]+a*y;}
}
```

Users of the mathematical program Matlab® (www.mathworks.com) have the advantage of a plethora of digital filters built ready to use into its "signal processing toolbox".

As an allustration, here are the few lines of code necessary to perform the same filtering in Matlab:

```
% Matlab "filtfilt"
% non-causal low-pass
% demonstration
% ----------------------
t=0:np-1;
x=square(2.*pi.*t./64);
ma=[1 -.7]
mb=[.3]
y=filtfilt(mb,ma,x);
plot(t,y)
```

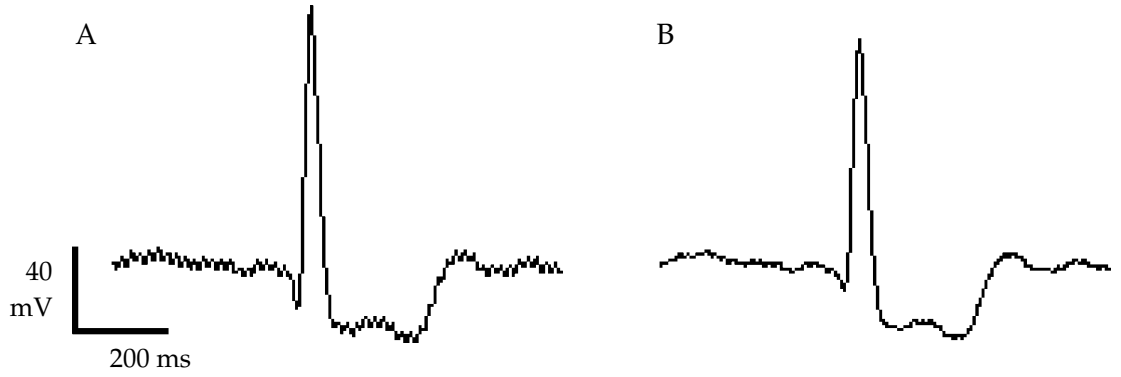


Fig. G-1 The ECG in A was filtered by the recursive low-pass filter at a roll-off frequency of 30 Hz (B).

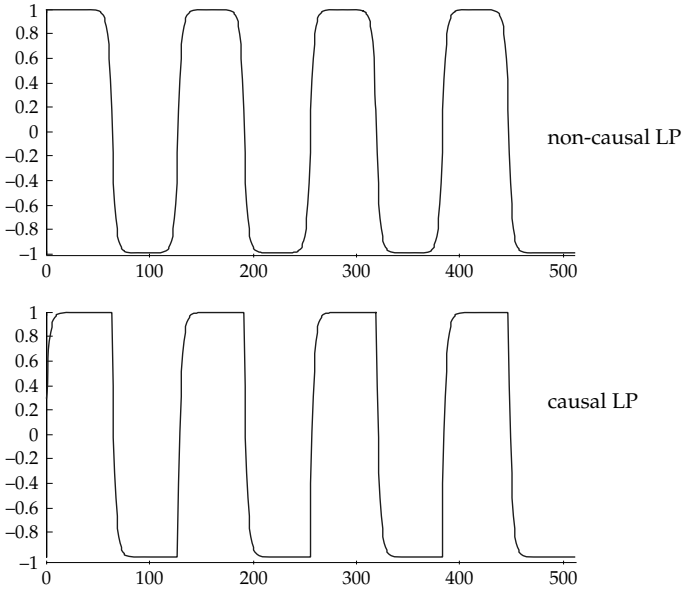


Fig. G-2 Response of a non-causal (top) and a causal (bottom) filter to a square wave.

The ouput of the last line is shown in Fig. G-2, top, and compared with the output of a causal filter (bottom):

H: PSEUDOCODE TO CALCULATE THE MACROSCOPIC CURRENT AND DWELL TIME DISTRIBUTIONS FROM A TRANSITION MATRIX

Routine FirstLatency(matrix A, array pnul, array S, int n, array evals, array amps)

```
{ # matrix A is the n*n transition matrix
  # array pnul of length n contains the probabilities to be in each state at t=0
  # array S indicates whether the state is closed (S[i]=0) or open (S[i]=1)
  # n is the number of states
  # The time constant of each exponential is returned in array evals
  # The amplitude of each exponential is returned in array amps
  # The routines Eigenvalues() and Invertmatrix() are standard routines
  call Routine FillDiagonal(A)
  # First modify matrix A to have transitions between closed states in the upper left corner
  call Routine SortMatrix(A, n, S, pnul)
  m=number of zeroes in S  # i.e. the number of closed states
  declare matrix evecs[m*m]
  # get the eigenvalues and eigenvectors of the upper left corner of A and return the result in
  # evals and evecs
  call Routine Eigenvalues(A, evals, evecs, m)
  # find scaling factors for eigenvectors
  declare array factors[m]
  call Routine Invertmatrix(evecs, pnul, m, factors)
  # Differentiate from infinity to 0, scale and sum the elements of the vectors
  # For symplicity, only isolated real roots are assumed. Differentiation of
  # repeated roots or complex roots is somewhat more elaborate.
  for (i=0 to i=m−1)
  { cum=0
    for (j=0 to j=m−1) cum=cum+evecs[j,i]
    amps[i]=-cum*factors[i]*evals[i]
  } # next i
} # end of routine
```

Routine Closedtimes(matrix A, array pnul, array S, int n, float t, array evals, array amps)

```
{ # matrix A is the n*n transition matrix
  # array pnul of length n contains the probabilities to be in each state at t=0
  # array S indicates whether the state is closed (S[i]=0) or open (S[i]=1)
  # n is the number of states
  # t is the integration time span
  # The time constant of each exponential is returned in array evals
  # The amplitude of each exponential is returned in array amps
  # The routines Eigenvalues() and Invertmatrix() are standard routines
  call Routine FillDiagonal(A)
  # First modify matrix A to have transitions between closed states in the upper left corner
  call Routine SortMatrix(A, n, S, pnul)
  declare matrix evecs[n*n]
```

```
  # get the eigenvalues and eigenvectors of the upper left corner of A and return the result in
  # evals and evecs
  call Routine Eigenvalues(A, evals, vecs, n)
  # find scaling factors for eigenvectors
  declare array factors[n]
  call Routine Invertmatrix(evecs, pnul, n, factors)
  m=number of zeroes in S  # i.e. the number of closed states
  # Integrate the open states over t seconds and store result in pnul
  # For symplicity, only isolated real roots are assumed. Integration of
  # repeated roots or complex roots is somewhat more elaborate.
  for (i=m to i=n−1)
  {  cum=0
     for (j=0 to j=n−1) cum=cum+factors[j]*evecs[i,j]*(1−exp(evals[j]*t))/evals[j]
     pnul[i]=cum
  }  # next i
  # get transition probabilities from open to closed states and store result in pnul
  for (i=0 to i=m−1)
  {  cum=0
     for (j=m to j=n−1) cum=cum+A[j,i]*pnul[j]
     pnul[i]=cum
  }  # next i
  call Routine FirstLatency(A, pnul, S, n, evals, amps)
}  # end of routine

Routine Opentimes(matrix A, array pnul, array S, int n, float t, array evals, array amps)

{  # The meaning of the input variables is as for routine Closedtimes()
   for (i=0 to i=n−1)
   {  if S[i]=1 then S[i]=0
      else S[i]=1
   }  # next i
   call Routine Closedtimes(A, pnul, S, n, t, evals, amps)
}  # end of routine

Routine Macrocurrent(matrix A, array pnul, array S, int n, array evals, array amps)

{  # matrix A is the n*n transition matrix
   # array pnul of length n contains the probabilities to be in each state at t=0
   # array S contains channel conductances for each state
   # n is the number of states
   # The time constant of each exponential is returned in array evals
   # The amplitude of each exponential is returned in array amps
   # The routines Eigenvalues() and Invertmatrix() are standard routines
   call Routine FillDiagonal(A)
   # Get the eigenvalues and eigenvectors of A and return the result in evals and evecs
   call Routine Eigenvalues(A, evals, evecs, n)
   # find scaling factors for eigenvectors
```

```
    declare array factors[n]
    call Routine Invertmatrix(evecs, pnul, n, factors)
    # scale vectors, sum their elements weighed by S
    for (i=0 to i=n−1)
    {  cum=0
       for (j=0 to j=n−1) cum=cum+evecs[j,i]*S[j]
       amps[i]=cum*factors[i]
    }  # next i
}  # end of routine

Routine FillDiagonal(matrix A, int n)

{  # matrix A is the n*n transition matrix
   # n is the number of states
   for (i=0 to i=n−1)
   {  cum=0
      for (j=0 to j=n−1)
      {  if i is not j then cum=cum+A[j,i]
      }  # next j
      A[i,i]=cum
   }   # next i
}  # end of routine

Routine SortMatrix(matrix A, int n, array S, array pnul)

{  # matrix A is an n*n matrix
   # pnul is an array of length n
   # array S indicates whether the state is closed (S[i]=0) or open (S[i]=1)
   s1=0,s2=0
   while s2<n
   {  while (s1<n) and (S[s1] is not 1) s1=s1+1
      s2=s1+1
      while (s2<n) and (S[s2] is not 1) s2=s2+1
      if (s1<n−1) and (s2<n)
      {  swap columns s1 and s2 of A
         swap rows s1 and s2 of A
         swap pnul[s1] and pnul[s2]
         swap S[s1] and S[s2]
      }  # end if
   }  # end while
}  # end of routine
```

I: REFERRED AND RECOMMENDED LITERATURE

Electricity and Electronics

Bach, M., Meigen, T. and Strasburger, H. (1997) Raster-scan cathode-ray tubes for vision research—limits of resolution in space, time and intensity. *Spatial Vision* 10, 403–414.

Brainard, D.H., Pelli, D.G. and Robson, T. (2002) Display characterization. In: J. Hornak (ed.) *Encyclopedia of Imaging Science and Technology* (pp. 172–188): Wiley, NY.

Horowitz, P. and Hill, W. (1989) *The Art of Electronics*. Cambridge University Press, Cambridge, etc.

Poynton, C.A. (1993) Gamma and its disguises. *J. Soc. Motion Picture Television Engineers* 102, 1099–1108.

Stanley, W.D. (1989) *Electronic Devices: Circuits & Applications*. Prentice-Hall, Englewood Cliffs, NJ.

Young, S. (1973) *Electronics in the Life Sciences*. Macmillan, London.

Electrochemistry

Bard, A.J. and Faulkner, L.R. (2001) *Electrochemical Methods: Fundamentals and Applications*. Wiley, New York.

Landolt, H. and Börnstein, R. (1923–1936) *Physikalisch-chemische Tabellen*. Springer, Berlin.

Parsons, R. (1959) *Handbook of Electrochemical Constants*. Butterworth, London.

Neurophysiology

Bullock, T.H. (1970) The reliability of neurons. *J. gen. Physiol.* 55, 565–584.

Bullock, T.H. (1976) Redundancy and noise in the nervous system: does the model based on unreliable neurons sell nature short? In: J.P. Reuben, D.P. Purpura and M.V.L. Bennett (eds) *Electrobiology of Nerve, Synapse and Muscle*. Rave Press, New York.

Hille, B. (1992) *Ionic Channels of Excitable Membranes*. Sinauer, Sunderland, MA.

Neher, E. and Stevens, C.F. (1977) Conductance fluctuations and ionic pores in membranes. *Annual Review of Biophysical Bioengineering* 6, 345–381.

Nicholls, J.G., Martin, R.A., Wallace, B.G. and Fuchs, P.A. (2001) *From Neuron to Brain*. Sinauer, Sunderland, MA.

Teunis, P.F.M., Bretschneider, F., Bedaux, J.J.M. and Peters, R.C. (1991) Synaptic noise in spike trains of normal and denervated electroreceptor organs. *Neuroscience* 41, 809–816.

Recording Methods

Berger, H. (1929) Über das Elektrenkephalogramm des menschen. *Arch. Psych. Nerverkr.* 87, 527–570.

Boulton, A.A., Baker, G.B. and Walz, W. (1995) *Patch-Clamp Applications and Protocols*. Humana Press, Totowa, NJ.

Cox, J.A. and Kulesza, J. (1984) Electrocatalytic oxidation and determination of arsenic (III) on a glassy carbon electrode modified with a thin film of mixed-valent ruthenium(III, II) cyanide. *Anal. Chem.* 56, 1021–1025.

DeFelice, L.J. (1997) *Electrical Properties of Cells: Patch Clamp for Biologists*. Plenum Press, New York.

Illinger, D. and Kuhry, J.-G. (1994) The kinetic aspects of intracellular fluorescence labelling with TMA-DPH support the maturation model for endocytosis in L929 cells. *J. Cell Biol.* 125, 783–794.

Levitan, E.S. and Kramer, R.H. (1990) Neuropeptide modulation of single calcium and potassium channels detected with a new patch clamp configuration. *Nature* 348, 545–547.

Lindau, M. and Neher, E. (1988) Patch-clamp techniques for time-resolved capacitance measurements in single cells. *PflügersArch.* 411, 137–146.

Machemer, H. (1987) *Uebungen zur Elektrophysiologie tierischer Zellen und Gewebe*. VCH Edition Medizin, Weinheim (D).
Neher, E. and Marty, A. (1982) Discrete changes of cell membrane capacitance observed under conditions of enhanced secretion in bovine adrenal chromaffin cells. *Proc. Natl. Acad. Sci. USA* 97, 6712–6716.
Neher, E. and Sakmann, B. (1995) *Single-Channel Recording*. Plenum Press, New York and London.
Standen, N.B., Gray, P.T.A. and Whitaker, M.J. (eds) (1987) *Microelectrode Techniques*. The Company of Biologists Limited, Cambridge.
Thomas, R.C. (1978) *Ion-Sensitive Intracellular Microelectrodes*. Academic Press, London, etc.
Wallis, D.I. (ed.) (1993) *Electrophysiology*. IRL Press at Oxford University Press, Oxford, UK.
Zhou, Z. and Misler, S. (1996) Amperometric detection of quantal secretion from patch-clamped rat pancreatic cells. *J. Biol. Chem.* 270, 270–277.

Signal Analysis

DeFelice, L.J. (1977) Fluctuation analysis in neurobiology. *Int Rev. Neurobiol.* 20, 169–208.
Dempster, J. (1993) *Computer Analysis of Electrophysiological Signals*. Academic Press, London.
Hamming, R.W. (1983) *Digital Filters* (second edition), ed. A.V. Oppenheim. Prentice-Hall, Englewood Cliffs, New Jersey (ISBN 0-13-212506-4).
Rieke, F., Warland, D., Ruyter van Steveninck, R. de and Bialek, W. (1997) *Spikes—Exploring the Neural Code*. MIT Press, Cambridge MA (ISBN 0-262-18174-6).
de Weille, J.R. (1983) Electrosensory information processing by lateral-line lobe neurons of catfish investigated by means of white noise cross-correlation. *Comp. Biochem. Physiol.* 74A, 677–680.
Wilson, H.R. (1999) *Spikes—Decisions and Actions*. Oxford University Press, Oxford UK (ISBN 0-19-852430-7).

Mathematics

Holden, A.V. (1976) *Models of the Stochastic Activity of Neurons*. Springer-Verlag, Berlin. (ISBN 3-540-07983-1).
MacKay, D.M. (1963) Psychophysics of perceived intensity: A theoretical basis for Fechner's and Stevens' laws. *Science* 139, 1213–1216.
Marmarelis, P.Z. and Marmarelis, V.Z. (1978) *Analysis of Physiological Systems, the White Noise Approach*. Plenum Press, New York (ISBN 0-306-31066-X).
Matlab: The Math Works, 3 Apple Hill Drive, Natick, MA 01760-2098; web address: www.mathworks.com.
Milsum, -J.H. (1975). *Biological Control Systems Analysis*. New York, McGraw-Hill. ISBN 0-07-042398-9.
Piguet, Y. (2005). *SysQuake: User Manual*. Calerga, Lausanne.
Rainville, E.D. and Bedient, P.E. (1981) *Elementary Differential Equations* (6th edition). Macmillan Publishing Co. Inc., New York (ISBN 0-02-397770-1).
Watkins, D.S. (2002) *Fundamentals of Matrix Computations* (2nd edition). John Wiley & Sons Inc., New York (ISBN 0-471-21394-2).

Index

This Page is Intentionally Left Blank