Sensors and Transducers

A guide for technicians

Ian R. Sinclair

Sensors and Transducers

A Guide for Technicians

Ian R. Sinclair

NEWNES

An imprint of Butterworth-Heinemann Ltd

Newnes
An imprint of Butterworth-Heinemann Ltd
Linacre House, Jordan Hill, Oxford OX2 8DP

PART OF REED INTERNATIONAL BOOKS

OXFORD LONDON BOSTON
MUNICH NEW DELHI SINGAPORE SYDNEY
TOKYO TORONTO WELLINGTON

First published by BSP Professional Books 1988
Reprinted by Butterworth-Heinemann Ltd 1991

British Library Cataloguing in Publication Data
Sinclair, Ian R. (Robertson), *1932* –
Sensors and transducers.
1. Electronic equipment. Transducers
I. Title
621.37'9

ISBN 0 7506 0361 5

Printed and bound in Great Britain by
Billing and Sons Ltd, Worcester

Preface

The purpose of this book is to explain and illustrate the use of sensors and transducers associated with electronic circuits. The steady spread of electronic circuits into all aspects of life, but particularly into all aspects of control technology, has greatly increased the importance of sensors which can detect, as electrical signals, changes in various physical quantities. In addition, the conversion by transducers of physical quantities into electronic signals and vice versa has become an important part of electronics.

Because of this, the range of possible sensors and transducers is by now very large, and most textbooks that are concerned with the interfaces between electronic circuits and other devices tend to deal only with a few types of sensors for specific purposes. In this book, you will find described a very large range of devices, some used industrially, some domestically, some employed in teaching to illustrate effects, some used only in research laboratories. The important point is that the reader will find reference to a very wide range of devices, much more than it would be possible to present in a more specialised text.

In addition, I have assumed that the physical principles of each sensor or transducer will not necessarily be familiar. To be useful, a book of this kind should be accessible to a wide range of users, and since the correct use of sensors and transducers often depends critically on an understanding of the physical principles involved, these principles have been explained in as much depth as is needed. I have made the reasonable assumption that electrical principles will not require to be explained in such depth as the principles of, for example, relative humidity. In order for the book to be as serviceable as possible to as many readers as possible, the use of mathematics has been avoided unless absolutely essential to the understanding of a device. I have taken here as my guide the remark by Lord Kelvin that if he needed to use mathematics to explain something it was probably because he didn't really understand it. The text should prove useful to anyone who encounters sensors and transducers, whether from the point of view of specification, design, servicing, or education.

I am most grateful to RS Components for much useful and well-organised information, and to Bernard Watson, of BSP Professional Books, for advice and encouragement.

Ian Sinclair
April 1988

Introduction

A sensor is a device that detects or measures a physical quantity, and in this book the types of sensors that we are concerned with are the types whose output is electrical. A transducer is a device which converts energy from one form into another, and here we are concerned only with the transducers in which one form of energy is electrical. The differences between sensors and transducers are often very slight. A sensor is performing a transducing action, and the transducer must necessarily sense some physical quantity. The shade of difference lies in the efficiency of energy conversion. The purpose of a sensor is to detect and measure, and whether its efficiency is 5% or 0.1% is almost immaterial, provided the figure is known. A transducer, by contrast, is intended to convert energy, and its efficiency is important, though in some cases it may not be high. Linearity of response, important for a sensor, may be of much less significance for a transducer. The basic principles that apply to one, however, must apply to the other, so that the descriptions that appear in this book will apply equally to sensors and to transducers.

The organisation of the book is in general by the physical quantity that is sensed or converted. This is not a perfect form of organisation – none is – because there are many 'one-off' devices that sense or convert for some unique purpose, and these have had to be gathered up in an 'assortment' chapter. Nevertheless, by grouping devices according to the sensed quantity, it is much easier for the reader to find the information that is needed, and that is the guiding principle for this book. In addition, some of the devices that are dealt with early in the book are those which form part of other sensing or transducing systems that appear later. This avoids having to repeat a description, or refer forward for a description.

Several points should be noted at this stage, to avoid much tedious repetition in the main body of the book. One is that a fair number of physical effects are sensed or measured, but have no requirement for transducers – we do not, for example, generate electricity from earthquake shocks. A second point is that the output from a sensor, including the output from electronic circuits connected to the sensor, needs to be proportional in some way to the effect that is being sensed, or at least to bear some simple mathematical relationship to the quantity. This means that if the output is to be used for measurements, then some form of calibration can be carried out. It also implies that the equation that connects the electrical output with the input that is being sensed contains various constants such as mass, length, resistance and so on. If any of these quantities is varied at any time, then recalibration of the equipment will be necessary.

Another point that we need to be clear about is the meaning of resolution as applied to a sensor. The resolution of a sensor measures its ability to detect a change in the sensed quantity, and is usually quoted in terms of the smallest change that can be detected. In some cases, resolution is virtually infinite, meaning that a small change in the sensed quantity will cause a small change in the electrical output, and these changes are detectable to the limits of our measuring capabilities. For other sensors, particularly when digital methods are used, there is a definite limit to the size of change that can be either detected or converted. It is important to note that very few sensing methods provide a digital output directly, and most digital outputs are obtained by converting from analogue quantities. This implies that the limits of resolution are determined by the analogue to digital conversion circuits rather than by the sensor itself. Where a choice of sensing methods exists, a method that causes a change of frequency of an oscillator is to be preferred, because frequency is a quantity that lends itself very easily to digital handling methods with no need for other analogue to digital conversion methods.

The sensing of any quantity is liable to error, and the errors can be static or dynamic. A *static error is the type of error that is caused by reading problems*, such as the parallax of a needle on a meter scale, which causes the apparent reading to vary according to the position of the observer's eye. Another error of this type is the interpolation error, which arises when a needle is positioned between two marks on a scale, and the user has to make a guess as to the amount signified by this position. The amount of an interpolation error is least when the scale is linear. One distinct advantage of digital readouts is that neither parallax nor interpolation errors exist. The other form of error is dynamic, and a typical error of this type is a difference between the quantity as it really is and the amount that is measured, caused by the loading of the measuring instrument itself. A familiar example of this is the false voltage reading measured across a high-resistance potential divider with a voltmeter whose input resistance is not high enough. All forms of sensors are liable to dynamic errors if they are used only for sensing, and to both dynamic and static errors if they are used for measurement.

Finally, two measurable quantities can be quoted in connection with any sensor or transducer. These are responsivity and detectivity, and though the names are not necessarily used by the manufacturer of any given device, the figures are normally quoted in one form or another. The responsivity is:

$$\frac{\text{output signal}}{\text{input signal}}$$

which will be a measure of transducing efficiency if the two signals are in comparable units (both in watts, for example), but which is normally expressed with very different units for the two signals. The detectivity is defined as:

$$\frac{\text{S/N of output signal}}{\text{size of input signal}}$$

where S/N has its usual electrical meaning of signal to noise ratio. This latter definition can be reworked as:

$$\frac{\text{responsivity}}{\text{output noise signal}}$$

if this makes it easier to measure.

Contents

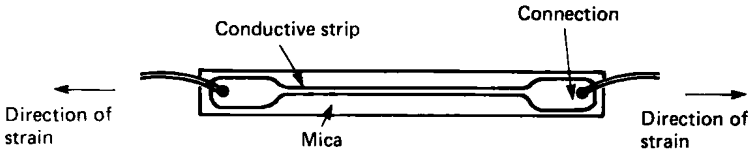
Chapter One
Strain and Pressure

Mechanical strain

The words stress and strain are often confused in everyday life, and a clear definition is essential at this point. Strain is the result of stress, and is a fractional change of the dimensions of an object. By fractional, I mean that the change of dimension is divided by the original dimension, so that in terms of length, for example, the strain is the change of length divided by the original length. This is a quantity which is a pure number, one length divided by another, having no physical dimensions. Strain can be defined for area or for volume in a similar way as change divided by original quantity. A stress, by contrast, is a force divided by an area. As applied to a wire or a bar in tension, for example, the tensile stress is the applied force divided by the area over which it is applied, the area of cross section of the wire or bar. For materials which can be compressed, the bulk stress is the force per unit area, which is identical to pressure applied, and the strain is the change of volume divided by the original volume. The most common strain transducers are for tensile mechanical strain.

Sensing tensile strain involves the measurement of very small changes of length of a sample. This is complicated by the effect of changes of temperature, which produce expansion or contraction that for the changes of 0°C to 30°C that we encounter in atmospheric temperature are of the same order of size as the changes caused by large amounts of stress. Any system for sensing and measuring strain must therefore be designed in such a way that temperature effects can be compensated for if necessary.

The commonest form of strain measurement uses resistive strain gauges. A resistive strain gauge consists of a conducting material in the form of a thin wire or strip which is attached firmly to the material in which strain is to be detected. This material might be the wall of a building, a turbine blade, part of a bridge, anything in which excessive stress could signal impending trouble. The fastening of the resistive material is usually by means of epoxy resins (such as 'Araldite'), since these materials are extremely strong and are electrical insulators. The strain gauge strip will then be connected as part of a resistance bridge circuit (Fig. 1.1). The effects of temperature can be minimised by using another identical unstrained strain gauge in the bridge as



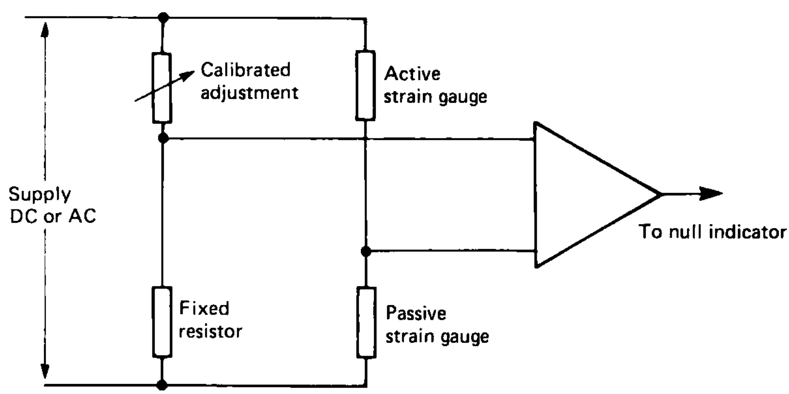


Fig. 1.1 Strain gauge use. (a) Physical form of a strain gauge. (b) A bridge circuit for strain gauge use. By using an active (strained) and a passive (unstrained) gauge in one arm of the bridge, temperature effects can be compensated if both gauges are identically affected by temperature. The two gauges are usually side by side, but with only one fastened to the cause of strain.

a comparison. This is necessary not only because the material under investigation will change dimensions as a result of temperature changes, but because the resistance of the strain gauge element itself will vary. By using two identical gauges, one unstrained, in the bridge circuit, these changes can be balanced against each other, leaving only the change that is due to stress. The sensitivity of this type of gauge, often called the piezoresistive gauge, is measured in terms of the gauge factor. This is defined as the fractional change of resistance divided by the change of strain, and is typically about 2 for a metal wire gauge and about 100 for a semiconductor type.

The change of resistance of a gauge constructed using conventional wire elements (typically thin Nichrome wire) will be very small, as the gauge factor figures above indicate. Since the resistance of a wire is proportional to its length, the fractional change of resistance will be equal to the fractional change of length, so that changes of less than 0.1% need to be detected. Since the resistance of the wire element is small, of the order of an ohm or less, the actual change of resistance is likely to be very small compared to the resistance of connections in the circuit, and this can make measurements very uncertain when small strains have to be measured. The use of a semiconductor strip in place of a metal wire makes measurement much easier, because the resistance of such a strip can be considerably greater, and so the changes in resistance can be correspondingly greater. Except for applications in which the temperature of the element is high (gas-turbine blades, for example), the semiconductor type of strain gauge is preferred. Fastening is as for the metal type, and the semiconductor material is surface passivated – protected from atmospheric contamination by a layer of oxidation on the surface. This latter point can be important, because if the atmosphere around the gauge element removes the oxide layer, then the readings of the gauge will be affected by chemical factors as well as by strain, and measurements will no longer be reliable.

Piezoelectric strain gauges are useful where the strain is of short duration, or rapidly changing in value. A piezoelectric material is a crystal whose ions move in an asymmetrical way when the crystal is strained, so that an EMF is generated between two faces of the crystal (Fig. 1.2). The EMF can be very large, of the order of several kV for a heavily strained crystal, so that the gauge can be sensitive, but the output impedance is very high and capacitive (Fig. 1.3). The output is not DC, therefore, so that this type of gauge is not useful

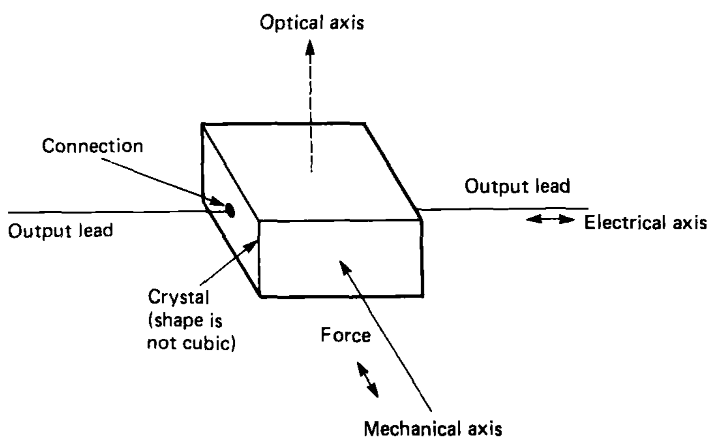


Fig. 1.2 Piezoelectric crystal principles. The crystal shape is not cubic, but the directions of the effects are most easily shown on a cube. The maximum electric effect is obtained across faces whose directions are at right angles to the faces on which the force is applied. The third axis is called the optical axis because light passing through the crystal in this direction will be most strongly affected by polarisation (see Chapter 3).

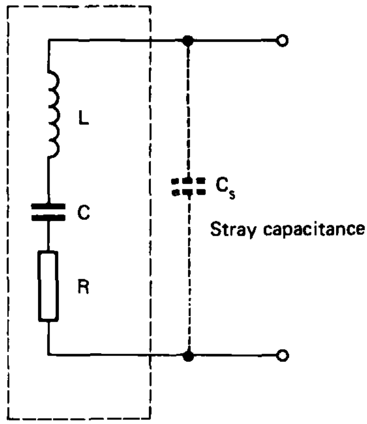


Fig. 1.3 The equivalent circuit of a crystal. This corresponds to a series resonant circuit with very high inductance, low capacitance and almost negligible resistance.

for detecting slow changes, and its main application is for acceleration sensing (see Chapter 2).

Two major problems of strain gauge elements of any type are hysteresis and creep. Hysteresis means that a graph of resistance change plotted against length change does not follow the same path of decreasing stress as for increasing stress (Fig. 1.4). Unless the gauge is over-stretched, this effect should be small, of the order of 0.025% of normal readings at the most. Over-stretching of a strain gauge will cause a large increase in hysteresis, and, if excessive, will cause the gauge to show a permanent change of length, making it useless. The other problem, creep, refers to a gradual change in the length of the gauge element which does not correspond to any change of strain in the

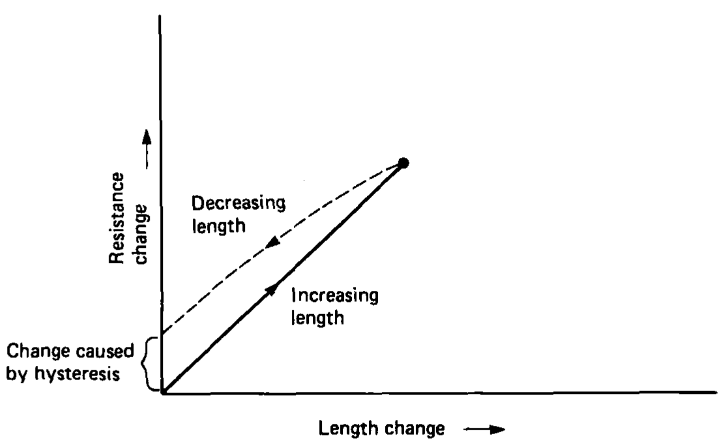


Fig. 1.4 The hysteresis effect on a strain gauge, greatly exaggerated. The graph is linear for increasing strain, but does not take the same path when the strain is decreasing. This results in the gauge having permanently changed resistance when the strain is removed.

material that is being measured. This also should be very small, of the order of 0.025% of normal readings. Both hysteresis and creep are non-linear effects which can never be eliminated but which can be reduced by careful choice of the strain gauge element material. Both hysteresis and creep increase noticeably as the operating temperature of the gauge is raised.

Interferometry

Laser interferometry is another method of strain measurement that presents considerable advantages, not least in sensitivity. Though the principles of the method are quite ancient, its practical use had to wait until suitable lasers and associated equipment had been developed, along with practicable electronic methods of reading the results. Before we can look at what is involved in a laser interferometer strain gauge, we need to understand the basis of wave interference and why it is so difficult to achieve with light.

All waves exhibit the effect that is called interference (Fig. 1.5). When two waves meet and are in phase (peaks of the same sign coinciding), then the result is a wave of greater amplitude, a reinforced wave. This is called constructive interference. If the waves are in opposite phase when they meet, then the sum of the two waves is zero, or a very small amplitude of wave, and this is destructive interference. The change from constructive to destructive

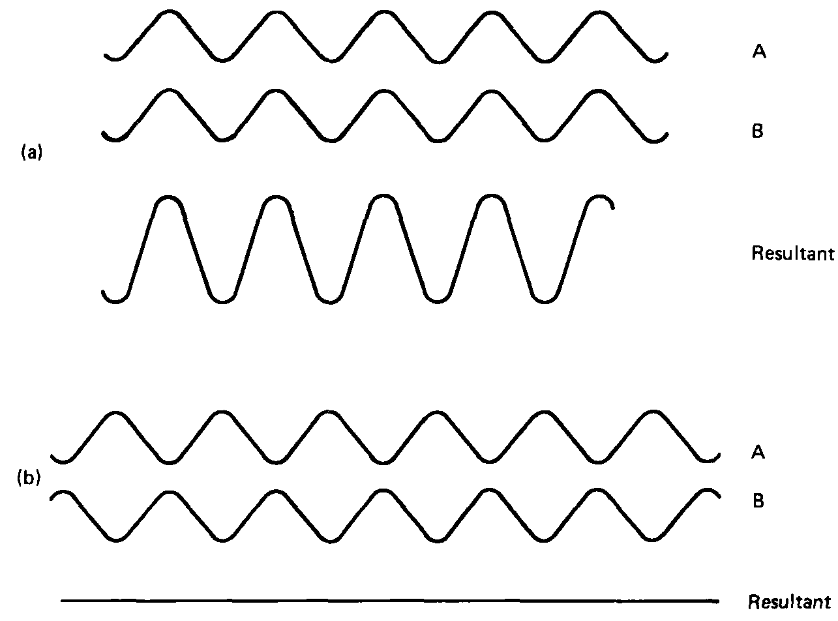


Fig. 1.5 Wave interference. When waves meet and are in phase (a), the amplitudes add so that the resultant wave has a larger amplitude. If the waves are in antiphase (b), then the resultant is zero, or of small amplitude.

interference therefore occurs for a change of phase of one wave relative to another of half a cycle. If the waves are emitted from two sources, then a movement of one source by a distance equal to half a wavelength will be enough to change the interference from constructive to destructive or vice versa. If the waves that are used have a short wavelength, then the distance of half a wavelength can be very short, making this an extremely sensitive measurement of change of distance.

The wavelength of red light is about 700 nm, that is 7^{-7} metres or 7^{-4} mm, so that a shift of half this distance between two sources of this light could be expected to cause the change between fully constructive and fully destructive interference – in practice we could detect a considerably smaller change than this maximum amount. The method would have been used much earlier if it were not for the problem of coherence. Interference is possible if the waves that are interfering are continuous over a sufficiently long period. Conventional light generators, however, do not emit waves continuously. In a light source such as a filament bulb or a fluorescent tube, each atom emits a pulse of light radiation, losing energy in the process, and then stops emitting until it has regained energy. The light is therefore the sum of all the pulses from the individual atoms, rather than a continuous wave. This makes it impossible to obtain any interference effects between two separate normal sources of light, and the only way that light interference can normally be demonstrated is by using light that has passed through a pinhole to interfere with its own reflection, with the light path difference very small.

The laser has completely changed all this. The laser gives a beam in which all the atoms that contribute light are oscillating in synchronisation; the type of light beam that we call coherent. Coherent light can exhibit interference effects very easily, and has a further advantage of being very easy to obtain in accurately parallel beams. The interferometer makes use of both of these properties as illustrated in Fig. 1.6. Light from a small laser is passed to a set of semi-reflecting glass plates and some of the light is reflected into a screen. The rest of the light is aimed at a reflector, so that the reflected beam will return to the glass plates and also be reflected to the screen. Now this creates an interference pattern between the light that has been reflected from the outward beam and the light that has been reflected from the returning beam. If the distant reflector moves by one quarter of a wavelength of light, the light path of the beam to and from the reflector will change by half a wavelength, and the interference will change between constructive and destructive. Since this is a light beam, this implies that the illumination on the screen will change between bright and dark. A photocell can measure this change, and by connecting the photocell through an amplifier to a digital counter, the number of quarter wavelengths of movement of the distant reflector can be measured electronically.

The interferometer, being very sensitive, is often too sensitive for many purposes. For example, the effect of changing temperatures is not easy to compensate for, though this can be done by using elaborate light paths in

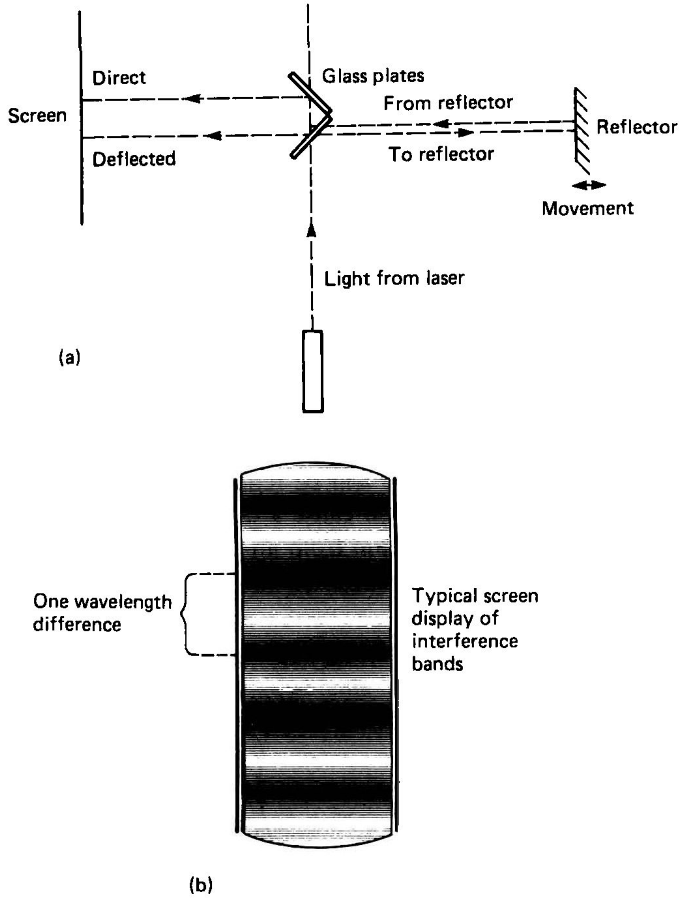


Fig 1.6 Principles of wave interferometry. The setup of laser and glass plates is shown in (a). The glass plates will pass some light and reflect some, so that both the reflector and the screen will receive some light from the laser beam. In addition, the light reflected from the reflector will also strike the screen, causing an interference pattern (b). For a movement of half of one wavelength of the reflector, the pattern will move a distance equal to the distance between bands on the screen.

which the two interfering beams have travelled equal distances, one in line with the stress and the other in a path at right angles. An advantage of this method is that no physical connection is made between the points whose distance is being measured; there is no wire or semiconductor strip joining the points, only the interferometer main body in one place and the reflector in another. The distance between the main part of the device and the reflector is not fixed, the only restraint being that the distance must not exceed the coherence distance for the laser. This is the average distance over which the light remains coherent, and is usually at least several metres for a laser source.

Fibre optic methods

Developments in the manufacture and use of optical fibres have led to these devices becoming used in the measurement of distance changes. The optic fibre (Fig. 1.7) is composed of glass layers whose refractive index is lower on the outer layer than on the inner. This has the effect of trapping a light beam inside the fibre because of the total internal reflection effect (Fig. 1.8). When a light ray is beamed straight down a fibre, the number of internal reflections will be small, but if the fibre is bent, then the number of reflections will be considerably increased, and this leads to an increase in the distance travelled by the light and hence to a change in the phase.

This change of phase can be used to detect small movements by using the type of arrangement shown diagramatically in Fig. 1.9. The two jaws will, as they move together, force the optical fibre to take up a corrugated shape in which the light beam in the fibre will be reflected many times. The extra distance travelled by the beam will cause a delay that can be detected by interferometry, using a second beam from an unchanged fibre. The sensor

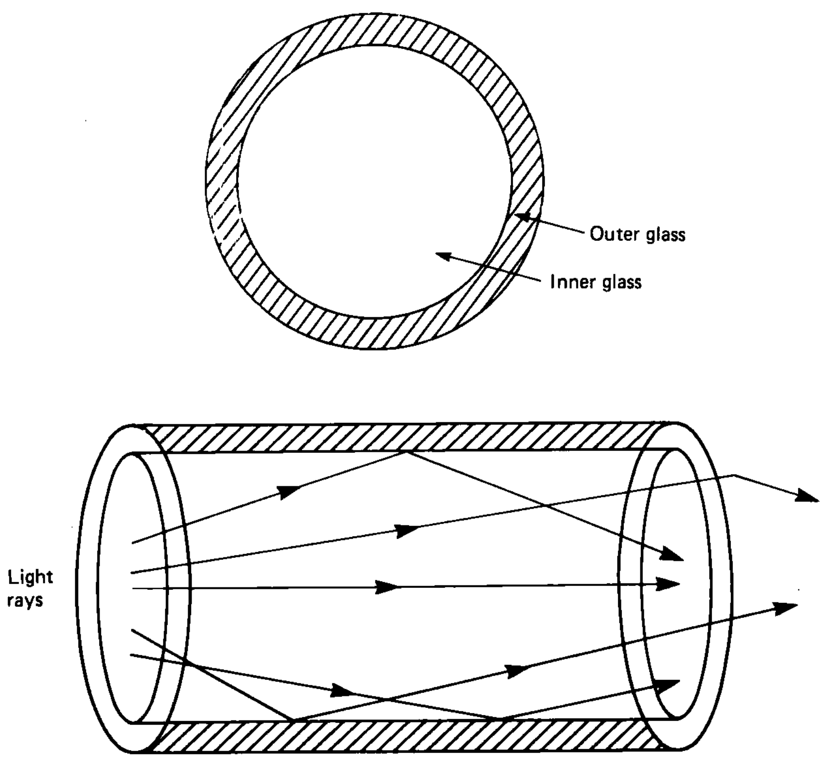


Fig. 1.7 **Optical fibre construction. The optical fibre is not a single material but a coaxial arrangement of transparent glass or (less usefully) plastics. The materials are different and refract light to different extents (refractivity) so that any light ray striking the junction between the materials is reflected back and so trapped inside the fibre.**

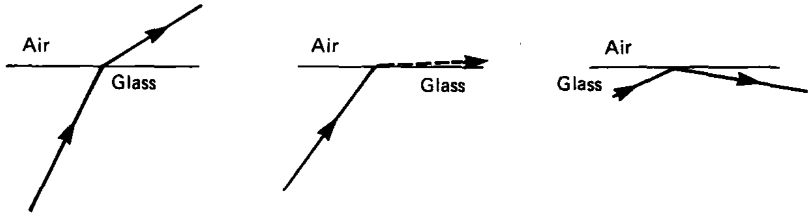


Fig. 1.8 Total internal reflection. When a ray of light passes from an optically dense (highly refractive) material into a less dense material, its path is refracted away from the original direction (a) and more in line with the surface. At some angle (b), the refracted beam will travel parallel to the surface, and at glancing angles (c), the beam is completely reflected. The use of two types of glass in an optical fibre ensures that the surface is always between the same two materials, and the outer glass is less refractive than the inner to ensure reflection.

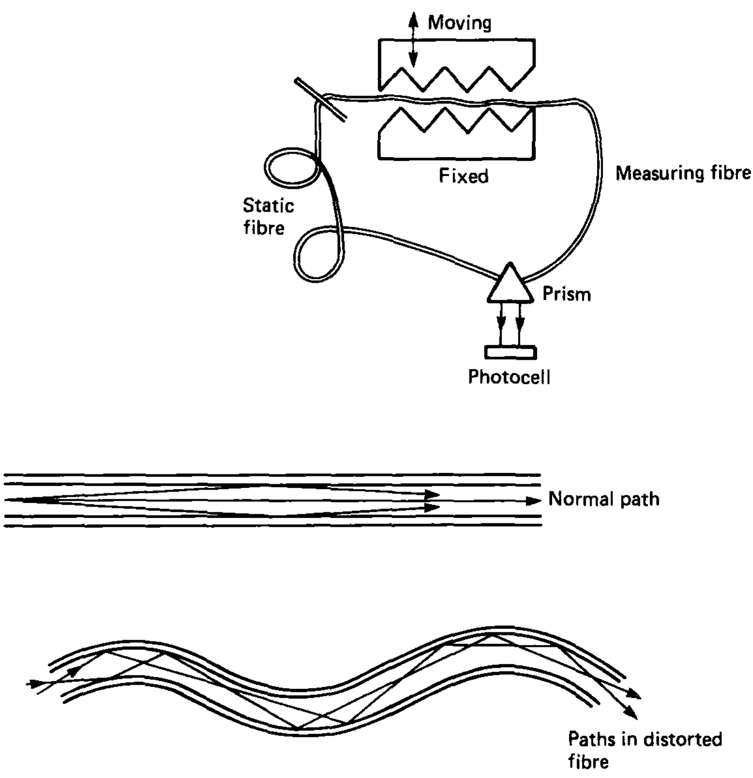


Fig. 1.9 Using optical fibres to detect small distance changes. The movement of the jaws distorts one fibre, forcing the light paths to take many more reflections and thus increasing the length of the total light path. An interference pattern can be obtained by comparing this to light from a fibre that is not distorted, and the movement of the pattern corresponds to the distortion of one fibre. The sensitivity is not so great as that of direct interferometry, and the use of fibres makes the method more generally useful, particularly in dark liquids or other surroundings where light beams could not normally penetrate.