AN ELECTRO-OPTIC INSHORE SURVEY AID

JOHN REGINALD GREENE

Submitted in fulfilment of the requirements for the degree of Master of Science in Engineering in the Department of Electrical Engineering at the University of Cape Town, April 1980.

ACKNOWLEDGEMENTS

Thanks are due to

Professor J.L.N. Besseling, for supervising this thesis, Plessey South Africa, for generously allowing me the time which made this thesis possible, and my wife, Biddy, for her constant help, encouragement and inspiration.

AN ELECTRO-OPTIC INSHORE SURVEY INSTRUMENT

ABSTRACT

ABSTRACT

The thesis is a proposal and a design study for an electronic instrument for inshore hydrographic survey operations. The operational requirements for such an instrument are stated and a detailed design specification evolved. A system study shows that for the economical realisation of the instrument certain parts are routinely within the state of the art, whereas others require more basic development work. In the former case solutions are proposed, and validated by experimental data in cases where insufficient information is available or the theoretical model is in doubt. In the latter case the need for a novel approach is justified by an extensive survey of earlier work and commercial precedent. Solutions are proposed, followed by detailed design and the presentation of experimental results. It is shown that the various subsystems together realise the required specification. Finally it is pointed out that some of the subsystems developed have potentially wider application, and possible directions for their further development are indicated.

The aim of the thesis is not to present a fully engineered design, but to identify a number of critical areas in which original work is required. Theoretical and experimental work is carried out in these areas, resulting in a number of experimentally validated solutions. The principal contribution of the thesis is the conception of a new integrated and economical measuring system based on phase-comparison methods, and in the proposal, development and testing of a novel absolute angle transducer as a crucial part of this measuring system.

CHAPTER 1
THE PROPOSAL

1.1 INTRODUCTION

In the course of a survey of the literature dealing with instrumental aids to maritime survey (1-14), as well as private discussion with a number of persons professionally involved in this area (15), it became evident that a hitherto unfilled need exists for a new type of electronic survey instrument, tailored to the requirements of short-range (or inshore) maritime survey. Such a need is strongly evidenced by frequent accounts of the novel use to which existing commercial equipment is put in harbours and estuary survey (3). This equipment, ranging from optical theodolites to sophisticated radionavigation systems is often highly unsatisfactory for short range maritime application. Theodolite-based systems, for example, are extremely awkward and tiring to use in conjunction with moving targets, and several highly skilled operators are constantly required. The radionavigation and radar systems are ill-matched to the requirements of short range and relatively high precision. They are, in addition, very expensive, often require skilled operation and are subject to occasional gross errors due to propagation anomalies. Various ad-hoc modifications have been attempted by a number of local distributors and end-users of such equipment to mitigate their defects with results that have been less than totally satisfactory. The case therefore seemed strong for the development of a new instrument based on a study of the specific requirement and drawing on the experience of electronic survey instrument development which has accumulated with particular rapidity over the past decade.

1.2 THE DETAILED REQUIREMENT

The instrument should be capable of determining the three-dimensional position of a reference point on a vessel relative to a co-ordinate system defined in relation to the shore. Reference to current marine survey practice suggests that, while metre precision is often adequate in respect of horizontal positioning (13), there are potential economic advantages in aiming at a significantly higher precision of vertical fixing (11), and the two aspects will be considered separately.

Positional information should be updated sufficiently rapidly to accommodate the vessel's motion without prejudicing accuracy, and should be available in the form of visual readout or printout either on shore or aboard the vessel or in the form of digital input to an on-line computing system.

The operational range over which the instrument should be capable of operating is determined by the size of typical estuaries and harbours and the desirability of covering reasonably large areas without relocating shore-based stations or reference points with the attendant difficulties of cross-referencing co-ordinate systems. There is inevitably a tradeoff between range and cost and it is likely that the former will also be dependent on local weather conditions. In the discussions cited a consensus seemed to emerge that, in order for the instrument to be really useful, it should have a range of at least 1km. Preferably, it should be possible to determine the position of a vessel over an area of approximately 2km x 2km, at least in conditions of good visibility. In conditions of reduced visibility reduction in operational range is acceptable, since the kind of survey operations envisaged are not likely to be carried out "blind". It was decided to settle for a 2km operational range when the visual range exceeds 4km.

Horizontal accuracy

Considering the size of the object being positioned (i.e. the vessel) and the precision involved in normal maritime survey practice, a standard deviation of 0,5m for a single determination of distance anywhere in the operating area seems a reasonable specification. Thus the major axis of the horizontal error ellipse for all positions within the operating area would not exceed one metre.

In specifying the accuracy of electronic survey instruments the danger of the uncritical use of statistical measures of precision and accuracy, such as standard deviation, is often overlooked. Their use involves certain implicit assumptions concerning the nature and distribution of errors,

which are not always valid. For example, errors in electronic survey instruments are part systematic and part random, and the systematic component may be part periodic or cyclic and part proportional. Moreover the requirement of the central limit theorem that errors be uncorrelated and of comparable magnitude may not apply. Error mechanisms are considered in detail in subsequent chapters. For the purpose of setting an initial deviation, with an additional constraint that in a large and representative population of measurements, fewer than 1% will be in error by more than three times the standard deviation. This also has the desirable effect of providing a reasonable guarantee of system integrity.

Vertical accuracy

In addition to determining the position of the vessel as described above it is desirable that a somewhat more precise readout be available of the instantaneous height of some reference point on the vessel relative to a fixed base datum. This is particularly important for dredger control. Discussions with persons involved in this work and a survey of the precision attained by competing approaches (such as a combination of tide gauges and vertical accelerometers) suggests that a standard deviation of 200mm is acceptable. Again the peak worst case error should not exceed three times this value at the 99% confidence level.

The overall effect of a single measurement with the system will be to fix the position of a point on the vessel within an error ellipsoid bounded by a horizontal circle of 1m diameter and having a polar axis of 0,4m.

It should be borne in mind that it is difficult to obtain a definitive statement of requirements for a proposed new instrument from its potential users who, lacking an adequate appreciation of the constraints and possibilities of the technology involved, are frequently unable to evaluate the complex performance/cost tradeoff relationship. The prospective user therefore quite understandably tends either to set an unrealistically ideal specification or alternatively, tends to anticipate the problems of the instrument

designer and modify the stated requirements accordingly. Neither approach is very helpful to the designer who, seeking a consensus, is frequently faced with enormous diversity and conflicting requirements.

In such a situation the only reasonable strategy seems to be for the designer to study the literature, discuss the problem with as many informed persons as possible, spend time observing the kind of activities to which this instrument will be applied, and in the end himself shoulder the responsibility of setting a specification, conscious that the marketplace will be the ultimate arbiter of the correctness of his choices. Such has been the strategy adopted in generating the specifications for the present instrument.

1.3 POTENTIAL APPLICATIONS AND ECONOMIC CONSIDERATIONS

It is frequently found that a new instrument which fills a real gap in measurement technique generates a wide range of unforeseen applications. Hence no prior listing of desired characteristics is likely to be exhaustive, and a large part of the art of instrument design is to incorporate sufficient versatility and excess performance capability to cope with these extensions as they arise, without prejudicing the primary functions for which the instrument is intended by the addition of unnecessary cost and complexity.

The main areas of application envisaged are the control of dredging operations in harbours and river estuaries, harbour survey and depth contouring and the provision of survey control for such tasks as harbour construction and maintenance, and the laying of underwater pipelines. Of these, dredging control is expected to be a major area. Very many harbours, especially those situated on or near estuaries, require virtually continuous dredging on account of silting, and dredging is an expensive operation. The costs can be greatly decreased by adequately precise position controls.

Two factors to be considered are capital and running costs. In the case of the microwave systems often used for such work, equipment cost may be several tens of thousands of rands, which places such equipment outside the reach of many small harbours and dredging contractors. Moreover such systems are not particularly easy to use and do not provide sufficient accuracy for short range work, especially in the vicinity of large conducting masses such as ships and cranes.

At the other end of the scale, positioning systems using shore-based theodolites are inexpensive in terms of capital outlay and provide adequate short range accuracy. However the methods are slow and inconvenient, and extremely expensive in respect of running costs, on account of the long-term commitment of highly-skilled operators.

This thesis will propose an instrument costing, in small-to-medium scale commercial manufacture, no more than three thousand rand and requiring only a single semi-skilled operator.

1.4 SUMMARY OF SYSTEM REQUIREMENTS

The proposal is for a low-cost instrument which can be used by a single semi-skilled operator to determine the position of a vessel in three dimensions.

In clear weather conditions the instrument should have an operational range enabling it to cover an operating area of about 2km x 2km.

The standard deviation of a single position-fix anywhere within this area should be no greater than 0,5m with a worst-case peak-error limit of 1,5m (99% confidence level).

The standard deviation of horizontal positioning relative to a horizontal projection from a shore-based datum should be no greater than 0,2m with less than 1% probability that the peak error will exceed 0,6m. This should apply over the whole operational range.

The following chapter will take a general look at the techniques available for realising this performance. Compatibility of the various possible subsystems will be considered, and a target specification for each of them evolved. Subsequent chapters will attend to the specific problems and the detailed design of each of the subsystems.

CHAPTER 2.
SYSTEM STUDY

2.1 POSSIBLE APPROACHES

It is possible at the outset to exclude some of the multitude of possibilities that suggest themselves. It can be seen from a survey of existing commercial equipment that many decades of development have failed to produce any radio-wave or sonar based instruments that meet the accuracy requirements laid down in the target specification. Nor is it clear how vertical position could be determined by such instruments.

On the other hand, experience with electro-optic instruments for land surveying applications suggests that the required range and accuracy could readily be achieved. Since the mid 1960's the availability of a small, moderately priced source of modulated light or infra-red radiation in the form of the gallium arsenide light-emitting diode (LED) has resulted in a wide range of compact lightweight distance measuring instruments. All of them, however, have been tailored to the requirements of cadastral land survey and are ill-suited to hydrographic use, although their occasional application in this connection has been reported. In contradistinction to the radio-wave instruments, it is evident that many possibilities exist whereby electro-optic instruments could be used to obtain information on vertical displacement.

There is a direct precedent for this in a number of commercial devices intended for land-based use. The Hewlett-Packard 3810 (16) for example contains a vertical tilt sensor which gives it a capability of crude trigonometric levelling, and a number of laser-based horizontal datum systems exist, of which the AGA 300 Geoplane (17) is but one example. More recently, a fully automatic system has become available in the Hewlett-Packard "total station". A similar system has been introduced by Kern.

In the light of these considerations, it was felt that an electro-optic based system was the one which offered the best chance of success, despite certain disadvantages and limitations, which are discussed below.

2.2 IMPLICATIONS OF THE CHOICE OF AN ELECTRO-OPTIC SYSTEM.

Beamwidth

In order to achieve a given range with an electro-optic link, it is necessary to achieve a certain concentration of power in the beam. This parameter is known as beam irradiance, and the relations governing range and beam irradiance are fully worked out in Chapter 3. Beam irradiance is the ratio of radiated power to beamwidth and it will be seen that in order to conserve power to meet the requirement of a small portable power supply, it is necessary to restrict the beamwidth as much as is possible without prejudicing operator convenience. Experience and commercial precedent suggest that a beamwidth of about $0,1^{o}$ - $0,5^{o}$ is acceptable and this is compatible with the effective source size of commercial gallium arsenide diodes and, it will be seen, with a reasonable size optical system.

Vessel or shore based instrument

Fractional degree beamwidth has the important implication of ruling out a vessel-based instrument due to roll and pitch motion. In principle the instrument could be gyro-stabilised or compensated by means of a damped pendulum but a survey of systems in which this approach has been attempted in closely analogous cases suggests that such a solution is impractical on grounds of cost and complexity.

In fact it seems there are no serious disadvantages, and some possible operational advantages, in a shore-based station, which enables the operator to perform his fairly exacting task in comfort in an uncluttered and stable environment. The fact that information is not available aboard the vessel concerning its position is of relatively minor significance since a simple telemetry link (electro-optic or radio) capable of operating over 2km will add little to the cost and complexity of the overall system.

Weather dependence

A further implication of the choice of an electro-optic system is a strong dependence on atmospheric conditions. However, while a navigation aid must be capable of all-weather operation, inshore hydrographic survey, for a variety of

reasons, is not carried out in conditions of fog. The subject of atmospheric effects is considered in detail in Chapter 3, and it will be found that the target specification in this regard can be met without undue difficulty provided we accept the reasonable restriction that full range operation will be required only in reasonably clear conditions (visual range of 4km or greater, corresponding to the absence of heavy haze or fog).

2.3 POSITION FIXING TECHNIQUES

Since the vessel will be free to move over an area of several square kilometres horizontally and in the vertical direction by only a few metres due to tidal and wave action, it is convenient to consider the two aspects of vertical and horizontal position fixing separately.

Approaches to determining two dimensional positions

Consider a point P constrained to move on a plane OXY. The position of P is classically represented in either rectangular or polar co-ordinates as either $(X ; Y)$ or $(R ; \theta)$.

There exists in general no simple way of directly measuring the orthogonal distances of the point P from the axes OX and OY. It is however possible, using appropriate instrumentation, to measure the distance R from a fixed point, and its azimuthal angle with respect to a fixed reference direction. We may then consider the fixed point to be the origin O, and the reference azimuthal direction as OX (or OY) and compute the rectangular co-ordinates.

(a) Direct approach - R, θ

This obvious approach to position fixing is frequently used in township layout work and other aspects of cadastral land surveying, the distance measurement being carried out by an electro-optic distance measuring instrument and the azimuthal measurement being performed by theodolite. There have even been sporadic reports of this technique having been adapted directly for inshore hydrographic survey, but instruments designed for landbased use have a number of disadvantages in the latter context:

i. The range of the electro-optic instrument often used in such a setup is rather limited, 1km being a maximum for reliable performance with a practical reflector array on the vessel. Recent laser-based instruments, such as the Hewlett Packard 3808A (18) overcome this problem, offering about 5km range at the expense of drastically reduced beamwidth. Such an instrument is far from ideal for maritime use and its cost is excessive for many users as a part of the overall system.

ii. The resolution of the instrument is unnecessarily great. This results in a number of trade-offs leading to high cost, inconvenient operation, slow speed and limited dynamic tracking capability.

iii. The theodolite is an awkward and inconvenient instrument for use with a moving target, requiring a high degree of both skill and sustained concentration on the part of the operator. The instrument has to be kept trained on the target and each reading has to be visually read and manually logged. The sextant is much more convenient in use but is by no means sufficiently accurate for submetre accuracy at several kilometres range.

The principal advantage of position fixing using both range and azimuth information is that determination of the vessel's position can be carried out by means of a single instrument occupying a single station and requiring only one operator. All of the other shore-based approaches require two or three shore stations with more or less severe constraints on their mutual positioning.

(b) Indirect systems

To appreciate the advantages of the range-azimuth approach it is necessary to review briefly some of the conventional alternative approaches to position fixing:

Dual-range systems

The vessel's position is computed from its measured distance from two stations situated at the ends of a known baseline. It can be seen that the requirements in respect of baseline length to achieve overall accuracy comparable to the individual accuracy of range measurement is quite modest. On the other hand both base stations need to be manned and range information from both of them is required before position can be computed.

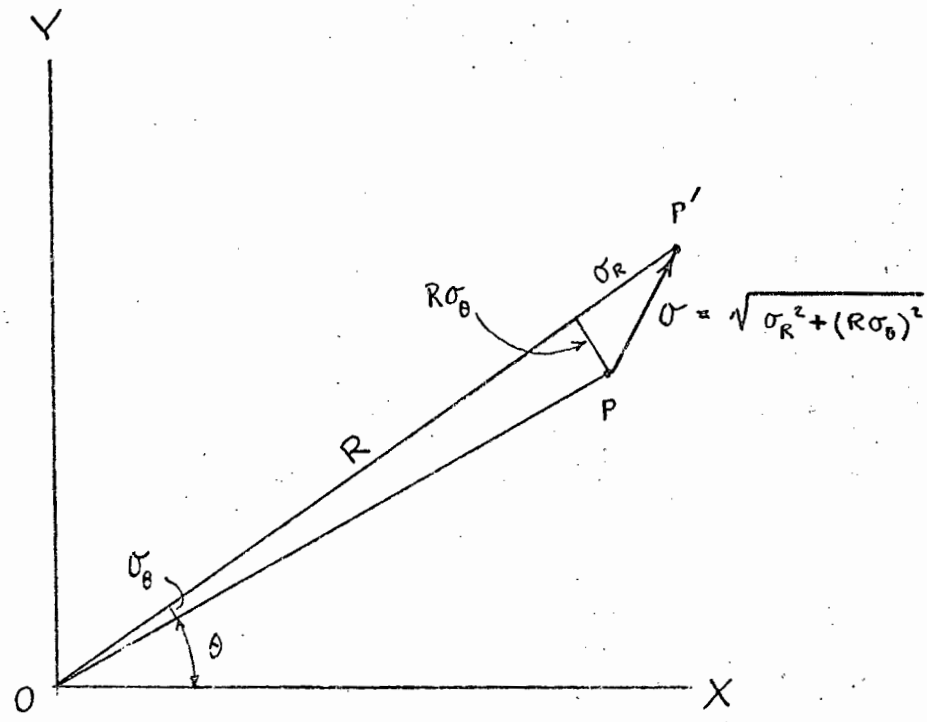
Hyperbolic systems

If time-synchronous signals are transmitted from two base stations of known location, the difference in time-of-arrival measured aboard the vessel positions the latter with respect to a set of confocal hyperbolic lines having the base stations as foci. To complete the position fix it is necessary to establish position with respect to another set of confocal hyperbolae, and for this purpose a third base station is required. The location of the three base stations relative to the operational area is subject to severe constraints if the accuracy sought is to be comparable to that of the time-difference measurement, due to a phenomenon called "lane expansion". This is a consequence of the growing separation between individual hyperbolae as distance from the baseline increases, and the increasingly oblique angle-of-cut between the families of hyperbolae.

Hyperbolic systems are convenient in that they are inherently multi-user, and only reception of signals is required aboard the vessel, but the limited accuracy and base station positional constraints rule out this approach in many cases. In the case of an electro-optic system with high directivity, the three stations would have to be manned, which would lead to excessively high running costs. In a harbour environment moreover, one will frequently not have sufficient freedom to position the base stations satisfactorily. Whatever the merits of hyperbolic systems as radio-wave navigation aids, these considerations appear to exclude this approach in the present case.

Azimuthal systems

Systems where position is determined by simultaneous angular observations from two fixed base stations have many characteristics in common with hyperbolic systems - in fact a hyperbolic system with widely separated very short baselines degenerates to the azimuth. A version of this system is classically used in hydrographic inshore work, the angles being observed by theodolite or sextant. Again, two stations are required, with intercommunication, and for good accuracy a long baseline is necessary.



COMBINED EFFECT OF ANGLE + DISTANCE ERROR

It can be seen that the indirect approaches offer no clear advantages and a number of decided disadvantages. Accordingly it was decided to adopt the direct range-angle approach. It remains to choose the techniques whereby these two independent parameters are measured.

2.4 INSTRUMENTAL PRECISION REQUIRED

Horizontal position

Let the r.m.s. error of position σ_t (i.e. half the major axis of the error ellipse) at the extreme range be 0,5m. Let the angular error be σ_θ radians, the distance error be σ_R and the maximum distance be R. Since the partial errors are independent we have:

$$\sigma = \left\{\sigma_R^2 + (R\sigma_\theta)^2\right\}^{\frac{1}{2}}$$

Now we require $\sigma = 0,5\text{m}$

when $R = 3 \times 10^3\text{m}$

Dividing the allowable error equally between range and angle,

$$\sigma_R = R\sigma_\theta = \frac{0,5}{\sqrt{2}}$$

$$\sigma_\theta = \frac{0,35}{3000} \text{ radian}$$

$$\doteq 24 \text{ arc seconds}$$

Vertical position

Let us assume that vertical positioning is to be achieved with a standard deviation of 200mm at the maximum operational range of 2 km. The standard deviation of vertical angle determination is then

$$\sigma_\theta = \frac{200 \times 10^{-3}}{2 \times 10^{-3}} \text{ radians or } 20 \text{ arc seconds}$$

Clearly when aiming for precision of this order it is necessary to make some assumptions regarding the motion of of the vessel. A small survey vessel may conceivably have a vertical component of motion of a substantial fraction of a metre due to wave-motion, even in a harbour. At a kilometre range this represents angular motion approaching a milliradian, or 3 minutes of arc, which is about nine standard deviations.

In such a situation the operator can average repeated determinations or, better, visually bisect the extremes of travel of the aiming point. Where ultimate accuracy is required, possible solutions would be to use a larger vessel, supplement the vertical angle determination with a recording vertical accelerometer aboard the vessel, or wait for calmer weather.

Similar considerations apply in the case of horizontal motion. If a horizontal position fix can be obtained once per second, a precision of 20 arc seconds at 1km distance requires that the vessel should not have moved laterally more than 0,1m. This would impose a maximum speed of 3km/hour or $\sim 1\frac{1}{2}$ knot. Since the vessel's horizontal motion is highly predictable, the operator can aim the instrument slightly ahead of the vessel and operate a "read" button as it crosses the telescope reticle.

We shall therefore set allowable limits on the errors in range and angle determination respectively:

$$\sigma_R = 0,3 \text{ metre}$$

$$\sigma_\theta = 20 \text{ arc seconds}$$

with less than 1% error probability at the 3σ level.

2.5 RANGE MEASURING TECHNIQUES

Since we have already settled on an electro-optic approach to range measurement, the choice remaining is between a pulse or continuous-phase mode of measurement:

(a) Pulse ranging: The atmospheric velocity of light is approximately 3×10^8m/sec. To determine distance to an accuracy of 0,3 metres we would need to determine the transit time of a pulse of light over the return path to an accuracy of:

$$\frac{0,3 \times 2}{3 \times 10^8} \text{ seconds}$$

$$= 2 \times 10^{-9} \text{ seconds}$$

To determine the transit time of a pulse of light with a precision of 2 nanoseconds is marginally within the state of the art. It is doubtful however whether it could be achieved

under the conditions of fading and scintillation encountered in an atmospheric optical channel, and it is reasonably certain that it would not be accomplished in a low cost, portable instrument.

(b) C.W. ranging: A simpler and better alternative exists in phase-comparison distance measurement, in which the beam of light is modulated at a relatively high frequency and the transit time inferred from the phase delay between the modulation envelopes of the outgoing and returning radiation. This approach will be discussed in detail in Chapter 4. It will be seen that the required precision presents no great difficulty.

2.6 ANGLE MEASURING TECHNIQUE

Electro-optic systems capable of automatically tracking the angular position of aircraft and missiles with 20 second accuracy have been reported (19,27). Typically they employ laser illumination and image-dissector tubes with follow-up servos. In this case such an approach is judged prohibitively expensive and unnecessarily complex. For a slow-moving target such as a ship the most straightforward approach seems to be a telescope system, with cross-hairs projected to infinity which the operator keeps trained on the vessel whose position is to be determined. This is mounted on a digital angle encoder, the two together constituting the azimuth system of a digital theodolite. For a precision of 20 arc seconds the telescope system presents few problems and any commercial theodolite telescope can be used. The problem of the digital angle transducer, however, is not so easily solved, as will be seen in Chapter 5. It will be seen that a new angle transducer was developed which easily meets the required performance specification and which moreover has the advantage of an intermediate output in terms of the phase angle between two low frequency electrical signals. As a result, an elegant and economical instrument is possible in which a great deal of the critical signal processing circuitry can be shared by the range-and angle-measuring subsystems. A further simplification occurs if the angle-measuring systems for azimuth and deviation determination are similar or identical. This possibility is discussed in the following section.

SUMMARIZED SPECIFICATION FOR

ANGLE TRANSDUCER

Range:	0 - 360° unambiguous
Accuracy:	10 arc seconds (rms)
Resolution:	2 arc seconds
Time for reading:	0,3 sec
Output:	phase displacement proportional to angle
General comment:	The design of the transducer should aim at minimising cost, bulk and power consumption. It should if possible improve on existing commercially available units in respect of cost and size.

Another aspect which must be considered is the requirement in respect of range of possible vertical angles. The surface of a body of water over an area of 2km square is of course very nearly a plane and in a harbour or estuary we would not usually expect wave or tidal motion in excess of a few metres. There is an advantage however in providing reasonable excess vertical angle range, in that this gives greater freedom of positioning the instrument in what is often a cluttered harbour environment. An elevated shore station guarantees a clear view of the operating area. If the height of the instrument above sea level is, say 10 metres, and the minimum range is 100m, then a tilt range of arc of -0,1rad or or -6^{o} is adequate. By the same token, if the instrument is at sea level and the aiming point on the vessel at mast-top height (which would be most undesirable because of the amplification of rolling and pitching motion) we might encounter a positive slope of a few degrees. Certainly a tilt range of $\pm 10^{o}$ seems to cater for all contingencies, but in the interests of flexibility of application it might be desirable to extend this to say $\pm 30^{o}$, as it appears that this can easily be achieved.

2.7 VERTICAL ANGLE REFERENCE

Since the telescope cross-hairs are kept trained on a specific reference point on the vessel the angle of elevation of the latter can be inferred from the angular tilt of the telescope relative to some datum. This can be measured by a transducer system. The only difference is that whereas the reference datum for azimuth angle is arbitrarily defined by the operator at the beginning of a measuring sequence, that for the vertical angle is invariably that of the local gravitational gradient. In a manual survey instrument such as a theodolite or simple level, an ordinary sensitive bubble may be used and it is the operator's task to ensure that the instrument is kept levelled. Most theodolites of recent design, however, incorporate some form of mechanically operated automatic compensation device which is capable, over a certain limited range (say 10 minutes of arc), of coping with instrument dislevelment.

The proposed instrument will be used continuously for protracted periods, and will thus be prone to inadvertent disturbance of level due to tripod subsidence. Moreover it is not considered good survey practice to re-adjust the level of the instrument during the course of an observation sequence. These factors, together with the requirement of unskilled or semiskilled operation, indicate the need for some form of automatic level compensation, or inbuilt vertical reference.

Accuracy: The calculated required accuracy for the horizontal angle determination has been shown to be about 24 arc seconds (section 2.4) and that of the vertical angle 20 arc seconds. The telescope/reticle system is of course common to both and there would seem to be a priori a case, in the interests of standardisation, simplicity and economy, for using an identical transducer system if possible to measure the displacement of the telescope in azimuth and elevation. An error in vertical reference is added directly to the output of the elevation transducer and, in view of the parity of precision required of the two systems, it is desirable that the error contributed by the vertical reference sensor be very small. Bearing in mind the single-second precision achieved by mechanical theodolite compensation systems it seems reasonable to set a target specification on the vertical reference system of ± 5 arc secs.

Range: The linear range of the vertical reference sensor (or the range of dislevelment over which the compensation system must operate) is governed by the precision with which the instrument can readily be set level by a semiskilled operator and the maximum expected tilt due to subsidence during an observation period of, perhaps, several hours.

Regarding the former requirement, it is a rapid operation requiring little skill to level an instrument to within about 30 arc seconds using a circular bubble. Concerning tripod stability, some simple tests and discussions with surveyors suggest that 20 arc seconds tilt is a reasonable maximum. Therefore assuming the effects are additive, we require 50 arc seconds range for the level sensor and 2% accuracy over this range. Increased linear range would of course be advantageous.

TABLE 1.

SUMMARISED ACCURACY SPECIFICATION FOR SUBSYSTEMS.

RANGE SUBSYSTEM:

Standard deviation for a single distance determination	0,3m
Peak error at 99% confidence level	0,9m
Unambiguous range	2km

ANGLE SUBSYSTEMS:

Standard deviation of single angle determination (azimuth or elevation)	20 arc seconds
Peak error at 99% confidence level	1 arc minute
Unambiguous range	360° (azimuth)
	$> \pm 10^{\circ}$ (elevation)

VERTICAL ANGLE REFERENCE:

Range	$\pm$ 50 arc seconds
Linearity	better than 10%
Peak error	5 arc seconds

POSITION UPDATE TIME:

Each subsystem (ignoring operator observation time) shall be capable of producing a measurement in 0,3 seconds with a complete update on the vessel's position being available once per second.

2.8 EFFECT OF TRANSVERSE TILT ON MEASURED HORIZONTAL ANGLE

The axis about which the instrument alidade rotates is nominally vertical. The component of tilt in the plane containing the vertical and the line of measurement (due to inadvertent dislevelment) is automatically compensated by the vertical reference system. It is also necessary to consider the transverse component which causes the telescope axis to deviate from its nominal projected orientation in the horizontal plane.

Suppose we are observing a distant point P at elevation α. The projection of OP onto the horizontal plane XOY makes an angle θ with an arbitrary reference direction in the plane. Elementary trigonometry shows that if the nominally horizontal axis is inclined a small angle γ, the observed azimuth angle will be in error by δ where:

$$\delta = \text{arc tan } (\tan \gamma \tan \alpha)$$

We can certainly invoke the small-angle approximations for α and δ, giving

$$\delta \doteq \gamma \tan \alpha$$

For observations near the zenith the error increases rapidly; in this instrument we are concerned only with near-horizontal observations. When $\alpha = 0^\circ$ there is no error on this account. For $\alpha = 10^\circ$, the maximum expected dislevelment of 50" yields a 10" error for δ. At the peak value envisaged, $\alpha = 30^\circ$, we have $\delta_{max} = 26''$. We shall see in Chapter 7 that, if operation at high elevations is required, it is possible measure and compensate for cross-axis tilt.

2.9 INSTRUMENT DESIGN PHILOSOPHY

In the light of these considerations, it was decided to investigate in detail the design of a portable, low cost electro-optic instrument capable of measuring range and angle to a passive, co-operative target, and having performance as set out in Table 1.

In the interests of economy and simplicity, both the range and angle-measuring subsystems would be based on phase-comparison techniques, enabling them to share signal processing circuitry. Morecver, having recourse to phase-comparison techniques greatly facilitates the accurate measurement of both range and angle and neither subsystem is stretched near the limits of attainable precision. This makes it possible to meet the target specification with relatively simple circuitry in the case of the range-measuring subsystem, and with relatively crude and inexpensive mechanical fabrication techniques in the case of the angle-measuring subsystem.

Whereas horizontal angle measurements are carried out with reference to an arbitrary datum, the instrument will contain a sensor of the local gravitational gradient as a datum for the measurement of vertical angle.

Physically, the instrument should consist of a small, tripod-mountable device equipped with a telescope provided with a reticle which the operator aligns with the vessel. Ideally the only other controls requiring the operator's attention should be an on/off switch and, perhaps, a "read" button. Output from the instrument will be range in metres, azimuth angle relative to a datum arbitrarily chosen by the operator and absolute angle of elevation. The output information can be manually read and recorded by the operator, automatically recorded on magnetic tape or in a semiconductor memory, fed into an on-line computer or telemetered to the vessel.

It will be seen that the realisation of an economical integrated instrument on the lines proposed depends upon the availability of suitable compatible subsystems for the determination of range, azimuth angle and vertical angle, and a phase measuring system capable of processing the output of these subsystems. Subsequent chapters will concern themselves with the development and design of these subsystems. Since however the viability of the proposed instrument depends critically on the feasibility of detecting optical radiation reflected from the vessel, a good place to begin is the electro-optic link.

CHAPTER 3

THE ELECTRO-OPTIC LINK

3.1 INTRODUCTION : THE NEED FOR A RETROREFLECTIVE LINK

It is possible in principle to determine the distance between two stations A and B by timing the arrival at one of the stations of an identifiable signal radiated by the other at a pre-arranged time. For metre-accuracy this presupposes the existence of clocks at A and B synchronised to within about 3×10^{-9}s. Although such systems are beginning to be feasible (using rubidium clocks) for navigation, they are obviously excluded for the present in the case of the proposed instrument. We may take it for granted that two-way radiation of signals between shore station and vessel will be required for the range determination.

This restriction leaves two basic approaches open - the signal radiated from the shore station may be received and re-radiated from the vessel or it may simply be returned directly by reflection. In the case of precise microwave position-fixing systems, the relatively long wavelength precludes the narrow beamwidth and target dominance required to define accurately the effective point of reflection. The technique of effective re-radiation is therefore used, as in the Tellurometer system. Conversely in the case of electro-optic systems, where narrow beamwidth and target dominance are easily achieved, the reflection technique is always used, with the sole exception of the very long-distance instrument described in reference (20).

In any event active re-radiation aboard the vessel would require a highly directive shipborne installation and an operator to keep it directed toward the shore station. This is precluded by the single-operator requirement.

Signal return can be effected either with or without a co-operative reflector aboard the vessel. If no reflector is used (as, for obvious reasons, in the case of military rangefinders) we have a LIDAR system, the optical equivalent of radar. Very large transmitted power is needed to contend with its fourth-power law and the wide-angle scattering of the target, and the method is compatible only with high power laser pulsed

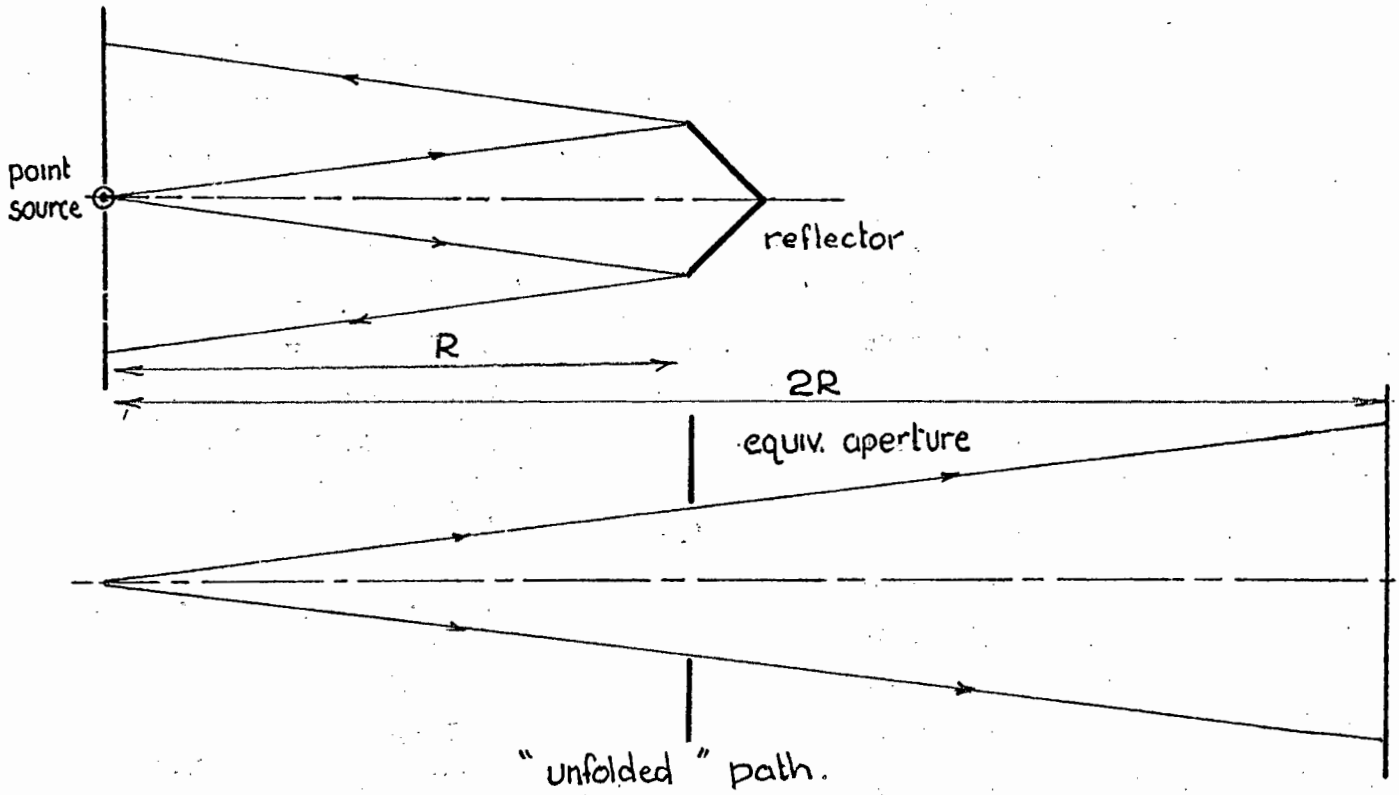
systems. If however a retroreflective (or high gain) target is installed on the vessel it is possible to achieve adequate returned power over kilometre ranges with only milliwatts of radiated power. Phase comparison distance measuring instruments (other than the extremely short range version used for automatic camera focusing) therefore invariably employ a co-operative retroreflector. The literature contains many analyses of the performance of electro-optic links with passive co-operative reflectors (21-23). The treatment below, though consistent with these analyses, will proceed somewhat differently, focusing attention on some aspects which are often overlooked.

3.2 THE REFLECTOR SYSTEM

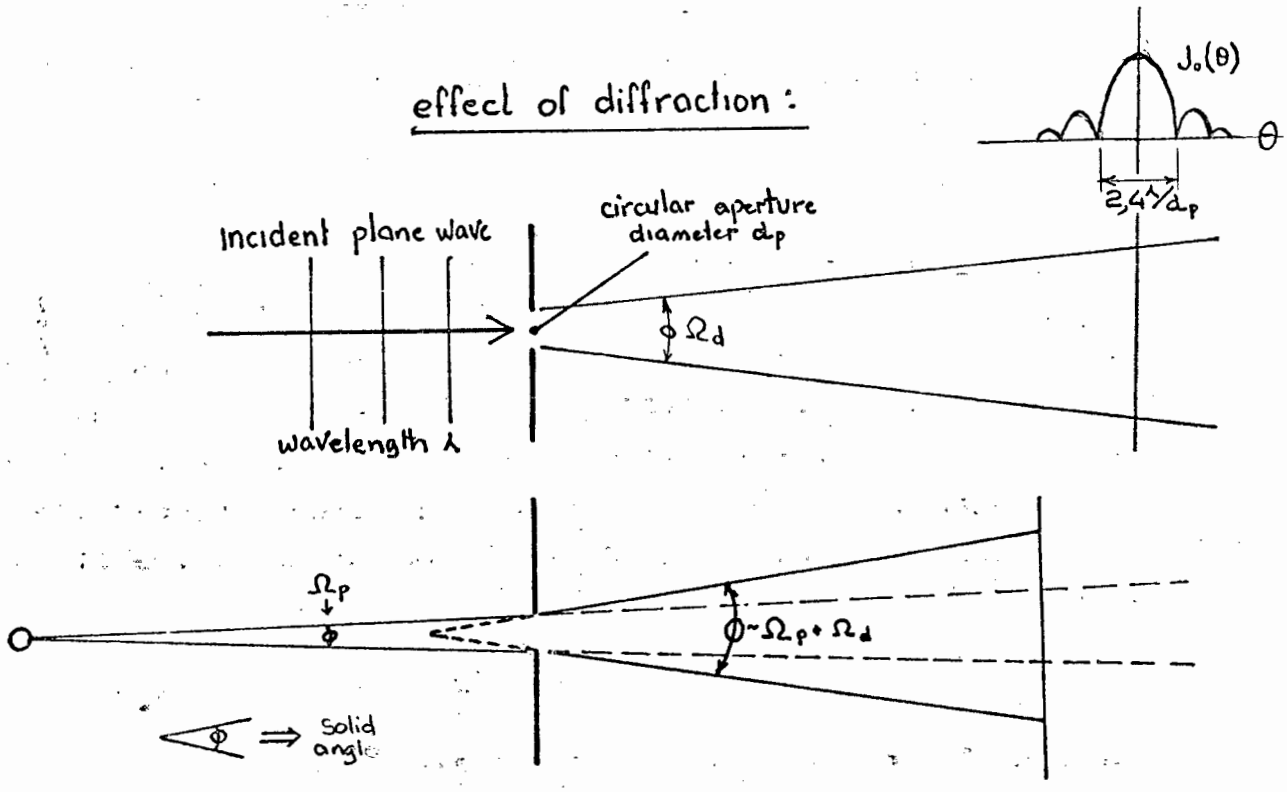
If a beam of radiation strikes a scattering target very little radiation is returned to the transmitter. The situation is dramatically improved if the target is replaced by a specular reflector - ideally the beam is then returned (deflected through 180°) without any loss in irradiance. An ordinary mirror is unsuitable as it would have to be kept aligned precisely perpendicular to the incident beam. A gain of several hundred times in the returned radiation over that due to a diffuse surface can be obtained without incurring alignment problems by coating the surface with partially retroreflective paint or cloth (such as the proprietory "Scotchlite") but a far better solution is to employ a cube-corner prism which virtually emulates a mirror without any critical requirement in respect of orientation. (A cube-corner prism can be rotated through many degrees with negligible loss in returned signal). The theory of the cube-corner retroreflector in an electro-optic link seems to be not well understood by engineers and some erroneous accounts have appeared in the literature (24).

The ideal retroreflector

An ideal retroreflector simply changes the direction of each incident ray of light by 180°. Thus, if the aperture is illuminated by means of a divergent bundle of rays, the bundle will be reflected without loss of energy or increased divergence and the radiant intensity will be unaltered. An optical link with such a reflector will exhibit an inverse-square-law dependence of received power as a function of range (neglecting propagation losses). In other words, as



A WAY OF VISUALISING THE ACTION OF A CUBE-CORNER REFLECTOR



seen in the accompanying diagram, the optical path can be 'unfolded' and the reflector considered as an aperture through which the beam travels unhindered. It is assumed that the geometry of the optics is such that all rays passing through the virtual aperture are collected by the receiver aperture.

If we let A_p be the effective area of the reflector aperture
d the distance from transmitter to reflector
Ω_r the solid angle of the reflected beam
I_t the radiant intensity of the transmitted beam (W/sr)

The solid angle of the returning radiation Ω_r is equal to that subtended by the prism aperture.

$$\Omega_r = \frac{A_p}{d^2} \qquad 3.1$$

The power intercepted by the prism is

$$W_p = \frac{I_t A_p}{d^2} \qquad 3.2$$

The radiant intensity of the reflected beam I_r is given by

$$I_r = \frac{W_p}{\Omega_r}$$

$$= \frac{I_t \times A_p}{d^2} \times \frac{d^2}{A_p}$$

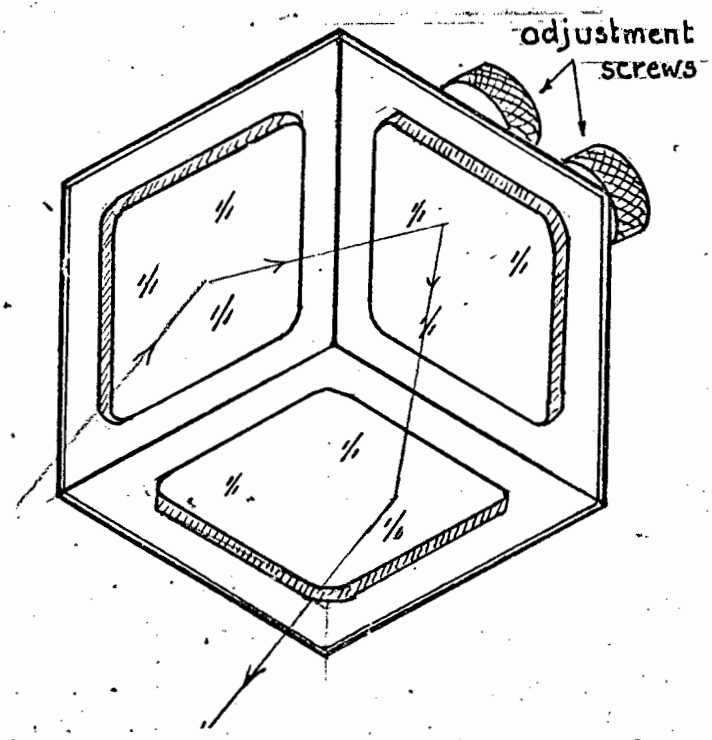
$$= I_t$$

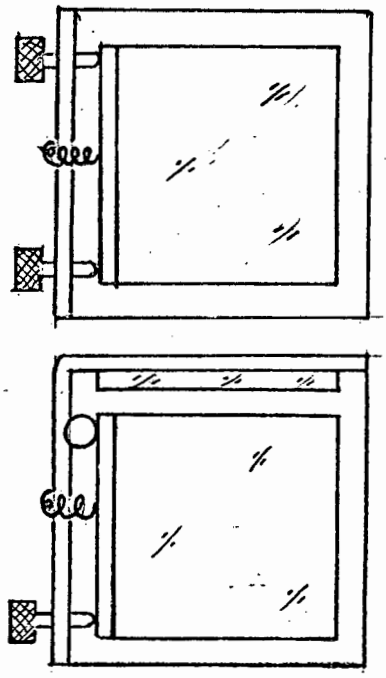
The imperfect retroreflector

In practice, the divergence of the reflected beam will exceed the angle subtended by the prism due to imperfections in the latter, and due to the diffraction spread of a beam passing through an aperture of finite size. The diffraction effect can be estimated by the angular distance, θ_d, between the first pair of nulls of the diffraction pattern for a circular aperture.

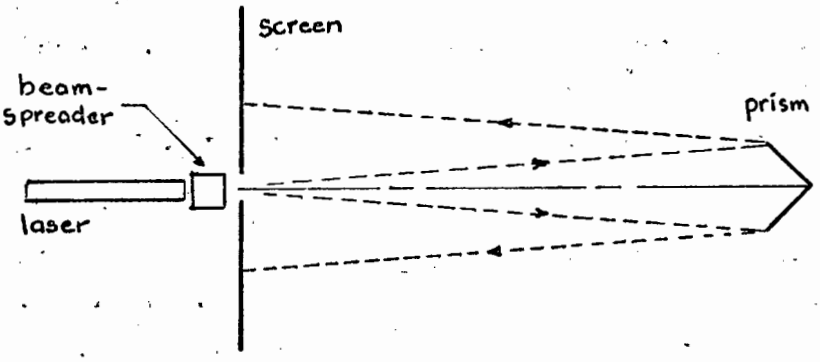
$$\theta_d = 2{,}4 \frac{\lambda}{d_p} \qquad 3.3$$

where d_p is the diameter of the aperture, and
λ is the wavelength of the radiation

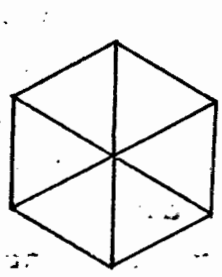




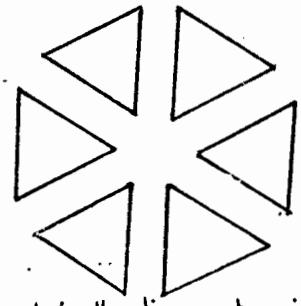
AN EXPERIMENTAL ADJUSTABLE RETROREFLECTOR.



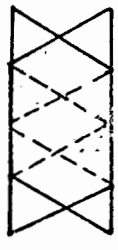
TEST SETUP



in perfect adjustment



symmetrically divergent prism



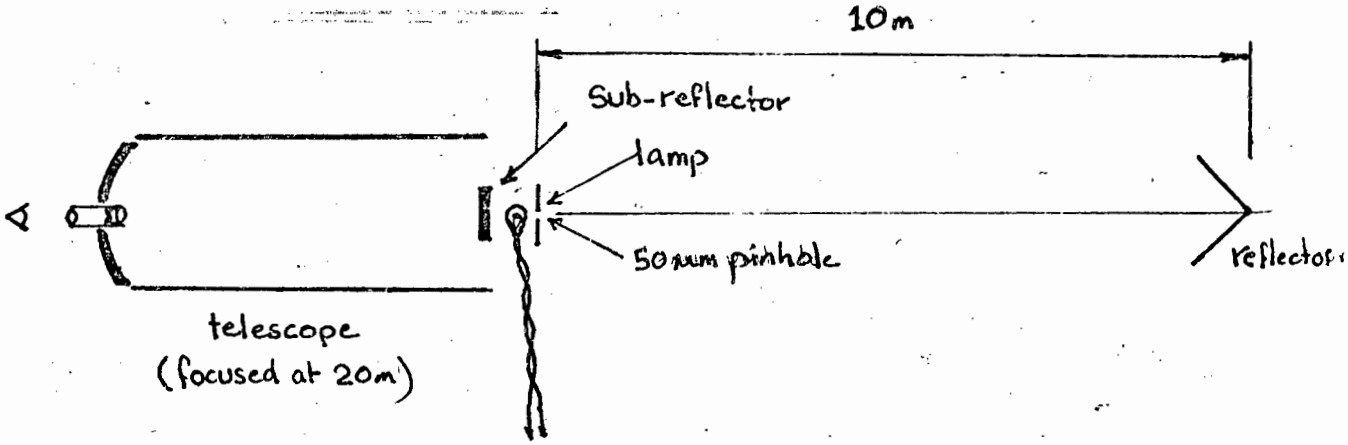
convergent in horizontal plane.

OBSERVED VISUAL PATTERNS

For the typical values $\lambda = 0,5\mu m$ and $d = 50mm$ this gives a main-lobe width of $2,4 \times 10^{-5}$ radian or 5 arc seconds.

Angular imperfections (deviations from orthogonality) cause each incident ray to be returned as a divergent bundle of six rays. In order to gain insight into the mechanism, a model retroreflector was constructed with three first-surface mirrors, one of which could be adjusted in angle. By illuminating this in a darkroom with a diverged laser beam it was easy to see the effect of angular errors. It can also be seen that a focusing effect is possible whereby at a given distance the returned radiation can be concentrated to yield increased irradiance over that due to an ideal reflector, thereby partially compensating beam divergence due to diffraction and scintillation. Negotiations by the writer with a firm of optical manufacturers (25) showed that it is, in fact, feasible to exploit this effect to achieve increased range in electro-optic instruments at little or no extra cost simply by modifying the way in which the manufacturing tolerances are written. This possibility does not seem to have been discussed in the literature. The effect is small but significant and will not be taken explicitly into account in the range predictions to follow, in the interests of allowing standard reflector prisms to be used. If it is exploited it can be considered to yield an extra safety margin of a few decibels.

Early experimentation on the range of electro-optic links gave rise to many puzzling anomalies and inconsistencies, and an unexpected sensitivity to the geometrical configuration of the optical system employed. Careful experimentation showed that, apart from the vagaries of atmospheric propagation, the major source of the trouble was retroreflector imperfections, frequently in excess of the specified limits stated by the manufacturer. Efforts were therefore made to find a way of testing prisms. Discussions with manufacturers did not prove fruitful - the standard test procedures involve the use of a large aperture interferometer (which was not available) and the test results are extremely difficult to interpret. After a number of attempts a suitable test method was devised. In essence a 160mm catadioptric astronomical telescope was used to observe the image of an illuminated 50µm pinhole reflected



perfect reflector

marginal reflector

poor reflector

VISUAL PATTERNS

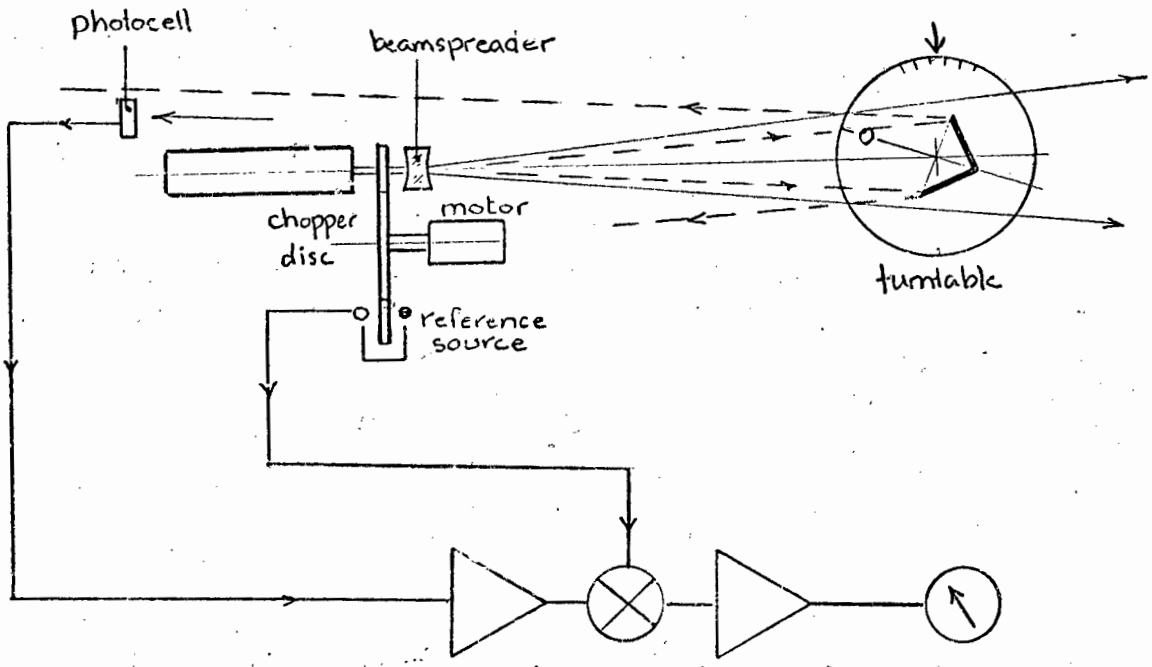
in the prism at a distance of 10m. The angular size of the image was therefore 5 micro-radians or 1,03 arc seconds. Angular error in the prism caused the image to split into up to six separate images. At a magnification of 120x the separation could be clearly perceived and estimated in relation to the size of each image, thus giving a direct measure of the deviation. This technique proved very satisfactory and several nominally perfect prisms (deviation less than 1 second) were chosen for initial range tests. This led to consistent and repeatable results and permitted the known reflector imperfections to be introduced and allowed for in a controlled manner. Prisms were also selected having nominal deviations of 3 arc seconds for final and representative range testing.

The combined effect of diffraction and angular deviation is complex but a reasonable approximation found in practice to be somewhat conservative is simply to use their sum. This gives $\theta = 8$ arc seconds. The solid angle is given to a good approximination by: $\Omega_p = \frac{\pi}{4}\theta^2$ or approximately 10^{-9}sr.

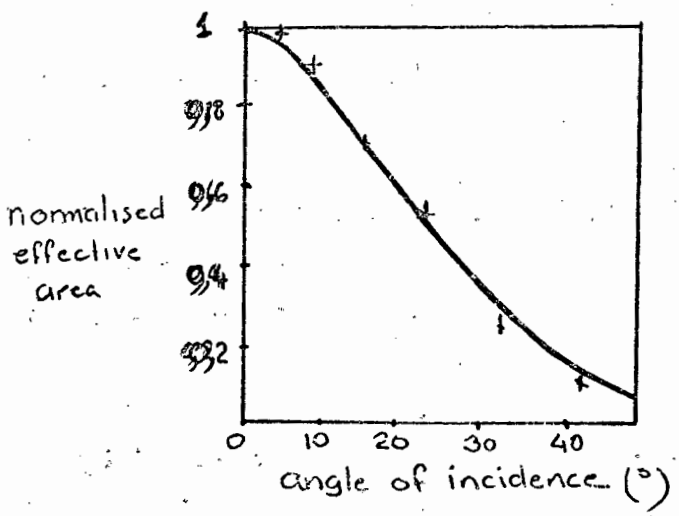
Theory of the imperfect retroreflector

The bundle of rays intercepted by a reflector is contained in the solid angle A_p/d^2. At short range where this is much greater than Ω_p the effect of prism diffraction and deviation can be neglected and the inverse square law holds as shown previously. If however the converse applies (i.e. at long range) the energy intercepted by the prism will be scattered into a solid angle Ω'_p. The most fruitful way of visualising this is in terms of the familiar radar equation with its fourth power law. This will be developed below. Note however that in comparison with a radar situation the retro-reflector exhibits extremely high directional gain. In the case of the prism considered above, the gain would be in the region of 180dB!

The transition distance at which we must consider prism imperfection can be determined by equating A_p/d^2 and Ω_p. This yields $d = (A_p/\Omega'_p)^{\frac{1}{2}} = 1,3$km.



SYSTEM FOR MEASURING POLAR RESPONSE A(p)



+ : experimental points.

For distances much in excess of this we shall see that A_p/Ω_p is a figure of merit for a prism which enters directly into the expression for range. In this case the following analysis is appropriate (again neglecting atmospheric losses).

Power intercepted by reflector $W_p = I_t \cdot \frac{A_p}{d^2}$ (watts)

This is re-radiated into the solid angle Ω'_p. Therefore the reflected beam irradiance is

$$I_r = \frac{I_t \cdot A_p}{d^2 \, \Omega'_p} \text{ (watts per sterradian)} \qquad 3.4$$

If we assume the reflected beam returned to the instrument is larger than the receiver collection area A_r square metres, the received power W_r is given by

$$W_r = I_r \frac{A_r}{d^2} \text{ (watts)}$$

$$= \frac{I_t \, A_p \, A_r}{\Omega'_p \, d^4} \text{ (watts)} \qquad 3.5$$

This has the form of the radar equation, modified by a gain term (A_p/Ω'_p) characterising the reflector. For a single prism as discussed above

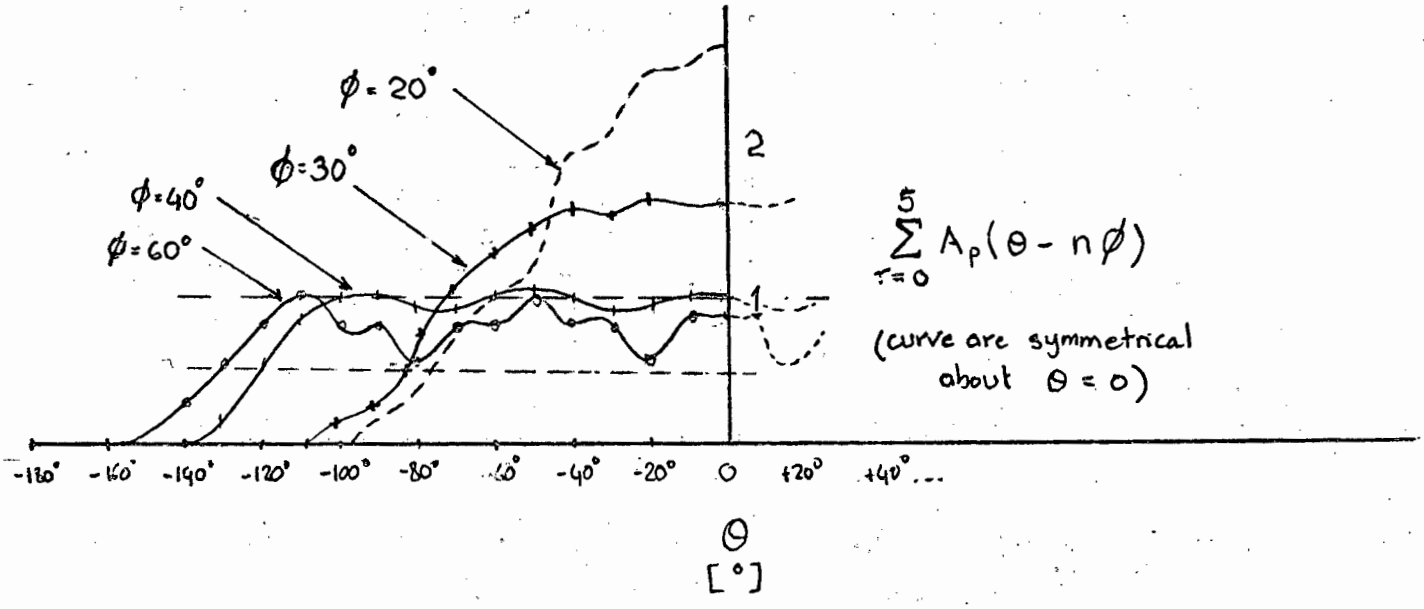
$$A_p/\Omega'_p = \frac{\pi (50 \times 10^{-3})^2}{4 \times 10^{-9}} = 2 \times 10^6 \text{ m}^2/\text{sr}$$

Therefore at the maximum range (neglecting atmospheric losses) of 2 km we have

$$W_r = 1{,}2 \times 10^{-7} \; I_t \, A_r \text{ (watts)} \qquad 3.6$$

Sensitivity to orientation

If a retroreflective prism is rotated relative to the incident beam, the energy reflected is reduced. The effective aperture of the prism $A_p(\theta)$ is plotted as a function of θ (following references 26 and 27). To verify this behaviour a prism was mounted on a turntable so as to rotate about its centroid and the returned energy measured by a large-area photocell. To improve signal-to-noise ratio and permit the measurement to be carried out in daylight, the laser beam was mechanically chopped and the photocurrent synchronously detected. The experimental points, which agree well with reference (26) are also plotted.



NORMALISED RETURN FROM AN ANGULARLY- DISPERSED CLUSTER OF 5 PRISMS.

A surprising aspect of the observed behaviour in the case of some of the prisms was a sharp assymmetry. The reason for this was traced to the loss of total internal reflection in the case of unsilvered prisms (see the following section) and is an additional reason for the use of silvered prisms, for which $A_p(\theta)$ is very nearly symmetrical with respect to θ.

In order to provide reasonable latitude for the vessel being surveyed to manoeuvre, there should be negligible reduction in $A_p(\theta)$ over some angular range θ_m, which should be at least 90° and ideally, 360°. This cannot be achieved with a single prism, but there is no difficulty (apart from expense) in providing a cluster of prisms. The number of prisms required will be determined by the total angular range required and the permissable peak-to-peak variation in $A_p(\theta)$.

Since we are dealing with incoherent radiation, the energy returned by each prism in a cluster can simply be added. The effective area of a retroreflective cluster, therefore, of n prisms singularly staggered by angle ϕ is given by

$$A_p(\theta) = \sum_{r=0}^{n} A_p(\theta - n\phi) \qquad 3.7$$

$A_p'(\theta)$ is sketched vs. θ for various values of ϕ. It can be seen that to cover 360° with an effective area no less than A_p (or that of a single prism on-axis) requires a cluster of 8 prisms.

If we are prepared to accept a 50% reduction in A_p (representing a 50% transmission loss) the number of prisms required drops to 6.

In most cases the line connecting the instrument to the vessel will be sufficiently nearly horizontal to entail negligible extra loss on this account. If it is required to work at an appreciable and varying angle of elevation (say greater than 10°) it will be necessary to accept the extra loss or to double the number of prisms in the cluster.

Practical considerations

In principle a cube-corner reflector is simply an assembly of three mutually perpendicular mirrors. In practice however the very stringent requirement for accuracy in orthogonality is more readily met if the reflector is fabricated from a block of fused silica or optical glass. Total internal reflection takes place at the glass/air interfaces, obviating the need for reflective coating, and yielding virtually perfect reflectivity (but limiting the acceptance angle as discussed above). The interfaces must moreover be kept scrupulously clean, posing severe reliability problems in a marine environment, even if attempts are made to seal the prism hermetically in a metal housing. A safer policy therefore is to sacrifice the total internal reflection and evaporate a silver coating on each reflecting surface. The reflectivity drops from unity to about 0,94, resulting in an overall transmission for the three reflections of $(0,94)^3$ or 0,83. Another slight loss (4%) occurs at the air-glass interface at the entrance aperture due to mismatch reflection. Antireflection coating is of course possible but is not likely to survive the frequent cleaning needed in a salt-spray environment. Thus the overall transmission of the prism will be about (0,83 x 0,96) or 0,8. The effect may seem small in relation to the uncertainties of, say, atmospheric transmission, but there are many such effects and they are cumulative. Experience has shown that good agreement between theory and practice requires careful attention to such detail.

The effect of finite transmission is easily accounted for. Let the prism transmission factor be 0,8. Its effect is to multiply directly the righthand sides of equations 3.3 and 3.4. Thus 2.6 becomes

$$W_r = I_t.A_p.10^{-7} \quad \text{(Watts)} \qquad 3.8$$

In summary, the great majority of applications for the proposed instrument will be served by a cluster of 50mm prisms as shown. The performance of such an array is conservatively described by equation 3.5 where (A'_p/Ω'_p) is about 2 x 10^6 m^2/sr with an optical transmission factor of 0,8.

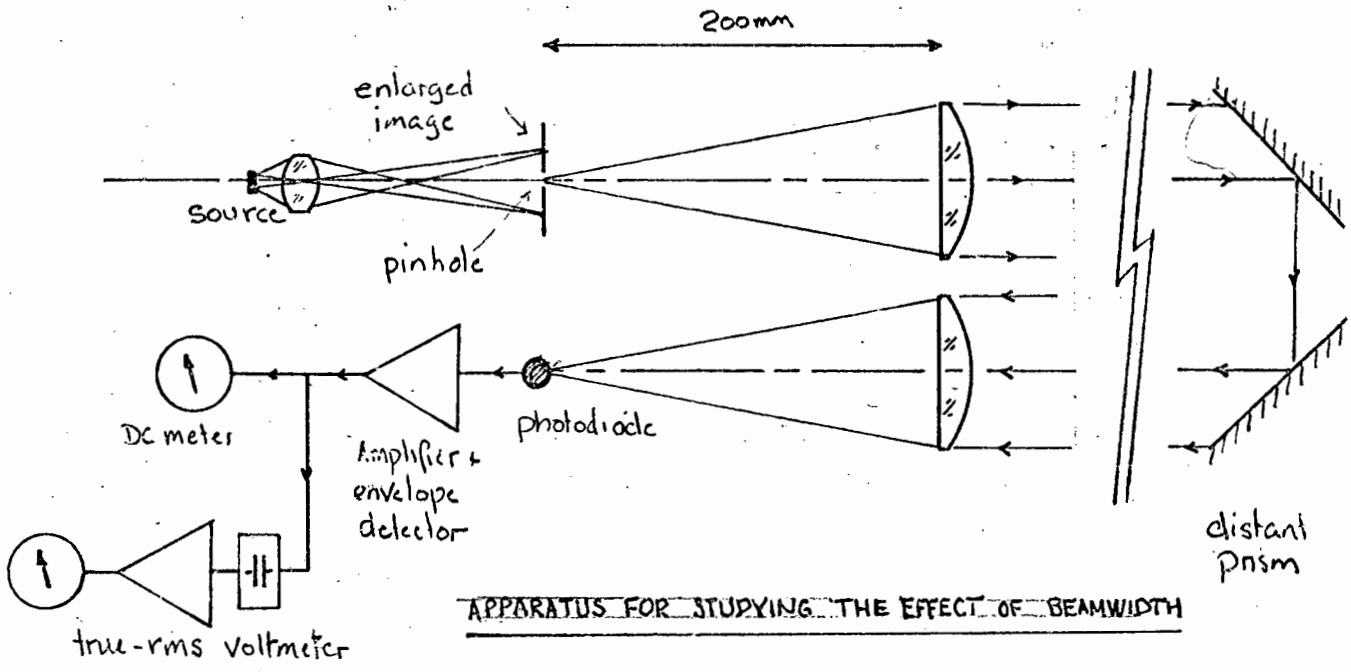
3.3 CHOICE OF BEAMWIDTH

In order to produce the near-parallel beam required, two elements are necessary - a source of radiation and collimation optics. The purpose of the latter is to concentrate the radiation into a near-parallel beam. An important parameter of the system is the angular divergence, or beamwidth of the transmitted beam.

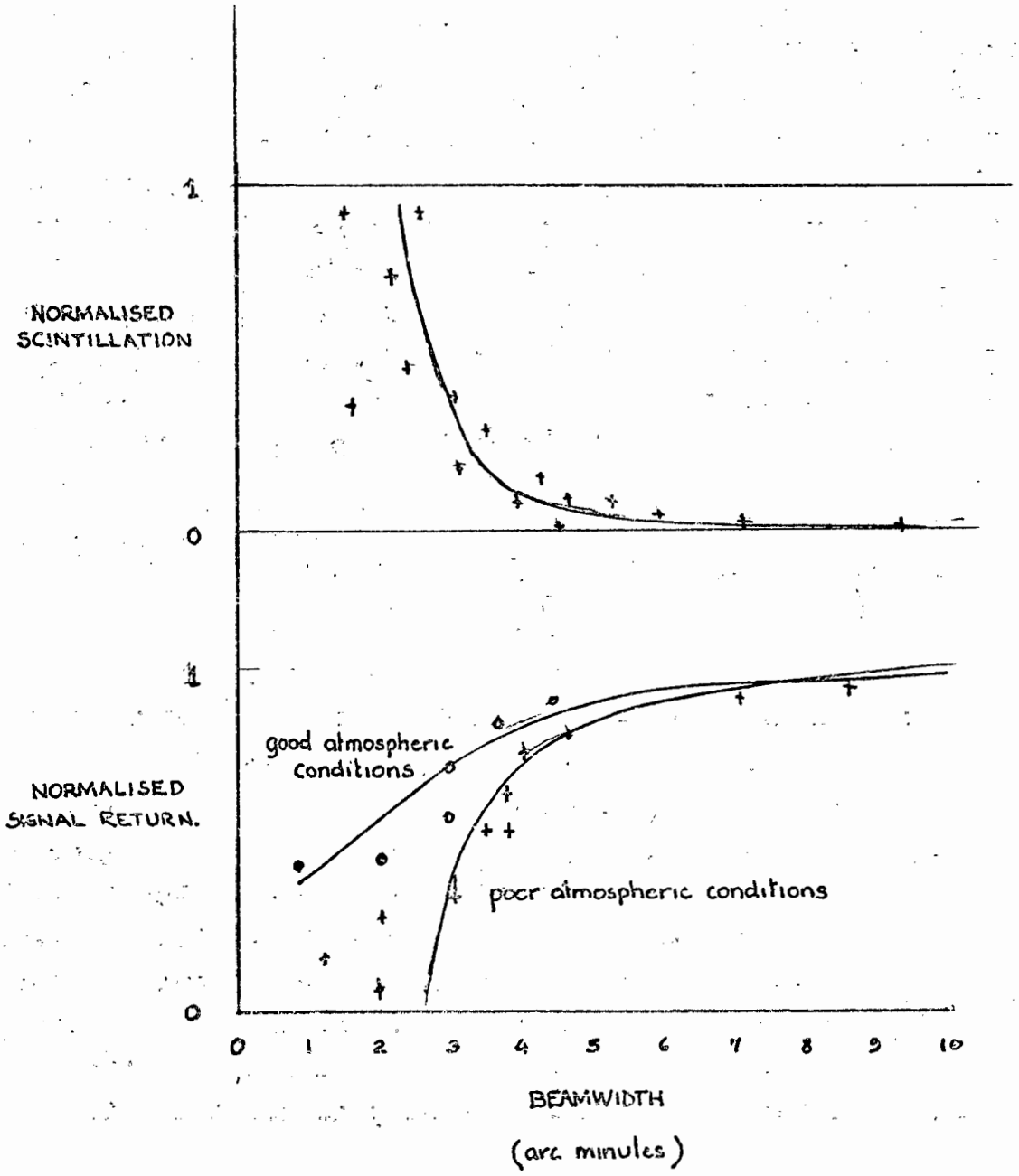
Required beamwidth

Beamwidth has been a rather controversial aspect of the design of electro-optical survey instruments. Virtually all the first generation of instruments based on the luminescent diode had beamwidths on the order of a quarter of a degree, or 15 arc minutes (28). However the quest for greater range led progressively to the use of physically smaller sources, with an attendent reduction in beamwidth, culminating in an instrument by Hewlett Packard with a beamwidth of about 2 arc minutes (18). Some of the commercial laser-based instruments, such as the Spectra-Physics Geodolite and the AGA Geodometer model 8, employ a beamwidth of 0,1 milliradian or 20 arc seconds.

One of the penalties of narrow beamwidth is a significantly greater difficulty in pointing the instrument for signal acquisition, particularly if the target cannot be clearly seen, as is frequently the case for distances in excess of a kilometre. Experience in the field shows that the difficulty escalates fairly rapidly for beamwidth less than 5 - 10 arc minutes. This however does not apply to the instrument under consideration, in which the target will be a vessel clearly silhoutted against the skyline or sea background. Indeed the angular measurement requires that the instrument be pointed within about 20 arc seconds of the vessel's position when a measurement is taken. Provided that boresight collimation between the telescope and emerging beam can be maintained within 20 seconds (a fairly easy requirement to meet) it might be agreed that there would be no operational disadvantage in a beam as narrow as, say, 40 arc seconds. There are however other considerations which point to the desirability of a wider beam. These relate to the effect of atmospheric "scintillation" which, since the topic is controversial, will be discussed below.



APPARATUS FOR STUDYING THE EFFECT OF BEAMWIDTH

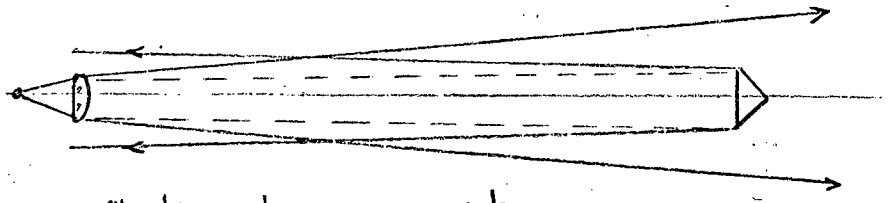


The effect of atmospheric scintillation

A beam of light propagating in the atmosphere undergoes an increase of beamwidth on the order of several seconds of arc on account of wavefront breakup due to turbulance-induced refractive index gradients. These effects have been investigated in detail by a number of workers (29-31). The effect depends on the nature of the terrain and estimates differ widely, but informal observations with a theodolite over grass-covered terrain support the order-of-magnitude cited. The fluctuation of arrival angle can have frequency components as high as several hundred hertz but the magnitude is too small to have much significance in the system under discussion. A more significant effect is beam-wander due to path curvature. This is caused by the vertical refractive index gradient due to pressure and temperature variation with altitude. Variations in the typical temperature lapse rate of $-6^{o}C/km$ can give rise to a diurnal variation of about 30μrad/km, or 6 arc seconds/km (32) but tests with a laser beam 1 - 2m above the ground showed that much greater beam-wander occurred. Again there are many variables and rigorous experimentation in this area is notoriously difficult and time consuming, but observations of a laser beam over a 1km path suggested that fluctuations of up to 300μ rad peak (or 1 arc minute) occur with a frequency centred about 1 - 5Hz. The effect of this on the received signal is to produce multiplicative noise (or deep-fade characteristics) with serious effects on signal-to-noise ratio. To avoid this, a beamwidth at least greater than 2 arc minutes is indicated.

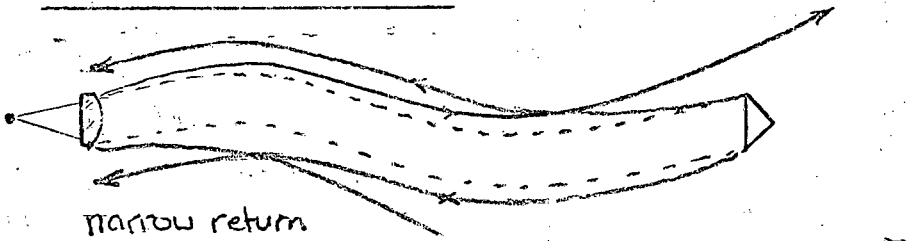
This was verified, again rather qualitatively, by the apparatus shown. A beam was formed by collimating radiation from the enlarged real image of a gallium arsenide diode, formed by an auxiliary optical system. In the plane of the real image, a set of calibrated pinholes, ranging from 50μm to 500μm defined the size of the effective source (and hence the beamwidth) without affecting irradiance. The focal length of the collimator was 200mm therefore the achievable range of beamwidths was 0,86 arc minutes to 8,6 arc minutes. The results are shown. At short range, or in 'quiet' atmospheric conditions the returned signal hardly depends on the beamwidth, but under conditions of high scintillation

clear atmosphere:



reflector returns narrow beam

turbulent atmosphere:



narrow return beam follows path of outgoing beam to receiver.

wandering beam

EFFECT OF BEAM WANDER

the returned signal drops off significantly for beamwidths of less than about 4 minutes of arc.

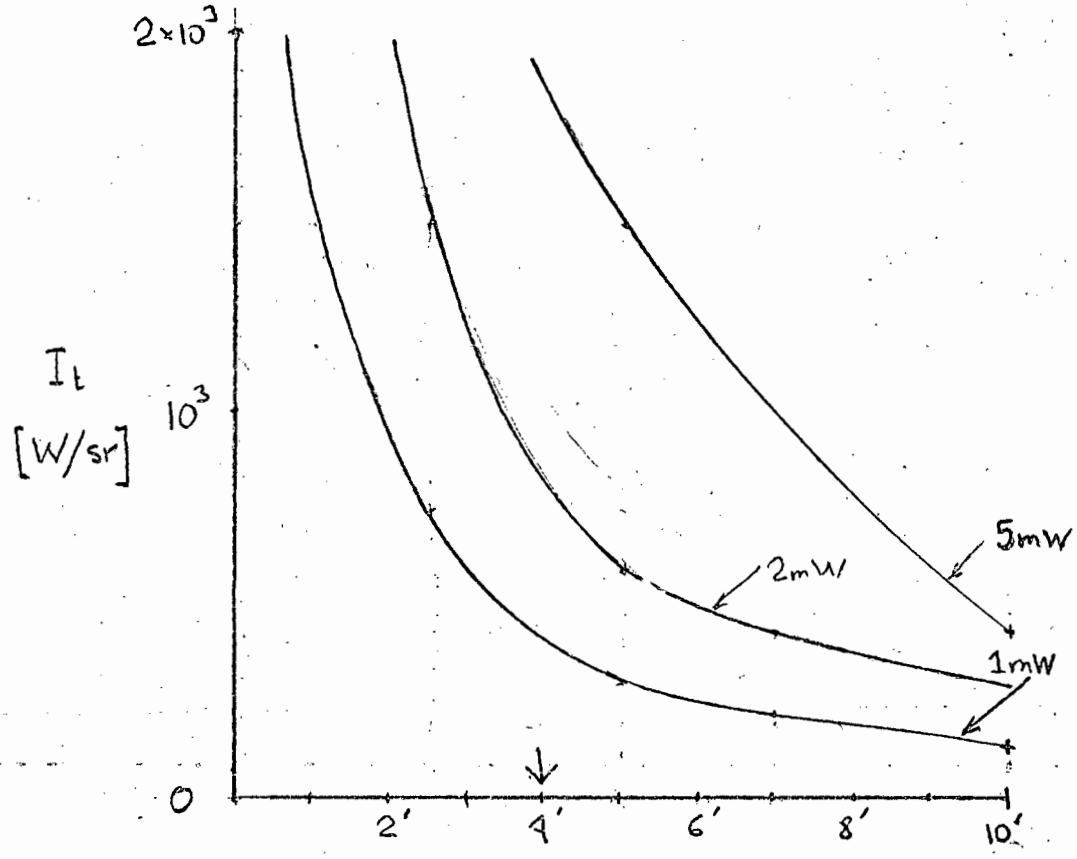
It has been erroneously argued that transmitted beamwidth cannot be important since the performance will be dominated by the much smaller divergence of the reflected beam. This overlooks the fact that atmospheric variations are slow, so that the atmospheric path is stationary in terms of the transit time of a light signal. Therefore all ray-paths are reversible and any rays (from any part of the transmitted bundle) which are intercepted by the reflector will be returned to the instrument (see diagram). Thus the performance is dominated by the transmitted beamwidth.

There is no objection in principle to a greater beamwidth, except that it implies optics of lower numerical aperture (and hence greater cost) and/or a waste of power (and thus excessive battery consumption). The optimum beamwidth would appear to be about 5 arc minutes (1,5 milliradian). It is interesting to note that this represents about the median value adopted in commercial instruments.

3.4 BEAM IRRADIANCE

Equation 3.5 shows that the irradiance of the transmitted beam I_t is one of the factors which determines the operational range of an electro-optic link. The purpose of this section is to relate I_t to the properties of the radiant source. It is necessary to distinguish two cases: if the radiation is available from a coherent ('dimensionless') source such as a laser, the transmitted irradiance can be directly calculated as the ratio of the available power and the solid angle occupied by the beam. If however an incoherent (extended) source such as a gallium-arsenide spontaneous emitter is used, this approach, though still formally valid, ceases to be fruitful. This is because of a hidden relationship between power and source area (and hence beamwidth). A parameter of the collimation optics - the area A_t - also enters directly into the determination of I_t.

In the case of the coherent source, only the diameter of the transmitting optics determines the beamwidth (due to diffraction).



RADIANT INTENSITY VS BEAMWIDTH
OF A LASER BEAM.

Case I. The coherent source

Let the total beamwidth be θ radians
the transmitted power be W_t watts

We can see immediately that

$$I_t = \frac{W_t}{\Omega_t}$$

$$= \frac{4W_t}{\pi\theta^2} \qquad \text{(watts/sterradian)} \qquad 3.9$$

To get a feel for magnitude let us calculate the beam irradiance obtained if we concentrate the power from a typical 1mW helium-neon laser (for which $\lambda = 0{,}63\mu m$) into the previously calculated 1,5mrad beam. Then,

$$I_t = \frac{4 \times 10^{-3}}{\pi \times (1{,}5 \times 10^{-3})^2}$$

$$= 5{,}6 \times 10^2 \text{ W/sr.}$$

The emergent beam diameter which would give rise to 1,5 milliradian beam spread is shown by equation 3.3 to be

$$\frac{2{,}4 \times 0{,}63 \times 10^{-6}}{1{,}5 \times 10^{-3}} = 1\text{mm}.$$

This is about the exit diameter of the beam of a typical laser which does indeed have a beam divergence on the order of a milliradian. The laser could therefore be used in this application without ancillary optics. I_t vs θ is plotted from equation 3.9 with W_t as a parameter. It can be seen that very high beam irradiance can be achieved for narrow beamwidths, permitting very long-range operation in conditions of low scintillation.

Case II. The extended source

Now we envisage a beam projected by situating a source of area A_s (m^2) and irradiance I_s (W/sr) at the focal point of a lens of effective area A_t (m^2) and focal length f (m). The analysis which follows is approximate but adequate for our purposes. A more rigorous treatment is given in Appendix 3.1.

The power transmitted W_t is given by the product of the effective source irradiance and the solid angle subtended by the lens at the source. It is necessary that I_s be an average value determined over this complete solid angle.

$$W_t = I_s \frac{A_t}{f^2} \quad \text{(Watts)} \qquad 3.10$$

The small-angle approximation implicit in taking A_t/f^2 as the solid angle subtended by the lens is well justified for numerical apertures 'slower' than f/2, for which the error would be only a few percent.

The angular divergence of the beam in the far-field will be represented by a solid angle equal to that subtended by the source at the focal length of the lens

$$\Omega_t = \frac{A_s}{f^2} \quad \text{(sterradian)}.$$

Therefore the transmitted irradiance is given by

$$I_t = \frac{W_t}{\Omega_t}$$

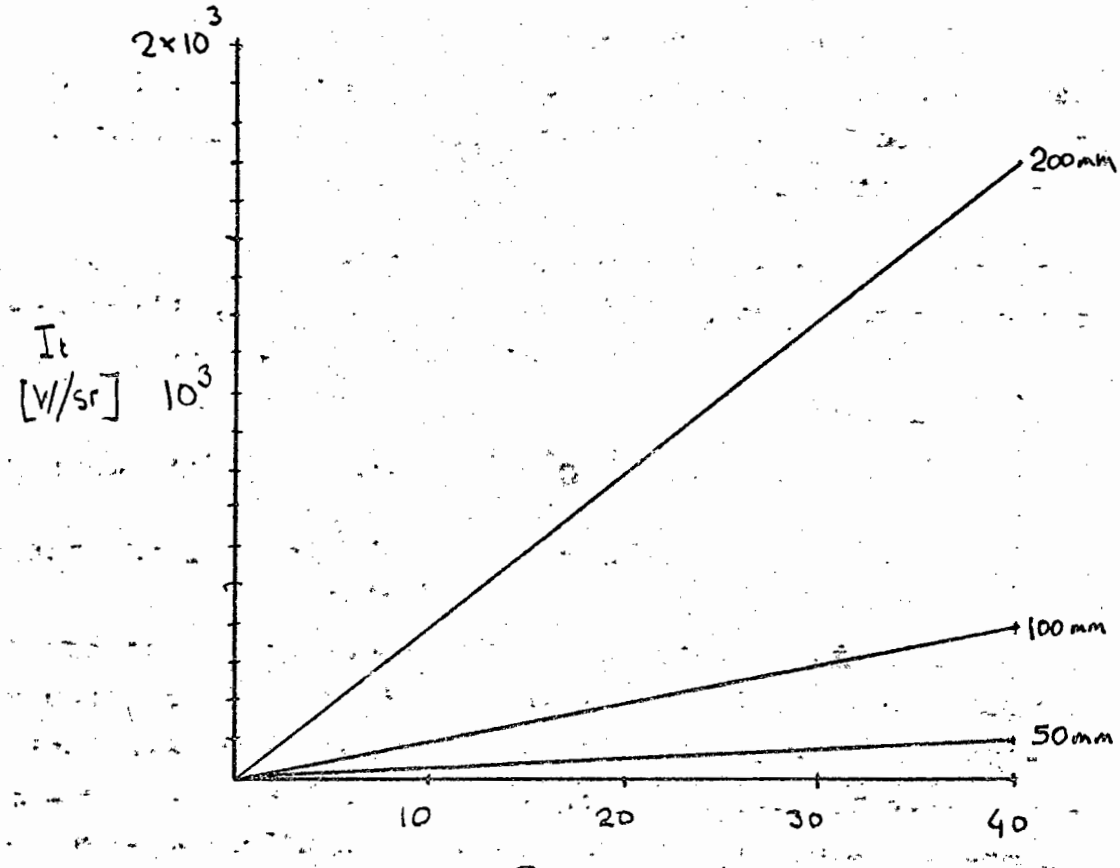
$$= I_s \frac{A_t}{A_s} \quad \text{(Watts/Sterradian)}. \qquad 3.11$$

One helpful way of interpreting this is to note that for an observer in the far field on the optical axis, the entire transmitting aperture appears to be as bright as the source, and the irradiance of the source is effectively multiplied by the ratio of the area of the transmitting aperture to that of the source.

Alternatively, we can write

$$I_t = N.A_t \quad \text{(Watts/Sterradian)} \qquad 3.12$$

where N is a parameter characterising the source and equal to I_s/A_s, or the power emitted into unit solid angle per unit source area. N is called the <u>radiance</u> of the source and it is traditionally quoted in the units watts per centimetre squared - sterradian.



lens diameter D_r as parameter

It is not always realised that, in situations where the beamwidth is unimportant, the radiance (and not the power) of the source is the property which determines the range. The area of the source however (along with the focal length of the lens) determines the beamwidth. For the same reason there is no advantage in terms of range in using an expensive optical system of low numerical aperture (f - number) in order to capture more of the flux from the nearly isotropically radiating source. More power will indeed be radiated in the beam but the beamwidth will be increased and the flux illuminating the target is unaltered.

For the purpose of comparison with the laser, we can calculate the irradiance produced by a typical modern gallium arsenide source having a radiance N_s of $5W/cm^2sr$, collimated by a lens of 50mm diameter.

$$I_t = 5 \times 10^4 \times \frac{\pi}{4} \times (50 \times 10^{-3})^2 \text{ (watts/Sr)}$$

$$= 10^2 Wsr^{-1} \text{(i.e. only about 1/60 that due to laser).}$$

In order to facilitate intercomparison between coherent and extended sources in electro-optic links, I_t is plotted vs N with diameter of the collimating lens as a parameter. This analysis does not unfortunately exhaust the interplay of parameters since there is also a correlation between source area and radiance for electroluminescent diodes (33; 34). This topic, and the choice of an optimum source will be treated below.

3.5 CHOICE OF A RADIATION SOURCE

Although the gallium-arsenide luminescent diode seems an obvious choice, its performance is somewhat marginal in respect of range, and a higher performance source would permit economies to be effected in other parts of the instrument (or alternatively could be traded off for greater range and increased versatility for the instrument). It was therefore considered that a systematic study should be undertaken of other possible approaches. First however it is necessary to make some assumption regarding the frequency at which the source must be capable of being modulated. This topic is considered in detail in a later chapter but experience suggests that, for the order of precision sought, a frequency

in the hundreds of kilohertz is appropriate. This feeling is confirmed by a plot of claimed or attained accuracy vs modulation frequency for a variety of commercial electro-optic instruments. Certainly a modulation frequency of 1MHz would be more than adequate.

(a) Conventional sources with external modulation

Considerations of atmospheric transparency (see diagram), eye-safety and availability of suitable detectors of high quantum efficiency suggest that a source in the visible or near-infra-red portion of the spectrum would be most suitable. The possibility exists of using any one of a number of high-radiance sources (zicronium compact-arc, high pressure mercury vapour, quartz-halogen incandescent or even helium-neon laser) followed by an external modulator. The hundreds-of-kilohertz region is unfortunately beyond the range of direct mechanical modulators (spinning wheel or piezo-electric tuning-fork choppers). There exists, however, a number of non-inertial devices such as the acousto-optic (Bragg effect), magneto-optic (YIG) and electro-optic (KDP and PLZT) modulators all of which are easily capable of modulating light at several tens of megahertz. All of them, however, require either high modulation power or high modulation voltages. Both of these are undesirable in a portable instrument and lead to severe (crosstalk) problems in phase-comparison distance-measuring equipment (see Chapter 4). Moreover, being of small physical and/or low numerical aperture, they require auxiliary optical systems to shape the input light beam. Yet another problem relates to spatial phase variations across their aperture which will be discussed below in the context of the gallium-arsenide diode. The most promising of the devices mentioned appears to be the relatively new 'ceramic' PLZT modulator, and experiments were carried out on a sample device supplied by the Allen Clarke Research Centre of the Plessey Company.

On balance, it was decided early on that the use of indirect modulation, although attractive in some ways, would inevitably lead to significant increases in bulk, cost and complexity of the instrument. It was therefore decided to limit the study to directly modulatable sources.

(b) Direct modulation of conventional sources

Incandescent sources are excluded due to thermal inertia. Low pressure rare gas discharge tubes cannot be modulated much in excess of a hundred kilohertz due to de-ionisation time, and moreover are very low radiance sources. High pressure vapour tubes have high radiance and again, an upper limit of modulation frequency not greatly in excess of 100kHz. Moreover Karplus, in a comprehensive review (35), points out that they tend to exhibit spatial wander and extreme non-homogeneity of modulation depth and phase. Karplus recommends the use of Z-modulation of a high intensity, low persistence cathode ray tube but this is hardly a practical approach for a portable instrument, on account of size, fragility and high voltage requirement. Spark sources have been considered but short electrode life, time jitter and generation of electromagnetic interference make them unsuitable in the intended application.

(c) CW Gas lasers

The obvious choice would be the helium-neon laser since this is the only small laser mass-produced at low cost. A series of tests on lasers purchased from Metrologic and Spectra-Physics showed that the light output is proportional to discharge current up to a frequency of 100 — 200 kHz, above which the modulation depth fell off rapidly, presumably on account of de-ionisation time.

It was noted that a commercial interferometer produced by Hewlett Packard (36) employed a laser in which the output of a specially designed laser was frequency-shifted by Zeeman splitting. It was conjectured that mode interaction between the Zeeman components would give rise to polarisation modulation which could be converted to amplitude modulation by means of a linear polarising filter, and it was decided to conduct an experiment to determine whether a standard commercially available laser could be modulated in this manner.

A Spectra-Physics 239 laser was subjected to an axial magnetic field of about 6×10^{-2} Tesla, while its cavity length was stabilised by a thermal feedback loop. It was found that modulation could be achieved at frequencies in the range of 1 kHz to 2 MHz, when the laser activity Q began to limit the modulation depth, and it was possible to phase-lock the amplitude oscillations to an external crystal-controlled

oscillator by controlling the intensity of the magnetic field. Arising out of this work, the present writer formulated a proposal for a new type of precise distance-measuring instrument which subsequently became the subject of a patent application.

It can be concluded that it is indeed possible to modulate a laser up to at least 1MHz in a reasonably simple and reliable manner. Moreover it is possible to do this in a way which avoids the presence of any high-level electrical signal synchronous with the modulation, a fact which has important implications for distance measuring in that it avoids the 'crosstalk' error mode discussed in the following chapter. It is felt that further work should be done to explore the potential of this mode of operation of a laser both in distance measurement and as a source for interferometry.

Despite the promising results attained it was decided not to pursue this approach for the present instrument, if a semi-conductor source proved to be adequate. Compared with the latter, the laser solution is complex and rather expensive, and the laser is relatively fragile, has a limited operating life and is inefficient, drawing typically 20mA at 800V (or 16W) of excitation power to produce 1 milliwatt of output (placing a heavy load on the battery operating the instrument). One other contra-indication should be mentioned, although its mechanism is not fully understood. It appeared from field observations that the laser beam was very much more susceptible to poor signal-to-noise ratio due to asmospheric scintillation than was a beam produced by an extended non-coherent source, even when the beamwidths were the same. It is conjectured that the effect is related to the coherence of the laser beam, being caused by micro-multi-path propagation in the beam resulting in a continuously moving pattern of interference nulls. If this conjecture is correct the effect can be compared with two well known analogous phenomena; one is the speckle pattern observed when a laser beam impinges on a diffuse reflector, and the other is the way in which stars 'twinkle' much more than planets on account of their spatial coherence.

There are applications in large-scale metrology for which the continuous gas laser is uniquely suited, and it has proved to be a most valuable component of electro-optic systems. However its disadvantages in terms of bulk, cost, fragility and power requirements lead us to prefer a solid-state alternative, should its performance be adequate. It was therefore decided at this point to investigate in more detail the suitability of gallium-arsenide electroluminescent devices.

(d) Semiconductor luminescent devices, coherent and non-coherent

These fall into two classes - non-coherent sources in which radiation takes place through spontaneous radiative recombination of holes and electrons in a semiconductor junction device, and coherent sources or semiconductor lasers in which the junction is part of an optical cavity and stimulated emission takes place. An intermediate class - the 'super-radiant' diode has also been reported in the literature but it offers no special advantage and has not been commercially exploited. Non-coherent emitters have been fabricated from a wide variety of compound semiconductors, emitting radiation of various wavelengths, but gallium-arsenide devices emitting in the near infra-red between 850 and 950nm have exhibited the greatest efficiency, and highest radiance and stimulated emission has been obtained only with gallium-arsenide. We shall therefore consider only GaAs devices.

The typical power radiated by a non-coherent GaAs source is in the range 0,3 - 3mW whereas semiconductor lasers produce continuous power in the range 1 - 50mW, but it is not here that the significant difference lies. The lasing diode is a source of vastly greater radiance, emitting the radiation from a much smaller area and into a much smaller solid angle. The radiance of a semiconductor laser may exceed that of a non-lasing device by up to six orders of magnitude, but this cannot be fully exploited in a beamwidth-limited situation since the very small emitting area would result in an unacceptably narrow beamwidth. Any attempt to magnify the effective area by optical means will rapidly reduce the effective radiance. An estimate of the relative advantage in an electro-optic link of a lasing device can be obtained by assuming that all the power from a medium-size 10mW device is concentrated uniformly

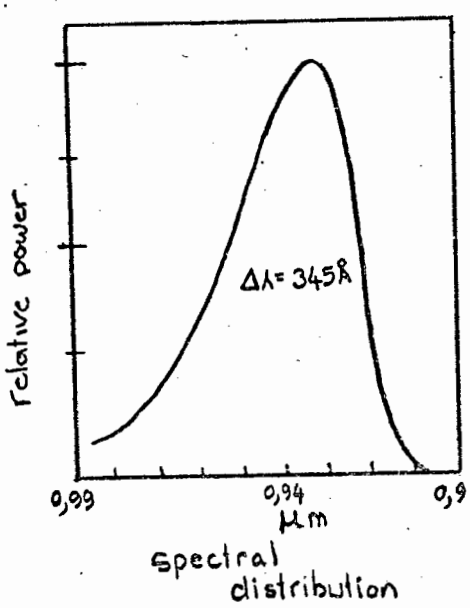
in the optimum beamwidth which we have determined to be 1,5 milliradian or 5 arc minutes. This gives a transmitted irradiance I_t of 5600W/sr (from equation 3.9). Comparing this with the situation considered in section 3.4 where a $5W/cm^2sr$ source was collimated by a 50mm diameter lens with I_t = 100 W/sr, the gain is 5,6 times. This is significant, but not spectacular - it would be matched, for example, by an increase in lens diameter, in the case of the incoherent source, of $\sqrt{5,6}$ or nearly 2,4 times although this would be an expensive solution if, as is likely, a lens of low numerical aperture and good resolution were required. The problem of shaping the beam to the required divergence when using a laser can be solved by coupling the radiation from the laser into an optical fibre and using the free end of the fibre as a source, which is placed at the focal point of the collimating optics. It will be seen in Chapter 4 that the use of a short length of light-guide or optical fibre in this way also has important advantages in distance measuring instruments in that measuring errors associated with phase variations across the emitting source are greatly reduced. Lasing diodes - rather surprisingly - are not immune to this source of error (37).

It is in coupling power into the light-guide that the lasing diode proves to be dramatically superior to the spontaneous emitter. The small physical size and narrow angular concentration of the former device yields coupling efficiencies of up to 60% in contrast with the few percent attainable using non-lasing sources (38-40). At the time that this phase of the work was being experimentally investigated it proved impossible to obtain suitable (hetero-junction) lasers capable of continuous room temperature operation, and there were (and to a much lesser extent still are) unanswered questions regarding their reliability and probable commercial cost. Accordingly all experimental work was carried out using non-lasing sources and these proved adequate, if only marginally so. The use of a semi-conductor laser offers improvements in I_t of at least an order of magnitude with a welcome increase in safety margins.

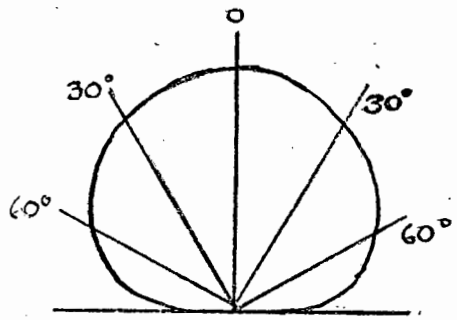
The state-of-the-art in semiconductor lasers and their coupling to optical fibres has advanced considerably in recent years due to intensive research activity directed toward the development of fibre-optic communication links, and a room-temperature continuous laser coupled to a short length of fibre can almost be regarded as a commercial component. Accordingly experimental work in this area was not considered a necessary part of this thesis. This is fortunate, since such devices are still remarkably difficult to acquire for development purposes if one does not have suitable industry contacts. Given the fact that non-lasing sources proved adequate, and the fluid situation regarding suitable lasing devices, this thesis will confine itself to a short paper study on the latter approach.

The non-coherent electroluminescent diode

It has been shown that, for all but the smallest diodes, the relevant parameter in an electro-optic link is radiance (the radiometric equivalent of the photometric term brightness which is sometimes used). This parameter is rarely quoted by manufacturers and it was necessary to devise suitable techniques for its measurement. These are detailed in Appendix 3.2. These methods were used to investigate a wide range of commercially available devices (Monsanto MV50, Fairchild FD100, General Electric SL1, Texas TIL27, Hewlett Packard 4400, Plessey GAL1 etc) with generally disappointing results. Values of radiance ranged from 0,02 and 0,3 W/cm^2sr. Perusal of the literature showed that, at least in the laboratory, values two orders of magnitude higher than this could be achieved (33). One promising commercial device was the Texas Instruments TIXL12. This is a high-power emitter actually intended for high-speed illumination, and has a power output of about 35mW which is unusually high. It achieves this by means of a hemispherical dome of gallium arsenide mounted on the junction. This frustrates the total internal reflection at the gallium arsenide/air interface which would otherwise allow only rays within 4^o of the normal to pass. Since the dome operates mainly by magnifying the source area and increasing the angle of emission, its effect on the radiance is small. Nevertheless several samples evaluated showed an impressive 10-12 W/cm^2sr.



spectral distribution



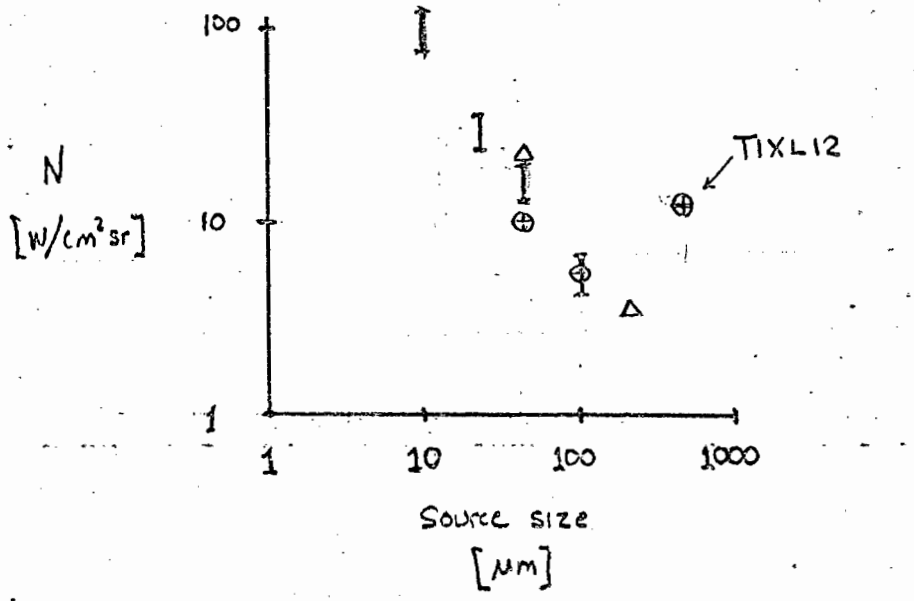
Polar diagram

max forward cont. current 300mA

on-axis irradiance 15mW/sr

effective source size ~ 0,5mm

TIXL 12 CHARACTERISTICS



GaAs DIODES — RADIANCE VS SOURCE SIZE.

I — reported by Burrus (33)

△ — reported by Davis

⊕ — measured by the writer

Note: of the above devices, only the TIXL12 is silicon-doped.

This is presumably due to the fact that it is silicon doped. Doping gallium arsenide with silicon has the effect of shifting the emission wavelength to 935nm, well away from the bulk absorption band (41;42). Most of the experimental work to follow was carried out with a TIXL12 measured at $N = 12W/cm^2sr$. The characteristics of the device are detailed opposite. Note that it is exceptional in that, due to its nearly hemispherical polar diagram, the average radiance as seen over large solid angles is virtually equal to the on-axis radiance.

An alternative and highly promising structure for high radiance devices is the Burrus-type gallium-aluminium-arsenide diode (43). The plot of N vs source diameter in section 3.4 follows a publication by Burrus in 1972 (33) and shows the strong negative correlation between source size and radiance. Included on the graph are some experimental values measured using samples of GaAlAs zinc-doped diodes supplied by the Allen Clarke Research Centre (44).

The modulation capability for these two types of diodes differs markedly. It has been shown that zinc-doped devices can typically be modulated at rates in the hundreds of megahertz (41;44). In the case of silicon-doped devices recombination time for spontaneous emission limits modulation rates to a few hundred kilohertz, with a possible 70% modulation depth at 1Mhz. This is adequate in the present application.

Fibre-optic sources

The desirability of using the free end of a short length of optical fibre as a source has been mentioned and will be discussed in the following chapter. Some simple experiments were carried out along these lines with disappointing results: for example the radiance measured at the free end of a 100mm length of 0,5mm guide butt-jointed to the dome of a TIXL12 was less than 1 W/cm^2sr, represencing a loss of an order of magnitude. Subsequently, however, experiments carried out at the Allen Clarke Research Centre of the Plessey Company (44) showed that a loss factor of only 2-3 in radiance was possible if a 100mm fibre was butt-jointed against the emitting area of a Burrus-type device. A further improvement could be effected by interposing between the emitting area and the fibre a sphere of high-refractive-index (chalcongenide) glass. Samples were provided with

$N = 3\ W/cm^2sr$ at the end of a 100mm fibre.

Extensive work has since been carried out on launching optical power into fibres. This has been well-documented (38-40) and it is possible to make a reasonably reliable prediction of the effective radiance that can be achieved.

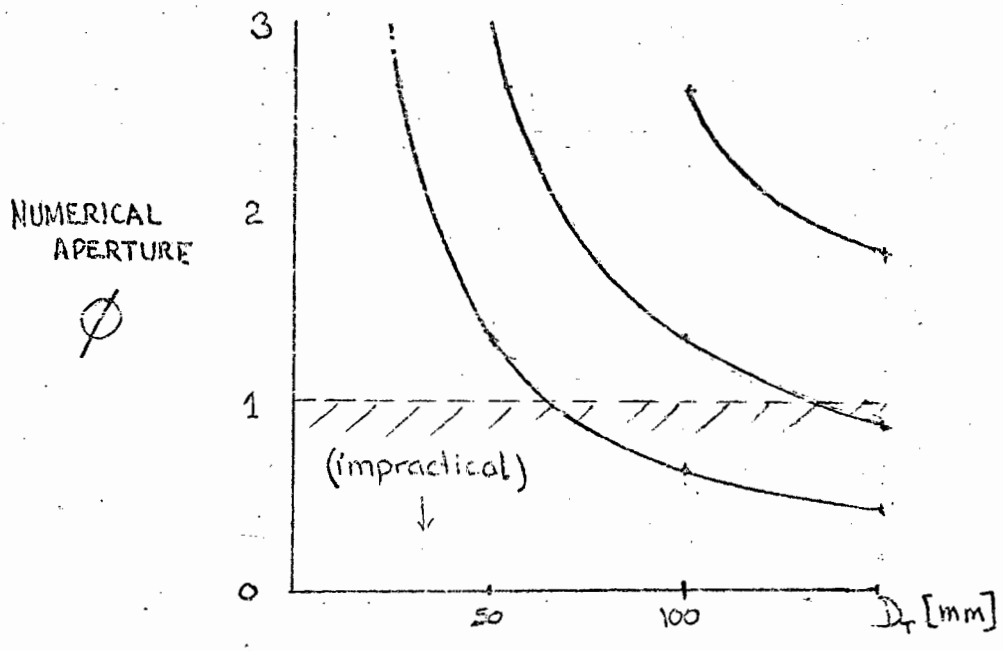
Consider a 100mm multimode step-index fibre with Δn (difference between core and cladding refractive index) of 1%. Using a high radiance Burrus-type diode it is possible to launch a power of about 0,5mW into such a fibre. With a double heterojunction laser as a source this might be increased to perhaps 5mW (45), using a 'selfoc' (46) or cylindrical lens to facilitate coupling. Light-emitting diodes and semiconductor lasers integrally coupled to short lengths of fibre-optic guide are now becoming available as a component in communication systems and their ready commercial availability cannot be far off. Let us estimate the radiance of the free end:

Since only a short length of fibre is needed (perhaps 0,1m) absorption losses are completely negligible. The numerical aperture of the fibre is given by $\sqrt{2n^2\Delta n}$ or 0,21 and the maximum angle that a meridional ray can make with the axis of the guide is arc sin (0,21/1,5) or 8^o. On leaving the guide, the semi-angle of the emerging cone of radiation will be arc sin (1,5 sin 8^o) or 12^o. Provided the numerical aperture of the collimator optics is lower than $(2 \tan 12^o)^{-1}$ or 2,3 substantially all the power launched into the fibre will be collimated into the beam.

Now suppose we employ a 30mm diameter collimating lens. For a 1,5 milliradian beam the focal length must be about 67mm. If we let the focal length be 70mm we have a numerical aperture for the lens of f/2,33 and, as we have seen above, substantially all the power is collimated in a 1,5 milliradian beam. Therefore we have approximately, (from 3.9),

$$I_t = \frac{4 \times (5 \times 10^{-3})}{\pi \times (1,5 \times 10^{-3})^2}$$

$$= 2,8 \times 10^3 \ W/sr$$



N.A. AS A FUNCTION OF SOURCE- & LENS DIAMETERS

(for a 5 arc minute beam)

Thus we see that the semiconductor laser, even when coupled via a fibre-optic guide, offers a potential advantage in terms of transmitted irradiance of between one and two orders of magnitude over incoherent light-emitting diodes radiating directly. At the same time the use of the light guide offers freedom from source-associated errors and flexibility in tailoring beamwidth. The only factors inhibiting its adoption are commercial availability and, possibly, cost, and it is likely that we shall soon see improvements in both.

3.6 COLLIMATION OPTICS

One of the merits of a directly modulatable source is the simplicity of the optical system required; in many cases a single refracting or reflecting element sufficing to collimate the radiation into a narrow beam. However we have seen that I_t is proportional to the area of the optics A_t. Also we have seen that high radiance indicates the use of a small source and this, combined with a minimum required beamwidth, leads to a requirement for good resolution and low numerical aperture optics. The cost of a large-aperture, high resolution, low numerical-aperture optical system can contribute significantly to that of the whole instrument, and it is worthwhile to seek out cost-effective solutions.

We can easily show the relation between source size, optical aperture and beamwidth. If the diameter of the source is d_s and that of the lens d_t, and the focal length of the lens is f, the optical beamwidth is

$$\theta \approx \frac{ds}{f} \quad \text{(radians)} \qquad 3.13$$

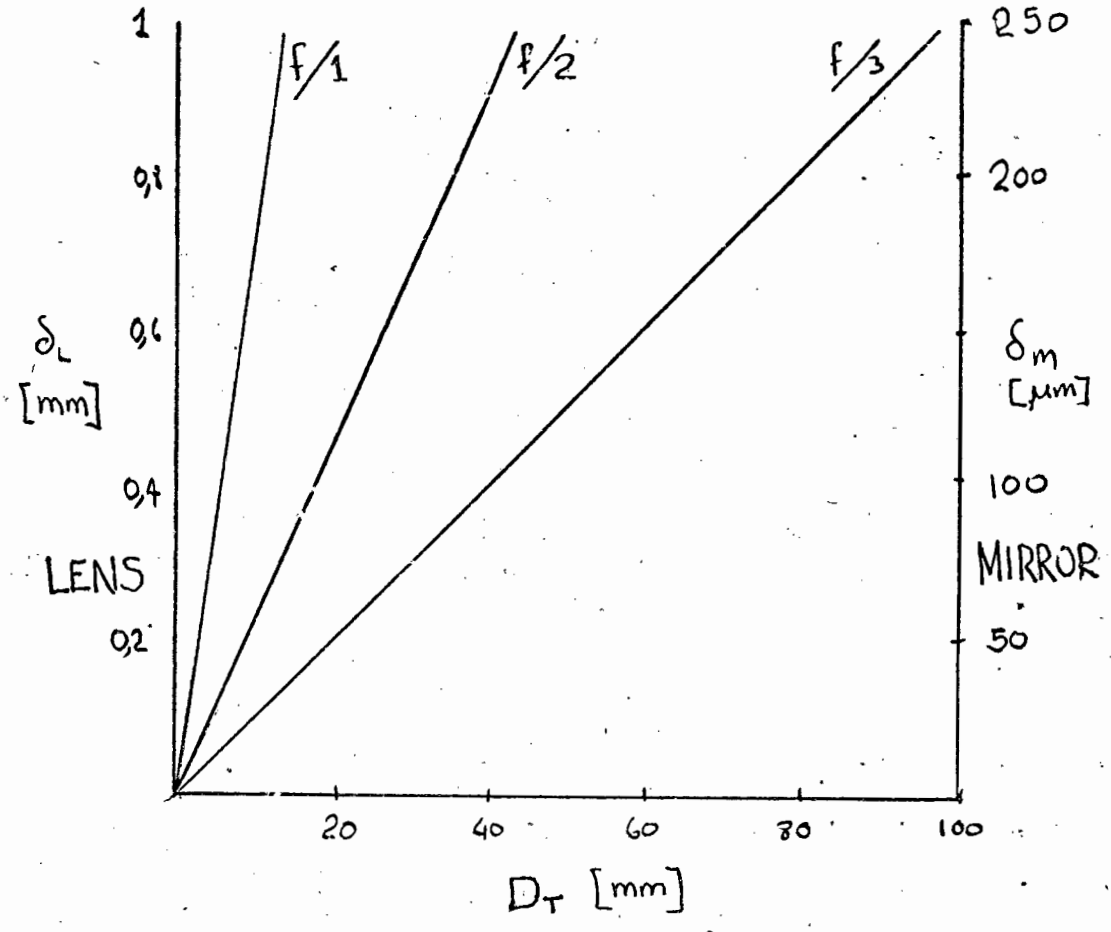
The numerical aperture (or f/number) of the lens is

$$\phi = \frac{f}{d_t} = \frac{ds}{d_t} \cdot \frac{1}{e} \qquad 3.14$$

For a 50µm source and a 50mm lens diameter, the predetermined beamwidth of 1,5 milliradian requires optics of numerical aperture

$$\phi = \frac{50 \times 10^{-6}}{50 \times 10^{-3}} \times \frac{1}{1,5 \times 10^{-3}} = 0,67.$$

The numerical aperture required as a function of optical aperture is shown with source size as a parameter. The beamwidth is assumed to be the optimum 1,5 milliradian or 5 minutes determined in section 3.3.



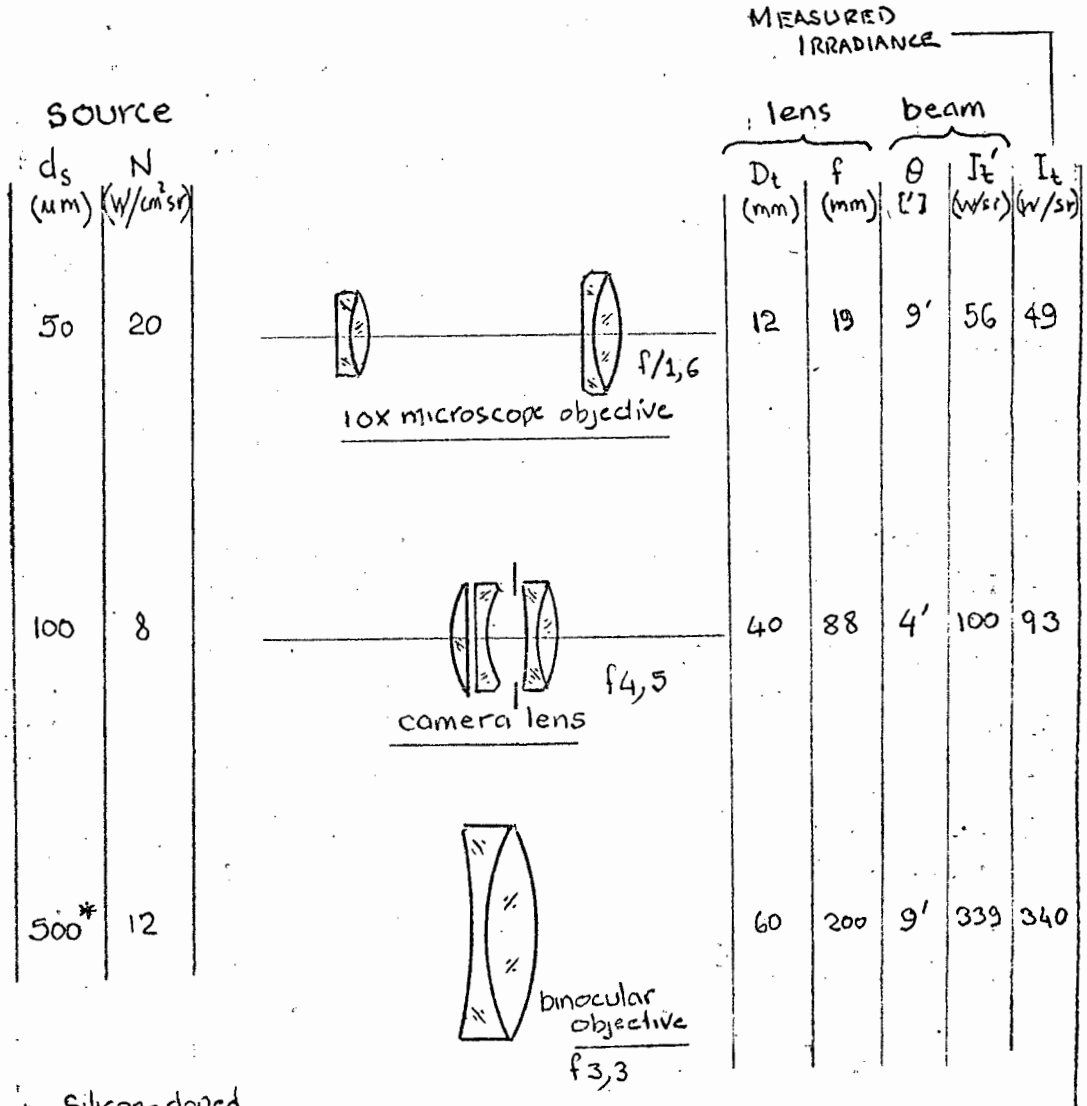
CIRCLE OF LEAST CONFUSION FOR A SIMPLE LENS & MIRROR

Optical resolution

The 1,5 milliradian beamwidth requirement implies that diffraction effects can be ignored, and geometrical considerations suffice in describing optical resolution. Moreover the optical system is required only to image radiation from an on-axis source, eliminating the need to consider off-axis parameters such as coma and field curvature, and the source is monochromatic, making chromatic aberration irrelevant. The only important parameters of the lens are therefore aperture, optical transmission and spherical aberration. The effect of spherical aberration is best described by the geometrical circle of confusion, which is the diameter of the smallest disc into which parallel-incident radiation can be focused. We will make the conservative assumption that energy density within the circle of confusion is uniform.

The relationship between aperture, numerical aperture and diameter of circle of confusion δ , is treated in Appendix 3 and summarised in a diagram for a single lens element optimally configured and for a spherical mirror. Although a mirror has about 4 times less spherical aberration than a simple lens, it is less convenient to use and aperture-blocking and mechanical complications arise from the inevitable path folding required. It is a useful element as part of a co-axial optical system and will be so considered in connection with the collection optics. A particularly cost-effective solution for high resolution low numerical aperture on-axis monochromatic collimation is the Mangin mirror. This is a rear-silvered mirror with non-concentric spherical surfaces, in which the spherical aberration of the mirror is highly compensated by refraction.

More complex optical systems yield vastly greater resolution. If a lens is split into n optimally configured elements, the reduction in the diameter of the circle of confusion is of the order of n^2. Several simple multi-element systems were designed and fabricated using 50mm spectacle blanks. (One of these designs, believed to be novel, exploits the freedom from on-axis aberration exhibited by the aplanatic points of a lens.) The purpose of this work was not to carry out systematic investigation in the specialised field of optical design, but to acquire insight into the cost/performance trade-off factors peculiar to this limited application of monochromatic on-axis collimation, and to acquire, within a limited



PROTOTYPE OPTICAL SYSTEMS

budget, suitable optical systems for experimentation.

The simple optical test methods devised are outlined in Appendix 4. It was found readily possible to make simple optical systems with numerical aperture of 2 or greater and a resolution (measured as the disc encircling 90% of the energy) down to about 100μm. Beyond these limits tolerance on fabrication (particularly centering) became rather critical. Readily available mass-produced optical components were also evaluated, including binocular objectives, and photographic and enlarger lenses.

Optical transmission losses

Because the path through the bulk material of the optical system is short, absorption losses can be totally neglected. There will however be a reflection mismatch loss at each air/glass or glass/air interface. If the glass is uncoated this will amount to

$$\left\{\frac{n-1}{n+1}\right\}^2 \quad \text{i.e. 4\% when } n = 1{,}5 .$$

Coating the surfaces with a $\pi/4$ layer of n = 1,22 will reduce this loss to negligible proportions. In practice it was found that the standard coating optimised for $\lambda = 0{,}55\mu m$ reduced the loss to 2%. Typically the optical path will contain 4 such interfaces and the 2 uncoated surfaces of the plane window on the electroluminescent device. The resulting transmission factor T_o given by

$$T_o = 0{,}98^4 \times 0{,}96^2 \approx 0{,}8 .$$

Prototype systems

The beam irradiance I_t due to several simple low-cost systems is shown opposite. Each lens is matched to an appropriate source. I_t was calculated from the measured radiance of the source and the lens area and estimated transmission, and measured directly using a large area photodetector at a distance greater than 100 times the focal length. Agreement was within 10% in each case.

Of the systems tested, the largest (and also the least expensive) was clearly the most effective despite its poor resolution, and the consequent need for a relatively large-area source. The simple binocular objective lens was used for subsequent work. It was found, however, and verified

experimentally, that a useful shortening of focal length could be achieved with negligible loss of performance by compounding it with a near-aplanatically-configured meniscus lens, which is available at negligible cost as a spectacle-lens blank. The lower f/number that results is no disadvantage with the wide emitting-angle of the TIXL12, and the instrument gains markedly in compactness.

3.7 ATMOSPHERIC PROPAGATION

We have already considered the wavefront-breakup and beam-wander due to local variations in atmospheric refractive index. Now we must consider the attenuation of a beam of light propagating in the atmosphere. The attenuation is due to two effects - Mie scattering by aerosol particles, and absorption. In the near infra-red the latter is almost exclusively due to water vapour (47).

In a uniform atmosphere the fractional power loss from the beam in unit path length is constant. If dW is the power loss from a beam of power W in path length dx we have

$$-dW = W\sigma dx \qquad 3.15$$

where σ may be referred to as the extinction coefficient (48).

Integrating 3.15 we get

$$W = W_t\, e^{-\sigma R} \quad \text{(watts)} \qquad 3.16$$

where W_t is the initial transmitted power (watts) and R is the total length of the atmospheric path (km). σ then has dimensionality km^{-1}. Clearly we can divide 3.16 by the solid angle of the beam Ω_t to obtain the relation between the transmitted beam irradiance and the irradiance after a transmission over a path R:

$$I(R) = I_t e^{-\sigma R} \qquad (W/sr) \qquad 3.17$$

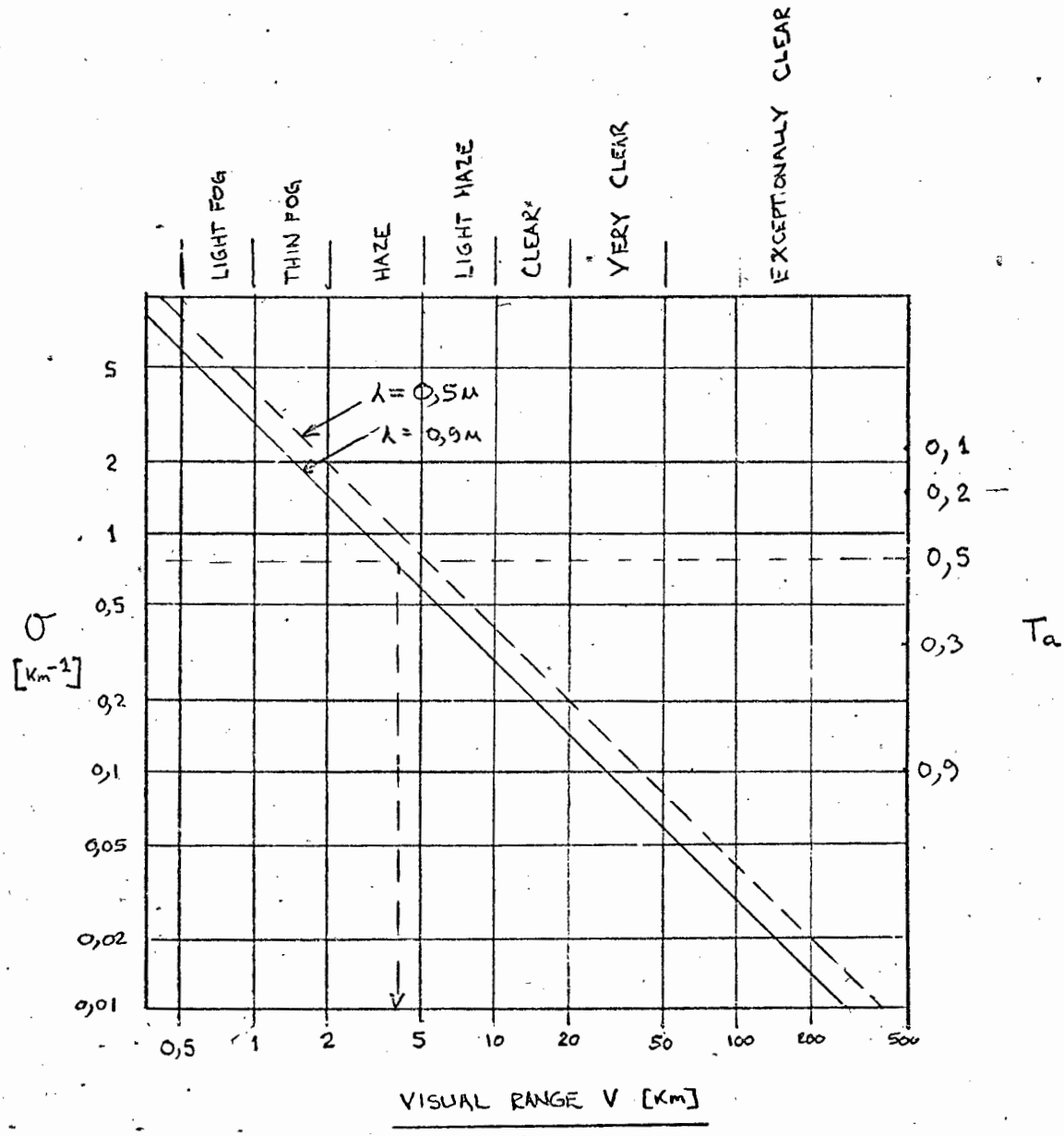
The extinction coefficient σ may be split into two parts, one (σ_s) due to scattering and the other (σ_a) due to absorption (48).

$$\sigma = \sigma_s + \sigma_a$$

therefore

$$I(R) = I_t e^{-R(\sigma_s+\sigma_a)}$$

$$= I_t\, T_s\, T_a \qquad (W/sr) \qquad 3.18$$


LIGHT FOG
THIN FOG
HAZE
LIGHT HAZE
CLEAR
VERY CLEAR
EXCEPTIONALLY CLEAR
λ = 0,5 μ
λ = 0,9 μ
σ [Km⁻¹]
5
2
1
0,5
0,2
0,1
0,05
0,02
0,01
0,1
0,2
0,5
0,3
0,9
Ta
0,5
1
2
5
10
20
50
100
200
500
VISUAL RANGE V [Km]

where T_s and T_a are dimensionless coefficients representing attenuation factors due to scattering and absorption respectively. For 1km of path length, σ and T are related by

$$T = e^{-R\sigma}$$

and

$$\sigma = -\frac{1}{R} \log_e T .$$

Our task now is to relate T_s and T_a to measurable and predictable parameters of the atmosphere.

Scattering

T_s may be calculated in terms of the effective visual range V, which is the accepted measure of atmospheric clarity. V may be expressed rather loosely as the distance at which a large dark object silhouetted against the daytime horizon can just be perceived, or more formally as "the horizontal distance for which the contrast transmission of the atmosphere in daylight (C_R/C_O) is 2% where C_O is the inherent contrast of an object against the horizon sky, and C_R is the apparent or perceived contrast at distance R". V may also be related to generally accepted verbal descriptions of weather conditions, as shown. It can be seen that V = 4km corresponds to an atmosphere free from fog or heavy haze.

The relationship between T_s and V is given by

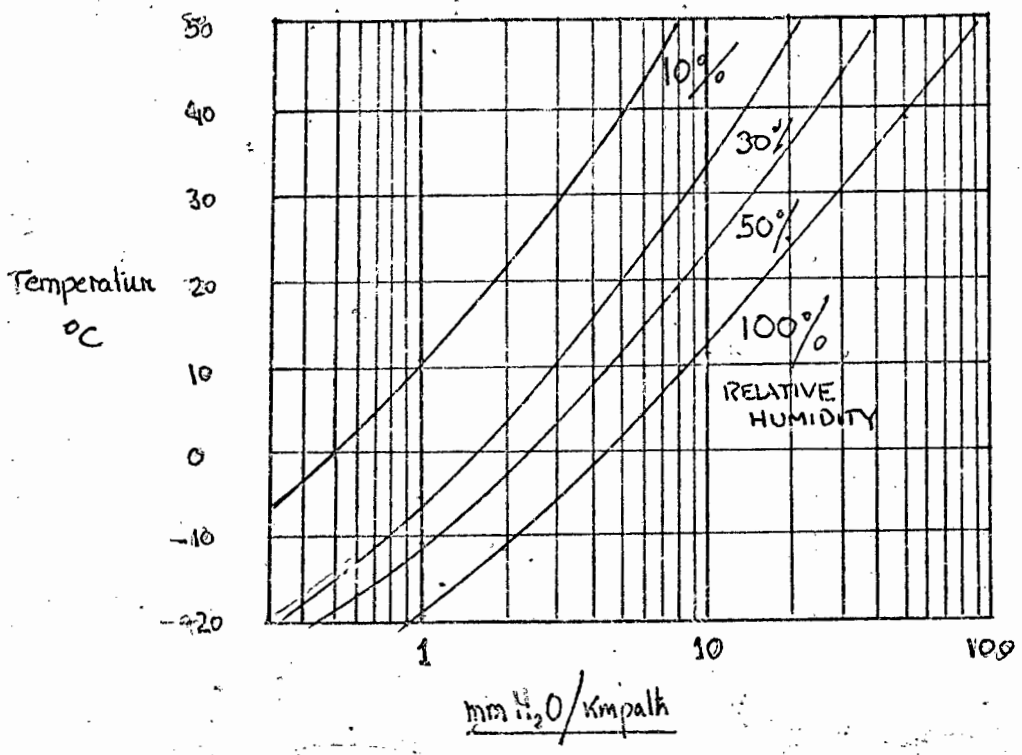
$$T_s = \exp\left\{\frac{-3,9}{V}\left(\frac{0,53}{\lambda}\right)^q\right\} \qquad 3.19$$

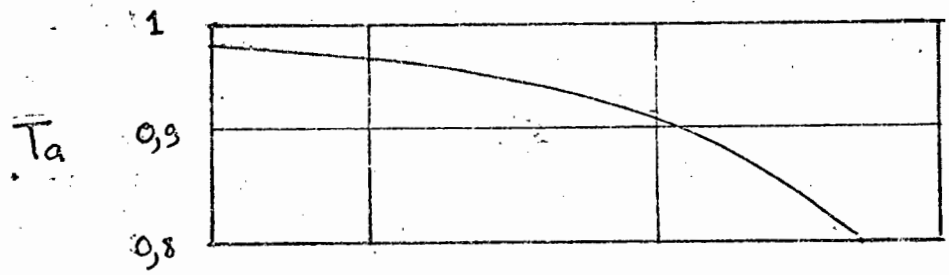
where q = 1,6 in conditions of exceptionally good visibility
1,3 in reasonably good visibility conditions, and
0,6 V when V = 6km (49-51).

T_a has been plotted from 3.19 as a function of V for visible light (0,55μm) and near infra-red (0,9μm). It can be seen that, for the latter, when V = 4km, T_s = 0,5 per km. This is the limiting condition for which the 2 km operational range is required, and pre-supposes the absence of fog or heavy haze. The worst case transmission factor for the return path will of course be T_s^4 or 0,06.

To facilitate comparison with other estimates given in the literature (52;53) it is interesting to note that this corresponds to an attenuation of about 5,6dB/km, or an overall power loss of 22,8dB. This is approximately equal to the

Precipitable H_2O in 1Km path length.





TRANSMISSION COEFFICIENT Ta

attenuation quoted by Ekberg (53) for "excessive rain" and medium snowfall. Most of the treatments in the literature of electro-optic links have limited their attention to rather clearer paths than this, typically assuming, for this type of instrument, an attenuation of no more than 2,5dB/km.

Absorption

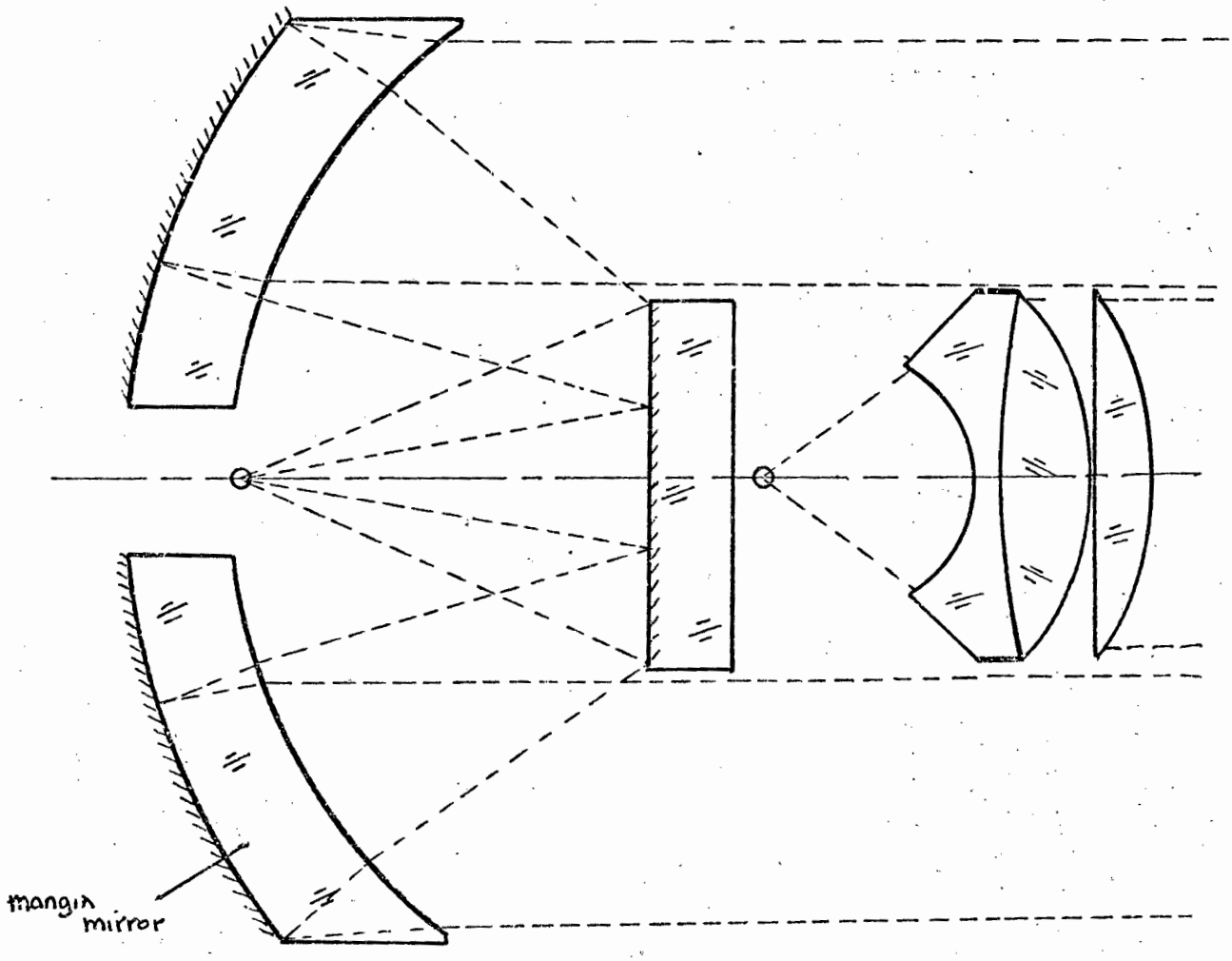
Many of the range predictions in the literature ignore absorption by water vapour (54) but it can be significant, especially in a situation where humidity is likely to be high. A straightforward treatment has been given by Hölscher (51), following Lange et al.

The value of T_a can be related to the number of precipitable millimetres of water, H, in the path length, according to a set of empirical equations developed by Lange. A plot of T_a vs H is given, adapting Hölscher. Lange also gives a means of computing H from readings of temperature and humidity. It can be seen that for t = 25°C and 100%RH, a 2km path will contain 50mm of precipitable water, corresponding to an overall transmission factor T_a of 0,8. This is in agreement with data taken from other sources (55).

3.8 COLLECTION OPTICS

The flux returned by the retroreflector must be collected by an optical system and focused onto the relatively small area photodetector. In order to determine the efficiency with which the radiation is collected it is necessary to consider the geometrical arrangement of the collimation (transmit) and collection (receive) apertures and the way in which the flux returned by the reflector is spatially distributed.

Returning to the simple model of retroreflector behaviour as that of an aperture (see diagram in 3.2), we see that the distribution of flux at the instrument due to a point radiator would be a uniformly illuminated disc with a diameter twice that of the reflector aperture. The radiating aperture of the collimator can be considered to be a continuous array of such point radiators. The nett flux distribution is therefore given by the superposition of the discs due to the array of virtual point sources. This two-dimensional function is a disc having a diameter $(d_t + 2d_p)$ where d_t and d_p



RELATIVELY LOW-COST HIGH-PERFORMANCE SYSTEM.
(150mm mangin mirror can resolve a 100M source.)

are the diameters of the transmit aperture and reflecting prism respectively. The form of the distribution is roughly conical, being given by the two dimensional convolution of the transmit and reflector aperture functions. Obviously, at close range the reflector aperture should be as large as possible and the collector aperture as near as possible to that of the collimator for good coupling efficiency. The optimum arrangement is a coaxial optical system. The collection optics can take the form of a folded optical system using, say, a Mangin mirror (see illustration). This configuration is particularly effective in that the inevitable aperture blocking associated with folded optical systems need not, in this case, occasion any additional loss. Obviously since the direction of each light ray is reversible, the roles of the lens and mirror (transmission and collection) can be interchanged.

The chief drawback of the coaxial optical system is its cost. The cost of optical elements is a strong function of this aperture and additional expenditure is involved in the mechanical complexity occasioned by the folded path. A much cheaper and simpler system (and one adopted in a number of commercial survey instruments such as the AGA12, Wild DK9, Tellurometer CD6 etc) is to use identical lenses for collimation and collection in a side-by-side binocular-like arrangement. This is less than optimum from the point of view of coupling and the usual remedy is to use an effectively larger reflector prism. The increased aperture is only required in the horizontal dimension (assuming that the transmitter and receiver are horizontally adjacent) and to save cost and reduce the bulk and weight of the prism it is suitably truncated. Nevertheless the use of these 'rectangular'prisms incurs an appreciable cost penalty.

In practice the problem hardly arises in the present instance. At short range the loss of signal due to poor coupling is no problem - it can be an advantage, reducing somewhat the dynamic range with which the instrument has to cope. At long range the additional spread of the returning beam due to prism imperfections gives rise to a fairly uniform flux distribution. At distances in excess of 1 km it was found unnecessary to take account of coupling losses.

It was decided therefore to use the cheapest and simplest system possible for experimental and prototyping purposes, in the form of two 60mm diameter binocular objectives mounted side by side. The effective collection aperture was therefore

$$A_R = \frac{\pi}{4} (60 \times 10^{-3})^2 = 2,8 \times 10^{-3} (m^2) .$$

The design problem posed by the collection optics is similar to that of the collimation optics and most of the points made in 3.6 apply here also. In one respect however an important constraint is released. In the case of the source, high radiance could be achieved at the expense of small size, and good resolution was required of the collimating optics. In the case of the receiver, little is gained by going to a very small area photodetector. The field of view is restricted and as a result the background radiation reduced but as we shall see in the following chapter, the noise advantage is not very great. Small area photodetectors have less capacitance but this is not critical where, as here, modulation frequencies are low. Moreover, for detectors of less than 0,5 - 1mm diameter, the capacitance is dominated by amplifier input capacitance and circuit strays. Finally, the problem of maintaining 'boresight' collimation of the transmit and receive optical systems is greatly eased if the receiver field of view is appreciably wider than the transmitted beam, and the dominance of the latter in determining pointing effectively precludes the possibility of error due to spatial variations in photodetector speed of response (see Chapter 4). It can therefore be assumed that the photodetector will have a diameter of at least 0,5mm and the resolution of the collection optics need be no greater than, say, 0,3mm for 90% encircled energy.

3.9 PREDICTED PERFORMANCE

We are now in a position to predict the worst-case returned power for our prototype setup, which consists of a source having $N = 12 W/cm^2 sr$ collimated by a lens of 60mm diameter and 0,8 optical transmission.

Therefore

$$I_t = N A_t$$
$$= 12 \times 10^4 \times \frac{\pi}{4} (60 \times 10^{-3})^2$$
$$= 340 \quad (W/sr).$$

Modifying 3.6 to account for transmission loss, we have, at a range of 2km when V = 4km, a returned power W_R where

$$W_R = 1,2 \times 10^{-7} \; I_t \; A_R \times T$$

when $T = T_s \; T_a \; T_p \; T_o$
$$= 0,06 \times 0,8 \times 0,8 \times 0,8$$
$$= 0,03$$

$\therefore$ $W_R = 1,2 \times 10^{-7} \times 340 \times 2,8 \times 10^{-3} \times 0,03$
$$= 3,4 \times 10^{-9} \; W \qquad 3.20$$

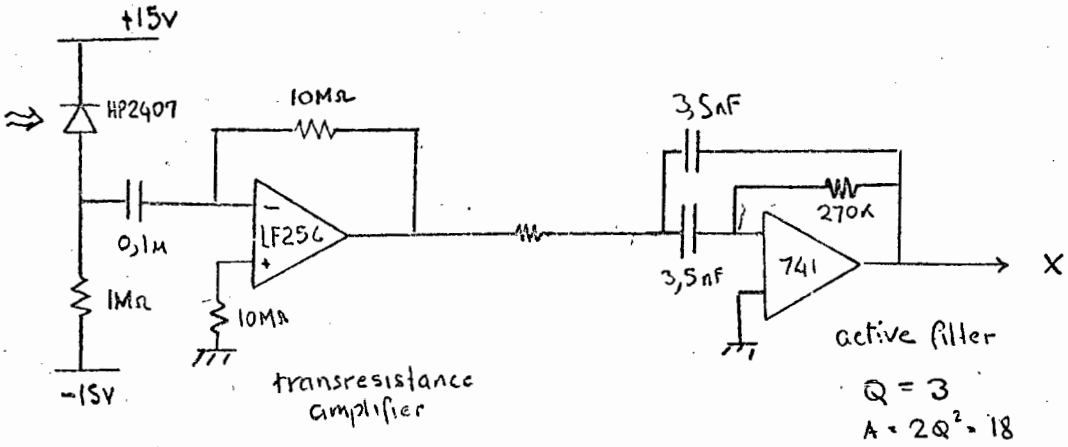
In the following chapter we shall consider whether a returned optical power of 3,4nW is sufficient to permit the measurement of phase to the required accuracy.

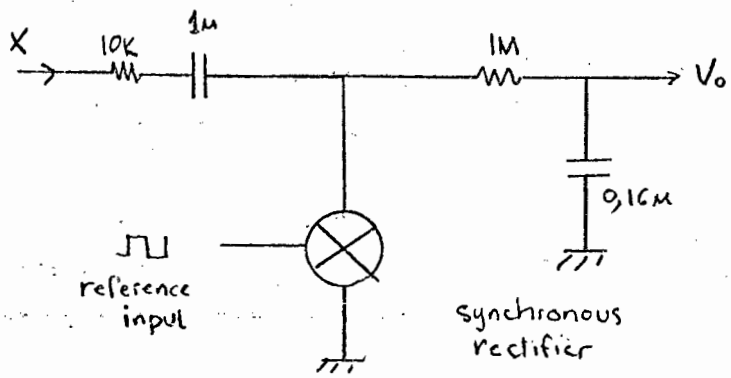
3.10 EXPERIMENTAL VERIFICATION

The foregoing range prediction clearly is based on a rather long chain of reasoning involving many assumptions. Although doubtful points were checked experimentally (e.g. diode radiance, reflector performance) or by means of numerous intercomparisons of published data (atmospheric transmission), it was judged prudent to carry out an experiment modelling as closely as possible the proposed system over the actual terrain envisaged. Although it would not be possible to verify the theoretical model with a high degree of precision (due chiefly to the variability of, and difficulty of specifying, atmospheric conditions), gross errors could be discovered by such an experimental setup, and the practical viability of the system convincingly demonstrated.

Experimental apparatus

The experimental work was carried out using the apparatus shown in the accompanying diagrams. A silicon doped (domed) GaAs source (TIXL12) with a measured radiance of $12W/cm^2sr$ was square-wave modulated at a frequency of 1kHz - sufficiently high to avoid atmosphere scintillation noise (e.g. sunlight





RECEIVE CHANNEL OF OPTICAL TRANSMISSION EXPT.

glinting off a choppy sea) and 1/f noise. Radiation was collimated by a 60mm diameter cemented doublet lens at f/4, yielding a beamwidth of 7 arc minutes, slightly in excess of the optimal 5 minutes. Reflection was accomplished by a 60mm diameter tested prism having a nominal 3 arc second deviation. Returning radiation was focused by a 60mm doublet identical to that in the transmission path, onto the sensitive area of a 1mm diameter silicon PIN detector. The resulting photocurrent was amplified by a FET operational amplifier in transresistance mode (current-to-voltage converter) and the output of the latter was synchronously detected by means of a shunt CMOS switch (56). The output of the synchronous detector was further amplified to drive a meter.

Calibration

The measurement of beam radiant intensity I_t has already been dealt with. It remained to determine the sensitivity and gain of the receive channel.

Manufacturers' data for the photodiode used - the HP4207 - quotes a responsivity of 0,33A/W at a wavelength of 0,9µm. A direct comparison of a number of such diodes with a bolometer-type radiometer (57) showed responsivities ranging from 0,3 to 0,4A/W. The device used in the experimental setup was close to the nominal 0,33A/W. The peak-to-peak photocurrent for square-wave modulated flux W_R (watts) is therefore $0,33W_R$ (amps).

The transimpedance amplifier used an LF356 FET operational amplifier with a feedback resistor of 10MΩ. The output of this amplifier was therefore a square wave of amplitude $3,3 \times 10^6 W_R$ (volts).

Because of the wide bandwidth at this point, the square-wave, for weak but perfectly usable signals, is completely buried in noise. In order to ease the linearity and dynamic range requirements of the synchronous detector switch, the transresistance amplifier was followed by a low-Q multiple-feedback active filter tuned to the fundamental of the square-wave drive to the light emitting diode (1kH). Limiting the Q of this filter to 3 minimises the likelihood of the synchronous detection sensitivity being altered due to phase-shift.

The effect of rejecting the harmonics of the received signal modifies the gain of the channel so that

$$V = A \times \frac{1}{2\sqrt{2}} \times \frac{4}{\pi} \times 3,3 \times 10^6 \, W_R \text{ (volts)}$$
$$= 2,7 \times 10^7 \, W_R \text{ (volts)} \qquad 3.21$$

where V = rms voltage at output of filter
A = filter voltage gain = $-2Q^2 = -18$.

For the limiting case derived in 3.20,

$$W_R = 3,4 \times 10^{-9} \text{ watts}$$

A rough check on the functioning of the instrument was obtained by using the source without collimating optics, facing the photodiode, also without optics, at a distance d (metres). The irradiance of the source can be calculated from the source radiance N and source diameter d_s (0,5mm).

$$I_o = N \frac{\pi}{4} (d_s)^2$$
$$= 12 \times 10^4 \times \frac{\pi}{4} \times (0,5 \times 10^{-3})^2$$
$$= 23,5 \times 10^{-3} \text{ W/sr} .$$

This agrees reasonably well with a measured 20mW/sr.

The solid angle Ω subtended by a 1mm photodiode is given by

$$\Omega = \frac{\pi}{4} \cdot 10^{-6} \times \frac{1}{d^2}$$

and the flux impinging on the photodiode is

$$W_R = \Omega \times I_o$$
$$= \frac{\pi}{4} \times 10^{-6} \times \frac{1}{d^2} \times 20 \times 10^{-3}$$
$$= \frac{1,6 \times 10^{-8}}{d^2} \text{ watts} .$$

Equating with 3.20, we find the distance d which, with no optics, yields the designed minimum power,

$$d = \sqrt{\frac{1,6 \times 10^{-8}}{3,4 \times 10^{-9}}} = 2,2\text{m} .$$

When the transmitter and receiver were set up facing one another without optics and care was taken to eliminate stray reflections, the received signal resulted in an output from the active filter of $0,15V_{rms}$, within 16% of the predicted value. Given the indirectness of the calculation and the

number of assumptions involved, this is held to be an acceptable order of agreement.

For subsequent tests the receiver was calibrated in terms of a bolometer radiometer standard. The post-detection gain was set to give a d.c. output of exactly 0,1V per nanowatt incident on the photodiode, using the inverse square law of distance to cross refer the two systems over the wide dynamic range involved. The calibration procedure is self explanatory. The validity of the use of inverse square law (provided adequate baffling is used to eliminate reflections) is shown by the logarithmic plot of power vs distance.

Experimental results

The instrument was reassembled with its optics and carefully collimated at a range of 1km. The signal received over 2km range was recorded over various types of terrain (including water) and in various weather conditions. As expected, the received power varied over a wide range - from <1 to 12nW. However under reasonably clear conditions (free from heavy haze or fog), when the visual range was judged to be at least 4km, the returned signal was never less than $0,1V_{rms}$, which corresponds to 4nW (eqn 3.21)(cf 3,4 nW; eqn 3.20).

Conclusion

The agreement between the predicted and measured performance of the electro-optic link, using a low cost and readily available system, was judged entirely satisfactory. The prototype optical system, though perhaps not optimum, was shown to achieve the required performance. There is no doubt that the performance can be further improved if fibre-coupled lasing sources are used.

CHAPTER 4

4.1 INTRODUCTION

This section deals with the theory and design of an electro-optic phase-comparison distance measuring system utilising a passive co-operative reflector and having an output in the form of the phase angle between a pair of low-frequency signals. A treatment of the historical development, principles of operation and parameters determing the performance of such instruments will be followed by design information for a unit to meet the present requirement.

The operating principles of such instruments are now well-established and it is easily possible to design and build an instrument meeting the requirement in respect of range and accuracy with a high degree of confidence that it will meet its target specification. The problem lies in doing so within the constraint of low cost. The operating constraints of such instruments tend to lead to relatively expensive components, fabrication and adjustment procedures, a fact which is borne out by observing the high cost of commercial electro-optic distance measuring instruments. The situation is considerably eased in the case of the present instrument as a result of the significantly reduced accuracy requirement. On the other hand, it is exacerbated by the more stringent demands in terms of operational range and operating conditions. The writer's experience with the design of comparable commercial equipment suggests that, in order to meet the present requirement of low cost, a careful study of critical error sources and cost-effective techniques for their elimination is required.

The evolution of instrumental technique for electro-optic phase-comparison distance measuring instruments follows a remarkably coherent pattern of development over many decades, each advance being characterised by increased phase resolution and/or immunity from systematic error. It is helpful to place the present state-of-the-art in perspective by a brief historically-based review (58-60). Articles based on the following section have been published by the writer. Similar reviews have been published (61) but this appears to be the first to take cognizance of both Western and Soviet streams of development.

4.2 ELECTRO-OPTIC INSTRUMENTS - EVOLUTION OF EXPERIMENTAL TECHNIQUE

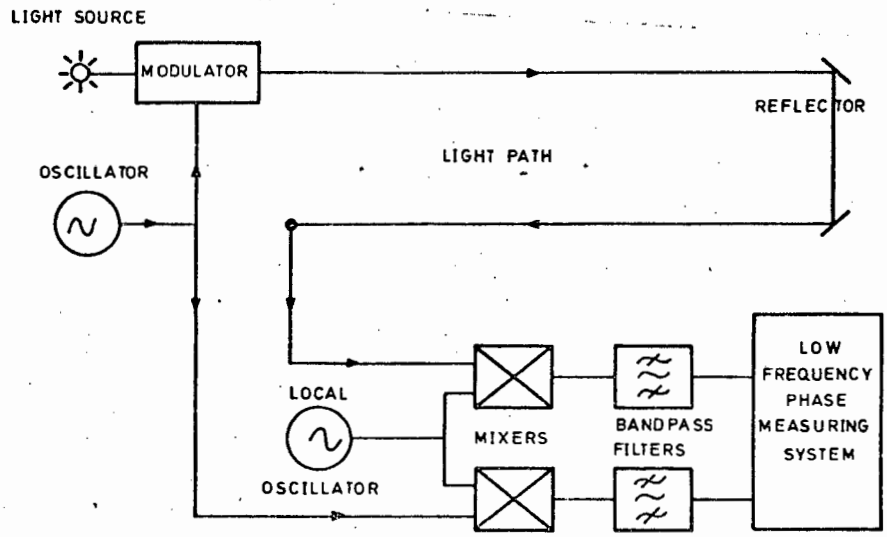
The development of electro-optic instruments for the determination of distance can be thought of as a continuous process beginning with a series of experiments by Karolus in the years 1925-1940 to determine the terrestrial velocity of light (62). Karolus and Mittelstaedt began by adapting the classical method of Fizeau, substituting for the toothed wheel an electro-optic modulator capable of modulating the intensity of the light at frequencies up to 10 MHz. As in the case of the toothed wheel the modulator functioned also as a synchronous shutter or phase-sensitive detector of the returning radiation, and the modulating frequency was varied to achieve a visual null. Subsequent workers improved the precision of the null detection by substituting a photodetector for the visual observation, and in 1935 Karolus proposed that, instead of using an electro-optic modulator to gate the returning radiation, the gain of the photodetector be modulated synchronously at the modulation frequency. In the case of a photomultiplier this can be achieved by applying a fraction of the alternating voltage across the modulator to the dynode structure of the photomultiplier.

A shortcoming of simple null schemes is that, at the critical null point, the signal disappears "into the noise", resulting in a rather broad and ill-defined null point, and better results can frequently be obtained in such a situation if a differential measurement can be substituted in which the amplitudes of two nearly equal signals can be compared, often by using synchronous or phase-sensitive detection. Such a refinement was proposed by Karolus in 1938 (35), and consists of a system in which the phase of the alternating voltage applied to the photomultiplier is periodically reversed by 180° at a low rate (say 50 Hz). In this case the mean photocurrent is unaffected by the reversal if and only if the light modulation envelope and the alternating voltage applied to the photomultiplier are in mutual phase quadrature. Thus the quadrature condition is sharply indicated by the disappearance of a photocurrent varying at the phase reversal rate.

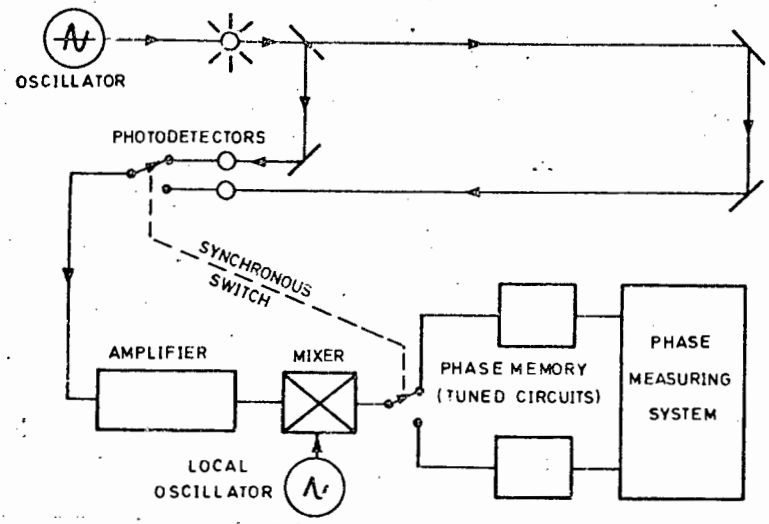
This led directly to the development, in the years 1940-1949, of the first effective portable electro-optic instrument designed specifically for geodetic distance measurement by E. Bergstrand (63), who was responsible for many notable advances in the state of the art. His instrument, which later became commercially available as the AGA Geodimeter used essentially the scheme proposed by Karolus in 1938, but the 180° phase reversal was elegantly accomplished by a periodic alteration of the bias voltage of the modulator, utilising its quadratic characteristic. Moreover the modulation frequency was fixed at 8 MHz, the measurement of arbitrary distance being accomplished by a calibrated variable delay line inserted between the modulation generator and the photodetector.

Despite Bergstrand's major contribution to the art of electro-optic distance measurement, and the fact that his was undoubtedly the first successful instrument commercially produced for the purpose, its claim to priority is not absolute. According to a comprehensive review by Andrianova et al (64), such an instrument was in existence at the Soviet State Optical Institute as early as 1936 but the review cites no source prior to 1956. The past and present extent of Russian involvement in electro-optic distance measurement is frequently overlooked in the West. Conversely, western sources are rarely cited in the extensive Russian literature on the subject. The first explicit disclosure of an electro-optic distance measuring instrument occurs in a United States patent granted to I. Woolf, an employee of RCA, in 1940. Woolf's instrument, which does not seem to have found commercial exploitation, anticipates many of the features of later instruments, such as phase detection in an electronic multiplier following photodetection, as well as fixed-frequency modulation together with Bergstrand's calibrated phase delay for the measurement of arbitrary distances.

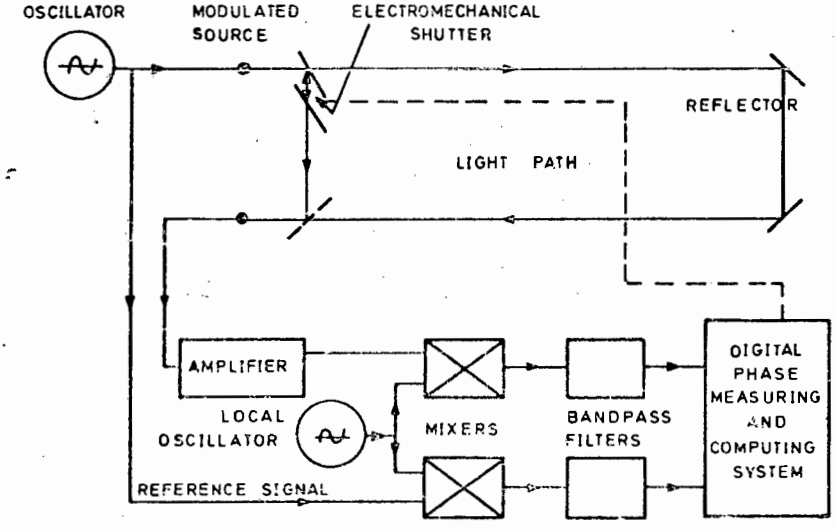
All the above may be described as homodyne techniques in that returning light signal is multiplicatively mixed, whether in an electro-optic modulator, photomultiplier or electronic multiplier, with a signal at the same frequency, resulting in a zero-frequency output which is nulled by an adjustment of the mutual phase of the two signals. This technique reached a



BJERHAMMER'S HETERODYNE METHOD.



HÖLSCHER'S MULTIPLEXING TECHNIQUE



GREENE'S VERSION

peak in 1967 with the introduction by Froome and Bradsell of the National Physical Laboratory, of the very accurate Mekometer (65), which represents a highly ingenious and elegant adaptation of the basic method of Karolus and Bergstrand. In the Mekometer, the light source is a pulsed Xenon flash tube, elliptically modulated by means of a crystal cavity. After traversing the distant path, the returning light passes through a variable light path, and is de-modulated by means of a second identical crystal mounted in the same cavity. The variable light path is adjusted for a null condition, permitting distance determination with sub-millimetre precision.

Heterodyne methods

The success of the Mekometer notwithstanding, the mainstream development which was to lead to the current widespread use of electro-optic distance measuring instruments did not involve the homodyne approach, chiefly on account of the difficulty of generating precisely known phase delays in the megahertz range. In the Mekometer this problem was circumvented by the use of a very high modulation frequency which permitted the use of an optical delay line (variable light path) of reasonable size.

In 1954, A. Bjerhammer, at the Royal Institute of Technology, Stockholm, introduced a technique (66) which was adopted by all successive instruments with the exception of the Mekometer, and the Russian GD314. Confronted by the problem of measuring the phase difference of high frequency signals representing the outgoing and returning light modulation, his solution was to mix the signals multiplicatively with a local oscillator, and to select the lower sideband, or difference frequency, which results. The process preserves the phase difference between the signals in the form of an identical phase angle between their low frequency counterparts. In effect, this amounts to a time expansion which greatly increases the achievable resolution.

Hölscher, in 1965, introduced a new technique which permitted electronic amplification of the returned signal before mixing down to the lower frequency at which phase is measured (67). The problem to be overcome involved the introduction of errors due to spurious phase delays in the high frequency amplifier following the photodetector. Delays due to this source would be indistinguishable from those due to the transit time of light. Hölscher solved the problem by including a fixed light path internal to the instrument, and time-multiplexing the signals resulting from the internal and external light paths through a common electrical channel. The phase lag due to the transit time then appears differentially. Electronic delays appear as a common mode term, and error on this account is eliminated.

In 1968, the present writer was engaged in engineering Hölscher's instrument for commercial production and decided to take his multiplexing technique one stage further. In order to eliminate the possibility of phase drift due to the use of two photodetectors in tuned amplifiers, an electromagnetically-operated optical switch was used to direct light from the internal and external paths alternately on a single photodetector, thus eliminating phase errors due to the detector itself. This approach is now almost universally adopted in the commercial electro-optic distance measuring instruments which have proliferated in the past decade. Much of this chapter draws on well-established techniques. The writer's contribution is also now an accepted part of the state-of-the-art.

4.3 RESOLUTION OF AMBIGUITIES

If we regard the modulated light beam as creating a set of fixed (equi-phase) reference points in space, it will be seen that the phase measurement serves the purpose of interpolating linearly between these points. If, however, the distance to be measured exceeds half the modulation wavelength , resolution of the total distance requires the determination of the number of complete cycles of phase angle along the path. Several strategies exist for determining this figure, but all of them involve changing the modulation frequency f_m in some way.

(It would, of course, be possible to determine the total distance by means of some independent method such as a pulse time-of-flight measurement, and to use the phase method to improve the precision of this "coarse" determination. Such an approach would however be somewhat inelegant, requiring extra circuitry due to the non-compatability of pulse and phase techniques.)

There are three principal approaches to ambiguity resolution:

1. Low-frequency modulation
   We may measure phase with a low modulation frequency chosen such that the unambiguous range exceeds the desired maximum range of the instrument.

2. Differential modulation frequencies

   We may measure phase at two slightly differing modulation frequencies. The result of subtracting the two phase delays corresponds to the phase delay which would have resulted had we modulated at a frequency equal to the difference between the two modulation frequencies.

   That is, it can be shown that if ϕ_1 and ϕ_2 are the phase angles obtained when measuring with modulation frequencies f_1 and f_2 respectively, $(\phi_1 - \phi_2)$ is the phase lag which would be obtained if the modulation frequency were $(f_1 - f_2)$.

   The practical significance of this technique is that the effect of very low modulation frequencies, and hence long unambiguous range, can be obtained without incurring the circuit complications involved in handling widely-spaced modulation frequencies.

3. Frequency modulation of the modulating frequency.

   While methods 1 and 2 are universally used in distance measuring equipment and are eminently suited to situations where the ratio between unambiguous range and instrumental resolution is large (typically 10^5), another possible method suggests itself where the ratio is smaller (e.g. 3×10^2 as in the present case).

We can see from 2 that deviation of the modulating frequency gives rise to a phase shift which is proportional to the amount of the frequency deviation and the total distance. Thus if the modulating frequency is itself frequency modulated by a known amount, the phase modulation of the returning signal will be a measure of the total distance.

The concepts discussed qualitatively above will be treated more formally in the following section.

Principles of operation

Let:

L be the distance to be measured
c the velocity of light in vacuo
n the atmospheric refractive index
t the time delay corresponding to the return path
f_m the modulating frequency
T_m the modulation period $= \frac{1}{f_m}$
ϕ the phase lag between the modulation envelope of the outgoing and returning radiation
v the atmospheric velocity of light

$$v = \frac{c}{n} \tag{4.1}$$

$$t = \frac{2L}{v}$$

$$= \frac{2Ln}{c} \tag{4.2}$$

$$\phi = 2\pi \frac{t}{T}$$

$$= \frac{4\pi n}{c} f_m L \tag{4.3}$$

It can be seen that the phase increases linearly with L. However since ϕ' is indistinguishable from $\phi + 2\pi n$, it can be seen that there will be a certain distance L' after which the phase will effectively jump from 2π to 0 and recycle. L' is the unambiguous range and is given by:

whence $L' = \frac{c}{2f_m n}$ 4.4

or $f_m = \frac{c}{2L'n}$ 4.5

Resolution:
from 4.3 and 4.4:

$$\phi = 2\pi \frac{L}{L'} \quad 4.6$$

or $$L = L' \frac{\phi}{2\pi} \quad 4.7$$

and $$\frac{d\phi}{dL} = \frac{2\pi}{L'}$$

$$\therefore \quad d\phi = 2\pi \frac{dL}{L'} \quad 4.8$$

Ambiguity resolution:
Let f_m successively be f_1 and f_2 and let the corresponding phase lags be ϕ_1 and ϕ_2.

From 4.3:

$$(\phi_1 - \phi_2) = \frac{4\pi n}{c} (f_1 - f_2) L \quad 4.9$$

in the limit,

$$d\phi = \frac{4\pi n}{c} df\, L \quad 4.10$$

or $$L = \frac{c}{4\pi n} \frac{d\phi}{df} . \quad 4.11$$

From 4.9 it can be seen that the effect of modulating with a low frequency can be obtained by differencing the phase delay obtained from two closely-spaced modulation frequencies. As was previously mentioned, this offers practical advantages in terms of circuit simplicity, and has been widely used. It appears superior to the use of direct modulation frequencies in the present instrument and would probably be adopted, but for the existence of an even simpler method which exploits the fact that, in the present case, an unusually small ratio between modulation frequencies is required.

The method depends on 4.11 from which it can be seen that $\frac{d\phi}{df}$ is a monotonic (and hence unambiguous) function of the total range L. Thus, if the modulating frequency is periodically perturbed a fixed amount (i.e. frequency modulated), the total range can be inferred by examining the amplitude of the resulting periodic phase deviation in the returning signal. The method can be thought of as a crude implementation of the difference-frequency methods in which the modulation frequency is rapidly and repetitively switched between the two relevant frequencies. Since the required difference effect is measured directly, a simple and crude determination will suffice, and the method offers further simplification and reduction of measuring time.

4.4 CHOICE OF MODULATION FREQUENCY

Two conflicting requirements dictate the choice of a modulation frequency: unambiguous range and resolution or accuracy. These respectively set upper and lower bounds on the frequency, and normally, in a general-purpose distance measuring instrument, the demands they make are incompatible, leading to the multiple-frequency arrangements discussed above. In the case of the present instrument, accuracy demands are moderate and it is worth considering whether a single-frequency instrument is a possibility.

Unambiguous range of 2km implies (from eqn 4.4) a modulation frequency no higher than 75kHz. Equation 4.8 shows that then, to achieve 0,3m accuracy, we would require a precision of phase measurement of at least 1,5 parts in 10^4. In the laboratory (we shall see in Chapter 6) there is no great difficulty in achieving it, but this order of precision is by no means easy to achieve in a distance measuring link in which the signal is subject to deep fading and wide dynamic range variations. The difficulties encountered in seeking very high phase resolution in distance measuring equipment are considered in detail in 4.7, but the fact that commercial instruments rarely expect (or achieve) a phase resolution precision better than 1 part in 3×10^3 is enough to make us cautious.

To achieve the required precision of 0,3m with 1 part in 3×10^3 phase resolution, we need a modulation frequency of 166,6kHz. For convenience we would choose 150kHz, giving a modulation wavelength of 1km. Ambiguity resolution is then a trivial problem with a fine/coarse frequency-ratio of only 2. Indeed for many purposes the coarse frequency could be dispensed with, since it would not be unreasonable to assume in the user an a priori knowledge of whether the vessel were, say, 100 or 1100m distant. Thus a single-modulation-frequency 150kHz instrument, although not strictly meeting the ambiguity requirement, is attractive in its simplicity.

It is by no means obvious that this represents the best design strategy - the problem of achieving the required accuracy would be greatly eased by adopting two modulation frequencies at, say, 75kHz and 750kHz. This could best be realised in the form of difference frequencies at (to yield exact metre increments) the following frequencies:

fine frequency	749 925Hz
coarse frequency	674 932,5Hz
difference	74 992,5Hz

At a modulation frequency of around 750kHz the phase accuracy requirement becomes an easy 1,5 parts in 10^3, very comfortably within normal commercial practice.

Although this design approach would have many advantages in commercial manufacture, due to its non-critical demands, it was not the approach adopted in the experimental work for this thesis, for a number of reasons. The writer had considerable experience with existing commercial instruments in this category, and the design of yet another similar one offered little in the way of challenge or potential new information. Nor was experimental work really required for a design so comfortably within the state-of-the art. On the other hand, it was felt that experimental work was required to settle a number of uncertainties. It was decided to see whether the specification could be met (except in respect of unambiguous range) by an instrument operating at a single modulation frequency of 150kHz. This would provide a searching test of the theoretical models employed and of the ultimate achievable accuracy. To have any

expectation of meeting the accuracy specification, comprehensive exploitation of error-cancelling techniques would be needed. Such a technique (affecting the modulation frequencies) is dealt with in the next section.

The use of "negative patterns"

In the case of a microwave distance measuring instrument, an active transponding instrument is used and there is no possibility of introducing the 'reference path' so easily provided in the case of an electro-optic instrument on account of the co-location of transmitter and receiver. As a result of this, phase-shift in the low-pass amplifiers following the heterodyne operation contributes directly to the measured phase shift, and instabilities in the electronic delay are a potential source of error. Wadley, in his "Tellurometer" instrument proposed an ingenious way of eliminating this source of error by the introduction of what he termed a "negative pattern" (68). In order to understand this terminology it is necessary to know that the "patterns" in the Tellurometer system are simply the set of equiphase points in space caused by the modulation frequencies. Normally the patterns are "positive". That is to say, the system is configured so that increasing distance results in an increase of measured phase shift. It is easily shown (see appendix 4.2) that for this to be the case, all that is required is that the reference (heterodyne) oscillator be below the modulation frequency, in which case the phase shift ϕ between the incoming frequencies will be reproduced as an identical phase shift ϕ at the low (or intermediate) frequency. If, on the other hand, the reference oscillator be located above the modulation frequency, the output phase shift corresponding to an input phase shift of ϕ, is $-\phi$.

If we subtract the phase-shift measured with the reference oscillator located below the modulation frequency from that obtained with the oscillator above the modulation frequency we shall obtain a result

$$\phi - (-\phi), \text{ or } 2\phi$$

Thus the resolution is doubled. But this is not the most important result. Suppose that the low-pass filtering and amplification in the two channels, following the mixing operation, introduce a differential phase shift. Without the use of a negative pattern the measurement would be perturbed by

δ yielding a measured phase-shift ϕ' where

$$\phi' = \phi + \delta$$

However δ is a constant, unaffected by the location of the reference oscillator above or below the modulation frequency. Therefore, using a negative pattern and computing the difference for the positive and negative patterns we arrive at a measured phase shift ϕ'' where

$$\begin{aligned}\phi'' &= (\phi + \delta) - (-\phi + \delta) \\ &= 2\phi.\end{aligned}$$

The use of a negative pattern therefore results in a doubling of resolution and a cancellation of error due to low-frequency circuit delays.

While negative patterns are invariably employed in microwave distance measuring instruments, this has not been the case for electro-optic instruments. In fact the only commercial instrument exploiting the technique appears to be the Hölscher/Greene Tellurometer MA100 (67). It is worth considering why so powerful (and simple) a technique has been overlooked.

It has been argued that low frequency phase delays δ in the electro-optic case are common to the internal measured phase-shift ϕ_i and the external measured phase-shift ϕ_e and hence are removed by the subtraction

$$(\phi_e + \delta) - (\phi_i + \delta) = \phi_e - \phi_i$$

obviating the need for a negative pattern. It is argued that subtraction of the negative pattern yields no extra information and is equivalent to simply doubling the phase shift measured using the positive pattern. This argument overlooks an important fact and its conclusions are erroneous.

The main purpose of the internal/external subtraction is to remove high-frequency delays. Since the high frequency circuitry is linear it is a tenable assumption that the delay is the same for the external and internal measurements. The low-frequency signal-processing however includes phase-detection which necessarily involves a non-linear operation. As a result it is amplitude-sensitive and the low-frequency delay is a function of signal-level (in some instruments this effect is the dominant error source, necessitating careful balancing of signal amplitudes). Now in general (and certainly in the proposed instrument which demands a minimum of operator intervention), while the amplitude of the signal transmitted over the internal path is constant, that transmitted over the

external path varies over a wide dynamic range. Hence it is necessary to postulate two different values of low-frequency delay, δ_e and δ_i.
Now the external internal subtraction yields

$$(\phi_e + \delta_e) - (\phi_i + \delta_i) = (\phi_e - \phi_i) + (\delta_e - \delta_i),$$

leaving an uncompensated differential error ($\delta_e - \delta_i$). This differential error is however independent of the reference oscillator situation. Therefore, remeasuring with a negative pattern and subtracting, we get ϕ''' where

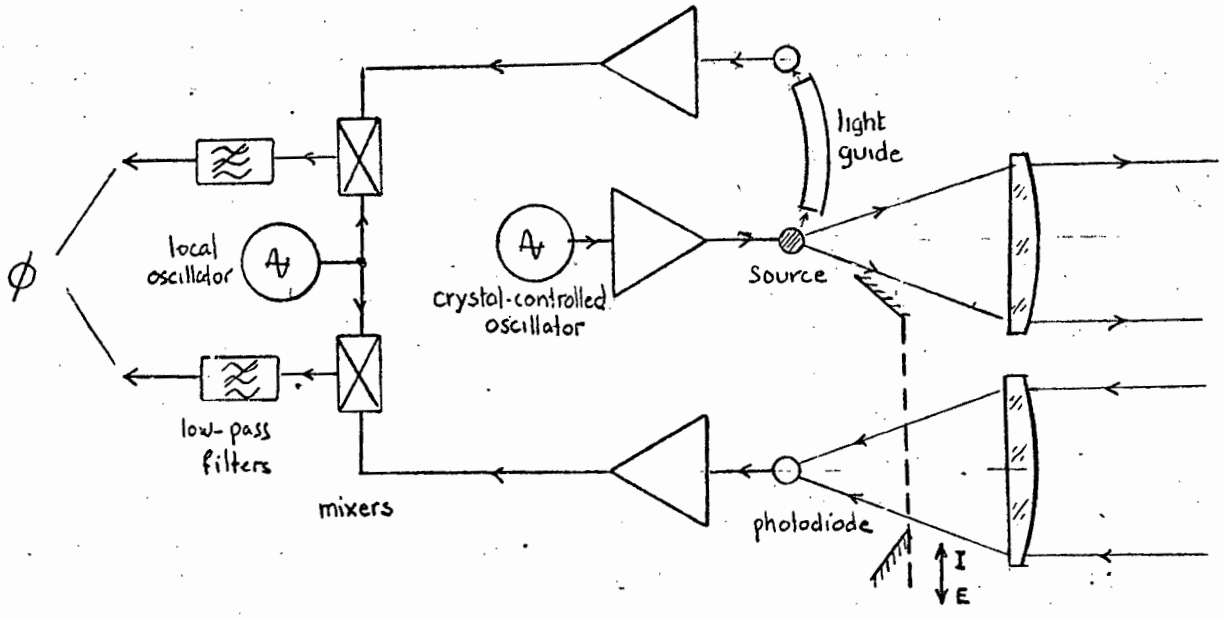
$$\phi''' = (\phi_e - \phi_i) + (\delta_e - \delta_i) - \left\{-(\phi_e - \phi_i) + (\delta_e - \delta_i)\right\}$$

and even the differential error is eliminated.

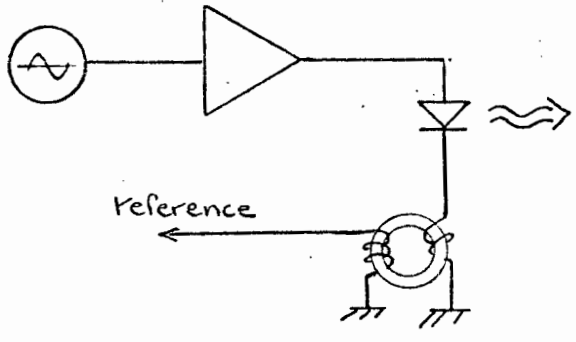
In order to verify this theory and investigate its significance, the writer examined the records of a large number of distance determinations carried out with a prototype Tellurometer MA100. The overall error determined in the normal mode of operation (i.e. using a negative pattern) was compared with the error using only the positive pattern (and doubling the result). The results showed a standard deviation of error in the second case 1,8 times that obtained in the first case (3,8mm vs 2,1mm). This validates the conclusion reached above and shows that, for that particular instrument, low-frequency amplitude-dependent phase-shift is a significant, if not dominant, portion of the overall error mechanism. The inclusion of a negative pattern is therefore fully justified for the present instrument. As a critical test of the technique and the achievable precision it was decided to carry out tests at half the above calculated pattern frequency - viz. 149 985 Hz to see whether the specification could be met. A practical advantage of this decision was that test equipment operating in this frequency range (phase measuring equipment, for example) was available.

4.5 AN EXPERIMENTAL PROTOTYPE SYSTEM

A block diagram of the experimental set up is illustrated opposite. A crystal-controlled oscillator at 149 985 Hz drives a modulator circuit which in turn drives a TIXL12 diode which is biased at an average current of 300mA. The low



PROTOTYPE SYSTEM BLOCK DIAGRAM



Unsatisfactory way of deriving reference.

modulation-frequency enables a high modulation depth (in excess of 90%) to be obtained. The radiation is collimated and the returning radiation collected by 60mm lenses as outlined in the previous chapter. The returning radiation is focused onto the sensitive area of a photodetector and the resulting photocurrent is turned into a signal voltage which is amplified. The amplified signal is mixed with a crystal-controlled reference oscillator at 149 985 Hz or 150 985 Hz. The lower-sideband signal is selected in each case by means of a bandpass filter centred at 1kHz and the resulting 1kHz signal is further amplified.

A sample of the outgoing radiation is obtained by means of a 100mm length of 0,5mm light-fibre positioned so that its end is illuminated by means of the "wasted" radiation emerging from the luminescence-diode. The free end of the light guide is butted up against the window of a second photodetector, and the photocurrent from this detector is processed in the same manner as that of the main photodetector, resulting in a second 1kHz (reference) signal.

This method of deriving the reference channel (i.e. via an optical link) though it may seem unnecessarily extravagant, has two important advantages. One relates to the superior isolation and freedom from spurious coupling or cross-talk between channels which results. It will be seen below that this is one of the most critical sources of error in this type of equipment and its elimination by shielding and de-coupling is one of the expensive areas in instrument fabrication (not to mention the special and critical testing required to ensure its absence). The second advantage is that it eliminates the phase instabilities first reported by Hölscher (51) between drive current and envelope of radiated power in the emitting diode.

To clarify this, consider an earlier experimental arrangement in which the transmit diode drive current was sampled by means of a current transformer (see diagram). It was soon discovered that when the instrument was switched on a rapid phase drift (several electrical degrees per second) was monitored. Needless to say, during this period the internal/external sampling period was far too long to negate the effect of the drift and gross error resulted. After a few minutes, as thermal

equilibrium is approached, the rate of phase drift is considerably reduced, but one is left with random fluctuations which, though small (typically $0,2^{o}$ rms) are sufficiently large to affect overall accuracy. When an optically derived reference signal is used these phase fluctuations are common to both channels and do not in any way affect the final result.

At this stage in the investigation no attempt was made to construct a phase-measuring system. The two 1kHz signals were simply fed into a commercial (Krohn-Hite) digital phase meter with a resolution of $0,1^{o}$. This gives a theoretical range resolution at 150kHz of 0,3m. Ideally one would prefer somewhat better resolution but this figure was judged more than sufficient to permit an exploration and, hopefully, elimination of the dominant errors in the system.

Aspects of the detailed design of the elements of the system will be discussed below in the context of projected and measured performance, and the elimination of systematic errors.

4.6 RESULT OF INITIAL TESTING

Test range

A set of retroreflective prisms was set up at distances of about 100, 200, 500, 1000 and 2000 metres and their distances determined within a few centimetres using a Tellurometer CD6. In order to determine the distance of the 1km and 2km stations with the CD6, the prism had to be augmented with a cluster of three and six prisms respectively. Reflector distances were determined (with a 3σ uncertainty of less than 0,06m) as

110,243; 205,179; 485,600; 1007,101 and 1973,990 metres.

For future reference these stations will be henceforth referred to as S1(110,243m) through S5(1993,990m) respectively. S0 will refer to the station occupied by the instrument.

Returned signal

The instrument was carefully focused and collimated for maximum returned signal using the reflector at S4. This was chosen because it is safely in the far-field region (at 10^{4} focal lengths) while the returned signal is sufficiently strong to dominate the residual noise and facilitate the maximisation process.

The gain of the receive channel was then set so that it was just short of saturation on the signal returned from S1, with a signal level of 1 V_{rms}. Under clear conditions, a signal of about 0,5 V_{rms} was then obtained from the reflector at S5. This is a dynamic range of 66dB, but it should be noted that it represents a dynamic range of only 33dB in the received power levels. This apparent anomaly causes a great deal of confusion to engineers who are unused to dealing with electro-optic systems (69). It is due to the fact that a photodiode is a square-law device whose (current) output is linearly related to optical input power.

A fundamental index of system performance is provided by the signal-to-noise ratio at the output of the receive channel. This is of course only meaningful if measured within a known bandwidth. It was decided to measure the noise in a 100Hz bandwidth. It is well known that for a simple two-pole resonant system the noise bandwidth B_N is related to the 3dB bandwidth Δf by the relation

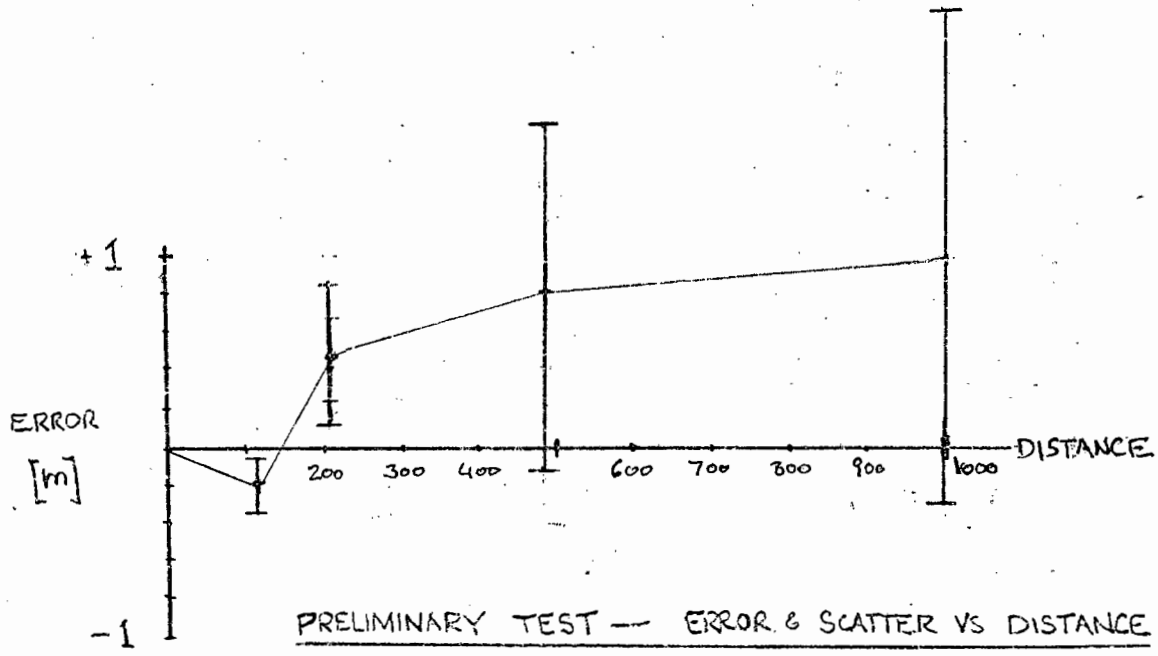
$$B_N = \frac{4}{\pi} f_{3dB}$$

A Sallen-and-Key bandpass filter with f_o = 1kHz and $\Delta f = 100 \frac{\pi}{4}$ = 78,5Hz (i.e. Q = 12,7) was therefore used at the output of the receive channel, followed by a sensitive true-rms voltmeter. For the reflector at S5 the signal-to-noise ratio was measured (under clear conditions) at 32dB. It was found that this signal-to-noise ratio was insufficient to permit a steady phase reading accurate to better than about 1^o unless the readings were averaged over several seconds. Thus at maximum range the system failed to meet the specification in terms of combined accuracy and update time.

Nevertheless it was decided to proceed with the evaluation of the instrument for a number of reasons.

(i) The system as developed exhibited an attractive simplicity and seemed to be worth studying in detail.

(ii) By means of various small modifications it should be possible to meet the specification. If, for example the detection sensitivity could be increased by a factor of about 5 by the use of an avalanche photo-



PRELIMINARY TEST — ERROR & SCATTER VS DISTANCE

detector and interference filter to limit background power, the target might be met. (This is discussed in section 4.8). An even simpler variant which would certainly suffice would be to replace the light-emitting diode by a semiconductor laser as discussed in section 3.5. (Such a device was not commercially available at the time.)

(iii) The apparatus, which appeared to be functioning correctly, afforded a valuable opportunity to investigate systematically the various possible sources of error.

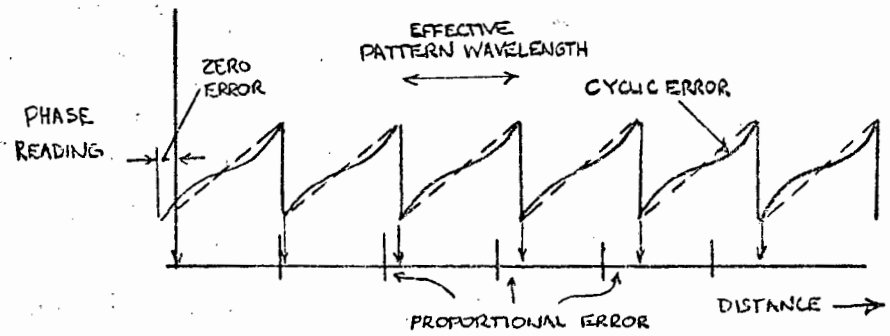
(iv) The possibility of validating theoretical models would yield a sound basis for the paper-design of future versions.

Up to this point it had been rather naively assumed, on the basis of previous experience with similar instruments, that 5nW would be an adequate level of returned optical power. This however failed to take account of the phase resolution demanded and the short measuring period (0,3s) specified. A rather belated analysis, to be presented subsequently, confirmed the order of magnitude of the discrepancy alluded to above.

Distance measurements

A set of measurements was carried out between S0 and S1 through S4, with the following results:

Station	S1	S2	S3	S4
$\phi_e - \phi_i$	39,6°	74,1°	175,1°	2,9°
Computed distance $\frac{1000}{360}(\phi_e - \phi_i)$(m)	110,0	205,7	486,4	1008,1
True distance (m)	110,243	205,179	485,600	1007,101
Error (m)	- 0,2	+ 0,5	+ 0,8	+ 1,0
Standard deviation of scatter (m)	0,3	0,7	1,8	2,6



CATEGORIES OF ERROR

The results are computed from a mean of ten determinations for each station, and for each station the standard deviation was computed. The average standard deviation of 1,4m is considerably in excess of the target specification, and the results clearly showed excessive random and systematic components of error. There does not seem to be any evidence of scale error, although too few points were involved for this to be said with much confidence.

There was no sign of gross malfunctioning in the instrument and it was decided to proceed with a detailed theoretical and experimental exploration of the possible sources of error.

4.7 CATEGORIES OF ERROR IN DISTANCE MEASURING INSTRUMENTS

It was found helpful, in attempting to disentangle the various error sources, to submit them to an analysis in terms of logical categories. The major subdivision is between <u>random</u>, or statistical, error and repeating or <u>systematic</u> error. The former may in principle be reduced by computing the average of a number of separate measurements of the same distance (or by increasing the observation time, which amounts to the same thing). If the errors are normally distributed we would expect the mean of n readings to have a variance reduced by a factor n, or a standard deviation reduced by $\sqrt{n}$. <u>Systematic</u> errors, on the other hand, are not reduced by averaging but since they bear a definite functional relationship to, say, the measured distance, it is always possible in principle to eliminate them by means of an appropriately devised observational or calibrating procedure. We shall treat these first.

Systematic errors may be subdivided into <u>scale errors</u> or errors proportional to the distance measured, <u>zero errors</u> or errors independent of the distance measured (but systematically dependent on some other variable, such as instrument pointing or signal level) and <u>cyclic</u> errors, or errors which repeat periodically with distance measured. The most likely period is of course that of the effective measuring wavelength, or 1km in the case of the prototype (500m after subtracting the negative pattern).

(a) Scale error

Scale error (or proportional error) arises when there is a discrepancy between the actual and intended effective measuring wavelengths. There appear to be only two possible causes, one instrumental and one environmental. The instrumental cause is drift in the modulating fequency, and the environmental cause is a discrepancy between the actual and expected values of atmospheric refractive index. In cadastral survey instruments the refractive index is computed from local measurements of atmospheric pressure and temperature and the writer has shown (70) that a precision of 1mBar and $1^{o}K$ suffices for a refractive index precision of 10^{-6}.

In the case of the present instrument, the required precision of 0,3m at a maximum range of 2km implies a precision of atmospheric index determination of 1,5 parts in 10^{4}, and compensation based on meteorological observations is hardly required. The chosen pattern frequency is based on $15^{o}C$ at sea level (1000mB) given a refractive index of 1,000274. At other temperatures and pressures the refractive index will be given by (See Appendix 4.1):

$$n = 1 + \left\{ 79{,}1 \ \frac{P\ [\text{mBar}]}{T\ [^{o}k]} \right\} . \ 10^{-6} .$$

Let us consider a fairly extreme case, where $T = -50^{o}C$. then we have $n = 1 + 79{,}1 \ \frac{1000}{223} \ 10^{-6} = 1{,}000354$. This represents a variation of only 8 parts in 10^{5}. Pressure variations are even less significant, justifying the contention that environmentally related changes of scale are not significant. By the same token, drift in the modulation frequency is not a factor of any importance. Even an indifferently designed crystal oscillator is unlikely to drift by more than a part in 10^{4} in the most extreme conditions of temperature.

Scale factor is therefore not a critical area in the present design and it is not surprising that proportional error is not in evidence in the initial test results for the prototype. Clearly, if the instrument were intended for use consistently in arctic or tropical conditions or at high altitude it would be only sensible to take this into account in setting the modulation frequency.

(b) Zero error

Again the present instrument differs from the cadastral version, for which an important source of zero error is centering, or positioning the instrument vertically above the point, and variations in the effective plane of reflection of the retroreflective prisms relative to their mechanical mounting. Clearly these factors are of no significance whatever when we are considering resolutions of a substantial fraction of a metre.

Variation in electrical phase shift as a potential source of zero error is almost entirely removed by the external/internal subtraction. What little remains due to level-dependent phase shift acting on the difference between internal and external signal levels is effectively dealt with by the differencing of positive and negative patterns as seen in Section 4.4. It was therefore thought that the instrument would be free from zero drift, but it was soon discovered that a rather subtle and elusive source of zero error remains:

Pointing-dependent zero error

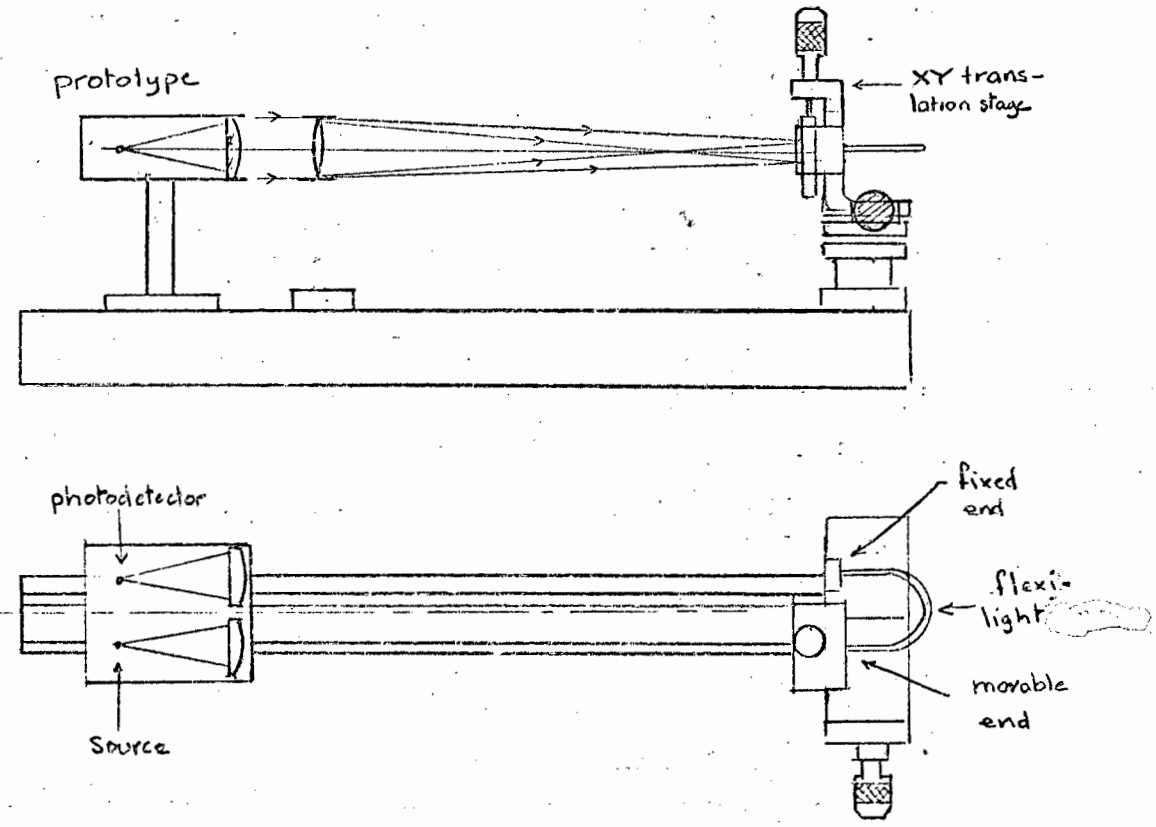
The time delay between the drive current to the gallium arsenide light-emitting diode and the resulting envelope modulation of the emitted radiation is not constant for the various elemental areas of the emitting junction. Since the instrumental beamwidth is, at all but the shortest ranges, considerably greater than the angle subtended by the reflector, the latter is illuminated by only a small part of the beam's cross-section. The radiation illuminating the reflector therefore comes from only a small part of the emitting area of the diode. Precisely which area contributes the radiation depends critically on the orientation of the instrument vis-a-vis the reflector, and a "pointing-dependent" zero error is introduced.

Curiously, this effect seems to have had no mention in the Western literature (other than a brief reference in papers by the present writer) (59;70) and only a perfunctory discussion in the Soviet literature (71), and it was judged necessary to investigate it in some detail as a major factor limiting the accuracy of electro-optic distance-measuring instruments.

Originally (72) it was thought to be a result of a transmission line type of transit delay of charge carriers in the semiconductor junction (see Appendix 4.3) and there is little doubt that, when the diode is being modulated near its high frequency limit, this is the dominant mechanism. It was therefore thought that the effect (being virtually a pure time delay) would be insignificant at low modulation frequencies and in an instrument where only a moderate resolution in terms of absolute distance was being sought. However the mechanism turned out to be more complex than this, one puzzling (and still unexplained) anomaly being that physically small junctions tended to exhibit as great an error as large ones. Moreover large errors were unexpectedly found at the low modulation frequency of 150kHz. Rather than a pure time delay, the error appeared to be intermediate in character between a pure time delay (which would yield constant distance error irrespective of modulation frequency) and pure phase delay (which would lead to increased distance error at low frequencies). It was conjectured on this basis that at least two different error mechanisms were simultaneously present: a phase delay along the lines discussed in the appendix and a tendency for certain areas of the emitting junction to "turn on" (begin emitting radiation) at specific threshold levels of drive current. Conceivably this could occur due to areas being"biased" off due to a combination of current-crowding and spreading resistance. It is not difficult to imagine that this effect could be frequency dependent, leading to a very complex interaction. Clearly this is an area that deserves further exploration but it was not felt to be germane to the present thesis, which is concerned with developing an accurate instrument using commercially available sources. Attention was therefore focused (with considerable success) on a series of palliatives to reduce the error.

Effect on accuracy

The effect of source-related phase error was dramatically in evidence in the prototype, resulting in apparent changes of measured distance of up to 1,5m for small angular changes (0-5 arc minutes) of instrument pointing, and it was found that this had accounted for a great deal of the spread in


prototype
XY translation stage
photodetector
fixed end
flexi-light
movable end
Source

the initial series of measurements. Improving the consistency of pointing by fitting a 20 power gunsight telescope to the instrument and exercising considerable care in pointing reduced the average standard deviation for a given point by a factor of three. Similar results could be obtained by pointing very carefully for maximum returned signal.

It was suspected at this point that some position-sensitive phase-shift was also occasioned by the photodetector. This was subsequently found to be the case, an error of about 1ns (corresponding to 0,3m in the return path) arising when the returning radiation was focused onto the edge of the photo-detector. Discussions with the supplier of the diode (the Hewlett Packard Company) revealed that the effect was probably due to the anomalously deep penetration of the photons impinging on the device near the junction edge. It was eliminated by fitting the 1mm junction of the photo-detector with a mask having a 0,5mm aperture.

In order to separate emitter and photodetector contributions to "pointing" error, as it came to be called, and to permit a systematic investigation including the plotting of phase contours for the diodes, a simple test rig was developed. The instrument was mounted on an optical bench and a 1-dioptre lens formed an enlarged image of the emitter diode, 2,5mm in diameter. This image was scanned by means of the free end of a piece of 50mm diameter light guide mounted on a micrometer-controlled XY translation stage. The other end of the light guide re-radiated the sampled radiation to another 1-dioptre lens which collimated it and returned it to the receiver. By inverting the instrument in the test rig the image of the photodetector rather than the photoemitter could be scanned. Using this setup the phase contours for the TIXL12 diode were plotted. Later the test-rig was used to investigate a number of other sources, as well as to determine the effectiveness of the palliative devices described below.

Approaches to "phase scrambling"

Several approaches were considered to attempt to make the source appear more homogeneous in respect of phase.

i. Defocusing. This seemed an obvious first attempt. It was hoped that a moderate degree of defocusing would impair the resolution (i.e. blur) of the source without exacting too great a penalty in increased beamwidth and loss of effective source radiance. This was not successful. Scrambling of a sort does occur, but it is on a micro rather than a macro scale. Each point on the source is effectively mapped into a disc in the image, with a roughly gaussian density profile, but contribution from, say, an edge of the diode is still overwhelmingly dominant in the image. Unacceptable loss of beam irradiance occurs before there is appreciable homogenisation of the source.

ii. Integrating sphere. The integrating sphere performs the task of scrambling almost perfectly, but at the cost of an unacceptable loss of radiance. Consider a 2mm diameter sphere (about the smallest feasible)mounted directly on a 0,5mm diameter junction. Neglecting absorption loss in the sphere and re-absorption by the junction we can calculate the loss of radiance approximately by noting that the radiation from the junction area $\pi/4\ (5 \times 10^{-4})^2 m^2$ is spread over the surface area of the sphere $4\pi \times 10^{-6}\ m^2$. Even if we assume (optimistically) that the radiation from the integrating sphere emerges with the same solid angle as that from the unmodified junction there will be a loss factor of the ratio of areas, or 64 times! The reality is worse than this, for a variety of reasons. Much of the radiation tends to become trapped in a lossy "whispering gallery" mode, although this can be frustrated by using an irregular cavity (e.g. crumpled aluminium foil) rather than a smooth sphere. Clearly however the integrating sphere is not a viable solution to the problem.

iii. Field lens. An approach is suggested by the well known solution to an anologous problem in astronomy. The task of measuring the radiance of a star by means of a photomultiplier located at the focal plane of a telescope is complicated by local variations in responsivity over the photocathode. The standard solution is to provide an auxiliary lens which images the photocathode onto

the main telescope aperture. Clearly all rays passing through the aperture will impinge on the photocathode but light rays from any given direction reaching the field lens will be distributed over the entire photocathode surface and transit of the star through the telescope field of view will not result in any movement of its image over the photocathode. This is exactly equivalent to the phase problem where we want radiation having the averaged phase of all the photo-emissive elements to be sent out along all directions contained in the transmitted beam. The problem is a practical one of obtaining a suitable lens. In order to image a 0,5mm source onto a 60mm aperture at f/3,4 we need a field lens with a focal length of approximately

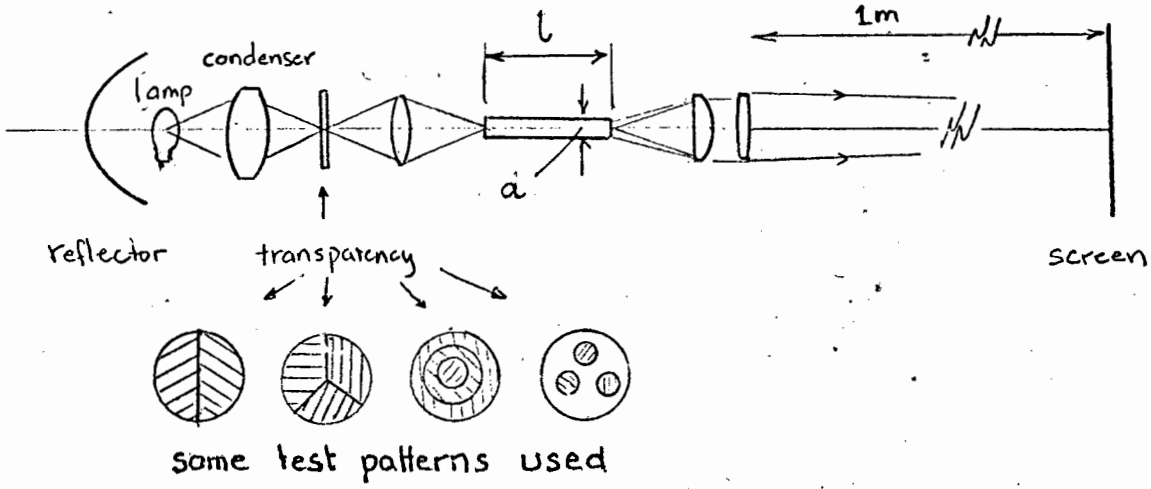
$$\frac{200 \times 0,5}{60} = 1,7\text{mm} .$$

The numerical aperture of the field lens should be at least equal to that of the main optical system, leading to a 1mm diameter lens. It was not possible to obtain a lens meeting these requirements exactly, the nearest approach being a 10x microscope objective lens. Tests were also carried out with a variety of home-made lenses made from droplets of molten glass and epoxy resin and turned from perspex rod and polished by exposure to heat ("fire-polished") or to the fumes of a solvent such as ether. It was reasoned that the optical quality of such a lens should be relatively uncritical due to its proximity to the source but in fact none of these home-made lenses was good enough.

Tests showed that the field lens effected a marked improvement in phase homogeneity, but this proved rather critically dependent on correct optical adjustment. When the field lens was correctly positioned (using the microscope objective and collimator lens set up on an optical bench) the phase error was reduced by a factor of 12, with a loss in radiance of only some 20%. However distance determinations carried out with the field lens installed in the instrument revealed an unanticipated increase in random error at short-range and a marked sensitivity of the distance reading to the orientation of the retroreflector. It was then realised that by imaging the source onto the collimation

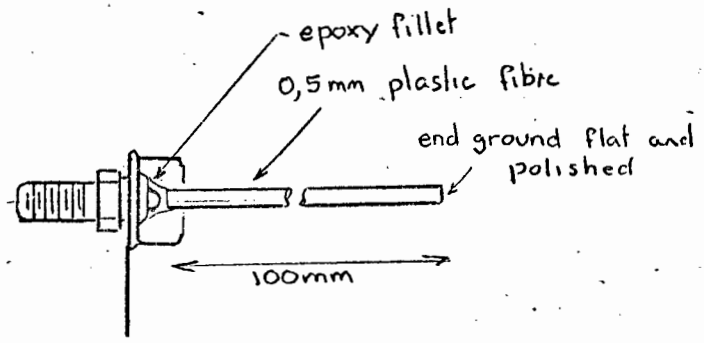
aperture, the field lens exchanges angularly related phase shifts in the outgoing beam (pointing error) for spatially related phase error across the collimation aperture ("aperture error"). Although the instrument is now insensitive to the location of the reflector in the transmitted beam, it becomes critically sensitive to the optical coupling from transmit aperture to receive aperture via the reflector. For example, suppose a small reflector were used, such that its aperture were smaller than that of the collimator lens. At short range the distribution of energy from the reflector would be such that only the inner adjacent edges of the lenses would be coupled and only half the emitting area of the source would contribute to a measurement. At longer ranges, prism deviation would cause coupling of the entire apertures. Thus if there happened to be a nett phase difference between the two halves of the emitting area this could appear as a systematic variation of instrument zero between short and long range measurements.

iv The light-guide scrambler. If radiation from the source is coupled into a short length of multi-mode optical fibre and the free end of the fibre used as a reconstituted source, a reduction in phase error will be effected by the multiple internal paths in the fibre. Clearly if the fibre is sufficiently long and the number of modes sufficiently high, any point on the free end will receive contributions from all elements of the source and effective phase averaging will occur. All the radiation coupled from the source onto the guide emerges at the far end without any increase in divergence, since at each internal reflector the angle of incidence is equal to the angle of reflection. Thus, if the numerical aperture of the fibre is at least as great as that of the collimator lens and coupling between source and fibre is optimised, there should be little loss of radiance. Attenuation loss is negligible due to the short length of fibre involved.



some test patterns used

TEST SETUP FOR INVESTIGATING 'SCRAMBLING'



TIXL12 EQUIPPED WITH LIGHT-FIBRE

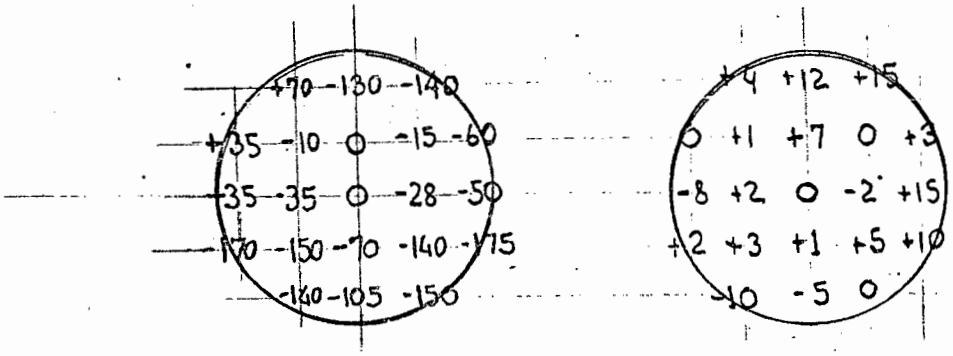


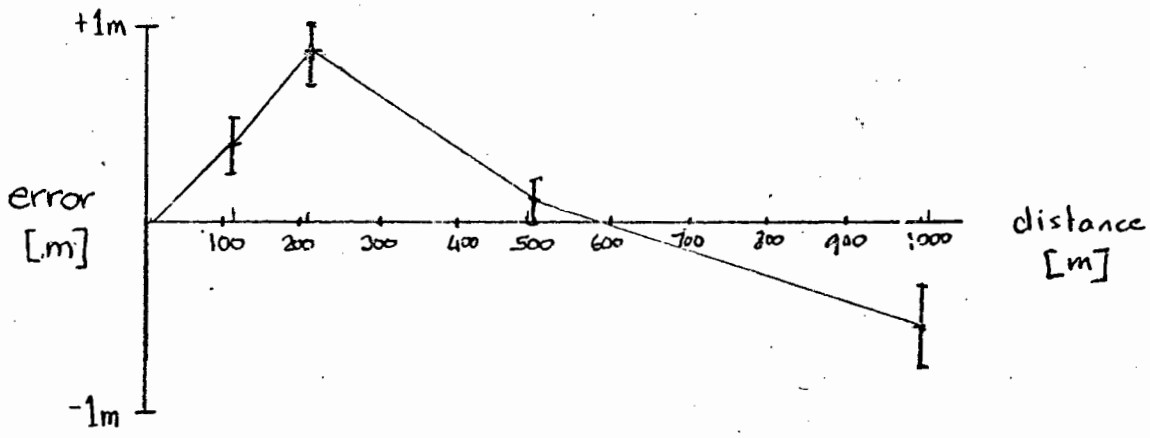
figure represents differential time delay in ns at 150kHz modulation.

RESULT OF PHASE-SCAN

In order to obtain a feel for the 'scrambling' process, an optical model was set up, using a length of 10mm diameter perspex rod with polished ends. Test targets using patterns of primary colours exhibiting various kinds of spatial symmetry were imaged onto one end of the rod. Light emerging from the free end was collimated by a 60mm diameter f/3,4 lens. The resulting beam of light was focused by a 1-dioptre lens to give a 50mm magnified image of the end of the rod on a screen. Scrambling was observed in the form of loss of definition in the image on the screen and it was found, using various lengths of rod, that very effective scrambling occurred for rod aspect rating (length divided by diameter) of 100 or greater.

On the basis of this finding, a 50mm length of 0,5mm fibre was butted against the source and the free end mounted at the focal point of the collimator lens. With this rather crude arrangement the radiance of the free end of fibre was measured to be about 1,5 W/cm^2sr, or about 8 x down on the source. This is not fundamental however, but rather reflects the crude facilities for optical fabrication available to the writer. As mentioned in the previous chapter, the subject of coupling radiation from semiconductor sources into light-fibre has recently been the subject of intensive research to meet the needs of the telecommunication industry and many published results have shown a high coupling efficiency to be possible (38-40).

A phase scan carried out on the free end of the fibre showed that the phase variations had been reduced about an order of magnitude. Despite the loss of radiance, the signal-to-noise ratio at 1km was still adequate provided random variations were averaged out by taking multiple phase readings. The distances S0 to S1 through S4 were re-observed, repointing (without using the telescope or taking any special care) 10 times on each reflector.


+1m
error
[m]
-1m
100
200
300
400
500
600
700
800
900
1000
distance
[m]

Station		S1	S2	S3	S4
error	(m)	+ 0,4	+ 0,9	+ 0,1	- 0,5
standard deviation	(m)	0,3	0,3	0,2	0,4

It can be seen that the variance on each station distance is greatly reduced, the mean standard deviation being a more acceptable 0,3m, and the dominant residual error is clearly systematic and functionally dependent on distance. It can be concluded that the pointing error problem can be satisfactorily solved by means of the light-fibre approach. Taking this finding in conjunction with the results of Chapter 3 and the need for extra range it is clear that a gallium-arsenide laser launching radiation into a short length of fibre (of about 200μm diameter) constitutes an ideal source for the projected instrument.

(c) Cyclic error

Obviously any gross error associated with a malfunctioning of the phase-measuring circuitry will give rise to an error which, like phase itself, is cyclic in character. Such errors however are readily detected. The dominant source of cyclic error in electro-optic distance measuring instruments is crosstalk between transmit and receive channels. For historical reasons this has sometimes been referred to in the literature as "contamination" (70). The problem is well known and has been widely discussed in the literature. Only a brief treatment need be given here.

The signal current in the photodetector will be a replica of the modulating current in the emitting diode, delayed in phase by ϕ. It will also be considerably smaller in amplitude - a minimum usable signal of say 10nW will give rise to a photo-current of about 3×10^{-10}A, about 170dB down on the rms transmit diode current of 100mA! Although often quoted this way, the figure expressed in dB is somewhat misleading since the impedance levels of photodetector and emitting diodes are very different. Nevertheless it remains true that the ratio of amplitudes of the two currents is $3,3 \times 10^{8}$. Clearly a very small degree of spurious electromagnetic coupling between the transmit and receive channels will suffice to perturb the phase of the returned signal.

The worse case is when the spurious signal is in quadrature with the received signal. If their amplitudes are respectively A_c and A_r the angular error that will result is δ_c max.

$$\delta_c \text{ max} = \text{arc tan } \frac{A_c}{A_r} .$$

when the two signals are in phase or 180° out of phase δ_c will be zero.

To see how the error varies with distance consider some distance for which the contaminating signal happens to be in phase with the returned signal. Clearly in this case no phase error results. Now as the phase of the returned signal changes by θ (that of the contaminating signal remaining constant) the resultant of the two phasors is perturbed by an angle δ. If for simplicity we allow the frame of reference to rotate so as to fix the orientation of A_r we have the simple expression

$$\delta = \arctan \frac{A_c \sin \theta}{A_r + A_c \cos \theta}$$

If the system is to be usable, A_c will be very much smaller than A_r, so we can write

$$\delta = \arctan \left(\frac{A_c}{A_r} \sin \theta\right) .$$

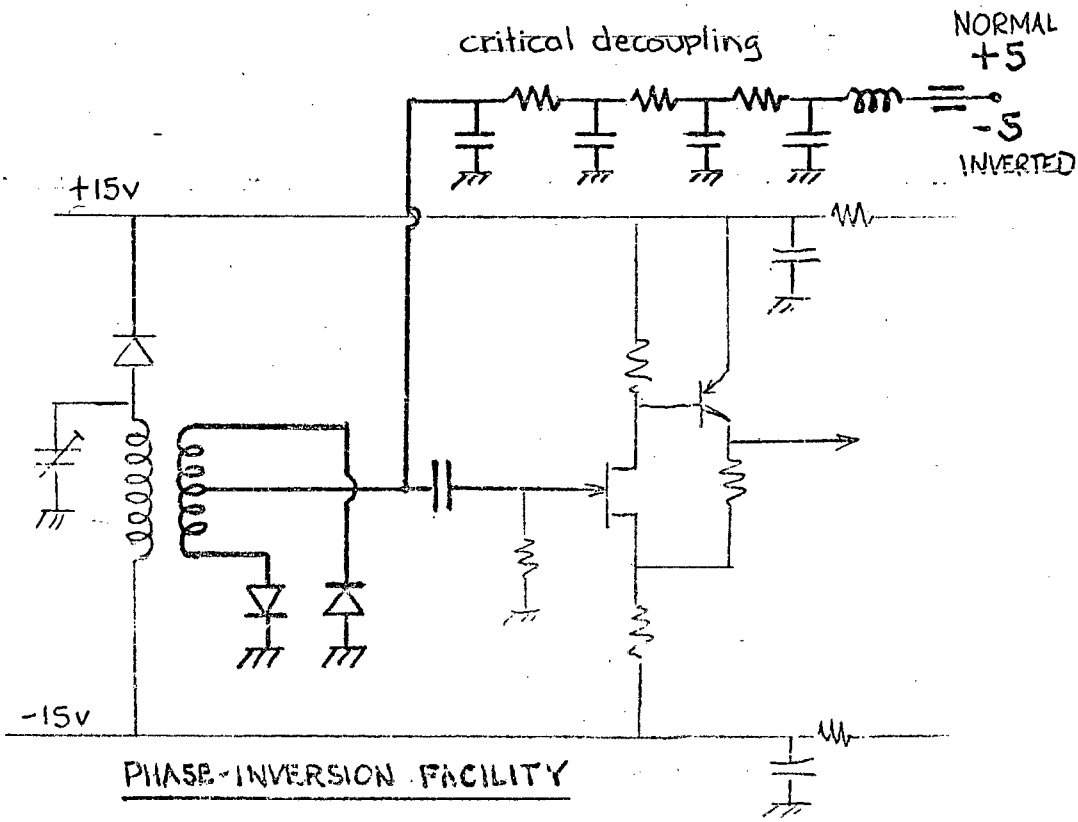
For the same reason δ must be small and we can invoke the small-angle approximation $\tan \delta \approx \delta$. Thus the error δ in degrees is given by

$$\delta' = \frac{180}{\pi} \frac{A_C}{A_R} \sin \theta$$

For a peak error of less than 0,3° we require $A_c < (5 \times 10^{-3})A_r$.

It should now be clear why the suppression of crosstalk is one of the most critical areas in the design of phase-comparison distance measuring equipment. It should be noted that subtraction of the internal signal does not remove this error, nor need we consider the effect of crosstalk on the internal signal itself, since its phase is constant.

Error reduction by phase reversal: In 1965 Hölscher (51) proposed a simple and effective technique for dramatically reducing the effect of crosstalk. Curiously, it does not seem to have been adopted in any of the commercial instruments (including the MA100 engineered by the present writer!).



PHASE-INVERSION FACILITY

The technique consists in inverting (adding π to) the phase of the received signal at the photodetector. The inversion must take place prior to any electromagnetic crosstalk entering the channel, so that the phase of the crosstalk signal is unaffected by the component.

If the error in the signal was δ , after inversion it will be (approximately) $-\delta$. Adding the original phase $\theta + \delta$ and that of the inverted signal $(\theta + \pi + \delta)$, and subtracting π it can be seen that δ is eliminated.

$$(\theta + \delta) + (\theta + \pi - \delta) - \pi = 2\theta$$

Again, as in the case of the negative pattern, error is eliminated by subtraction with an effective doubling of resolution. Actually the treatment above is exact only for θ a multiple of $\pi/2$. A more careful analysis (Appendix 4.4), in the general case yields a residual error δ', given by

$$\delta = \frac{\sin 2\theta}{\left(\frac{A_R}{A_C}\right)^2 - \cos^2 \theta}$$

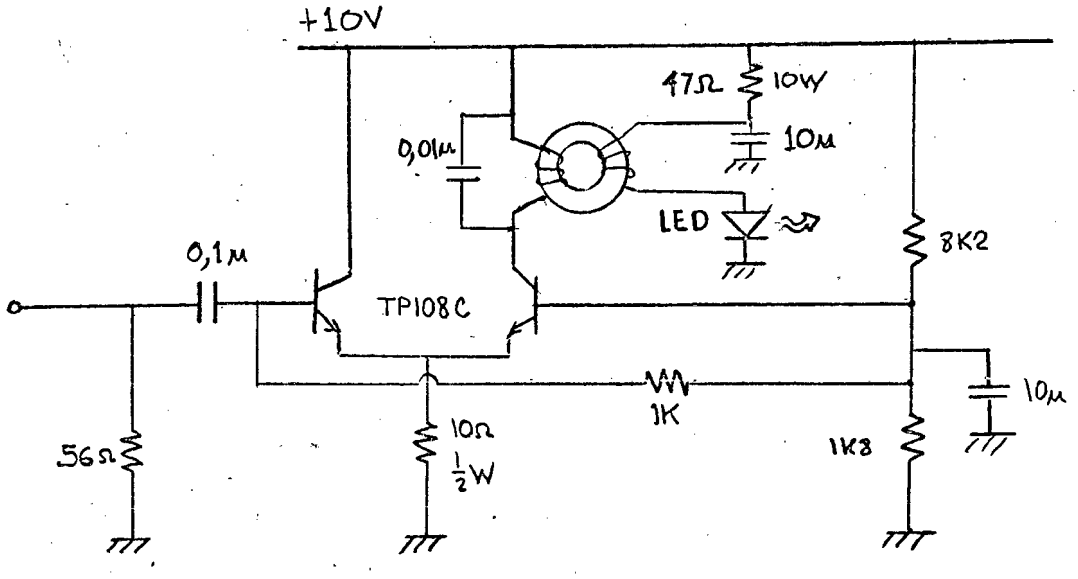
which has a peak value δ'max when $\theta = 45^{\circ}$, where

$$\delta'_{max} \doteq \left(\frac{A_C}{A_R}\right)^2$$

If δ = 1% the residual error is reduced from 0,6° to 0,34 arc minutes.

Practical details

The implementation of the scheme is straightforward. The tuned circuit which forms the load for the photodetector is provided with a bifilar centre-tapped secondary winding, and a diode switch earths either of the ends of the secondary, thus providing the required phase inversion. It is crucial to note, however, that contaminating signals induced by stray flux in the input tuned circuit will also have their phase reversed and will not be eliminated. The screening of this transformer is therefore a critical design problem. In the first prototype a multiple screen was used with a permalloy layer. A totally closed screen of conducting material such as copper (acting as a "shorted turn" against alternating magnetic fields) is not sufficient at the low frequency of 150kHz on account of the large skin depth.

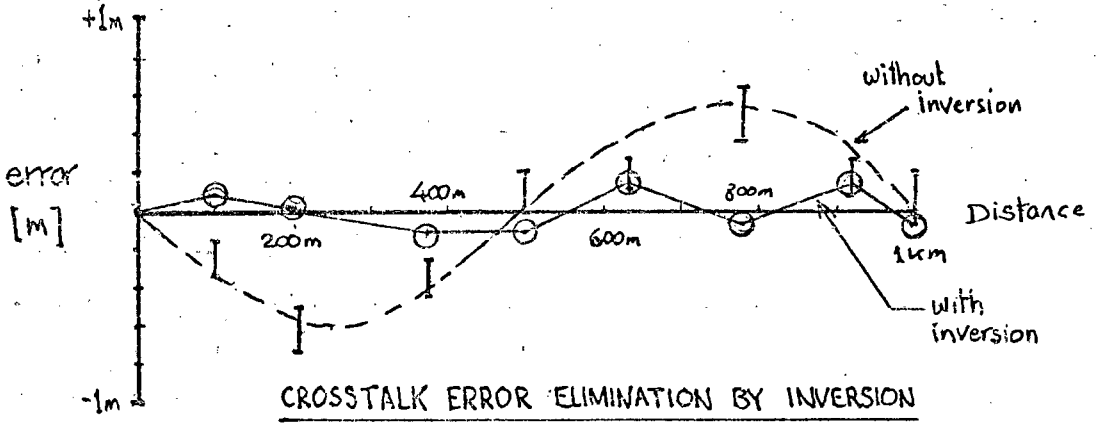


LED MODULATOR CIRCUIT

Cancellation by phase inversion depends upon the assumption that the level of crosstalk is small to begin with - say 1% of the weakest usable signal. This demands a level of spuriously induced input current in the photodetector load which is about 3×10^{10} times less than the current in the transmitting diode only some 100mm away. Accordingly the instrument makes the most stringent demands on design and practice in respect of shielding and decoupling. It was found that, based on fairly extensive previous experience in this area, the required order of performance could be achieved by ordinary good practice - thick (2mm) copper shielding boxes with soldered lids, single point earthing for each box, double-shielded coaxial interconnections and heavy supply decoupling with multiple π-section filters optimised for maximum transmission loss around 150kHz. Some areas however are particularly vulnerable and require special attention.

(i) Radiated fields from the transmitter :
In order to facilitate matching the 1-2Ω resistance of the emitting diode, to conserve batter power and to improve the envelope waveform of the modulated radiation, it is desirable to use a resonant circuit in the collector of the modulating transistor. This causes the problem of leakage flux which tends to couple into the receive channel resonant circuits. The problem is exacerbated by the fact that circulating currents, Q times as great as the diode current, flow in the tuned circuit. The only solution appears to be to minimise leakage flux by the use of high-permeability toroidal cores and to resort to nested high-permeability shields. In the prototype these were improvised using tape from a permalloy wound-tape core.

The problem of transmitter supply decoupling is greatly eased by adopting a balanced configuration which minimises the modulation-frequency component flowing into the supply. This costs a factor of two in supply current, but is amply repaid by crosstalk reduction. The circuit diagram shows the configuration finally adopted for the transmitter.



CROSSTALK ERROR ELIMINATION BY INVERSION

(ii) Conducted crosstalk at the receiver front end:
In addition to the supply line, which can be heavily decoupled, two other conductors lead directly to the most sensitive part of the instrument, the photodetector preamplifier. One of these is the photoconductor bias line, but it was soon discovered that, at low modulating frequencies such as 150kHz there is no merit in providing a special high-voltage supply for a PN photodiode. The separate line was retained so that an avalanche photodetector, which requires about 150V bias potential, could be tried. The other is the switching line for the phase inversion. It is very important that these be adequately decoupled, and multi-section filters were used, housed in compartmented shields. The instrumentation described in the next section was used to verify the attenuation factor of these filters as being greater than 150dB.

Measured results:
The level of crosstalk can be reliably inferred by measuring a series of known distances spanning a spatial wavelength or pattern distance. This was done with the results shown. In order to preserve the sinusoidal shape of the perturbations, it is necessary to maintain a reasonably constant signal amplitude, since the magnitude of the error is inversely proportional to received signal strength. This was done by using a cluster of three reflector prisms to minimise random scatter at the longer distances. For shorter distances the signal level was maintained constant by interposing a series of optical attenuators between the lens and the emitter diode. The results were plotted with and without phase-inversion, showing that, by means of the latter, crosstalk at this level of signal strength is reduced to insignificant proportions. Some residual periodicity can be discerned in excess of that predicted by the theory. This is almost certainly due to spurious signals induced in the input tuned circuit, probably due to flux leaked from the transmitter tank circuit, and inverted along with the signal. The nett result was encouraging, especially when it is noted that the increased radiation anticipated with a lasing source gives rise to increased signal levels without increased crosstalk (or possibly with a reduction in crosstalk, if drive-current levels are lower) and hence offers contamination-free gain.

Although the method of detecting crosstalk outlined above is the ultimate criterion of its presence, it is not a convenient method during development and it provides no feedback of information when one is attempting to reduce crosstalk by piecemeal modification of circuitry or construction techniques. A better method is to measure directly the level of spurious signal present at the output of the receive channel when no optical input signal is present. The problem is that it is necessary to measure a signal in the presence of perhaps 100dB of noise (depending on the bandwidth at that point) and this is sufficient to tax the linearity and dynamic reserve even of commercial lock-in detectors. Furthermore these do not normally operate at frequencies higher than 50 or 100kHz, precluding the possibility of investigating the presence of 150kHz spurious signals at the output of the receiver amplifier prior to the mixing operation. It was therefore decided to develop an instrument capable of detecting the presence of a coherent high frequency signal in more than 100dB of noise. The instrument was not designed only to assist in the development of the present instrument but as a universal piece of test equipment for phase comparison distance measuring equipment. Versions have subsequently been in use in research and development and production testing by the Tellurometer division of the Plessey Company.

The principle of the crosstalk detection instrument has been described elsewhere (see Appendix 4.5). Briefly, an input signal (consisting mainly of noise plus a small amount of coherent signal) is mixed with a crystal-controlled local oscillator at the frequency of the coherent component. Prior to the mixing operation, the phase of the local oscillator is progressively stepped through increments of 120° at a rate of 33 steps per second. The mixer output contains a 3-step sampled approximation to a sinusoidal signal at 1Hz. This is filtered by a three stage n-path filter with 0,1Hz bandwidth. The output of this filter, which uses reed switches for maximum linearity and switching signal isolation, is further bandpass filtered and precision rectified to yield a DC signal proportional to the coherent portion of the input signal and - unlike the case of a conventional synchronous detector - independent of the phase relation between it and the local oscillator. Thus an effective bandwidth of 0,1Hz is obtained without the need for a coherent reference signal.

This circuit arrangement was originally suggested by the "Synchrohet" lock-in detector marketed by the Princeton Applied Research Corporation (73). It differs however in several important respects, including that of phase-insensitivity and is believed to be novel. It has proved invaluable in the development of distance measuring instruments. With the aid of this instrumentation, periodic error due to crosstalk was reduced to negligible proportions - that is to say it disappeared into the noise even at the lowest usable signal levels. A plot of error vs distance then shows that the error is dominated by random scatter which decreases with the square root of the number of readings averaged. Systematic error has therefore been effectively eliminated and it remains to consider noise and its effect in producing errors that are random in character.

d. Random error

The effect of finite signal-to-noise ratio in the received signal on the precision with which phase can be measured is discussed in Appendix 4.6. It is shown that the phase perturbation of the signal can be expressed by a standard deviation of error given by

$$\delta \doteq \frac{180}{\sqrt{S/N}\ \sqrt{B.t}\ \pi}$$

In section 4.6 we saw that a signal-to-noise ratio of 32dB measured in a noise bandwidth of 100Hz resulted in a phase scatter of about 1^{o} (or roughly an order or magnitude larger than desired). The averaging time of the digital phase detector used was 0,1 seconds, and the signal-to-noise ratio of 32dB corresponds to a signal-to-noise amplitude ratio $\sqrt{S/N}$ of 40. Thus the above expression yields a predicted standard deviation of phase determination of δ' where

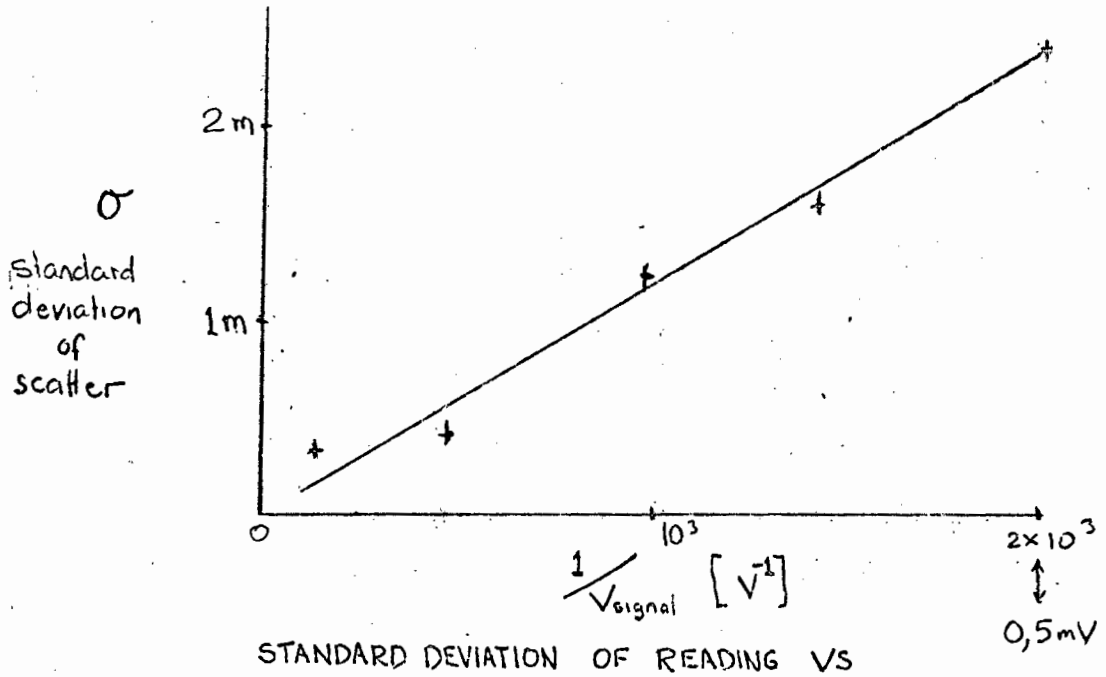
$$\delta' = \frac{180}{40\times\sqrt{10}\times\pi} = 0,5^{o}$$

which is in good agreement with the informal observation of about 1^{o} scatter.

We see therefore that in order to achieve a standard deviation of $0,1^{o}$ we would require a signal-to-noise ratio of

$$5 \times 40 = 200, \text{ or } 46\text{dB}$$

measured in a 100Hz bandwidth. Because of the linear relation between optical power and photocurrent this implies an increase in the received optical power (or reduction in the



STANDARD DEVIATION OF READING VS RECIPROCAL SIGNAL LEVEL

overall transmission attenuation) of only 5 times, or 7dB. This would be readily achieved by the projected order-of-magnitude increase in effective source radiance implied by the use of a fibre-coupled coherent radiation source as discussed in chapter 3.

The anticipated linear relationship between reciprocal signal amplitude and standard deviation of phase scatter was verified by attenuating the signal, observing for each attenuator setting a series of phase readings, and computing the standard deviation. It was concluded that the assumption of the previous paragraph, that the instrumental specification in terms of random error and measuring time would be met given the availability of a suitable source, was fully justified.

Measuring time: We have assumed above a measuring time of 0,3s. However it might be objected that in seeking to eliminate systematic error, we have segmented this period into a number of shorter periods - periods of positive and negative patterns to eliminate low frequency level dependence and of direct and inverted phase to reduce the effect of crosstalk - and the random error will therefore be increased. However it is easily shown that this effect is cancelled when the partial results are combined, and only the overall measuring period is significant in determining randomness. If the measuring time t is segmented into N parts, the effective measuring bandwidth associated with each partial measurement will be increased by N, as will the effective noise power. The random error of each partial measurement will therefore increase by $\sqrt{N}$. But the effect of averaging N statistically with statistically independent errors is to reduce the variance by N or the standard deviation by $\sqrt{N}$. Thus in the averaged result we are left with the same random variation that would have obtained had the overall measuring period not been segmented.

We have hitherto neglected the time required for the "internal" mode phase measurement. This is because the internal signal may be up to 60dB stronger than the weakest usable external signal. Accordingly, even if only 10% of the available time is used for measuring the internal phase (i.e. 30ms) the contribution to the random phase error from the internal signal will be only 1% and can be totally neglected. The overall measuring cycle could be apportioned thus:

Internal	External			
	Positive Pattern		Negative pattern	
	Normal	Inverted	Normal	Inverted
20ms	70ms	70ms	70ms	70ms

To eliminate any possibility of phase drift between the time of the internal and external measurements, it might be better to distribute the measurement:

External		Internal	External	
Positive			Negative	
Normal	Inverted		Normal	Inverted
70ms	70ms	20ms	70ms	70ms

It is important to ensure that the nature of the phase measurement is such that the internal phase can be measured in 20ms (i.e. only 20 cycles of the comparison frequency) without round-off error.

Receiver noise considerations

Combining the results of sections 3.9 and 4.6 it can be seen that a power level of about 5nW incident on the photodetector results in a measured signal-to-noise ratio of 32dB in a noise bandwidth of 100Hz at the receiver output. To provide a basis of comparison with comparable designs in the literature, to permit a comparison of achieved and theoretically predicted receiver sensitivity and to study the implications of receiver design, it is desirable to refer this noise to the photodetector. This will give us the noise equivalent power (NEP) or level of incident radiation at the photodetector that yields an rms value of photocurrent equal in magnitude to the noise in 1Hz bandwidth. The NEP can be calculated from the received power W_r and the signal-to-noise ratio S/N measured in a noise bandwidth B from

$$NEP = \frac{W_R}{\sqrt{S/N} \times \sqrt{B}} \text{ watts}/\sqrt{\text{Hz}}$$

$$= \frac{5\times10^{-9}}{40\times10} = 1{,}2\times10^{-11} \text{ W}/\sqrt{\text{Hz}}$$

Laboratory measurement

To confirm this result in the laboratory the following experiment was undertaken. The transmitter and receiver were

removed from the instrument and placed facing one another on the laboratory bench, a variable distance apart. The photodetector was illuminated only by a direct ray from the source diode, reflected multipath rays being frustrated by the use of suitably placed baffles. The modulated component of on-axis irradiance was measured directly at 16mW/sterradian using a large area photodetector calibrated against a bolometer, as described previously. From the known area of the photodiode in the receiver, the incident power could then be calculated as a function of distance between transmitter and receiver. At the output of the receiver the signal-to-noise ratio was measured by means of the previously described combination of active filter and true-rms voltmeter. The noise bandwidth was 100Hz The photodiode was simultaneously illuminated by means of an incandescent lamp, in order to simulate the effect of background radiation. The level was adjusted to provide a steady photocurrent of about 10μA. The incident signal power on the 0,5mm diameter photodetector at a distance r is clearly

$$W = (16 \times 10^{-3}) \times \frac{\pi}{4} (0,5 \times 10^{-3})^2 \ / r^2 \text{ (watts)}$$

$$= (3,1 \times 10^{-9}) \ / r^2 \text{ (watts)}.$$

If the effective noise power at the photodetector is W_n in a 100Hz bandwidth and the signal-to-noise amplitude ratio at the output is $\sqrt{S/N}$ we have

$$\sqrt{\frac{S}{N}} = \frac{3,1 \times 10^{-9}}{r^2 W_n}$$

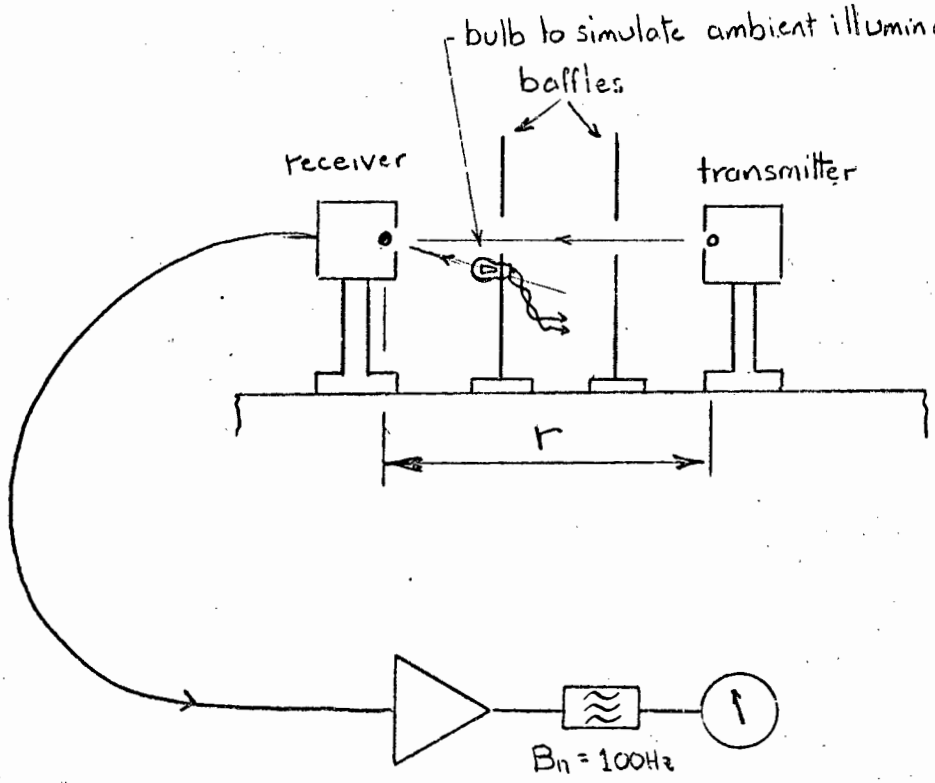
$$\therefore W_n = \frac{3,1 \times 10^{-9}}{r^2 \sqrt{\frac{S}{N}}}$$

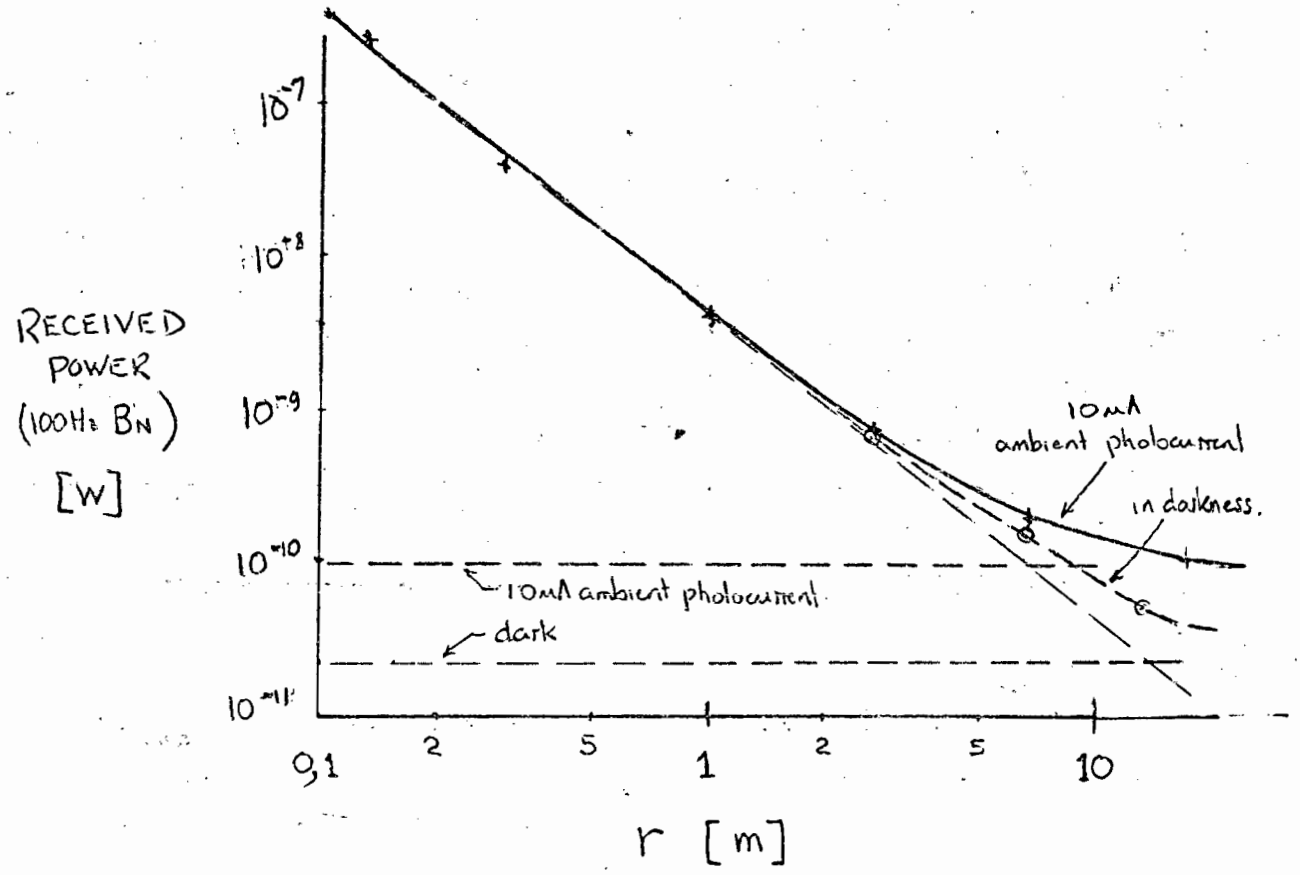
The NEP is simply the input noise in unit bandwidth $= \frac{W_n}{\sqrt{100}}$

$$\therefore \text{NEP} = \frac{3,1 \times 10^{-10}}{r^2 \sqrt{\frac{S}{N}}}$$

A 32db ($\sqrt{\frac{S}{N}} = 40$) signal-to-noise ratio was measured for a transmitter/receiver separation of 0,9m, giving

$$\text{NEP} = \frac{3,1 \times 10^{-10}}{(0,9)^2 \times 40} = 9,5 \times 10^{-12} \text{ W}/\sqrt{\text{Hz}}.$$





RECEIVED POWER vs DISTANCE r

Again we get a figure for NEP of about 10^{-11} W/$\sqrt{Hz}$. The apparent discrepancy between the 0,9 metre separation for 32dB signal-to-noise ratio and the 2,1m separation for the roughly equivalent 3,4nW received signal in section 3.9 is accounted for by the differing photodetector areas. The unmodified photodiode used in Chapter 3 had a 1mm diameter sensitive area. In the work described in this section it had been replaced by a device fitted with a 0,5mm mask or aperture in order to avoid phase delays (see section 4.7b). The 4:1 difference in areas accounts for the approximately 2:1 difference in distance for the same subtended solid angle.

If the signal + noise at the output of the receiver channel is plotted versus r on log-log graph paper a line is obtained with a negative slope corresponding to the inverse square law. At distances in excess of a few metres the signal contribution is negligible and the output tends to a fixed value (somewhat dependent on ambient illumination) corresponding to an input power of the photodetector of $\sqrt{B}$ x NEP or $\sim 10^{-10}$W. If the ambient illumination producing the 10 μA photocurrent was removed, the noise level dropped to $\sim 0,2 \times 10^{-10}$, indicating the dominance of the photodiode shot noise component, and the near optimality of the receiver preamplifier.

4.8 RECEIVE CHANNEL DESIGN

The design of the receive channel is dominated by the need to achieve adequate sensitivity, or a sufficiently low noise-effective-power (NEP). This section discusses the factors influencing the design, and shows that the theoretically predicted NEP for a simple photodetector, agrees closely with the figure of about 10^{-11}W/$\sqrt{Hz}$ determined experimentally in the preceeding section.

Photodetector noise

The topic of photodetector NEP has been very extensively treated in the literature (74-76) particularly in recent years on account of the resurgence of interest in electro-optic communications (77). Many of the treatments are very complex, being based on photon statistical treatments. This complexity is quite unnecessary for the present purposes and it will be seen that a very simple model suffices and agrees well with the experimental results.

Photodetector requirements for distance measurement differ in two important ways from those for optical communication:

(i) Although high-frequency modulated light may have to be detected, in the case of distance measurement there is no wide bandwidth requirement as there necessarily is when information is being transmitted (the rate of information transfer in the case of a static distance determination is strictly infinitesimal). Thus a resonant load can be used for distance measurement and the capacity of the detector and input circuitry 'tuned out'. The limitation of required bandwidth allows the use of relatively high photodetector load impedances.

(ii) In the case of a fibre-optic data link the noise limit is set either by the shot noise associated with the detector leakage or dark-current, or ultimately by the quantum fluctuations of the light-signal itself. This cannot be achieved in a distance measuring application, where the optical link is an open atmospheric one, and the noise limit is set by the shot noise associated with the photocurrent due to background radiation - particularly scattered sunlight.

A combination of (i) and (ii) implies that we are almost inevitably in a background limited noise situation in the case of distance measurement. The high load impedance and high background radiation together ensure that shot noise dominates both Johnson and excess receiver noise (the latter is minimised in any event by the relativity high modulation frequency involved). The situation can be modified somewhat by using an optical filter to decrease the photocurrent due to background radiation reaching the detector. The minimum filter bandwidth that can be used is dictated by the linewidth - and more significantly wavelength vs temperature - of the source. Reference to the literature (78) indicated that this was 0,4μm. Since this is about one tenth the half-power spectral bandwidth of the detector alone, one would expect a reduction in background radiation and resulting photocurrent of about an order of magnitude. A simple measurement with a commercial Bausch and Lomb 400 Angstrom filter confirmed that this was the case,

background photocurrent falling from a typical 3μA to a typical 0,3μA with the filter in front of the photodetector. The signal was attenuated some 10% or 1dB.

Avalanche photodetection

It was thought that the use of a narrow-band interference filter might make it possible to exploit the relativity noise-free current gain available from the use of an avalanche photodetector (79), which is the solid-state equivalent of a photomultiplier, and a series of experiments was carried out using a Texas Instruments TIXL56 avalanche photodiode. Further details are given in Appendix 4.7 and here we will give the conclusion and point out some features of interest.

1) The main conclusion is that the use of avalanche photodetection is only of marginal utility (a) if the bandwidth requirement is less than many magahertz (b) if an appreciable part of the noise at the receiver input is due to shot noise resulting from photodetector dark current or ambient illumination. This is basically because such noise is amplified along with the signal by the avalanche multiplication process. This is borne out by results presented in the literature (80). The result of the experimental work was to show that, with an optimum 400 Angstrom filter, in average conditions of ambient illumination (moderately bright sunlight scattered off a choppy sea) the improvement in signal-to-noise ratio resulting from avalanche detection of 150kHz modulated light was about 3dB. This represents an increase of range of some 40% in conditions of near perfect visibility and considerably less (perhaps 10%) in more typical conditions. Such a marginal improvement is certainly outweighed by the attendant complications and disadvantages of the avalanche device. This would <u>not</u> necessarily apply in the case of an instrument designed for high resolution distance measurement and employing modulation frequencies in the tens-of-megahertz region, or in the case of an instrument for which considerably extended night-time range was an operational advantage.

2) In order to achieve the required uniformity over the junction and avoid excess noise due to micro-avalanching, avalanche photodetectors invariably have small photosensitive areas. The TIXL56 has a 0,25mm diameter sensitive area. This is not necessarily a disadvantage but it does make the requirements on optical performance and mechanical alignment more stringent. The device showed the same kind of spatial variation of phase delay between the modulation envelope of the radiation and the resulting photocurrent as the PIN junction photodiode. Due to the small junction area and rather awkward construction it was not possible to fit a mask to improve the situation. It seems that a custom-made device with a mask fitted by the manufacturers might be needed for accurate distance measurement.

3) A problem in using the avalanche photodetector is that it requires a large, critical, temperature-dependent bias potential of around 150V. The problem of making this potential vary with temperature in the required manner is usually solved by fabricating two identical diodes on a common silicon chip and using one of them exclusively as a reference for the bias voltage. This works well, but the dual device is prohibitively expensive, and commitment to its use further restricts the already limited choice of possible devices. A better method was suggested by a proposal in the literature (81). It depends on the fact that, for a wide range of conditions, optimum signal-to-noise ratio is obtained for a definite value of avalanche current multiplication factor - typically about 100. To set the bias potential at a value that will yield this value, a control loop is used. The magnitude of noise (which is also magnified by the avalanche gain) in a band of frequencies adjacent to the modulation frequency is monitored and the bias potential is automatically increased until the noise contributed by the photo-detector just exceeds the circuit noise contributed by the preamplifier to which the detector is coupled. This is the operating point which yields nearly the optimum signal-to-noise ratio. Moreover the system

is fully adaptive, compensating both for varying temperature and varying background illumination. The method is closely analogous to, but not identical with, that proposed by Raines et al (81). It worked extremely well and seems to be a very promising technique for use generally in electro-optic distance measuring instruments.

4) The voltage dependent gain of the avalanche photodetector permits its use as a combined amplifier/mixer. This opens up new circuit possibilities. If the bias potential is modulated by the super-imposition of a small alternating voltage from a local oscillator, the photocurrent will contain the mixing products of the input signal and the local oscillator waveform, and the photodetector can in effect be tuned over a wide range simply by varying the local oscillator frequency. The effect has been discussed in the literature (82) and adopted in practice for at least one commercial instrument (16). The mode of operation is advantageous when several different high modulation frequencies have to be accommodated, as in high resolution distance measuring instruments. Among other things, it avoids the complication of having multiple or switched resonant circuits as the photodetector load impedance - a critical point in the circuit especially vulnerable to induced crosstalk.

Thus it can be seen that, although avalanche photodetection is highly advantageous in some forms of distance measuring equipment, it is at best marginally so in the present instrument and any argument there may be for its adoption is outweighed on grounds of cost, circuit complexity and additional phase error. Similar considerations apply to the case for incorporating an interference filter in front of a PIN detector. As we shall see, the signal-to-noise improvement is small, and the cost is considerable. Simple filters such as the Kodak gelatin filter "Wratten 87C" (which is optimised for GaAs wavelengths) are inexpensive and perhaps worth including but the signal-to-noise improvement is very small. Typically ambient photocurrent is reduced by a factor 2-3 and a signal loss of 30% is experienced.

Intrinsic PIN photodetector performance

Responsivity:

The PIN photodetector is a PN junction in which the depletion layer capacitance has been reduced by including a region of compensated intrinsic semiconductor between the P and the N regions. This is significant only at frequencies in excess of a few hundred megahertz (83). For our purposes it may be considered identical to a simple PN junction. Incident photons create electron-hole pairs and the charge carriers are swept out by the high field in the depletion region. The fraction of photons which actually creates electron-hole pairs is fairly constant at about 0,6. Responsivity R in amps of photocurrent per watt incident on the detector can be calculated from

$$R = \frac{\eta \lambda q}{hc} \quad A/W$$

where λ = radiation wavelength ($0,9 \times 10^{-6}$ m)
h = Planck's constant ($6,6 \times 10^{-34}$ J s)
q = electronic charge ($1,6 \times 10^{-16}$ C)
c = velocity of light (3×10^{8} m/s)

giving
$$R = \frac{0,6 \times 0,9 \times 10^{-6} \times 10^{-19}}{6,6 \times 10^{-34} \times 3 \times 10^{8}}$$
$$= 0,44 \text{ A/W}.$$

Due to surface reflection losses the actual responsivity is slightly less than this. For the diode used (HP4207) the manufacturer quotes a responsivity of 0,35A/W. An experiment was carried out in which the diode was illuminated by a GaAs diode and the photocurrent measured using a 1MΩ load and synchronous detector. A calibrated bolometer was then substituted for the detector and the incident radiation measured directly, yielding the following result:

Photodetector area = $(\pi/4)(0,5 \times 10^{-3})^2 = 2 \times 10^{-7}$ m^2
Power incident on bolometer (100mm^2 area) = $1,3 \times 10^{-4}$ W
Power incident on photodetector = $2,6 \times 10^{-7}$ W
Voltage developed across 1MΩ load resistor = 1,0V
∴ Photocurrent = $1,0 \times 10^{-6}$ A
∴ Responsivity = 3,8 A/W

Shot noise

The rms noise associated with a photocurrent I_t in a bandwidth B is given by

$$I_n = \sqrt{2qI_tB} \text{ (A)}$$

or, in 1Hz bandwidth,

$$5{,}6 \times 10^{-10} \sqrt{I_t} \text{ (A)}$$

The photocurrent I_t is made up of the steady component of modulated radiation (completely negligible for weak signals, which is the case that concerns us here), photodetector dark current, which is extremely small for the HP4207 (2×10^{-9}A) and can also be neglected, and photocurrent due to background radiation.

The literature abounds in vague and often contradictory measures of typical background radiation (84-86). In most cases figures are quoted for the visible range and simply extrapolated without explicit justification into the infra-red region. It was thought advisable to investigate the matter of background radiation experimentally. The instrument was pointed at a wide variety of targets (including a choppy and foaming sea surface in conditions of bright sunlight) and the resulting photocurrent measured. The current ranged from 2 to 8μA corresponding to incident radiant power of about 6 to 23μW. Thus we have for the worst case shot noise

$$I_{n\ max} = 1{,}6\text{pA} .$$

At frequencies above a few kilohertz, excess noise is negligible so we have for the photodiode alone, an effective NEP given by that incident power which would yield a photocurrent equal in magnitude to the shot noise in 1Hz bandwidth

$$\text{i.e. NEP} = 1{,}6 \times 10^{-12} \times 3{,}8 = 6 \times 10^{-12} \text{ W}/\sqrt{\text{Hz}} .$$

Although this is the true NEP for the detector itself, the NEP which we would infer from S/N measurements at the output of the receiver channel (post mixing) would be twice as great. This is because the receive channel is a heterodyne system, with OdB image rejection since desired and image response are only 2kHz apart and both are well within the input bandpass range. We therefore have an excellent correlation between predicted and measured NEP (10^{-11} W/$\sqrt{\text{Hz}}$). It may seem rather surprising that all the noise is contributed by the photodiode, and the receiver preamplifier appears to have a OdB noise figure. It is however quite reasonable, on account of the

high background radiation level, causing shot noise which swamps all other noise sources. If we were to operate in an environment with less background radiation, or incorporate a narrowband optical filter, we should begin to see the effects of receiver noise.

The assertion that photodiode shot noise dominates the total noise contribution is demonstrated in the next section.

Circuit considerations

Two approaches have been followed in the design of photodetector amplifiers for electro-optic communications. In an approach advocated by Personick at Bell Telephone Laboratories (87) the load for the photodetector is the very high input impedance presented by a field-effect transistor (FET) (88). Frequency distortion due to the capacitive component of the load is compensated subsequently by equalisation. The traditional (and more widely employed) approach is to minimise the effect of capacitance by coupling the photodetector current into the virtual-earth point of a transresistance amplifier. Both approaches yield similar signal-to-noise ratio performance.

In the case of distance measuring instruments the usual approach has been the former version, as load capacitance can be tuned out, exploiting the high-frequency narrowband nature of the signal. If the total load capacitance is C and the allowable Q-factor of the input tuned circuit is Q, the realisable load resistance will be

$$R_{Load} = \frac{Q}{2\pi f C} .$$

Input tuned-circuit coupling offers certain other advantages. A low D.C. bias source resistance is presented to the diode, avoiding large changes in bias potential as ambient illumination varies, and phase-inversion for crosstalk error cancellation is readily achieved by means of a bifilar secondary winding. A major disadvantage is the vulnerability to induced contaminating signal by stray flux coupling.

A tuned preamplifier

In the early stages of development, when emphasis was on crosstalk error rather than NEP, it was thought essential to use a resonant input circuit, and the measurements recounted in section 4.6 were carried out with such a receiver.

It was later realised however that the receiver performance was far from optimum. The reason is that it is not possible to realise a high Q factor simultaneously with a tuning capacitance of say 10pF, on account of the stray capacitance associated with the rather large inductor required. In fact the best achieved was a Q of about 10 with 20pf capacitance. This represents a load resistance of

$$R_L = \frac{10}{2\pi \times 150 \times 10^3 \times 20 \times 10^{-2}} \quad \text{(ohms)}$$

$$\approx 0{,}5\ \text{M}\Omega \ .$$

The noise current contributed by a FET of transconductance g_m for a 1MΩ source resistance is given approximately by (88):

$$I_n = \sqrt{\frac{2kT}{R}\ \frac{0{,}7}{g_m}} \quad \text{(amps)}$$

Taking g_m equal to at least 1mS we have

$$I_n \leq \sqrt{\frac{2 \times 4 \times 10^{-21}}{10^6}} \cdot \frac{0{,}7}{10^{-3}} = 2{,}3 \times 10^{-12}\ \text{A}.$$

Thus the FET noise current exceeds that of the photodiode shot noise, even when the photocurrent is 8μA (2,3 pA vs 1,6pA).

The Johnson noise in 1Hz bandwidth associated with the ½MΩ resistor is negligible:

$$I_{NJ} = \sqrt{\frac{4kT}{R}} \quad \text{(amps)}$$

$$= \sqrt{\frac{4 \times 4 \times 10^{-21}}{0{,}5 \times 10^6}} \quad \text{A}$$

$$= 1{,}8 \times 10^{-13} \quad \text{A} \ .$$

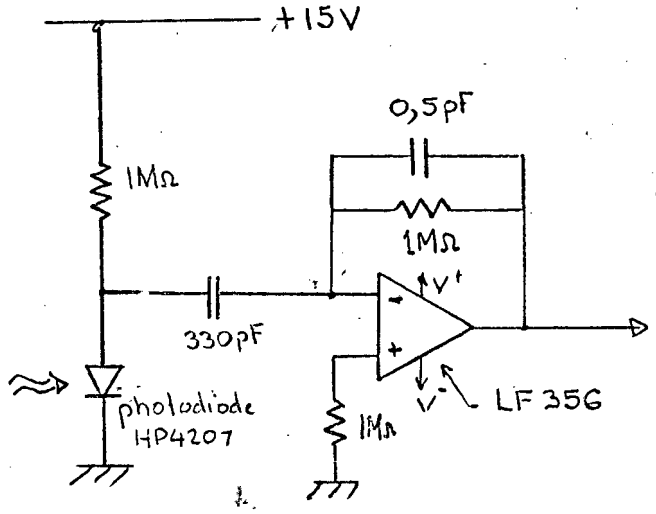
The noise figure of the preamplifier is given by

$$NF = 20 \log \frac{2{,}3 + 1{,}6}{1{,}6} \quad \text{dB}$$

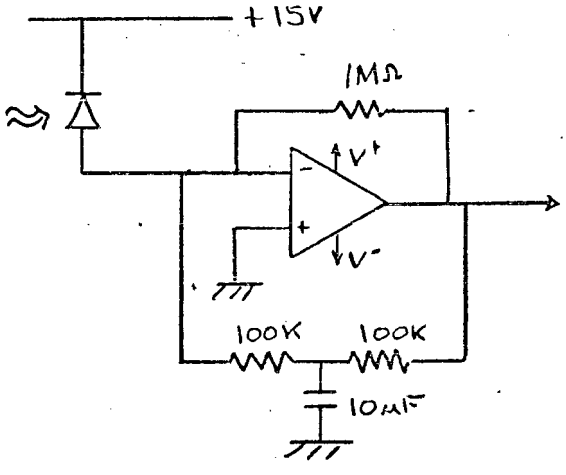
$$= 20 \log 2{,}4$$

$$= 7{,}7\ \text{dB}$$

and the effective NEP is degraded to $2{,}4 \times 10^{-11}\ \text{W}/\sqrt{\text{Hz}}$. To this would have to be added the referred noise from subsequent stages.



A TRANSRESISTANCE PREAMPLIFIER.



ALTERNATIVE ARRANGEMENT.

(less variation in photodiode bias voltage due to ambient photocurrent)

A transresistance preamplifier

For a modulation frequency as low as 150kHz the transresistance approach offers effective competition to a tuned stage. If we assume that the capacitance of the photodetector, amplifier input and circuit strays total 10pF we can calculate the maximum effective resistance that must be presented to the photodiode as

$$R_{max} = \frac{1}{2\pi . f. \ C}$$

$$= \frac{1}{2\pi \times 150 \times 10^{3} \times 10^{-11}} \text{ (ohms)}$$

$$= 10 \ k\Omega .$$

If an amplifier is used with a gain/bandwidth product of 15MHz the gain at 150kHz will be approximately $100\angle 90^{o} = j100$.

Let the feedback resistor be R_F. The impedance from the virtual earth point to ground will be approximately

$$R_{IN} \doteq \frac{R_F}{j100}$$

$$= - jR_F/100 .$$

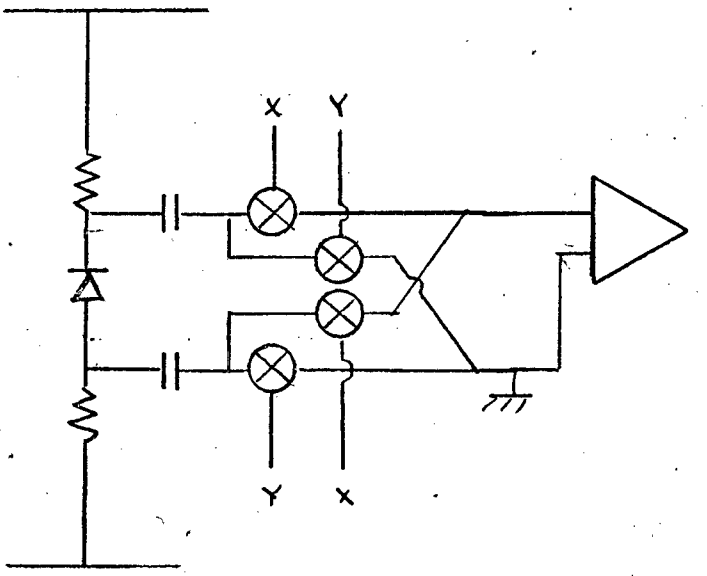
The input is therefore inductive with a magnitude $R_F/100$. The maximum allowable value of R_F is therefore 1 MΩ . The noise current associated with feedback resistor R_F

$$= \sqrt{\frac{4 \ kT}{R}} \quad (A/\sqrt{Hz})$$

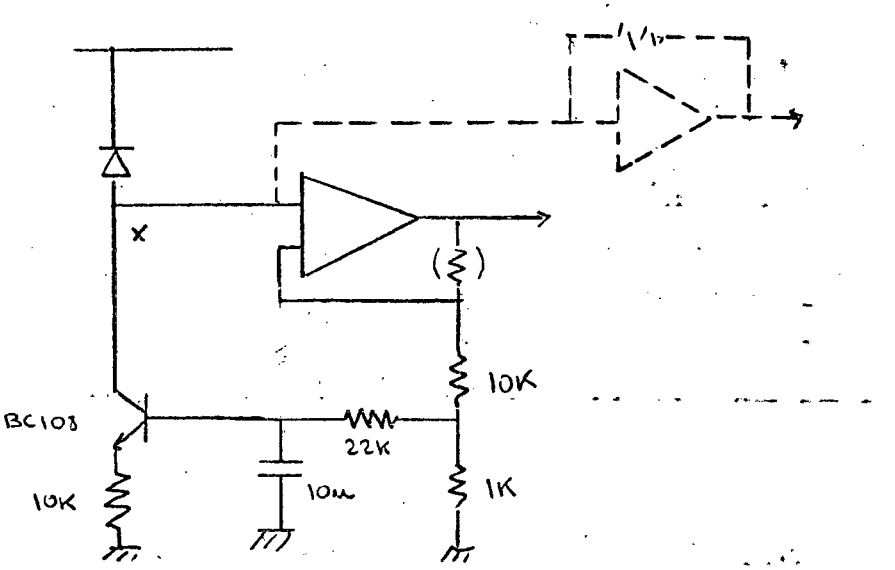
$$= \sqrt{\frac{4 \times 4 \times 10^{-21}}{10^{6}}}$$

$$= 1,3 \times 10^{-13} \quad A/\sqrt{Hz}$$

while the noise input current for a typical FET input operational amplifier, say the LF 356 is about $10^{-14} A/\sqrt{Hz}$. Thus we would expect the shot current noise to be dominant and indeed this proves to be the case. This was the configuration for which an effective NEP of $10^{-11} W/\sqrt{Hz}$ was measured in section 4.7.



SIGNAL INVERSION WITH A CMOS BILATERAL SWITCH



ACTIVE PHOTODIODE BIAS CONTROL

(Simulates large inductance at x)

Comparative evaluation

It can be seen that, in terms of noise performance, the transresistance amplifier has an advantage over the tuned amplifier. Lacking an input resonant circuit, it is also less vulnerable to inductively radiated crosstalk and seems in many ways the superior choice. The variation of bias potential with ambient illumination can be overcome by providing a decoupled D.C. feedback path, but a major problem is how to incorporate the signal inversion feature. It might be argued that the elimination of an input tuned circuit greatly reduces the problem of crosstalk and the direct/inverted measurement procedure might not be necessary in this case, but it still seems desirable to incorporate so powerful an error-cancelling feature if at all possible. It is, of course, possible to effect phase inversion by coupling the photodiode via a DPDT reversing switch to the amplifier input. The switch can be realised using diode or CMOS switches. It is however necessary to couple capacitively between the photo-diode and amplifier when using the switch, and we encounter the problem of providing a D.C. bias path to the photodetector without shunting the signal path with a resistor which will impair the signal-to-noise ratio.

The problem can be solved by providing a transistor in a feedback loop which sinks the direct photocurrent provided by the photodetector while presenting a very high A.C. impedance to the photodetector. The circuit can also be considered as a simulated inductor, which presents a low D.C. and high A.C. load impedance. The circuit was breadboarded and performed well, the only drawback being an additional source of noise current, which was measured at about 1,4pA in 1Hz bandwidth. Thus the receiver noise figure is impaired some 3dB.

Conclusion

Three viable approaches to the design of the photodetector preamplifier have been presented. Two of them offer good performance with a noise penalty on the order of 3dB (i.e. the tuned load and the system described immediately above). A third alternative - a straight transresistance amplifier with a decoupled feedback path for D.C. - offers superior noise performance but does not offer the possibility of signal

inversion. The purpose of this thesis is not to present a completed design but to study the possibilities and demonstrate the viability of the proposed system. Thus it is not felt necessary to make a final choice between the competing systems. Given the availability of the fibre-coupled lasing device discussed in Chapter 3, any of these approaches would perform well with a more than adequate margin of safety. In all cases satisfactory agreement was obtained between predicted and measured quantities and sufficient data has been amassed and interpreted to enable a paper design of a distance measuring instrument to be undertaken with confidence.

CHAPTER 5

THE AZIMUTH-ANGLE MEASURING SUBSYSTEM

5.1 INTRODUCTION

The measurement of angle differs in a fundamental way from that of length, in that angle is dimensionless, being defined in terms of the subdivision of the circle, rather than an arbitrary standard of reference. It will be seen in this section that this fact makes possible the use of some particularly elegant and powerful techniques for its measurement leading to a simple and easily made angle transducer capable of a very high accuracy.

Traditionally, the measurement of angle in land survey has been accomplished by means of the transit, or the more accurate theodolite, which is an opto-mechanical instrument containing an optically divided glass circle. The resolution and accuracy attainable is of the order of one arc second, or somewhat better than one part per million, which greatly exceeds the accuracy of most electrical-output angle transducers of comparable size.

In view of the importance of angle measurement in fields ranging from metrology to automatic industrial control and the aerospace industry, there has been a long history of development of electrical angle transducers, variously referred to as resolvers, shaft position indicators, encoders, rotary transducers, shaft digitisers and synchros. Attempts have been made to adapt these devices to survey applications in order to provide for the theodolite the considerable advantage of an automatic digital readout. These attempts however have thus far failed to produce an electronic theodolite which can compete in terms of accuracy, size or cost with the classical theodolite.

A survey of the literature and of commercially available transducers disclosed that, while transducers of single second accuracy do exist, their very high cost and large size preclude their use in survey instruments except in very special circumstances. Even in the 10-20 arc second accuracy category the cost and bulk of available transducers is such as to render them far from ideal for the proposed instrument, and it was decided to develop a suitable transducer, having if possible

an output in the form of the phase of an electrical signal. This would render it compatible with the distance-measuring sub-system and make possible a sharing of much of the signal processing and readout circuitry of the latter, with consequent minimisation of component count and power consumption, in addition to the obvious saving of both cost and space.

5.2 PRESENT REQUIREMENT

Accuracy

The overall azimuth-angle-measuring subsystem consists of a manually-aimed telescope and angle transducer. Field experience and a study of the performance of gunsight telescopes suggest that a reasonable compromise between angular resolution and angular field-of-view (for rapid target acquisition) is afforded by an optical system with a magnification between 10 and 20. The precision with which such a telescope can rapidly be pointed at a well-defined aiming point was determined to be about 50mm standard deviation at 1km distance, which corresponds to a standard deviation of angle of 10 arc seconds. Assuming that angular error due to pointing is uncorrelated with that due to the transducer, the precision required of the latter (to meet the overall specification of 20 arc seconds) is about

$$\sqrt{20^2 - 10^2} \text{ arc seconds}$$

$$= 17 \text{ arc seconds}.$$

However, since at least part of the error due to the transducer is likely to be periodic rather than random, it is probably safer to aim at a root-mean-square error for the transducer of better than 10 arc seconds.

To facilitate automatic recording and minimise the likelihood of gross error it is desirable that the azimuth angle transducer should have a complete 360° unambiguous range.

It is worth noting that if an identical transducer system were to be adopted for the vertical angle determination, rather greater precision would be desirable to allow for error in the vertical reference system. To compensate for this, however, only a limited range (say $\pm 30^\circ$) would be required.

Speed of response

Since the target specification calls for a complete update at least once per second, and the phase-measuring and readout circuitry will be shared by the distance measuring subsystem and the azimuth and vertical angle-measuring subsystems, it is reasonable to divide the available time evenly and call for a 0,3 second response time from each subsystem. This leaves 0,1 sec for system housekeeping and computation.

Form of output

To achieve compatibility with the distance measuring subsystem, the output should be in the form of the mutual phase displacement between two periodic signals, having a frequency of about 1kHz. The phase displacement should be a linear function of the angular input to the transducer.

5.3 SURVEY OF ANGULAR DISPLACEMENT TRANSDUCERS

Scope of the survey

Many of the devices described in manufacturers' catalogues and in the literature are no more than indicators of small angular displacement; for example the differential interferometer, the optical lever, the auto-collimator, and the rotary differential transformer. Others again are low-resolution indicators of angular position based on devices such as the strain-gauge and the rotary potentiometer. Such devices have no relevance to the present requirement and are disregarded in the survey which follows. The survey attempts an orderly exposition of all transducers described in the literature which have a potential accuracy commensurate with the present requirement, in order to facilitate and justify the approach taken.

Previous surveys

No truly comprehensive review of angle transducers seems to have been published in the past twenty years, although many techniques are discussed in standard texts on metrology and electronic instrumentation. A paper published by Brewer in 1963 (89) covers some of the techniques but a more complete treatment is given by Sydenham (90). Instrument technology in the USSR (91) differs significantly from Western practice in several ways, and two papers by Yeliseyev (92;93) are of considerable interest.

Much of the material that follows was gleaned from the primary sources cited in these papers and from manufacturers' literature.

Categories of angle transducer

The survey will discuss primary or static transducers, which contain no moving parts other than the shaft representing the angular input, and secondary or dynamic transducers which contain a uniformly rotating element. The primary transducers will be subdivided into analogue and digital. High resolution transducers typically produce an ambiguous output (i.e. an output which cycles through its range for an angular input less than 360°). This is not a serious limitation, however, since the ambiguity can easily be removed by referring to a simple visual scale (a protractor) or automatically by incorporating a low resolution transducer on a common shaft and logically combining the outputs of the two transducers.

Primary transducers (analogue)

The Resolver

One of the oldest and best known angle transducers is the resolver or two-phase synchro. This is really a precisely constructed two-phase wound-rotor induction motor operated under locked rotor conditions. The orthogonal stator windings are driven in mutual time quadrature, giving rise to a spatially rotating magnetic field. The e.m.f. induced in the rotor by this field has a phase relationship to the drive waveform which is a linear function of the angular position of the rotor.

The resolver merits special consideration as an angle transducer in the present application since its output is precisely in the form required. In its conventional form, however, it is deficient in accuracy. Most commercial resolvers have an accuracy of about one thousandth of a rotation or 20 arc minutes. The most accurate resolvers achieve an accuracy of from 3-6 arc minutes. The accuracy seems to be limited by mechanical imperfections such as bearing eccentricity and by the difficulty of achieving the accurately sinusoidal spatial distribution of gap flux which is required for a pure harmonic-free rotating field. The cost of resolvers accordingly rises steeply with demanded accuracy.

It should be noted that even in the case of a perfect two-phase resolver (i.e. one free of error), the one-to-one correspondence between angular rotation and phase angle would demand of the phase measuring circuitry an accuracy of better than 10/(360 x 3600) or about 1 part in $1{,}3 \times 10^5$. It will be shown that, contrary to general opinion, this order of precision can actually be achieved, but it certainly represents a considerable challenge, and a transducer which made less stringent demands on the phase-measuring system would be preferable.

Modifications to improve accuracy - the Multipole Resolver

The variation of phase output with angular rotation may of course be increased by increasing the spatial frequency of the rotating field by increasing the number of pole-pairs, and multipole resolvers having up to 24 poles have been made. However, while increasing the resolution of the system, this approach does not necessarily improve the accuracy - in fact multipole synchros tend to have larger errors than the two-phase resolver. An indirect approach whereby a highly accurate multipole resolver can be made, is however possible.

The Vernier Resolver

In 1963, Kronecker (94) described a synchro-resolver having one more pole on the rotor than on the stator, and employing an interference effect analogous to the operation of a vernier scale to obtain enhanced resolution. Moreover the accuracy was improved by spatial averaging to a reported 10 arc seconds. This approach was seriously considered, but eventually rejected for two reasons:

i) The signal processing required precise analogue circuitry involving accurate amplitude comparisons and

ii) The transducer required accurate machining and it was felt that the mechanical fabrication on a one-off basis would present great difficulties.

For these reasons, and the non-compatibility of output, this otherwise promising approach was not pursued further, but it should be noted that in view of the paucity of precise and economic angle transducers and the frequent requirement for them, the lack of exploitation of Kroneckers' elegant solution is surprising.

The Capacitive Multipole Resolver

The task of fabricating a precise multipole resolver is greatly eased if the device is based on capacitive, rather than magnetic, principles. Highly accurate patterns can be produced on glass discs using photo-lithographic techniques. A pair of such sinusoidal tracks in mutual space quadrature is adjacent to an output track. The sinusoidal tracks are driven in time quadrature and electrostatically coupled to the output track by a set of radial sectors whose angular position represents input to the sensor. It is easily shown that the output phase will be a linear function of the input angle θ

$$\phi = n\theta$$

where n is the number of complete cycles around the circle.

This transducer exploits a powerful principle of spatial averaging. Since the angle about a point is always precisely 2π, it follows that the algebraic sum of the location errors of circular graduation marks must always be zero. Thus, in so far as the signal contributions from various parts of the disc are evenly weighted, position errors in the patterns do not lead to gross output non-linearity. It is necessary to add the qualification "gross" because within a single period $\frac{360^{o}}{n}$ the linearity will of course depend on the accuracy of the sinusoidal pattern, as well as second order effects such as fringing fields.

This transducer was first described by Radenbush in 1958 (95). A precision of about 6 arc seconds was claimed for the prototype, but no details of size were given. Careful consideration was given to this approach and a model system constructed for evaluation but it was felt that the difficulties of manufacture to the required mechanical precision were too great. It should be noted that the transducer is today widely used in numerically controlled machines and the cost of a single-second version is several thousand rand. More recently a similar transducer using inductive coupling between photo-lithographically-produced discs has been developed and marketed under the name "Inductosyn".

Other Analogue Transducers

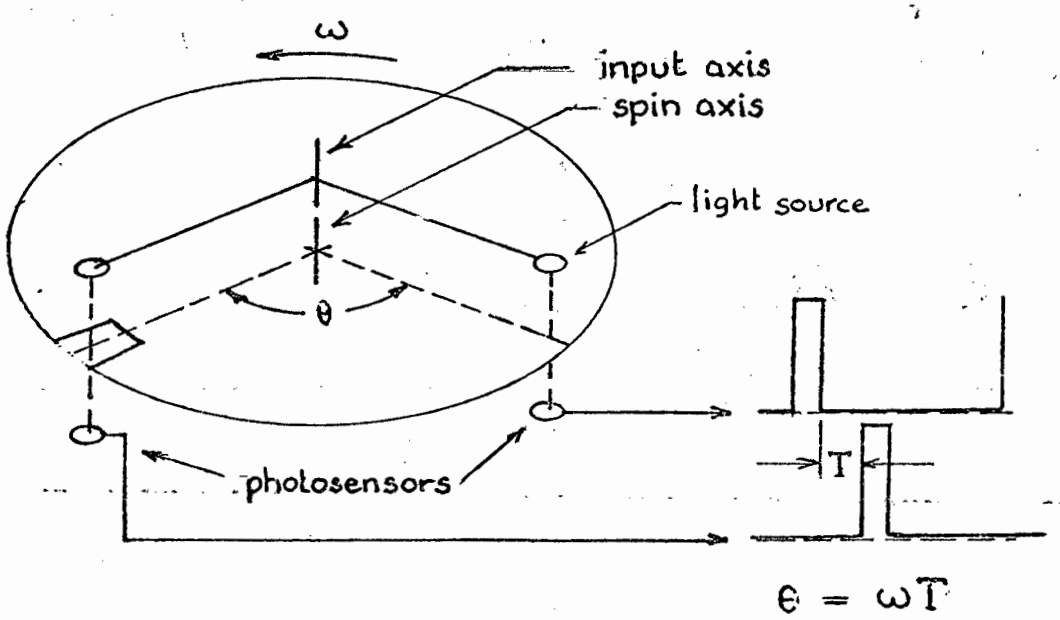
Two other devices of high accuracy deserve mention. Both employ spatial averaging techniques:

A complex optical angle transducer employing spatial averaging was described by Belyy and Ivanov in the Soviet Journal of Optical Technology in 1973 (96). The method is ingenious and an accuracy of 0,2" is claimed. The transducer is however extremely bulky (650mm diameter) and obviously very difficult to manufacture. It seems clear that it has no application in a portable survey instrument and is included here only for completeness and as another instance of the power of spatial averaging techniques.

Another device employing spatial averaging is a precision rotary capacitor transducer developed in Canada by Makov et al (97). The method depends on the precise measurement of differential capacitance (to a precision of 10^{-7}pF). Again the technique is bulky and incompatible with the present requirement but the fact that a rotary capacitor can be constructed so as to yield single-second accuracy yields some insight into the powerful error-cancelling mechanisms that can be invoked in the case of angle measurement and is eloquent testimony to the power of symmetry.

Primary Transducers (Digital)

Optical digital encoders are frequently used for high resolution angle determination. The principles of operation are well known and will not be treated here. Most small encoders (up to 50mm diameter by 50mm long) have an accuracy of no better than 20 arc minutes. Gray-coded discs having a resolution of 1 in 2^{16} are made, corresponding to a least count of about 20 arc seconds. However, 2^{-16} resolution requires 65 536 divisions and necessitates a large disc (typically 20cm diameter) if the divisions are to be large enough to pass a reasonable amount of light. Some improvement and a moderate degree of interpolation between divisions may be obtained at the expense of increased cost and complexity by resorting to multiple optical sensor arrays and Moiré techniques. An ingenious interpolation technique has been developed by the Hewlett-Packard Company (98) but it is very complex and depends on extremely precise manufacture. Extreme mechanical precision is also required to prevent eccentricity errors due to bearing imperfections.



YELISEYEV'S PROPOSAL

Digital transducers were rejected for the present application on grounds of high cost, large size and non-compatibility of output. It is worth noting however that in recent years digital transducers with analogue interpolation have appeared with 22 bit resolution (Itek, Digisec, Zeiss) and genuine single second accuracy. They are all of large diameter (typically greater than 250mm) and cost around R10 000.

Secondary Transducers

The static angle transducers described above may be considered as primary transducers in that angle is converted directly to electrical output, whether analogue or digital. There exists in addition a class of secondary transducers in which time is involved as an intermediate variable, a continuously rotating element within the transducer permitting time-domain interpolation between circle divisions. This important class of transducers has received scant attention in the technical literature and little in the way of commercial exploitation.

Two forms of the transducer may be distinguished:

The Chronometric measurement of angle:

In 1964, Yeliseyev (93) discussed the possibility of an angle measuring system in which the spatial graduation marks on a fixed circle were replaced by time-marks generated by the transition of a single mark on a uniformly rotating disc past a suitable sensor. The time displacement between the sensor's output pulses, and those of a fixed reference sensor is a function of the angular displacement of the two sensors around the spin axis of the disc, and speed of rotation. If the disc rotates at a rate of ω radians per second, the time for a complete rotation will be $2\pi/\omega$. The time delay T between the pulses produced by the two sensors when they are separated by an angle θ will be given by $T = \theta/\omega$. Yeliseyev states that accuracy down to a single arc second should be possible, but gives no indication of size, and makes extremely stringent demands on constancy of speed of rotation. His error analysis is incomplete, and no experimental results are given.

The Dual Modulator Angle Transducer

Angle measuring systems which have, in common with the chronometric transducer, the provision of a continuously rotating element have been described by De Bey and Webb (99) and Sydenham (100). These transducers for which Sydenham has coined the name "dual-modulator" transducers consist basically of two independent a.c. generators having their rotors driven at a constant rate. The phase displacement between the electrical outputs of the generators is a linear function of the relative angular displacement of the stators. Clearly if the generators each produce, say, 360 cycles of output for each rotation, a mutual phase displacement of 1^{o} between stators will produce 360^{o} of electrical phase shift. Such a device will in itself be incapable of determining the number of integral degrees of displacement (this must be accomplished by a coarse backing system) but it can be seen that a quite modest phase resolution of 1 part in 3600 will suffice to yield an angular resolution of 1".

Versions described in the literature:

The first transducer to use time-domain interpolation obtained by means of a rotating element was that described by de Bey and Webb in 1958 (99). The transducer consists of a toothed rotor spinning inside a pair of similarly toothed stators, one of which is fixed and the other free to rotate about the spin axis. A fixed bias potential is applied between the rotors and stator, and an alternating potential is produced on rotation by the capacitance variation.

The device reported by Sydenham eight years later (100) is similar, but uses radial optical gratings to produce the alternating signals. The gratings are illuminated by a diffuse light source and the light passing through the grating detected by a set of large area photodetectors positioned behind a fixed matching grating.

Brief mention has also been made in Sydenham's review article of a magnetic variable reluctance version, embodying in addition Kronecker's vernier principle. This device was marketed under the trade name Vernisyn but seems to be obsolete and no account appears in the open literature.

Critical evaluation

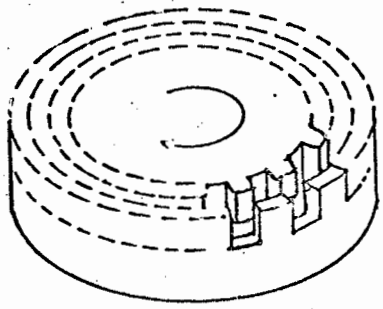
The chronometric and dual-modulator transducers share the feature of interpolation in the time domain. They differ in that the output of the former is in the form of a time interval between unique events, whereas the dual-modulator transducers employ both spatial and time averaging, having an output in the form of the mean phase delay of a signal which is produced by the integrated contributions of a large number of elements distributed around the circumference of the input stator. These differences, as will be seen below, lead to important differences in performance, design constraints and sources of error.

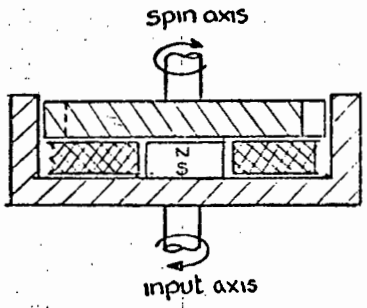
Conclusion

The high potential accuracy, inherent error-cancelling properties leading to non-critical fabrication constraints, and the output in terms of phase, constitute a very strong a-priori case in favour of some form of dual modulator transducer in the present requirement. Accordingly a fairly extensive theoretical and experimental investigation of such devices was undertaken. This work is summarised in the appendices and in the sections which follow.

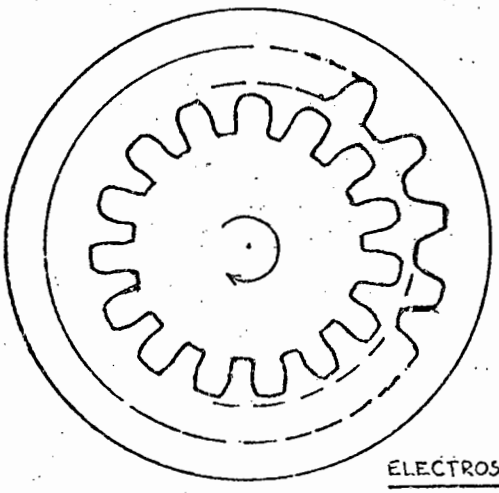
5.4 STUDY OF DUAL MODULATOR TRANSDUCERS

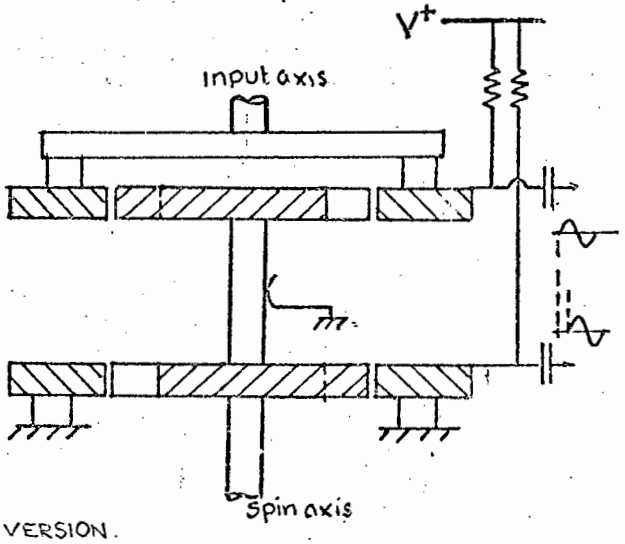
In order to gain physical insight into the error mechanism of this class of transducer, and in order to provide hardware against which theoretical models could be verified, it was decided to build capacitive, magnetic (variable reluctance) and electro-optic versions. Mechanical work was minimised by designing compatible systems which could fit interchangeably on a basic mechanical test-bed which consisted of a base containing a small d.c. motor and bearing-supported shaft, and an alidade mounted on a precision bearing system and free to rotate concentrically with the motor driven shaft. Special methods were developed to test the quality of the bearing system and it was determined that a bearing concentricity (peak radial runout) of 10μm and peak-to-peak tilt of about 5 arc seconds was achieved. Later in the course of the development a more comprehensive mechanical test-bed became available, having provision for independent adjustment of tilt and eccentricity in addition to concentric rotation, but the design of this instrument was carried out with the help of a mechanical engineering colleague and does not form



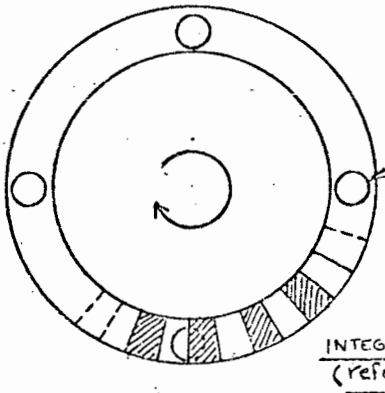


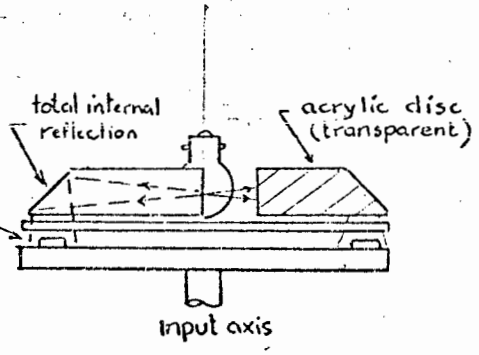
VARIABLE RELUCTANCE VERSION





ELECTROSTATIC VERSION.





INTEGRATING PHOTOELECTRIC VERSION
(reference sensors not shown)

part of this thesis.

Details of experimental systems

Some of the hardware is shown in the accompanying diagrams. The magnetic system used a 90-tooth mild steel gear wheel of 70mm diameter running inside a mild steel drum furnished with 90 radial slots. The transducer was completed by a 1000 turn solenoid and ceramic magnet. Relative rotation of gear and drum produced reluctance variations with a resultant change in flux linking the coil and, therefore, produced an electrical output.

The capacitive system used a matching internal and external gear of 80 teeth machined to leave a 150μm gap. Provision was made for attaching one stator to the base and the other to the alidade (normal operation) or alternatively for mounting both stators in fixed relative orientation on the alidade. This "locked stator" test proved to be a very powerful and important one. Clearly under these conditions the phase displacement between the signals from the two stators should have a relative phase displacement which is fixed and unaffected by rotation of the alidade as well as by small tilts and lateral translations. Deviations from this condition of phase invariance are a sensitive indicator of error, permitting the transducer to be evaluated for absolute phase accuracy without recourse to any external standard or angle-measuring instrument of known high accuracy. This is another example of invoking the unique self-checking properties of angle transducers mentioned in the opening paragraph of this chapter.

The electro-optic versions are closely patterned on Sydenham's version, except that attempts were made to devise a more efficient illuminating system.

Results of initial experimentation

The aim of this early work was familiarisation with the potential and limitations of the technique and the evaluation of a theoretical model which would permit the rational design of a transducer. Several possible error sources which are not discussed in the literature immediately came to light:

i) Crosstalk: The effect of crosstalk, or spurious electrical coupling between outputs, is to superimpose a sinusoidal perturbation upon the intended linear relationship between angle and phase. The effect is particularly pronounced in the magnetic and capacitive versions due to the difficulty of preventing coupling through stray and fringing fields. The effect is discussed in Appendix 4.4, where it is shown that a high degree of cancellation is possible if the phase of a signal can be inverted by 180° without inverting the phase of the spurious component. This could readily be done in the case of the capacitive version by changing the polarity of the stator bias potential. The mean phase measured with the stator biased alternately positive and negative would be relatively free of cross-talk error.

An alternative approach to eliminating this source of error is to use a hybrid system, deriving the two signals by different methods (e.g. capacitive/magnetic, magnetic/electro-optic, etc.). It was soon realised that the reference sensor, being fixed relative to the spin rotor, is entirely uncritical and the simplest electro-optic system will suffice, whereas the requirements for the moving sensor system, which is subject to second-order tilts and displacements as the alidade rotates, are much more stringent. The adoption of this hybrid approach unfortunately aggravates the effect discussed in the next paragraph.

ii) Frequency dependent phase shift: If the source impedance of the sensor systems is complex, there will be a phase shift of the sensor output which is dependent on the speed of the rotor. The effect is discussed further in Appendix 5.2. It will be seen that if the transfer functions of the two sensor systems are matched, rotor speed variation will not result in any relative (differential) phase shift. Clearly it is considerably easier to achieve such a match if the two systems are identical. This is a contra-indication in respect of the hybrid system mentioned above. The problem is not a particularly serious one, however, since rotor speed can be closely controlled, if required, by means of a pulse-tachometer feedback servo, or the use of a

small synchronous motor.

Phase offset

If long term absolute stability is required of the transducer (i.e. if it is required to reproduce a given phase displacement for a given direction of the alidade relative to the base), it is possible that slowly varying parameters of the transfer function of the sensor systems will cause a drift in the relation between angular displacement and phase difference.

$$\phi = k\theta + \delta$$

It was realised that an elegant and simple technique exists which eliminates this error completely, as well as doubling the resolution of the system. If the direction of rotation of the motor be reversed, clearly the sense of phase displacement with respect to alidade rotation is reversed, while the phase error δ is unchanged. Thus we have:

$$\phi' = -k\theta + \delta$$

On subtracting we have an expression from which the error has been eliminated:

$$\phi - \phi' = 2k\theta$$

Thus the mean of two readings with motor reversed is free of phase offset error. This is potentially a powerful technique, but care must be taken that the rotor speed is identical in both directions and that the transducer is not disturbed by the reaction torque on motor reversal. The technique is particularly relevant in respect of an angular transducer for the determination of vertical angle, since in this case the determination is an absolute one, the reference direction being the local gravitational gradient. Drift in the apparent reference direction due to electrical phase shift is completely eliminated. It is also worth noting the effective doubling of resolution. This technique does not seem to have been described previously in the literature.

Eccentricity

The effect of alidade bearing eccentricity is a complex function of sensor properties and transducer geometry and is discussed in greater detail in Appendix 5.3. Their significance can best be summarised, and design of the transducer facilitated, by drawing attention to the fact that there exists a number of binary choices:

Disc or drum: The geometry of the transducer may be planar or cylindrical. It is shown in Appendix 5.3 that the cylindrical configuration has some advantages in respect of eccentricity tolerance.

Proximity or intersection sensing: An important distinction in terms of performance and sources of error relates to the mode of action of the sensor. Capacitive sensors generate their peak output at the instant of nearest approach to the rotor index mark, whereas an electro-optic sensor has a definite "line-of-action" and the significant event in this case is the crossing of this line by the index mark. Proximity type sensors have the limitation that their output amplitude is a function of distance and is therefore affected by alidade bearing eccentricity. Intersection type sensors yield an output which is not a function of lateral displacements. If, however, the line-of-action is not radial with respect to the rotor, phase errors will be generated by eccentric movement and tilt.

Discrete point or integrating sensors: The two limiting cases here are the electro-optic sensors on the one hand, and the capacitative sensor with ring-geometry stator on the other. It can be seen from Appendix 5.3 that complete cancellation of errors arising from alidade eccentricity can be achieved by averaging the output of a diametrally-disposed pair of sensors. If the averaging is performed by simple analogue summing of the sensor output, however, the cancellation depends upon precise amplitude balance. In the capacitive integrating case one is critically dependent on the balance of the output of proximity-type sensors, and hence on accuracy of gap width.

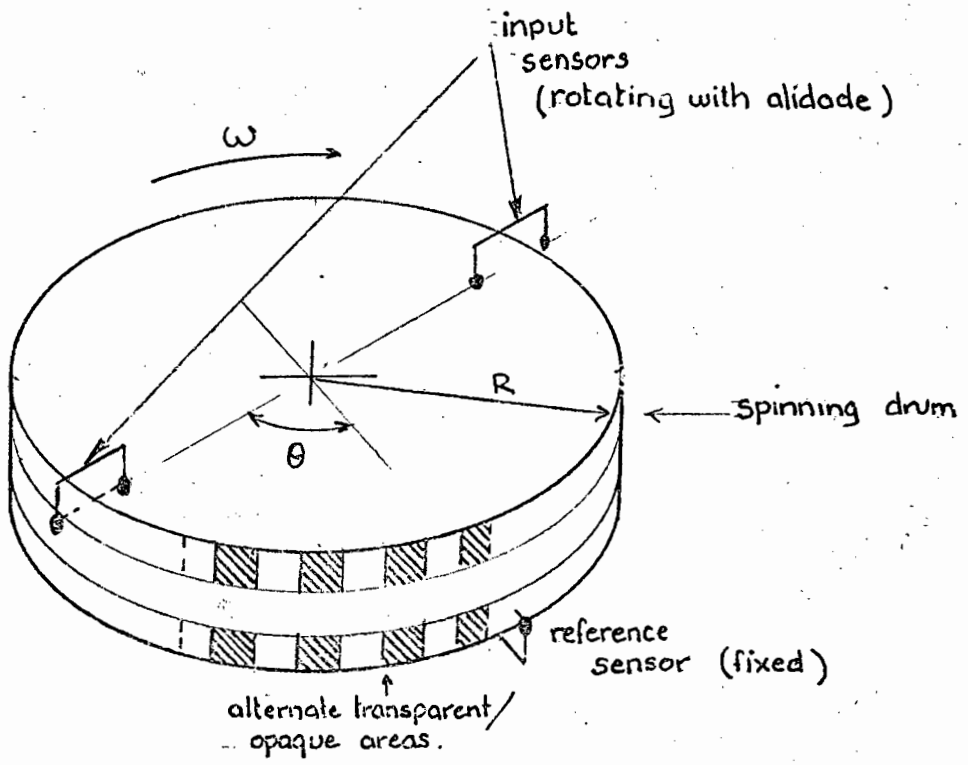
In practice, severe problems due to eccentricity were experienced with the integrating capacitive device. This was exacerbated by the small gap width necessary to obtain appreciable output, on account of fringing effects. The variable reluctance device was somewhat more tolerant in respect of gap width, but here, too, eccentricity problems were experienced. In addition, spurious phase shifts were experienced due to permanent magnetism and external magnetic fields, and the inherently reactive source impedance led to a stringent requirement in respect of motor speed constancy.

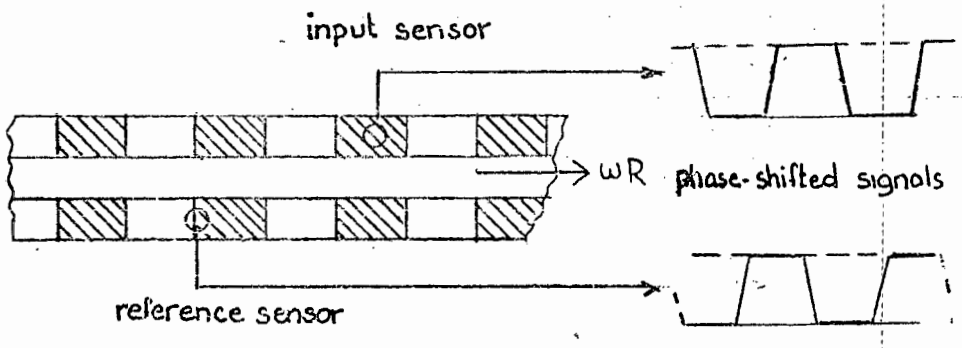
Comment on published versions

In the light of the analysis and experimentation, it was evident that neither de Bey and Webb nor Sydenham had fully exploited the potential of the dual modulator transducer. In terms of the analysis presented above, de Bey and Webb's device can be described as having cylindrical geometry and integrating proximity-type sensing. Accordingly it requires accurate machining and alignment. Sydenham's version is a compromise. It uses large area photodetectors and a diffuse light source. Accordingly it is partly integrating, and the sensors have properties intermediate between proximity and intersection type devices. Unfortunately it inherits the weaknesses of both. The amplitude is fairly strongly a function of the spacings between the discs, and the phase of the signal generated depends upon directionality in the illuminating source. It is very difficult to achieve truly diffuse illumination and this is no doubt the reason for Sydenham's use of a multiplicity of incandescent bulbs in a diffusing cavity - an arrangement wasteful of both space and power.

In view of these considerations, it was felt that an optimum dual modulator for the present application would have cylindrical geometry and discrete electro-optic sensors. Early experiments showed that, even in this case, amplitude balance was excessively critical, and it was decided to measure the phase displacement of each diametrally-opposed sensor separately with respect to the fixed reference sensor and average the results numerically, using for the tests an on-line minicomputer. This strategy proved highly successful, resulting in complete cancellation of the primary eccentricity error, and is the subject of a patent application. It should be emphasised that the principle involved here is the averaging of the outputs of discrete sensors in a way which is not sensitive to their output amplitude. There are many ways of doing this, and the approach is not restricted to digital numerical averaging, which is merely one such method, particularly well suited to the early experimental phase of the work in that it is a versatile and highly transparent method, permitting comprehensive evaluation of sources of error.

It was therefore decided to proceed with the experimental design and testing of the transducer along these lines.





THE PROPOSED TRANSDUCER.

5.5 DESIGN OF A NEW ANGULAR DISPLACEMENT TRANSDUCER

Description of the proposed transducer

The transducer consists of a cylindrical rotor carrying an equi-spaced set of graduations or index marks, which rotates with a constant angular volocity ω radians/sec. The index marks consist of alternately transparent and opaque bars parallel to the axis about which the rotor spins (called the "spin axis"). An electrical signal having the form of a trapezoidal wave is generated by the bars as they pass between a fixed sensor (the "reference sensor") consisting of an infra-red-emitting diode and phototransistor. The line joining the centroid of the emitting area to that of the photosensitive region of the phototransistor is called the "line-of-action" of the sensor, and it is radial with respect to the rotor - that is to say the line-of-action when produced passes through the spin axis.

A pair of identical sensors A and B arranged on a diameter of the rotor is free to rotate about an axis which is nominally collinear with the spin axis. This axis is called the "input axis" and rotation of the sensor system about this axis constitutes the angular input to the transducer. To provide error checking capability a further set of sensors C and D was arranged on a diameter orthogonal to AB.

Principle of operation

Consider a single input sensor and assume that the angular displacement of this sensor about the spin axis, relative to the reference sensor, is θ.

The reference and input sensors will each produce a periodic electrical output signal of frequency:

$$f = \frac{n\omega}{2\pi} \quad \text{Hz},$$

where n is the number of rotor divisions.

The signals will have a relative time displacement τ where

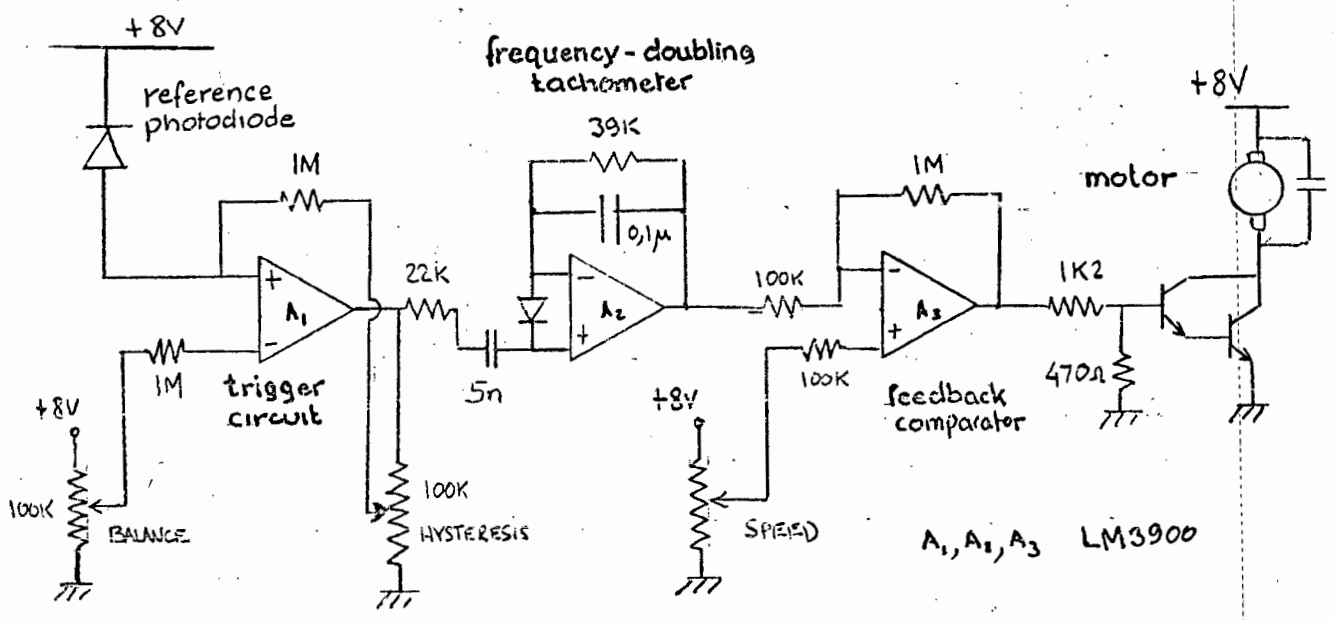
$$\tau = \frac{\theta}{\omega} \quad \text{(seconds)}.$$

We may write this as a phase delay ϕ where

$$\phi = 2\pi \frac{\tau}{T}$$

$$= 2\pi \frac{\theta}{\omega} \frac{n\omega}{2\pi}$$

$$= n\theta .$$



PRECISISION MOTOR SPEED CONTROL

Thus the output phase angle is linearly related to the angular displacement between the two sensors, and the constant of proportionality (or magnification factor) is simply n, the number of rotor divisions. Since ϕ is periodic in 2π, the output will be unambiguous for input angles which are a multiple of $(2\pi/n)$.

Practical implementation

The rotor was made by attaching a suitably divided film strip with 200 divisions to a metal drum of 50mm diameter. This was spun at about 2000 rpm by a small permanent magnet d.c. motor, yielding output frequencies in the vicinity of 1000 Hz.

The drum and sensors were mounted on the mechanical test bed assembly described above. Two orthogonal sets of input sensors were used. The infra-red emitting diodes were biased at 50mA and spaced 10mm from the phototransistors, yielding a peak photocurrent of about 1mA.

The signals were processed by a Krohn-hite digital phase meter, and later, by a specially developed microprocessor-based phase-measuring system. The processing aspects are discussed in more detail in Chapter 6.

Difficulties were at first experienced with erratic speed variations in the motor, which was intended for use in a tape recorder. These were due to the action of the governor, which was removed and replaced by a feedback control system, using a pulse-counting tachometer actuated by the signal from the reference sensor.

The signals from the sensors were about $1V_{rms}$ when the rotor was stationary, or the light path blocked. Noise was about $100\mu V$. The resulting signal-to-noise ratio of 80db was adequate for a phase measurement to $0{,}006^{o}$. In practice, rotor eccentricity and graduation error gave rise to a periodic phase perturbation of about 1^{o}. However, by gating the phase measuring system so that measurements were averaged over an integral number of rotor revolutions, this error was completely eliminated. One of the problems inherent in the fabrication of a drum-type transducer is in the discontinuity which results from the inevitable discrepancy between the length of the film strip and the circumference of the supporting drum. This results in a phase discontinuity in each rotation of the drum.

Experiments were carried out on the possibility of providing a resilient element with provision for adjusting the effective diameter of the drum. All this, however, proved unnecessary. By taking reasonable care in manufacture, and allowing for the thickness of the film strip when turning the supporting drum, the error was reduced to negligible proportions. In full scale production it is unlikely that film strip would be used. The graduated cylinder would most likely be produced directly on a sensitized glass or perspex drum using an indexing mechanism and photographic methods, and so the problem would not arise.

In any event the effect of the phase-jump (provided it is less than half a modulation period) is merely to complicate somewhat the signal processing. No error in angle determination results as long as the phase measurement is averaged over an integral number of drum rotations.

Initial Evaluation

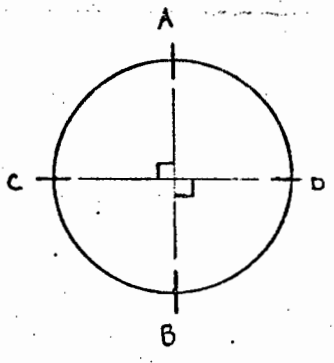
Resolution and repeatability

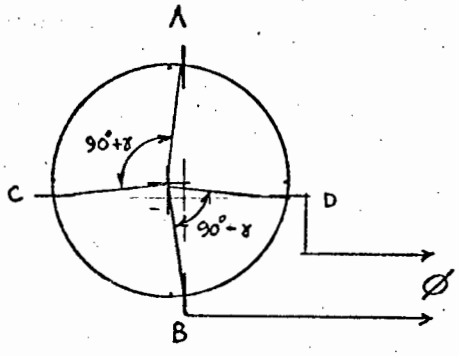
Initially, digital phase readings on the Krohn-hite phase-meter were wildly erratic. To establish the cause, the analogue output of the phase-meter was examined using an oscilloscope. It was found that the phase angle between the reference signal V_R and any one of the signal sensors exhibited a peak periodic swing of about 5^o, suggesting a rotor eccentricity ε_r given by

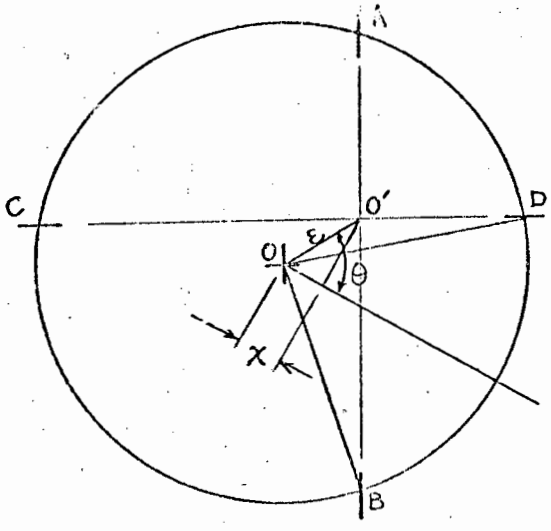
$$\varepsilon_r = (25 \times 10^{-3}) \times \tan\left(\frac{5}{200}\right)^o = 10^{-4}\text{m}$$

Eccentricity of this order is hardly surprising - indeed it is remarkably small given the method of manufacture. This explanation was confirmed by noting that the peak value of the sinusoidal swing was itself a cosinusoidal function of the angular orientation between sensors.

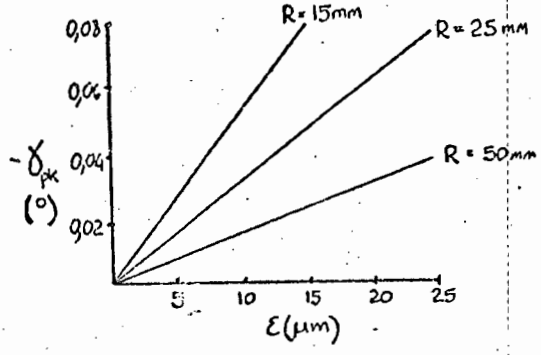
It is shown in Appendix 5.3 that provided the phase measurement is averaged over an integral number of rotor revolutions, no error results from this eccentricity. Initially, averaging was carried out simply by low-pass filtering the output of the Krohn-hite phase-meter. For later work the averaging was accomplished digitally, measuring phase over an even integral number of rotor revolutions.



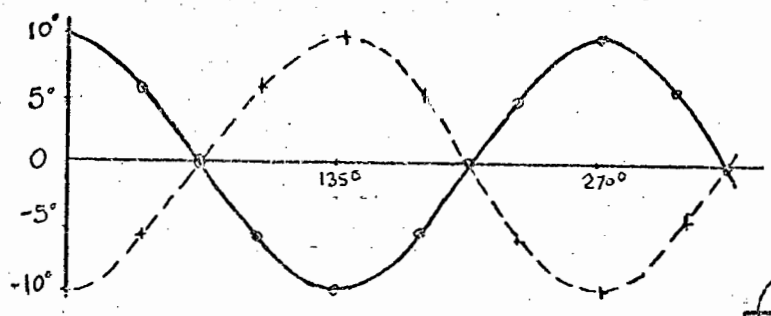


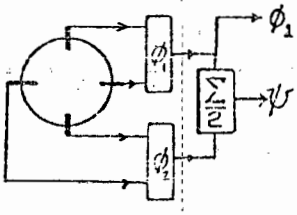


$$\chi = \varepsilon \cos\theta$$
$$B\hat{O}D = 90 - \gamma$$
$$\gamma_{max} = 90 - 2\arctan\left\{\frac{1}{1-\frac{\varepsilon}{R}}\right\}$$

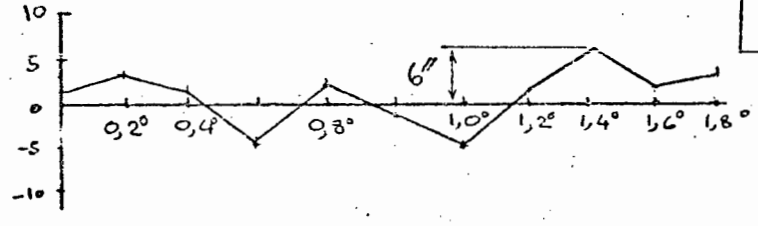


$\phi_1 = N\gamma$ (N = 200)





$\psi = \frac{1}{2}\{\Delta P_{AC} + \Delta P_{BD}\}$ (arc seconds)



RESULTS OF EXPERIMENTAL TEST FOR CONCENTRICITY.

This resulted in a repeatability of 0,05°, corresponding to a mechanical displacement of about 1 arc second. It was clear that random variations of this order were caused, not by signal-to-noise ratio but by phase-meter resolution, and, in the case of the digital system, numerical truncation errors. The error from this cause was, in any event, well within the specified limits.

Input axis eccentricity

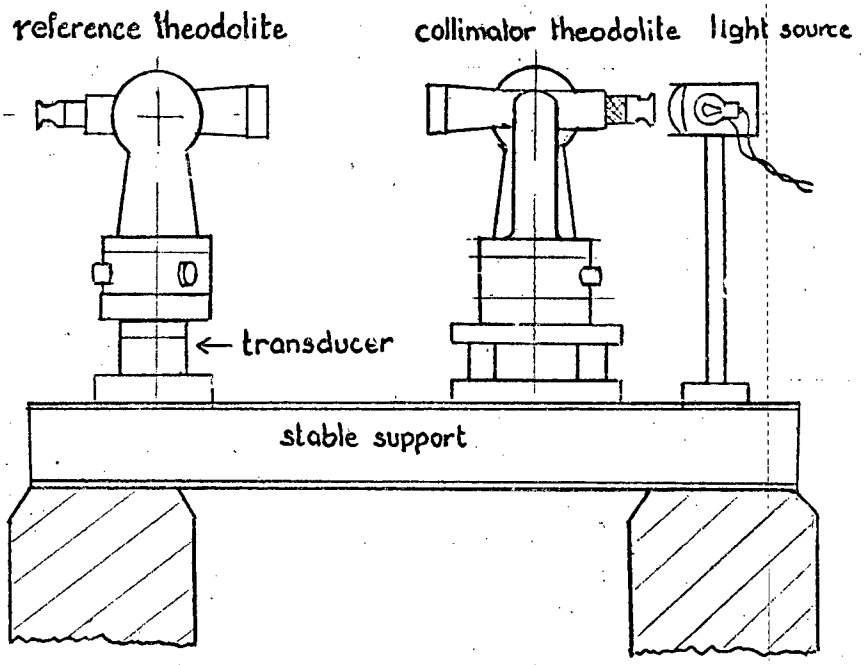
Eccentricity between input and spin axes, on the other hand, is an important potential source of error (see Appendix 5.3) and although bearing runouts had been determined as less than 10μm, it was felt appropriate to perform an in situ measurement of input/spin axis eccentricity.

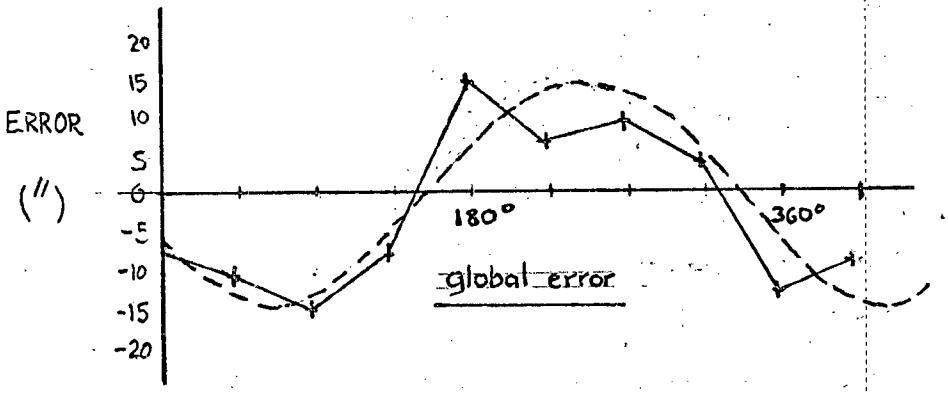
A straightforward method is shown in the accompanying figure. In the event of perfect concentricity the angle AOC or DOB remains constant (close to 90°) as the sensor system is rotated. Eccentricity gives rise to a sinusoidal perturbation of this angle, as plotted. The geometry of the situation is analysed, and it can be seen that an eccentricity of greater than 15μm exists. Such a degree of eccentricity leads to an error, for a single sensor, of 2′ and, although theoretically, perfect error cancellation exists between opposing sensors, there is the possibility of significant second order errors arising due to sensor imperfections.

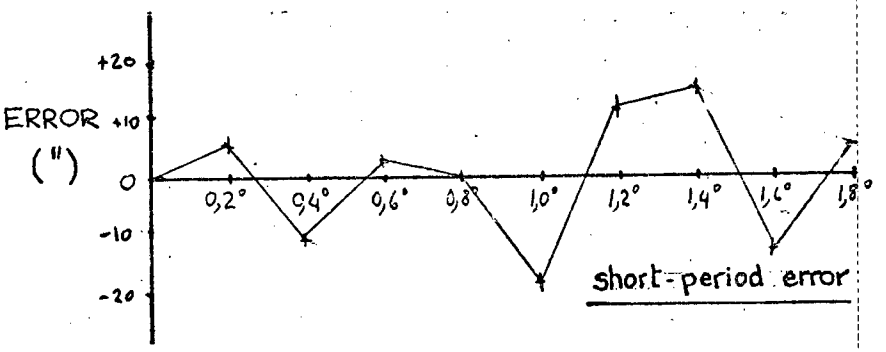
It was therefore thought desirable to minimise the eccentricity error, and this was accomplished by mechanical adjustments. The test was then repeated, with the results shown. It can be seen that the error is greatly diminished and that the errors for the two angles appear in antiphase, as might be expected.

The antiphase variation of ΔP_{AC} and ΔP_{BD} leads one to expect that the mean $\frac{1}{2}(\Delta P_{AC} + \Delta P_{BD})$ would be relatively constant. This angle is shown plotted to a greatly expanded scale. The residual peak error of 6" is an informal guide to the order of accuracy which can be expected of the transducer.

On the basis of this rather hopeful outcome, it seemed justified to proceed with more complete testing.


reference theodolite
collimator theodolite
light source
transducer
stable support


ERROR
(")
20
15
10
5
0
-5
-10
-15
-20
180°
360°
global error


ERROR
(")
+20
+10
0
-10
-20
0,2°
0,4°
0,6°
0,8°
1,0°
1,2°
1,4°
1,6°
1,8°
short-period error

Overall testing

The transducer was tested by the methods outlined in Appendix 5.4. The test setup is as shown. It can be seen from both incremental and global tests that peak error is less than 15". Root-mean-square error or standard deviation is less than 10 arc seconds.

Reference to manufacturer's literature and an intercomparison between three available T2 theodolites suggested that the instrument had a standard error of at least 2". Repeatability of observations did not however justify any attempt to disentangle transducer and theodolite errors by the methods referred to in Appendix 5.4 and references (101 and 102).

Clearly the performance of the transducer readily meets the specified requirement. A study of the nature of the errors, however, suggested that the design had in some ways been excessively conservative, and it was decided to design and test a MkII version.

In particular, it was considered that there were two modifications which were worth making leading to (possibly) reduced but still adequate performance in respect of accuracy, with attendant advantages such as smaller size and simpler manufacture.

5.6 DESIGN OF MkII PROTOTYPE TRANSDUCER

The design followed the same lines as that of the MkI version, with certain simplifications and a reduction in size.

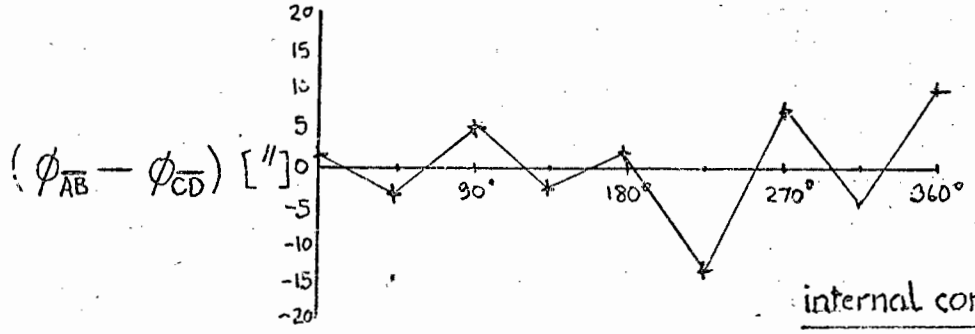
(a) It was found on the MkI prototype that the error pattern of sensors C and D was virtually identical to that of A and B. This was not surprising - indeed the extra sensors were originally incorporated only as a check. Accordingly a production model transducer need incorporate only two sensors. It should be noted however that the redundancy implied by the provision of two orthogonal sensor pairs gives a very powerful self-checking capacity to the transducer and it seems that this feature should not be lightly dispensed with.

(b) The diameter of the rotor (50mm) was considered excessive, leading to a somewhat bulky transducer of about 120mm diameter. Eccentricity-related errors are inversely proportional to rotor diameter, and study of the relative magnitudes of periodic and random error suggested that little if any accuracy would be sacrificed by a 40% reduction in rotor diameter. Further reduction would still be possible in terms of the required specification, but would lead to a device in which fabrication, adjustments and maintenance became somewhat difficult. It was therefore decided that the rotor should have a diameter of 30mm.

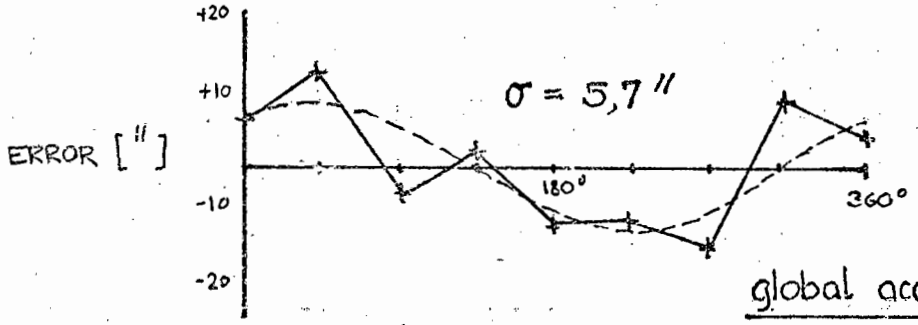
(c) The provision of 200 index marks on the rotor leads to an unambiguous range of only $1,8^{o}$. This in turn leads to a rather severe requirement in respect of the coarse 'back-up' system, and unnecessarily so, since the performance of the transducer if far from limited by the available precision of phase measurement. Moreover the marks are already, of necessity, rather small, demanding correspondingly small apertures in the electro-optic sensors. This situation would be exacerbated by the proposed reduction in rotor diameter.

To meet these objections and yet retain sufficient precision, it was decided to set the number of divisions at 32, giving an unambiguous range of $11,25^{o}$. 32 was chosen rather than, say 30 or 36 (which would have given a 12^{o} or 10^{o} range rather than the somewhat odd-looking $11,25^{o}$), to facilitate binary numerical operations when a microprocessor is used, for example, to fit together coarse and fine determinations.

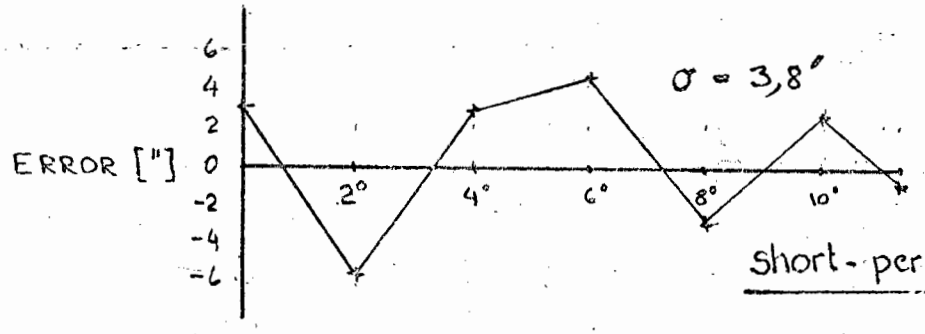
In order to provide information for ambiguity resolution an additional pair of sensors (one stationary, one moving) was provided to monitor the transit of a single index mark. The relative timing of the pulse received from these two sensors provides "coarse" information on angular displacement without any ambiguity over the complete 360^{o}. This was observed to operate correctly. The accuracy requirement on the coarse determination of angle in order to avoid an ambiguity error of $11,25^{o}$ is that the error should be less than half the period of the fine measurement, i.e. $5,625^{o}$. In practice a precision of say 2^{o} affords an adequate safety margin and this of course presents no problems. There is no need even



internal consistency test



global accuracy test.



short-period error

RESULTS OF FINAL TRANSDUCER TESTS.

for a pair of sensors since the expected eccentricity error with a single moving sensor is less than

$$\text{arc tan } (30 \times 10^{-6}/15 \times 10^{-3}) \text{ or } 0{,}11^{\circ}.$$

Fitting together the coarse and fine measurements to produce an overall result is a trivial operation in modular arithmetic which can be performed by a simple digital system. Due to extensive prior experience with such systems it was judged unnecessary actually to construct one. In any event the complete instrument would probably be under microprocessor control, and the requisite operations could readily be carried out numerically by means of a simple algorithm.

Evaluation

Testing followed the same sequence as that adopted in the case of the MkI prototype. Eccentricity determination was carried out by measuring the phase between signals from adjacent sensors. This transducer had no means provided for adjusting the relative concentricity of spin and input axes, and reasonably careful machining resulted in a net eccentricity of no more than 50μm.

An internal consistency check was carried out in the following manner. Refer to diagram. The phase of V_A, V_B, V_C and V_D was measured against the reference signal V_R.

Let the result be ϕ_A, ϕ_B, ϕ_C, ϕ_D. Now the average phases $\frac{1}{2}(\phi_A + \phi_B)$ and $\frac{1}{2}(\phi_C + \phi_D)$ were computed. Let these be $\phi_{\overline{AB}}$ and $\phi_{\overline{CD}}$.

As shown in Appendix 5.3, $\phi_{\overline{AB}}$ and $\phi_{\overline{CD}}$ are free from eccentricity error.

Hence $(\phi_{\overline{AB}} - \phi_{\overline{CD}})$ should be constant within the accuracy of the transducer.

This is a necessary but not sufficient requirement. It is however a valuable first check on transducer performance. The results are shown.

Finally, the transducer was tested against a T2 theodolite. The results, as can be seen, are well within the required specification.

5.7 CONCLUSION

The transducer which was finally developed easily meets the requirement for a small, precise angle transducer requiring no very critical fabrication. It is able to achieve this, and to outperform previously published versions based on similar principles because of careful attention to detail based on a new and comprehensive error analysis. Its principal novel (and patentable) feature is the use of a multiplicity of discrete sensors with digital (or other amplitude-insensitive) averaging. The principle could be extended to transducers of much greater (sub-second) precision and it is hoped subsequently to carry out work in this direction.

CHAPTER 6
PHASE MEASUREMENT

6.1 INTRODUCTION

The precise measurement of electrical phase angle is crucial to the measurement of both range and angle using the subsystems described in previous chapters. This chapter will consider the performance required of the phase measuring system by each of the subsystems, review phase instrumentation techniques, analyse sources of error, propose a design for a suitable system and present experimental evidence of the performance of the proposed design.

A note on the definition of phase:

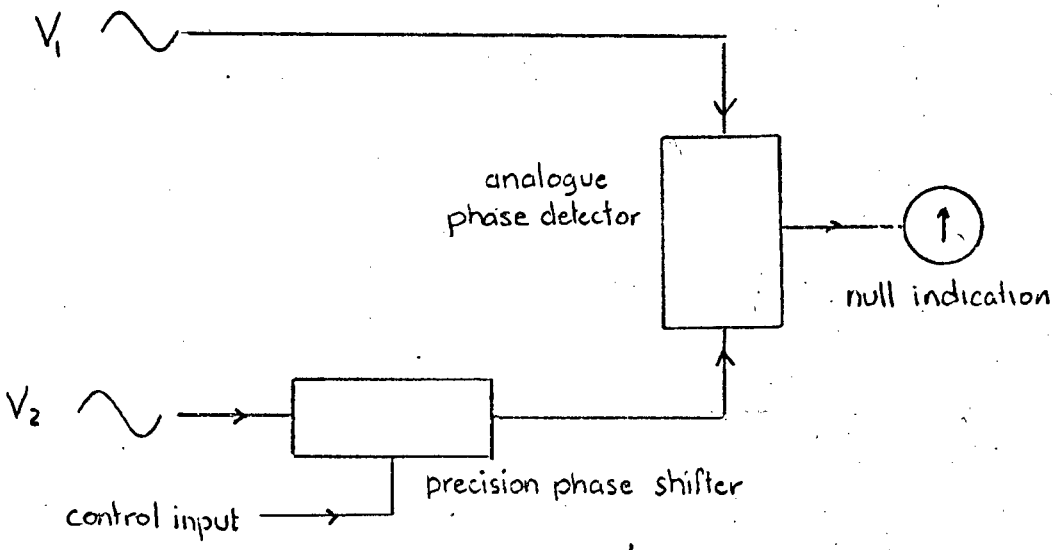
Strictly, relative phase is defined in electrical engineering only for two sinusoidal signals of the same frequency. For example the phase angle between two signals $A \sin \omega t$ and $A \sin (\omega t + \phi)$ is given by the angle ϕ. Clearly $\phi + 2n\pi$ is indistinguishable from ϕ therefore phase is periodic in 2π radian intervals. It is, however, possible to generalise the concept to cover the case of any periodic signal undergoing pure transport lag (linear phase shift).

Consider any repetitive function of time with period T, i.e. $f(t) = f(t + T)$. Consider the signal $f(t)$ delayed by time τ. If the signals are identical in all respects other than their mutual translation by time τ we may say that they have a mutual phase displacement ϕ where:

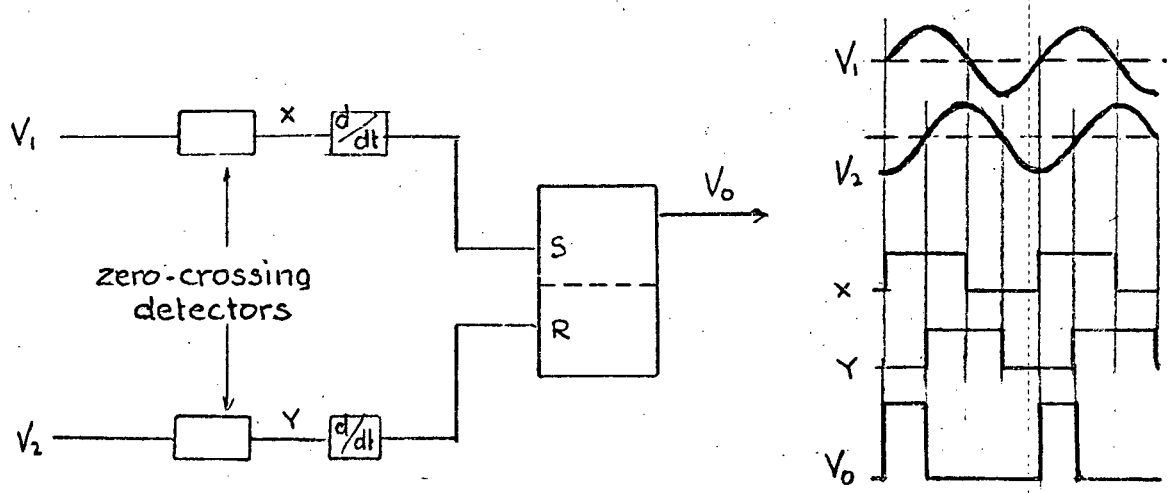
$$\phi = 2\pi \frac{\tau}{T} .$$

Clearly when $f(t) = A \sin \omega t$, phase as defined above has the normal meaning ϕ. It we determine phase in relation to the instants of zero-crossing (as is conventional) we may still speak of phase $2\pi \frac{\tau}{T}$ when the signals differ in amplitude.

The only limitation is that the waveforms be identical so that the instant of zero-crossing is consistently defined for the two waveforms. Thus it is always possible to ascribe a meaningful phase shift to any waveform which undergoes pure translation in time. Such phase shift may be measured by successive comparison with any arbitrary waveform of the same zero-crossing rate.



COMPENSATION / NULL PHASE DETECTION



PHASE / DUTY-CYCLE CONVERSION.

In much of what follows, particularly that relating to angle measurement, it is the broader interpretation of "phase" which is intended, and we are not limited to sinusoidal signals. In the case of distance measurement, waveform is not invariant and we are restricted to the more usual interpretation of phase shift - i.e. as that of the fundamental Fourier component of a signal.

6.2 PHASE MEASUREMENT REQUIREMENTS

Range measurement

In the section on the choice of modulation frequency, it was assumed that a precision of phase measurement of 1 part in 3×10^3 was attained. This amounts to $0{,}1^\circ$. The measurement should be complete in 0,3 seconds. A further requirement which must be considered is that of dynamic range, since the returned signal is a strong function of distance. Experience suggests that a dynamic range of 60dB might be needed, but this may readily be reduced to, say 20db by AGC methods (see Section 6.6).

Angle measurement

For the case considered (32 drum divisions), a precision of 10 arc seconds requires a phase precision of 1 part in 4×10^3. However since the signals are internally generated there is no dynamic range problem, and it is found that the phase measuring requirements of the angle measuring system are trivial in comparison with those of the range measuring subsystem. It is the latter which actually dominates the design of the phase measuring circuitry, although the former dictates the precision demanded.

6.3 TECHNIQUES OF PHASE MEASUREMENT

Analogue methods

An obvious starting point is the classical phase detector, which usually takes one of two forms - differential quadrature vector addition or analogue multiplication. Both have the advantage (unlike methods depending on zero-crossing detection) of using all the information contained in the input waveform, and are capable of great sensitivity. The output is now, however, a linear function of phase angle, the relationship usually being co-sinusoidal. The output is, moreover, critically dependent on input signal amplitudes. Recently, analogue techniques of considerable sophistication and

precision have been developed (103-105) but they are not really ideal in an instrument requiring a digital readout.

Compensation, or null methods

In order to achieve linearity and insensitivity to amplitude, an analogue detector may be used in a null system (see diagram). Some means is then required for phase shifting one of the input signals by a precisely known amount. Frequently the phase shifting element is a two-phase induction generator operating under locked rotor conditions, and known as a "resolver". This approach has frequently been used in the past in ranging instrumentation, but precision is usually limited to about 1 or 2 parts in 10^3. Another approach is to use a shift register, counter or analogue delay line to effect the compensating delay.

Duty cycle methods

These methods are now widely used where linear conversion of phase is required. The signals are squared, or hard-limited, to define the instant of zero-crossing, and the limited signals differentiated and caused to set and reset respectively a two-state device such as a RS flip-flop. The duty cycle τ/T of the waveform appearing at the output of the flip-flop is then linearly related to the phase angle ϕ, by

$$\phi = 2\pi \frac{\tau}{T} .$$

The output waveform may be voltage limited and low-pass filtered to yield an analogue voltage output which may subsequently be digitised by normal A to D conversion, or a direct conversion into the digital domain may be achieved by gating and ratio-computing operations.

Since a linear, digital output was required in the present instance, the duty cycle method is appropriate, and the direct digital conversion method was chosen.

Phase meters based on the above principles have been described in the literature by a number of workers (106-111), and a comprehensive review of digitisation methods has been presented by Kuznetskii and Chmykh (112). It is of interest to note that Soviet approaches to this problem differ significantly from Western practice, and are little known in the West.

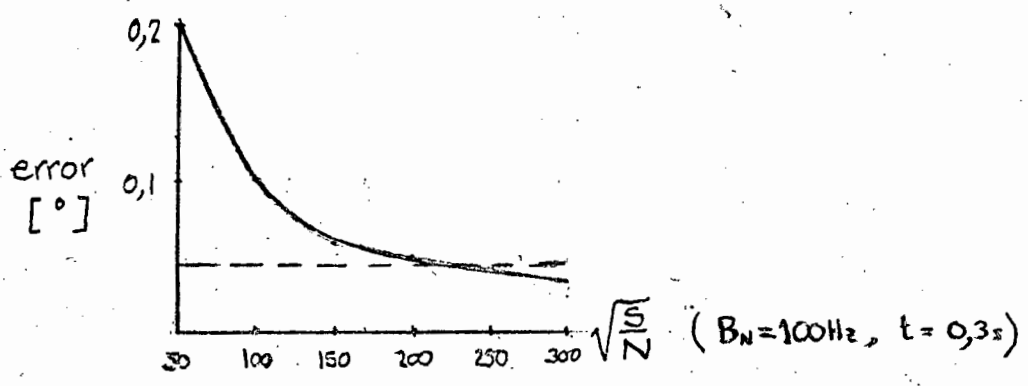
For the range of frequencies and dynamic range specified in the present requirement, the accuracy claimed is usually about $0,1^{o}$ (or 3×10^{-4}) but recently rather complex and expensive instruments have become commercially available having an accuracy of $0,05^{o}$. The most accurate phasemeter reported in the literature has a stated precision of $0,01^{o}$ (110). Thus it can be seen that the specified requirement is not unreasonable and the development effort, drawing on all the available published material, was aimed at producing a simple, economical and reliable system.

6.4 SOURCES OF ERROR - AN OVERVIEW

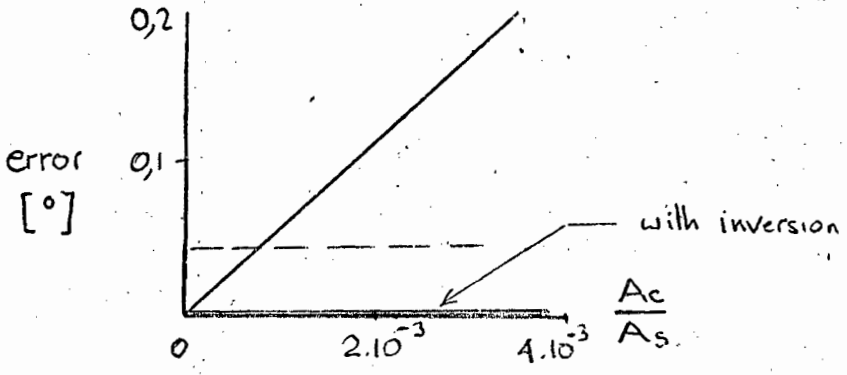
Errors may be grouped into analogue and numerical categories. Analogue errors are associated with the squaring or hard-limiting operation, and numerical errors with truncation and roundoff errors in the digital and numerical processing. Experience has shown that, even where the specified requirement is not very stringent, reliable and economical production of instruments based on phase measurement depends on a comprehensive error analysis and choice of methods insensitive to the known error sources.

Analogue errors are those affecting the accuracy with which the relative times of zero-crossing of the input waveforms are transformed into the duty cycle of a pulse train. They are caused by finite signal-to-noise ratio in the input waveforms, the presence of harmonics and drift in the threshold level of the comparator or signum operation which defines the instant of zero-crossing. These errors contribute a static offset which can normally be calibrated out of the system but, more significantly, they give rise to an apparent phase shift which is level-dependent. This can be reduced to acceptable levels by careful circuit design, by pre-processing with a low phase-shift automatic gain control (AGC) amplifier and by adopting certain error cancelling techniques making use of information contained in both leading and trailing edges of the input waveform. These will be considered in greater detail below. A theoretical analysis of these sources of error is contained in Appendix 6.1.

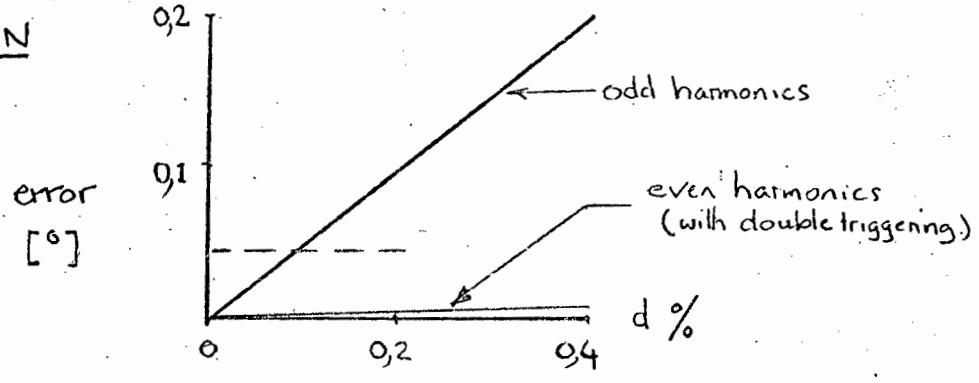
NOISE



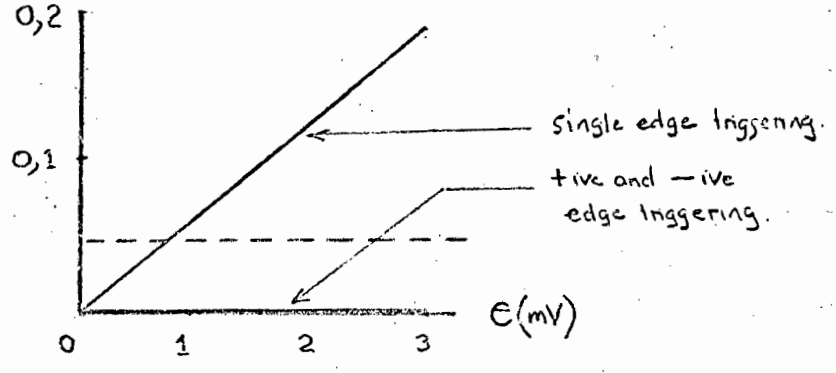
CROSSTALK



DISTORTION



DRIFT:
(comparator offset.)



Other sources of analogue error are electrical crosstalk, which causes a sinusoidal perturbation of the measured phase, and differential phase-shift in the two signal-processing channels, which gives rise to sensitivity of the measured phase to frequency shift. Straightforward techniques, based on signal inversion, exist whereby these effects can be cancelled almost completely. The theory has already been dealt with in sections 4.7(c) and 5.4.

The dependence of error on signal-to-noise ratio, crosstalk, harmonic distortion and comparator drift is summarised in the accompanying figures. Assuming that these effects are all uncorrelated, and dividing the error equally among them, we can see that an error of about 0,045° or $1{,}25 \times 10^{-4}$ can be tolerated in each case. The ordinate corresponding to this degree of precision is indicated, and it can be seen that the following levels are tolerable:

$\sqrt{S/N}$	~220
Crosstalk	8×10^{-4}
Distortion	0,1 %
Drift	< 1mV

Digital errors

The remedy for errors produced by numerical truncation is obviously the provision of more bits. There is one source of digital error however which does not fall in this category. It relates to the conversion of the mark-space ratio bearing the phase information into a digital number. There are many ways of doing this but they all amount to determining the ratio of the number of clock pulses occuring during a "mark" to the total number of clock pulses in a complete cycle of the incoming signal. Suppose M clock pulses are produced in a complete cycle at the comparison frequency, and N pulses are gated during a 'mark' period. In order for the ratio N/M to reflect the phase-shift with a precision 1:p we clearly need at least p or p/2 clock pulses per comparison frequency period (depending on whether or not the clock frequency is constrained by a feedback loop to produce exactly M pulses per period. If this is the case, the truncation will occur only for N.) Since partial gated clock pulses are counted as clock pulses if they exceed the response time of the counting circuitry (say 10ns for TTL), the error has a rectangular probability density function with an average value or bias of one-half clock pulse.

Techniques to improve accuracy

There is a fairly extensive literature of phase measurement, and many ingenious systems have been proposed. These include techniques to improve resolution using the phase of high order harmonics of a pulse train (113) - a technique which appears in many forms in the Russian literature, and hardly at all in the West - and those achieving an effective multiplication of phase-shift by successive stages of frequency multiplication and frequency translation by heterodyning (114). In many ways the use of such techniques is equivalent to increasing modulation frequencies in distance measuring instruments, increasing as it does the resolution while reducing unambiguous range. Their applicability is however limited by the non-linear elements which are an inevitable concomitant of frequency multiplication. The presence of these nonlinear elements greatly increases the likelihood of phase error due to the level variations which are also invariably present. It is conceivable however that with a very "stiff" automatic gain control (AGC) system in the receiver they could be used with profit, especially since the use of a 'negative pattern' greatly reduces susceptibility to level-dependent phase shift. In the case of angle measurement the signal levels are constant and frequency multiplication can be used without any serious problems, but here we are not really resolution-limited and there is considerably less need for the technique.

It should be borne in mind that we have already exploited one such technique - that of heterodyning. By translating signals from f_1 to f_2 while preserving phase-shift we effectively realise a time expansion equal to f_1/f_2.

6.5 METHODS OF TESTING PHASEMETERS

Several accounts of phasemeter test methods have been published, especially in the Russian literature (115-117). For the present purposes, however, two simple techniques were evolved.

Direct comparison

If a reference phasemeter is available, the system under test can be calibrated directly. A signal undergoes an adjustable phase-shift in a second-order all-pass network and the pair

of phase-shifted signals at the output and input to the network is simultaneously measured by the system under test and the reference system.

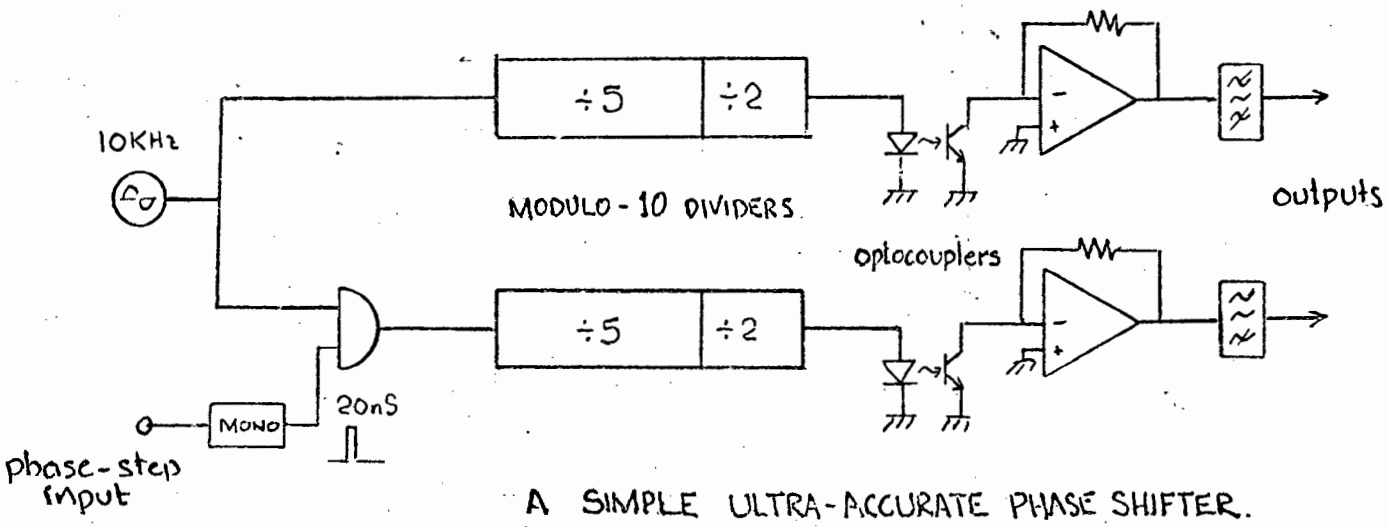
An important point to note is that the all-pass circuit does not have a linear-phase characteristic and does not preserve the signal waveform as would a transport lag, and the two systems being compared will in general respond differently to the variously-phase-shifted harmonic components in the signal, unless it is a pure sinusoid. Even then great care must be taken that harmonic components are not introduced by nonlinear distortion in the all-pass network.

Phase generator method

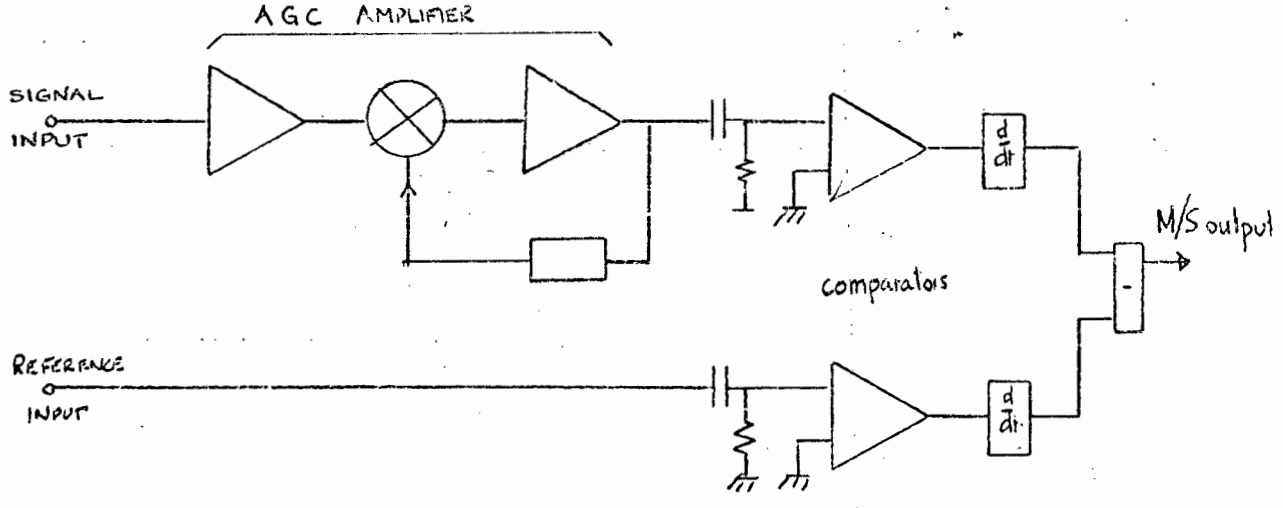
A technique found more generally useful in development work was a digital instrument capable of generating a pair of identical waveforms with arbitrary relative time-shift. The circuit embodiment of such a system can take many forms and the logic and circuit details are of no special interest. It can also be emulated in software using a microprocessor driving a pair of A/D converters and low-pass filters, an approach which offers considerable flexibility.

One of the merits of this approach is that waveform is totally independent of epoch, and delay generated by the system exactly simulates pure transport-lag. The nature of the simulated signal is therefore identical with that of the actual signals produced in the angle transducer. In order to simulate the signals obtained in range-measurement, quasi-sinusoidal signals can be generated and further improved by post-bandpass filtering. The amplitude of the signals can also easily be varied.

Simple logic circuitry using TTL can be designed to exhibit clock-skew and other timing errors on the order of 10^{-8}s. If the signals are being simulated at a frequency of 1kHz, this represents a phase error of 1 part in 10^5 which is well within the precision required in this thesis. However it seems worth recounting in slightly greater detail a very simple system which offers practically unlimited accuracy - for example signals having a mutual phase relationship accurate to within 1 part per million at frequencies of several megahertz can readily be generated. The writer has not come across any comparably accurate technique in the literature, despite a fairly extensive search.



A SIMPLE ULTRA-ACCURATE PHASE SHIFTER.



PHASE-MEASURING SUBSYSTEM
showing analogue signal processing

If two counters modulo M are driven from a clock at rate f and the last stage of each counter is a symmetrical binary divider, we shall have at the outputs to the dividers two square waves which can take up in relation to one another M distinct phase states. For clarity suppose M = 10. If we add a pulse to (or delete one from) one of the counters we shall cause their mutual phase relationship to change by 36^{o} exactly. The word exactly may seem imprecise but if we assume no coupling between the counters it is difficult to see how the phase steps can be anything but exact. In so far as we are free from any spurious coupling between the counters, so are we free from systematic error, and precision is limited only by random and periodic phase jitter. These are caused respectively by finite signal-to-noise ratio and timing errors, and both of them average to zero over a sufficiently long time.

Two such phase-shift generators were constructed and compared. Opto-isolator coupling was used to control spurious cross-coupling. The phase shift of the two units agreed to a standard deviation of better than 2 parts in 10^{-6}, at which point the measurement precision was limited by random scatter. The technique is a powerful one somewhat analogous to the angle-testing procedures of Chapter 5. In the opinion of the writer it deserves to be better known.

6.6 DESIGN OF A PHASE-MEASURING SUBSYSTEM

The basic structure of the phase-measuring subsystem chosen is shown as a block diagram. The incoming signals are compressed in dynamic range by means of an automatic gain control (AGC) circuit. The signals are then applied to AC-coupled zero-crossing detectors. The outputs of these are in turn differentiated and used to set and reset a two-state element (i.e. an R-S flip-flop). The output from the flip-flop is a rectangular wave containing, in the form of a duty cycle (τ/T), information about the relative phase displacement ϕ of the original signals. In order for the relationship between (τ/T) and ϕ to be linear within the required 1 part in 3000, it is necessary to give close attention to the design of the analogue signal-processing section, and this is where the bulk of activity was directed.

There are many approaches to producing a numerical output from (τ/T), and the subject is discussed in Appendix 6.2. This section of the work was overtaken by events. Early experiments were carried out with a simple stochastic system implemented in MSI TTL. It worked extremely well but was not ideally suited to the very rapid determination ideally required, especially for the internal signal where there is no noise problem and hence no need for temporal averaging. Later a microprocessor-based data-acquisition system became available and this was used to compute the ratio (τ/T) directly. All that was required in addition was a pair of highspeed counters resident on the microprocessor bus to accumulate the gated and ungated clock pulse streams over the measurement period. The microprocessor would then read the two counters (and reset them) and compute the ratio. In addition the microprocessor was available to control the measuring sequence and to scale and average data. The microprocessor equipment required to do this was already available from another project (118) and does not form part of this thesis. Such a system would presumably form part of any new survey (or other) instrument including the present one. Its design however is a routine engineering task which does not fall within the scope of this thesis, which is limited to a comprehensive design study and presentation of theoretical and experimentally validated solutions to specific problems which arise.

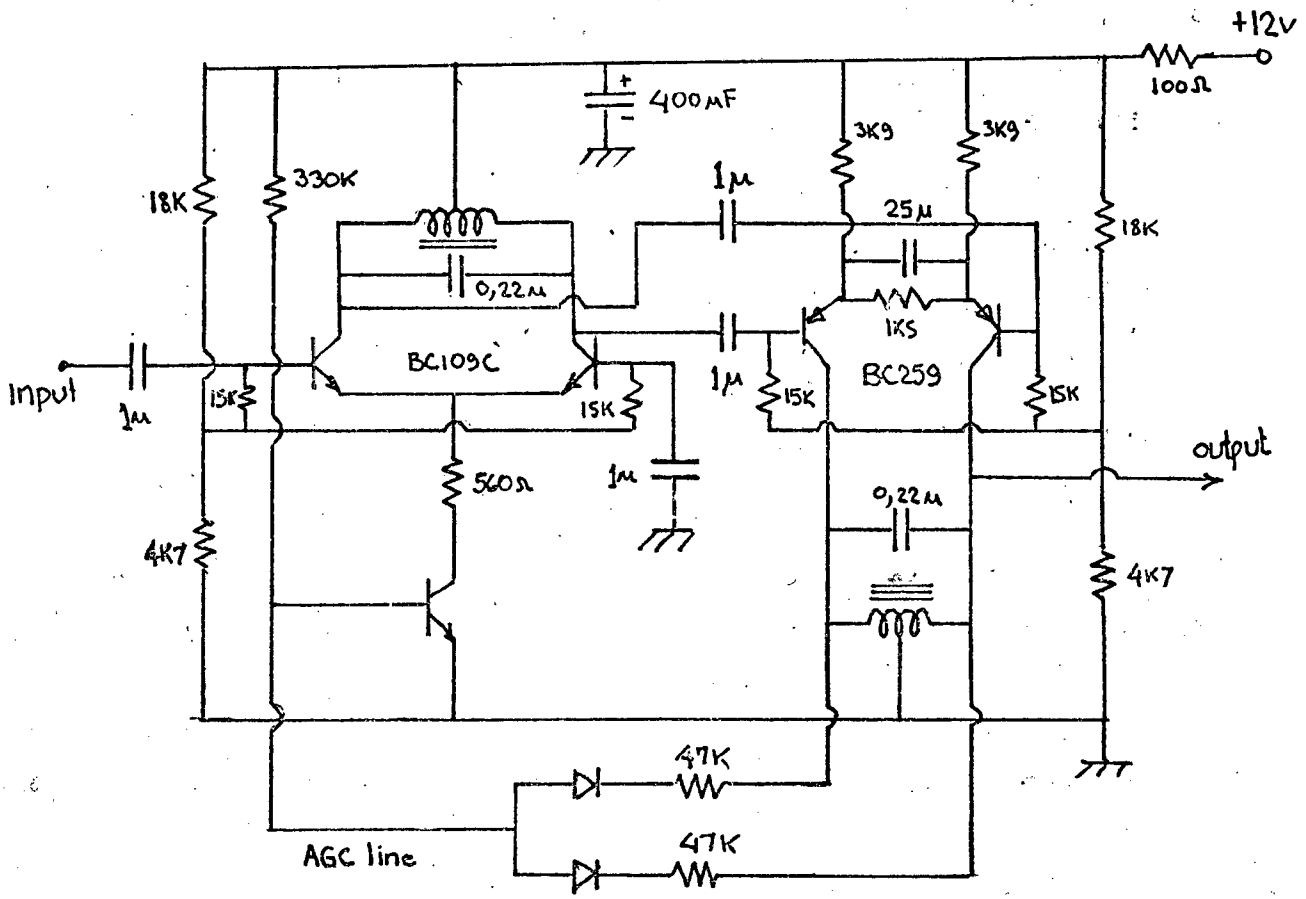
The Automatic Gain Control System

In order to facilitate the zero crossing detection operation on the heterodyned received signal it is desirable to compress its 0,5mV- $1V_{rms}$ dynamic range into a more readily handled compass - say 0,5V-1V. This represents 60dB of automatic gain control, and it is necessary to accomplish it without

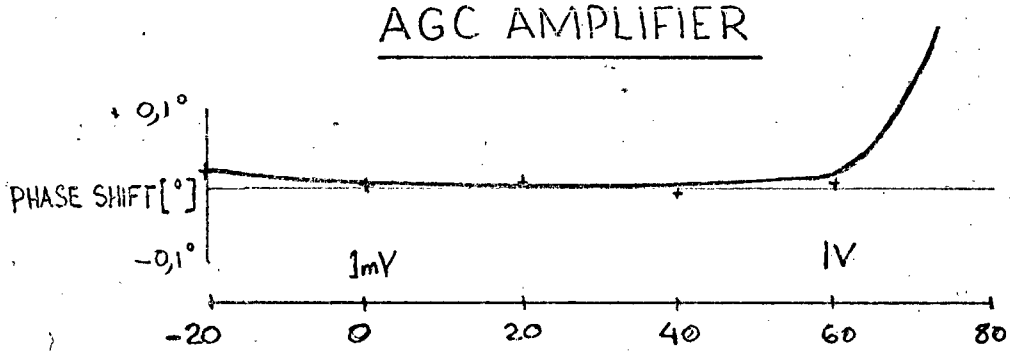
i) introducing level dependent delay or phase-shift greater than 1 part in 3000 (i.e. 330ns in the 1kHz signal).

ii) introducing an unacceptable level of distortion.

Several approaches were tried: a fast gating technique analogous to the pulse-width multiplier was reasonably successful but was complex and introduced noise due to switching signal breakthrough. Controllable potential



AGC AMPLIFIER



PHASE SHIFT VS INPUT LEVEL
[dBm]

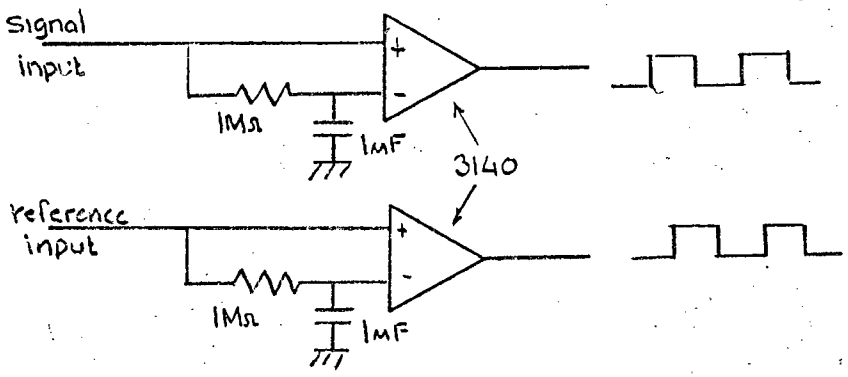
dividers using light-dependent resistors and field effect transistors caused appreciable distortion even when feedback compensation was attempted. Two methods which appeared promising were an attenuator based on the photon-coupled FET (119) and one based on a balanced pair of FET's in a new analogue-multiplier configuration (120). These were not however pursued once it was discovered that an adaptation of a simple two-stage amplifier using emitter-coupled pair stages with tuned collector loads originally developed for another purpose (121), performed more than adequately - better, in fact, than any of the other circuits investigated.

AGC range was in excess of 70dB with less than 1 part in 5000 (i.e. $0{,}07^{o}$) differential phase shift. Excess phase shift for very strong input signals was found to be due to capacitive base/collector coupling competing with the drastically reduced transconductance of the transistors for these input levels. If an even wider dynamic range were required it could probably be achieved by replacing the individual transistors by cascode pairs. It is interesting to note that this circuit also performed well (and with little phase shift) as a signal compressor acting directly at 150kHz. At this frequency it far outperformed the competing versions.

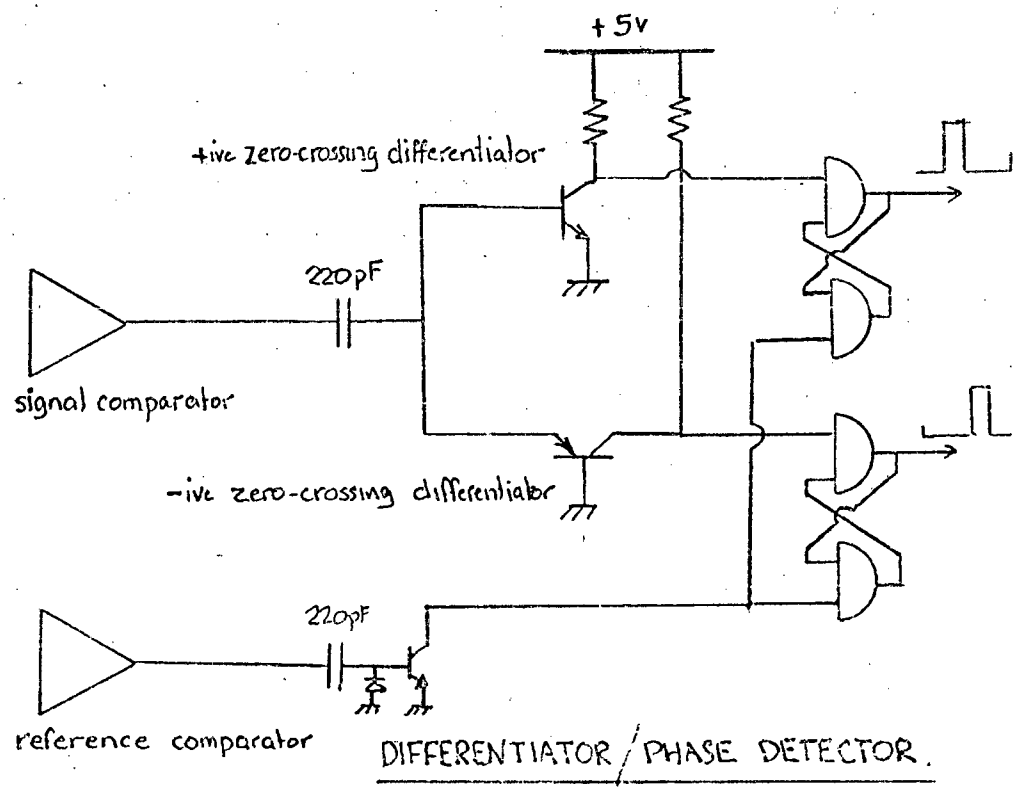
Zero-crossing detection

A straightforward approach to determining the instants of zero crossing is to apply the gain-compressed signal to a zero-referenced high gain comparator. There are two principal sources of error:

i) Saturation of successive stages in the comparator causes a delay between zero crossing at the input and output level transition, which is a function of input level. Normally the delay "saturates" or tends to a constant value for a specified value of input level or "overdrive factor" and the provision of AGC makes it easy to realise this. Although it is not required in the present context it is worth noting that delay variation due to the progressive saturation of successive stages can be effectively controlled by using low Q bandpass rather than direct coupling between stages (122).



ZERO-CROSSING DETECTORS



DIFFERENTIATOR / PHASE DETECTOR.

ii) Input voltage drift at the comparator input, or input voltage offset resulting from input offset current and dc source resistance. This causes the zero-crossing time to be replaced by the time at which the signal crosses some arbitrary non-zero level. If however we detect both the positive- and negative- going zero crossings their mean will be free from error on this account (110). An additional advantage of this scheme is that error due to even harmonic distortion is also removed.

Other ingenious systems for minimising zero-crossing errors have been proposed (108;109), but in fact the earlier gain compression circuit used here is so effective that the input to the comparator hardly varies, and a simple approach is adequate. A BIFET operational amplifier serves as the comparator, and its reference input is derived from the average dc level of the input signal via a low-pass filter. It is easily shown that a signal formed from the difference of a direct and a low-passed input signal is equivalent to a high phase version of that signal. The low-pass filter has a long time constant (~1 sec) and the signal is effectively limited, output transitions occuring within a few millivolts (due to amplifier voltage offset) of the input zero crossings.

To eliminate the effect of the offset, the comparator output, which slews at about 10^7V/s, is leading edge- and trailing edge- differentiated in two separate differentiators, one a common-emitter and the other a common-base stage. A zero transition produces a 30nS pulse which sets or resets a RS flip-flop formed by cross coupling a pair of TTL Schottky NAND gates.

Two such flip-flops are used, one set from the pulses occurring on positive, and the other from pulses occurring on negative, zero-crossings of the input waveform. The output from both flip-flops is processed as outlined above to produce a digital numerical version of their duty cycles, and the mean of the separate outputs is free of error due to comparator drift.

Experimental results

Outputs from the digital phase generator (which had previously been shown to be free of systematic error down to 1 part in 10^6) were used to drive the inputs of the system described. The outputs of the flip-flops gate 10MHz pulses from a crystal oscillator into one of two parallel 24-bit counter chains. The contents of the counter can be loaded into the microprocessor data acquisition system mentioned earlier. A simple program was written to control the count sequence and compute the quotient of the contents of the counters. The microprocessor scaled the result by multiplying by 360 and presented the result directly:

Phase increment:	0	36	72	108	144	180	216	252	288	324	360 (degrees)
Error:	+0,01	-0,02	-0,02	-0,01	+0,01	+0,02	0	+0,01	+0,01	-0,01	-0,02 (degrees)

Peak error is $0{,}02^{\circ}$ or 1 part in 18000.

A check was also carried out, using a precision attenuator, on the sensitivity of phase to level:

Signal level	1mV	10mV	100mV	1V	3V
error	+0,05	$-0{,}02^{\circ}$	$-0{,}01^{\circ}$	0	0

The peak error at 1mV level is 1 part in 7200.

Clearly the phase measuring system performs more than adequately, and the initial assumption of a phase resolution of 1 part in 3000 can be seen to have been conservative, especially when we take into account the fact that, due to the use of negative patterns, cancellation will reduce even the small amount of level dependence present.

CHAPTER 7

THE VERTICAL REFERENCE SENSOR

7.1 INTRODUCTION

We have seen the feasibility of developing a simple, low-cost transducer capable of a standard deviation of error substantially less than 10 arc seconds. This transducer was intended to measure the angular displacement of the alidade of the instrument relative to an arbitrary datum in the horizontal plane. There is no reason why a similar (or identical) transducer should not be used to measure the vertical deflection of the telescope, and hence determine the elevation of the vessel. This dual use of a major sub-system element can be expected to save both development and production costs.

In the case of a vertical angle measurement, however, a new element is required. Vertical angles in survey practice are determined not in relation to an arbitrary datum, but with respect to the local vertical, or direction of the gradient of the earth's gravitational field. A sensor is therefore required to determine this direction in space.

Accuracy required

The precision of vertical angle fixing demanded (see 2.4) is 20 arc seconds - hardly different from that in respect of a horizontal angle. Several sources of error and uncertainty exist external to the transducer itself, however, and in order to achieve the required accuracy it is necessary to tighten somewhat the vertical transducer accuracy specification and set a suitable limit on the errors of the reference device to be discussed in this chapter.

In the first place, the path followed by a ray of light from vessel to instrument will not follow a straight line. The effect of the earth's curvature and refraction by the resulting curved strata of the atmosphere is to cause the ray on average to follow a path having a radius 7 times that of the earth (133). Its effect therefore is merely that of a small reduction (14%) in the apparent vertical displacement of the vessel due to the earth's curvature itself. This amounts to $D^2/4R$ when R is the earth's radius, 6370km, or rather less than 1/3 of a metre for D = 2km. These effects are totally negligible. We have seen however in 3.3 that

variations in the temperature lapse rate close to the surface of the sea can cause a more significant - and unpredictable - diurnal beam-wander of up to about 6 arc seconds.

Secondly, we have stated that there is a probable standard deviation of telescope pointing of some 10 arc seconds. This assumes calm conditions, in which the short-term vertical motion of the vessel does not exceed, say, 0,1m. The combined uncertainty of pointing on account of visual and optical effects may then be estimated as $(10^2 + 6^2)^{\frac{1}{2}}$ or about 12 arc seconds. This leaves, for the combined error due to the transducer and the vertical reference, an error of $(20^2 - 12^2)^{\frac{1}{2}}$ or 16 arc seconds. Since we can readily achieve a standard deviation of 10 arc seconds for the transducer, the precision of the vertical reference should be (in terms of standard deviation) no worse than $(10^2 - 6^2)^{\frac{1}{2}}$ or 8 arc seconds. Actually this is not a sufficiently stringent requirement, since the error of the vertical reference may be a systematic zero drift rather than a randomly distributed variable. We shall rather arbitrarily but conservatively specify a peak absolute error for the vertical reference no greater than 5 arc seconds.

As we have seen in 2.4, the range of the vertical reference device should be such as to be able to cope with dislevelment of the instrument over at least 50 arc seconds. The required linearity over the working range is thus 10%.

Classical survey practice

Classical practice is represented in the well-known theodolite. The instrument is set up by the operator so that the alidade rotates about a truly vertical axis, using a conventional spirit-level type bubble as reference. The disadvantages of such a manual system are obvious and there is a clear trend, even in modern 10 arc second engineering theodolites, to incorporate some automatic form of compensation for instrument dislevelment. The case for doing so in the present instrument is thought very strong, as was argued in 2.7.

An alternative approach

An alternative exists to the use of a limited range compensator to refer transducer-determined telescope tilt relative to the alidade to an absolute vertical reference. This is to employ a transducer covering the entire tilt range of the

instrument (i.e. ± 10° or ± 30°)and operating directly with reference to gravity. At least two such transducers are known to exist:

i) For use in the aerospace industry several manufacturers (Kearfott, Hamlyn, Singer) have produced horizontal accelerometers, in which a pendulum in a force-balance servo is automatically returned to a datum position. The output of the transducer is proportional to the electrodynamic restoring force, and hence to the horizontal component of acceleration. Clearly in a static situation tilt is equivalent to horizontal acceleration.

Clearly, the specification demands of the transducer a zero stability better than 5 arc seconds, a linearity over the 10° range of 0,014%. This is a formidable challenge but it is achieved by several of the commercial units - at considerable cost (e.g. several thousand rands). There are other disadvantages to this approach: the output is a sinusoidal rather than linear function of tilt, and is in the form of an analogue voltage. Extension to a 30° tilt range would hardly be possible.

ii) An ingenious transducer developed by the Hewlett-Packard Company for use in a survey instrument (16) uses a toroidal cavity fabricated from a machinable ceramic. The cavity is half-filled with a conducting liquid, the free surface of which defines a horizontal plane. As the toroidal cavity rotates due to tilt of the instrument, electrodes deposited on the inside of the cavity are covered and uncovered and the resulting resistor ratio variation is converted into a voltage in an operational amplifier circuit. A peak error of 10 arc seconds is claimed for this device over a tilt range of -20° to +50° with enhanced resolution for more limited ranges.

In many ways this type of transducer is an ideal one for the instrument under consideration. Its principal drawback is that it is not commercially available as a component and the extreme precision and cleanliness required in its manufacture (16) were far beyond the resources of the writer. It is rumoured that even the Hewlett-Packard Company had serious difficulties in achieving reliable operation of this device. Its critical nature is not surprising when one

considers that sub-micron accuracy and stability are required, and the device depends on reproducible behaviour at a critical liquid/solid interface. Angle of contact between a liquid and a solid is notoriously sensitive to microscopic chemical contamination and temperature variation. A relatively minor disadvantage is the nature of the output, which is incompatible with phase methods and requires precise analogue signal-processing circuitry.

It was concluded that the direct-tilt transducer approach was not advantageous, nor indeed viable in the case of the present instrument, and a limited range level sensor should be used in conjunction with a version of the angle transducer developed in chapter 5.

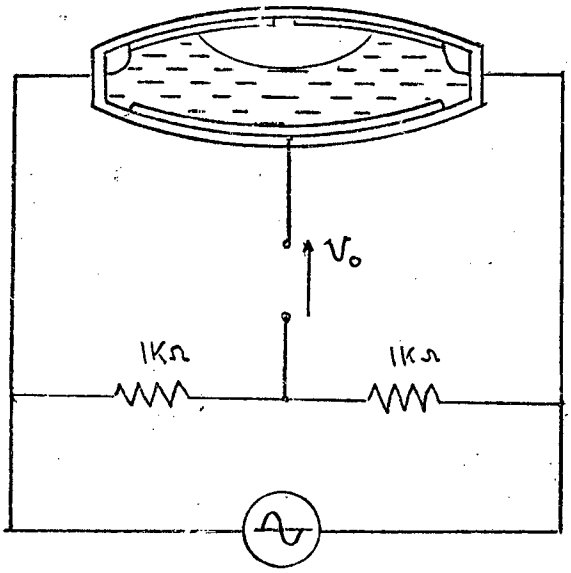
7.2 SURVEY OF LEVEL SENSORS

In order to become familiar with the possibilities and limitations of the various approaches, a comprehensive literature survey was undertaken, as well as some simple experiments.

Level sensors of 5 arc second or greater accuracy are used in a number of applications:

1) Engineering metrology. As a substitute for a simple spirit level, various devices such as the Nivéltronic Level and Tally Level electronic levels are used having single-second resolution and an electrical output proportional to tilt over a range of at least several minutes of arc. These devices normally use a pendulum transduced by a linear voltage differential transformer. They meet the accuracy required in the present instrument but are ruled out on grounds of cost, weight and bulk.

2) Engineering construction monitoring. Devices such as the Electrolevel consist of a bubble-type element in which the usual ethanol fluid is replaced by an electrolytically conductive one. Electrodes are deposited on the inside of the phial and movement of the bubble unbalances a bridge giving an electrical output proportional to tilt over a limited range (123). The device is intended for building into large civil engineering projects such as dams and bridges, to monitor subsidence. Again the performance is adequate



AN ELECTROLYTIC BUBBLE SENSOR

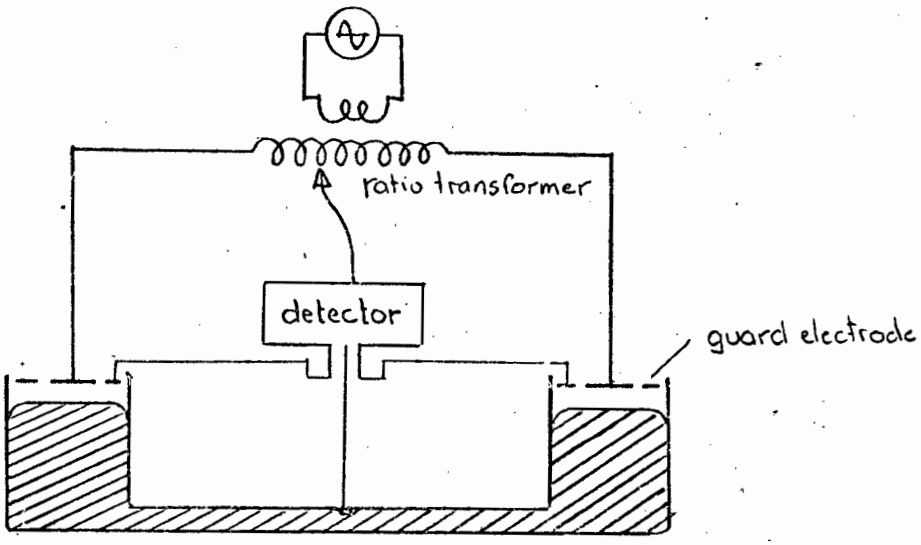
for the required application but the device is too large and too expensive for inclusion in the instrument.

3) The Aerospace industry. In addition to the pendulous force-balance transducers described above, a requirement also exists in the aerospace industry for limited-range tilt sensors for use in the static levelling of platforms for inertial navigation. They are miniaturised versions of the electrolytic bubble devices discussed above and are available at moderate cost from the Hamlyn Company. They are not very comprehensively characterised in the manufacturer's literature, so several units were purchased for evaluation. They were driven at 400Hz as recommended by the manufacturer to avoid electrolytic migration and polarisation effects and mounted on a stable platform in an environmental chamber. Unfortunately their zero stability left much to be desired, some units showing a drift of zero position of up to 3 arc seconds per degree celsius. It was reluctantly concluded that this otherwise attractive solution did not offer adequate accuracy.

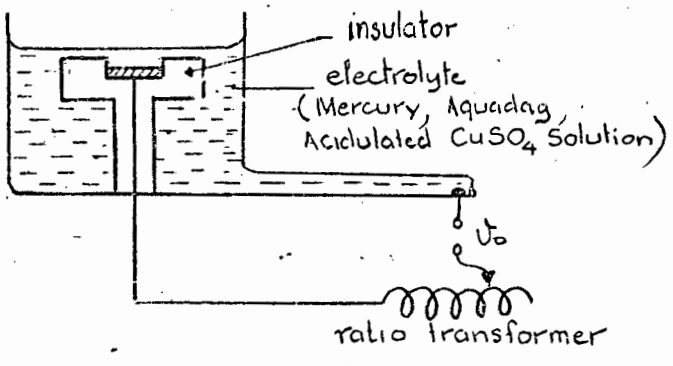
4) Special purpose applications. Several highly accurate level sensors intended for various scientific applications have been described in the literature. Although not directly applicable to the present requirement, they were considered worth studying in the quest for a reliable, low-cost device.

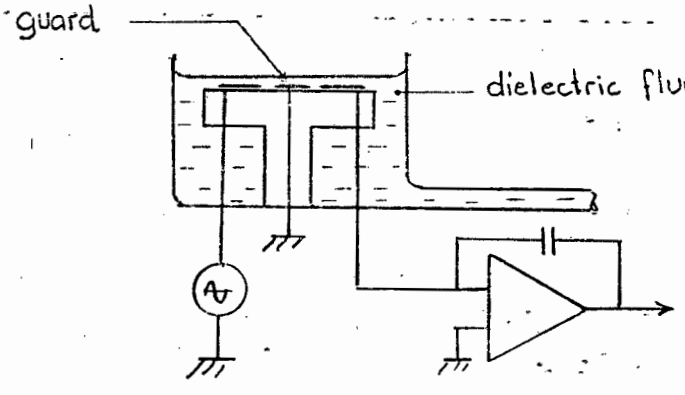
Electromagnetic levitation

One of the most sensitive tiltmeters or level sensors is described by Simon et al (124). It consists of a sliver of pyrolytic graphite diamagnetically levitated in the field of a permanent magnet. Motion of the graphite as the assembly tilts is detected by a pair of photocells. The main drawback is that, owing to the weakness of the diamagnetic effect, a massive and bulky magnet is required. Ferromagnetic levitation would of course be much superior in this respect but active stabilisation is required and the suspended particle is very sensitive to stray magnetic fields. Electrostatic levitation - perhaps in a dielectric liquid - seems a distinct possibility but no reference to such a device could be found, although the electrostatic levitation of gyroscope rotors



A MERCURY TILTMETER (AFTER STACEY)





EXPERIMENTAL SUBMERGED-ELECTRODE SYSTEMS.

has been reported (125).

Mercury Pool Tiltmeters

Stacey et al (126) have described a sensitive tiltmeter featuring 140dB dynamic range . Two mercury cisterns (of sufficiently large diameter to make meniscus effects negligible) are interconnected by a tube, the diameter of which can be selected to provide critical damping for the system. The mercury forms the moving plate of a differential capacitor.

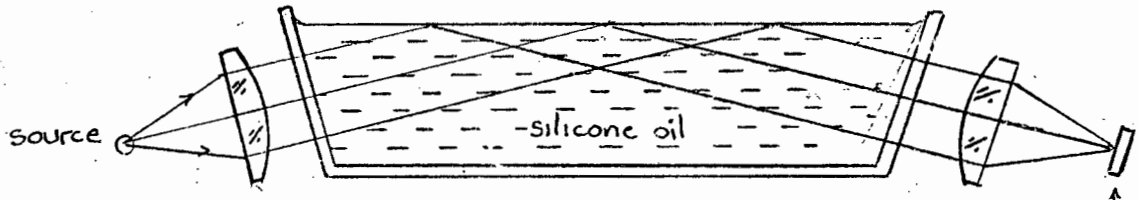
This is an attractively simple system but experimental tests showed that it is a difficult one to adapt to portable use. Zero stability after gross over-tilting is impaired by the adhesion of minute mercury droplets to the electrode system, although it is possible that this could be overcome by a suitable choice of electrode material, scrupulous cleanliness and the use of an inert or reducing atmosphere. Although bulk oscillation of the mercury is easily damped, vibrations can set up standing waves on the mercury surface, with consequent error.

Experimental Submerged-electrode Versions

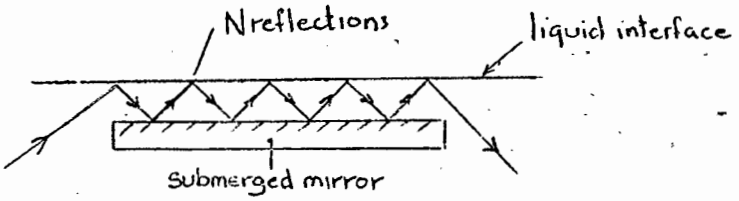
Attempts were made to devise analogous systems using conducting or high dielectric fluids with totally submerged electrode systems. This completely solves the overload recovery problem, but other problems proved intractable - notably the effect of vibration, the attainment of adequate sensitivity and the variation of sensitivity with temperature, due to expansion of the liquid/vessel system. The choice of a suitable conducting fluid is also problematic. At the level of sensitivity required one is plagued by either polarisation (even at high drive frequencies) in the case of an electrolyte or by local inhomogeneity and migration in the case of a colloid.

Free Liquid/Air Interface Systems

A newtonian liquid is incapable of sustaining shear force and its surface therefore defines a gravitational, equipotential surface. This has been exploited in the vertical angle indexing system of the Kern single second theodolite. The vertical index mark is reflected by a free silicone oil surface, using total internal reflection. Care has to be taken to allow for the refraction which also occurs, and its variation with

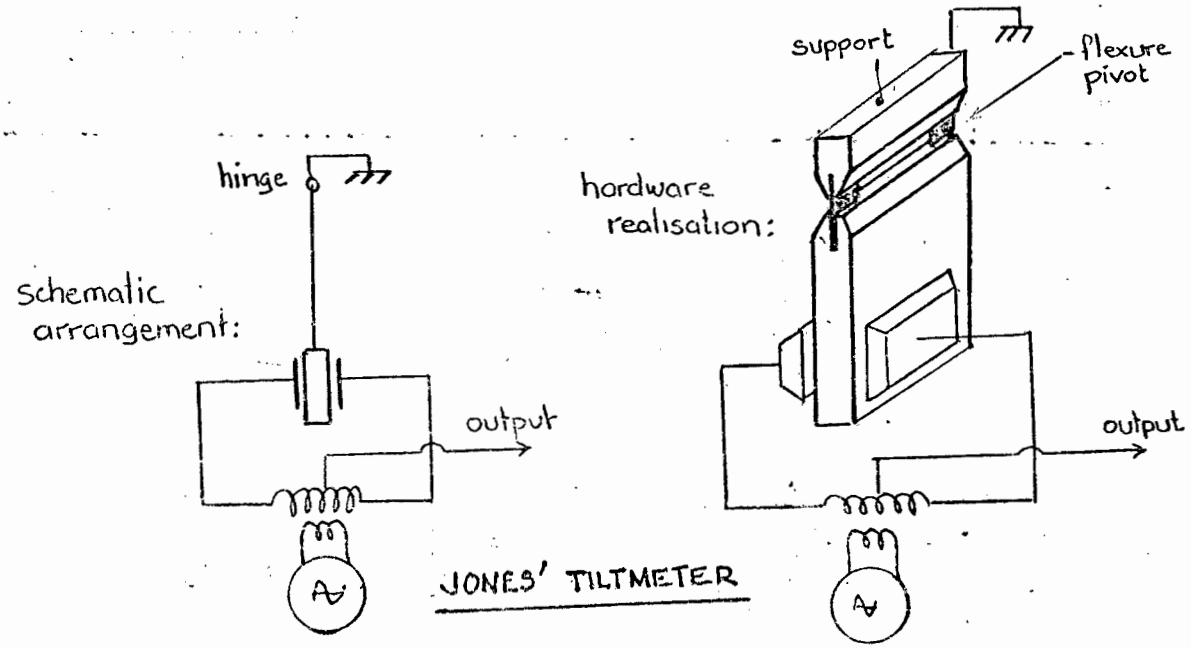


EXPERIMENTAL FREE-INTERFACE ELECTRO-OPTIC SENSOR



MULTIPLE-REFLECTION VERSION

(magnification 2^N)



JONES' TILTMETER

temperature-induced change in refractive index. A system free from this complication is used in the Hewlett Packard "Total Station" (98). Here an autocollimator observes a free mercury surface. The mercury is covered with a silicone oil layer to provide damping and prevent oxidation of the surface.

Although these systems clearly can be made to work, they are not without serious difficulties. Once again, although bulk motion of the fluid is readily damped, higher order ripples induced by vibration are less easy to suppress, and cause distortion of optical images. A second difficulty is that the displacement of the image in a miniature system is extremely small (say 0,1μm per arc second) and it is difficult to maintain positional stability of optical (and electro-optical) components to this order over long periods of time. Light-emitting diodes are particularly bad in this respect. An attempt was made to magnify the deflection by multiple reflection but this system was hopelessly susceptible to vibration.

Free-liquid systems using resistive or capacitive sensing are also possible and were tried, but they are subject to the limitations discussed above.

Although many of these methods appeared promising, none of of them performed entirely satisfactorily despite considerable experimentation.

Precise pendular systems

A very sensitive and stable device capable of detecting earth tides has been described by Jones (127). It is remarkably simple, consisting of a flexure-supported pendulum, the bob of which forms the moving element of a differential capacitor. With this device coupled to simple synchronous detector circuitry, Jones reports that displacements of the bob of as little as 10^{-14}m can be reliably measured, with drift of the order of 10^{-9} radians per day. (128).

7.3 A CRITIQUE OF THE TECHNIQUES SURVEYED

Broadly, the methods fall into two classes - those which do and do not employ some form of magnification or mechanical advantage to increase the effect of the very small displacements

involved in tilts measured in arc seconds. The latter are best exemplified by the free-liquid interface and pendular versions. They are inherently accurate and free from spurious effects but their exceedingly small output places stringent demands on the signal-processing circuitry, particularly in respect of drift. Sensors in the former class, on the other hand, have a large, easily processed output but the very mechanism which contributes the magnification frequently renders them vulnerable to spurious effects - e.g. surface-tension effects in the case of a bubble, and vibration in the case of devices employing multiple internal reflection from a liquid surface. The sensors with magnification also tend to be less tolerant of errors in mechanical fabrication. This does not however necessarily mean that they will be prohibitively expensive. A ball-bearing is a component made to the closest attainable mechanical tolerances, but it is inexpensive on account of a very sophisticated manufacturing process justified by high volume production.

An analogous situation exists in the precision bubble, which is inexpensive despite extremely accurate and ultra-clean fabrication using multiply-distilled fluid (usually ethyl alcohol)of the highest purity. It was thought that a possible cost-effective solution to the level sensor problem would be to modify such a bubble so as to produce from it an electrical output.

It was decided to experiment with two approaches - one from each of the catagories discussed above. The first would be a pendular sensor, capacitively transduced, and essentially an adaptation of Jones' transducer. The second would be a standard bubble equipped with electrodes and producing an electrical output due to change of capacitance in response to bubble movements.

7.4 DESIGN OF A PENDULAR SENSOR

The vertical compensation system of most theodolites and automatic levels (Kern being the exception) uses a pendulous prism in the optical path, and, as was pointed out above, a pendulum and displacement transducer forms the basis of electronic levels used in engineering metrology. It was

decided to investigate whether a simple low cost miniature version could be designed for use in the present instrument. One of the problems encountered is the displacement of the necessarily lightweight pendulum due to the spurious force on it caused by the LVDT displacement sensor. It is necessary to keep in mind the extremely small forces involved - a simple pendulum of mass 10g tilted from the vertical through an angle 10 arc seconds experiences a restoring torque of only 5×10^{-6}Nm. This places very stringent demands on both the displacement transducer and the supporting pivot. In both respects Jones' version (127) is a model of good design.

The present requirements differ from those faced by Jones, who was developing an instrument for fixed laboratory installation. A survey instrument however calls for two important features related to portability:

i) small size
ii) excellent zero stability after recovering from severe mechanical shock and tilt overload.

It was originally thought that a suitably compliant flexure support might be deficient in respect of shock tolerance, and alternatives were considered such as a quasi-kinematic support (129) and a miniature precision ball race, following current practice in Japanese theodolites (130). In the end it was decided that the best approach was a highly compliant flexure pivot, and a displacement measuring system tolerant of lateral compliance.

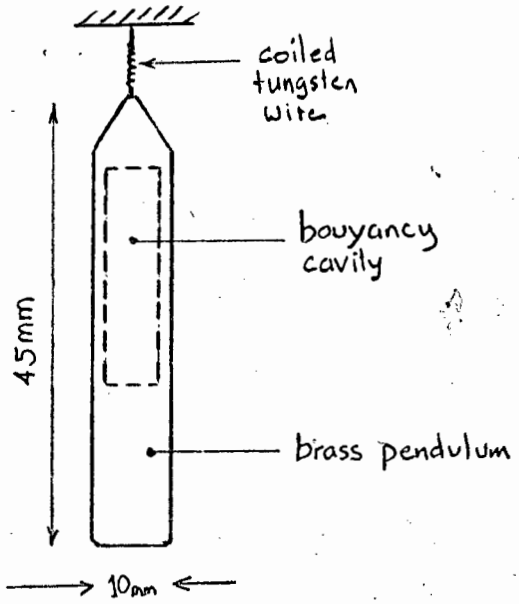
The proposed design is shown in the accompanying diagram. To meet the requirements four special features are incorporated:

i) The pendulum is fluid damped using silicone oil.

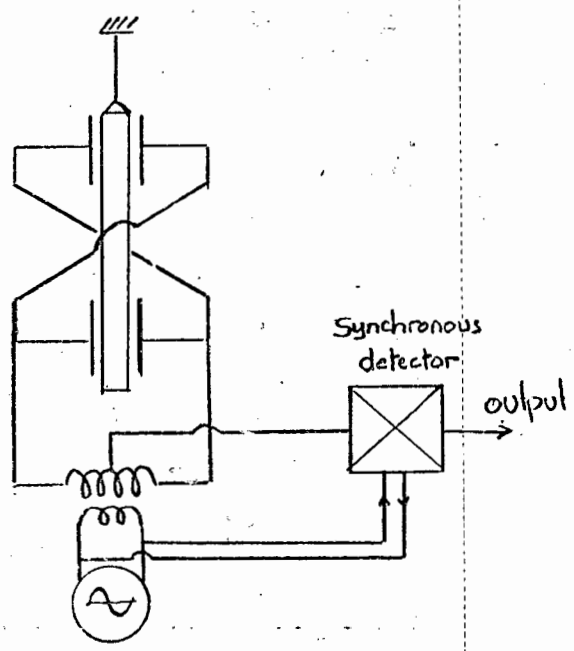
ii) It is made partially bouyant to reduce the effect of shock. The bouyancy cavity is of course located above its centroid so that the pendulum retains a strong dipole action with respect to the gravitational gradient.

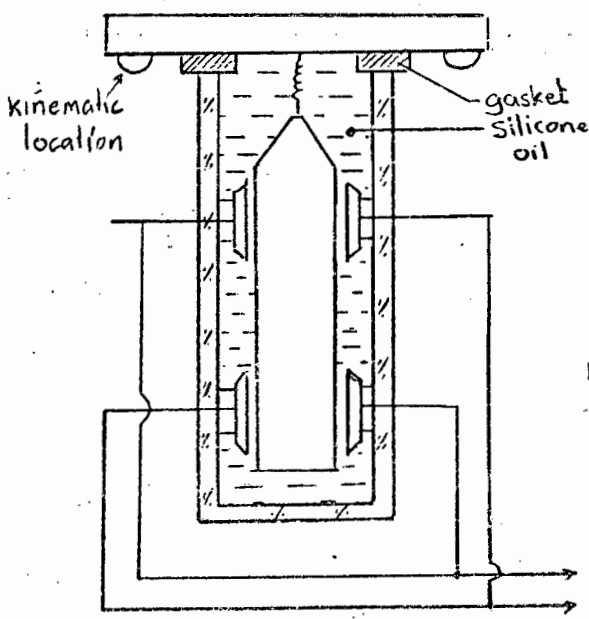
iii) A highly flexible and compliant flexure support is used in the form of a coiled tungsten wire (taken from the filament of an incandescent lamp). This has three beneficial effects: there is little likelihood of

detail of pendulum



Schematic arrangement of proposed system :





HARDWARE REALISATION OF PROPOSED TILT SENSOR

angular deflection of the pendulum due to spurious torque in the support, the restoring spring action is small, ensuring good sensitivity and the high axial compliance obviates the possibility of overstressing in the event of high vertical acceleration or shock.

iv) As a side-effect of (iii), the support offers little lateral constraint, and lateral shift of a mere fraction of a micrometre could cause significant error. To overcome this, the pendulum deflection is measured by a bridge-configured set of four electrodes. The output is, to first order, sensitive only to <u>angular</u> displacement of the pendulum relative to the electrode system.

Since the pendulum is free to rotate in two mutually orthogonal planes, an additional orthogonal set of electrodes (driven at a different, non-commensurate frequency) can be used to determine cross-axis tilt (see 2.8).

The sensitivity of the transducer can readily be estimated. Assume a cylindrical pendulum 10mm in diameter and 30mm long, and electrodes each 5mm x 10mm, with a nominal gap of 0,5mm. The capacitance between an electrode and the pendulum is roughly 1pF. A tilt of 10 arc seconds causes a change of differential capacitance at the lower pair of electrodes of some 3×10^{-3}pF. This is indeed a small capacitance change but it has been conclusively shown by Jones et al (127) and by Skalski (131) that by using appropriate guarding techniques and synchronous detection, capacitive changes of this order can be reliably and stably measured. In the experience of the present writer this order of precision is readily achievable.

Although the capacitive pendular transducer appears to offer the best potential for a really accurate and reliable portable tilt sensor, and would seem to be the obvious choice if, say, single second performance or better were required for an electronic theodolite, its fabrication does require reasonably accurate machining and it would not be particularly inexpensive. It was therefore decided to proceed first with the alternative version using a bubble-sensor. In the event, this latter version proved to be cheap, simple and entirely adequate and the pendular sensor was not constructed in final form, although various aspects of its operation were experimentally investigated.

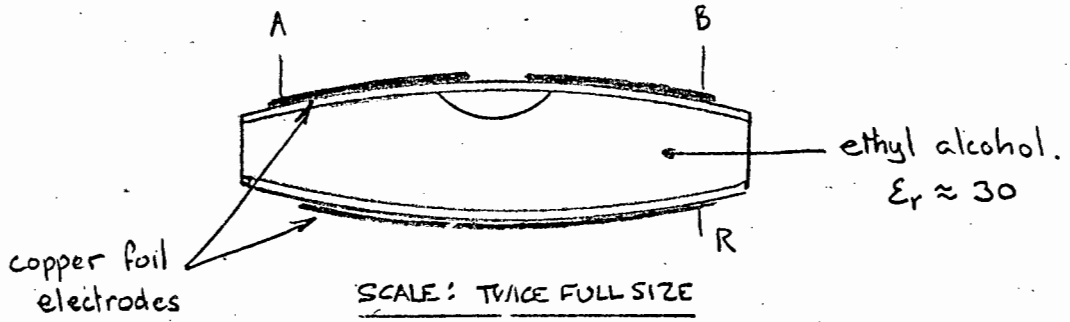
7.5 A CAPACITIVELY TRANSDUCED BUBBLE SENSOR

A conventional bubble sensor consists of a glass vial whose bore is accurately ground to a barrel-shaped surface of large radius of curvature along the length of the vial. It is nearly fully filled with ethyl alcohol, which has a relative permittivity of about 30. When the vial is tilted the liquid moves under gravity, resulting in motion of the bubble in the opposite direction along the vial length. The long radius of curvature affords large mechanical advantage, and typical sensitivities are 2mm of bubble movement for 40 arc seconds tilt. For small displacements the relation is linear, and 10 arc seconds tilt gives rise to 0,5mm displacement of the bubble.

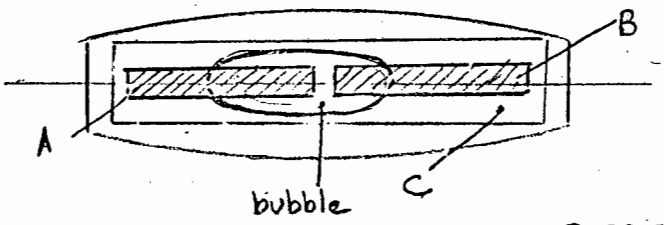
It was conjectured that because of the large dielectric constant of the fluid, it might well be possible to obtain an electrical output signal corresponding to this degree of movement. Some simple experiments were undertaken to establish the viability of the approach and to identify possible problem areas.

A replacement bubble from a builder's level (sensitivity 1mm per 20 arc seconds) was modified by the addition of three copper foil electrodes, which were simply glued to the vial, as illustrated. The unit was mounted on a pivoted table which could be tilted by means of a micrometer screw. As a preliminary test to establish that the bubble would respond without 'stiction' to very small tilts, the bubble edge was observed with a microscope (x 300) while the table on which it was mounted was rocked through an angle of a few arc seconds. There was no evidence of hysteresis or backlash.

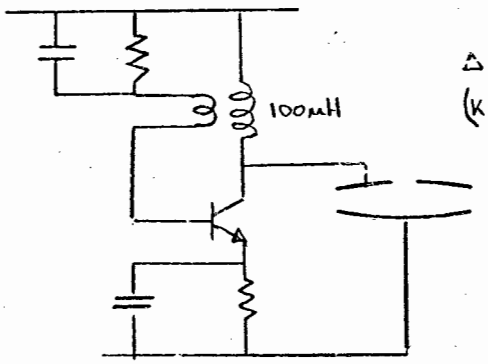
The effect of bubble movement on interelectrode capacitance was monitored by connecting two of the electrodes as part of the resonant circuit of a 1MHz transistor oscillator, having a tuning inductance of 100mH. The oscillator frequency was plotted vs bubble displacement. From this it is readily calculated that 1mm of bubble movement produces a capacitance change of 0,05pF. This is some 17 x greater than the differential capacity change predicted for the pendular sensor and is easily measurable. The increased sensitivity is of course due to the magnification of the bubble movement relative to the movement involved in the tilting.

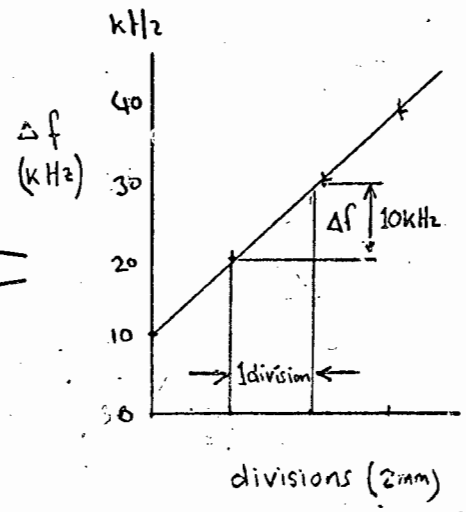


SCALE: TWICE FULL SIZE



BUBBLE WITH ELECTRODES





$f = 3{,}8\,\text{MHz}$.

$C = \dfrac{1}{4\pi^2 f^2 L}$

$= 18\,\text{pF}$

$\dfrac{\Delta C}{C} = 2\,\dfrac{\Delta f}{f}$

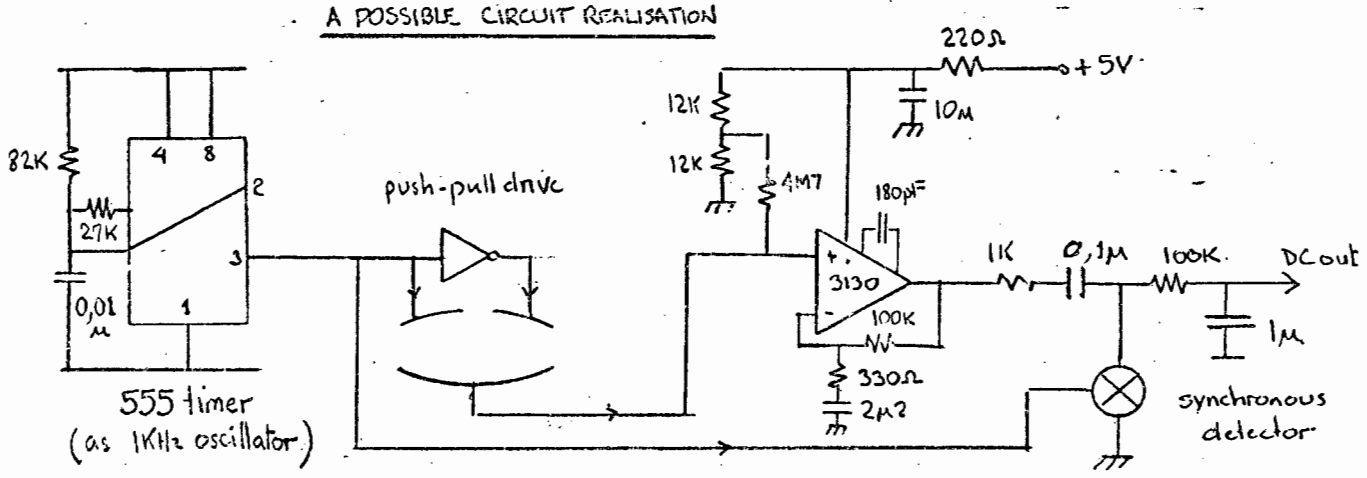
$= 5{,}3\times10^{-3}/\text{division}$

$\therefore \Delta C = 0{,}1\,\text{pF/div.}$

or $0{,}05\,\text{pF/mm}$

CIRCUIT FOR MEASURING CAPACITANCE CHANGE.

A POSSIBLE CIRCUIT REALISATION



A potential problem was thought to be change in sensitivity due to variation in bubble size with temperature. A simple experiment (placing the vial in a refrigerator) showed that bubble length (nominally 6mm) decreases by a factor of about 2 for 20° fall in temperature. However this was easily overcome. The unit was driven in push-pull so that the differential output was a function only of bubble position, and not of size. This makes the tacit assumption that the glass wall thickness of the vial is constant. This was found to be the case within the accuracy required - had it not been, the electrodes could have been (empirically) shaped to compensate for the variation. As a last resort, of course, temperature compensation and/or linearisation could be applied to the electrical output but in the event this proved unnecessary.

The linearity was within 7% over a 60 second tilt range both at normal (15°) and elevated (40°C) temperatures. Low temperature tests are difficult to carry out due to condensation, but there is no reason to think the results would differ appreciably. Zero stability over the 25°C temperature range appeared to be a few arc seconds but it is difficult to know how much confidence can be placed in this due to the possibility of spurious tilting of the supposedly stable base on heating. The zero stability was therefore verified by mounting two identical units on a common base and comparing their outputs at normal and elevated temperatures. One of the units was then rotated through 180° (to obviate masking due to a fortuitously identical drift in the two units) and the test repeated. Agreement between the two units was within 5 arc seconds throughout the test. The problem of providing a cheap, simple and accurate level sensor can therefore be considered solved.

Extension of the principle to two-axis tilt-sensing for cross-tilt compensation (see 2.8) is straightforward. Two orthogonally mounted units could of course be used but a better approach might be to use a bubble vial with circular symmetry and employ two independently driven orthogonal sets of electrodes. Because of the moderate precision required, quadrature drive could be used for the two pairs of electrodes, but if crosstalk is a problem, different frequencies

could be used. Detailed tests were not carried out on this mode, since it is probably not required in most applications, and a suitable sensitive circular bubble was not available. The system was however constructed with a low cost bubble having a sensitivity of 1mm for 10 arc minutes. Apart from the reduced sensitivity, the system performed satisfactorily, with less than 3% spurious cross-axis coupling.

The effectiveness of this simple and low cost method of level sensing is a vindication of the strategy of adapting, where possible, precision components which are inexpensively available due to the sophisticated manufacturing processes possible in high-volume production. It is thought that the system developed may well have wider application.

CHAPTER 8
CONCLUSION

The purpose of this chapter is briefly to draw together the contents of the foregoing chapters, show that all the elements required for the proposed instrument have been realised, and point out where it is believed a useful contribution has been made.

The first chapter identified a gap in the range of existing survey instrumentation, formulated the operational requirements of a new instrument and cast them in the form of a specific target specification for its design. The second chapter addressed the problem of how this specification could best be achieved, arguing in terms of the logic of the requirement and commercial precedent. It was concluded that in order to meet the constraint of low cost, an integrated measuring system was indicated in which as much of the circuitry as possible was shared between the various functions. Subsequent studies showed that the two principal functions - the measurement of range and angle - could most easily and economically be carried out to the required precision using electrical phase-shift as an intermediate variable, and the basis was laid for an instrument centred around phase measurement.

Chapter Three dealt with the problem of a reflected electro-optic atmospheric link, and close attention was focused on the practical behaviour of retro-reflective targets as a frequent source of anomalous behaviour in such links. Possible sources were investigated, as well as the economical realisation of suitable optics - potentially an expensive component of the instrument. The propagation of radiant flux in the system was considered in detail and a model of atmospheric propagation proposed from a wide range of sources in the literature. On this basis experimental predictions of the returned energy were made, and the model was validated by the presentation of experimental results. The techniques evolved in this chapter for the testing of retro-reflectors appear to correlate better with the actual performance of the reflectors than those reported in the literature or generally employed in the industry.

Chapter Four outlines the history, theory and practice of range measurement by electro-optic phase-comparison. Sources of error and techniques for their detection and minimisation are extensively treated - more completely, it is thought, than they have been

previously in the literature. The development of special test equipment which should be of general utility in the design of comparable instruments is described. The foregoing treatment is applied to the design of a 150kHz distance-measuring system and test results are presented showing that, with certain qualifications, the target specification can be met. Critical aspects of the circuit design are discussed in some detail.

Chapter Five surveys existing angle transducers and establishes a case for the development of a new transducer. A comprehensive literature survey discloses the existence of a powerful but little-known technique for angle measurement, and an adaptation of the technique leads to a novel transducer with powerful error-cancelling properties. Two versions of the transducer are developed and tested. The second version is shown to meet the requirements of the present instrument. The present requirement however by no means exhausts the potential of the transducer and it is believed to have much broader application to metrology generally. Special procedures are developed for testing angle transducers.

Chapter Six deals with the technique of digital phase measurement, based on a literature survey and confirmatory experiment. It is shown that the demands of both the range and angle measuring sub-systems in terms of phase measurement are readily met by a fairly simple measuring system. A novel and extremely accurate test method for phasemeters was developed and applied.

Chapter Seven deals with a vertical compensator system which converts the angle transducer of Chapter 6 into a vertical angle-measuring system. Commercial and published approaches to tilt sensing are extensively reviewed and a cost-effective technique proposed, based on the adaptation of an inexpensively available component. Experimental data are presented showing that the required performance is readily achieved.

The comprehensive engineering of an instrument of this complexity is an ambitious task which is properly the work of a team of engineers and technicians over a period of several years, and is beyond the scope of this thesis. What is presented here is the outcome of the research-and-development phase which must precede production design. A novel integrated survey instrument based on phase has

been proposed and the critical design areas identified. In each case a solution has been formulated and experimentally validated and the proposed instrument has been demonstrated to be viable. A comprehensive document has been produced as a guide to subsequent engineering design activity.

In many instances the subsystems and measurement procedures developed exhibit a potential accuracy and versatility which far exceeds the modest requirements of the present instrument. These techniques could with advantage be applied to the more stringent needs of general survey and metrological instrumentation.

REFERENCES

1. Radio Aids to Aeronautical and Marine Navigation. Special Issue Proc. I.E.E. vol. 105 Part B Suppl. 9 (1958)
2. Advances in Marine Navigational Aids. I.E.E. Conference Publication No. 87. London (1972)
3. Allen A: A shore-based aid to marine navigation. Ibid.
4. Bond J: Acoustic transponders as a navigational aid. Ibid.
5. Garrett P.H: River and harbour aid to navigation system. Ibid.
6. Isbister E.J: The use of radar transponders in harbour systems. Ibid.
7. Lee K. and Walden N: Electronic line-of-sight navigational and survey aids. Ibid.
8. Matthews B. and List J: Harbour surveying systems. Ibid.
9. Williams W.P: Manoeuvering and berthing aids. Ibid.
10. Schofield W: Engineering surveying vol. I. Butterworths, London (1972)
11. Marshall A.G: Accurate Measurements save time, money, effort. World Dredging (Dec. 1975)
12. Ridge M: Position Fixing by E.D.M. in the Port of Southampton. Plessey Company Internal Publication (Undated)
13. Marshall A.G: Position System Aids Dredging. Plessey Tellurometer Division Publication 2105/1.
14. Laurita S: Electronic Surveying and Navigation. Wiley (1976)
15. Marshall A., James D. (Plessey S.A.), Baranello R., Romaniello C. (Tellurometer U.S.A.), Harries D., Jackson J. (Survey Dept., Fort Hare) et al - private communications.
16. Hewlett Packard Journal Vol. 27 No. 8 (1976) pp 2 ff.
17. Schöldström R. in Bjerhammer A: Course Lasers in Surveying and Construction. Royal Inst. Tech. Stockholm. (1973) Paper 18.
18. Hewlett Packard Company: 3808A Medium-Range Distance Meter. Technical Data (May 1978)
19. Ehling E.H. (Ed.) Range Instrumentation. Prentice Hall (1967)
20. Froome K.D. and Russel G. The NPL Geoponder: An Experimental Optical Transponder for Electromagnetic Distance Measurement. Survey Review Vol. XXIV No. 184 (1977) pp 51-65.
21. Koechener W: Optical Ranging Employing High-Power Injection Laser Diode. Trans. Aerospace and Electronic Systems. AES No. 1 (1968) pp 81-91.
22. Goldstein B.S. and Dalyrymple G.F: Gallium Arsenide Injection Laser Radar. Proc. I.E.E.E. Vol. 55 No. 2 (1967) pp 181-188.
23. Ekberg J: The Focal Length of the Optics in Optical Communications Systems. Proc. I.E.E.E. Vol. 57 No. 1 (1969) pp 88-94.
24. Henriksen S.W: Proc. I.E.E.E. Vol. 63 No. 5 (1975) pp 813-814.
25. Precision Optics, Calif., U.S.A. Private communication.
26. RCA Electro-Optics Handbook. (1968) Section 12-5.
27. Lehr C.G: Laser Tracking Systems in Laser Applications. Vol. 2. Ed. M. Ross. Academic Press (1974)

28. Plessey S.A. Tellurometer MA100 Manual.

29. Miller S.E. and Tillotson L.C: Optical Transmission Research. Proc. I.E.E.E. Vol. 54 No. 10 (1966) pp 1300-1311.

30. Kerr J.R. et al. Atmospheric Optical Communications Systems. Proc. I.E.E.E. Vol. 58 No. 10(1970) pp 1691-1709.

31. Kazarian R.A. et al. Measurement of the Average Structural Characteristic of the Atmospheric Refractive Index. Proc. I.E.E.E. Vol. 58 No. 10 (1970) pp 1546-1547.

32. Lawrence R.S. and Strohbehn J.W: A Survey of Clear-Air Propagation Effects Relevant to Optical Communications. Proc. I.E.E.E. Vol. 58 No. 10 (1970) pp 1523-1545.

33. Burrus C.A: Radiance of Small Area High Current Density Electroluminescent Diodes. Proc. I.E.E.E. Vol. 60 No. 2 (1972) pp 231-232.

34. Davis R. The Allen Clark Research Centre, The Plessey Company, Caswell, U.K. Private Communication.

35. Karolus A: Physical Principles of the Electro-optical Determination of Distances. J. Geophys. Res. Vol. 65 No. 2 (1960) pp 394-403.

36. Dukes J.N. and Gordon G.B: A two hundred foot yardstick with graduations every microinch. Hewlett-Packard J. Vol. 21 No. 12 (1970)

37. Barkalov S.S. and Deryagin V.N: A Study of the Modulation Phase Distribution over the Emitting Surface of the P-N Junction of a Semiconductor Laser. Sov. J. Opt. Tech. translated Vol. 35 (1969) pp 487-488.

38. Botez D. and Ellenberg M: Comparison of Surface and Edge-emitting LED's for use in fibre-optic communications. I.E.E.E. ED26 No. 8 (1979) p 1230ff

39. Dierschke E.G: Infra-red emitter diodes for optical fibre communications. Int. Conf. on Communications I.C.C. 79 pt. 111 Boston 1979.

40. Gambling W.A. University of Southampton. Private communication.

41. Gooch C.H: Injection Electroluminescent Devices. Wiley (1973).

42. Optoelectronics - Theory and Practice. Texas Instruments Ltd. Ed. Chappell A. (1976)

43. Burrus C.A: Efficient Small Area GaAlAs Heterostructure Electroluminescent Diodes Coupled to Optical Fibers. Proc. I.E.E.E. Vol. 59 No. 8 (1971) pp 1263-1264.

44. Goodfellow R. The Allen Clark Research Centre, The Plessey Company, Caswell, U.K. Private communication.

45. Chinone N. et al. Highly Efficient GaAlAs buried heterojunction lasers with buried optical guide. App. Phys. Lett. Vol. 35 No. 7 (1979) pp 513-516.

46. Gambling W.A. University of Southampton. Private communication.

47. RCA op. cit. Section 7-1.

48. Judge M.E. Radio and Electr. Eng. Vol. 49 No. 11 (1979) pp 545 ff.

49. RCA op. cit. Section 7-4.

50. Kruse, McGlaughlin and McQuistan. Elements of Infrared Technology. Wiley and Sons (1962).

51. Hölscher H.D. in E.D.M; Oxford Symposium, 1955; published Hilger and Watts, London (1967).

52. RCA op. cit. Section 7-4.

53. Ekberg J. op. cit.

54. Richter H. and Wendt H: Range and accuracy of the EOS Electro-optical telemeter; in EDM op cit.

55. RCA op cit. Section 7-4.

56. Richards J.C.S. Capacitor-Fed Parallel Chopper as a Phase-Sensitive Demodulater. Proc. I.E.E. Vol. 118 No. 12 (1971) pp 1723-1728.

57. Hewlett Packard Radiant Flux Meter. Manufacturer's Data.

58. Greene J.R. Precise Measurements Conference. Pretoria (1975)

59. Greene J.R. Geophysical Prospecting Vol. XXV No. 2 (1977) pp 269-279.

60. Greene J.R. Electron Eng. Vol. 45 No. 593 (1977) pp 51-55.

61. Bjerhammar A.E: The New Generation of Electro-Optical Distance Measuring Instruments in Course:Laser in Surveying and Construction. Royal Inst. Tech. Stockholm (1973)

62. Froome K.D: The Velocity of Light and Radio Waves. Academic Press. London (1969)

63. Bergstrand E. Archiv f. Matematik, Astronomi och Fysik, 29a, 1 (1943)

64. Adrianova I.I. et al. Phase Optical Ranging and Optical Radiation Modulation. Sov. J. Opt. Tech. Vol. 37 No. 4 (1970) pp 250-260.

65. Bradsell R.H: The Electronic Principles of the Mekometer III Survey Review Vol. XXI No. 161 (1971) pp 112-118.

66. Bjerhammar A.E: Über Distanzmessungen mittels elektro-optischer Methoden. Zeit. f. Vermessungswesen, Heft. 11;12 (1956)

67. Hölscher H.D: A Short Range Highly Accurate Electro-Optic Distance Measuring Equipment. Conference on Precision Measurements. Graz, Austria (1970)

68. Wadley T.L: Electronic Principles of the Tellurometer. Trans S.A.I.E.E. (May 1958)

69. Midwinter J.E: Optical Fibres for Transmission. Wiley (1979) p 397.

70. Greene J.E: Accuracy Evaluation in Electro-Optic Distance Measuring Instruments. Surveying and Mapping Vol. XXXVII No. 3 (1977) pp 247-256.

71. Zhiltsov A.E. et al. An installation for the Investigation of Radiation from Light-Emitting Diodes that has been Modulated in the Microwave Range. J. Exp. Tech. No. 2 Translated (1974) pp 168-9.

72. Davis R. The Allen Clark Research Centre, The Plessey Company, Caswell, U.K. Private communication.

73. P.A.R. The Synchrohet Commercial Brochure.

74. Anderson L.K. and McMurty B.J. High Speed Photodetectors. Proc. I.E.E.E. Vol. 54 No. 10 (1966) pp 1335-1349.

75. Melchior H et al. Photodetectors for Optical Communication Systems. Proc. I.E.E.E. Vol 58 No. 10 (1970) pp 1466-1486.

76. Vlasov V.G. and Lazneva E.F: Threshold Sensitivity of the Heterodyne Method for Detection of Amplitude-Modulated Radiation with a Photodiode. Sov. J. Opt. Tech. translated (1969) pp 395-397.

77. Personick S.D: Receiver Design for Digital Fibre-Optic Communication Systems. Part 1. B.S.T.J. Vol. 52 (1973)

78. Gooch C.H. op cit. p 138.

79. Johnson K.M. et al. Photo diode Signal Enhancement Effect at Avalanche Breakdown Voltage. Digest of Tech. Papers Internat. Solid State Circuits Conf. Vol. 7 (1964) pp 64-65.

80. Gooch C.H. op. cit. pp 179 ff.

81. Raines J.A. et al. Stabilisation of the Operating Point of an Avalanche Photodetector. J. Physics. E: Sci. Inst. Vol 3 (1970) pp 621-623.

82. Davis Q.V. and Kulezyk W.K: Optical and Electronic Mixing in an Avalanche Photodiode. Electr. Let. Vol. 6 No. 2 (1970) pp 25-26.

83. Hülscher H.D. Semi-conductor Devices for Detecting VHF-modulated Infra-red Radiation. N.I.T.R. Internal Report (undated).

84. Ross M. Laser Receivers. Wiley, New York (1965) pp 54-57.

85. RCA op cit. Section 6-8.

86. Kopeika N.S. and Bordogne J. Background Noise in Optical Communication System. Proc. I.E.E.E. Vol. 58 No. 10 (1970) pp 1571-1577.

87. Personick S.D. op.cit. pp 843-886.

88. Goell J.E. An Optical Repeater with High Impedance Input Amplifier. B.S.T.J. Vol. 53 No. 4 (1974) pp 629-643.

89. Brewer R.C: Transducers for Positional Measuring Systems. Proc. I.E.E. Vol. 110 No. 10 (1963) pp 1818-1828.

90. Sydenham P.H: Linear and Angular Transducers for Positional Control in the Decametre Range. Proc. I.E.E. Vol. 115 No. 7 (1968) pp 1056 ff.

91. Yeliseyev S.V: The Present-Day Period of Development in Geodetic Instrument Construction. Geod. and Aerophot. (1963) pp 51-56.

92. Yeliseyev S.V: Instruments for Precise Measurements of Long Distances and Geodetic Instrument Construction. Ibid Vol. 1 (1965) pp 48-50.

93. Yeliseyev S.V: Precise Chronometric Systems for Obtaining Directions and Angles. Ibid. Vol. 1 (1964) pp 62-64.

94. Kronecker G: Design, Performance and Application of the Vernier Resolver. B.S.T.J. Vol. 36 (1957) pp 1487-1500.

95. Raudenbusch D.H: A High Precision Digital Shift-Position Indicator. I.R.E. Nat. Conv. Rec. Vol. 16 (1958) pp 211-216.

96. Belyy A.V. and Ivanov V.A: Angular Motion Pickup. Sov. J. Opt. Tech. Vol. 40 No. 6 (1973) pp 299-302.

97. Makov D.M: New Devices for Precise Angle and Length Measurements. Conference on Precise Electro-magnetic Measurement. London (1974) I.E.E. Conf. Publ. 113.

98. Furst M.R. The Hewlett Packard 3820A Electronic Total Station. CONSAS Conference, CapeTown (1978)

99. De Bey L. and Webb R: A Shaft-Position Digitiser System of High Precision. I.R.E. Nat. Conf. Rec. Vol. 6 (1958) pp 204-210.

100. Sydenham P.H: An Optical Shaft Resolver using Coupled Light-Modulators. J. Sci. Inst. Vol. 44 (1967) pp 269-270.

101. Hume K.J. an d Sharp G.H. Practical Metrology. Vol. 4 McDonald, London (1962) pp 18 ff.

102. Kissam P: Optical Tooling for Precise Manufacture and Alignment. McGraw-Hill (1962)

103. Middlebrook R.D: Improved Accuracy Phase Measurement. Int. J. Electr. Vol. 40 No. 1 (1976) pp 1-4.

104. Maxwell D.E: A 5 to 50 MHz Direct-Reading Phase Meter with Hundredth-Degree Precision. I.E.E. Trans. on Inst. and Meas. Vol. IM-15 No. 4 (1966) pp 304-310.

105. Frater R.H: A Precision Phase Meter. Ibid Vol. IM-15 No. 1-2 (1966) pp 9-19.

106. Bell E.C. and Leedham R.V: Digital Measurement of Mean Phase Shift. Electr. Eng. Vol. 34 No. 416 (1961) pp 54 ff.

107. Ehret R.L. et al. Linear Integrated Circuit Phase Meter. I.E.E. Trans. on Inst. and Meas. Vol IM-18 No. 3 (1969) pp 157-159.

108. Paull C.J: Phase Measuring with Increased Accuracy. Electr. Eng. (July 1971) pp 52-55.

109. Barnes R.N. and Williams C.C: A Precision Digital Phase Meter. Conference on Digital Instrumentation. London 1973. I.E.E. Conf. Pub. No. 106.

110. McKinney J.E: Digitised Low-Frequency Phasemeter Assembled from Logic Modules. J. of Res. of the N.B.S. - C. Eng. and Inst. Vol. 71C No. 3 (1967) pp 227-238.

111. Rudkin A.M: A New Approach to Phase Measurement. Marconi Instr. Vol. 14 No. 5 (undated) pp 105-108.

112. Kuznetskii S.S. and Chmykh M.K: Digital Methods of Measuring Phase Shift. J. Exper. Tech. translated (1970) pp 1253-1265.

113. Shevelenko V.D. et al. Use of Pulse Spectra to Increase the Sensitivities of Phasemeters. Izmeritel'naya Tekhnika translated (1976) pp 997-1000.

114. Noble F.W. and Cook P.W: Electrical Phase Shift Multiplier. Rev. Sci. Inst. Vol. 36 No. 7 (1965) pp 971-973.

115. Pyatin S.I: Automatic Modulation Methods for Checking Phase Meters. Izmeritel'naya Tekhnika translated (1976) pp 521-523.

116. Makievskii A.E: An instrument for checking Phasemeters. Ibid. (1975) pp 1804-1808.

117. McKinney J.E. op. cit. pp 236.

118. Liadsky J. Private communication.

119. The photofet - Siliconix application note.

120. Trofimerkoff F.N. and Smallwood R.E: Four quadrant JFET Multipliers. I.E.E. J. Solid State Circuits. Vol. SC12 No. 3 (1977) pp 316-318.

121. Plessey S.A. Ltd. MA100 Manual.

122. Frank R.L: Phase-Shift Compensation of Limiter Amplifiers. Proc. I.E.E.E. Vol. 58 No. 2 (1970) pp 257-258.

123. The Principles and Applications of The Electrolevel. British Aircraft Corp. (1967)

124. Simon I. et al. Sensitive Tiltmeter Using a Diamagnetic Suspension. J. Sci. Inst. Vol. 39 (1968) pp 1666-1671.

125. Knobel H.W: The Electric Vacuum Gyro. Control Eng. Vol. 11 (1964) pp 70-73.

126. Stacey F.D. et al. Displacement and Tilt Transducers for 140dB range. J. Sci. Inst. Series 2. Vol. 2 (1969) pp 945-949.

127. Jones R.V. and Richards J: The Design and some applications of Sensitive Capacitance Micrometers. J. Sci. Inst. Vol. 6 (1973) pp 589-600.

128. Jones R.V: The Pursuit of Measurement - 60th Kelvin Lecture. Proc. I.E.E. Vol. 117 (1970) pp 1185-1191.

129. Whitehead T.N: Instruments and Precision Mechanism. Dover (1954)

130. Topcon Theodolites. Commercial Brochure.

131. Skalski C.A: Capacitance Distance Transducer. Proc. I.E.E.E. Vol. 56 No. 1 (1968) pp 112-113.

132. Lusty J. 4-digit phase-meter. B.Sc. Thesis, University of CapeTown (1975)

133. Schofield W. Engineering Surveying. Vol. 1. Butterworth (1972) pp 30 ff.

134. Carter E.A. and Faulkner S.F: New Phase Sensitive Phase Rectifier Circuit. Electron. Let. Vol. 13 No. 15 (1977) p 439.

135. P.A.R. Lock-in Amplifiers. (1975)

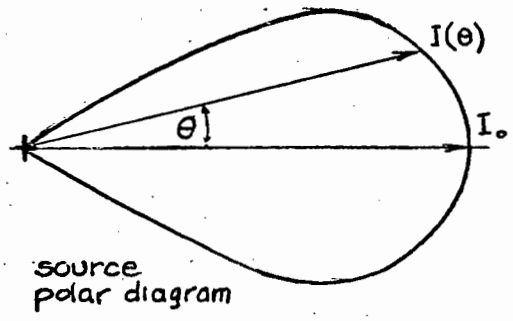
APPENDIX 3.1

RANGE OF AN ELECTRO-OPTIC LINK WITH AN EXTENDED SOURCE

1. Source polar diagram

The irradiance of the source as a function of angle, $I(\theta)$ can be modelled by the relationship (reference 23)

$$I = I_o(\cos\theta)^p$$



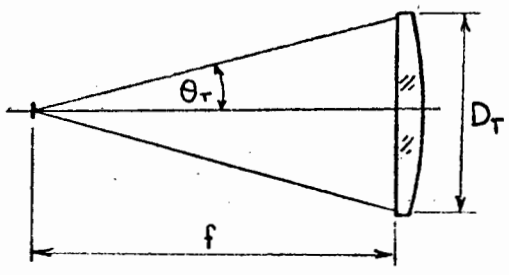
where I_o is the on-axis irradiance. If the full beamwidth between half intensity points is $2\Delta\theta$, p is given by

$$p = \frac{\log_{10} 1/2}{\log_{10} \cos\Delta\theta} .$$

In the case of a TIXL 12 the polar diagram is virtually hemispherical and for $\theta < 180°$, $I(\theta) \doteqdot I_o$. More typically the source is approximately lambertian, with p = 1.

2. Transmitted power in a cone of semi-angle θ_T

Suppose the radiation from a source is collinated by a lens of diameter D_T with a focal length f and a transmission factor T. The power transmitted by the lens will be



$$W_T = T\int_0^{\theta T} I(\theta)\{2\pi \sin\theta\}d\theta$$

For a Lambertian radiator $I(\theta) = I_o \cos\theta$ and we have

$$W_T = T I_o \pi \sin^2\theta$$

$$= T I_o \left[\frac{(\pi/4) D_T^2}{D_T^2/4 + f^2} \right]$$

$$\doteq T I_o A_T \frac{1}{f^2} \left(\frac{1}{1 + 1/4\phi^2} \right) \qquad (1)$$

where A_T is the lens area and ϕ is the numerical aperture f/D_T. For the case where D_T = 60 mm and f = 200 mm, ϕ = 3,33. The term in the brackets is therefore very nearly unity (0,98)

$$\therefore W_T \doteq T I_o A_T / f^2 \qquad (2)$$

Clearly A_T/f^2 is very nearly the solid angle subtended by the lens.

3. Beamwidth and beam intensity

In the far-field, D_I (size of the source image at a distance R) is given by $D_s . R/f$, where D_s is the source diameter.

Therefore the power density $P(W/m^2)$ in the image is given by

$$P = \frac{W_T}{(\pi/4) D_I^2}$$

$$= \frac{W_T 4 f^2}{\pi D_s^2 R^2} \qquad (3)$$

Substituting 2 into 3,

$$P = \frac{T I_o A_T 4 f^2}{f^2 \pi D_s^2 R^2}$$

$$= T I_o \frac{A_T}{A_S} \frac{1}{R^2}$$

where A_s is the source area.

It should be noted that I_o/A_s is the source radiance $N_S (W/m^2 sr)$

$$\therefore P = T N_s A_T \cdot \frac{1}{R^2} (W/m^2).$$

4. Received power

If a receiver lens system of transmittance T_R and area A_R intercepts the beam at distance R, the received power W_R is given by

$$W_R = A_R P T_R$$

$$= T T_R N_S A_T A_R \frac{1}{R^2} .$$

If the minimum required received power is W_{Rmin}, the maximum range is given by

$$R = \sqrt{\frac{T T_R N_S A_T A_R}{W_{Rmin}}} .$$

APPENDIX 3.2

MEASUREMENT OF SOURCE RADIANCE N_S

Method 1

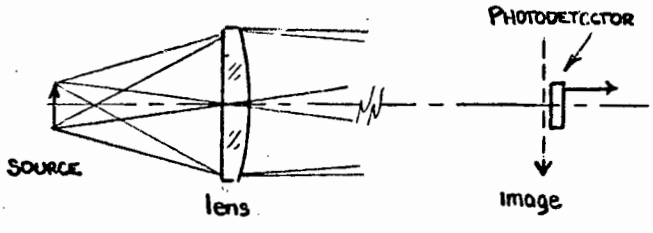
N_s is calculated from measured values of irradiance I and source area A_s.

$$N_s = I/A_s \ .$$

A_s can be determined by visual inspection, or using an infra-red viewer to observe its projected image. Most accurately it can be determined radiometrically, scanning the imaged source with a moving photodiode, and plotting the irradiation profile. Correlated values of mean irradiance and effective source area can then readily be obtained.

Method 2

A magnified image of the source is formed by a lens system of known magnification and transmittance. Radiance can then be directly determined from the power falling on a small photo-detector located within the image.



Let
- A_s be the source area
- f be the lens focal length
- R be the distance to the image
- T_L be the lens transmittance
- A_p be the photodetector area
- W_p be the power incident on the photodetector
- A_L be the lens area

$$W_p = \frac{N_s A_s A_L}{T_L f^2} \cdot \frac{A_R}{A_s (R/f)^2}$$

$$= \frac{N_s A_L A_R T_L}{T R^2}$$

$$\therefore \quad N_s = \frac{W_p R^2}{A_L A_R T_L}$$

If MKS units are used, N will have the units $Wm^{-2} sr^{-1}$. Conventionally N is expressed as N' W/cm^2sr

$$\therefore \quad N'_s = \frac{10^{-4} W_p R^2}{T_L A_L A_R} \quad W/cm^2sr \ .$$

In the experiments, a 60 mm diameter lens was used at an image distance of 1 m. The photodetector had 1 mm diameter, and a responsivity of 0,4A/W. The lens transmittance was measured at 0,8.

$$N_s = \frac{10^{-4} \cdot W_p \cdot 1}{\pi/4 (60 \times 10^{-3})^2 . \pi/4 (10^{-6}) . 0,8}$$

$$= 5,6 \times 10^4 W_p \ W/cm^2sr.$$

In terms of the photocurrent I_p,

$$N_s = 2,2 \times 10^4 \ I_p \ W/cm^2sr.$$

An advantage of this method is that the mean effective radiance over the solid angle subtended by the lens is automatically obtained. Because it is a focused system photocurrents on the order of 1 m A are obtained. These can be measured directly without the need for modulation and synchronous detection.

APPENDIX 3.3

RELATIONSHIP BETWEEN APERTURE, NUMERICAL APERTURE AND CIRCLE OF CONFUSION

Consider a lens system of aperture (effective diameter) D and focal length f. It is customary to speak of the numerical aperture $\phi = f/D$.

If the lens is a simple, optimally configured element used to concentrate incident plane waves at a focal point, spherical aberration will prevent the radiation from being focused into the diffraction-limited Airey circle pattern. The minimum diameter of the circle into which the radiation can be concentrated is called the circle of (least) confusion, and is easily shown by geometrical optics to be

$$d = 0{,}07 \frac{D^3}{f^2}$$

$$= 0{,}07 \, \phi^{-2} D .$$

Thus for $\phi = 3{,}33$, $d/D = 6{,}3 \times 10^{-3}$, and a simple lens of 60mm diameter will have d = 0,4mm.

The optimum configuration for this application (assuming optical material having a refractive index of 1,5) is a biconvex lens with a ratio of curvatures of 6:1, with the more curved surface toward the plane wave. This is very closely approximated in performance by a plane convex lens. It can therefore be seen that at $\phi = 3{,}33$ a 60mm plano-convex lens is just adequate for a 0,4mm source size.

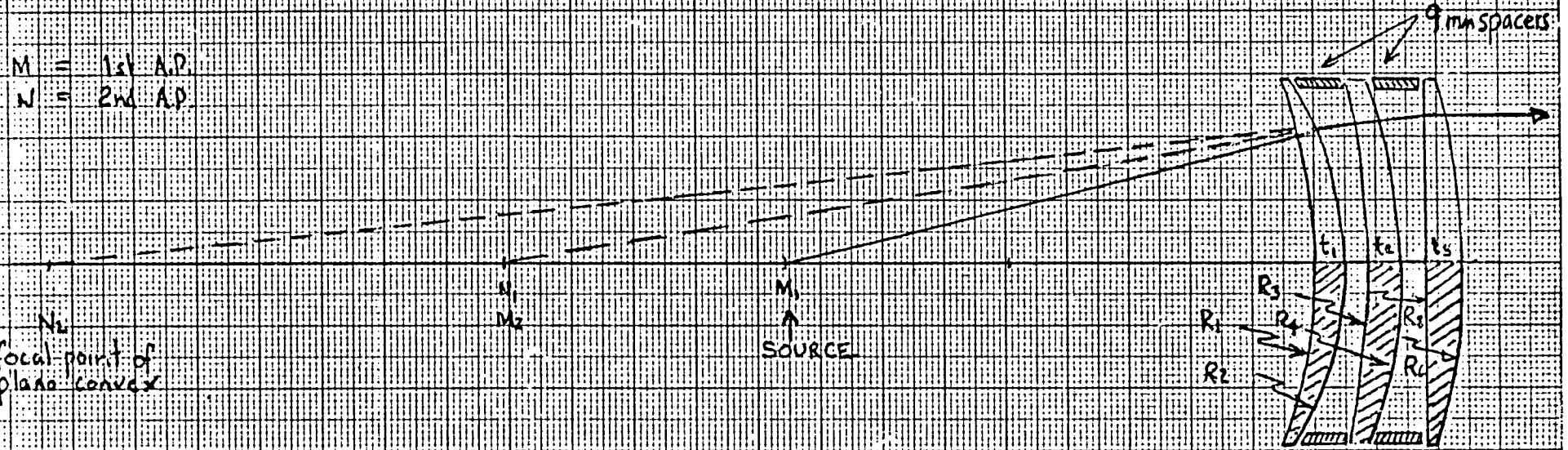
If a lower numerical aperture is required, or a smaller circle of confusion for the same size lens, it is necessary to employ a more complex optical system. If the lens is split into n elements, with the power of the lens optimally distributed (the more strongly convergent lens toward the focal point) the diameter of the circle of confusion is reduced by a factor of about n^2. The lens configuration progresses from plane-convex to meniscus in the direction of the focal point.

A novel approach to the design of a simple collimator lens is to use nearest the focal point a lens whose inner surface is concentric with the focal point of the entire lens system.

If the curvature of the second surface is appropriately chosen at 1,66 times that of the inner surface (for n = 1,5) the source or detector at the focal point is magnified 1,5 times without any spherical aberration. We are here exploiting the freedom from aberration exhibited by the aplanatic points of a spherical surface. The process can be repeated in subsequent elements but the final one must of course be plano-convex, and in this element alone (which has a much higher ϕ) we incur spherical aberration. An experimental design along these lines is shown. A simple model using readily available spectacle blanks was tested, and the principle verified. The performance was not quite as good as predicted (d = 200mm rather than the predicted value of 120mm) but this is probably due to the fact that lenses having the exact calculated values were not available.

The aplanatic design is strictly limited to the on-axis performance required by a collimator, since it exhibits severe coma and field curvature. For the purpose in hand, however, it represents an effective, economical and non-critical approach which is believed to be novel. It is particularly useful if a low numerical aperture system is required in the interests of compactness.

EXPERIMENTAL DESIGN OF AN APLANATIC COLLIMATOR.



Estimated diameter of circle of least confusion = $\frac{0,07}{(1,5)^2}\frac{d^3}{f^2}$

$\frac{0,07}{(1,5)^2}\frac{(58)^3}{(225)^2}$

$= \underline{120\ \mu m}$

$n = 1,523$

$(n = 1,5$ at $0,9\ \mu m)$

R_1	85,39 mm	$t_1 = 5mm$
R_2	54,39 mm	
R_3	139,47 mm	
R_4	87,17 mm	$t_2 = 5mm$
R_5	∞	
R_6	116,22 mm	$t_3 = 6mm$

APPENDIX 3.4

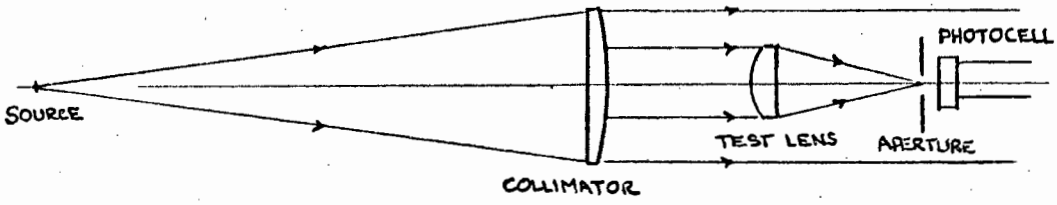
SIMPLE OPTICAL TEST METHODS

Subjective visual estimates of the resolution capability of an optical system show little correlation with its performance in an electro-optic system. The estimates are invariably optimistic on account of the signal-processing capacity of the eye-brain system which has a considerable ability to compensate for lens defects.

A simple radiometric system is therefore required to evaluate lenses and the obvious measure of performance is the circle of least confusion. In reality however, as opposed to the geometrical idealisation, the image of a disc formed by a lens will have a fuzzy, rather than a well-defined boundary. The most appropriate measure is therefore the diameter of the disc which encircles a definite proportion - say 90% of the incident power.

METHOD 1

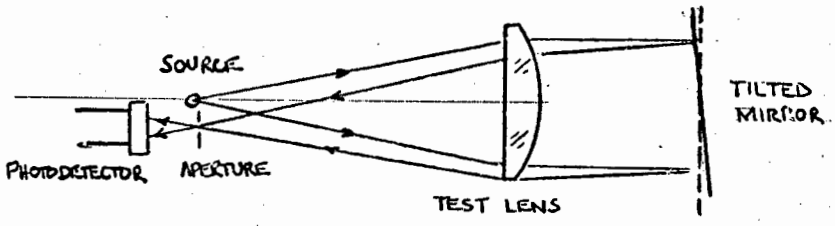
This relies on an accurate collimator to produce a near parallel beam of diameter at least equal to that of the test lens.



The test lens focuses the incident radiation onto an aperture behind which a large area photodiode collects the radiation which has passed through the aperture. Various precision pinholes form the aperture, ranging from 10μm to 1mm. To test the energy encircled by, say a 100μm diameter circle, the photocurrent is measured with a large aperture in place. A 100μm aperture is then substituted, the lens carefully focused for maximum photocurrent and the photocurrent determined as a fraction of that with a large aperture.

METHOD 2

The first method cannot be used unless a suitable collimator lens is available, having an aperture greater than that of the test lens and superior performance, although its performance requirement is facilitated by the fact that it can be of arbitrarily long focal length. It is also possible to use the test lens in an autocollimator mode:



Due to the double passage through the lens the effective measured circle of confusion is effectively doubled.

APPENDIX 4.1

ATMOSPHERIC REFRACTIVE INDEX

The refractive index of the dry atmosphere at standard temperature and pressure (0°C and 760mm Hg) is given by the Barrel and Sears (1939) formula adopted by the IAG in 1963. Strictly this applies only in the visible range, but the effect of dispersion is slight at the near-infrared end of the spectrum, and it seems well justified to apply the formula at the near-infrared wavelengths used in instruments with a gallium arsenide source, where λ = 0,90 to 0,93μm.

The Barrel and Sears formula for phase refractive index gives

$$n_\phi = 1 + (287{,}604 + 1{,}6288\lambda^{-2} + 0{,}0136\lambda^{-4})10^{-6} \quad (1)$$

However, due to the effect of dispersion, this is not the velocity with which a change in light amplitude is propagated. Therefore, for instruments employing modulated light, one must use the group refractive index n_g given by,

$$n_g = n_\phi - \frac{d n_\phi}{d\lambda} \quad . \quad (2)$$

From (1) and (2)

$$n_g = 1 + \{287{,}604 + (3 \times 1{,}6288)\lambda^{-2} + (5 \times 0{,}0136)\lambda^{-4}\}10^{-6} \quad (3)$$

This gives, for

λ = 0,9μm $\quad$ n_g = 1,0002937

λ = 0,93μm $\quad$ n_g = 1,0002933 .

It can be seen how slight is the effect of dispersion. It is unnecessary to know the exact emission wavelength of the instrument. The refractive index of the dry atmosphere is proportional to barometric pressure and inversely proportional to the absolute temperature.

If one writes

$$n_g = 1 + 10^{-6} N_g$$

and takes N from the above discussion as 293,3 at S.T.P., it can be easily shown that

$$N_g = 105{,}5 \frac{P}{(273 + t)} \quad (4)$$

where

P is the barometric pressure in mm Hg, and

t is the atmospheric temperature in °C.

Effect of Water Vapor

The above analysis refers to the dry atmosphere. The effect of water vapor is to add a small negative term N to equation (4) above. This is given by,

$$N_g' = N_g + N$$

where

$$N \doteq -15{,}02 \frac{e}{(273 + t)}$$

where

e is the partial water vapor pressure in mm Hg, and

t is once again the temperature in °C.

That the effect of N can usually be neglected can be seen from the following table which relates N to wet and dry bulb thermometer readings.

Magnitude of N: (sign is negative)

Dry Bulb Temperature °C	-10	0	10	20	30	40
Wet Bulb Depression °C						
0	0,1	0,3	0,5	0,9	1,6	2,6
10			0	0,2	0,6	1,3
20					0	0,4

APPENDIX 4.2

POSITIVE AND NEGATIVE PATTERNS

Consider two signals having the same frequency f and a relative phase angle ϕ. We can represent them by

$$\cos 2\pi ft \text{ and}$$

$$\cos (2\pi ft + \phi).$$

To facilitate the measurement of ϕ it is convenient to translate them in frequency by multiplying each by $\cos 2\pi f't$, by giving

$$\cos 2\pi ft \cdot \cos 2\pi f't, \text{ and}$$

$$\cos (2\pi ft + \phi) \cdot \cos 2\pi f't$$

which we can write as

$$\tfrac{1}{2}\{\cos 2\pi(f - f')t + \cos 2\pi(f + f')t\} \text{ and}$$

$$\tfrac{1}{2}[\cos\{2\pi(f - f')t + \phi\} + \cos 2\pi\{(f + f')t + \phi\}].$$

Selecting the lower sideband by low-pass filtering and putting $f - f' = \Delta f$ we are left with

$$\tfrac{1}{2} \cos 2\pi \Delta ft \text{ and}$$

$$\tfrac{1}{2} \cos (2\pi\Delta ft + \phi).$$

If $f > f'$, $\Delta f > 0$ and the low frequency signals have the same relative phase angle ϕ as the original pair of signals.

If $f < f'$, $\Delta f < 0$. To see the effect on the phase displacement note that

$$\cos(-2\pi\Delta f t + \phi)$$

$$= \cos\{-(2\pi\Delta f t - \phi)\}$$

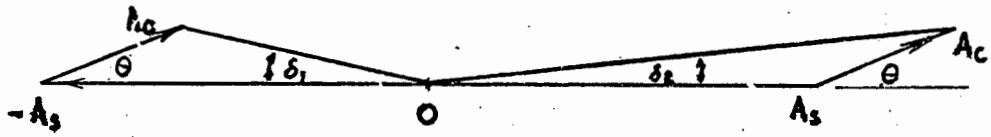
$$= \cos(2\pi\Delta f t - \phi) .$$

Thus the sense of phase displacement is reversed relative to that of the original signals.

NB. Appendix 4,3 bound out of order (after 4.6)

APPENDIX 4.4

CROSSTALK ERROR REDUCTION BY SIGNAL INVERSION



Clearly $\delta_1 \neq \delta_2$ and the error cancellation is not exact. The residual error $\Delta = \delta_1 - \delta_2$ is given by

$$\Delta = \frac{A_C \sin\theta}{A_R - A_C \cos\theta} - \frac{A_C \sin\theta}{A_R + A_C \cos\theta}$$

$$= \frac{A_C \sin\theta}{A_R} \left\{ \frac{1}{1 - \frac{A_C}{A_R}\cos\theta} - \frac{1}{1 + \frac{A_C}{A_R}\cos\theta} \right\}$$

$$= \frac{2\gamma^2 \sin\theta \cos\theta}{1 - \gamma^2 \cos^2\theta} \text{ where } \gamma = \frac{A_C}{A_R} \ll 1$$

$$= \frac{\sin 2\theta}{\gamma^{-2} - \cos^2\theta}$$

Δ has a peak value of approximately γ^2 at $\Theta = 45^\circ$.
It cycles at twice the rate of θ with a periodicity of π.

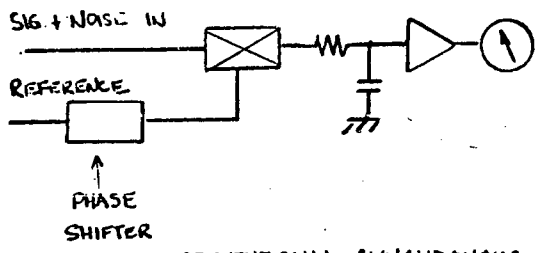
If γ = 1%, $\Delta_{pk} \doteqdot 10^{-4}$ which is negligible.

APPENDIX 4.5

A NEW SYNCHRONOUS DETECTOR FOR CROSSTALK DETECTION

The phase-sensitive detector, or lock-in amplifier, is an important tool in both scientific research and engineering practice for the detection and measurement of signals deeply buried in noise: progressive circuit refinement has led to the commercial availability of instruments permitting the recovery of signals accompanied by wideband noise having an amplitude thousands of times as great (134).

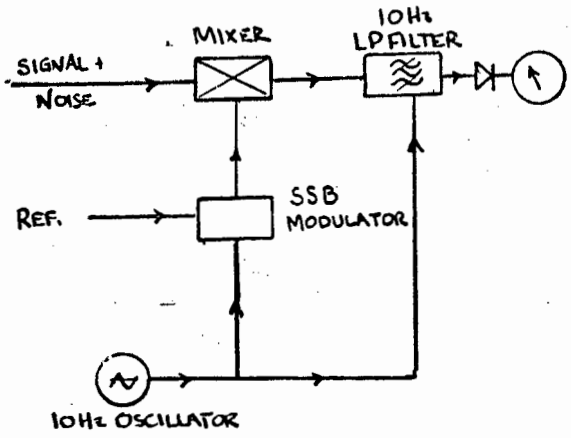
The basic principle of a phase-sensitive detector is simple. It depends on the availability of a reference signal known to be coherent with the signal being sought - a condition that is almost always fulfilled in the case of instrumental work. The incoming signal+noise is mixed multiplicatively in a switching mixer with the reference signal, and the DC resultant due to the component of the input waveform (i.e. the signal) is extracted by means of low-pass filtering. In essence the technique amounts to translating the incoming signal+noise to the vicinity of zero frequency, where the bandwidth can be arbitrarily restricted by simple R-C filtering. Ultimate performance limits are set by DC drift at the mixer output, and intermodulation between signal and noise due to mixer non-linearity.



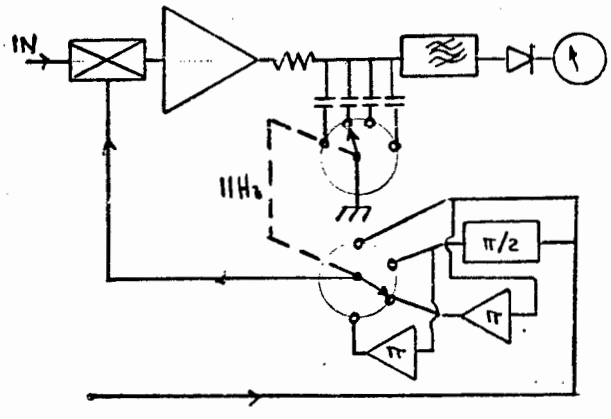
CONVENTIONAL SYNCHRONOUS DETECTOR

Despite the high order of performance attained, dynamic range is limited at frequencies greater than 100kHz, and, of course no output is obtained if the signal and reference happen to be in phase quadrature.

The need for a fixed and known phase relationship can be a serious problem in narrow-band systems where phase varies drastically in response to small frequency shifts. The problem can be solved by the provision of parallel in-phase and quadrature channels with vector summation of their outputs, but there is an alternative approach which seems to offer an excellent and highly cost-effective solution to the problems of both high-frequency performance and sensitivity to phase-shifts.



Instead of translating the input to zero frequency (thereby incurring problems of 1/f noise and drift) we translate it to a band of frequencies centred on, say, 10 Hz. This is achieved by sidestepping the reference by 10Hz in a SSB modulator which, for the single reference frequency, is easily implemented in the form of a progressively-switched phase-shifter. The output of the mixer is processed in an ultralinear narrow-band tracking filter. At the low frequency of 10Hz an almost ideal tracking filter can be realised as a commutated or N-path filter using reed-relays as switching elements. The conflicting requirements in respect of drift and linearity can now be met due to a separation of functions: linear mixing at low level is followed by high-gain low-pass amplification. The N-path filter provides a highly linear filtering operation, even at high drive levels, and the low frequency and high signal levels ensure that breakthrough from the switching to the signal circuit (the most serious limitation of N-path filters) is entirely negligible. Note that the phase relationship between the signal and reference is preserved in the form of a corresponding phase difference between the 10Hz sidestep frequency and the output of the N-path filter.



A simple version of the proposed system as shown in the accompanying diagram has been built and shown to work well. No attempt has yet been made to optimise the design or to determine ultimate performance limits. It should be pointed out that in some respects the proposed system resembles a commercial instrument developed by Princeton Applied Research and marketed as their 'Synchrohet Model 186A' (135). However the underlying principle differs, and the PAR instrument does not have the feature of insensitivity to phase-shift.

APPENDIX 4.6

ON THE EFFECT OF FINITE SIGNAL-TO-NOISE RATIO ON MEASURED PHASE

Let S be the signal power
N be the noise power in bandwidth B
t be the measuring time

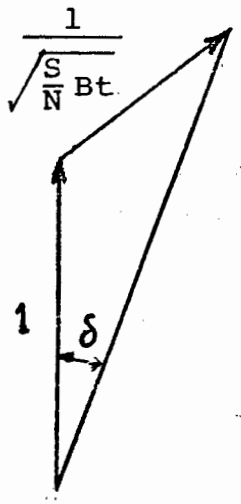
The effective measurement bandwidth will be proportioned to the reciprocal of the measuring time t. The constant of proportionality will depend on the system and the nature of the averaging process (e.g. first-order averaging by low-pass filtering, or integration). For a digital phase detector, however, we have simple numerical averaging over a fixed time and a constant of proportionality of unity is a reasonably accurate assumption.

Thus the S/N (power) ratio in the measurement bandwidth is simply

$$\frac{S}{N} \cdot Bt$$

and the ratio of signal and noise amplitudes is

$$\sqrt{\frac{S}{N}}\ \sqrt{Bt}$$



One effect of the noise will be to perturb the phase of the received signal by an angle δ. Normalising the rms magnitude of the signal to unity, the magnitude of the random-phase noise vector will be

$$\frac{1}{\sqrt{\frac{S}{N}}\ \sqrt{Bt}}$$

A well known result gives, gives for the standard deviation of δ, the value of δ given by a quadrature disposition of the signal and noise vectors.

Thus we have

$$\tan \delta = \frac{1}{\sqrt{\frac{S}{N}}\sqrt{Bt}} \quad \text{(radians)}$$

or approximately,

$$\delta = \frac{180}{\sqrt{\frac{S}{N}}\sqrt{Bt}\ \pi} \quad \text{(degrees)}$$

Thus the product $\sqrt{S/N}\ \sqrt{B}$ is given by $\frac{180}{\delta \sqrt{t}\,\pi}$

for $0{,}1^{o}$ standard deviation in 0,3s measuring time,

$$\sqrt{S/N}\ \sqrt{B} \doteqdot 10^{3}$$

therefore in a 100 Hz noise bandwidth we would require a signal-to-noise <u>amplitude</u> ratio greater than 100.

APPENDIX 4.3

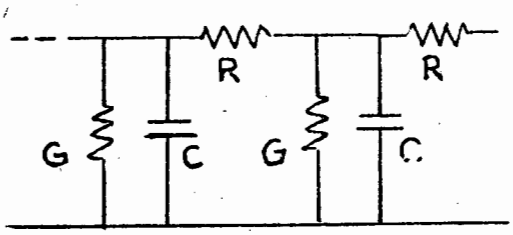
ON TRANSMISSION-LINE DELAY IN A GALLIUM ARSENIDE JUNCTION

It is of interest to see whether significant delay is to be expected due to the distributed resistance and capacitance in the LED junction. A highly simplified model is used, with numerical data supplied by R. Davis of the Allen Clarke Research Centre.

In a simple RC distributed transmission line the velocity of propagation v of a signal of angular frequency ω is given by

$$v = \frac{\omega}{\beta}$$

where β is the phase co-efficient.



If $\omega < G/C$

$$\beta = \frac{\omega C}{2}\sqrt{\frac{R}{G}}$$

$$\therefore v = \frac{2}{\omega C}\sqrt{\frac{G}{R}} .$$

For a gallium arsenide junction, typically we have

R (resistance per unit length) = $3,3 \times 10^{5}$ Ω /m

C (capacitance per unit length)= 4×10^{-6} F/m

G (shunt conductance per unit length)= 4×10^{4} Ω^{-1}/m

$$v = \frac{2}{4 \times 10^{6}} \sqrt{\frac{4 \times 10^{4}}{3,3 \times 10^{5}}}$$

$$= 2 \times 10^{5} \text{ m/s}$$

$\therefore$ time to travel 100μm is = 0,5ns .

This corresponds to a distance error of $\sim$15cm.

APPENDIX 4.7

AVALANCHE PHOTODETECTION

The merits of avalanche multiplication in a photodetector have been exhaustively treated in the literature. It is well-known that avalanche multiplication is advantageous in a situation where the Johnson noise of the photodetector load resistor and subsequent amplifier dominates the system noise. In an atmospheric link, however, the photodetector is likely to be background-limited. The dominant noise is then contributed by the shot noise associated with the background photocurrent and there is little advantage in using an avalanche device.

This can be seen from the expression for the noise-effective power (NEP) of a generalised photodetector. For m = 1 we have a non-multiplying photodetector such as a PN junction photodiode and for m>1 we have a model for an avalanche detector (79).

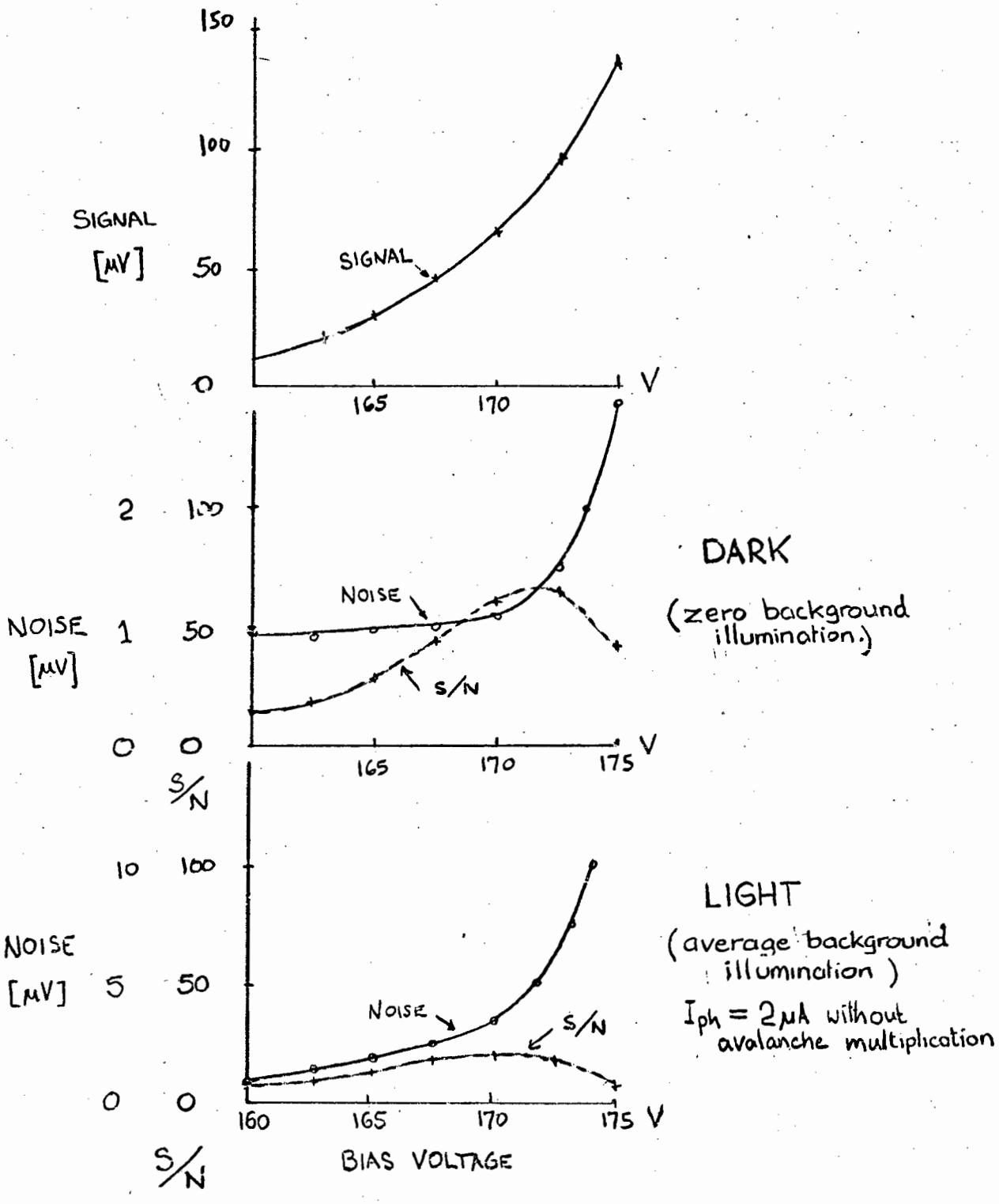
$$(NEP) = \frac{1}{S}\left\{2eI_o + \frac{4kT}{m^2R}\right\}$$

where
- S = photodetector responsivity (A/W)
- e = electronic charge
- I_o = mean photocurrent
- R = load resistance
- m = avalanche multiplication coefficient .

The effect of m is to reduce the loadresistor noise contribution and it has little effect if the first term dominates. The expression is only approximate and its implication that m can be increased indefinitely without penalty is an oversimplification resulting from the neglect of excess noise due to the multiplication process. The optimum value of m depends on I_o but even in total darkness it will rarely exceed 100.

Calculation shows that the advantage to be gained by avalanche multiplication is at best marginal for the background levels anticipated in the present instrument, especially since the source optical bandwidth and wavelength shift with temperature preclude optical filtering narrower than about 400 Angstrom.

Tests were however carried out in which signal and noise levels were plotted as a function of avalanche bias voltage and the signal-to-noise ratio calculated for each point. It can be seen that the improvement in signal-to-noise ratio is dramatic in conditions of darkness but marginal (~3dB) in conditions of ambient illumination simulating the typical operating conditions of the instrument.

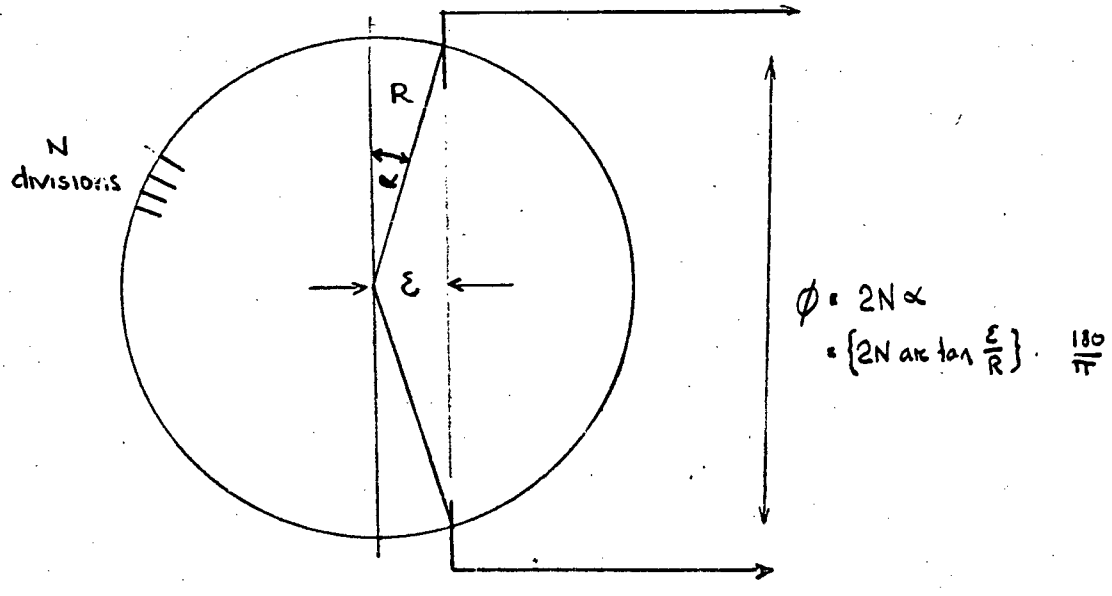


APPENDIX 5.1

A METHOD FOR TESTING BEARINGS FOR ANGLE TRANSDUCERS

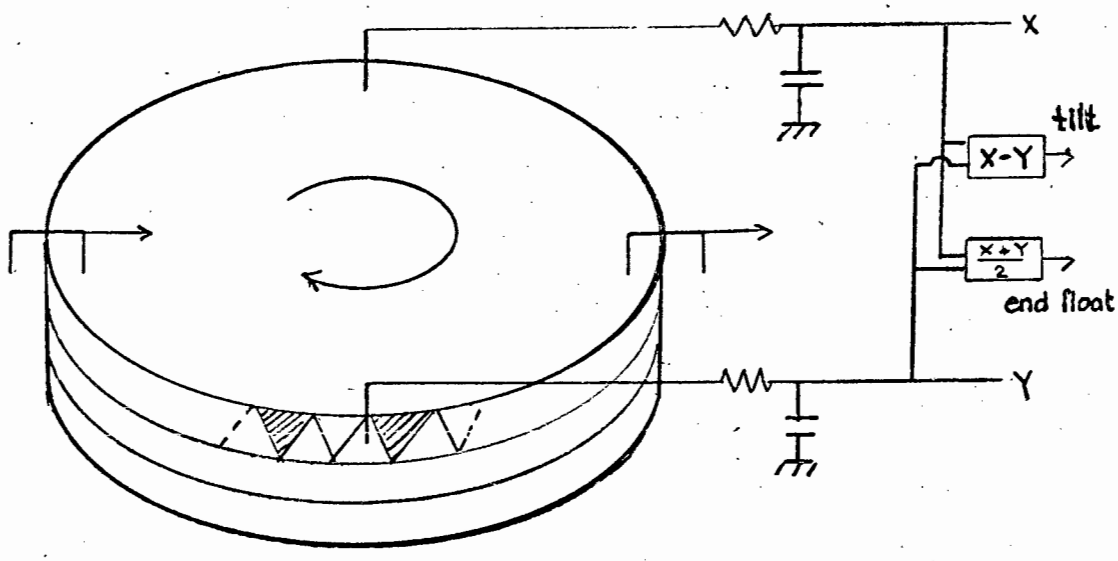
The performance of the bearing system supporting the input shaft is crucial in angle transducers since bearing imperfections such as radial runout and tilt place an ultimate limit on the precision of the transducer. Even in the case of low cost bearings routinely used for domestic and industrial machinery the magnitude of these errors is small, typically measured in micrometres and seconds of arc, and measuring them is difficult unless one has access to highly specialised and very expensive equipment (such as 'Tallyrond').

In the course of developing the angular transducer, it was noted that, by suitably processing the outputs from four intersection sensors distributed around a spinning disc or drum, it is possible to infer eccentric displacement of the input axis relative to the spin axis as well as angular rotation. All that is required is to compute the difference in phase of the signals derived from a pair of diametrally opposed sensors. As can be seen from the diagram, if the spinning element has a radius R and N circumferential divisions, an eccentric motion ε will give rise to a difference in electrical phase $\phi = 2N\{\arctan \varepsilon/R\} \times (180/\pi)$. Since $\varepsilon \ll R$ we have, approximately $\varepsilon = \phi\,(\pi/180)(R/2N)$.



If a drum/type sensor is used and the parallel divisions replaced by a trapezoidal or triangular ones, displacement of the sensor parallel to the spin axis can be inferred from the duty cycle or mark-space ratio of the signal from the sensor. The difference of this displacement related output from a diametral pair of sensors is a measure of tilt. The <u>average</u> change for a diametral pair is of course a measure of end float. Thus, by suitable processing of the signals from a set of four sensors around a suitably configured spinning drum, displacement in respect of all six possible degrees of freedom can be determined.

Angular displacements down to 0,1 arc second and linear displacements in any of the axes of 0,1µm are easily resolved.



APPENDIX 5.2

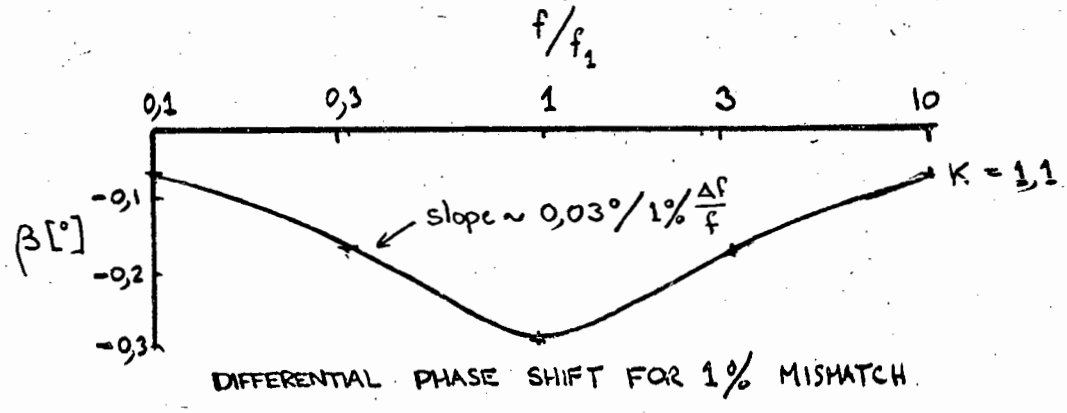
FREQUENCY SENSITIVITY DUE TO MISMATCH OF CHANNEL TRANSFER FUNCTIONS

If the phase/frequency characteristics of the signal and reference channels in an angle transducer are not identical, variations in the rotational speed of the rotor will give rise to differential phase shift and consequent error.

It is a simple matter (e.g. by means of a capacitor across the photodetector transresistance amplifier feedback resistor) to ensure that the dynamic response of both channels is determined by a single dominant time constant. If the channels are not identical assume the 3dB cutoff frequencies are f_1 and kf_1 (for a 10% mismatch $k = 1{,}1$).
If both channels are driven at a frequency f, there will be a differential phase-shift at the output to the channels given by

$$\beta = \arctan (f/f_1) - \arctan (f/kf_1)$$

From this equation β can be plotted as a function of k, with relative frequency variation as a parameter. A simple tachometer servo will easily keep the motor speed (and hence frequency) constant to within 1%, and we can see that for a phase change of 1 part in 10^4 we require $k \approx 1{,}01$. Thus the channel time constants should be matched to within ~1%,



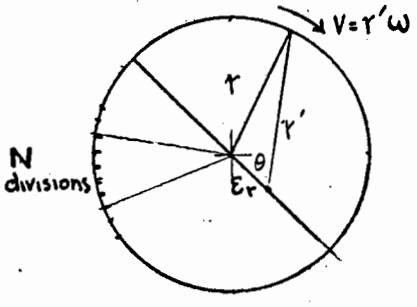
DIFFERENTIAL PHASE SHIFT FOR 1% MISMATCH

APPENDIX 5.3

ECCENTRICITY-RELATED ERROR IN DUAL-MODULATOR TRANSDUCERS

1. Rotor eccentricity

(Non-coincidence between the centroid and the centre-of-rotation of the rotor). Let this eccentricity be ε_r



Nominal frequency $f = \frac{N\omega}{2\pi}$.

Instantaneous frequency $f' = \frac{N\omega}{2\pi}\frac{r'}{r}$

but for $\frac{\varepsilon_r}{r} \ll 1$, $r' \doteq r\left\{1 + \frac{\varepsilon_r}{r}\cos\theta\right\}$

$$\therefore f' = \frac{N\omega}{2\pi}\left\{1 + \frac{\varepsilon_r}{r}\cos\theta\right\}.$$

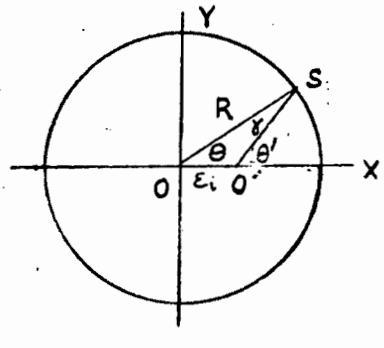
This represents a narrow band fm signal, with a peak deviation (at $\theta = 90^\circ$) of ε_r/r radians. The relative amplitude of the sidebands is then ($\varepsilon_r/2r$) and the energy loss in the sidebands neglibible.

We can see from considerations of symmetry that there will be no nett perturbation of the average phase, provided this is measured over an integral number of complete rotations of the rotor.

2. Input shaft eccentricity (ε_i)

In the case of a cylindrical rotor the radial distance from the centre of rotation is defined by the spinning drum. As we shall see, cancellation of eccentricity error in the mean output of two diametrally-disposed sensors is then exact. This is not the case for a planar rotor, as has been shown by Whitehead (129) for the case of an optically-read circle. We will analyse the superior cylindrical configuration.

There is no loss of generality in aligning the X-axis with the direction of eccentric displacement.



$$\theta' = \theta + \gamma$$

In $\Delta OSO'$, $\frac{\varepsilon_i}{\sin\gamma} = \frac{R}{\sin\theta'}$

$$\therefore \gamma = \arcsin\left\{\frac{\varepsilon_i}{R}\sin\theta\right\}$$

since $\varepsilon_i \ll R$ and $\sin\theta' \leq 1$

we can write $\gamma = \frac{\varepsilon_i}{R}\sin\theta'$.

For the MkI prototype, taking R = 25mm, ε_i = 10μm,

$\gamma_{pk} = 4 \times 10^{-4}$ radian or 1,4 arc minutes.

Consider now a system in which two sensors are mounted at opposite ends of a diameter. The mean error

$$\frac{\gamma_1 + \gamma_2}{2} = \frac{1}{2}\left\{\frac{\varepsilon_i}{R}\sin\theta + \frac{\varepsilon_i}{R}\sin(\theta + \pi)\right\}$$

$$= 0 .$$

If however the averaging process is carried out in an analogue manner, the mean will be weighted by the amplitudes A_1 and A_2 of the sensor outputs. If these are not identical there will remain a residual error given by

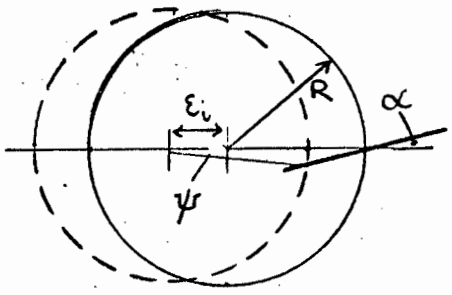
$$\gamma' = \frac{\varepsilon_i}{R} \cdot \frac{2|A_1 - A_2|}{A_1 + A_2} .$$

If however the phase angle of the signal due to each sensor is first converted to a digital number and the averaging accomplished digitally or numerically, we are independent of sensor amplitudes and the cancellation is exact.

The above assumes that the sensors are located on a diameter. It is necessary to consider the case where line joining each sensor to the input axis makes an angle $180^\circ - \delta$ where δ is typically $< 1^\circ$, due to mechanical tolerances. In the case of a cylindrical rotor, the effective circumferential displacement of the sensor is defined by the point of intersection of the circumference and the sensor line-of-action. Thus the angular displacement of a single sensor from its intended diametral location can be resolved into an effective tilt of the line of action of the sensor through δ, and a perpendicular displacement of the line joining the two sensors. This latter effect results in an apparent eccentricity of R tan $(\delta/2)$, which adds to the input eccentricity ε_i. A numerical solution of the complete trigonometric equation relating circumferential displacement to δ confirms the correctness of this simpler approach, which yields greater physical insight into the error mechanism.

3. Effect of tilt angle

The effect of relative tilt in the line of action of the sensors is to produce an effective circumferential displacement of the point of intersection in response to displacement of the input axis along the line joining the sensors.



Clearly,

$$\psi = \arctan\left\{\frac{\varepsilon_i}{R}\tan\alpha\right\}$$

writing $\tan x \doteq x$ for x small,

$$\psi \doteq \left\{\frac{180}{\pi}\,\frac{\varepsilon_i}{R}\,\alpha\right\}^{\circ}$$

If the lines of action of each of a pair of sensors are adjusted to parallelism, error due to tilt cancels. Strictly, exact cancellation depends on the absence of orthogonal eccentric error but the coupling between these effects is extremely weak and discrepencies introduced by treating them independently are in the second order of small quantities. This has been confirmed by a numerical cancellation using the complete trigonometrical expression.

If $\alpha = 1^{\circ}$, $\varepsilon_i = 10\mu m$, then for R = 25mm ψ = 1,5 arc seconds.

In the case of the present instrument therefore, the demands in terms of precision of manufacture are extremely modest.

APPENDIX 5.4

TEST METHODS FOR ANGULAR TRANSDUCERS

1. Direct comparison with standard transducer.

   If a reference transducer is available, and known to be significantly more accurate (say 10 times) than the specified requirements of the test transducer, the transducers may be coupled to a common shaft and the outputs compared for a sufficient number of input angles. Even in this straightforward case caution must be observed:

   (a) the input angles chosen must be well distributed around the circle in such a way as to avoid the inadvertent suppression of high-order periodic errors.

   (b) difficulty may be experienced in ensuring collinearity of the shafts without strain, since the combined shaft will be overconstrained by the redundant bearing system.

If a large diameter flexible coupling (able to tolerate lateral and angular shaft misalignment with sufficiently small angular transmission error) is available, the best solution is to mount the transducers rigidly and couple the input shafts using the flexible coupling. Suitable couplings are made down to 1 arc second accuracy by Itek Measurement Systems, but they are very expensive.

An alternative approach which was used successfully during development work for intercomparing transducers, is to couple the shafts of the transducers rigidly, and to mount one of the transducers rigidly, and the other on a flexural translation device. This is the dual of a gimbal system, providing high torsional stiffness with high lateral compliance. Careful attention to symmetry is required in its design, or lateral deflection may give rise to spurious torsional deflection.

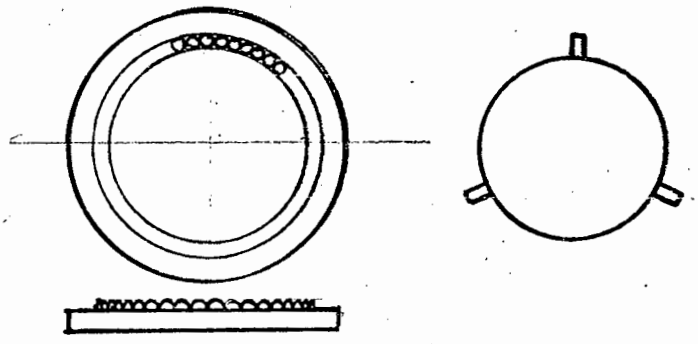
Although this is the most convenient and rapid of all tests to perform, it could not be used in the early development work due to the lack of a suitable reference transducer. Its use was restricted to transducer inter-comparisons.

2. Direct comparison with angular reference unit.
The reference device can take two forms.

(a) The precision polygon. This is a standard piece of apparatus in optical metrology. It consists of regular polygon in quartz or steel, and having from 5 to 12 sides. The angles between the faces, which are accurately plane and highly polished, are identical within, typically, 1, 3 or 10 arc seconds.

(b) Ultradex (or similar) standard angle generator. This is an ingenious device wherein two discs resembling crown gear-wheels engage one another in one of, say, 100 mutual angular orientations. The wheels have accurately shaped triangular teeth and are ground and lapped together so that all teeth are in contact irrespective of angular orientation. Elastic averaging ensures that the angular displacement between adjacent positions of engagement is accurate, typically, to within 0,1 arc second. In some versions a micrometer tangent screw is provided for interpolation between the indent-positions.

Since neither of the above devices was available (or obtainable in time for early development work on the project) a simple device was designed and built, exploiting the high precision and ready availability at low cost, of 10mm steel balls.



The device is constructed as in the diagram. A suitably dimensioned groove was machined in a 10mm aluminium-alloy plate. Steel balls 10mm $\pm$ 1µm were placed on the groove and forced into it in a hydraulic press. Examination under a microscope showed that all balls were in contact.

Another aluminium-alloy disc was provided with three accurately drilled holes, into which three pieces of 5mm silver steel ground stock rod were inserted.

Exact analysis of the probable angular error of such a device is complicated but elementary considerations suggest that it will be of the order of

$$\frac{1 \times 10^{-6}}{150 \times 10^{-3}} \times \frac{180}{2\pi} \times 3600 \text{ arc seconds}$$

or about 1 arc second.

In fact, direct comparison with a T2 single second theodolite gave a standard deviation of comparison of 4 arc seconds, which was adequate for the early development work. The device proved extremely useful. It is interesting that a device which rivals the accuracy of the best dividing heads can be built in this manner at minimal cost. It is felt that there may well be other potential applications.

Use of reference device

The reference device is mounted as concentrically as possible with the transducer and a mirror mounted on the reference device is observed by an autocollimator to establish a reference direction. The reference device is indexed and the reference direction restored by rotating the transducer through an angle which is read off and compared with the indexing angle.

3. Comparison with theodolite.

This method was used extensively. A theodolite (Wild T2 'single second') was mounted on the transducer being tested. The telescope of the theodolite was aimed at a point at virtual infinity provided by another theodolite with illuminated cross-hairs used as a collimator. The transducer/theodolite combination was rotated through the required angle and the theodolite horizontal tangent screw used to restore collinearity of lines of collimation of theodolite and collimator. The angle through which the transducer had been rotated could then be read from the horizontal circle. In order to cancel errors due to circle eccentricity it is necessary to observe on "both faces" i.e. to observe angles with the telescope in both the normal and inverted position.

Effect of errors in the reference device

In all the direct comparison tests the measured errors are of course a combination of the errors in the reference and test unit. Provided however the error pattern is sufficiently repeatable, separation of errors is possible by indexing the reference device relative to the test device and obtaining comparative rotation date at different relative orientations.

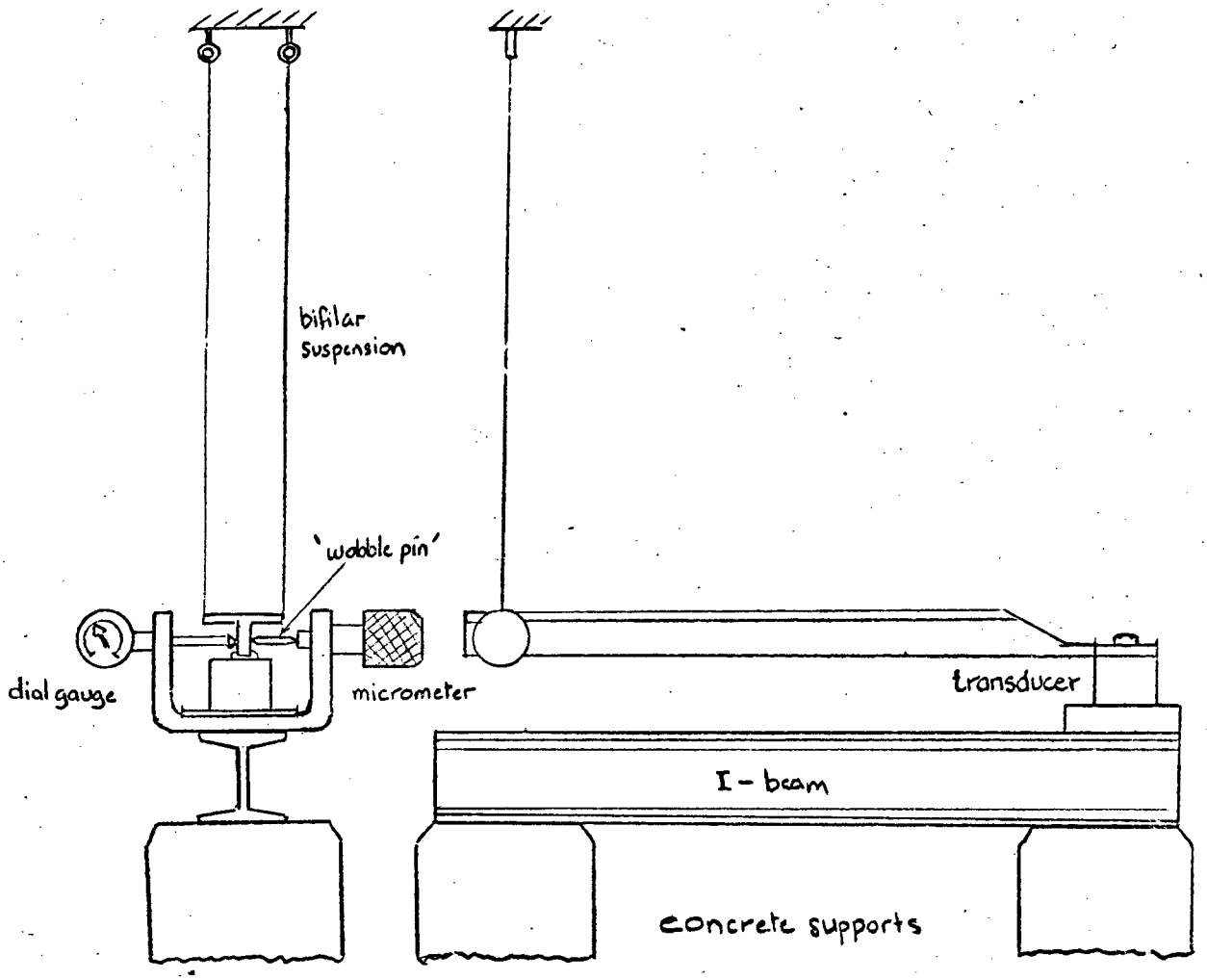
The theory of such error separation techniques has been exhaustively covered in connection with the calibration of glass circles for astronomical use (101) and is treated in several standard works on metrology (102). The mathematical theory is fairly complex although its implementation amounts to a simple numerical procedure. The chief limitation of these techniques is that they are time consuming, in that a very large number of readings must be taken to ensure validity.

Determination of short-range accuracy

The transducers developed in this thesis typically have a non-ambiguous range of only a few degrees. The fine structure of error within a single unambiguous period can be determined using the methods outlined above, but this puts a severe strain on accuracy and resolution requirements on the part of the reference device. Moreover random errors can easily swamp the systematic errors under investigation.

Accordingly a method was developed for generating small angles with a precision on the order of a few arc seconds and a repeatability exceeding one arc second.

The principle is very simple. (See diagram below). A ribbed aluminium alloy arm one metre long was pivoted in a horizontal plane. The vertical axis system of a transducer was used as pivot at one end, ensuring true rotation with lateral play of less than 2µm. The weight of the arm was supported by bifilar wires, and the free end was driven by a differential micrometer screw and its deflection monitored by a dial gauge reading directly to 1µm. Errors periodic with the rotation of the screw were obviated by a short pointed coupling rod known as a "wobble pin". 1µm displacement corresponded to an angle of 1µ rad or 0,2 arc second.



TRANSDUCER SMALL-ANGLE TEST RIG

APPENDIX 6.1

ZERO CROSSING ERRORS IN PHASE DETECTORS

The following simplified treatment is adapted from a number of more comprehensive accounts in the literature. (108-110)

We will consider the measured phase-shift (inferred from zero crossings) between a perfect sinusoid and one contaminated by additive noise, dc offset and harmonic distortion. The case of random noise has already been covered in Appendix 4.6

Effect of DC offset:

To determine the displacement of the zero crossing caused by a D.C. offset e volts in the comparator we need only note that the zero crossings will be given by the solution of the equation

$$E_1 \sin \omega t + e = 0$$

where E_1 is the amplitude of the fundamental.
Therefore, for $e \ll E$, the phase will be perturbed by

$$\Delta\phi = -\frac{e}{E_1} \text{ radians}.$$

Typically if E_1 = 1Vpk and e = 5mV (typical for an untrimmed comparator)

$$\Delta\phi \doteq 0{,}3^{\circ}.$$

To improve on this figure it would be necessary to trim the offset. A better way, discussed by various workers (109;110) is to use the zero crossing of both the positive- and negative-going portions of the waveform. The result is to make the mean zero crossing instant independent of offset. This method was adopted in the present instrument.

Harmonic distortion:

The amount of displacement of the zero-crossing due to harmonic distortion is a complicated function involving the relative amplitudes and phases of the various harmonics. McKinney has shown that a (usually pessimistic) upper bound is given by (110).

$$|\Delta\phi| \leq \sum_{q=2}^{Q} |E_q|/E_1 \quad .$$

Paull (108) points out that when both positive- and negative-going zero crossings are used and the mean computed, the result is free from error due to even harmonics, which tend to constitute the dominant distortion at low levels.

Barnes and Williams (109) carry out a more complete analysis and show that if both zero crossings are used the worst case phase errors due to a harmonic E_n are

$$\frac{90n}{\pi}\left(\frac{E_n}{E_1}\right)^2 \text{ degrees for n even, and}$$

$$\frac{180}{\pi}\left(\frac{E_n}{E_1}\right) \text{ degrees for n odd.}$$

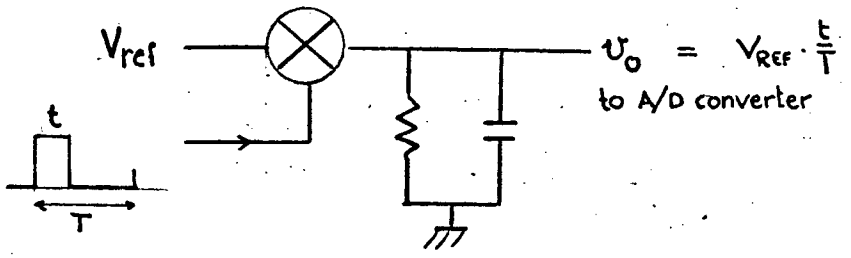
Data extracted from (109) are used in the text.

APPENDIX 6.2

DIGITAL PHASE MEASUREMENT

Many possibilities exist for producing a digital or numerical output proportional to the duty cycle of a rectangular pulse train.

1. Conversion via the analogue voltage domain



A precision (e.g. MOS) switch under the control of the signal bearing the phase information connects an accurately defined reference voltage to a low-pass filter. The average direct voltage across the capacitor is proportional to phase shift. Conversion to the digital form is carried out with a conventional A/D converter or a digital voltmeter.

The method has the advantage of simplicity but it is difficult to achieve high accuracy. For $0{,}1^{o}$ precision the errors introduced by variations in the reference voltage, offset due to the switch, and A/D converter offset and non-linearity must not exceed $2{,}8 \times 10^{-5}\ V_{ref}$, or 1,4 mV when $V_{ref} = 5V$.

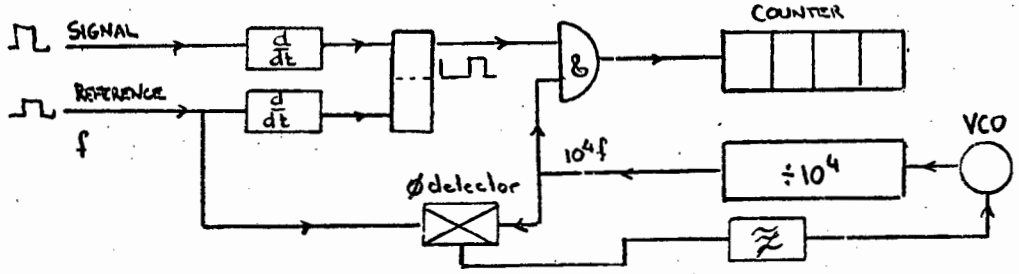
Another advantage is that the smoothing of random variations is easily achieved by the output time constant. Periodic perturbation of phase such as those caused by rotor eccentricity in the angle transducer are less easily dealt with. They can however be eliminated if an integrating A/D converter is used with its integration period synchronised with the rotation rate.

2. Conversion by direct ratio count

In principle if a high-speed clock is gated separately by the intervals T and t (see diagram above) into two counters, the quotient of the contents of the counters represents the desired phase shift. To measure to $0{,}1^{o}$ precision, the clock frequency should be at least 10^{4} times the frequency of the incoming signals to prevent gating- and clock-pulse truncation-error.

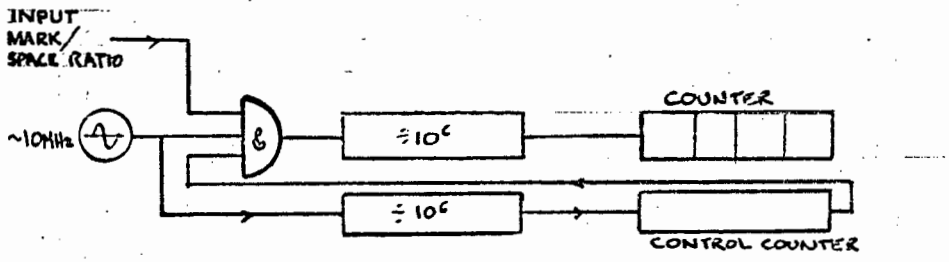
Given the likely availability of a microprocessor in the final instrument, the division is a simple software operation and the method is attractive in its straightforward simplicity. If however computing power is not available, matters can be contrived so that the divisor (i.e. the count accumulated during the period T) is an exact power of 2 or 10 (depending on whether the logic is binary- or BCD- organised). The division then becomes trivial. Many methods for achieving this have been proposed:

i) Phase-lock loop method.



An oscillator can be controlled in a phase-lock loop to produce exactly, say, 10^4 pulses in period T. The content accumulated by the other counter in period t then represents the phase-shift as a decimal fraction of 360^o. This is essentially the method used by Barnes and Williams (109).

ii) A stochastic method.



Here, a finite train of, say, 10^6 pulses is applied to an AND gate. The input duty cycle is asynchronously applied to the other input of the gate. Provided a satisfactory relationship exists between the measuring period, the input frequency and the clock-pulse repetition rate, the number of pulses passing through the gate and accumulating in a following counter will be a faithful measure of the duty cycle. In this case there is an

additional error to be considered - that of asynchronous truncation of the last of the 't' periods. It is necessary to ensure that the number of periods involved in the measurement is such as to constrain the error within acceptable limits. The method works best when the signals are of reasonably high frequency. It has been employed by the writer with excellent results in the measurement of phase at 10kHz to an accuracy better than 2 parts in 10^4 (121).

With measuring times of less than 1 second and measurement frequencies as low as 1kHz asynchronous truncation becomes a serious problem with the stochastic method. In 1974 the writer proposed a modification in which the pulses missing due to truncation were separately accumulated. The error can then be compensated using a simple digitally implemented algorithm. This proposal was subsequently executed under the writer's supervision by J. Lusty (132). Accuracies of 1 part in 10^4 were obtained over the frequency range 100Hz - 100kHz.

All the above approaches were used at various stages of the work recounted in this thesis. As soon as a microprocessor-based data-acquisition system became available however, direct ratio computation seemed the obvious approach, and further work along these lines was discontinued.