Getting the most out of C.O.A.S.T.

David Buscher
Mullard Radio Astronomy Observatory
Cavendish Laboratory
Cambridge

and

Pembroke College

November 1988

PREFACE

The work described in this thesis was carried out under the supervision of Dr P.A.G. Scheuer between October 1985 and October 1988 at the Cavendish Laboratory, Cambridge. The work is my own except where otherwise stated and was not done in collaboration with others. It is the not the same as any other dissertation submitted for a degree, diploma or any other qualification at any university. This thesis does not exceed 60 000 words.

PUBLICATION

Chapter 3 is a modified form of a paper currently in press in *M. Not. R. Ast. Soc.* [11].

ORIGINALITY

Chapter 1 is a review of developments in optical aperture synthesis and is not original work.

Chapter 2 considers the techniques estimating of fringe visibilities from photon-noise-limited data and is mostly a review of previous work, although some of the analysis is my own.

Chapters 3–6 are original except where otherwise stated.

The delay line described in chapter 7 is based on the design developed by Connes [16], but the specific implementation is my own.

Chapter 8 describes experiments which are a continuation of the work of Baldwin *et al.* [6] and Haniff *et al.* [37], but the specific work descibed here is my own.

Chapter 9 is original except where otherwise stated.

David Felix Buscher
November 1988

ACKNOWLEDGEMENTS

I would like to thank everyone in the radioastronomy group who helped to provide a congenial atmosphere in which to work. In particular I would like to thank my supervisor Peter Scheuer, Donald Wilson, Peter Warner, and John Baldwin for many useful discussions, for their constructive comments on the manuscript and for their respective senses of humour.

I am also indebted to Chris Haniff for lessons in practical interferometry with a hammer and chisel, to Graham Woan for interesting discussions on ideas which may or may not have been obvious, and to Alan Matthews for continually reminding me that there was life beyond radioastronomy.

I would like to thank the staff at the Roque de Los Muchachos Observatory in La Palma for their assistance during my observing run there, and Jan Noordam for providing the 'photon-tagging' software for the IPCS.

I am grateful for a three-year studentship from the S.E.R.C., and for addtional support from Pembroke College.

Contents

List of Figures

Chapter 1

Introduction

In modern times, the capabilities of the astronomical optical telescope have increased dramatically: its sensitivity, spectral resolution and spectral coverage have increased by large factors, giving rise to the current explosion in astronomical research. Since the nineteenth century, however, the angular resolving power of ground-based telescopes has remained relatively fixed at about one arcsecond. This limit is imposed by random fluctuations in the refractive index of the atmosphere which perturb the phase of light waves as they travel towards the Earth's surface. As a result, the image formed by the largest telescopes is as blurred as that formed by a telescope with a diameter comparable to the scale length of the atmospheric fluctuations (about 10 centimetres).

At radio wavelengths, in contrast, high resolution images are routinely made in conditions where similar large phase fluctuations are present, using aperture synthesis telescopes and so-called 'closure phase' or 'self-calibration' techniques. It is proposed that these techniques can be carried over directly to optical wavelengths to give many orders of magnitude improvement in the angular resolution of optical images; accordingly the Mullard Radio Astronomy Observatory and the Institute of Astronomy in Cambridge have begun the design and construction of the Cambridge Optical Aperture Synthesis Telescope (COAST) with the aim of making images of astronomical objects with detail on milli-arcsecond scales.

The scientific benefits of this thousandfold improvement in resolution hardly need emphasising: among the current research areas in astrophysics that are limited by the lack of information on such scales we can list close binary systems, the inner regions of active galaxies, stellar mass loss processes, star formation, and the calibration of the cosmic distance scale; a glance at the history of radioastronomy shows the impact that similar improvements in resolution have had in increasing our understanding of known phenomena and in discovering previously unsuspected phenomena.

This thesis is concerned with problems of the technique of optical aperture syn-

thesis, with particular emphasis on the design and performance of COAST. In this chapter we shall review the basic principles of the technique and briefly describe the overall design of COAST. From this basis we can then introduce the specific problems that are dealt with in subsequent chapters.

1.1 Radio VLBI

The approach to the ideas of optical aperture synthesis can be made from many angles because it unites several disciplines, notably optical imaging theory, optical speckle interferometry and radio aperture synthesis. Here we shall start from the techniques used in radio astronomy and proceed to the implementation of these ideas in the optical.

The making of images in radio astronomy has always been fundamentally different in technique from optical imaging, because of the vast difference in the wavelength of the electromagnetic radiation being observed. This means that to acquire an image of the same resolution requires a radio telescope many orders of magnitude larger then its optical counterpart. Instead of building very large monolithic telescopes, radio astronomers have made explicit use of the van Cittert-Zernike theorem, fundamental to all imaging, which relates the spatial distribution of the emission from a source to the mutual coherence of the radiation received at two points separated by a distance $\mathbf{d}_{12}$. For a source at infinity, the relationship is simply a Fourier transform. Radio astronomers measure this coherence or 'fringe visibility' as a function of $\mathbf{d}_{12}$ by correlating the signals received by two antennas and varying their separation. Inversion of the Fourier transform then yields an image of the observed object. The angular resolution of the image is the same as that of a monolithic aperture of diameter d_{max}, where d_{max} is the maximum separation of the antennas, and so this technique is known as 'aperture synthesis'.

The fringe visibility is a complex quantity, possessing both amplitude and phase. As radiostronomers increased d_{max}, eventually to intercontinental distances, they found that the phase of the measured visibilities became corrupted by ionospheric refractive index fluctuations and instrumental effects. However it was realised that these errors could be assigned to the individual *antennas* and that if there were many antennas observing at the same time, then there were more measured quantities than independent errors. To put it quantitatively: if we are observing with M antennas then there will be $M-1$ independent phase errors (because we only measure the differential phase of the signal), but there are $\frac{1}{2}M(M-1)$ observable visibilities. Thus it is possible to derive (for $M \geq 3$) $\frac{1}{2}(M-1)(M-2)$ error-free quantities which are dependent only on the object phase.

The quantities normally chosen are the *closure phases*, the sum of the visibil-

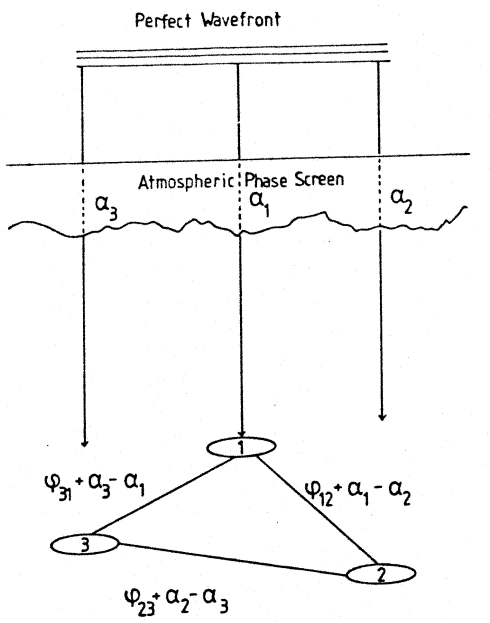

Figure 1.1: *A 'triangle' of antennas observing an object in conditions of large antenna-based phase errors*

ity phases around a loop of (usually) three antennas. To see that these are error-free, we can consider the 'triangle' of antennas in figure 1.1. If the visibility phases, $\{\phi_{12}, \phi_{23}, \phi_{31}\}$, on the three baselines are subject to 'antenna errors' $\{\alpha_1, \alpha_2, \alpha_3\}$, then the measured phases are

$$\begin{aligned}
\phi'_{12} &= \phi_{12} + \alpha_1 - \alpha_2 \\
\phi'_{23} &= \phi_{23} + \alpha_2 - \alpha_3 \\
\phi'_{31} &= \phi_{31} + \alpha_3 - \alpha_1
\end{aligned}$$

but the measured closure phase

$$\begin{aligned}
\phi'_c &= \phi'_{12} + \phi'_{23} + \phi'_{31} \\
&= \phi_{12} + \phi_{23} + \phi_{31} \\
&= \phi_c
\end{aligned}$$

is uncorrupted. It was found that it was sufficient to measure these closure phases and the (usually less corrupted) visibility amplitudes in order to be able to reconstruct a reliable image of the object using iterative methods [69]. Intercontinental radio interferometry (VLBI) routinely produces images of undisputed quality using this technique.

1.2 Application to optical astronomy

Returning to the optical regime, we can see that these ideas can be applied to the phase variations introduced by the atmosphere, though the scales of the fluctuations are very different from those in radio interferometry: whereas in radio VLBI the phase error at any one antenna remains stable for minutes, at optical wavelengths the characteristic timescale, t_o, for these fluctuations is about 10ms, and the characteristic horizontal distance over which the phase error changes by one radian, known as r_o, is about 10cm in the optical compared to a few kilometers at radio wavelengths. Nevertheless, we know that the physics of electromagnetic radiation is invariant between these scales (indeed the principles of radio aperture synthesis were first demonstrated in the optical regime [60]) and so it should in principle be possible to construct the optical equivalent of a radio interferometer and use VLBI techniques to construct high resolution optical images despite the atmospheric perturbations. Such a telescope would in fact obtain better angular resolution than conventional optical telescopes would be able to achieve even in the absence of the atmosphere, which has implications for the future of space-based astronomy.

Following this line of thought, we can now describe the major components of an optical interferometer in terms of their equivalents in radio interferometry.

1.3 The Components of an Optical Interferometer

- **Antennas** In the optical, these become small individual telescopes, typically of order r_o in diameter (the reasons for this will be discussed in later chapters). The 'primary beamwidth' (i.e. the angular resolution of the indivdual array elements) is therefore approximately 1 arcsecond. It is interesting to note that, because of the relatively large focal ratios of optical telescopes, it is possible to observe many unaberrated 'beams' simultaneously in the focal plane of the telescope. However, for acceptable accuracy in the active tilt-correcting systems that must be used (see Chapter 5), it is best to observe objects that are smaller than the primary beamwidth.

- **Amplifiers and Mixers** There is no equivalent to these in an optical interferometer. What is required is a *phase-preserving, broadband* system for transporting the signals from the 'antennas' to the correlator. This is done in radio interferometry by mixing the incoming signal down to some intermediate frequency (I.F.) and sending the electrical signal down a cable. Amplification is used to raise the signal above the background electrical noise level in the cables.

  In the optical it is possible to mix the astronomical signal down to electronic

frequencies (as has been demonstrated in the infra-red by Johnson [47]), but the I.F. bandwidth required for reasonable sensitivity is about 30THz — well beyond the range of current electronics. Hence the signal must be sent to the correlator 'at R.F.' i.e. as an optical beam. Amplifiers are not needed to raise the signal above the noise because the thermal noise present at optical frequencies is negligible.

One possible application for amplifiers, though, would be to boost the signal from each telescope so that it could be split amongst many correlators. This could be done using lasers, but at the current state of technology, lasers are limited to bandwidths of a few GHz. Furthermore, the detectors used in the correlators are quantum-limited (i.e. their thermal noise is negligible), whereas optical lasers are very far from satisfying this specification at present.

Hence the optical signal received by each telescope is simply reflected, unamplified, down a tube (which may be evacuated to remove the effects of atmospheric fluctuations inside the instrument) to the combining optics.

- **Delay lines** Because of Earth rotation, the relative distances travelled by light rays from an astronomical object to the different telescopes is continuously changing. In order to maintain temporal coherence between the beams when they are combined, interferometers must incorporate 'path compensators' which add delays to the signals from different telescopes such that the total travel time of the signal from the object via the telescopes to the correlator is the same (to within the coherence time of the radiation) for all telescopes. This can be done electronically in a radio interferometer, but in the optical this must be done by making the the beam from each telescope travel an extra optical path whose length can be varied. The specification of these optical delay lines is quite stringent: the total delay that has to be introduced to cope with the range of zenith angles involved (about $\pm 30°$) is half the maximum baseline — many tens of metres for a useful interferometer, but this delay must be stable (at least during the exposure time) to a fraction of a wavelength.

- **Correlator** In essence this is very simple — the incoming beams are interfered with one another on detectors to produce $\frac{1}{2}M(M-1)$ fringe patterns; the strength of modulation of these patterns is proportional to the amplitude of the spatial coherence function and the postion of the peak of the modulation with respect to some fixed point gives the phase.

- **Exposure Time** This must be short enough to freeze the motions of the atmosphere which might blur out the fringes. Hence it is of order of the coherence time, t_o, i.e. about 10ms.

- **Observing Wavelength** For highest sensitivity, it is best to use the longest wavelength at which photon-noise-limited detectors are available, because the upper limits to the aperture size and integration time are set by the atmospheric parameters t_o and r_o, and these increase with increasing wavelength.

- **Observing Bandwidth** This is limited initially by the uncertainty in the atmospheric and instrumental path length errors, but if this can be overcome by, for example, tracking the white light fringe, then the only limitation is the bandwidth smearing of objects not at the centre of the field of view. A typical bandwidth used to construct an image 10 pixels across might therefore be of order 10% of the observing frequency. It is relatively simple in the optical regime to spectrally disperse the fringe pattern and thus to observe at many wavelengths simultaneously. This not only allows the sensitivity of the telescope to be increased, but also allows the study of the variation with wavelength of the spatial distribution of the optical emission from an astronomical object.

1.4 A Brief Review of Past and Current Work in Optical Aperture Synthesis

1.4.1 Single Telescope Work

In 1890 Michelson [60] was the first person to successfully use aperture synthesis methods for astronomical measurements, using two apertures mounted on a single telescope to determine the angular diameters of the moons of Jupiter. In later experiments he made the first ever direct measurement of the angular diameter of a star apart from the sun [61].

Thereafter the technique remained largely dormant at optical wavelengths until Labeyrie [53] pioneered astronomical speckle interferometry. In this method the atmospherically perturbed wavefront across the whole pupil of a large telescope, which can be regarded as an array of r_o-sized sub-pupils each with a random phase perturbation, gives rise to multiple fringe patterns in the image plane. This technique gave improved observing efficiency over Michelson's scheme and has been much used, especially for resolving sub-arcsecond binary systems [58] and determining stellar diameters [55]. Both schemes suffer from only being able to recover information about the modulus of the object visibility, and thus true imaging is possible in only a minority of cases.

A number of methods of recovering phase information in speckle interferometry have been tried [48, 2], the most promising of which, the 'speckle masking' method [57], is closely related to the closure phase method in radio interferometry [18]. An-

other method of phase recovery which is even closer to radio astronomical techniques involves blocking off the pupil of the telescope except for an array of small holes. This effectively converts the telescope into a separate-element interferometer so that the closure phases can be measured directly [6] (experiments using this method are described in chapter 8). It is worth noting that both methods have been able to produce true images of astronomical objects in practice as well as in principle [37, 40].

1.4.2 Separate-Element Interferometers

For the highest resolutions, arrays of separate telescopes must be used. Hanbury-Brown and Twiss [36] used the technique of intensity interferometry to determine the angular diameters of stars using baselines of up to 200 metres, but the technique was limited to stars brighter than magnitude 2, and only the visibility amplitude could be measured.

Direct interference between the wavefronts from independent telescopes was first achieved by Labeyrie [54] in 1974. This two-telescope system with 25 centimetre apertures is now installed at CERGA and is making stellar diameter measurements at optical and near-IR wavelengths on North-South baselines of up to 67 metres [50]. Also at this site is an interferometer with two much larger (1.5 metre) telescopes, which is under development.

At Mount Wilson, several two-element interferometers have been built with mainly astrometric goals in mind [81, 14], although some stellar diameter measurements have been made [43]. The latest version, the Mark III interferometer, has a choice of four baselines up to 20 metres in length. This instrument has proven that an optical interferometer can be made to be highly automated in operation — it can acquire and track the fringes from over 150 stars in one night [14].

The University of Sydney is constructing a stellar interferometer with North-South baselines up to 640 metres [25]. The main intention is to measure the modulus of the fringe visibility in order to determine stellar diameters and binary star separations, but the design has been 'left open' to allow for the light from more than two array elements to be combined, and hence for closure phase measurements. The prototype instrument, with a fixed 11.4 metre baseline, has been used to determine the angular diameter of Sirius [26].

A heterodyne interferometer working at 10 micron wavelengths is beginning operation at Mount Wilson [22]. Using CO_2 lasers as local oscillators, this system has relaxed mechanical tolerances because of its heterodyne mode of operation, but the bandwidth is currently limited to a few GHz.

Many other groups around the world (e.g. ESO, Georgia State University, NOAO) have proposals for the construction of interferometers, but to date none of these are

fully funded.

1.5 COAST

All the separate-element interferometers described above were designed primarily as two-telescope devices. For phase recovery with closure phase techniques, a minimum of three telescopes must have their beams combined simultaneously. Furthermore, a good distibution of baselines in 2 dimensions is essential if proper images are to be made. COAST is the only interferometer currently being built with imaging as its primary objective. Consequently, it is a four-element array with movable elements. The possible element positions are arranged in a 'Y' to give good 2-D 'snapshot' coverage (see figure 4.1). A brief summary of its parameters are given below.

- **Observing wavelength** 800nm initially, with operation at 1600nm planned.
- **Bandwidth** 10% of the observing frequency.
- **Maximum Baseline** About 100 metres, giving a resolution of better than 2 milliarcseconds.
- **Telescopes** Siderostats followed by 40 centimetre diameter beam-reducing telescopes. These are mounted on platforms which can be moved to any of 13 foundations.
- **Delay lines** Reflectors on trolleys running on tracks 25 metres long. These will be housed along with the combining optics in an earth-covered tunnel for thermal stability.
- **Correlator** Beam combination system undecided at present, but it will make use of photon-counting avalanche photodiodes as detectors. These have peak quantum efficiencies of greater than 50% and negligible dark count rates when cooled to liquid nitrogen temperatures.

1.6 Outline of Thesis

In this thesis we shall consider the basic design of an astronomical optical aperture synthesis instrument for best performance, with particular emphasis on the design of COAST. Thus the path of enquiry we shall follow will be determined in many cases by the practical constraints inherent in the initial COAST proposal, but the investigations will not be entirely restricted to ideas that are of immediate applicability in the current instrument.

The common theme running through chapters 2–5 is the interaction of two of the problems that most affect ground-based optical interferometry: atmospheric perturbations to the optical wavefront and the low light flux levels from astronomical sources. The former problem reduces the allowable aperture sizes, integration times and optical bandwidths which in turn exaggerate the problems of low flux levels. Thus an optical interferometer is severely limited in the signal-to-noise ratio available on any individual measurement.

In chapter 2, the estimators that must be used in measuring interferometric signals at low light levels are reviewed and formulae for the performance of these estimation methods are obtained. We go on to consider how to maximise the signal-to-noise ratio of these measurements in the presence of atmospheric perturbations: in chapter 3, we determine what aperture size and integration time should be used and in chapter 4, we consider how many array elements the instrument should have and how best to arrange them.

In order to increase the signal-to-noise ratio, active compensation of atmospheric perturbations is necessary, but compensation systems are themselves subject to the same measurement problems at low light levels. In chapter 5 we investigate these systems and how well they can be expected to perform.

The following chapter is a design study for the optical correlator subsystem in COAST, with particular reference to the design parameters developed in the previous chapters.

To gain experience of the problems caused by the requirement for extreme mechanical stability in the instrument, a prototype path compensator was built, and the results are reported chapter 7. These have a bearing on the validity of the assumption in the previous chapters that the system will be dominated by atmospheric rather than instrumental path-length variations.

Because COAST is not yet operational, experimental verification of the ideas of optical aperture synthesis and an appreciation of the problems inherent in real rather than idealised instruments were gained using a scaled down interferometer on a single large telescope. Chapter 8 reports the results of experiments using this ‘aperture masking’ method on the Isaac Newton Telescope.

The final chapter brings together the results of the previous chapters by estimating a limiting magnitude for COAST, and this is used in discussing the kind of astronomy the system will be capable of.

Chapter 2

Photon Noise

2.1 The Problem

It will be shown in chapter 5 that, for most practical schemes, the data from the correlator can always be represented as information about the spatial intensity distribution across some notional 'fringe pattern', $I(\mathbf{x})$. The information we require is the value of the spatial coherence function on each of the $\frac{1}{2}M(M-1)$ baselines sampled by the M telescopes, although we know in advance that there will be $(M-1)$ arbitrary 'antenna' phase errors.

In an ideal world, we could simply use very small apertures ($\ll r_o$) and take very short ($\ll t_o$) exposures. Then, at high light levels, the value of the object coherence function η_{ij} on a baseline ij can be simply obtained from the complex value of the corresponding fringe at frequency $\mathbf{u}_{ij}$ in the detected intensity pattern, $d(\mathbf{x})$:

$$\begin{aligned} D(\mathbf{u}_{ij}) &= \int\int_{-\infty}^{\infty} \exp(2\pi i \mathbf{u}_{ij} \cdot \mathbf{x}) d(\mathbf{x})\, dx\, dy; \\ \eta_{ij} &= M \frac{D(\mathbf{u}_{ij})}{D(\mathbf{0})}\,, \end{aligned}$$

where M is the number of interfering light beams, taken to be of equal intensity, in the fringe pattern.

However in the real world, life is not so simple. It is true that real optical detectors can be found that are virtually noise-free, but with small apertures, short integration times and narrow optical bandwidths, the amount of light received in one exposure is very small. To take a practical example, consider the fringe pattern formed from superposing the light beams from a 4-element array observing 3C273, the brightest quasar, which has an I magnitude of 12. If the array elements have 10cm diameter apertures, the exposure time is10ms, the observing bandwidth is 10% centred on 800nm and the total system efficiency (including detector quantum efficiency) is 20%, the amount of light detected in one exposure will be approximately 4 photons.

Thus the observed fringe pattern will consist of a small number of 'photon events' whose place of arrival will depend on the classical (high light level) fringe pattern, but only in a probabilistic sense. Any attempt to derive information about the classical fringe pattern will therefore be subject to 'photon noise' which is a fundamental quantum limit to the observation, even when perfect detectors are used. In this chapter, we will look at the methods of estimating interferometric parameters in the presence of photon noise, and the signal-to-noise ratios of these estimates.

2.2 Definitions

For the purposes of this investigation we shall consider a noise-free detector with very high spatial resolution sampling a spatial intensity pattern. The results, as shown in chapter 5 can be applied to most detection schemes with little modification.

We shall adopt the analysis of Goodman and Belsher [33] who make use of the semiclassical model of the photoelectric process [32, pp. 466-67]. In this model, the observed intensity pattern will consist of a finite number of 'photoevents', each occurring in a very small region of space and time, which can be represented as narrow impulses of finite energy $h\nu$, where ν is the frequency of the (supposed nearly monochromatic) radiation. The probability of N such events occurring in a region S of the detector in a time τ is given by

$$\Pr(N) = \frac{\lambda^N \exp(-\lambda)}{N!}\,,$$

where

$$\lambda = \frac{\eta}{h\nu} \int\int_S \int_\tau I(\mathbf{x}, t)\, dt\, dx\, dy\,,$$

where $I(\mathbf{x}, t)$ is the classical radiation intensity at position $\mathbf{x} = (x, y)$ on the detector at time t and η is the detector quantum efficiency.

Now if we consider an exposure of fixed length τ and define an aggregate energy density

$$I(\mathbf{x}) \equiv \int_t^{t+\tau} I(\mathbf{x}, t)\, dt\,,$$

then the probability density function for the position of a given photoevent is

$$p(\mathbf{x})\, dx\, dy = i(\mathbf{x})\, dx\, dy\,, \tag{2.1}$$

where

$$i(\mathbf{x}) = \frac{I(\mathbf{x})}{\int\int_{-\infty}^{\infty} I(\mathbf{x})\, dx\, dy}\,,$$

and the mean number of photon events in one exposure will be

$$\overline{N} = \frac{\eta}{h\nu} \int\int_{-\infty}^{\infty} I(\mathbf{x})\, dx\, dy.$$

With this background we can now calculate some quantities of interest. The following analysis will in general be heuristic, with little attempt being made to prove that our solutions are optimal. Instead we shall examine simple *ad-hoc* estimators whose frequency distributions can be easily analysed, in order to gain some understanding of the noise processes at work. These simple estimators also have the advantage of being computationally cheap, which is an advantage if real-time data reduction is envisaged.

2.3 A First Order Estimator

As a first attempt we can examine using our high-light-level estimator on the photon-limited data; that is, we estimate the complex visibility using the Fourier Transform:

$$D(\mathbf{u}) = \int_{-\infty}^{\infty} \exp(2\pi i \mathbf{u}\cdot\mathbf{x}) d(\mathbf{x})\, dx\, dy.$$

If we represent the photon events as Dirac delta-functions

$$d(\mathbf{x}) = \sum_{j=1}^{N} \delta(\mathbf{x} - \mathbf{x}_j)\,,$$

where $\mathbf{x}_j$ is the position of the jth photon and N is the number of photons a given frame, then

$$D(\mathbf{u}) = \sum_{j=1}^{N} \exp(2\pi i \mathbf{u}\cdot\mathbf{x}_j). \tag{2.2}$$

We can now find the expected value of this estimator:

$$\begin{aligned} E[D(\mathbf{u})] &= E\left[\sum_{j=1}^{N} \exp(2\pi i \mathbf{u}\cdot\mathbf{x}_j)\right] \\ &= \left\langle \sum_{j=1}^{N} \int! \int_{-\infty}^{\infty} \exp(2\pi i \mathbf{u}\cdot\mathbf{x}_j) p(\mathbf{x}_j)\, dx_j\, dy_j \right\rangle \end{aligned}$$

where the order of the expectation over the space of all possible photon positions in a given fringe pattern and the summation over the different photons has been interchanged, and where $\langle\rangle$ denotes expectation over the space of all possible classical intensity patterns produced by atmospheric perturbations to the incoming beams.

Substituting equation 2.1 and assuming that the number of photons in a fringe pattern is independent of its shape, we have

$$\begin{aligned} E[D(\mathbf{u})] &= \overline{N}\left\langle \int\int_{-\infty}^{\infty} \exp(2\pi i \mathbf{u}\cdot\mathbf{x}) i(\mathbf{x})\, dx\, dy \right\rangle \\ &= \overline{N}\langle V(\mathbf{u})\rangle, \end{aligned} \tag{2.3}$$

where $V(\mathbf{u})$ is the visibility of the fringe at frequency $\mathbf{u}$ in the high-light-level intensity pattern

$$V(\mathbf{u}) = \int\int_{-\infty}^{\infty} \exp(2\pi i \mathbf{u}\cdot\mathbf{x}) i(\mathbf{x})\, dx\, dy$$

Thus $D(\mathbf{u})$ is an unbiased estimate of the expected high-light-level complex visibility.

Now let us consider the variance of this estimate. For a complex quantity like this the noise in the data may have different variances along different directions in the complex plane. This can be determined for any arbitrary direction from just two parameters, $\mathrm{var}_1(Q)$ and $\mathrm{var}_2(Q)$, which are defined for a complex quantity Q as

$$\mathrm{var}_1(Q) \equiv E(QQ^*) - E(Q)E(Q)^* \tag{2.4}$$

$$\mathrm{var}_2(Q) \equiv E(QQ) - E(Q)E(Q). \tag{2.5}$$

With these definitions, the variance of Q along a direction at an angle θ to the real axis is

$$\mathrm{var}(Q,\theta) = (1/2)[\mathrm{var}_1(Q) + Re\{\mathrm{var}_2(Q)e^{-2i\theta}\}]. \tag{2.6}$$

Thus $\mathrm{var}_2(Q)$ represents the part of the noise which varies with direction in the complex plane and $\mathrm{var}_1(Q)$ represents the circularly symmetric term in the noise. These quantities are calculated for $D(\mathbf{u})$ in appendix A, giving

$$\mathrm{var}[D(\mathbf{u}),\theta] = (\overline{N}/2)\left(1 + Re\left[\langle V(2\mathbf{u})\rangle e^{2i\theta}\right]\right) + \overline{N}^2\mathrm{var}[V(\mathbf{u}),\theta]$$

Notice that because the high-light-level signal has not been assumed to be constant, we automatically include an 'atmospheric' noise term $\overline{N}^2\mathrm{var}[V(\mathbf{u}),\theta]$, due to the fluctuations in the fringe signal caused by atmospheric phase perturbations. This atmospheric noise will be ignored in this chapter but will be dealt with in detail in the next chapter. We also notice a phenomenon pointed out by other authors and which occurs throughout the following analysis, that the variances of photon-noise-limited estimators of fringe parameters depend on the value of the high-light-level Fourier components at higher frequencies, here the component $V(2\mathbf{u})$. To simplify the analysis we will assume that in most practical cases,

$$|\langle V(2\mathbf{u})\rangle| \ll 1$$

and so

$$\mathrm{var}[D(\mathbf{u}),\theta] \simeq \frac{\overline{N}}{2}$$

The estimator $D(\mathbf{u})$ can therefore be represented as the sum of the 'true' complex visibility and a circularly symmetric complex noise process as represented diagramatically in figure 2.1. The probability distribution of this noise process is approximately

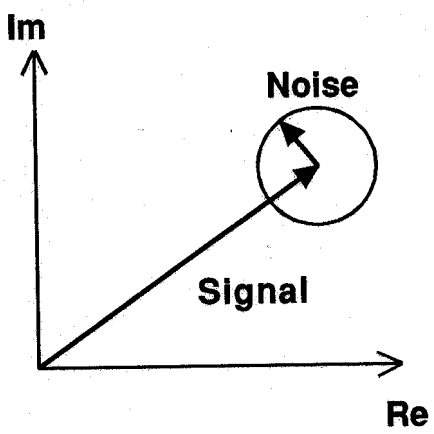

Figure 2.1: *Schematic representation of the signal plus noise*

Gaussian if the number of photons is large, as can be seen by considering equation 2.2 as a 'nearly random' walk in the complex plane.

The signal-to-noise ratio will be given by

$$\begin{aligned} SNR &= \frac{|E[D(\mathbf{u})]|}{(\mathrm{var}_1[D(\mathbf{u})])^{1/2}} \\ &\simeq |\langle V(\mathbf{u})\rangle|\overline{N}^{1/2} \end{aligned} \tag{2.7}$$

Now let us calculate the value of the signal-to-noise ratio for an astronomical observation. Using our previous example of the observation of 3C273 with a 4-telescope array and assuming that all the beams are combined into one pattern, then $\overline{N} \simeq 4$. If we assume for the moment that there is no reduction of the fringe visibility due to the atmosphere and that the nucleus of 3C273 is unresolved, then the visibility of the fringe corresponding to a baseline ij will be

$$|\langle V(\mathbf{u}_{ij})\rangle| = 1/4.$$

Thus

$$SNR \simeq 0.5\,,$$

so that even given very optimistic assumptions, we are in the region of low signal-to-noise ratios. Our analysis from here onwards will centre on this area, which is in contrast to the radio-astronomical case, for which the signal-to-noise ratio is usually much larger than unity after one integration time.

The situation may seem hopeless from this viewpoint, but with 10ms exposures we can accumulate 10^4 fringe parameter estimates in a few minutes' observation, and we can hope to average these estimates in some way to provide a final estimate whose signal-to-noise ratio is much greater than unity. Essentially, we must make use of

our knowledge that the underlying amplitudes and closure phases change only slowly with time — our total integration time is limited only by earth rotation and by the timescale on which the object itself changes.

It is immediately clear that our first estimator $D(\mathbf{u})$ cannot be averaged over many frames — the atmosphere will change the phase of the fringe by many radians between exposures and the mean value of the averaged estimator will be zero. Instead we must split the information into amplitudes and closure phases which can be separately averaged. We shall examine this approach in the next two sections.

2.4 Amplitude

The most obvious way to estimate the amplitude would be to take the modulus of the complex estimate $D(\mathbf{u})$ and find the sample mean over n exposures

$$A_1 = \frac{1}{n}\sum_{k=1}^{n} |D_k(\mathbf{u})| \,,$$

where $D_k(\mathbf{u})$ is the value of $D(\mathbf{u})$ for the kth exposure. Unfortunately manipulation of this expression to determine the performance of this estimation method is not trivial and most authors [33, 20] have instead chosen to study the mean square modulus

$$A_2 = \frac{1}{n}\sum_{k=1}^{n} |D_k(\mathbf{u})|^2.$$

These authors show that this estimate is biased and so construct another, unbiased, estimate:

$$A = \frac{1}{n}\sum_{k=1}^{n} (|D_k(\mathbf{u})|^2 - N_k) \,, \tag{2.8}$$

whose mean is

$$E(A) = \overline{N}^2 \langle |V(\mathbf{u})|^2 \rangle.$$

Dainty and Greenaway [20] show that the signal-to-noise ratio of this estimator is

$$SNR_A = \frac{\overline{N}\langle |V(\mathbf{u})|^2 \rangle}{(1 + 2\overline{N}\langle |V(\mathbf{u})|^2 \rangle + \overline{N}^2 \mathrm{var}\{|V(\mathbf{u})|^2\} + \langle |V(2\mathbf{u})|^2 \rangle)^{1/2}} \,, \tag{2.9}$$

Once again we notice an atmospheric noise term $\overline{N}^2 \mathrm{var}\{|V(\mathbf{u})|^2\}$ and a double frequency noise term $\langle |V(2\mathbf{u})|^2 \rangle$. As before, both of these will be assumed to be negligible. When the single frame signal-to-noise ratio is low (i.e. $\langle |V(\mathbf{u})|^2 \rangle \overline{N} \ll 1$) the signal-to-noise ratio for this estimator goes as roughly the square of the signal-to-noise ratio in equation 2.7.

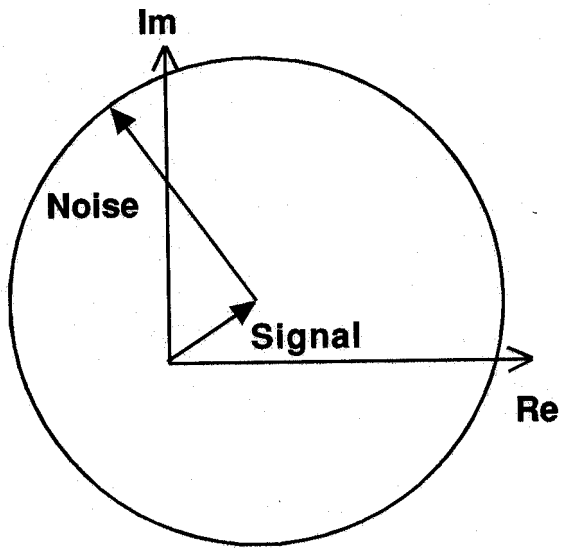

Figure 2.2: *Schematic representation of very low signal-to-noise ratio data*

In order to see whether this is an artefact of the estimator used, a Bayesian solution to the estimation problem can be developed. In Bayesian statistics, probability distributions are used to represent our *state of knowledge* given the data that we have rather than to describe the frequency distribution of some random variable (such as an estimator). Thus a Bayesian analysis will tell us how much relevant information is present in the data without requiring us to explicitly search for an optimal estimator. The analysis for this problem is described in appendix B and it is shown there that the variance of the Bayesian *a posteriori* amplitude distribution is very similar to that of our *ad-hoc* estimator A. The latter estimator is therefore to be preferred since it is much simpler to compute.

Thus at low signal-to-noise ratios, there is an extra loss of information when we are forced to incoherently average our amplitude data. We can give a qualitative explanation for this by comparing the schematic diagrams in figure 2.1 and figure 2.2: we can see that for low signal-to-noise ratios a decrease in the amplitude of the 'true' signal *increases* the amplitudes of many of the measured realisations. If we could average coherently, that is if the phase of the underlying signal was the same for all the frames, then these spurious realisations would be removed because they would have the wrong phase, but in our case we can use only the amplitude information and thus can only rely on the data whose magnitude is slightly greater than the noise, requiring that we take many more samples to accurately determine the result.

2.5 Closure phase

2.5.1 The Statistics of Phase

The statistical treatment of 'phase-like' quantities, that is quantities whose values are only known modulo 2π, for example closure phase, is not widely understood and so we will begin this section by looking at these quantities in a general way and only later on look at closure phase in particular.

Consider a measurement of a phase which gives the result as a value on $(-\pi, \pi)$. If we simply take the average of several such measurements as an estimate of the true phase, the result will be biased, the bias depending on the true phase and the scatter in the measurement. We can see this by considering a true phase which is close to π. Any noise in the measurement will produce many measurements whose value is approximately $-\pi$, giving a mean near to zero. In particular a symmetric distribution centred on π will have a mean of zero. This difficulty can be circumvented by using the modal value as an estimate, but the analysis of such a scheme (for example to determine any bias in our estimate) would be be difficult.

Another point to note is that the central limit theorem does not hold for large phase variances. For instance a measurement whose frequency distribution is uniform on $(-\pi, \pi)$, i.e. one from which no information can be extracted, has a variance of $\pi^2/3$, and one might be tempted to think that this variance could be reduced to a small value by averaging many measurements. Furthermore it is possible to find phase distributions which do contain information which have a larger variance than this, so in general, the variance of a phase is not a very useful quantity.

The 'Unit Vector' Method

Let us however consider another method for averaging phases. Say we are given n measurements $\{\theta_k, k = 1..n\}$, the sum of a fixed phase ϕ and a noise process ϵ_k whose probability distribution is given by $p(\epsilon)$ for all the measurements, and where the noise is uncorrelated between frames. Then let us take as an estimator for ϕ

$$\hat{\phi} = \arg\left(\sum_{k=1}^{n} e^{i\theta_k}\right). \tag{2.10}$$

We shall hereafter call this the 'unit vector method' as what we are doing is assigning the phases to unit vectors in the complex plane and summing these vectors. The behaviour of this estimate can be determined as follows: define

$$\begin{aligned} Re^{i\hat{\phi}} &= \sum_{k=1}^{n} e^{i\theta_k} \\ &= e^{i\phi} \sum_{k=1}^{n} e^{i\epsilon_k}. \end{aligned}$$

Now if $E(\sin\epsilon) = 0$ (for instance if $p(\epsilon)$ is symmetric) and $E(\cos\epsilon) > 0$ (satisfied by any distribution which has more area near the origin than near $\pm\pi$) then

$$\begin{aligned} E(Re^{i\hat{\phi}}) &= e^{i\phi}(nE(\cos\epsilon) + inE(\sin\epsilon)) \\ &= e^{i\phi} \times (\text{a positive real number}). \end{aligned}$$

Thus under this very general set of conditions $\hat{\phi}$ will be an unbiased estimate of ϕ modulo 2π (in the sense that for large enough n, $\hat{\phi}$ will be distributed arbitrarily tightly around ϕ — we must be cautious about talking about the *mean* of $\hat{\phi}$).

Without loss of generality, we will hereafter set ϕ to be zero, in order to look at the variations of $Re^{i\hat{\phi}}$ parallel and perpendicular to the mean vector. We shall define a 'unit vector' as the value of one of the terms in the summation in equation 2.10, i.e. $e^{i\theta_k}$. For this quantity the expected value of the real and imaginary parts are

$$\begin{aligned} \overline{c} &= E(\cos\epsilon) \\ \overline{s} &= E(\sin\epsilon) \end{aligned}$$

respectively and their variances and covariance are

$$\begin{aligned} {\sigma_{cc}}^2 &= E(\cos^2\epsilon) - E(\cos\epsilon)^2 \\ {\sigma_{ss}}^2 &= E(\sin^2\epsilon) - E(\sin\epsilon)^2 \\ {\sigma_{cs}}^2 &= E(\cos\epsilon\sin\epsilon). \end{aligned}$$

If $p(\epsilon)$ is symmetric,

$$\overline{s} = {\sigma_{cs}}^2 = 0.$$

and unit vector can then be described as the sum of a fixed vector along the real axis of length $\overline{c}$ and a complex zero-mean noise process with independent real and imaginary parts with variances of ${\sigma_{cc}}^2$ and ${\sigma_{ss}}^2$ respectively. Adding n such vectors together will give a vector of length $n\overline{c}$ and variances $n{\sigma_{cc}}^2$ and $n{\sigma_{ss}}^2$, and for large n these distributions will be Gaussian in shape. Furthermore, if n is large enough such that $\overline{c} \gg \sigma_{cc}/\sqrt{n}, \sigma_{ss}/\sqrt{n}$ then the variance of $\hat{\phi}$ will be approximately

$$\mathrm{var}(\hat{\phi}) \simeq \frac{\sigma_{ss}^2}{\overline{c}^2 n}.$$

Hence we can define a quantity called the 'phase error' of a symmetric phase distribution:

$$\alpha \equiv \frac{\sqrt{E(\sin^2\epsilon)}}{E(\cos\epsilon)}. \tag{2.11}$$

If $\alpha > 1$ the phase is distributed almost uniformly in $\{-\pi, \pi\}$ but we can 'average' n measurements to give a phase distribution with phase error α'

$$\alpha' = \frac{\alpha}{\sqrt{n}},$$

and if $\alpha' \ll 1$ then the variance of this averaged phase will be given simply by α'^2. It is hence α which is the most useful measure of the spread in a phase measurement, and not the variance.

The 'Amplitude-Weighted Vector' Method

We can seek an improvement on the unit vector method by adding the unit vectors with different weights in accordance with any extra information we may have about the quality of each measurement. In many cases the initial measurement is actually of a complex number, the argument of which is the phase we are trying to average. In such cases it is often true that measurements with higher amplitudes have less error in the phase part and so a mathematically tractable way of including this information is to weight the unit vectors with the measured amplitudes. This is equivalent to simply adding the measured complex numbers

$$\hat{\phi}' = \arg\left(\sum_{k=1}^{n} d_k e^{i\theta_k}\right), \tag{2.12}$$

where $\{d_k, \theta_k\}$ are the measured amplitudes and phases.We shall hereafter call this the 'amplitude-weighted vector method' and a single term in the summation we shall call 'the amplitude-weighted vector'.

The analysis proceeds as before, but with a new phase error defined as

$$\beta \equiv \frac{\sigma_{yy}}{\overline{x}}, \tag{2.13}$$

where $\overline{x}$ is the length of the mean of the amplitude-weighted vector, $\overline{x} = |E(d\, e^{i\theta})|$, and ${\sigma_{yy}}^2$ is the variance of the amplitude-weighted vector perpendicular to the mean vector, i.e. ${\sigma_{yy}}^2 = E(d^2 \sin^2\theta)$ if it is assumed that the mean vector lies along the positive real axis. The variance of $\hat{\phi}'$ will be, for $\beta \ll 1$,

$$\operatorname{var} \hat{\phi}' = \beta^2,$$

as expected.

It is shown in reference [98] that, for a complex signal with circularly symmetric complex Gaussian noise, the unit vector and amplitude-weighted vector phase errors of a given measurement are related by

$$\alpha = \eta\beta,$$

where

$$\eta \simeq \begin{cases} 1 & \text{if } \beta \ll 1 \\ 2/\sqrt{\pi} & \text{if } \beta \gg 1 \end{cases}$$

and

$$\beta = \frac{1}{\sqrt{2}\,SNR}\,, \tag{2.14}$$

where SNR is the signal-to-noise ratio of the complex data. This means that the unit vector and amplitude-weighted vector methods are equivalent for high signal-to-noise ratio data, but that the amplitude-weighted vector method has phase errors which are about 10% smaller in the case of low signal-to-noise ratio data.

2.5.2 Addition of Phases

We now consider what happens to the phase noise when we add phase measurements, for example when we determine a closure phase. The results of this section are also applicable to phase reconstruction algorithms such as the Knox-Thompson algorithm that explicitly add measured phase differences in order to determine an integrated phase [48].

Say we are given measurements of m different phase quantities $\{\theta_k, k = 1..m\}$ such that $\theta_k = \phi_k + \epsilon_k$, where the $\{\phi_k\}$ are the 'true' phases and the $\{\epsilon_k\}$ are realisations of m independent symmetric zero-mean noise processes with frequency distributions $\{p_k(\epsilon_k)\}$. The sum of these measurements

$$\theta_+ = \sum_{k=1}^{m} \theta_k$$

will have a phase noise with distribution

$$p_+(\epsilon_+) = p_1 * p_2 * \ldots p_m\,,$$

where $*$ denotes convolution. Now if the initial phases are known only modulo–2π then their sum can only be known modulo-2π and so when the individual phase errors are large their sum will 'wrap round' and cause a large increase in the total phase error. We can calculate the phase error on the sum in terms of the phase errors on the individual phases and appendix C shows that when the individual phase errors are small,

$${\alpha_+}^2 \simeq \sum_{k=1}^{m} {\alpha_k}^2\,,$$

where α_k and α_+ and are the phase errors of the individual phases and their sum respectively. This is similar to what we would expect intuitively when we are adding independent distributions. When the phase errors of the individual phases are large, however, we have the result

$${\alpha_+}^2 \simeq \frac{1}{2}\prod_{k=1}^{m} 2{\alpha_k}^2.$$

Thus the phase error of the sum increases very rapidly at low signal-to-noise ratios, imposing a strong constraint on the lowest signal-to-noise ratio that can be used.

The above analysis ignored any amplitude information present, i.e. it was a 'unit vector method' analysis. We can examine also the performance of the amplitude-weighted vector method, although in this case there is no single amplitude we can attach to the phase sum. Mathematical convenience suggests the use of the product of the individual amplitudes as our weighting factor for the unit vector, i.e. we take the product of the individual amplitude-weighted vectors

$$R_+ e^{i\theta_+} = \prod_{k=1}^{m} d_k e^{i\theta_k}$$

In this case we get a similar result for the phase error of the sum (see appendix C):

$${\beta_+}^2 \simeq \begin{cases} \sum_{k=1}^{m} {\beta_k}^2 & \beta_k \ll 1 \forall k \\ \frac{1}{2}\prod_{k=1}^{m} 2{\beta_k}^2 & \beta_k \gg 1 \forall k \end{cases}.$$

2.5.3 Application to Closure Phase

It is clearly better to use the amplitude-weighted vector method for averaging closure phases, since we have amplitude information. Say we have measurements of the complex visibilities of the fringes corresponding to a triangle of baselines $D_{12,k}$, $D_{23,k}$, $D_{31,k}$ where k refers to the exposure number. Then, if these measurements come from three *separate* fringe patterns (so that the noise is uncorrelated between fringes), our estimator for the closure phase is

$$\hat{\phi}_{123} = \arg\left(\sum_{k=1}^{n} T_{123,k}\right)$$

where $T_{123,k}$ is the 'triple product' [57] for the kth exposure

$$T_{123,k} = D_{12,k} D_{23,k} D_{31,k}.$$

In the case of three fringe measurements from the *same* fringe pattern, we can no longer assume that the noises are will be uncorrelated as the same photons contribute to the noise on each fringe. Wirnitzer [97] shows that in this case the simple-minded triple product estimator is biased and constructs an unbiased triple product

$$\begin{aligned} T_{123,k} &= D_k(\mathbf{u}_{12}) D_k(\mathbf{u}_{23}) D_k(\mathbf{u}_{31}) \\ &\quad - |D_k(\mathbf{u}_{12})|^2 - |D_k(\mathbf{u}_{23})|^2 - |D_k(\mathbf{u}_{31})|^2 \\ &\quad + 2N_k \end{aligned} \tag{2.15}$$

where $\mathbf{u}_{12}, \mathbf{u}_{23}, \mathbf{u}_{31}$ are the fringe frequencies corresponding to the baselines between a set of three telescopes such that $\mathbf{u}_{12} + \mathbf{u}_{23} + \mathbf{u}_{31} = 0$.

Because of this latter relationship, we can define the triple product as a function of *two* spatial frequencies

$$T_k(\mathbf{u}_{12}, \mathbf{u}_{23}) \equiv T_{123,k}$$

since $\mathbf{u}_{31} = -\mathbf{u}_{12} - \mathbf{u}_{23}$. This function is known as the *bispectrum* [57] and will be used interchangeably with the term 'triple product'.

The variance of the triple product estimates for the separate fringe pattern and the common fringe pattern cases are calculated in full in appendix A but the results are in general agreement with one another and with the analysis of section 2.5.2. We shall hereafter assume that the fringes are all in the same pattern, and for this case the phase error is given by

$$\beta \simeq \frac{(1/2 + \ldots + \overline{N}^3 \mathrm{var}_{yy}[V(\mathbf{u}_{12})V(\mathbf{u}_{23})V(\mathbf{u}_{31})])^{1/2}}{\overline{N}^{3/2}|\langle V(\mathbf{u}_{12})V(\mathbf{u}_{23})V(\mathbf{u}_{31})\rangle|}\,, \tag{2.16}$$

where only the noise terms important at the low and high light level extremes have been retained, and where $\mathrm{var}_{yy}[V(\mathbf{u}_{12})V(\mathbf{u}_{23})V(\mathbf{u}_{31})]$ is the variance due to atmospheric noise of the high light level triple product $V(\mathbf{u}_{12})V(\mathbf{u}_{23})V(\mathbf{u}_{31})$ perpendicular to the direction of its mean.

2.5.4 Correlations between closure phases

For an array of M telescopes, there are $\frac{1}{6}M(M-1)(M-2)$ different baseline triangles, but the number of linearly independent *object* closure phases (i.e the number of constraints on the phase of the object Fourier transform) that can be measured is only $\frac{1}{2}(M-1)(M-2)$. In other words, in a noise-free instrument there would be no point in computing all the possible closure phases since they could all be derived from a linearly independent subset. In the presence of noise, however, it would be worthwhile computing the full set closure phases if the errors on the different closure phases were uncorrelated, since we would then gain in the number of independent constraints on the object closure phases. To see if this is the case, we must calculate the correlation coefficient between the triple products in the presence of noise.

We shall consider first a simple model where the visibility measurement on each baseline is the sum of a complex signal S_{ij} and a circularly symmetric complex noise process n_{ij} which is uncorrelated between baselines. The measured triple product on a triangle between telescopes 1, 2 and 3 is then

$$\begin{aligned} T_{123} &= (S_{12} + n_{12})(S_{23} + n_{23})(S_{31} + n_{31}) \\ &= S_{12}S_{23}S_{31} \\ &\quad + S_{12}S_{23}n_{31} + S_{23}S_{31}n_{12} + S_{31}S_{12}n_{23} \\ &\quad + S_{12}n_{23}n_{31} + S_{23}n_{31}n_{12} + S_{31}n_{12}n_{23} \\ &\quad + n_{12}n_{23}n_{31}. \end{aligned}$$

Clearly the mean value of this quantity is the object triple product $S_{12}S_{23}S_{31}$ and the rest of the terms in the expression are zero-mean noise terms. Using the formalism presented in equations 2.4–2.6 we can then show that the variance of the triple product is the same in all directions in the complex plane and given by

$$\begin{aligned}\mathrm{var}[T_{123},\theta] &= |S_{12}S_{23}S_{31}|^2(\gamma_{12}^{-2}+\gamma_{23}^{-2}+\gamma_{31}^{-2} \\ &\quad +\gamma_{12}^{-2}\gamma_{23}^{-2}+\gamma_{23}^{-2}\gamma_{31}^{-2}+\gamma_{31}^{-2}\gamma_{12}^{-2} \\ &\quad +\gamma_{12}^{-2}\gamma_{23}^{-2}\gamma_{31}^{-2}),\end{aligned}$$

where γ_{ij} is the signal-to-noise ratio on a baseline

$$\gamma_{ij} = \left(\frac{S_{ij}S_{ij}^*}{\langle n_{ij}n_{ij}^*\rangle}\right)^{1/2}.$$

This illustrates the result mentioned in previous sections that the variance of the triple product scales as the sum of the variances on the individual baselines when the signal-to-noise ratios are high and scales as their product when the signal-to-noise ratios are low.

The covariance between the triple products on different closure phase triangles sharing a common baseline can also be calculated. In doing this we must remember that the triple product is a complex quantity and so what we want to calculate is the covariance of the errors on the triple products resolved along lines in the complex plane which are perpendicular to their respective mean vectors (this will determine the covariance of the *phase* errors when we have averaged enough samples such that the resulting phase errors are small). Once again this can be done in terms of two covariances:

$$\mathrm{covar}_1[A,B] \equiv E(AB^*) - E(A)E(B^*) \tag{2.17}$$

$$\mathrm{covar}_2[A,B] \equiv E(AB) - E(A)E(B) \tag{2.18}$$

$$\begin{aligned}\mathrm{covar}[A,B,\theta_A,\theta_B] &= \frac{1}{2}Re\{\mathrm{covar}_1[A,B]e^{-i(\theta_A-\theta_B)} \\ &\quad +\mathrm{covar}_2[A,B]e^{-i(\theta_A+\theta_B)}\},\end{aligned} \tag{2.19}$$

where $\mathrm{covar}[A,B,\theta_A,\theta_B]$ is the covariance of two complex variables A and B along lines at angles of θ_A and θ_B to the real axis. With this formalism we can calculate the covariance of two triple products at right angles to the mean vectors

$$\mathrm{covar}[T_{123},T_{234},\phi_{123}+\pi/2,\phi_{234}+\pi/2] = |S_{12}S_{23}^2S_{31}S_{34}S_{42}|\gamma_{23}^{-2}.$$

We see from this that the covariance is proportional to the the variance of the noise on the shared baseline. We have already seen, however, that at low signal-to-noise

ratios the variance of the triple product increases as the *product* of the individual noise variances and so in this region the correlation between the errors on different closure phases becomes small. To put this quantitatively we can calculate the correlation coefficient

$$\begin{aligned}\mu_{123,234} &\equiv \frac{\text{covar}[T_{123}, T_{234}, \phi_{123} + \pi/2, \phi_{234} + \pi/2]}{(\text{var}[T_{123}, \phi_{123} + \pi/2]\text{var}[T_{234}, \phi_{234} + \pi/2])^{1/2}} \\ &= \frac{1}{3 + 3\gamma^{-2} + \gamma^{-4}}\end{aligned}$$

where it has been assumed for simplicity that all the signal-to-noise ratios are the same i.e. $\gamma_{ij} = \gamma$. Hence

$$\mu_{123,234} \simeq \begin{cases} 1/3 & \text{at high signal-to-noise ratios} \\ \gamma^4 \simeq 0 & \text{at low signal-to-noise ratios} \end{cases}$$

The above results are recalculated in appendix A for the case where the baseline noise is photon noise and where atmospheric noise is included, with very similar results.

Thus in the presence of small amounts of noise, the noise on the closure phases is strongly correlated so that the full set of closure phases contains no new information compared to a linearly independent subset (this can be rigorously proved by considering the rank of the covariance matrix). For low signal-to-noise ratio data, however, the noise on the closure phases becomes decorrelated and so computing the full set of closure phases will mean that we increase the number of independent constraints on the object closure phase. It is important to note that this result holds only if we average many samples of data; it is easy to show that for any one exposure, the closure phase errors are not independent. The resolution of this dilemma is that a lack of correlation only implies *independence* when the variables involved have Gaussian distributions, and this occurs only in the limit where we have averaged many triple products.

2.6 Conclusions

[1]There is a qualitative loss in information when we are forced to incoherently average visibility data at low signal-to-noise ratios. This sets a strong limit to the lowest light levels that an optical interferometer can work at.

[2]At low signal-to-noise ratios, it is worth our while averaging all the possible closure phases in order to better constrain our final map. However, most currently extant hybrid mapping methods, intended for radio VLBI, implicitly assume that the signal-to-noise ratios of the measurements are high and hence do not make use of the 'extra' closure phases. Optical aperture synthesis mapping programs will therefore

have to be written which fit the model map directly to the measured triple products, taking into account their covariances.

Chapter 3

Optimising the aperture size and the integration time

We have seen in the previous chapter how the short coherent integration times and small aperture sizes imposed by the small scales of atmospheric perturbations to the optical phase mean that the amplitude and closure phase measurements are seriously affected by photon noise. If we try to gather more light by increasing the aperture size or exposure time, there will be fluctuations of many radians in the optical phase across the apertures and during the exposure, with two detrimental consequences: firstly, the measured amplitudes and closure phases will fluctuate because there is no longer a single phase error associated with each aperture, that is there will be an increase in 'atmospheric noise', and secondly the mean fringe visibility will be decreased, affecting the signal-to-noise-ratio of photon-noise-limited measurements.

Clearly at there is an optimum trade-off between collecting more light and these atmospheric effects: for moderately low light levels, we can increase the aperture size etc. until the atmospheric noise becomes larger than the photon noise; for very low light levels, where the increased atmospheric noise is negligible compared to the photon noise, it will become unprofitable to increase the aperture size and integration time when the loss in fringe contrast reduces the signal-to-noise-ratio more than the gain in number of photons increases it. In this chapter, we compute these optimum points in relationship to the scales sizes of the seeing, r_o and t_o.

3.1 Atmospheric Model

We shall use here the standard model of atmospheric phase perturbations [31, 75]. In this model, phase perturbations are induced across a wavefront incident on the atmosphere by fluctuations in the refractive index of the air. This in turn arises because of the mixing of layers of warmer and colder air by turbulence in the lower atmosphere.

Because the detailed motion of a turbulent fluid is inherently unpredictable, it can best be described statistically. In the following discussion it will be assumed that the variables describing the turbulence are second-order processes, that is that it is sufficient to know their means, variances and covariances in order to fully describe their statistics. This is equivalent to the assumption that all the distributions concerned are Gaussian. Furthermore it will be assumed that the statistics of the turbulence are isotropic and homogenous i.e. invariant under a shift of origin or rotation of axes, and that the turbulent process is pseudo-stationary i.e. that the overall parameters of the turbulence change only slowly with time.

Under these assumptions it is convenient to talk about the *structure function* of a random variable x

$$D_x(t) \equiv \langle [x(t') - x(t'+t)]^2 \rangle.$$

The structure function is thus the mean square difference in the value of x at two points as a function of their separation in space or time.

The model of the turbulent motions that cause this mixing is due to Kolmogorov [51]: in this model there exists some unspecified mechanism of turbulent energy input into a fluid (the atmosphere) which occurs on some large scale L_o. The large scale motions are unstable and break up into smaller and smaller eddies. At some small scale l_o, however, viscous effects become important and motions on this scale are rapidly dissipated. It is assumed that between these scales is a so-called 'inertial range' where viscous dissipation is unimportant and the energy in the turbulence on one scale is continually being fed to turbulence on smaller scales, such that in equilibrium there is a constant cascade of turbulent energy from the 'outer scale' L_o to the 'inner scale' l_o.

For astronomical viewing conditions l_o is a few millimetres in the lower atmosphere. Little is known about the outer scale, which depends on the particular meteorological conditions which are 'stirring' the turbulence; experimentally estimated values have ranged between 5 metres [19] and a few kilometres [15]. Thus for describing the fluctuations across individual array elements this model is entirely adequate, but its application across the distances between the outermost telescopes in a large array is less defensible.

For this region, $l_o \ll r \ll L_o$, Tatarski [92] shows that the structure function of the refractive index fluctuations is

$$\begin{aligned} D_n(\mathbf{r}) &\equiv \langle |n(\mathbf{r}') - n(\mathbf{r}'+\mathbf{r})|^2 \rangle \\ &= {C_n}^2 |\mathbf{r}|^{2/3}, \end{aligned}$$

where $n(\mathbf{r})$ is the value of the refractive index of the air at a point $\mathbf{r}$ and ${C_n}^2$ is a measure of the strength of the refractive index fluctuations, which is a function of the

strength of the turbulent motions and the temperature difference between the layers of air being mixed.

An initially plane light wave propagating through these refractive index inhomogenaities will receive perturbations to both amplitude and phase. We can write the complex electric field amplitude of the radiation on a plane a distance z into the atmosphere as

$$E(\mathbf{r}) = E_o(\mathbf{r})e^{\psi(\mathbf{r})}$$

where $E_o(\mathbf{r})$ is the electric field that would have existed had the atmosphere not been present and $\psi(\mathbf{r})$ is a complex quantity

$$\psi(\mathbf{r}) = l(\mathbf{r}) + i\phi(\mathbf{r}).$$

Thus $l(\mathbf{r})$ represents perturbations to the log amplitude of the wave (i.e. 'scintillation') and $\phi(\mathbf{r})$ represents phase perturbations. Theoretical arguments and experimental evidence suggest that l and ϕ have Gaussian frequency distributions and Tatarski shows that they have structure functions of the form

$$D_l(\mathbf{r}),\ D_\phi(\mathbf{r}) \propto \int_0^z {C_n}^2(z)\, dz\, |\mathbf{r}|^{5/3}$$

where in all cases it is assumed that $l_0 \ll |r| \ll L_o$ and that C_n does not vary significantly over distances smaller than L_o.

All the calculations in this chapter of the effect of the atmosphere on imaging systems can be shown to depend only on $D_l(\mathbf{r})$ and $D_\phi(\mathbf{r})$. However, under astronomical seeing conditions the contributions of the fluctuations of the log amplitude to the degradation of interferometric measurements will be small compared to the effects of phase fluctuations [75] and so for the rest of this chapter we shall neglect scintillation effects i.e. we shall make the 'near field' approximation

$$D_l(\mathbf{r}) \simeq 0.$$

We shall further assume that the observed object is smaller in angular size than the *isoplanatic patch*, that is, that the light rays arriving at a point on the Earth's surface from different parts of the object will have been perturbed by the same amount. For typical astronomical observing conditions this patch is a few arcseconds in size [64, 82] and so this assumption holds true for most objects of interest.

It is conventional [31] to define the scale parameter r_o, the 'seeing cell' size, such that

$$D_\phi(\mathbf{r}) = 6.88\,(|\mathbf{r}|/r_o)^{5/3} \tag{3.1}$$

at some wavelength, for instance under 1 arcsecond seeing conditions r_o is approximately 10cm. This parameter is then sufficient to describe the seeing conditions

at that wavelength. Furthermore, if the refractive index fluctuations are weak so that dispersion effects can be neglected, then r_o scales with wavelength as $\lambda^{6/5}$ so that knowledge of r_o at one wavelength is sufficient to characterize the seeing for all optical wavelengths.

The most popular model for the temporal evolution of the wavefront phase perturbations is Taylor's 'frozen turbulence' hypothesis, in which the temporal variations are seen as being due to the bulk motion of the turbulent cells across the telescopes. This is equivalent to the statement that the time taken for the wind blow a turbulent cell past a fixed point is smaller than the timescale for the evolution of that cell. Although this clearly cannot apply to 'dome seeing', where there is no local wind and so convective evolution of the refractive index inhomogenaities inside the telescope dome will play the dominant role, there is good evidence for this hypothesis for layers of turbulence at heights of 2km and above [12]. What is also clear is that the temporal evolution consists of the contributions from many different layers of turbulence moving at different velocities [94, 12]. If there were only one layer of turbulence dominating the phase perturbations, then we might expect that the lifetime of a particular interference pattern would be approximately given by the time taken for a point in the turbulence to be blown across the telescope, but large telescope speckle interferometry has shown that this lifetime is actually much smaller than this [86], and corresponds more closely to the time taken by a point in one layer of turbulence to cross an r_o-sized patch in another layer of turbulence.

Given this model, the temporal structure function of the phase variations at a given point in space will be given by

$$D_\phi(t) \equiv \langle |\phi(t'+t) - \phi(t')|^2 \rangle = (t/t_o)^{5/3}\,, \tag{3.2}$$

the equation serving to define a coherence time t_o which is typically of order 10 milliseconds in size. Note that the phase *difference* between two well separated points will have a structure function twice as large.

This model agrees with experimentally determined phase structure functions for spatial scales up to 2 metres [9, 77] and for timescales from 10 milliseconds to more than 10 seconds [14].

In this chapter we shall consider separately the effects of temporal and spatial phase perturbations. We shall not make any assumptions about the cross-spectrum between the temporal and spatial fluctuations; the study of the effects of such cross-correlations will have to wait until there is firmer experimental evidence as to the form of these correlations.

3.2 The effect of temporal fluctuations

In this section we shall consider an interferometer consisting of several point-like apertures spread over large distances so that the temporal variations of the phase are effectively uncorrelated between apertures. We shall further assume that the optical bandwidth is small so that temporal coherence effects are negligible.

Under these assumptions the complex visibility of a given fringe after an exposure time T will be

$$V(T) = \frac{V(0)}{T} \int_t^{t+T} \exp\{i[\phi_1(t') - \phi_2(t')]\}\, dt' \,, \tag{3.3}$$

where $V(0)$ is the visibility that would be measured with a very short exposure and $\phi_1(t)$ and $\phi_2(t)$ are the phase perturbations over the two array elements whose beams are being interfered.

3.2.1 Amplitudes

It is easy to see from the above equation that the modulus of the visibility after a finite integration time will always be smaller than that of the instantaneous visibility (qualitatively we can say that the fringes become blurred because the fringe has moved during the exposure). Clearly we must compensate the measured visibilities for this loss, but, because the actual phase fluctuations over an integration cannot be measured (because otherwise we would have been able to take a shorter exposure), we must treat the problem statistically.

A simple approximation to the real phase fluctuations is that the phase remains roughly constant over one coherence time, but changes randomly between successive periods of length t_o, such that the phase is uncorrelated between successive periods. Equation 3.3 simplifies to

$$V(nt_o) = \frac{V(0)}{n} \sum_{k=1}^{n} \exp\{i[\phi_{1_k} - \phi_{2_k}]\} \,, \tag{3.4}$$

where ϕ_{i_k} are the phase errors on telescope i in time interval k. The summation can be seen as a random walk in the complex plane, and therefore the r.m.s. visibility modulus after a coherent integration time nt_o is $|V(0)|/\sqrt{n}$.

A more rigorous result can be derived from the work of O'Donnell & Dainty [67] who show that the mean square visibility will vary as

$$\langle |V(T)|^2 \rangle = \frac{2|V(0)|^2}{T} \int_0^T \left(1 - \frac{t}{T}\right) C(t)\, dt \,, \tag{3.5}$$

where $C(t)$ is the correlation function of the fringes

$$C(t) \equiv \frac{\langle V(0,t') V^*(0,t'+t) \rangle}{\langle |V(0)|^2 \rangle} \,,$$

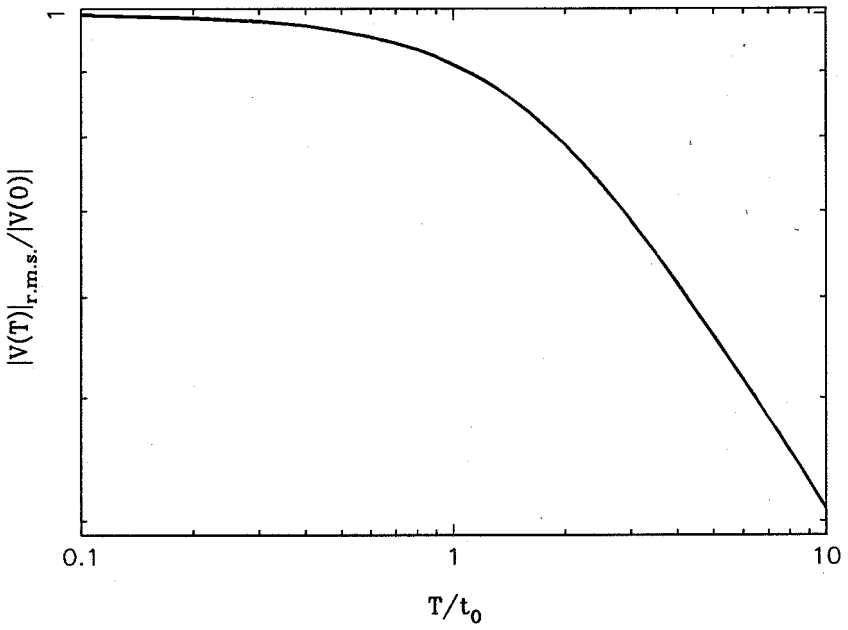

Figure 3.1: *R.m.s. visibility as a function of exposure time.*

where $V(0,t)$ is the instantaneous visibility of the given fringe at time t. For the Tatarski spectrum of fluctuations we can show that

$$C(t) = \exp\left[-(t/t_o)^{5/3}\right] ,$$

if we assume that the phase fluctuations are uncorrelated between different array elements. Hence for long exposure times ($T \gg t_o$),

$$\sqrt{\langle |V(T)|^2 \rangle} \simeq \sqrt{\frac{1.79 t_o}{T}} |V(0)|.$$

Thus the long-exposure limit is of the same form as given by the simpler analysis (with the modification that the effective coherence time is $1.79 t_o$). For shorter exposure times, equation 3.5 can be integrated numerically to give a graph of visibility against exposure time which is shown in figure 3.1.

Clearly, if we need to use long exposure times we must take many exposures and then compensate the sample r.m.s. visibility using the value given in figure 3.1. However, there are complications: t_o is variable, changing according to meteorological conditions, and may not be measurable directly; in addition there may be instrumental phase errors which further reduce the visibility. Therefore it is better to calibrate the measured visibility by comparing it with the visibility measured on a nearby unresolved reference star. Measurements of source and reference should preferably be made in quick succession and the source and reference should be close in the sky

because seeing conditions can change with time and zenith angle. In order that any residual differences in t_o have the least effect on our calibration it is desirable that the change in r.m.s visibility for a given change in t_o is small. This is clearly a function of the gradient of the curve in figure 3.1 and hence it is desirable to work with short integration times from this point of view. Beyond an exposure time of about $2t_o$ however, a 2% change in t_o will always give about a 1% change in the r.m.s. visibility.

In addition to biasing the amplitude, the atmospheric phase fluctuations will also add noise to the measurement of the amplitude. We can employ the methods used in deriving the expression for the mean squared amplitude to derive a closed-form expression for the variance of the (squared) amplitude as a function of exposure time, but this involves the evaluation of a four-dimensional integral, and for higher-order parameters that will be needed later on, such as the second moments of the triple product, the dimensionality of the required integral will be even larger. A better way of evaluating these quantities is to construct a numerical simulation of the Tatarski model phase fluctuations: having done this, we can then set up a computer simulation of an interferometric array and derive the means and variances of all the required parameters from the sample means and variances of the simulated measurements. This method will converge more rapidly than numerical evaluation of the closed-form integrals for the region of low to moderate atmospheric noise, which is the main region of interest. Another advantage of developing a numerical simulation is that it can be used as a general-purpose tool for studying more complex problems such as real-time phase-tracking.

To generate a model realisation of the phase variations we simply have to generate discrete 'white noise' i.e. a sequence of independent pseudorandom numbers, 'filter' the noise to give the correct power spectrum (in fact it is computationally simplest to generate the white noise in frequency space, since the Fourier transform of a white noise process is also white noise), and then use the resulting sequence of numbers to represent the phase perturbation above a telescope for a set of equally-spaced time intervals. Care must be taken in setting the limits of this discrete approximation to a continuous process: in particular, the power spectrum of this process has a pole at the origin and so we must be careful about the long-wavelength cut-off in the simulation. Luckily the quantities of interest in the simulation depend only on the phase structure function, which is relatively insensitive to the long-wavelength cut-off, but the cut-off must still be at wavelengths much larger than t_o (see appendix D).

Figure 3.2 shows the result derived from such simulations for the signal-to-noise ratio of atmospheric-noise-limited amplitude measurements. It confirms the result that would be expected from the simple model of the phase fluctuations introduced at the beginning of this section i.e. that the signal-to-noise ratio tends to unity for long exposure times, and allows us to decide the optimum trade-off between photon

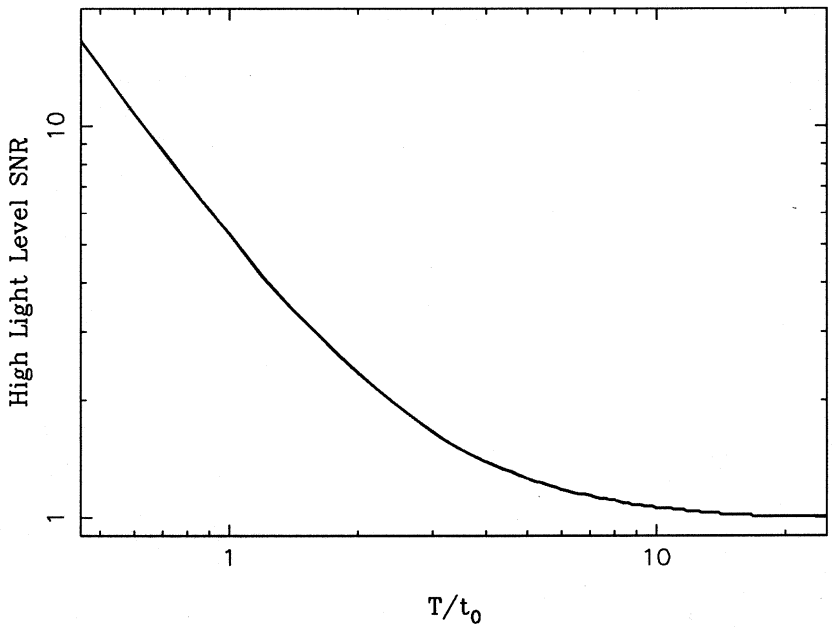

Figure 3.2: *The signal-to-noise ratio of high light level (i.e. atmospheric noise limited) measurements of fringe amplitude as a function of exposure time. The signal-to-noise ratio is defined as the the mean squared visibility modulus divided by the standard deviation of the squared modulus.*

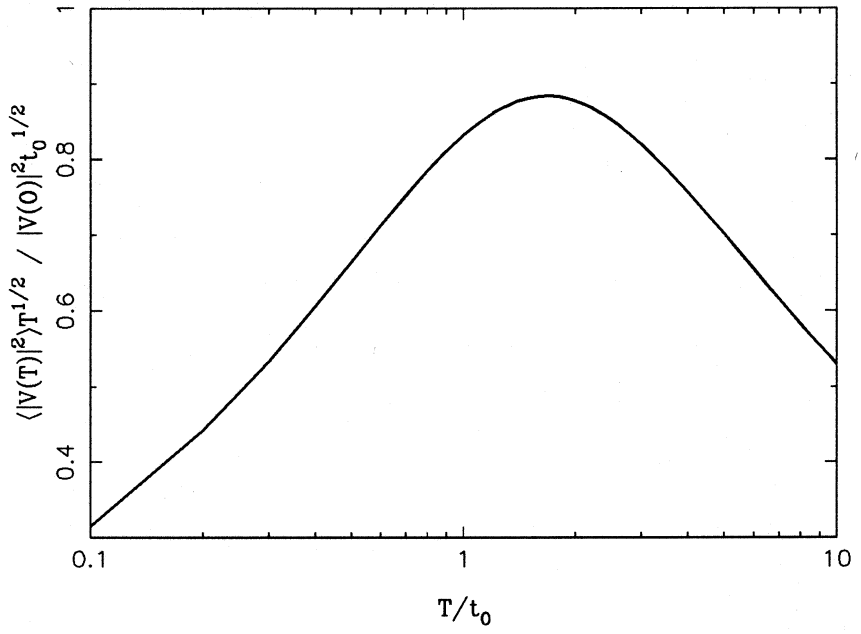

Figure 3.3: *The signal-to-noise ratio of photon-noise-limited amplitude measurements as a function of exposure time.*

and atmospheric noise when working at high light levels. We shall not give detailed recipes for such a trade-off here as the answers will depend on the light level. A good rule of thumb, however, would obviously be to increase the exposure time until the atmospheric noise approximately equals the photon noise.

For low light levels we expect the integration time to be optimum at the point where losses in fringe visibility offset any gain in total light collected. This question is discussed by O'Donnell & Dainty [67] but the analysis is repeated here to derive results for the Tatarski spectrum of fluctuations and as an illustration of the analysis used in later sections.

The amplitude signal-to-noise ratio given in equation 2.9 simplifies for low light levels (such that $\langle|V|^2\rangle\overline{N} \ll 1$) to

$$SNR_A(T) \simeq \langle|V(T)|^2\rangle\overline{N},$$

where $V(T)$ is the visibility of a given fringe after an exposure of length T. Thus, since $\overline{N} \propto T$

$$SNR_A(T) \propto \langle|V(T)|^2\rangle T.$$

However, it is not sufficient to maximise the signal-to-noise ratio for a single exposure because number of exposures that we are able to take in a fixed total observation time decreases as $1/T$. Since the photon noise will be uncorrelated between exposures, the final signal-to-noise ratio after incoherent averaging will therefore be

$$SNR_{final} \propto \langle|V(T)|^2\rangle\sqrt{T}.$$

Figure 3.3 shows a plot of this function against T and we can see from this that the optimum exposure time is approximately $1.6t_o$. Note that the atmospheric signal-to-noise ratio at this integration time is much greater than unity, so that we are fully justified in neglecting atmospheric noise.

3.2.2 Closure phases

We expect that atmospheric phase changes will not bias the closure phase measurements (since there will be no preferred direction towards which the triple product could be biased), but that it will give rise to noise in the high-light-level closure phase. Neglecting photon noise for the moment, the phase error in the triple product will be

$$\beta = \frac{[\mathrm{var}_{yy}(V(\mathbf{u}_{12})V(\mathbf{u}_{23})V(\mathbf{u}_{31}))]^{1/2}}{|\langle V(\mathbf{u}_{12})V(\mathbf{u}_{23})V(\mathbf{u}_{31})\rangle|}.$$

We can get a rough idea of what to expect by using the simple model of the phase fluctuations used to derive equation 3.4. Readhead [72] has analysed the perturbations to the triple product with this model. He shows that the atmospherically perturbed

triple product phase is an unbiased estimate of the closure phase, and we can derive from his results that

$$\beta(nt_o) = \left[\frac{(n-1)(n-2)}{2n}\right]^{1/2}. \tag{3.6}$$

Notice that β is less than 1 for coherent integration times shorter than $5t_o$, indicating the resistance of the triple product to phase perturbations. However we also notice that this model is an unreliable guide for short exposure times as it seems to indicate that for exposures of length $2t_o$ or less, there is no error in the closure phase.

Figure 3.4 shows the more accurate prediction of the triple product phase error as a function of exposure time obtained using simulations of the Tatarski spectrum phase perturbations. At the limit of large exposure times, the curve follows the $\sqrt{T}$ law expected from equation 3.6 but the phase error remains below 1 for exposure times of up to $11t_o$, indicating again an 'effective coherence time' slightly longer than t_o.

For low light level observations, we are again limited by the reduction of the mean fringe visibility with exposure time. Applying the analysis of the previous subsection, the triple product phase error obtained in a fixed total observation time will be

$$\begin{aligned} \beta_{final} &\propto \sqrt{T} \cdot \frac{1}{T^{3/2}|\langle V^{(3)}(T)\rangle|} \\ &= \frac{1}{T|\langle V^{(3)}(T)\rangle|}, \end{aligned}$$

where $V^{(3)}(T)$ is the high light level triple product $\langle V(\mathbf{u}_{12})V(\mathbf{u}_{23})V(\mathbf{u}_{31})\rangle$ after an exposure time T.

Clearly β_{final} must be *minimised* to achieve the highest possible closure phase accuracy. Figure 3.5 is a plot of this expression as a function of T, obtained using the results of simulations. It can be seen from this graph that the optimum exposure time for measuring the closure phase is about $2t_o$, with a sharp increase in the error on either side of this time.

3.3 The effect of spatial fluctuations

In this section we shall consider an interferometer consisting of several widely spaced apertures of size comparable with the seeing cell size, but where exposure time and optical bandwidth effects can be neglected.

We can show (see chapter 6) that all the proposed methods of beam recombination are essentially the same in their sensitivity to atmospheric fluctuations: if we write the distribution of the phase perturbations across the ith aperture as $\phi_i(\mathbf{x})$ where $\mathbf{x}$ is measured relative to the centre of the relevant aperture, then the normalised

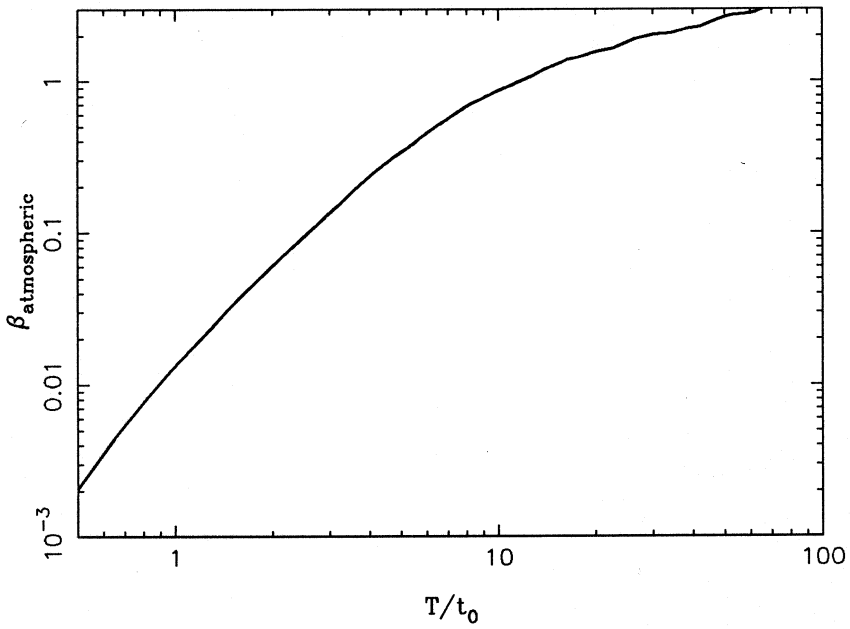

Figure 3.4: *The atmospheric triple product phase error as a function of exposure time.*

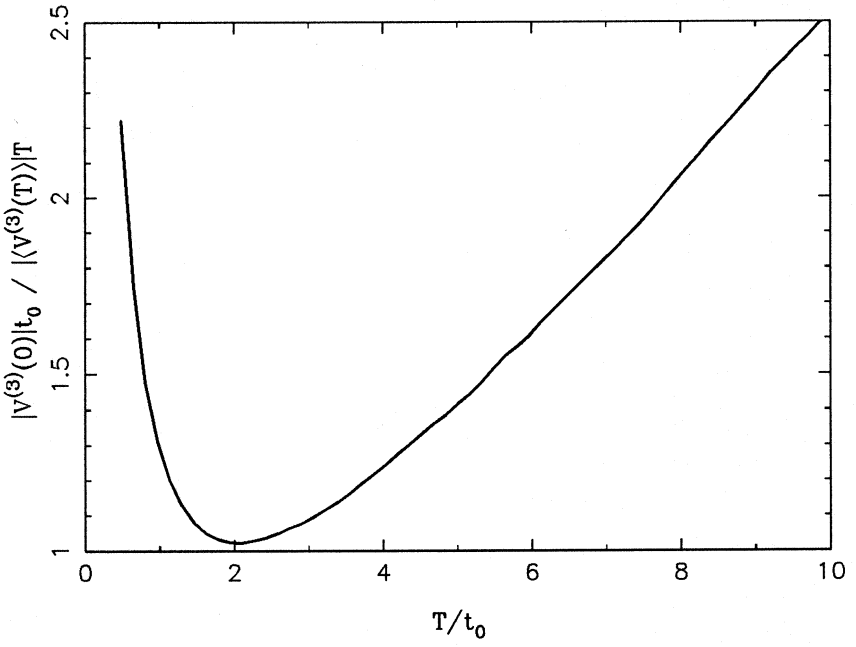

Figure 3.5: *The triple product phase error for photon-noise-limited measurements as a function of exposure time.*

complex amplitude of the fringe corresponding to the interference of apertures 1 and 2 measured by any of the beam recombination methods will be given by

$$V(S) = \frac{V(0)}{S} \int\int_{aperture} \exp\{i[\phi_1(\mathbf{x}) - \phi_2(\mathbf{x})]\}\, dx\, dy, \tag{3.7}$$

where the apertures are assumed to be of the same shape and of area S. Immediately we see that the expression for the perturbed visibility is of the same form as for the case of finite exposure times (see equation 3.3), the only difference being that the spatial case is 2-dimensional; in fact we can apply the results from the previous section directly for the case of slit apertures (with a modification of the constants).

A discrete model of the phase perturbations can again be used to get a qualitative idea of the results we expect: in this case we assume that the apertures can be divided into n sub-apertures or 'seeing cells' of area r_o^2 across which the phase fluctuations are small, but between which the phases are uncorrelated. For this model, equation 3.7 reduces to

$$V(nr_o^2) = \frac{V(0)}{n} \sum_{k=1}^{n} \exp\{i[\phi_{1_k} - \phi_{2_k}]\}\,, \tag{3.8}$$

where ϕ_{i_k} is the phase perturbation to the kth sub-aperture in aperture i.

3.3.1 Amplitudes

Noticing the similarity between equation 3.8 and equation 3.4, we can immediately conclude that

$$\sqrt{\langle |V(nr_o^2)|^2 \rangle} = V(0)/\sqrt{n}$$

for the simple model.

For a more accurate calculation, we can apply the analysis of Korff [52] to show that the mean square visibility of the fringe formed by two well separated apertures with aperture transmission function $P(\mathbf{x})$ ($\mathbf{x}$ measured relative to the centre of the relevant aperture)

$$|P(\mathbf{x})| = \begin{cases} 1 & \text{where the aperture is clear} \\ 0 & \text{where the aperture is opaque} \end{cases}$$

will be

$$\langle |V|^2 \rangle = \frac{|V(0)|^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L(\mathbf{r}) \exp[-D_\phi(r)]\, dr_x\, dr_y}{\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |P(\mathbf{r})|^2\, dr_x\, dr_y\right]^2}\,, \tag{3.9}$$

where L is the overlap integral

$$L(\mathbf{r}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P^*\left(\frac{\rho + \mathbf{r}}{2}\right) P\left(\frac{\rho - \mathbf{r}}{2}\right)\, d\rho_x\, d\rho_y.$$

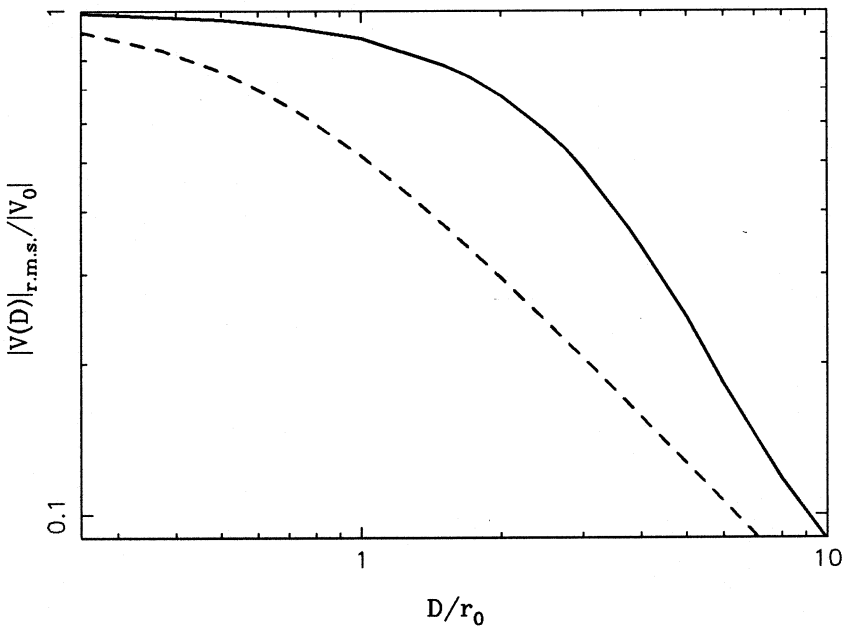

Figure 3.6: *R.m.s. visibility as a function of aperture diameter. The full line is for tilt-corrected apertures and the dotted line is for uncorrected apertures.*

For circular apertures of diameter D,

$$L(\mathbf{r}) = \frac{D^2}{2}\left(\arccos(|\mathbf{r}|/D) - (|\mathbf{r}|/D)[1-(|\mathbf{r}|/D)^2]^{1/2}\right).$$

The integral in equation 3.9 can then be evaluated numerically to give the graph of visibility loss against aperture diameter shown as the dashed line in figure 3.6.

We can see that the loss in visibility for apertures of very moderate size is appreciable; however Fried [31] and others [65, 39] point out that most of the instantaneous phase fluctuations across the aperture can be assigned to a mean tilt across the wavefront i.e. a shift in the centre of the image. If these tilts can be removed by an active optical system in real time, then we can hope to reduce the visibility loss. Clearly the idea of correcting for atmospheric perturbations can be extended to higher-order abberations of the wavefront (defocus, astigmatism etc.), but this requires much more complex active optics, and the additional reduction in the magnitude of the residual perturbations is not as large as that achieved by tilt correction. Such systems are not considered further here.

We cannot easily calculate the visibility loss for a tilt-corrected interferometer using Korff's method because the residual fluctuations are not homogenous, that is to say that the r.m.s. difference in residual phase between two points is a function not only of the distance between the points but also of their mean distance from the centre of the aperture [39]. We can instead use a numerical simulation of the phase

fluctuations to derive the required loss. The method of generating realisations is essentially the same as for the temporal case except for the fact that the fluctuations are 2-dimensional. This latter point means that the problem of achieving an adequate representation of the long-wavelength fluctuations is more severe because it becomes computationally expensive to generate large enough realisations. However for the case of tilt-corrected apertures, the problem eases somewhat because the long-wavelength fluctuations contribute mostly to the wavefront tilts.

The full line in figure 3.6 shows the results from simulations where the tilts across each aperture were removed by subtracting the best fit (in the mean square sense) plane from the phase perturbations. This shows that there is indeed a substantial gain in visibility for a tilt-corrected aperture when the aperture diameter is of order a few r_o, but that for larger apertures this gain is smaller because the the contribution of tilts to the visibility loss becomes less important as compared with the contribution of higher order aberrations of the perturbed wavefront. Note that this curve differs from that presented by Tango and Twiss [85, Figure 4] because they assumed that the residual phase perturbations were homogenous. The results presented here show an increase in the r.m.s. visibility of about 8% for the region $D = r_o$ to $3r_o$ compared with those obtained by Tango and Twiss. However, these results are in any case an upper limit, since errors in the tilt-correction servo will tend to reduce the mean fringe visibility. In particular, as the aperture diameter increases above about $3r_o$, the image formed by the tilt-correction system will begin to break up into 'speckles', greatly complicating the task of determining the tilt (see chapter 5).

In determining the atmospheric noise on the amplitudes and other 'higher order' parameters (e.g. the mean and second moments of the triple product), it is even less practical to evaluate the closed-form solutions than in the case of the temporal fluctuations because the dimensionalities of all the integrals increase by a factor of two. The results presented hereafter were all obtained from simulations, even those for non-tilt-corrected interferometers. The latter were obtained by modifying the spectrum of the generated perturbations so that the structure function of the perturbations for scales less than and approximately equal to r_o were as close as possible to the required value as given in equation 3.1 (See appendix D for details). It is best to regard these results as a rough comparison with the more accurate results for tilt-corrected interferometers.

Figure 3.7 shows the atmospheric-noise-limited signal-to-noise ratio for tilt-corrected and uncorrected apertures as a function of diameter. We can see that, as expected, the atmospheric noise is reduced by tilt correction. Surprisingly, though, the large-aperture signal-to-noise ratio is slightly smaller than unity in both cases. It is not clear how to interpret this in terms of a modification to the simple 'seeing cell' model.

For photon-noise-limited observations the signal-to-noise ratio of the amplitude

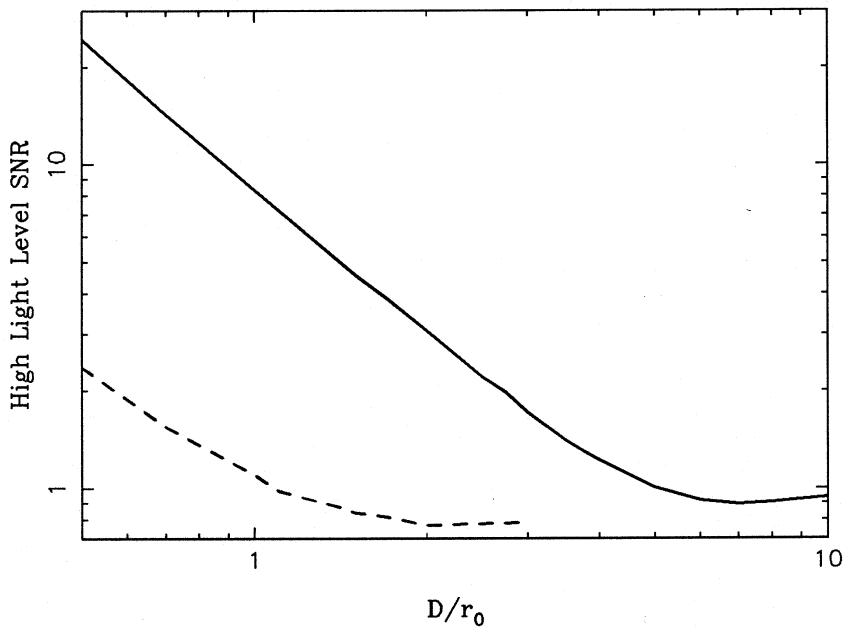

Figure 3.7: *The signal-to-noise ratio of high light level (i.e. atmospheric noise limited) measurements of fringe amplitude as a function of exposure time. The full line is for tilt-corrected apertures and the dotted line is for uncorrected apertures.*

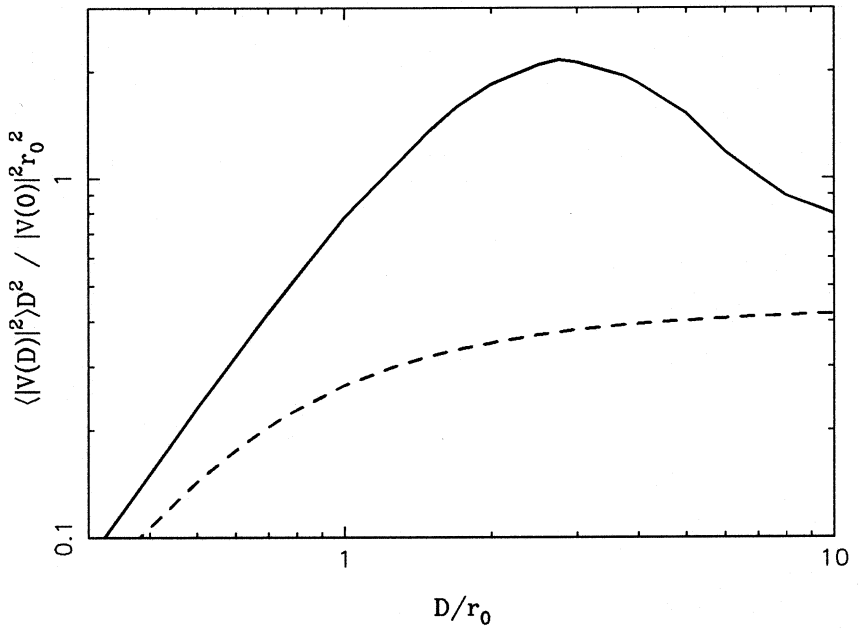

Figure 3.8: *The signal-to-noise ratio of photon-noise-limited amplitude measurements as a function of aperture diameter. The full line is for tilt-corrected apertures and the dotted line is for uncorrected apertures.*

measurements with an aperture of diameter D will be given by

$$SNR_A(D) \propto \langle |V(D)|^2 \rangle D^2.$$

Assuming that there are no other constraints to the aperture size (e.g. financial considerations) then the optimum diameter is that which maximises this quantity. We can see from figure 3.8 that the optimum aperture when no tilt correction is envisaged is in fact infinitely large (we can use Korff's results for large apertures [52, equation 37] to show that $\lim_{D\to\infty}\langle |V(D)|^2 \rangle D^2 = 0.435|V(0)|^2 r_o^2$). In practice there is not much point using an aperture diameter larger than about $3r_o$ which is worse than an infinite aperture by only 15%.

For tilt-corrected interferometers, however, there is a significant advantage (more than a factor of two in signal-to-noise ratio) in using array elements of diameter $3r_o$ rather than much larger ones. We can attribute this to the fact that the contribution of the wavefront tilt to the visibility loss is less significant for large apertures.

3.3.2 Closure phases

Figure 3.9 shows the atmospheric triple product phase error as a function of aperture diameter, obtained using simulations. It shows that the error in the closure phase for one exposure will be less than a radian for aperture diameters less than $2r_o$ in the case of uncorrected apertures and $7r_o$ in the case of tilt-corrected apertures. The latter figure is in sharp contradiction with simple first-order theory which might lead us to imagine that there were more than 30 'seeing cells' inside the the area of an aperture of diameter $7r_o$.

For low-light-level observations we can show that the photon noise phase error will be

$$\beta(D) \propto \frac{1}{D^3 |\langle V^{(3)}(D) \rangle|}$$

Readhead [72] shows that for the 'seeing cell' model the modulus of the mean triple product vector is

$$|\langle V^{(3)}(\text{Aperture Area} = nr_o^2) \rangle| = |V^{(3)}(0)|/n^2. \tag{3.10}$$

We can see from this that we do not expect a large aperture to be optimal for low-light-level closure phase measurements even for uncorrected apertures, since

$$\beta(\text{Aperture Area} = nr_o^2) \propto n^{1/2}$$

and the graphs calculated using simulations (figure 3.10) bear this out: the optimum aperture diameter is about $1.2r_o$ for uncorrected apertures and about $2.8r_o$ for tilt-corrected apertures.

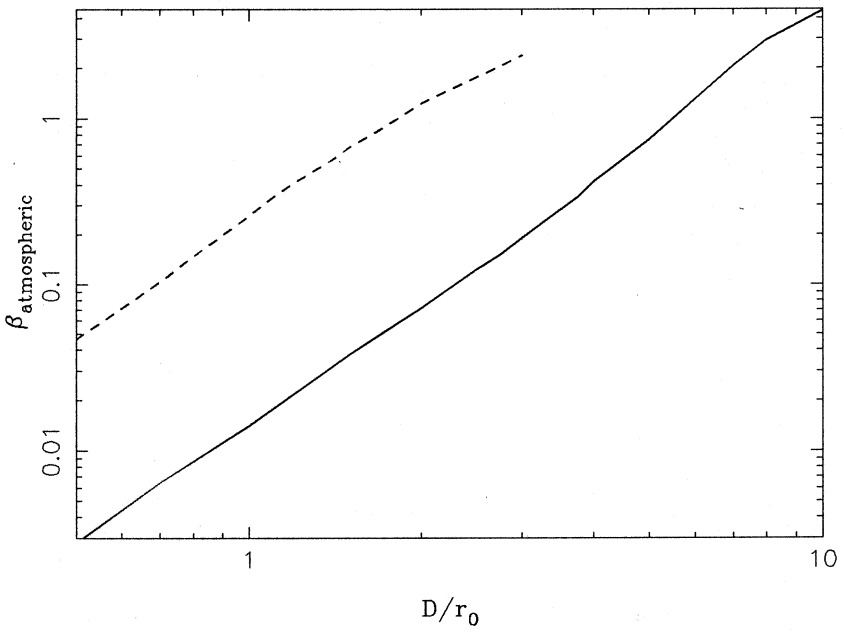

Figure 3.9: *The atmospheric triple product phase error as a function of aperture diameter. The full line is for tilt-corrected apertures and the dotted line is for uncorrected apertures.*

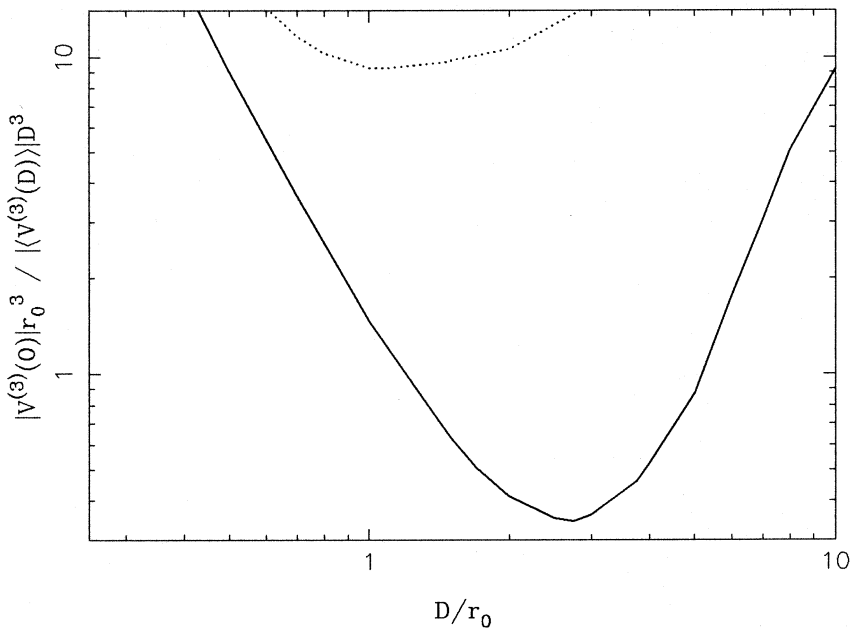

Figure 3.10: *The triple product phase error for photon-noise-limited measurements as a function of aperture diameter. The full line is for tilt-corrected apertures and the dotted line is for uncorrected apertures.*

3.4 Conclusions

The above results show that to minimise photon noise at low light levels one should use tilt-corrected apertures approximately $3r_o$ in diameter, and an exposure time of about $2t_o$. This of course assumes that we know the values of t_o and r_o, but since they both vary (albeit slowly) with time, some thought will have to go into developing methods for determining these parameters in real time and into making the aperture sizes and exposure times adjustable.

Reference to figures 3.2,3.4,3.7 and 3.9 shows that if these aperture sizes and integration times are used then the error on the amplitudes and triple products due to atmospheric noise is small, which means that these parameters are nearly optimal even at higher light levels at which the atmospheric noise must be taken into account.

In practice, however, it is best to use smaller apertures and integration times in order to minimise the effect of changing seeing conditions on the calibration of the amplitudes, as discussed in section 5.1. In a system where photon events are time-tagged the same data can be processed twice, once with short integration times to yield well-calibrated amplitudes and once with integration times of about $2t_o$ in order to derive the closure phases, which do not require calibration. The same cannot be done in the case of aperture size. Although in extreme cases it might be useful to do two consecutive observing runs, one with small apertures for determining amplitudes and one with larger apertures to determine the closure phases (or in fact one could have two instruments with different aperture sizes observing simultaneously), in practice a compromise aperture size would make the most efficient use of instrument time where extreme calibration accuracy is not required. For example, if we are prepared to allow the photon noise on the amplitudes and closure phases to increase by a factor of $\sqrt{2}$ (i.e. we are prepared to average for twice as long), then reference to figures 3.8 and 3.10 shows that the aperture size can be reduced from $3r_o$ to $1.7r_o$. Reference to figure 3.6 then shows that using the smaller apertures reduces the effect of changes in r_o by a substantial amount — a 1% change in r_o will change the r.m.s. visibility by only 0.5% as compared with 1% for the larger aperture.

Chapter 4

Array Configuration

The question of how many elements an aperture synthesis array should have and what configuration they should be in is an old and much-discussed topic in radio interferometry, but although many results applicable to the design of an optical aperture synthesis array exist in the literature of radio astronomy, the optical regime presents certain new features which mean it is worthwhile re-examining this question. The first and most important of these is that, because of the lack of amplifiers in the optical, increasing the number of array elements can actually decrease the signal-to-noise ratio of the measurements on a given baseline. Secondly, we are not working with a phase-coherent interferometer, which is the main subject of the design studies in radio interferometry. This means that we are working with closure phases, so that the number of constraints on the image we are measuring scales differently with the number of telescopes compared with connected-element interferometry. Furthermore, we are often working with large numbers of low signal-to-noise ratio exposures, as opposed to the radio VLBI case where the signal-to-noise ratio per sample is usually high and only few samples are taken at each point in the Fourier plane, which means that there are more independent closure phase measurements than would be encountered in radio VLBI (see section 2.5.4).

4.1 Design Criteria

In the following discussion, we shall ignore the constraints imposed by the cost and complexity inherent in large arrays. This may seem to invalidate the whole exercise since financial constraints are usually the most important considerations in array design, but they are particularly hard to model (how much money is a better signal-to-noise ratio worth?). Instead we shall try to determine how the performance of the array changes with increasing numbers of array elements in an idealised model. These results can then be used to determine the trade-off between the number of array

elements and other areas of the design, for instance whether better performance could be achieved at the same cost by improving the quality of the optics or the detectors.

We shall take as our criterion of performance the time taken by the array to produce a map of a given quality, 'quality' being defined as the dynamic range of the map, i.e. the ratio between the flux of the brightest source and the lowest believable feature in the reconstructed image. The experience in radio interferometry suggests that this is a function of the completeness of the coverage of the Fourier plane [1] (hereafter called the 'u-v plane') and of the number and signal-to-noise ratios of the constraints on the measured Fourier components (i.e. the amplitudes and closure phases).

The first criterion means that we must arrange to have a reasonably even coverage of the u-v plane. In radio interferometry, earth rotation is normally used to give 2-dimensional coverage from fixed linear arrays of telescopes, but this is less practical in the optical because we cannot use a full 12-hour integration to allow the baselines to sweep out all positions angles on the sky. In practice we are limited to observing objects less than about 45 degrees from the zenith so as to minimise the air mass we are looking through, and this means that for a typical object the projected baseline will sweep out on the sky about one half of the circumference of an ellipse. The array must therefore have a 2-dimensional arrangement, and that adopted for COAST is a 'Y' configuration (see figure 4.1). With a symmetrical arrangement of 4 telescopes we get a coverage of the u-v plane as shown in figure 4.2. Clearly, however, there are large gaps in the coverage and this must be remedied by moving the telescopes to different stations and making further observations. This is time-consuming and it would be much easier from this point of view to have a larger number of telescopes — figure 4.3 shows the coverage obtainable from an array which has all thirteen proposed stations occupied by telescopes.

However, increasing the number of telescopes may conflict with the second requirement, i.e. of high-precision amplitude and closure phase measurements. If we combine the beams from M telescopes in a single fringe pattern, the photon rate will increase as M but the visibility of the fringe on a given baseline will fall as $1/M$, so that the signal-to-noise ratio per exposure defined in equation 2.7 falls as

$$\begin{aligned} SNR &= |V|\,\overline{N}^{1/2} \\ &\propto M^{-1/2}. \end{aligned}$$

A glance at equations 2.9 and 2.16 shows that errors on the measurements of amplitude and closure phase will also increase with the number of telescopes and so that there will be a competition between u-v plane coverage and signal-to-noise ratio.

[1] Lannes [56] puts the relationship between the size of the 'holes' in the u-v plane coverage and the errors on the reconstructed map into a quantitative form

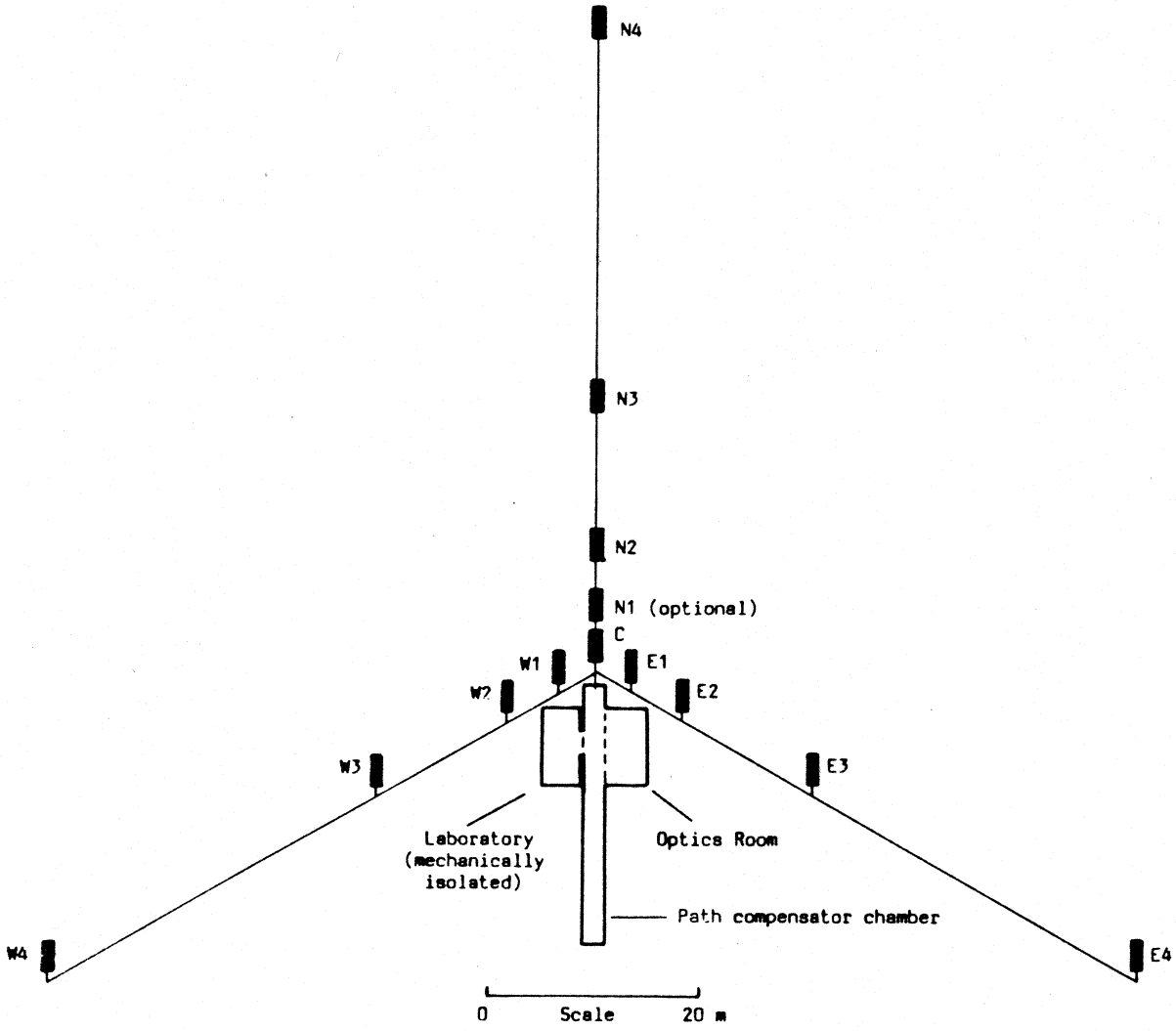

Figure 4.1: *The proposed layout of telescope stations for COAST.*

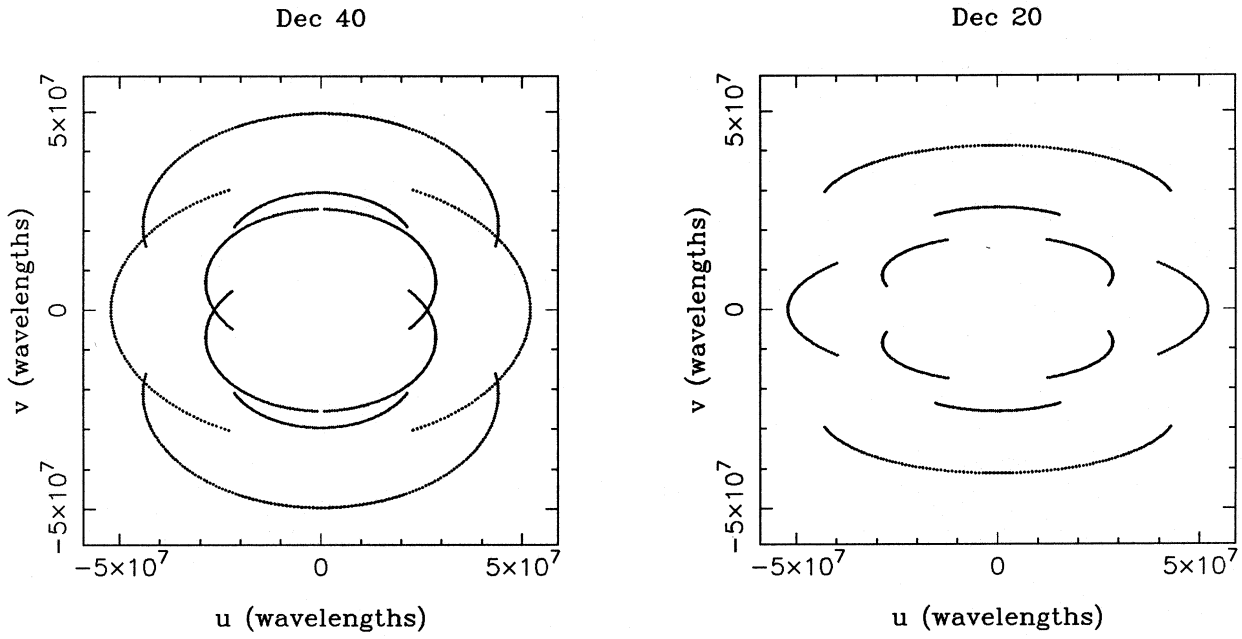

Figure 4.2: *The u-v plane coverage obtained with telescopes occupying stations C, N3, E3 and W3 in figure 4.1. The figures shown the coverage obtained with zenith angle limits of* $\pm 45°$ *and with the source at declinations of (a)* $40°$ *and (b)* $20°$.

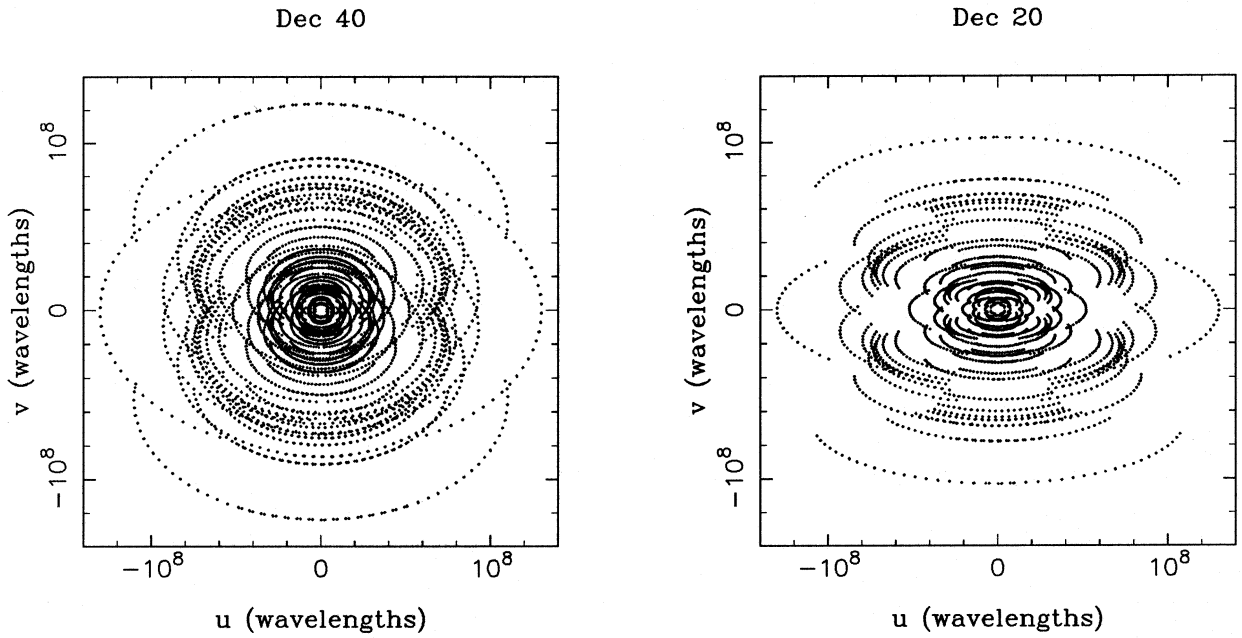

Figure 4.3: *The u-v plane coverage obtainable by making use of all the stations in figure 4.1. The zenith angle limits and the source declinations are the same as for figure 4.2*

To evaluate the optimal trade-off between these two factors we shall assume that we can get acceptable coverage of the u-v plane through the provision of rapidly movable array elements and a large number of possible stations. Thus a smaller array can be compared in performance with a larger array if it can measure the amplitudes and closure phases to an acceptable accuracy more quickly than the larger array, so that there is time to reconfigure the array and cover the u-v plane in the same total observation time as needed by the larger array.

What comprises an acceptable accuracy is slightly harder to define. Ideally we would like some measure of how well a given set of amplitudes and closure phases with a given set of errors constrains the reconstructed image, or at least how the image-plane constraints change when we vary the number of independent measurements and their errors. In phase-coherent interferometry, where the transformation between the measured data and the final map is linear, it is relatively easy to show that the r.m.s. noise per resolution element in the map is of order $\sigma/\sqrt{n}$ where σ is the noise on an individual visibility measurement and n is the number of independent visibility measurements. In hybrid mapping, where the transformation between map and measurements is non-linear, the noise behaviour is less well understood. Cornwell [17] states that for self-calibration methods with high signal-to-noise ratio data, the noise on the map scales roughly as in the phase-coherent case, with a reduction factor due to the lost antenna phase information. When the signal-to-noise ratios of the amplitude and closure phase measurements are small, however, self-calibration methods fail to produce maps which resemble the source, even when there are measurements at a large number of different u-v points [38]. This may be due to the implicit assumption in most hybrid-mapping programs that the phase errors on the closure phases are small, and an improvement may occur if direct fitting to the triple product is used [18], but little has yet been demonstrated in this direction.

To simplify the analysis we shall ignore the above objections for the time being and assume that a large amount of low signal-to-noise ratio data constrains the map just as well as a small amount of high signal-to-noise ratio data. We then have the further problem of deciding the relative weights of amplitude and closure phase data, because at low signal-to-noise ratios there is no longer a fixed ratio between the noise on the amplitudes and those on the closure phases. Unfortunately this is not a well-posed problem because the amplitudes and closure phases constrain different properties of the image and therefore the answer is highly image-dependent. For instance there will be little information in the closure phases if the object is symmetric, and equally well amplitude information is of no use in determining which way round an asymmetric source is. Therefore we shall consider two separate figures-of-merit for an array: one for amplitudes and one for closure phases. For an array which measures n_a amplitudes with signal-to-noise ratios after averaging for unit time

of $\{\gamma_k,\, k = 1 \ldots n_a\}$, our 'amplitude observing efficiency' will be defined as

$$\eta_a = \sum_{k=1}^{n_a} \gamma_k^2 \tag{4.1}$$

and if it measures n_c independent closure phases with phase errors $\{\beta_k,\, k = 1 \ldots n_c\}$, then our 'closure phase observing efficiency' will be defined as

$$\eta_c = \sum_{k=1}^{n_c} \beta_k^{-2}. \tag{4.2}$$

Clearly, these figures-of-merit are ad-hoc, but we can justify their use by noting that in the case where all the errors on the amplitudes are the same then the time taken to observe a given number of amplitudes to a given accuracy will be inversely proportional to the amplitude observing efficiency (assuming that it takes a negligible amount of time to reconfigure the array); a similar statement can also be made about the closure phase observing efficiency.

In the following discussion we shall ignore any dependence of the observing efficiency on the shape of the object by assuming it to be unresolved and we shall assume an idealised instrument with point-like apertures and short integration times.

4.2 Non-Redundant Configurations

If the M telescopes are arranged non-redundantly i.e. all of the baselines sampled at any one time are different, and all the M beams are interfered in a single fringe pattern, the visibility of all the fringes will be M^{-1} and the total number of photons in each exposure will be $M\overline{N}_0$ where $\overline{N}_0$ is the number of photons entering each aperture. Referring to equation 2.9 we can see that at high light levels, the signal-to-noise ratio of the squared amplitudes will be $(\overline{N}_0/2M)^{1/2}$. However, the number of amplitudes measured simultaneously will rise as $M(M-1)/2$ and so the amplitude observing efficiency will be

$$\eta_a = \frac{\overline{N}_0 M(M-1)}{4M}.$$

Thus the observing efficiency will rise linearly with the number of telescopes when the number of telescopes is large. There will come a point, though, where the loss of signal-to-noise ratio due to the large number of telescopes will mean that we are no longer in the high signal-to-noise ratio region and this formula will no longer apply.

Similarly, for the closure phases at high light levels, equation 2.16 shows that the phase error will increase as $(3M/\overline{N}_0)^{1/2}$ but the number of independent closure phases will be $(M-1)(M-2)/2$ so that the closure phase observing efficiency will be

$$\eta_c = \frac{\overline{N}_0 (M-1)(M-2)}{6M},$$

again showing a linear rise in observing efficiency with M for large M.

For faint objects the signal-to-noise ratio of the amplitude measurements will fall off as $\overline{N}_0/M$ (see equation 2.9). Thus the observing efficiency will be

$$\eta_a = \frac{\overline{N}_0^2 M(M-1)}{2M^2}.$$

which means that the observing efficiency will saturate at large M. For the closure phase measurements, the falloff in accuracy will be even more rapid, equation 2.16 showing that the phase error will rise as $(M\overline{N}_0)^{3/2}$. However, as shown in section 2.5.4, the number of independent closure phase measurements in this regime is larger than in the high light level case and will be $M(M-1)(M-2)/6$. This gives an observing efficiency of

$$\eta_c = \frac{\overline{N}_0^3 M(M-1)(M-2)}{6M^3},$$

which once again saturates.

Thus there is little to be gained from combining all the beams from a large number of array elements when observing faint sources (which is precisely when observing efficiency is most crucial). An alternative strategy, however, might be to observe with many different arrays simultaneously. Clearly the observing efficiency of such a method will always be K times better than that of a single array of the same size, where K is the number of separate arrays. The question then arises as to what the most efficient use of M telescopes is — is it better to split them up into a large number of sub-arrays each containing a small number of telescopes or would it be better to use a smaller number of large arrays? If each sub-array has m elements then there will M/m sub-arrays (assuming that M is divisible by m) so that the total observing efficiency will be

$$\eta_a = \frac{M}{m} \cdot \frac{\overline{N}_0^2 m(m-1)}{2m^2}$$

for the amplitudes, and

$$\eta_c = \frac{M}{m} \cdot \frac{\overline{N}_0^3 m(m-1)(m-2)}{6m^3}.$$

Figure 4.4 shows a plot of these efficiencies as a function of m, which show that the optimum number of elements in a sub-array is 2 if we are solely interested in measuring amplitudes and 5 in the case of closure phases.

4.3 Redundant Arrays

In a redundant array, one or more of the baselines measured at any one time is repeated. If the beam recombination method is non-redundant, i.e. the repeated

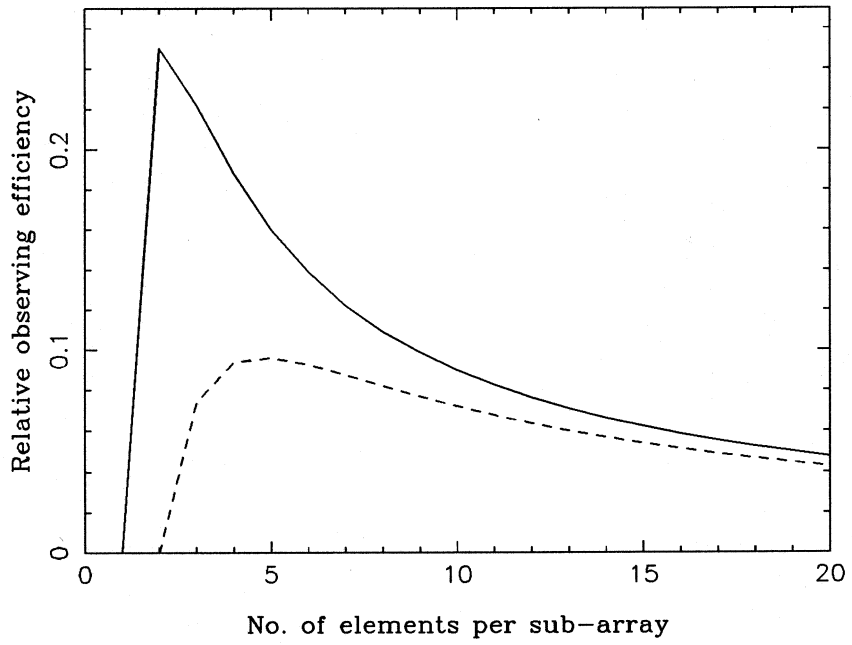

Figure 4.4: *The relative observing efficiencies of a fixed (large) number of telescopes when split up into non-redundant sub-arrays of different sizes. The full line shows the amplitude observing efficiency and the dashed line shows the closure phase observing efficiency.*

baselines appear at different frequencies in the fringe pattern, then we increase the number of independent measurements of the relevant visibilities, but this is exactly compensated for by the reduction in the number of different baselines that would have been measured by a non-redundant array of the same size, and so the results of the previous section will apply. If however the beam recombination scheme is redundant, so that the fringes from identical baselines appear at the same fringe frequency, the signal-to-noise ratio of the fringe measurements will be affected in two ways: at high light levels, the addition of fringes at the same frequency but with different random atmospheric phase shifts will mean that the measured visibility fluctuates between frames, even when the apertures are point-like and the integration times small; at low light levels when this extra atmospheric noise is unimportant, redundancy may be beneficial because the mean signal is larger than in the non-redundant case. The resemblance of this last statement to the considerations relevant when optimising the aperture sizes is not accidental; we can consider a non-redundant array of large apertures as a redundant array of r_o-sized sub-apertures. However this is only one type of redundancy — we shall consider here another form of redundancy in which the apertures are evenly spaced on a regular grid rather than clustered around isolated points in u-v space.

We shall restrict ourselves to considering an array of elements arranged along a line with unit spacing between them. This will hopefully bring an insight into the relative merits of redundant arrays without having to consider every possible form of redundancy. We shall assume that the redundancy of the beam combination method exactly reflects the redundancy of the array, i.e. that every member of a set of repeated baselines contributes to the same fringe frequency component and that this is true for all possible sets. This arrangement has been partially analysed by Hofmann [41], but we extend the analysis here to include a computation of the observing efficiency and a discussion of the measurement of the amplitudes.

In this scheme, the complex visibility of the fringe corresponding to a baseline ij will be made up of the sum of the contributions from all the n_{ij} pairs of telescopes which sample that baseline

$$\hat{V}_{ij} = \sum_{p=1}^{n_{ij}} V_{ij_p} \exp[i(\phi_{i_p} - \phi_{j_p})] \,, \tag{4.3}$$

where the subscript p denotes different pairs of telescopes and where ϕ_{i_p} and ϕ_{j_p} are the atmospheric phase errors associated with telescopes i_p and j_p respectively. It will be assumed that these phase errors vary randomly between $-\pi$ and π and are uncorrelated between telescopes.

It is then easy to show that the mean squared visibility amplitude will be

$$\begin{aligned} \langle|\hat{V}_{ij}|^2\rangle &= \sum_{p=1}^{n_{ij}} \langle|V_{ij_p}|^2\rangle \\ &= n_{ij}/M^2 \end{aligned} \tag{4.4}$$

where we have made use of our assumptions that the object is unresolved and that the apertures are point-like, so that the visibility contribution from each repetition of the baseline is $1/M$. Hence the mean squared visibility increases linearly with the baseline redundancy. For the linear redundant array we are considering, the redundancy of a baseline of length L units will be

$$n_{ij} = (M - L). \tag{4.5}$$

At high light levels, the noise on the amplitude will be dominated by the fluctuations due to the random atmospheric phases in the summation in equation 4.3. We can show that the signal-to-noise ratio of the amplitudes will be

$$\begin{aligned} SNR_a &= \frac{\langle|\hat{V}_{ij}|^2\rangle}{(\mathrm{var}[|\hat{V}_{ij}|^2])^{1/2}} \\ &= \left(\frac{n_{ij}}{n_{ij}-1}\right)^{1/2} \end{aligned}$$

which tends to unity for heavily redundant baselines.

When considering the closure phases, it may at first seem as though no closure phase measurements are possible because there are more than two antenna-based errors per baseline. However, when we make use of the triple product as an estimator, the closure phase is recovered providing we average sufficient exposures so that the atmospheric noise terms become small. The reader is referred to Roddier [76] and Cornwell [18] for a discussion of this. The results of Readhead *et al.* [73] can be used to show that the mean triple product on a triangle consisting of baselines ij, jk, and ki will be

$$\langle\hat{V}_{ij}\hat{V}_{jk}\hat{V}_{ki}\rangle = \sum_{p=1}^{n_{ijk}} \langle(V_{ij}V_{jk}V_{ki})_p\rangle \tag{4.6}$$

where the subscript p is used here to denote different *closed triangles* consisting of the three baselines, and n_{ijk} is the number of such triangles. For the given array,

$$n_{ijk} = 2(M - L) \tag{4.7}$$

where L is the length of the longest baseline in the triangle. The factor of 2 is due to the existence of 'mirror image' triangles in the array, which add constructively to the total triple product.

Readhead *et al.* and Roddier [78] also show how the atmospheric noise increases as a function of the number of non-closing triangles which can be made from the relevant baselines, and Roddier calculates the point of cross-over between atmospheric-noise-dominated and the detector-noise-dominated regimes. However it is clear that for high enough light levels, a non-redundant scheme is to be preferred since the atmospheric noise is then small.

At low light levels, the increase in the mean visibility due to redundancy will increase the signal-to-noise ratio on a baseline when compared with a redundant array with the same number of apertures. However the number of baselines sampled will be less. We shall consider the measurement of amplitudes first; for the redundant array at low light levels, we can combine equations 2.9, 4.4 and 4.5 to show that the signal-to-noise ratio on a baseline of length L will be

$$SNR_a = \overline{N}_0(M-L)/M$$

and the amplitude observing efficiency will therefore be

$$\begin{aligned} \eta_a &= \overline{N}_0^2 \sum_{L=1}^{M-1} (M-L)^2/M^2 \\ &= \overline{N}_0^2 (M-1)(2M-1)/6M. \end{aligned}$$

We immediately notice that the observing efficiency does not saturate when the number of telescopes is large as is the case for a single non-redundant array.

For the closure phases at low light levels, combining equations 2.16, 4.6 and 4.7, we see that the phase error of the closure phase on a triangle whose maximum baseline is L will be

$$\beta = M^{3/2}/2(M-L)\overline{N}_0^{3/2}$$

and there will be $(L-1)$ such triangles. Thus the observing efficiency will be

$$\begin{aligned} \eta_c &= \sum_{L=2}^{M-1} 4\overline{N}_0^3 (L-1)(M-L)^2/M^3 \\ &= \overline{N}_0^3 (M-1)^2(M-2)/3M^2 \end{aligned}$$

which once again rises linearly with M when M is large.

We can also look at the combination of the array elements in sub-arrays, as in the non-redundant case. Here we shall assume that all the sub-arrays consist of m equally-spaced telescopes. For an array of M telescopes arranged in sub-arrays of m telescopes each, the observing efficiencies will be

$$\eta_a = \frac{M}{m} \cdot \frac{\overline{N}_0^2 (m-1)(2m-1)}{6m}$$

and

$$\eta_c = \frac{M}{m} \cdot \frac{\overline{N}_0^3 (m-1)^2 (m-2)}{3m^2}.$$

Examination of these formulae shows that it is best not to subdivide the array in the redundant case, if our sole criterion is observing efficiency, although there will be little loss in efficiency if the sub-arrays are large.

4.4 Discussion

To summarize: for a non-redundant array at high light levels, the observing efficiency rises linearly with the number of telescopes until the point at which the loss of signal-to-noise ratio due to the large number of telescopes means that the signal-to-noise ratio on a baseline falls below unity. At low light levels, the observing efficiency of a non-redundant array saturates with large numbers of telescopes and so it is best to split it up into a number of sub-arrays with about 5 elements in each. Redundant arrays with redundant beam combination are best not used at high light levels because of the atmospheric noise caused by the redundancy, but at low light levels they are preferable because the observing efficiency does not saturate.

Even when compared with non-redundant schemes consisting of small sub-arrays, the observing efficiency of the redundant array is superior: the ratio of the amplitude observing efficiencies of a non-redundant array with 5-element sub-arrays to that of a redundant array with the same total number of elements will be

$$\frac{\eta_a(non-redundant)}{\eta_a(redundant)} \simeq 0.24,$$

where we have assumed that the total number of array elements is large. The ratio of the closure phase observing efficiencies will be

$$\frac{\eta_c(non-redundant)}{\eta_c(redundant)} \simeq 0.048,$$

However, what we are concerned about in the end is map quality rather than raw observing efficiency. Other considerations come into effect if, for instance, our mapping method can only make use of high signal-to-noise ratio data. In this case the observing efficiency may be misleading if the array measures a large number of u-v points but with a relatively low signal-to-noise ratio. It may take longer observe a source with such an array than indicated by the observing efficiency because we need to average sufficient samples to make the final signal-to-noise ratio greater than unity at each u-v point.

If the number of telescopes available is small and the time taken to reconfigure the array is not negligible, the choice of the best observing strategy would be weighted

towards those schemes which maximise the u-v plane coverage that can be achieved without reconfiguring the array. These schemes would inevitably be non-redundant.

If we are intending to follow the white light fringe either actively or passively (see chapter 5) then it is difficult to use redundantly combined beams to do so because the path-length information from several baselines will be 'scrambled'. This could, however, be overcome by having separate beam combination systems for fringe tracking and for object visibility measurements (perhaps using light from a different part of the source spectrum [21]), but at the expense of extra complexity.

One attraction of redundant schemes is the fact that the set of closure phases measured is complete enough to allow us to solve explicitly for the 'antenna phase errors' on the measured baselines (although there will still be a small number of unrestricted degrees of freedom) and hence directly determine the phases of the object coherence function [34]. The conventional wisdom in radio astronomy, however, is that self-calibration with non-redundant arrays is sufficiently robust a technique that a given number of telescopes is probably better employed in increasing the u-v coverage than in increasing the constraints on a restricted set of baselines [69]. In other words, indeterminacy in the problem of calibrating the antenna phases is similar in magnitude to the problem of interpolating and extrapolating unmeasured Fourier components.

Chapter 5

Active control systems at low light levels

It can be seen from chapter 3 that the blurring of the fringe patterns by atmospheric phase perturbations reduces the signal-to-noise ratio of the fringe parameter estimates at both high and low light levels. If these perturbations can be measured and corrected in real time by moving optical elements in the system so that the perturbations are nullified, we can hope not only to increase the signal-to-noise ratio but also to reduce the uncertainties in calibrating the visibility measurements. The effectiveness of such systems is ultimately limited by how well the perturbations can be measured. For perturbations occurring inside the instrument, laser beams can be used to measure the optical path length fluctuations, but to measure the perturbations which have occurred to the light beam on its path from the top of the atmosphere to the telescopes, we must use light from the source itself for our measurements. It would be possible to use the light from another source as our reference if it occurred within the isoplanatic patch of the observed source, but the chances of finding such a source are remote: if for instance we need a reference source with a visual magnitude brighter than 10 and the isoplanatic patch is, say, 2 arcseconds in radius then there will be less than one chance in 10^4 of finding a reference source near enough to a given point in the sky, even at low galactic latitudes [3]. Schemes have been suggested for producing an artificial reference source by scattering a yellow laser beam off sodium atoms in the mesospheric sodium layer about 100 km above the Earth's surface [30], but even in such a case the received flux will be small.

Thus the active optical system will suffer from low signal-to-noise ratio problems similar to those encountered in the measurement of the interferometric signals. In this chapter we shall consider the effect of these limitations on the performance of the system. We shall consider here only the problems of measurement rather than the problems of effecting a correction once the measurement has been made, although an

idea of some of the latter technical problems can be found in chapter 7.

The wavefront perturbations across each aperture can be decomposed into components corresponding to a 'piston' effect i.e. a shift in the mean overall phase of the wavefront, a wavefront tilt corresponding to a shift in the centre of the image formed if the wavefront is brought to a focus, and higher order abberations corresponding to defocus, astigmatism and so forth [65]. We shall restrict ourselves to considering the correction of the lowest order terms i.e. the piston and tilt terms, since these contribute most to the loss of signal and are the simplest to correct.

5.1 Fringe Tracking

The piston component of the phase error is by far the largest, but by adopting the visibility and the closure phase as our interferometric signal estimators, we remove completely the lowest order effects of this term. The remaining effects come about because of the finite integration time and the finite bandwidth used.

The effects of a finite integration time on the fringe measurements have been discussed in chapter 3. Active correction of the piston term on sub-t_o timescales would allow the use of longer integration times without the corresponding loss of visibility and increase in atmospheric noise. A fundamental defect of such an approach, though, is that if we are able to measure the phase well enough to correct it then the signal-to-noise ratio is already high enough to mean that there is little gain in extending the coherent integration time. Having said that, there are advantages to be gained if the measurements for error correction and the astronomical fringe measurements are made in different correlators. This is the approach suggested by the French A.S.S.I. group [21] where the error correction system uses the light from the near infra-red region of the source spectrum while the astronomical fringes are observed in the optical. The two channels can be optimised for their required function: the IR channel has a wide optical bandwidth and a short sampling time, while the optical channel has high spectral dispersion in order to study the variation of source structure with wavelength and, because it uses a CCD as a detector, is integrated for long periods so as to minimise the effects of read-out noise. Furthermore, the resolution in the IR channel will be much lower than in the optical channel on the same baseline so that there will be high-visibility fringes in the IR channel even when the object is resolved in the optical channel.

The main disadvantage of such systems though is that the short integration times and high signal-to-noise ratio per exposure required in the correction channel in order to be able to cancel out the fringe motions mean that this method will only work for relatively bright sources.

5.1.1 Slope Removal

An alternative technique is to try and remove only the long timescale fringe motions. For instance, if we make an analogy between the temporal and spatial phase perturbations, we can see that much of the power in the temporal variations is in 'tilts' i.e. phase ramps in time. By applying 'temporal tilt correction' we could therefore obtain increases in usable exposure times in a similar manner to the way that spatial tilt correction allows us to increase aperture sizes. The way to do this is known in radio astronomy as 'fringe fitting' [87] and has been suggested for use in optical interferometry in space [49].

The basis of the method is to notice that the complex visibility of a fringe perturbed by a constant phase ramp can be described as a complex oscillation of constant frequency. Thus if we take a Fourier transform of the measured fringe signal as a function of time we will see a peak at the position corresponding to the slope of the perturbing phase ramp.

Photon Noise

As usual, this technique is affected both by photon noise and by atmospheric noise. To understand the photon noise it is helpful to consider the data as a two-dimensional fringe pattern, one dimension x containing the instantaneous distribution of light in the fringe pattern and the other dimension t being time. A two-dimensional fringe will be running across this pattern, the spatial frequency s_0 being fixed by the optics and the temporal frequency f_0 varying according to the slope of the fringe motions. In other words, if we consider the Fourier domain the position of peak amplitude will move along a line of constant spatial frequency (see figure 5.1).

In searching for this peak we are looking for the Fourier component with the largest amplitude; we are not concerned about the phase. We can therefore apply the analysis presented in section 2.4. However what we are interested in is not measuring the height of the peak but in knowing its position. Thus we are more interested in the peak height relative to noise *at other frequencies*. If we assume that all the power in the Fourier transform of the classical intensity distribution is concentrated around the points $(0, 0)$ and (s_0, f_0) then the photon noise elsewhere in the plane will have a constant variance of $\overline{N}^2$ where $\overline{N}$ is the number of photoevents occurring during the total slope measurement period T. To determine the position of the peak it is then simply a matter of setting some threshold above the noise and looking for points along $s = s_0$ which are above this threshold. For this to be reliable, the peak must rise to a significant level above the noise within one coherent integration time (i.e. T) - there is little to be gained in incoherently averaging successive observations since the fact that we were forced to terminate coherent integration indicates that the slope has

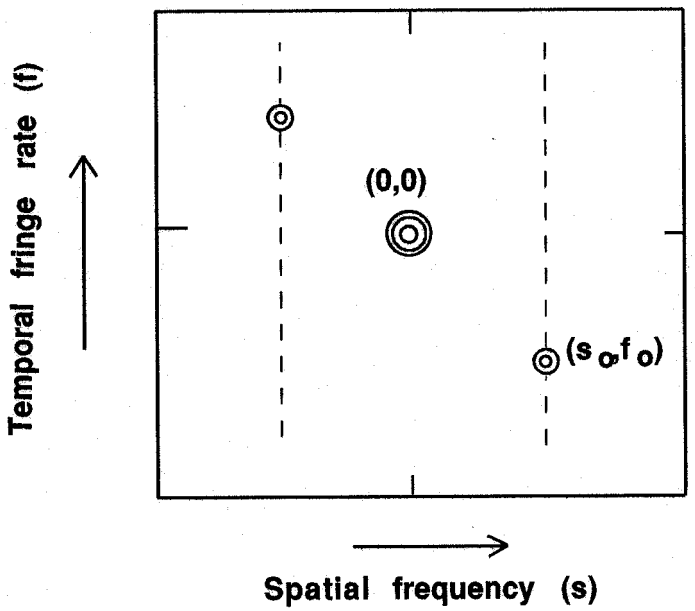

Figure 5.1: *Schematic diagram of the space-time power spectrum of the fringe pattern measurements. The position of the peak moves along the indicated line depending on the mean slope of the temporal phase perturbations.*

changed significantly and hence that the peak frequency f_0 has changed. We shall take a signal-to-noise ratio of 4 as being the minimum for reliable detection. It is easy to show that the mean peak signal is $\langle |V(f_0,T)|^2 \rangle \overline{N}^2$ where $V(f_0,T)$ is the visibility of the strongest Fourier component in the classical spatio-temporal intensity pattern and so the signal-to-noise ratio is

$$SNR_d = \langle |V(f_0,T)|^2 \rangle \overline{N}\,,$$

where the subscript d refers to detection. Note that because this signal-to-noise ratio refers to noise at different frequencies to the signal it has the same functional form at high and low light levels.

Atmospheric noise

In the case of radio VLBI or optical interferometry in space it is a good approximation to take the phase error variations in time as being pure slopes over long periods, since they are caused by local oscillator frequency offsets and drifts in the telescope separations respectively. In the case of atmospheric phase perturbations, we are merely attempting to fit a slope to a random time series, making use of the fact that it is dominated by long wavelength modes. This has two consequences: firstly the relative height of the peak (i.e. its visibility) reduces as we increase the integration

time because of the residual higher-order perturbations and secondly power begins to appear at other frequencies as the best-fit slope begins to change. The optimum integration time is one that balances these effects against the increase in the number of photons collected. If we ignore the peak-broadening effects the optimum integration time will maximise the quantity

$$SNR_d \propto \langle |V(f_0, T)|^2 \rangle T. \tag{5.1}$$

The mean squared visibility $\langle |V(f_0, T)|^2 \rangle$ was computed as a function of integration time using the simulations of the spectrum phase perturbations described in chapter 3. The peak position was determined in these simulations by finding the best fit slope to the phase perturbations - this will only necessarily be the position of the peak when the residual perturbations are small, but this is the main region of interest, and anyway we will see in the next section that it remains a good approximation even for quite large residual perturbations. The results are shown in figures 5.2 and 5.3. We can see that the optimum integration time in terms of pure photon noise signal-to-noise ratio is $T \simeq 9t_o$, but that it might be preferable to use a shorter integration time, say $5t_o$, which reduces the fluctuations of the high light level visibility while only sacrificing a small amount of photon signal-to-noise ratio.

Despite the increased integration time and higher visibilities compared to uncorrected exposures, in order to get a detection the light level must be quite high: we can define a 'canonical signal-to-noise ratio' as

$$SNR_0 = \{\langle |V(0)|^2 \rangle \overline{N}_0\}^{1/2} \tag{5.2}$$

where $\langle |V(0)|^2 \rangle$ is the mean squared visibility that would be observed in the absence of temporal phase fluctuations (but including such effects as visibility losses due to finite aperture size etc.) and $\overline{N}_0$ is the mean number of photoevents occurring in time t_0. With this definition we can see from figure 5.2 that we shall only detect the slope reliably when the canonical signal-to-noise ratio is greater than about 1.

The drawbacks to this method are common to most error-correction schemes: firstly, it works only at quite high light levels where the benefit in terms of signal-to-noise ratio is least and secondly the accuracy of the error correction and hence the measured visibilities are dependent on the source flux so that, when calibrating the visibilities on a reference object of different brightness, account must be taken of this effect.

5.1.2 Fringe envelope tracking

The second and more serious effect of the piston term in the phase perturbations is the loss in temporal coherence between the beams from different telescopes. Under

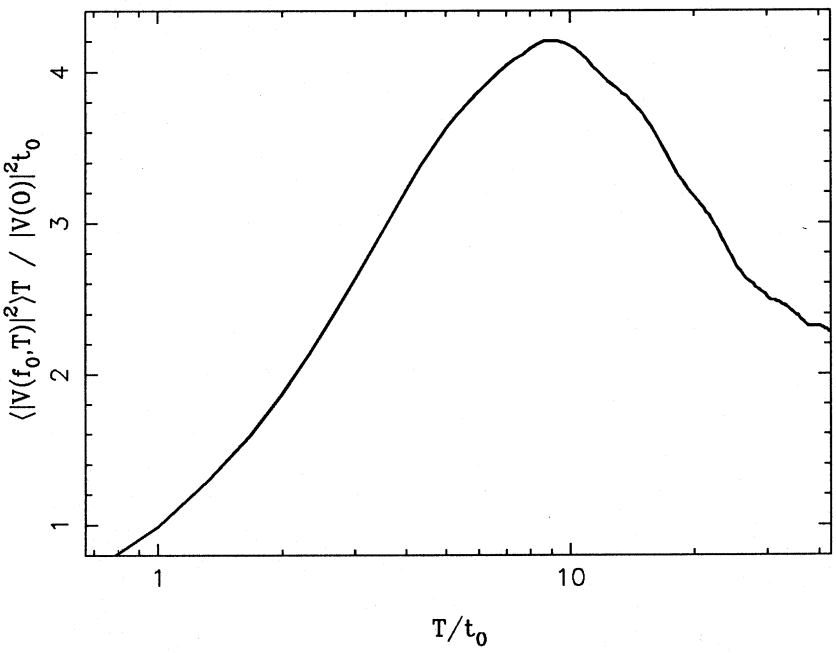

Figure 5.2: *The signal-to-noise ratio for determining the linear component of the atmospheric temporal phase fluctuations under photon-noise-limited conditions, as defined in equation 5.1.*

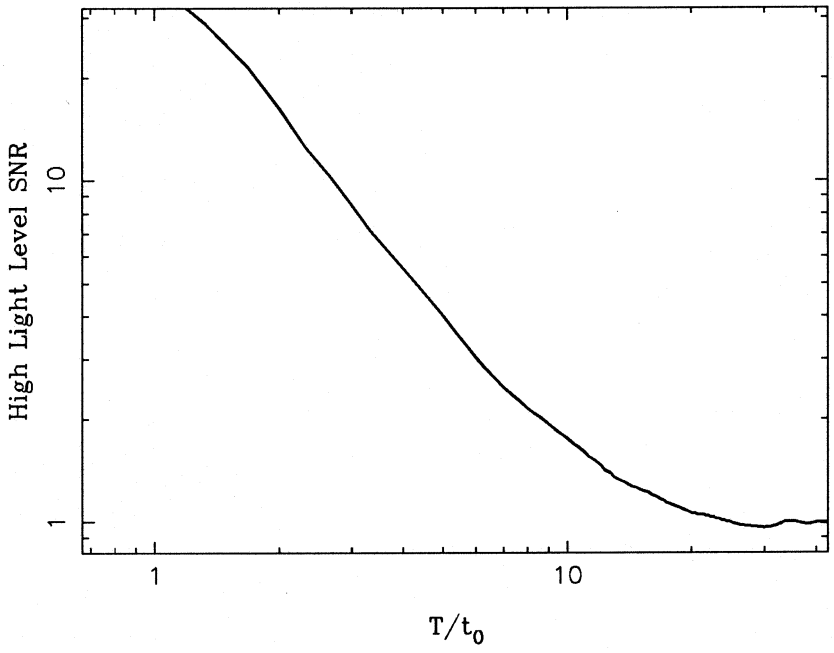

Figure 5.3: *The high-light-level signal-to-noise ratio of fringe amplitude measurements when the best-fit phase slope has been removed from the atmospheric temporal phase variations.*

1 arcsecond seeing conditions the relative atmospheric optical path delay between beams arriving at telescopes separated by 100 metres will have an r.m.s. fluctuation of about 66 microns (in other words, the position of the 'white-light fringe' is uncertain by this amount). If we are observing at a wavelength of 800nm with a fractional bandwidth of 10% then the coherence length of the radiation will be 8 microns and hence our chances of observing any fringes at all are small. If we do not attempt to actively find the white light fringe position then we must reduce the bandwidth by a factor of 10 or more in order to stand a reasonable chance of seeing the fringe. If we place the additional constraint that the uncertainty in the fringe visibility due to coherence losses is less than say 1% then we must reduce the bandwidth to less than 0.4% (this is the approach adopted by Davis *et al.* [26] who use a 0.3nm bandpass centred around 446nm). Thus the photon rate is reduced by a factor of about 25 compared to the bandwidth allowable if we were constrained by 'bandwidth smearing' requirements only. We can compensate for this by having many channels observing at different wavelengths and adding the results incoherently, but it is easy to show that the overall signal-to-noise ratio at low light levels is still reduced by large factors.

Thus we need a system for finding the white light fringe position and for tracking its motion over time. The requirements of such a system are not as severe as for a system which removes the phase perturbations completely, since the allowable error in our knowledge of the white light fringe position is of the order of the coherence length of the radiation rather than being of the order of a wavelength. We shall use the term 'fringe envelope tracking' for this type of system because we are trying to follow the position of the fringe amplitude envelope that is caused by temporal coherence effects.

Group delay fringe tracking

One way of doing this is to scan the optical path delay rapidly so that the fringe envelope moves past our detector and look for the position of maximum fringe visibility. However this is not only difficult mechanically because of the high speeds required but is also wasteful in terms of signal-to-noise ratio considerations because we spend a lot of time observing the low-contrast portion of the envelope. What we can do instead is to invert the principle of the Fourier transform spectrometer: that is we can determine where the fringe envelope is by observing a dispersed spectrum.

In this technique, called 'group delay fringe tracking' and first proposed by Labeyrie, we disperse a one-dimensional fringe pattern (whose phase increases along the x direction) so that the fringes formed by light of different wavelengths are separated in y direction (see figure 5.4). If we consider the instantaneous fringe phase as a function of wavelength, we can see that it increases linearly with optical frequency, the gradient being proportional to the relative path delay between the two interfering beams

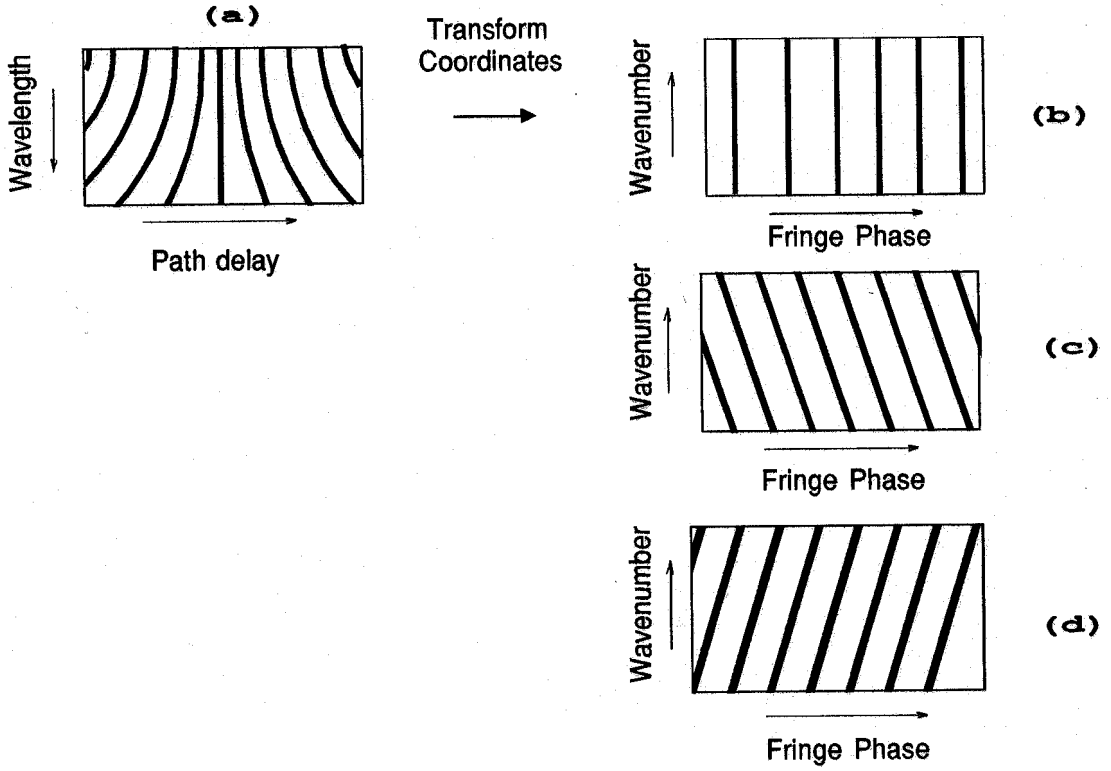

Figure 5.4: *Schematic diagram of a dispersed fringe pattern when the relative path delay between the incoming beams is (a) zero (b) positive & (c) negative.*

at the phase centre of the correlator. Hence if the detected fringe pattern is mapped onto a new set of coordinates such that the new x coordinate is proportional to phase and the new y coordinate is proportional to optical frequency, then we will see sloping fringes. A Fourier Transform of this pattern will show a peak at a point whose spatial frequency is fixed by the optics and whose 'spectral frequency' corresponds to the optical path delay. The resemblance of this to temporal fringe fitting is obvious and much of the analysis follows through. Indeed it has been suggested that these two methods should be combined so that at each point in time we can determine both the path delay and the fringe drift velocity from a 3-dimensional spatio-temporal-spectral (!) fringe pattern [49]. We shall concentrate here on the group delay method alone because it has a greater impact on the limiting magnitude of the system.

From here on, we shall restrict the problem to one dimension by considering a fixed spatial frequency (i.e. that corresponding to the fringe spatial frequency set by the optics) and considering the variation of the complex fringe visibility at this frequency with optical wavenumber - we shall call this the 'spectral visibility function'. The Fourier conjugate space we shall call 'group delay space' since the position of the peak amplitude in this space will indicate the optical path delay.

The convolution theorem tells us that the peak in group delay space will be con-

volved with a function whose shape depends on the spectral visibility function at zero optical path delay. This depends both on the observed source and on the instrumental spectral response. For example the width of the peak will be depend on the source spectrum and on the spectral bandpass: generally the wider the bandpass, the better we can determine the white light fringe position, but if the bandpass is too wide, different optical wavelengths on the same baseline will sample different parts of the source coherence function (or in fact the source's shape may change with wavelength) and the resulting variation of visibility with wavelength will tend to broaden the peak.

The convolution theorem also tells us that the response of the system cuts off at white-light fringe offsets greater than the coherence length set by the dispersion of the spectrum, which is what we might expect intuitively.

Rough estimate of the signal-to-noise ratio

The major limitation, though, is once again the available signal-to-noise ratio. Much of the formalism carries over from the discussion of temporal slope correction. In this case, however, we shall assume that no correction of the short or medium term temporal phase variations is available, so that the optimum exposure time is simply that given in section 3.2, i.e. about $1.6t_o$. The change in group delay will take much longer than this: for a system with a spectral bandpass $1/\alpha$ times the centre frequency, the white-light fringe must move by α wavelengths (i.e. $2\pi\alpha$ radians) in order for the group delay peak to move by its own width. The time this will take will be of the order of $(2\pi\alpha)^{6/5}t_o$. For a 10% bandwidth, this corresponds to $143t_o$ or about 90 exposures. We might expect to be able to integrate incoherently for about half this time without smearing the peak too badly and so we would get a factor of 6 improvement in signal-to-noise ratio compared with a single exposure. If we set the minimum signal-to-noise ratio for reliable detection of the peak to be 4, then we require SNR_d for an exposure to be 0.6. We can derive from the results of chapter 3 that this implies a canonical signal-to-noise ratio of 0.8. This result does not depend very strongly on the bandpass assumed: if we increase the bandwidth, we decrease the amount of time the peak takes to move its own width. We could increase the limiting magnitude at the expense of less accurate knowledge of the optical path delay by smoothing out the group delay peak and then averaging incoherently for a longer period, but this possibility will not be considered further here.

Simulations: [1] Passive tracking

To test these predictions, a numerical simulation was set up of a group delay fringe tracking system. The simulated correlator consisted of 32 spectral channels equally spaced in wavenumber across a 10% fractional bandwidth. The Nyquist sampling

frequency therefore corresponds to the white-light fringe being ±160 wavelengths from the phase centre of the correlator. Each channel had a square bandpass equal in width to the channel separation so that the full bandwidth was covered. It was assumed that the variations of the object intensity and the sampled object coherence function across the bandwidth were negligible. The fringes in each channel were perturbed by simulated atmospheric path delays with a Kolmogorov spectrum, with the path delay (but not the *phase* delay) at any one instant being the same for all channels. The fringes in each channel were coherently integrated for $1.5t_o$ (t_o defined at the centre wavelength) and the resulting visibilities were then further reduced to take account of the finite bandwidth of each channel when the white-light fringe offset was large. The spectral visibility function was then Fourier transformed using an FFT and the squared modulus computed. Figure 5.5 shows the resulting output as a function of time. We can see how the modulus of a given Fourier component rises and falls as the peak moves through the sampled frequency, with a typical timescale of about $50t_o$, although there are some much longer periods corresponding to quiet atmospheric conditions. We can see rapid fluctuations in amplitude due to atmospheric noise on individual exposures and also longer periods where the peak is between group delay samples and so no sample is high. This latter could be avoided if we oversampled in group delay space, but this was not considered here because it would increase the computation time and complicate the analysis, since the noise in different channels is then correlated.

Photon noise was then added to the simulation. It was assumed that the fringes in each spectral channel were sampled in the spatial direction at twice the Nyquist rate, i.e. 4 samples per fringe, and simulated photoevents were generated in each channel according to the derived classical fringe visibility for each exposure. Figure 5.6 shows the resulting output for canonical signal-to-noise ratios of 0.5, 0.7, 1.0 and 1.4 (corresponding to $V(0) = 0.25$ and $N_0 = 4$, 8, 16 and 32 photons per t_o). The output has been incoherently averaged in the time direction using a low-pass filter with a time constant of $75t_o$. We can see that at low signal-to-noise ratios the signal only rises above the noise during the periods when the atmospheric perturbations are 'quiet'. When the canonical signal-to-noise ratio is above 0.7, a good enough signal is present often enough so that the human eye can interpolate between the bad patches (it helps if you squint when looking at it!). Thus 'passive tracking' is possible at these light levels — that is, if the coherence length of the individual spectral channels is long enough so that we can see fringes most of the time, then we can hold the instrumental path delay constant and work out after the observation what the path delays were. This knowledge can then be used to coherently combine the fringe measurements in different channels and achieve the same signal-to-noise ratio as if the internal path delay had been moved to track the white light fringe.

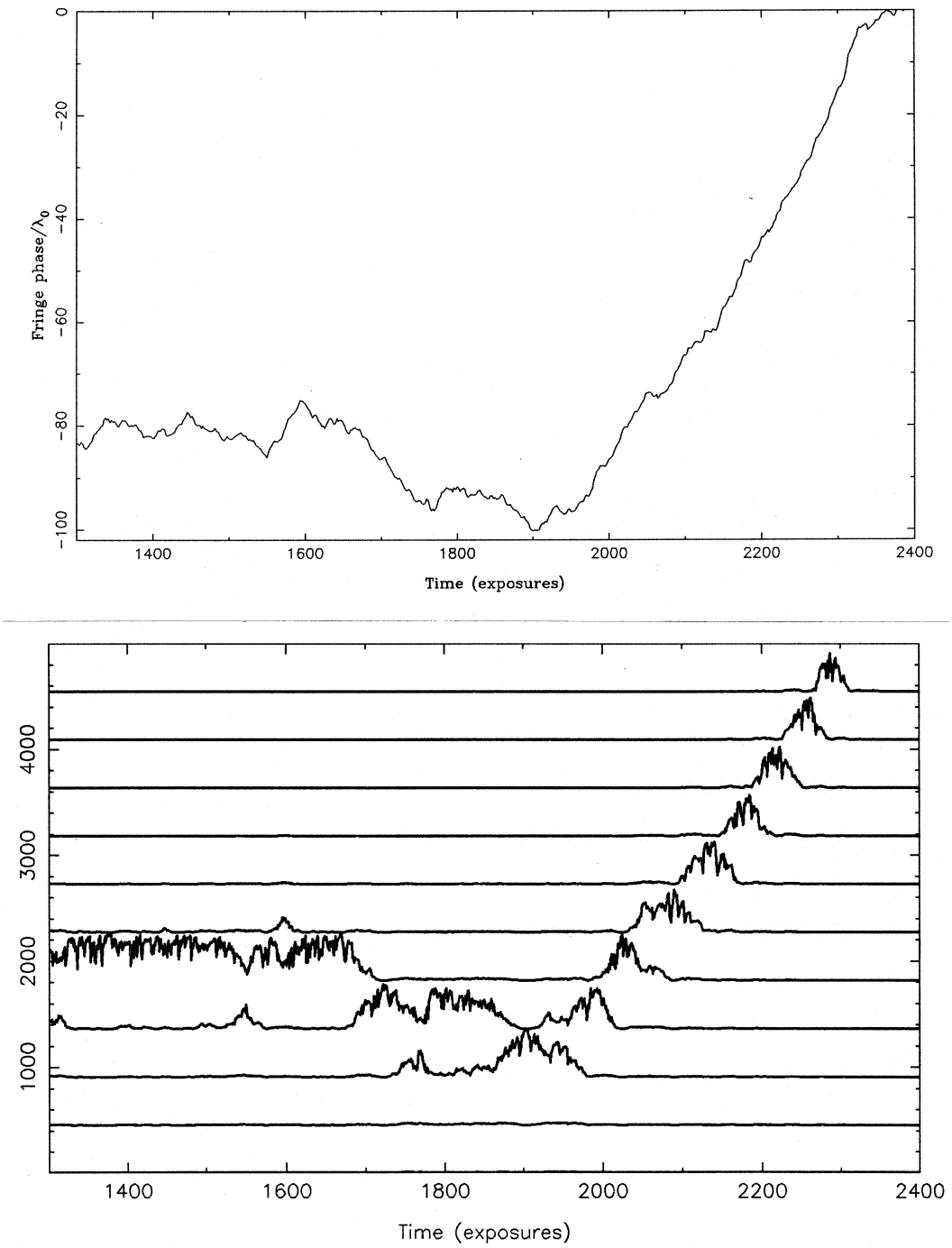

Figure 5.5: *The high-light-level response of the group delay envelope tracking system to simulated atmospheric path length fluctuations. Figure (a) shows the 'true' path-length variations and figure (b) shows the response of the system. Each line in (b) represents the power in one group delay channel as a function of time.*

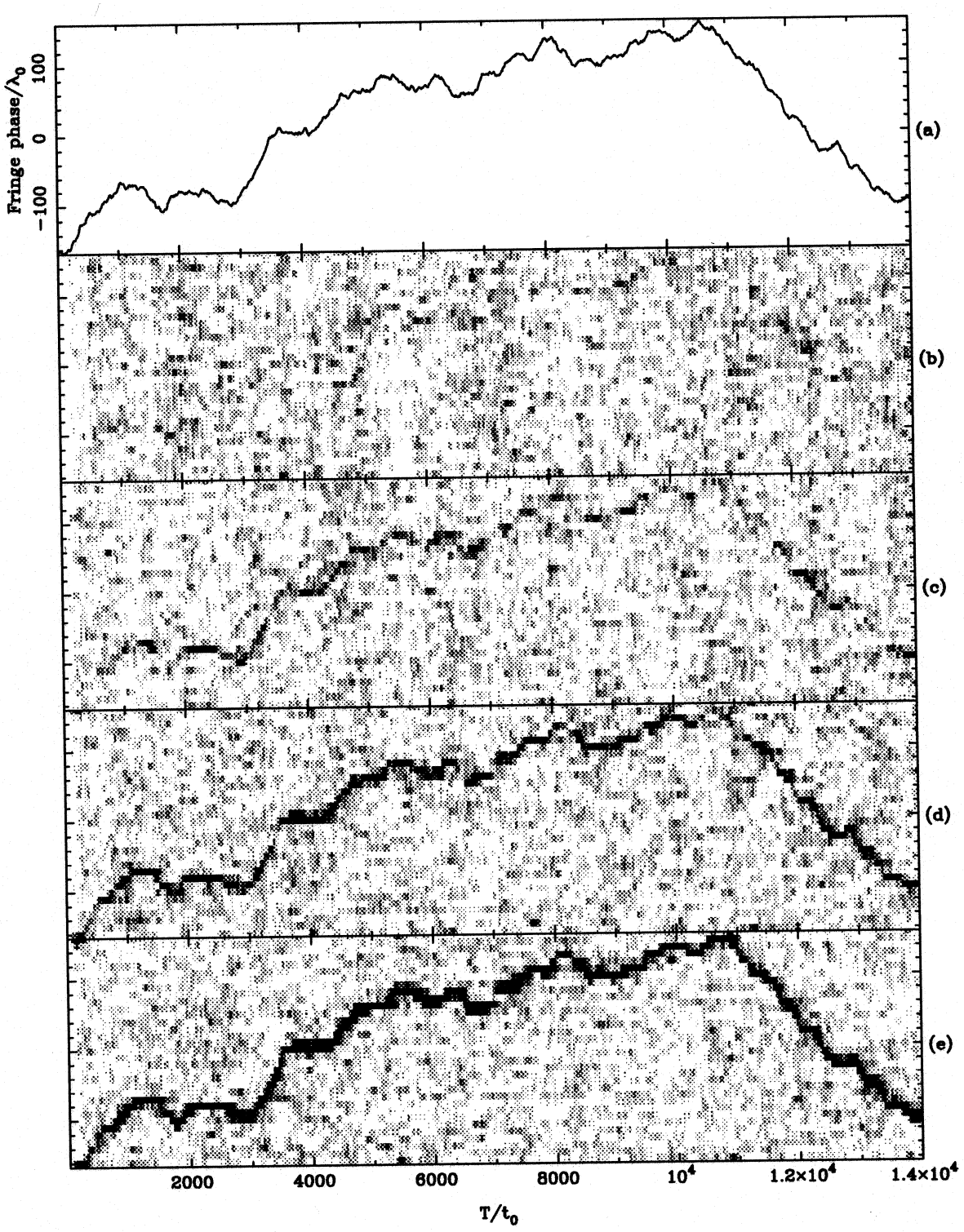

Figure 5.6: *The response of the group delay envelope tracking system to simulated atmospheric path length fluctuations at low light levels. Figure (a) shows the 'true' path-length fluctuations and figures (b), (c), (d) and (e) show the response at simulated photon rates of 4, 8, 16 and 32 photons per t_o respectively. The instantaneous fringe visibility in all cases was 0.25 (see text).*

[2] Active tracking

The disadvantage of this method is that it requires a large number of channels if the total bandwidth is to be large, which is a problem if the correlator is to be made out of discrete detectors — in COAST the detectors of choice are avalanche photodiodes operated in geiger mode, and these cannot be used in contiguous arrays because of stray photons generated in the avalanche process. The alternative is to use a small number of wide-bandwidth channels and move the internal delay to keep the white-light fringe within the coherence length of the channels. The signal-to-noise ratio for this method is inherently no different than for a larger number channels providing that we can keep track of the white-light fringe. If however we do lose track of it , we must start again from scratch in searching for it because once the fringes have disappeared due to loss of coherence, we have no information to tell us in which direction to search. In the passive mode we have the benefit of hindsight and can backtrack as well as forward track to estimate the group delay even when the signal has disappeared into the noise.

From figure 5.6(d) it would appear that, when the canonical signal-to-noise ratio is of order 1, the signal is strong for enough of the time to allow us to actively track the group delay with a small number of channels. It was decided to see if the tracking process could be automated and still cope with such low signal-to-noise ratios. It was decided to simulate a system with 5 spectral channels, giving in group delay space one centre channel and 4 error channels — two positive and two negative. This allows us two chances to see the white-light fringe delay moving away before it is 'lost' — if the group delay peak moves more than two channels away, aliasing will cause it to reappear on the opposite side, and unless the software is very intelligent a catastrophic loss of track will occur. It was decided not to try to deal with such an occurrence since this would be of limited value at low signal-to-noise ratios.

Algorithm

The algorithm used for tracking the fringes was mostly ad-hoc; it attempts to reproduce what one is able to do 'by eye' when presented with figure 5.6. First the data is temporally smoothed with a first order filter. A suitable time constant for this is about $40t_o$ — any longer and the system is too slow to respond when steep phase ramps appear. Then we reject spurious noise peaks by placing a threshold on the data: if no channel appears above the threshold (set to 2.5 times the r.m.s. noise level) then the system does nothing and waits for the signal to reappear.

Thirdly we reject peaks which are too far from our last good estimate. This is done by assigning a probability distribution to the peak position. It starts out as being tightly constrained around our last good estimate ('good' being defined as the

peak being above some threshold — here defined as 4 times the r.m.s. noise level) but then the width of the distribution is increased as time goes on until a new 'good' peak appears. The standard deviation should increase as $t^{6/5}$ but this was approximated as simply t to save computation. The probability of a given peak being the true group delay position is then, according to Bayes' Theorem, weighted by this 'prior probability'. We can make a simplified model of this by assuming that the noise is signal-independent and that the peak is fixed in height and concentrated in one channel at a time. Let us write the value at a given point in time of the n independent group delay channels as $\{z_k,\, k = 1...n\}$. If the peak is in delay channel d and is of height h then the probability of a given set of measurements is

$$p(\{z_k\} \mid d, h) \propto \exp\left(\frac{-(z_d - h)^2}{2\sigma^2}\right) \prod_{k \neq d} \exp\left(\frac{-z_k^2}{2\sigma^2}\right) ,$$

where it has been assumed that the noise in each channel is independent and normally distributed with variance σ. If we now write the prior probability for d in the form $\exp\left(-(d - d_0)^2/2\sigma_d^2\right)$ where d_0 is the position of the last good estimate and σ_d is the predicted r.m.s. movement of the peak, then the *a posteriori* probability of the peak being at position d is

$$p(d \mid \{z_k\}, h) \propto \exp\left(-\sum_{k=1}^{n} z_k^2/2\sigma^2 - h^2/2\sigma^2 + hz_d/\sigma^2 - (d - d_0)^2/2\sigma_d^2\right).$$

It would be possible at this point to marginalise over some prior distribution of peak heights $p(h)$, but we shall assume for simplicity that h is known. If we now drop the factors which are independent of d we get

$$p(d \mid \{z_k\}, h) \propto \exp\left(hz_d/\sigma^2 - (d - d_0)^2/2\sigma_d^2\right).$$

The most probable value of d is the one which maximises the exponent. This latter has the advantage of being very easy to compute.

Having decided on the best estimate of the group delay, the simulator's path compensator is moved to make this delay zero. This part of the simulation is unrealistic in that the path compensator is able to move instantaneously in discrete steps (integer multiples of the group delay sampling distance). This was so that when the path compensator delay was changed, it was simple to continue using the information accumulated in the group delay channels by simply 'shuffling' the channel values along. In reality one would move the delay smoothly, but this would need some algorithm for interpolating the old channel values onto the new sampling points.

Results

This algorithm was found to work successfully at a canonical signal-to-noise ratio of 1 and to track for long periods (of the order of a thousand exposures) when this was

reduced to 0.7. Failure usually occurred when the atmospheric phase perturbations entered a steep phase ramp, since the peak then stayed only a short time in each channel and if a mistake was made because of a spurious noise peak then there was little time to recover from it. This is perhaps a good argument for adding temporal slope correction to the system, since it would give extra information and help to boost fringe visibilities just at the time when this was needed.

The main ommission from the current algorithm is a method of making use of past fringe motions to predict the future motions e.g. some form of Kalman or Wiener filter. This would clearly increase the performance of the system, but some care would have to be taken to introduce information about the different qualities of the group delay estimates at different times.

The system was always started up close to the white-light fringe, as no fringe-searching algorithm had been written, but such an algorithm would not be hard to implement since at these signal-to-noise ratios there is little doubt as to whether fringes are present, even if their exact location is a little harder to pinpoint.

5.1.3 Phase Tracking

As a comparison, it was decided to investigate an alternative way of tracking the white-light fringe called phase tracking. This technique does not in itself allow one to tell where the white-light fringe is, it merely keeps one in the same place relative to the fringe envelope for long enough that some other method can be used for determining the offset from the white-light fringe, for example by slowly moving relative to a reference point in the envelope and noticing the visibility changes.

The method is very simple both in terms of hardware and of software. A single spectral channel is used — this channel can have a wide bandwidth to increase the signal-to-noise ratio but it must be remembered that the fringes will begin to blur if the bandwidth is so wide that the object coherence function sampled at different parts of the bandpass is substantially different. The technique relies on the assumption that the phase difference between two consecutive exposures is never more than 180° in magnitude. The difference between the measured phases on consecutive exposures is made to conform to this requirement by suitable additions or subtractions of 360° and the path compensator is moved in the deduced direction of the phase change. By keeping track of the fringe motions over many successive exposures, phase excursions of many cycles can be 'unwrapped' and the path delay kept constant to better than a wavelength.

The disadvantage of this method is its sensitivity to noise. It relies for its success on choosing the right 'unwrapping' for the phase change on every exposure. A wrong decision will cause the path delay to 'slip' by 360° and there will be no method of

telling this has happened apart from secondary indications such as change in the mean fringe visibility. Loss of track for this method is as catastrophic here as it is with the group delay method, because if a wrong decision causes a movement away from the peak of the fringe envelope, the lowered visibilities so caused will make the chances of another wrong decision even higher. The signal-to-noise ratio requirements are therefore quite severe: the combined effects of the atmospheric phase change and the photon noise must never (or at least very rarely) produce a phase change that is greater than 180°. The atmospheric requirement forces us to use very short exposures, which means that it is even harder to meet the photon noise requirements.

To find out the limiting signal-to-noise ratio for this method, a numerical simulation of this scheme was set up and was run with the same atmospheric phase perturbations as used for testing the group delay simulator. It was found that even with no photon noise, the tracker failed to follow the perturbations unless the exposure time was shorter than $0.5t_o$. With this exposure time the tracker worked successfully at a signal-to-noise ratio per exposure of 2.5 (the criterion of success used here was tracking for more than $6000t_o$ — this corresponds to about a minute for typical seeing timescales). This corresponds to a canonical signal-to-noise ratio of 3.5. Thus if the total bandwidth allowed for the fringe tracking system is the same for this method as for the group delay method, then the group delay method will work for objects 2.7 magnitudes fainter than for the phase tracking method. Of course, no use has been made of methods which allow the prediction of fringe motions from their past motions, but there would have to be a substantial improvement in performance for the phase tracking method to become comparable to the group delay method.

5.2 Tilt Correction

Chapter 3 showed the substantial gain in signal-to-noise ratio and calibration accuracy that can be achieved if the tilt component of the wavefront perturbations is removed. This however assumed perfect compensation of the tilts. In practice, the tilt correction will be imperfect because of inaccuracies in the measurements of the tilt, both because of photon noise and because of 'atmospheric noise': in the case of tilt determination, the atmospheric noise will be in the form of changes in the shape of the image we are trying to track.

To gain an idea of the magnitude of the problem, we must first work out how well we need to track the tilts. We shall express the tilts in terms of the offsets they cause in the image plane, since most tilt detection systems operate in this plane. If the differential tilt error between two beams is θ and the radius of the Airy disc formed by these beams is θ_0 then we can show, by application of equation 3.9 in the case where the phase errors involved are small, that the r.m.s. visibility loss due to the

tilt error when the beams are combined will be

$$\eta \simeq 1 - 1.8\langle(\theta/\theta_0)^2\rangle.$$

Thus for visibility losses less than 10% we need the tilt correction to be better than 0.23 times the Airy disc radius.

Tango and Twiss [85], in an excellent review of the tilt correction problem, show that for a Kolmogorov spectrum of fluctuations, the total power in the atmospheric tilts is

$$\langle\theta_u^2\rangle = 0.457(D/r_o)^{5/3}\theta_0^2(D)$$

where θ_u is the *relative* tilt between two apertures, assumed to be widely separated, and D is the aperture diameter. They derive the power spectrum of these fluctuations (see their Appendix B) which has a power law dependence of $f^{-2/3}$ at low frequencies and $f^{-11/3}$ at high frequencies. The 'knee' frequency between these regimes is given by

$$f_0 = v_T/\pi D$$

where v_T is some characteristic transverse windspeed. From the definitions of r_o and t_o given in chapter 3, we can see that this windspeed is of order $0.31r_o/t_o$ and so

$$f_0 \simeq (10[D/r_o]t_o)^{-1}.$$

Hence for a value of t_o of 10ms and an aperture with diameter $2r_o$, the knee occurs at 5Hz.

Tango and Twiss then consider the effect of the introduction of a tilt-correcting servo with a finite bandwidth. The servo will be unable to correct tilt fluctuations faster than some cut-off frequency f_1. They show that the fractional power in the residual tilts will be

$$\frac{\langle\theta_c^2(f_1)\rangle}{\langle\theta_u^2\rangle} = 0.184\int\left[1 - \frac{f_1/f_0}{\{t^2 + (f_1/f_0)^2\}^{1/2}}\right] t^{-5/3}J_1(t)\,dt$$

where θ_c is the residual tilt error after correction by the servo. If the servo cut-off frequency is $f_1 = f_0$ the reduction in tilt power is just adequate for there to be less than 10% visibility loss for an aperture diameter of r_o. However, this neglects the effects of noise in the measurement of the tilt, which we shall now consider.

5.2.1 Tilt measurement

We shall consider a tilt measurement system which determines the tilt from the displacement of the image formed by bringing the incoming beam to a focus. We shall assume that the tilt correction is fast enough and accurate enough so that the image is always close to the nominal centre of the image detector, 'close' being defined as much less than an Airy disc radius away (as we have seen, the visibility losses will be unacceptable if this is not the case).

Optimal linear estimators

Given noisy measurements of the image intensity distribution, we need some estimate of the position of the image centre. Clark *et al.* [13] introduce the idea of 'cross-correlation' estimators; we shall derive their results here from a different angle so that they can be applied to detectors which are not photon-noise-limited and which have pixels which are not small when compared to the image size.

Say we are given the measurements of n pixel intensities $\{\hat{I}(x_k, y_k),\ k = 1...n\}$ where (x_k, y_k) is the coordinate of one corner of the pixel, assumed here to be rectangular and of dimension $(\Delta x_k, \Delta y_k)$. We shall restrict ourselves to *linear* estimators of the image position

$$\begin{aligned} \hat{x} &= \sum_{k=1}^{n} h_x(x_k, y_k)\hat{I}(x_k, y_k) \\ \hat{y} &= \sum_{k=1}^{n} h_y(x_k, y_k)\hat{I}(x_k, y_k)\,, \end{aligned}$$

where h_x and h_y are what Clark *et al.* call 'cross-correlation functions': this is something of a misnomer in that we do not evaluate the cross-correlation of these functions with the measured image at all possible displacements but only at the origin. We shall concentrate here on the estimator for the x-component as the results for the y component will be the same.

If the true image intensity distribution is $I_0(x, y)$ and the image is displaced a small distance u_x from the origin then the resulting change in the true intensity in a given pixel will be

$$\Delta I(x_k, y_k) \simeq u_x \Delta_k$$

where

$$\Delta_k \simeq \left.\frac{\partial I_0}{\partial x}\right|_{(x_k, y_k)} \Delta x_k \Delta y_k$$

if the pixels are small. Hence if we had only a single pixel measurement, our estimate for the displacement would be

$$\hat{x} = \Delta\hat{I}(x_k, y_k)/\Delta_k.$$

If the noise in each pixel is independent we have n independent estimates of the image displacement. In combining them linearly, we can apply the standard result that the minimum variance estimate will be the average of the individual estimates weighted by the inverse of their variances

$$\hat{x} = \frac{\sum_{k=1}^{n}(\Delta_k/\sigma_k)^2(\Delta\hat{I}(x_k, y_k)/\Delta_k)}{\sum_{k=1}^{n}(\Delta_k/\sigma_k)^2}$$

where σ_k is the r.m.s. noise on each pixel. Hence

$$h_x(x_k, y_k) = \frac{\Delta_k/\sigma_k^2}{\sum_{k=1}^{n}(\Delta_k/\sigma_k)^2}. \tag{5.3}$$

For photon noise,

$$\sigma_k^2 = \overline{N} I_0(x_k, y_k)\Delta x_k \Delta y_k \Big/ \iint_{-\infty}^{\infty} I_0(x, y)\, dx\, dy$$

where $\overline{N}$ is the mean number of photons in the image and it has been assumed that a photon contributes unit flux to the image. This can be used to derive equation 9 in the paper of Clark *et al.*.

Readout-noise-limited measurements

For COAST, it is planned to use a CCD as the detector for tilt correction. This offers a much higher quantum efficiency then most photon-counting detectors, especially in the red (DQE of about 40% compared to about 1% for the PAPA camera discussed in Clark *et al.*'s paper) and enables the same detector to be used for aquisition as well as tilt correction because of the large number of pixels avaliable. The disadvantages of CCDs in this application are the relatively long ($\sim 80\mu$s/pixel) readout time and the readout noise ($\sim$ 6 electrons/pixel). Both problems can be reduced by reading out a smaller number of larger pixels, but the question then arises as to how few can be read out without losing in terms of signal-to-noise ratio. We shall now show that reading out the minimum usable number of pixels i.e. 4 large pixels in a square (the 'quad cell' arrangement) is near optimal, provided that we are readout-noise-limited and that we restrict ourselves to linear estimation methods.

If the readout noise per pixel is σ_r then the variance of our estimator will be

$$\begin{aligned} \mathrm{var}(\hat{x}) &= \sigma_r^2 \sum_{k=1}^{n} h_x^2(x_k, y_k) \\ &= \sigma_r^2 \Big/ \sum_{k=1}^{n} \Delta_k^2. \end{aligned} \tag{5.4}$$

When the pixels are large, the change in flux in the pixel per unit image displacement must be written in full i.e.

$$\Delta_k = \int_{y_k}^{y_k+\Delta y_k} [I_0(x_k + \Delta x_k, y) - I_0(x_k, y)]\, dy.$$

If we now subdivide one of the pixels along either the x or y direction, to get two new pixels denoted by k_1 and k_2, we have the relationship

$$\Delta_k = \Delta_{k_1} + \Delta_{k_2}.$$

Hence

$$\Delta_k^2 = \Delta_{k_1}^2 + \Delta_{k_2}^2 + 2\Delta_{k_1}\Delta_{k_2}.$$

Substitution of this formula into equation 5.4 shows that the variance of our estimate will increase when we subdivide the pixel unless Δ_{k_1} and Δ_{k_2} are of different sign. This can only occur if $I_0(x, y)$ has gone through a turning point (i.e. a local maximum, minimum or saddle point) inside the pixel we have subdivided. In the case of the quad cell and an Airy disc, the peak of the main lobe will occur between the pixels, and so it is only worth subdividing the pixels at or after the first minimum. Since the flux beyond this minimum is a small fraction of the total, there will be little gain in doing so. We shall assume hereafter that the detector consists of four semi-infinite pixels i.e. it is a Hartmann or 'knife edge' detector, since what we are doing is trying to have half the image flux on either side of a knife edge (there are two knife edges at right angles in the 2-D case).

5.2.2 Atmospheric noise

The problem with the analysis so far has been the assumption that the image shape is known. For apertures much larger than r_o the residual (i.e. non-tilt) phase fluctuations will distort the shape of the image from exposure to exposure and cause errors in our estimates of the image position even at high light levels. Figures 5.7(a) through 5.7(f) show numerically generated images of point sources seen through simulated phase screens. We can see that for an aperture diameter of r_o, the image is essentially diffraction-limited, with only small variations in peak height. The main effect of the perturbations is seen in the sidelobes. As the aperture diameter increases, we see that the flux in the main lobe begins to fluctuate and more and more power occurs in the sidelobes, till at a diameter of $4r_o$ the characteristic 'speckle' structure is beginning to appear.

What will the effect of this be on the tilt correction? Firstly, is the definition of the image 'centre' used by the tilt correction system adequate on such hopelessly asymmetric images? To see if this is the case, it was decided to determine if there was any reduction of the fringe visibility when different definitions of the image centre are used to determine the tilts. This was done numerically by generating simulated phase screens and using them to generate high-light-level image profiles like those in figure 5.7. The nominal centres of these images were determined using two different criteria, namely (i) the 'knife edge' criterion i.e. the 'centre' is the point which subdivides the flux into two equal halves, and (ii) the centroid criterion which equates the centre with the centre of gravity of the image. The centre positions determined by these criteria were used in turn to correct the tilts on two separate apertures, and the resulting visibilities and their variances were compared with those obtained when

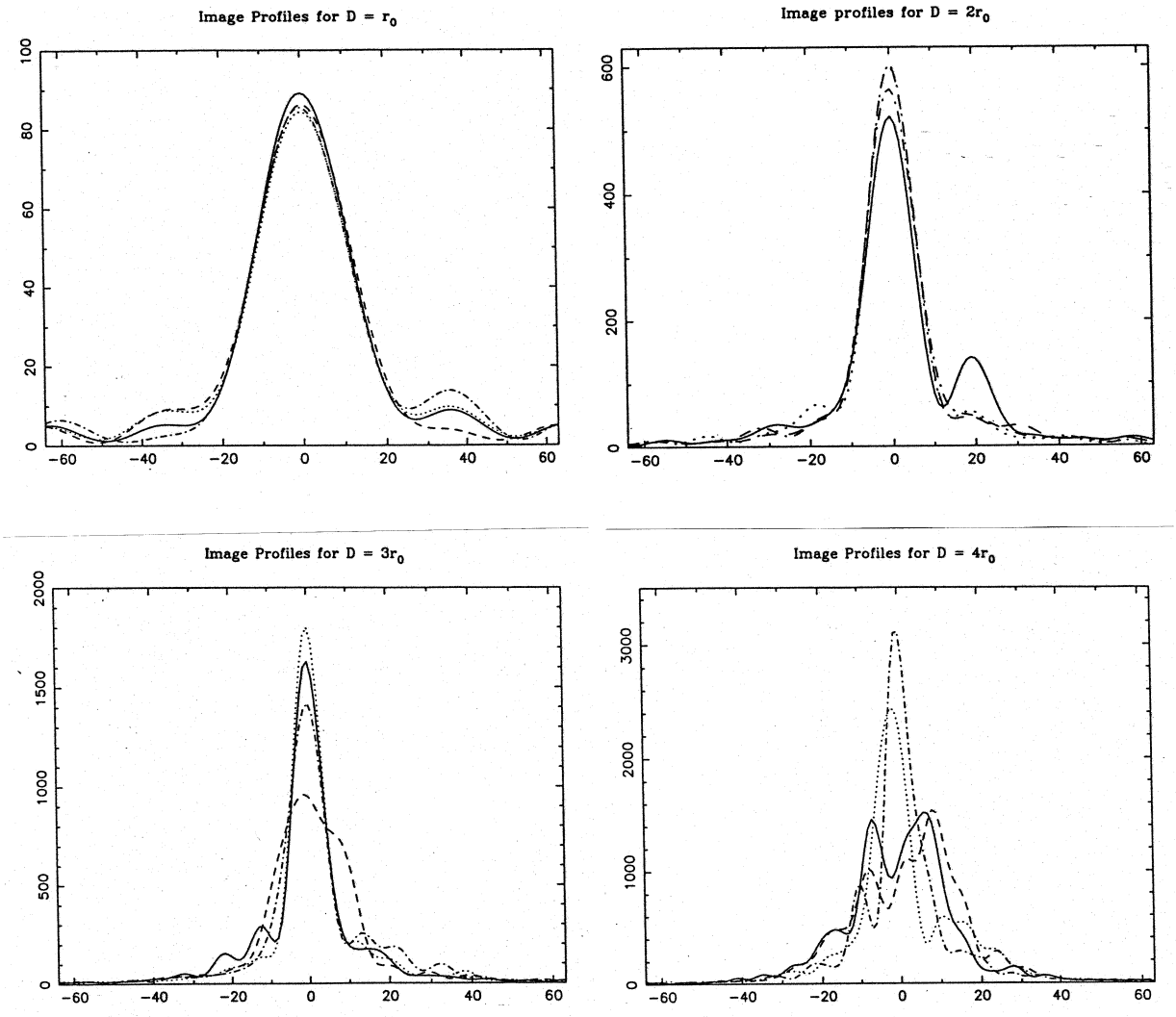

Figure 5.7: *Images of a point source seen through simulated atmospheric phase screens for apertures of various diameters. The lines plotted show the distribution of the flux in 6 different short exposures for each value of the diameter. The two-dimensional images were projected onto a line by integrating the flux along the orthogonal dimension.*

the tilts were determined directly i.e. by finding the least-squares best fit plane to the wavefront perturbations.

It was found that there was no substantial difference between the r.m.s. visibilities for the various methods over a remarkably wide range of aperture diameters. The centroid definition fared the worst: when the diameter was $4r_o$ the r.m.s. visibility of the fringes when the tilts were corrected using this criterion was lower by a factor of 10% than the visibility in the case where the least squares best fit plane was removed, but this improved at larger and smaller diameters. The knife edge definition was never worse than the least-squares plane determination (to within the statistical errors); there is in fact some evidence that the knife edge definition is *better* than the direct definition for diameters larger than $8r_o$ (the largest diameter tested was $12r_o$). The atmospheric noise on the visibility was always roughly the same for all three methods.

We can explain these results qualitatively as follows: for small aperture diameters the image is diffraction-limited and all definitions of the centre coincide; for medium-sized images most of the perturbations appear in the sidelobes and this is accentuated by the centroid definition which weights the sidelobes more than the centre; for the largest apertures, the higher-order perturbations begin to cause significant visibility losses and the knife-edge definition in some way takes account of this — possibly it works because the knife edge will tend to place the 'centre' near to the brightest speckle, and if the brightest speckles in two images are superposed we get a high fringe visibility.

The next effect of the atmospheric noise in the tilt correction system will be because it alters the flux density in the image compared to the diffraction-limited case. In a knife-edge system, the flux change in each pixel per unit displacement of the image is, to first order, directly proportional to the integrated flux along the knife edge - we shall call this the 'knife-edge flux'. Atmospheric noise will change the value of this flux and so change the proportionality constant between measured flux difference and image displacement. This can be compensated for by measuring the seeing, but any errors in the seeing measurement will translate into errors in tilt correction (note that it is potentially disastrous therefore to measure the seeing using the estimated power in the tilts!). Furthermore, because the mean flux density is always reduced by the atmospheric perturbations, the the detector signal-to-noise ratio will fall because the amount of flux crossing the knife edge per unit image displacement will fall. Lastly, the correct proportionality constant to use will fluctuate as the image shape changes and this will introduce a multiplicative error unless we have enough signal-to-noise ratio and pixels to determine the correct constant on an exposure-by-exposure basis.

To investigate these effects, further simulations were run with the aim of deriving statistics about the knife edge flux. In these simulations then knife edge position was again taken as being at the centre of the image as defined by the knife-edge criterion

— no account was taken of the variation in flux due to displacements away from the centre. Figure 5.8 shows the variation in mean knife-edge flux as a function of aperture diameter. This flux is normalised by the flux that would be expected from a diffraction-limited aperture of the same diameter and so the graph can be read as showing the knife-edge flux from a fixed-size aperture as a function of the seeing. From this point of view, we can see that we are most insensitive to seeing changes for aperture sizes below about $2r_o$ and that for apertures above this diameter, the knife edge flux increases linearly with r_o. We can interpret the latter result in terms of the image width for larger apertures being proportional to λ/r_o.

Figure 5.9 shows the percentage r.m.s. fluctuation in the knife edge flux from exposure to exposure. As expected, this increases with increasing aperture size for small apertures, but surprisingly the fractional variation begins to fall for apertures larger than $6r_o$. This can be explained in terms of a simple model of the image formed by large apertures: in this model, the image consists of a number of speckles randomly distributed over a patch of angular dimension λ/r_0. The speckles are modeled as being of fixed height and width, the width being given by the diffraction-limited resolution of the aperture, i.e. λ/D. The number of speckles in the image increases as $(D/r_o)^2$ and so the number of speckles lying across the knife edge will be given by a Poisson distribution with mean D/r_o. From this it is easy to see that the fractional fluctuation in the knife edge flux will be of order $(D/r_o)^{-1/2}$, and indeed the tail of the curve in figure 5.9 appears to follow this power law.

Thus the error in our determination of the tilt due to the rapid changes in the image shape will usually be $\lesssim 30\%$ of the actual error, so that if the tilt correction is rapid enough that the actual error never becomes large, we may neglect this effect.

5.2.3 Limiting flux level

With this analysis in hand, we can now calculate the expected tilt error for a typical system. We shall take an aperture size of $2r_o$ and assume that the exposure time has been chosen so that the r.m.s. tilt errors expected at high light levels due to the residual high-frequency components of the atmospheric tilt are roughly the same as the r.m.s. tilt errors due to photon/readout noise. We can make use of Tango and Twiss' equation 5.8 to show that for the quad-cell system the tilt error in one direction due to detector noise will be

$$\langle\theta_d^2\rangle = [(0.9656\theta_0/\langle k\rangle)(\sigma/F)]^2$$

where θ_0 is the diffraction-limited Airy disc radius, $\langle k\rangle$ is the degradation of the mean knife-edge flux due to the atmosphere (this can be read off figure 5.9) , F is the total collected flux per exposure and σ is the detector noise per pixel (this includes readout

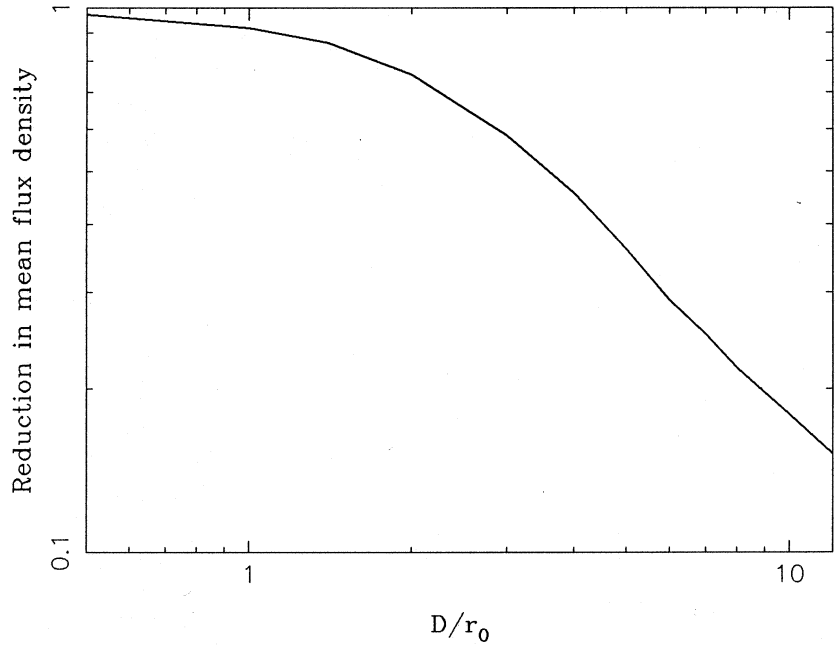

Figure 5.8: *The mean 'knife-edge flux' (see text) as a function of the aperture diameter, normalised by the value for a diffraction-limited aperture of the same diameter.*

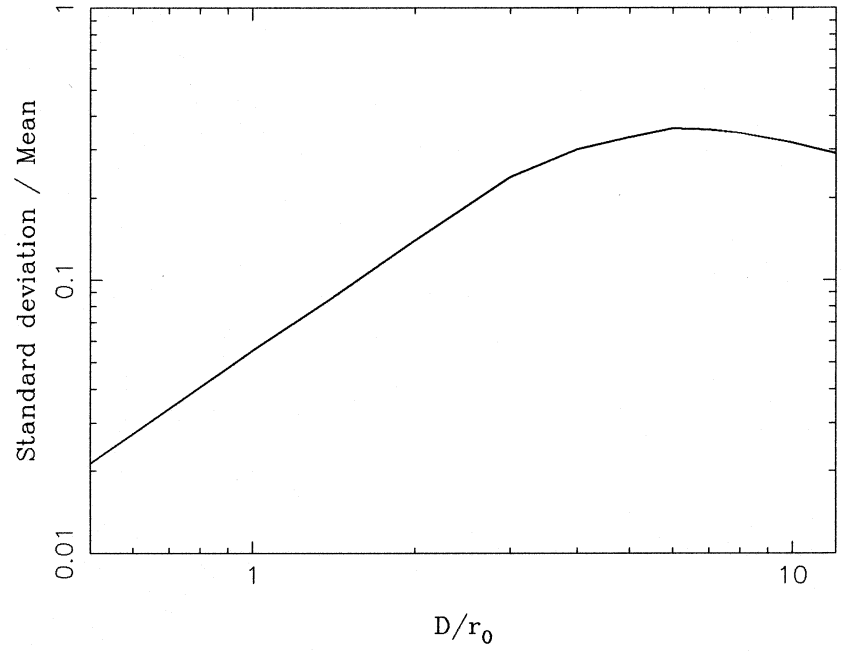

Figure 5.9: *The fractional variance of the 'knife-edge flux' (see text) as a function of the aperture diameter.*

noise and photon noise). This tilt variance must be multiplied by 4 to take account of the fact that there are two orthogonal tilt directions and two different apertures to tilt-correct. We must double this again to include the contribution of the residual atmospheric tilts to the total tilt error. Thus if we require the total error to be less than $(0.23\theta_0)^2$ (so as to give a visibility loss of less than 10%) and the aperture diameter is $2r_o$, we need

$$F/\sigma \simeq 16.$$

Thus if the readout noise per pixel is 6 electrons, then we need a flux of 132 detected photons per exposure in the tilt-correction system. Note that at these photon rates, the readout noise and photon noise are comparable in magnitude.

This is a relatively high photon rate compared to those acceptable in the fringe tracking system, but the tilt correction system gains in several ways. Firstly, the allowed integration time is slightly longer. If we choose our servo cut-off frequency as being twice the 'knee' frequency, f_0, of the tilt power spectrum then the Nyquist sampling time will be $2.5t_o(D/r_o)$ so that for an aperture size of $2r_o$ the integration time for the tilt correction image will be about 3 times as long as the fringe exposure time.

Secondly, the measured tilt is independent of wavelength as long as the air paths travelled by the light of different wavelengths are substantially the same. If this is so, then we can use a wide bandwidth in a different part of the source spectrum to correct the tilts in the fringe observation channel. Several authors have discussed this topic [42, 96] and show that the wavefront aberration, and especially the tilt component thereof, is highly correlated between wavelengths of 500nm and 1000nm for zenith angles less than 60°. Thus we could use the wavelength interval 500nm to 700nm for tilt correction and observe the fringes in a bandpass from 760nm to 840nm. Note however that to use such a large bandwidth, the system must include atmospheric dispersion correctors to reduce the dispersion of the image seen by the tilt correction system and to make its mean position independent of the source spectral type. The other effect of a large bandwidth is that the image shape will change with wavelength — the 'red' image may be diffraction-limited while the 'blue' image is breaking up into speckles. The resulting integrated image will be an average of the two, however, so the fluctuations in the image shape will not be too large.

Lastly, we can use a larger aperture to determine the tilts on a smaller concentric aperture — Greenwood [35] shows that we can use a tilt measurement aperture perhaps twice as large as the aperture we are trying to correct without the tilts becoming too decorrelated. A larger aperture will however suffer from higher atmospheric noise.

Combining these results, we can see that there will be about 3 times as many photons per exposure in the tilt correction system as in an exposure in the fringe measurement system (calculated on the assumption that the integration time for the

tilt correction system is 3 times longer, that the effective bandwidth is twice as large, that the aperture area used for the tilt correction system is twice as large and that there are 4 telescopes contributing to the measured fringe pattern compared to a single telescope for each tilt correction image).

5.3 Discussion

In the above sections we have shown that we can correct the slope component of the fringe wander with a canonical signal-to-noise ratio (defined in equation 5.2) of about 1 and that we can track the larger-scale motions of the white-light fringe at a similar signal-to-noise ratio. We have also calculated a limiting flux for acceptable tilt correction. We can now ask the question as to which active correction system breaks down first as we go to lower and lower light levels, and what effect this has on the rest of the system.

Clearly the slope correction and envelope tracking systems will cease to operate at similar light levels. If they are used in tandem, then a failure of either one of them will reduce the signal-to-noise ratio and thus rapidly bring about the failure of the other. Both systems depend on the high fringe visibilities brought about through tilt correction and so it is desirable that the tilt correction should fail last of all. To determine if this is so, we can use the value for the flux collection advantage of the tilt-correction system calculated at the end of section 5.2.3, but we must also have a value for the fringe visibility since this affects the fringe measurements and not the tilt measurements. We can see from figure 3.6 that with perfect tilt correction the r.m.s. visibility of a point source when using an aperture diameter of $2r_o$ will be reduced by 32% due to atmospheric perturbations across the aperture. We must reduce this by a further 10% to take account of imperfect tilt correction. The effects of temporal fluctuations are taken into account in our definition of the canonical signal-to-noise ratioand so we need only to decide on values for the coherence function of the source and the visibility loss due to instrumental aberrations. We shall assume here that most of the source flux is unresolved on the maximum baseline we are using (e.g. in an observation of the narrow-line region surrounding an active galactic nucleus), but that the instrumental aberrations cause a 35% loss in visibility (this value is justified in more detail in chapter 9). Hence the visibility observed in a 4-telescope system is about 0.1, where we have used the definition of visibility introduced in chapter 3.

Combining all these results shows that when the tilt-correction system fails due to lack of light, the canonical signal-to-noise ratio in the fringe measurement system is about 0.66. At this level even passive fringe tracking is difficult and thus we can be reasonably confident that the tilt-correction system will be adequate for any source on which we can see fringes. Of course it is possible to revert to the other extreme and

use very narrow bandwidths in the fringe detection system. In this scheme (called an 'absolute' interferometer because the variation in pathlength due to instrumental effects must then be smaller than the atmospheric effects), the fringe signal will eventually rise above the noise if we average for long enough, but the averaging time may prove to be prohibitive: if we use individual spectral channels with a bandwidth of 0.4% of the centre frequency and incoherently average 25 channels (i.e. for a total bandwidth of 10%), the closure phase error on the faintest object for which the tilt correction system will work will be about 60 radians for a single exposure. Thus it would take about 5×10^5 exposures i.e. about 1.4 hours averaging to achieve a closure phase error of 5° for a single set of baselines. We would need therefore need to observe for a total of about 23 hours with this array if we want to make a map with 10×10 pixels. Given that a source is only observable when it is within 30° of the zenith (since we want to minimise the air mass we are looking through), we can only observe most sources for about 4 hours in any one night and so the observation would take about 6 days. For many astrophysical situations (e.g. close binaries, stellar surfaces), the timescale on which the observed source changes may be shorter than this, and this is probably the factor which most limits the absolute interferometer.

Chapter 6

The Design of the Optical Correlator

The optical correlator subsystem lies at the heart of the interferometer. It takes as its input the (suitably path compensated) beams from the M telescopes and produces as its output estimates of the complex visibilities on the $M(M-1)/2$ possible baselines. Such systems have been built before, but only for two-telescope systems [81, 24, 50]. The new requirement for COAST is that the system has to be able to combine the beams from at least 3 telescopes simultaneously so that the closure phase can be determined. In this chapter we shall discuss the overall design of the correlator and attempt to specify the performance required of individual elements in the system.

6.1 Design specifications

We shall assume the following specification for the desired performance of the correlator:

[1] The system must be able to combine the beams from up to 4 telescopes and measure the visibilities on all the possible baselines simultaneously. It is essential that the correlator should work efficiently with only 2 or 3 telescopes when the system is being initially tested and it would be desirable if the system could be easily extended to cope with a larger number of telescopes.

[2] The correlator should be as efficient as possible. That is, losses of light and fringe visibility occurring in the correlator should be kept to a minimum. It should be borne in mind that a 10% loss in fringe visibility has the same effect on the signal-to-noise ratio as a 20% loss in the number of photons collected.

[3]In keeping with the previous criterion, the correlator should be able to use discrete detectors. The motivation for this is the desire to use avalanche photodiodes as detectors in COAST. These detectors offer photon-counting performance and very

good quantum efficiency at long wavelengths ($\sim$ 50% at 800nm), but they must be used as discrete devices. This is because the avalanche process which occurs on the detection of a photon itself generates stray photons, which would wreak havoc if the diodes were not packaged separately. Because of this, the correlator design should try and keep the number of devices needed to a reasonable number and must take account of the need to feed the light into separate detectors, preferably through optical fibres.

Given that discrete detectors will always tend to have better performances than their array counterparts, adopting such a design restriction will have advantages in many situations, for example in infra-red interferometry, and not just in COAST.

[4] The correlator should be able to work with a relatively wide bandwidth, of the order of 10% of the centre wavelength. We have seen in chapter 4 that splitting this bandwidth up into several spectral channels would allow us to track the white-light fringe at very low light levels, but even if this method of tracking is not used, a system with a number of spectral sub-bands has quite a few advantages. Firstly, we are more easily able to tolerate a variation with wavelength in the instrumental response; for example problems due to spectral dispersion are reduced because we can compensate for this in software by recombining the spectral channels with different phase offsets. Secondly, our ability to determine the white-light fringe position at low light levels will be limited to an accuracy of the order of the width of the fringe envelope corresponding to the total bandwidth. If we use channels of smaller bandwidth for determining the fringe visibility then we can reduce the uncertainty due to temporal coherence effects in our visibility calibration. Lastly, in many astrophysical situations it is important to be able to determine the source structure in a particular narrow waveband, e.g. a spectral line. In a system with several channels, one of the channels can be restricted to the desired waveband and the rest be used for tracking the white-light fringe.

6.2 Pairwise or All Together?

In the optical regime, a correlator is simply implemented: we combine the incoming wavefronts to form interference fringes of some sort. However there are many different ways of doing so, and so we must make several design decisions at an early stage or we shall get lost in the wealth of possibilities.

The first decision to be made is whether to determine the $\frac{1}{2}M(M-1)$ visibilities by forming $\frac{1}{2}M(M-1)$ separate fringe patterns by splitting each beam $(M-1)$ ways and combining each of the possible pairs of beams in a different place, or alternatively we can combine all the beams in one fringe pattern such that the interference of each pair of beams produces a fringe at a different spatial or temporal frequency. We can show that the latter scheme is preferable in terms of its efficiency with the following argument.

The signal-to-noise ratio of the mean square amplitude of a fringe at frequency $\mathbf{u}_{ij}$ at low light levels will be given (see equation 2.9) by

$$SNR_A \simeq |V(\mathbf{u}_{ij})|^2 \overline{N}$$

Writing the visibility loss due to loss of spatial coherence and atmospheric and instrumental effects as η_{ij}, the visibility of a fringe in a pattern containing the beams from M telescopes is

$$V(\mathbf{u}_{ij}) = \frac{\eta_{ij}}{M}$$

Assuming for simplicity that the mean number of photons collected per aperture per exposure is 1 then

$$\overline{N} = M$$

and so

$$SNR_A(all\_together) = \frac{\eta_{ij}^2}{M}. \tag{6.1}$$

If the beams are combined pairwise then the mean number of photons in each pattern is

$$\overline{N} = \frac{2}{M-1}$$

and

$$V_{ij}(\mathbf{u}_{ij}) = \frac{1}{2}\eta_{ij}\,,$$

where $V_{ij}(\mathbf{u}_{ij})$ denotes the visibility of the fringe in the fringe pattern consisting of beams i and j. Hence

$$SNR_A(pairwise) = \frac{\eta_{ij}^2}{2(M-1)}. \tag{6.2}$$

Comparison of equations 6.1 and 6.2 shows that combination of all the beams into one pattern is better for the determination of fringe amplitudes at low light levels. Examination of equation 2.9 will convince the reader that this result is in fact valid at all light levels, providing that we can neglect the 'double frequency' noise term.

A similar proof can be constructed to show that combination into a single fringe pattern is also better for the determination of closure phases at low light levels, but extension of this result to high light levels is less straightforward because of the large number of 'double frequency' noise terms.

Thus, at least for low light level work, the single fringe pattern scheme is preferable. It is also preferable in terms of the calibration of the closure phases. In the separate fringe pattern scheme, there will be systematic baseline-dependent phase errors unless we know the relative phase delays between the phase centres of the three fringe patterns that go into the closure phase. These can be calibrated out, for example by using a laser interferometer, but we must beware of thermal/atmospheric drifts in

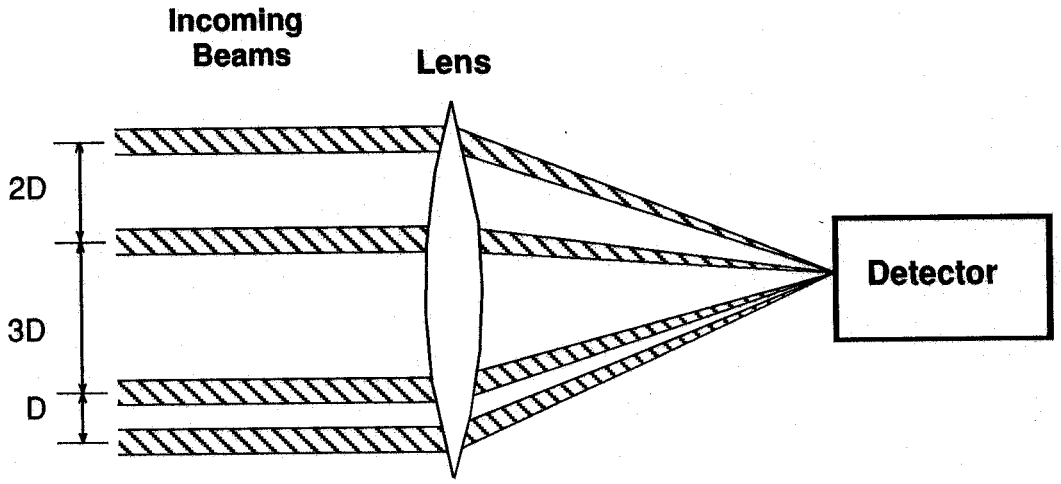

Figure 6.1: *Image plane beam combination.*

the path delays: a change of less than 12nm corresponds to a closure error of 5° and systematic errors of this size can lead to large effects in the map plane [95]. In the single fringe pattern case this problem does not arise, because any path delay affects all the fringes in the same manner.

6.3 Aperture Plane versus Image Plane Combination

There still remain a large number of plausible ways of combining the beams. These can be classified according to whether the beams are interfered in the 'image plane' or the 'pupil plane' (see figures 6.1 and 6.2) and whether the fringe pattern is sampled temporally or whether a spatial fringe pattern is formed.

We can show that these methods are all equivalent from the signal-to-noise ratio point of view in that the amount of corruption by atmospheric fluctuations across the apertures is the same and the resulting signals can all be represented in terms of a notional spatial fringe pattern, thus allowing us to show their equivalence in terms of photon noise. We shall do this by comparing two specific cases: a pupil plane, temporally sampled scheme and an image plane, spatially sampled scheme. The extension of these results to any general case is then straightforward.

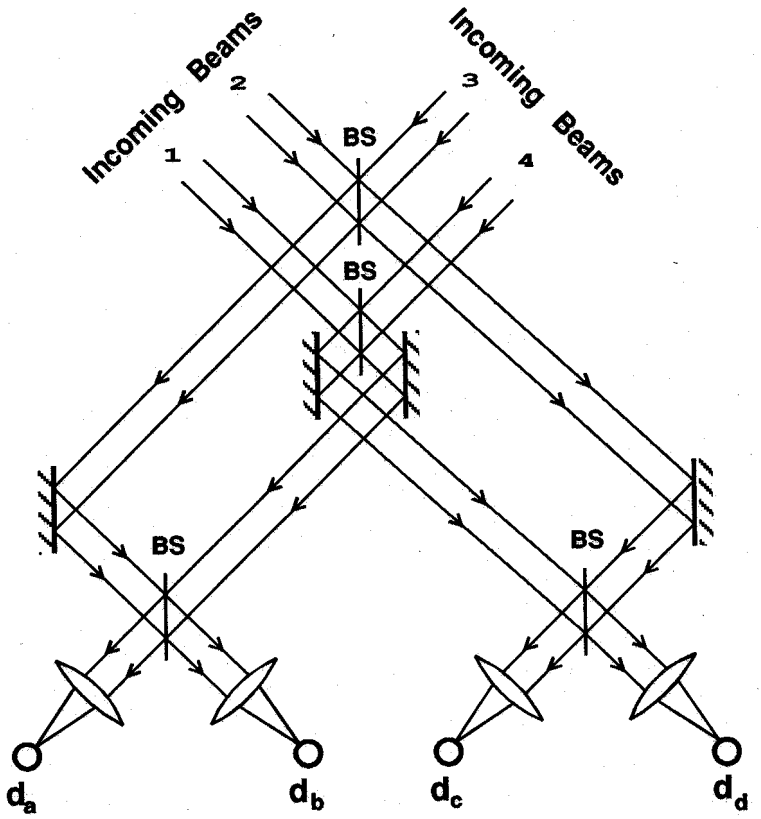

Figure 6.2: *Pupil plane, temporally sampled beam combination (optical path delay scanning mirrors not shown). Legend: BS = beamsplitter; d_a, d_b, d_c, d_d = detectors.*

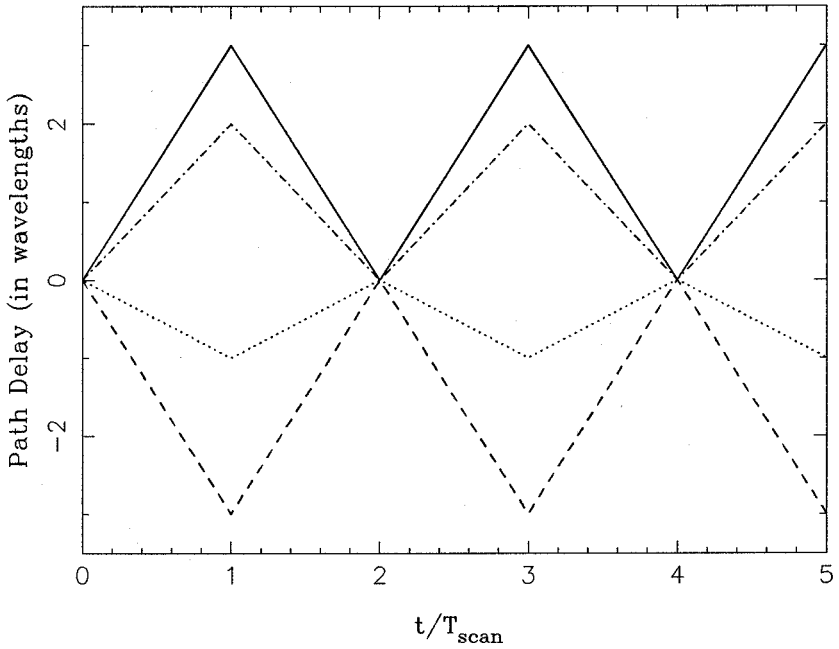

Figure 6.3: *Optical path delay as a function of time for a 4-beam temporally-scanned beam combiner. Each line represents the path delay in one arm. The fringes appear at 1,2,3,4,5 and 6 times the fundamental fringe frequency.*

6.3.1 Image Plane, Spatial Fringe Pattern Scheme

In this scheme, as represented by e.g. the interferometric proposal for the VLT (ESO 1987, chapter 12), the wavefronts from the M telescopes are brought to a common focus on a detector which measures the spatial distribution of the incident radiation intensity, as in figure 6.1. We can write the spatial distribution of the complex amplitude of the radiation field (in a given polarisation state) incident on the focussing lens as

$$P(\mathbf{x}) = E_1(\mathbf{x} - \mathbf{d}_1) + E_2(\mathbf{x} - \mathbf{d}_2) + \ldots + E_M(\mathbf{x} - \mathbf{d}_M)$$

where $\mathbf{x} = (x, y)$ is a vector in the plane of the lens, $E_j(\mathbf{x})$ and is the distribution of amplitude across beam j with $\mathbf{x}$ measured relative to the centre of the relevant beam, and $\mathbf{d}_j$ is the displacement of the beam centre from some origin. The resulting distribution of the classical (high light level) intensity on the detector will then be the Fraunhofer diffraction pattern

$$i(\mathbf{r}) = \left\langle \left| \int\int_{-\infty}^{\infty} \exp\left(2\pi i \mathbf{r}\cdot\mathbf{x}/\lambda f\right) P(\mathbf{x})\, dx\, dy \right|^2 \right\rangle ,$$

where λ is the wavelength of the incident radiation, f is the focal length of the lens and $\langle\rangle$ denotes averaging over a time long compared to the coherence time (i.e. the inverse of the bandwidth) of the radiation. Taking the Fourier Transform of this pattern gives

$$I(\mathbf{u}) = \langle \int\int_{-\infty}^{\infty} P(\mathbf{x}) P^*(\mathbf{x} - \lambda f \mathbf{u})\, dx\, dy \rangle ,$$

which is the sum of the 'fringes' generated by all pairs of points in the input plane of the lens with separation $\lambda f \mathbf{u}$. Thus we expect the maximum fringe signals to be at spatial frequencies $\mathbf{u}_{ij} = \mathbf{d}_{ij}/\lambda f$, where $\mathbf{d}_{ij}$ is the separation vector of two beams, $\mathbf{d}_i - \mathbf{d}_j$. If the beams are combined non-redundantly such that the $\mathbf{d}_{ij}$'s are all different by at least one aperture diameter, then the normalised complex amplitude will be

$$V(\mathbf{u}_{ij}) = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \langle E_i(\mathbf{x}) {E_j}^*(\mathbf{x})\rangle\, dx\, dy}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \langle |E_1(\mathbf{x})|^2\rangle + \langle |E_2(\mathbf{x})|^2\rangle + \cdots + \langle |E_M(\mathbf{x})|^2\rangle\, dx\, dy}. \tag{6.3}$$

If the perturbations to the incident radiation are pure phase perturbations $\phi_j(\mathbf{x})$ then this reduces to

$$V(\mathbf{u}_{ij}) = \frac{\eta_{ij}}{M} \cdot \frac{1}{S} \int\int_{Aperture} \exp\{i[\phi_i(\mathbf{x}) - \phi_j(\mathbf{x})]\}\, dx\, dy ,$$

where S is the area of an aperture and η_{ij} is the value of the spatial coherence function of the observed object on baseline ij.

If the detector does not have infinite spatial resolution and coverage, but rather consists of a number of discrete detectors, we can still model this as a modified fringe

pattern $i'(\mathbf{r})$ falling on an ideal detector. For example if the discrete detectors are squares of dimension w,

$$i'(\mathbf{r}) = [i(\mathbf{r}) * H_w(\mathbf{r})] \times \sum_{k=1}^{n} \delta(\mathbf{r} - \mathbf{r}_k)$$

where * denotes convolution,

$$H_w(\mathbf{r}) = \begin{cases} 1 & 0 < r_x, r_y \leq w \\ 0 & \text{otherwise} \end{cases}$$

and $\{\mathbf{r}_k, k = 1..n\}$ are the positions of the detectors. The resulting 'effective visibility' will thus be attenuated by approximately

$$1 - \frac{1}{6}\left(\frac{\pi w|\mathbf{d}_{ij}|}{\lambda f}\right)^2 \tag{6.4}$$

for $(\pi w|\mathbf{d}_{ij}|/\lambda f) \ll 1$.

6.3.2 Beamsplitter Combination, Temporal Sampling

In this scheme, represented in the special case of a two-telescope interferometer by the Mount Wilson interferometers (Shao and Staelin, 1980), the wavefronts of the parallel beams from the telescopes are superposed using beamsplitters (see figure 6.2) and the total intensity of each of the emergent beams is measured. The intensity of one of these beams, assuming ideal beamsplitters, will be

$$i_a = \frac{1}{4}\left\langle \left| \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [E_1(\mathbf{x})\exp(i\alpha_1) + \cdots + E_M(\mathbf{x})\exp(i\alpha_M)]\, dx\, dy \right|^2 \right\rangle ,$$

where α_j is the phase delay introduced into beam j between the aperture and detector a. This expression can be rewritten as

$$i_a = I_0 + \sum_{j>i} I_{ij}\cos(\alpha_i - \alpha_j + \theta_{ij}) ,$$

where

$$I_0 = \frac{1}{4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \langle|E_1(\mathbf{x})|^2\rangle + \cdots + \langle|E_M(\mathbf{x})|^2\rangle\, dx\, dy ,$$

$$I_{ij} = \frac{1}{2}\left| \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \langle E_i(\mathbf{x}){E_j}^*(\mathbf{x})\rangle\, dx\, dy \right|$$

and

$$\theta_{ij} = \arg\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \langle E_i(\mathbf{x}){E_j}^*(\mathbf{x})\rangle\, dx\, dy \right) .$$

Thus if the optical path delay in each arm of the interferometer is 'scanned' rapidly, using for example a mirror mounted on a piezoelectric actuator, so that the phase delay in one arm during one scan is

$$\alpha_j(t) = 2n_j\pi(t/T_{scan})$$

where n_j is an integer (see figure 6.3), then the output of the detector will be a 'temporal fringe pattern'

$$i_a(t) = I_0 + \sum_{j>i} I_{ij}\cos(2n_{ij}\pi[t/T_{scan}] + \theta_{ij})\,,$$

where $n_{ij} = n_i - n_j$. This fringe pattern will have peaks at frequencies $\nu_{ij} = n_{ij}/T_{scan}$ with visibility

$$V(\nu_{ij}) = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\langle E_i(\mathbf{x})E_j{}^*(\mathbf{x})\rangle\,dx\,dy}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\langle|E_1(\mathbf{x})|^2\rangle + \langle|E_2(\mathbf{x})|^2\rangle + \cdots + \langle|E_M(\mathbf{x})|^2\rangle\,dx\,dy}\,, \tag{6.5}$$

where it has been assumed that all the n_{ij} are different. Comparison with equation 6.3 immediately shows that the two methods have the same sensitivity to atmospheric spatial phase perturbations. If the temporal scanning time is short so that $T_{scan} \ll t_o$ then both methods will also be the same with respect to the temporal fluctuations. Furthermore, the (fixed) differences in the phase delay from a given aperture to the different detectors can be determined experimentally and used to add coherently the complex fringe amplitudes derived from all the detectors, hence yielding the same signal-to-noise ratio as for the image plane case.

Another way to determine the fringe parameters is to 'step' the path delays by multiples of a quarter wavelength: the intensity measurements will then be a set of linear equations in $I_{ij}\cos(\theta_{ij})$ and $I_{ij}\sin(\theta_{ij})$, and the sequence of path length steps can be arranged so that these equations can be inverted to determine the fringe visibilities. We can view these measurements as sampling the temporal fringe pattern at quarter wavelength intervals.

6.3.3 Technical Considerations

The above argument has shown that there is little to choose from a theoretical point of view between the various methods of beam combination. Thus our decision must be based on how easy they are to implement in practice. The main practical difficulty in implementation is our decision to use discrete detectors in the design. Spatial fringe patterns are at an immediate disadvantage here because they require a theoretical minimum of two detectors per baseline in order to be able to determine the two components (i.e. the real and imaginary parts) of the fringe visibility. In practice we

want to gather most of the incoming light and this requires many more detectors. For instance, if we are combining beams in the image plane the minimum possible fringe frequency (corresponding to placing two beams side-by-side in the plane of the lens) will have 2.4 fringe cycles across the main lobe of the Airy pattern. To lose less than 10% fringe visibility because of the finite pixel size (see equation 6.4) we must have 4 detectors per fringe cycle, so that even if we were to discard the light from the Airy sidelobes we would require at least 10 detectors for the lowest fringe frequency. If we have to combine four beams in the same pattern then we must use 6 independent fringe frequencies separated by at least this minimum fringe frequency. Thus we need at least 60 detectors, compared with 4 detectors for a 4-beam temporally sampled scheme. When we add the requirement for, say, 4 spectral channels this translates to 240 detectors compared with 16. The problems with building and reading out such a large number of detectors is clearly formidable, not to mention the difficulty of maintaining relative calibration between them.

The temporally-sampled scheme therefore has a large advantage over spatially-sampled schemes when discrete detectors are being used. However it does have some disadvantages: its successful operation depends on having a fast, accurate scan of the path delay and, because the simplest temporal scanning system uses beamsplitters to combine the light, the beamsplitters may cause problems if their properties vary with wavelength and polarisation, because this would reduce the fringe visibility. In the next sections we shall examine these problems in turn to decide on the technical specifications required for the performance of the system to be acceptable.

6.4 Fringe Scanning

For the temporally-scanned system, there must be some form of rapidly-variable path delay in each incoming beam. Typically this might be in the form of a movable mirror [81] or an elasto-optic cell [74]. While the latter has a good frequency response ($\gtrsim$ 80kHz), the former is preferable because of its high efficiency and the linearity of its motion. However its frequency response is poor - a typical piezo-electric actuator may have its first resonance at 5kHz and loading by a mirror may reduce this. If we require a sawtooth path delay waveform as shown in figure 6.3 with strokes of several wavelengths, the maximum practicable scan frequency may be of the order of a few hundred Hertz.

If the scan time is too long however, changes in the optical path length due to the atmosphere during the scan may cause power to 'leak' from one fringe frequency into another. This not only increases the total noise on a fringe estimate, but will also bias the fringe amplitude measurements. This bias will depend on the amplitude of the fringes at adjacent frequencies and hence will be difficult to 'calibrate out' by

observing reference sources. We must therefore determine the maximum acceptable scan time in terms of the atmospheric time constant t_o.

Let us consider a temporal fringe pattern consisting of fringes which in the absence of perturbations would appear at frequencies of $\{k\omega,\ k = 1 \ldots n\}$ where $\omega = 2\pi/T_{scan}$. The atmosphere will introduce an error $\epsilon_k(t)$ into the linear phase modulation during the scan, giving a normalised intensity pattern

$$i(t) = \sum_{k=0}^{n} \left(V_k e^{i[k\omega t + \epsilon_k(t)]} + V_k^* e^{-i[k\omega t + \epsilon_k(t)]} \right)$$

where V_k is the visibility of the fringe at frequency $k\omega$. Note that $V_0 = 1$. The measured visibility at fringe frequency mw will be

$$I_m = (1/T_{scan}) \int_0^{T_{scan}} i(t) e^{-im\omega t}\, dt$$

we can show (see appendix E) that the mean squared fringe visibility will be

$$\left\langle |I_m|^2 \right\rangle = \sum_{k=1}^{n} |V_k|^2 \left(L_{|k-m|} + L_{|k+m|} \right) ,$$

where L_j is a 'leakage coefficient'

$$L_j = \frac{2}{T_{scan}} \int_0^{T_{scan}} (1 - t/T_{scan}) \exp[-D_\epsilon(t)] \cos[j\omega t]\, dt\,, \tag{6.6}$$

where $D_\epsilon(t)$ is the temporal structure function of the fringe phase perturbations. L_0 is clearly just the loss in mean squared fringe visibility due to the finite integration time. The leakage coefficients for the first four harmonics were calculated by determining the integral in equation 6.6 numerically and are plotted in figure 6.4. From this we can see that if we want less than 1% leakage it is acceptable to have a scan time of $0.4t_o$.

In practice we would often add several successive scans coherently and this will change the relative amount of leakage. If the leaked signal was uncorrelated between scans then we would expect the relative amount of leaked power to fall as the inverse of the number of scans. Unfortunately the scans are very close together in time and so this is not a very good approximation. On the other hand, alternate scans occur in opposite directions and this will cause the part of the leaked signal due to perturbations which are correlated between scans to have the opposite phase relative to the fringe signal on two successive scans, and hence coherent integration will cause part of the leakage signal to cancel out. This intuitive idea can be made more formal by calculating the leakage for a two-scan integration. In this case the measured fringe visibility is

$$\begin{aligned} I_m &= (1/2T_{scan}) \int_0^{T_{scan}} i(t) e^{-im\omega t}\, dt \\ &\quad + (1/2T_{scan}) \int_{T_{scan}}^{2T_{scan}} i^*(t) e^{+im\omega t}\, dt \end{aligned}$$

and we can show (see Appendix E) that the leakage coefficient is now

$$\begin{aligned} L_j &= (1/T_{scan}) \int_0^{T_{scan}} (1 - t/T_{scan}) \exp[-D_\epsilon(t)] \cos[j\omega t]\, dt \\ &\quad + (1/4\pi k T_{scan}) \int_0^{T_{scan}} \{\exp[-D_\epsilon(T_{scan} - t)] + \exp[-D_\epsilon(T_{scan} + t)]\} \times \\ &\qquad \sin[k\omega(T_{scan} - t)]\, dt. \end{aligned} \tag{6.7}$$

These leakage coefficients were evaluated numerically and are presented in figure 6.5; we can see that the relative amount of leakage is indeed reduced by integrating for two scans.

The maximum acceptable scan time can now be determined if we can set a maximum acceptable leakage. This is difficult because this will vary according to the source being observed. For a barely resolved source where all the fringe visibilities are comparable, a power leakage of 1% is acceptable because this will cause only about 1% bias in the fringe visibilities. Hence a scanning time of $0.4t_o$ would be acceptable in this application. However for a heavily resolved source, with one fringe visibility of the order of 1/100th of that at a nearby fringe frequency, the leakage coefficient would need to be 10^{-6} for the same accuracy, since the leaked power is proportional to the *square* of the fringe visibility. The latter case would imply a scanning rate of the order of 1000 scans per t_o, which would be difficult to achieve with mechanical scanners. We can circumvent this problem in a number of ways: for example, the allocation of fringe frequencies can be made in such a way as to keep fringes of widely different amplitudes as far apart in frequency space as possible. Another method would be to determine the the leakage coefficients empirically by measuring the apparent visibilities at frequencies where there is no 'true' fringe power. This is complicated by the presence at a given frequency of number of harmonics generated by the fringes on different baselines; from this point of view it is best to choose a scanning frequency which is high enough such that the leakage coefficient drops rapidly with harmonic number. Alternatively the leakage can be measured with only a single strong fringe signal present, but this will reduce the amount of observing time available.

If we choose a compromise scan rate of say 10 scans per t_o, then the power leakage coefficient for the first harmonic (for a coherent integration of two scans — things should get better if we integrate for longer) is about 4×10^{-4}, which means that nearby fringes can have visibilities different by a factor of 10 and the leakage will still only amount to about a 4% error. This is a significant amount, but small enough that it should not be difficult to remove with suitable calibration.

If t_o is 10ms, then the scan time will be 1ms, or in other words the frequency of the scanning waveform will be 500Hz (since one period of the scanning waveform consists of two scans, one forward and one backward). This scanning frequency is used in

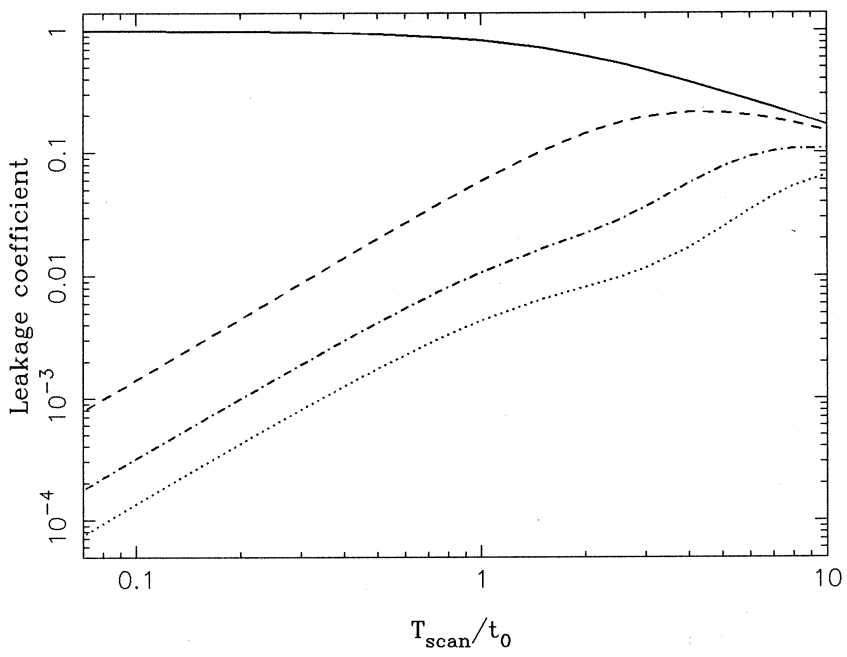

Figure 6.4: *The leakage coefficient L_j (see text) for a temporally-scanned fringe pattern as a function of the scan time, for the case when the coherent integration time is equal to the scan time. The full line is the zeroth harmonic, i.e. the loss in mean square visibility, and the next three harmonics are shown as the dotted lines (the leakage coefficient decreases with increasing harmonic number).*

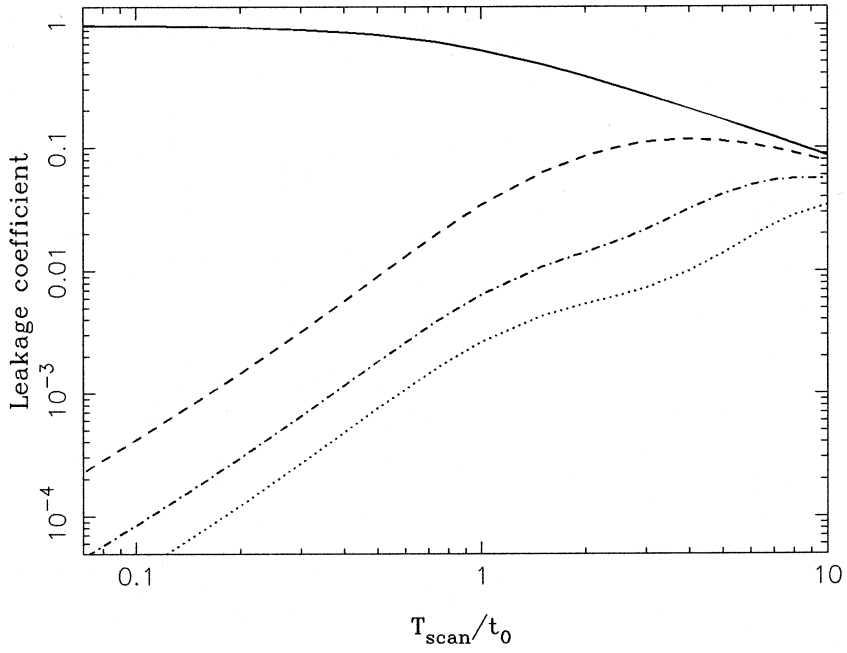

Figure 6.5: *The leakage coefficient as a function of the scan time, for the case when the coherent integration time is twice the scan time, and successive scans are in opposite directions.*

the interferometer at Mt. Wilson [81] and there a scan with very good linearity is achieved using a mirror mounted on piezoelectric actuator. Admittedly, the stroke of the scan in this system is only one wavelength, but it does seem as though scanning at such frequencies should not be too technically demanding.

6.5 Beamsplitter tolerances

If enough care is exercised in the design of the system, the paths of the light beams from the telescopes to the correlator can be made sufficiently symmetric that all the amplitude and phase changes along the path (e.g. due to reflections off mirrors) are the same for all beams. This symmetry is broken if the correlator contains beamsplitters because each outgoing beam will be the sum of one of the incoming beams which has undergone reflection at the beamsplitter and the other incoming beam which has been transmitted through the beamsplitter. If the beamsplitter introduces phase and amplitude changes which are different for transmission and reflection then the visibility of the observed fringes may be reduced. We shall now calculate how much variation in beamsplitter response can be tolerated.

We shall assume that the beamsplitters used are made from dielectric films and hence are lossless, that they are symmetric in that the reflection coefficient for a beam incident from one side of the beamsplitter is the same in both amplitude and phase as that for a beam incident from the other side of the beamsplitter (this can be achieved by sandwiching a symmetric semi-reflective film between two pieces of glass) and that all the beamsplitters in the system are identical. From energy conservation arguments we can then show that the transmission and reflection coefficients, t and r, are related by

$$|t|^2 + |r|^2 = 1.$$

If we now consider two coherent beams incident on opposite sides of the beamsplitter, the interference of the beams will produce output beams with powers

$$\begin{aligned} P_1 &= 1 + Re(rt^* e^{i\theta}) \\ P_2 &= 1 + Re(r^* t e^{i\theta}) \end{aligned}$$

where the power in each of the input beams is taken as being unity and where θ is the phase difference between the incoming beams. For energy to be conserved for all θ, we have

$$rt^* = -r^* t. \tag{6.8}$$

Hence we can write these coefficients in the form

$$r = [(1-\delta)/2]^{1/2} e^{i\phi} \tag{6.9}$$

$$t = [(1+\delta)/2]^{1/2} e^{i(\phi+\pi/2)} \tag{6.10}$$

where δ is the fractional deviation (assumed small) of the beamsplitter power transmission coefficient from the ideal value of 50%.

It is easy to see that the visibility modulus, and hence the signal-to-noise ratio is only affected by the moduli of r and t and not by their phases. If however these phases change with wavelength or polarisation, then the fringes at different wavelengths or polarisations will combine to form a 'blurred' fringe with reduced modulus. We shall consider the amplitude and phase effects separately and restrict our attention to the four-beam correlator shown in figure 6.2.

Table 6.1 shows the complex visibilities of the fringes on the six baselines in each of the four detectors and table 6.2 shows the resulting visibility losses due to an incorrect amplitude reflection coefficient. We can see that the worst loss is of order $(1-\delta)$ (e.g. baseline 1–3 at detector a). If however the fringe measurements from different detectors can be combined coherently (e.g. if we can combine the signals for baseline 1–3 from detectors a and c) then the visibility loss on this baseline can be reduced to about $(1-\delta^2/2)$. The worst visibility loss is now $(1-\delta^2)$ (e.g. on baseline 1–2) which means that we can tolerate beamsplitter power transmission coefficients in the range 45% to 55% if we require less than 1% visibility loss.

In order to be able to coherently combine the signals from different detectors, we must know the systematic phase differences between the fringes seen by different detectors — combining the complex fringe visibilities with a systematic phase error of $\Delta\phi$ will degrade the resulting visibility by approximately $(1-\Delta\phi^2/2)$. Thus if we require a visibility loss of less than 1% due to this effect we need to know the relative path delays inside the correlator to better than 0.14 radian, i.e. $\lambda/45$. Whether this precision can be achieved depends on how fast the relative delay changes in time (e.g. due to thermal drifts), since the longer this takes, the more time is available in which to calibrate the system. Because of the symmetry of the design, only temperature *differences* across the correlator will change the relative path delays. If the correlator is about 1 m across and situated on an optical table with a thermal expansion coefficient of $10^{-5}/^{\circ}$C, and $\lambda = 800$nm, we can tolerate a net *change in temperature difference* between the two arms of $(4\times10^{-3})^{\circ}$C. This involves the change in temperature of large masses of metal and so we might hope for the timescale for such a change to be of the order of a few minutes. Over this timescale, averaging the differences between the fringe phases measured at different detectors would allow the path delays to be determined while a source is being observed, providing that the signal-to-noise ratio per exposure is greater than about 1 on at least one baseline.

In the case of the phase part of the reflection and transmission coefficients, it has already been emphasised that it is only variations in the fringe phase as a function of wavelength or polarisation which affect the fringe visibility modulus. Examination of table 6.1 and equations 6.9 and 6.10 will convince the reader that in fact the phase

	a	b	c	d
1–2	$tt(rr)^*$	$tr(rt)^*$	$rr(tt)^*$	$rt(tr)^*$
1–3	$tt(tr)^*$	$tr(tt)^*$	$rr(rt)^*$	$rt(rr)^*$
1–4	$tt(rt)^*$	$tr(rr)^*$	$rr(tr)^*$	$rt(tt)^*$
2–3	$rr(tr)^*$	$rt(tt)^*$	$tt(rt)^*$	$tr(rr)^*$
2–4	$rr(rt)^*$	$rt(rr)^*$	$tt(tr)^*$	$tr(tt)^*$
3–4	$tr(rt)^*$	$tt(rr)^*$	$rt(tr)^*$	$rr(tt)^*$

Table 6.1: *The complex transfer coefficients for the fringes on each baseline (1–2, 1–3 etc.) at each of the four detectors (a, b, c, d) in the temporally-sampled scheme shown in figure 6.2. The symbols t and r are the complex transmission and reflection coefficients of the beamsplitters.*

	a	b	c	d
1–2	$(1-\delta^2)/4$	$(1-\delta^2)/4$	$(1-\delta^2)/4$	$(1-\delta^2)/4$
1–3	$(1-\delta)(1-\delta^2)^{1/2}/4$	$(1-\delta)(1-\delta^2)^{1/2}/4$	$(1+\delta)(1-\delta^2)^{1/2}/4$	$(1+\delta)(1-\delta^2)^{1/2}/4$
1–4	$(1-\delta)(1-\delta^2)^{1/2}/4$	$(1+\delta)(1-\delta^2)^{1/2}/4$	$(1+\delta)(1-\delta^2)^{1/2}/4$	$(1-\delta)(1-\delta^2)^{1/2}/4$
2–3	$(1+\delta)(1-\delta^2)^{1/2}/4$	$(1-\delta)(1-\delta^2)^{1/2}/4$	$(1-\delta)(1-\delta^2)^{1/2}/4$	$(1+\delta)(1-\delta^2)^{1/2}/4$
2–4	$(1+\delta)(1-\delta^2)^{1/2}/4$	$(1+\delta)(1-\delta^2)^{1/2}/4$	$(1-\delta)(1-\delta^2)^{1/2}/4$	$(1-\delta)(1-\delta^2)^{1/2}/4$
3–4	$(1-\delta^2)/4$	$(1-\delta^2)/4$	$(1-\delta^2)/4$	$(1-\delta^2)/4$

Table 6.2: *The amplitude transfer coefficients for the fringes on each baseline at each of the four detectors in the temporally-sampled scheme shown in figure 6.2. See equations 6.9 and 6.10 for the definition of δ.*

	a	b	c	d
1–2	$\phi_{t_L}+\phi_{t_R}-2\phi_{r_L}$	$\phi_{r_R}-\phi_{r_L}$	$2\phi_{r_L}-\phi_{t_L}-\phi_{t_R}$	$\phi_{r_L}+\phi_{r_R}$
1–3	$\phi_{t_L}-\phi_{r_L}$	$\phi_{r_R}-\phi_{t_R}$	$2\phi_{r_L}-\phi_{r_R}-\phi_{t_R}$	$\phi_{r_L}+\phi_{t_L}-2\phi_{r_R}$
1–4	$\phi_{r_L}-\phi_{t_R}$	$\phi_{t_L}-\phi_{r_R}$	$2\phi_{r_L}-\phi_{t_R}-\phi_{t_L}$	$\phi_{r_L}-\phi_{t_R}$
2–3	$\phi_{r_L}-\phi_{t_R}$	$\phi_{r_L}-\phi_{t_R}$	$\phi_{t_L}-\phi_{r_R}$	$\phi_{t_L}-\phi_{r_R}$
2–4	$2\phi_{r_L}-\phi_{r_R}-\phi_{t_R}$	$\phi_{r_L}+\phi_{t_L}-2\phi_{r_R}$	$\phi_{t_L}-\phi_{r_L}$	$\phi_{r_R}-\phi_{t_R}$
3–4	$\phi_{r_L}-\phi_{r_R}$	$\phi_{t_R}+\phi_{t_L}-\phi_{r_R}-rl$	$\phi_{r_R}-\phi_{r_L}$	$2\phi_{r_R}-\phi_{t_R}-\phi_{t_L}$

Table 6.3: *The phase transfer coefficients for the fringes on each baseline at each of the four detectors for systems employing asymmetric beamsplitters. The phase transmission and reflection phase shifts are ϕ_{t_L} and ϕ_{r_L} for beams entering from the left hand side of the beamsplitters and ϕ_{t_R} and ϕ_{r_R} for beams entering from the right hand side. Equation 6.8 means that we have the relationship $\phi_{r_L}+\phi_{r_R}-\phi_{t_L}-\phi_{t_R}=\pi$ for lossless beamsplitters.*

of the fringes at the four detectors is independent of ϕ, the beamsplitter phase shift, because the phase shifts in different beams cancel each other out. This property comes about because of the symmetry of the beamsplitter construction and so we need worry about beamsplitter phase constraints only if they are asymmetric. For the latter case we can calculate the allowable fringe phase shifts as a function of wavelength and polarisation; table 6.3 then shows how the fringe phase shifts depend on the phases of the beamsplitter transmission and reflection coefficients.

If we assume that the fringe phase shift with wavelength is approximately linear across the spectral passband, then the allowable phase change is

$$[\phi(\lambda + \Delta\lambda) - \phi(\lambda)]^2 \lesssim 6\delta V/V$$

where $\delta V/V$ is the allowable fringe visibility loss. For the phase shift between polarisations we have

$$[\phi(s) - \phi(p)]^2 \lesssim 2\delta V/V$$

where s and p represent parallel and perpendicular polarisations. Hence for a visibility loss of less than 1% in each case, we require fringe shifts of less than 0.24 radians across the passband and 0.14 radians between polarisations.

Having determined the required tolerances, it is then a matter of seeing if it possible to attain this performance in practice. Beamsplitters can be found in optical catalogues with the required amplitude splitting ratio for both polarisations over small passbands [59], but the phase properties of the beamsplitters are not quoted. Thus it may be necessary to have custom beamsplitters made. The phase tolerances can then be attained by constructing symmetric beamsplitters, but achieving an amplitude response which is acceptable for both polarisations may be more difficult. We can minimise the magnitude of the problem by reducing the angle of incidence of the beams at the beamsplitters (i.e. by 'concertina-ing' the design in figure 6.2) since the difference between parallel and perpendicular polarisations will always disappear near normal incidence. The drawback of such an approach is that the size of the correlator will increase as the angle of incidence is decreased, and this will worsen the temperature stability of the system.

6.6 Fibre Optics

The design in figure 6.2 could be built using monomode optical fibres instead of air paths and fibre couplers instead of beamsplitters (see [79]). Among the advantages of such a scheme are: (1) no internal alignment is needed beyond launching the incoming beams into the fibres; (2) the whole correlator can be a few centimetres in size rather than requiring a large optical table and (3) monomode fibre couplers with excellent bandwidth and polarisation properties are readily available.

These must however be balanced against the disadvantages. Firstly, there may be large losses associated with launching light beams into fibres. A survey of the available literature seems to indicate that it is hard to get much more than 50% of the theoretical coupling efficiency [89, 99]. Secondly, the coupling optics must be optimised for a particular r_o and are difficult to change quickly if we wish to adapt to changing seeing conditions. However, the loss in signal-to-noise ratio due to lack of optimisation may not be that large: we can use the results of Shaklan & Roddier [80, figure 5] to show that using optics optimised for 2 arcsecond seeing under 1 arcsecond conditions is only 30% worse in terms of the mean coupled power than using a system optimised for 1 arcsecond conditions.

Lastly there appears to be no simple and efficient way of spectrally dispersing the fringes within a fibre-optical system. Wavelength-division demultiplexers are available, but designed for separating light at widely differing wavelengths (e.g. 1300nm and 1500nm). To get the closely spaced spectral channels required for fringe tracking one might have to decouple the light from the fibre, disperse it with a prism and then recouple the resulting beams into the multimode fibres which feed the detectors. This clearly reduces the advantage of compactness and simplicity offered by fibre optics.

6.7 Conclusions

The work in this chapter has provided at least part of the theoretical framework to the overall design of the correlator. It has indicated the types of design that are promising and the necessary performance that must be achieved. The major part of the correlator design, however, will depend on experimental tests of the various components. In particular a comparison of the performances of prototype systems using fibre optics and using beamsplitters will be needed to resolve the questions of practical performance which cannot be decided from pure theory.

Chapter 7

A Prototype Delay Line

As explained in the introduction, any astronomical interferometer whose collecting elements are fixed to the Earth's surface must incorporate a variable internal optical path to compensate for the geometric delay in the incoming light beams due to Earth rotation. A further use of such a delay line is to compensate for any random variations in the lengths of paths internal to the instrument due to e.g. thermal expansion or mechanical vibrations.

The specification of such a delay is very demanding. While observing, the delay must be moved sufficiently accurately to keep within the 'fringe envelope' - about 8μm wide for a system observing at 800nm with a 10% fractional bandwidth. However it need not have an absolute accuracy as good as this because the atmospheric phase perturbations will make the white light fringe position uncertain by more than 60μm (calculated for seeing conditions characterised by $r_o = 10$cm at $\lambda_0 = 500$nm and for a baseline of 100m) and so all that is needed is sufficient absolute accuracy to start off the search for the fringe envelope in the right area. A more stringent limitation is the allowable *variation* of the delay line position error during the exposure time, since such motions will blur the astronomical fringes. If the delay is subject to errors whose variance during the exposure time is δ^2 (assumed to be small) the r.m.s. fringe visibilities will be reduced by

$$1 - \frac{1}{2}\left(\frac{2\pi\delta}{\lambda}\right)^2 ,$$

so that for a loss of less than 5%,

$$\delta < \lambda/20 = 40\text{nm}$$

for $\lambda = 800$nm. If the errors are in the form of a drift whose velocity is approximately constant over the exposure period (e.g. thermal expansion, servo velocity errors) then

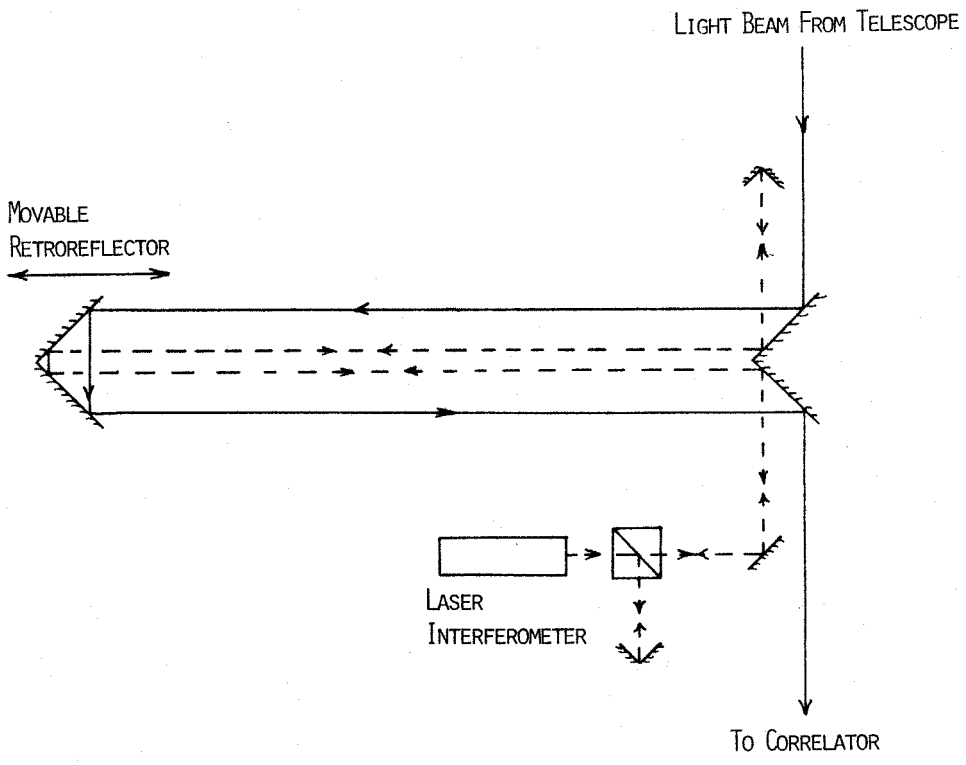

Figure 7.1: *Diagram of a possible delay line configuration*

the r.m.s. visibilities will fall by

$$1 - \frac{1}{6}\left(\frac{2\pi\tau}{\lambda}\right)^2 \langle v^2 \rangle \,,$$

where τ is the exposure time, $\langle v^2 \rangle$ is the mean square velocity error and it has been assumed that this loss is small. For less than 5% loss in fringe visiblity, we therefore require that

$$\langle v^2 \rangle < (0.087\lambda/\tau)^2 = (7.0\mu\mathrm{m/s})^2 \,,$$

for $\lambda = 800$nm, $\tau = 10$ms. In comparison with these accuracies, the total range of the delay line is huge: for an element spacing of 100m, the delay must be varied by 50m in order to observe at zenith angles of up to 30°. With an East-West baseline and observing an object at the meridian, the delay must move smoothly at 7.2mm/s, and in order to slew between source and reference objects it may have to move at 1m/s, though it need not move so smoothly.

Thus the delay line must have a positional accuracy of 4 parts in 10^{10} over short periods and better than 1 part in 10^6 absolute accuracy. Furthermore, it must have a velocity error of less than 1 part in 10^3.

It is proposed that such a delay line can be built as an 'optical trombone' (see figure 7.1) consisting of a movable mirror on a trolley running along a 25m long track.

The required accuracy can be achieved with an active system which continuously corrects the mirror position using measurements from a laser interferometer, one of whose arms runs parallel to the astronomical light beam for the length of the delay. It was decided to build a prototype of such a system in order to study the problems of attaining such high accuracy.

7.1 The Laser

The accuracy of the delay line is ultimately limited by the accuracy with which the optical path can be measured, which in this case is limited by the wavelength stability of the light emitted by the laser. The light emitted by a laser consists of one or more narrow lines or 'modes' whose wavelength is determined by the spacing between the mirrors at the ends of the laser cavity. Thermal expansion and mechanical stresses cause this separation to change with time so that the modes drift across the 'gain profile' of the lasing material, about 1GHz in width for a He-Ne laser operating at 633nm. Thus the long term stability of the wavelength of even a single mode laser is only about 2 parts in 10^6 and it is therefore necessary to provide active stabilisation of the laser cavity length.

7.1.1 Stabilisation Method

Many methods have been proposed for stabilisation of He-Ne lasers [8, 5, 93], but it was decided to try using transverse Zeeman stabilisation [62] because of its inherent resistance to optical feedback [10], which is caused when a small amount of the laser light (as little as 10^{-7} of the output) is scattered back into the laser cavity, causing amplitude and frequency changes in the emitted light.

Transverse Zeeman stabilisation makes use of the complex phenomena which occur when a magnetic field is applied transverse to the axis of a He-Ne laser cavity [28]. Space does not permit detailed explanation of these phenomena but the result is that if a certain characteristic magnetic field is applied to a two-mode He-Ne internal mirror laser, the modes collapse into a single mode which consists of two orthogonal linearly polarised 'sub-modes' with a small ($\sim$ 50kHz) frequency difference between them. This frequency difference varies with the position of the mode in the gain profile of the lasing material, and can be measured as the 'beat frequency' observed when the two polarisations are mixed with a linear polariser. Hence a negative feedback servo can be implemented to correct changes in the laser wavelength by, for instance, heating the glass envelope of the laser so that thermal expansion alters the end mirror spacing.

7.1.2 Implementation

The laser used for these experiments was a Spectra-Physics model 133, which has a gas ballast tank which is separate from the lasing tube. This means that the lasing tube is quicker to heat, but has less pointing stability. A coil of bifilar (to avoid generating stray magnetic fields) resistance wire was round round the tube to provide electrical heating. The transverse magnetic field was provided by two rows of permanent magnets glued to the laser cover so that they straddled the tube. The field was adjusted by adding or subtracting a number of thin steel strips acting as pole-pieces. The light from the back end of the laser is passed through a polariser aligned at 45° to the magnetic field and detected with a photodiode. The beat frequency signal from the photodiode is amplified, filtered and passed to a frequency-to-voltage converter. The resulting voltage is compared to a reference voltage and the difference used to determine the current to a heating coil. An integrator was incorporated into the error amplifier so as to reduce the effect of long-term ambient temperature changes.

7.1.3 Results

It was found that, after some adjustment of the magnetic field, the two modes of the laser did collapse to a single mode, as evidenced by measuring the strength of the intermode beat at 550MHz. The coherence length of the laser light was thus increased from one laser cavity length (27cm) to an estimated value of 7km, set by the frequency difference of the sub-modes.

The range of beat intra-mode beat frequencies as the single mode was tuned across the gain profile was about 0 to 70kHz, but it was found that the variation of beat frequency with mode frequency was not the smooth form predicted and measured by Ferguson and Morris, but was instead an irregular variation (see plate 1) which depended on the detailed arrangement of the pole pieces. Umeda *et al.* [93] have reported a similar disagreement with Ferguson and Morris' results, though their results were less chaotic in nature, and have suggested that this was due to their having used a different isotopic mixture for the lasing gases. However it is suspected in our case that the these irregularities are due to an inhomogenous magnetic field — for instance, it was found that the steel clip holding the lasing tube had become magnetised.

Even so, it was found that there was still a usable region in the beat frequency *vs.* mode frequency graph (indicated in plate 1) that could be used for locking the laser frequency. The servo was therefore adjusted to stabilise in this region. It was found that the system achieved equilibrium very rapidly (in a few seconds) once the laser envelope had been heated above ambient temperature by an amount sufficient to make the available heating and cooling rates approximately equal. This lock was

found to be very stable in the long term, but again it was necessary to raise the equilibrium tube temperature adequately above the range of ambient temperature variations, to about 60° C.

The frequency stability of the laser was measured by observing the changes in the intra-mode beat frequency and using the measured slope of the beat frequency *vs.* mode frequency curve to estimate the corresponding mode frequency change. This of course assumes that no other effects (changing magnetic fields, changing plasma conditions) affect this curve over the period of the observation. It was found that in the short to medium term (1 to 30 seconds) the laser frequency stability in laboratory conditions was better than 200kHz (4 parts in 10^4) and that the long term (overnight) stability was better than 5MHz (1 part in 10^8). However on very short timescales ($\sim$ 10ms) the beat frequency displayed significant periodic fluctuations implying laser frequency oscillations of about 1 part in 10^8. This was found to be due to mains frequency ripple in the plasma discharge current, which may not actually be affecting the end mirror separation, but rather affecting the state of the plasma discharge.

The beat frequency between this laser and another laser stabilised by a different method[1] was observed in order to gain independent confirmation of these results, but it was found that the variations in frequency of this second laser (as evidenced by the variations of the beat frequency between two identical lasers of this sort) were too large (about 1 part in 10^8 on timescales of a few minutes) to set any stringent limits on the stability of the Zeeman-stabilised laser.

All these results were taken when the effects of optical feedback had been reduced to a minimum — placing a piece of black plastic foam on the white wall 2 metres away where the main laser beam was pointing increased the laser stability by a factor of 10.

An annoying property of the Zeeman laser that was noticed was the fact that the output beam contains a ripple component at the beat frequency even when the polarisations had not been mixed with a polariser. This may be due to non-linearities in the properties of the lasing plasma and also possibly inhomogeneities in the magnetic field. This ripple, which is about 10% of the mean intensity, reduces the accuracy of obtainable on short-timescale intensity measurements (e.g. in measuring fringe intensities in an interferometer) and blurs out oscilloscope traces when parameters other than the beat frequency are being measured. Filtering this ripple from the detected intensity signal is impractical because it is intended to measure fringes moving at comparable frequencies.

[1]This was a He-Ne laser stabilised by the method of Bennet et al kindly lent by Dr. Bob Butcher

7.1.4 Conclusions

The wavelength stability of the current laser on the critical timescale around 10ms (the exposure time for the astronomical fringes) is clearly not adequate for the final instrument, but it will provide sub-micron accuracy over paths of many tens of metres and as such is a useful test instrument.

The stability could be improved primarily by using a laser with a stabilised power supply. Improving the homogeneity of the magnetic field would facilitate setting up the servo and might reduce the amount of ripple in the output beam. There may be some advantage in adapting the system for a green He-Ne laser, since this means that the laser wavelength would be further from the observing wavelength proposed for COAST.

Given the above-mentioned experience with the piece of black foam, it is clear that even with Zeeman stabilisation the utmost care must be taken over avoiding optical feedback if the maximum frequency stability is to be achieved.

Finally it is interesting to note that even the worst frequency variations reported here correspond to a change in the average temperature of the glass envelope of $(10^{-3})^\circ$ C, an indication of the difficulty of achieving these stabilities by passive means.

7.2 The Interferometric Path Length Measurement System

By making the delay line part of the optical path of one arm of a Michelson interferometer (see figure 7.1), the change in length of the optical delay can be measured in units of half the wavelength of the light source by electronically counting the number of fringes that pass a point in the interference pattern. The fractional part of the path length change in these units can be derived by measuring the intensity of the light at this point and using a knowledge of the variation of the fringe intensity over a cycle. Absolute delay measurements can be made by moving the delay to a known reference point and then keeping track of the change in delay thereafter.

7.2.1 Implementation

This scheme was implemented with the stabilised laser as a light source and with the modification suggested by Peck [70] of replacing the plane mirrors at the end of the interferometer arms with retroreflectors (see figure 7.2), with two resulting advantages: (i) The retroreflectors always return the beams parallel to the direction they arrive in, irrespective of the alignment of the retroreflectors and so the interferometer

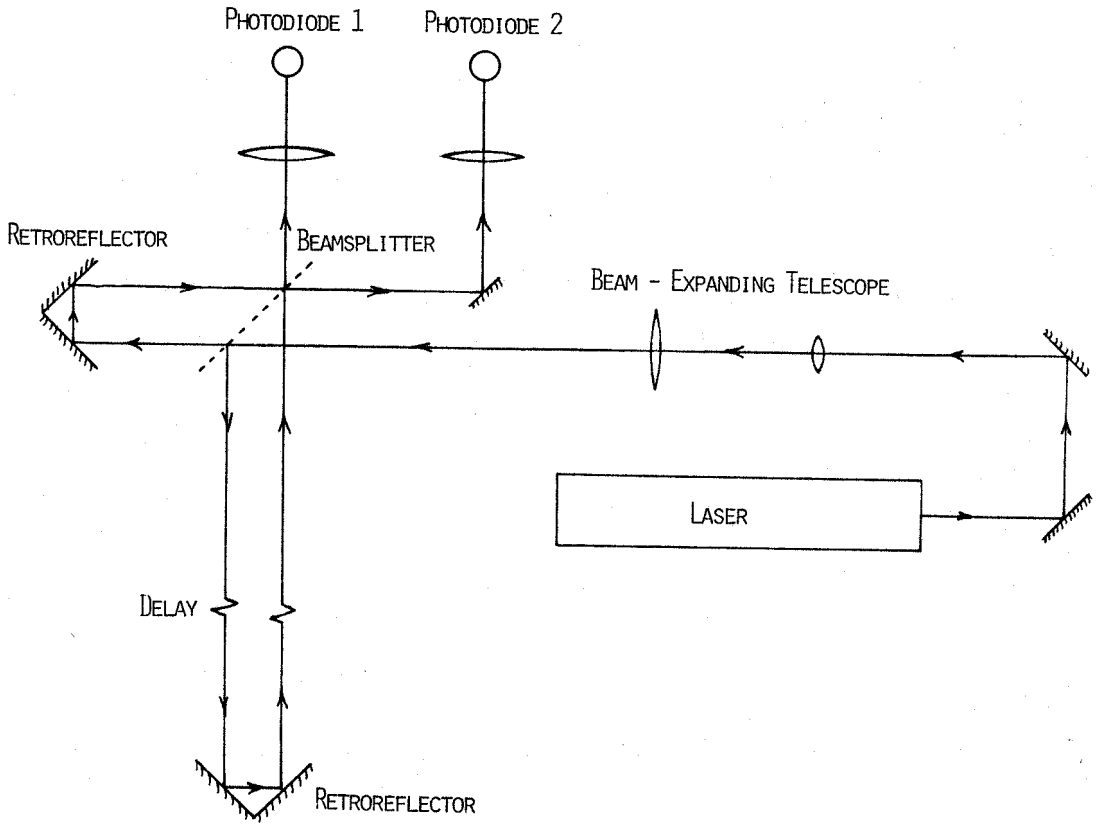

Figure 7.2: *The layout of the distance-measuring interferometer*

can be aligned in minutes rather than hours; (ii) The returned beams are displaced from one another so that two complementary fringe patterns can be observed (see figure 7.2). The beamsplitter coating is a layer of aluminium of the correct thickness [91, appendix 4.1] such that the fringe phase at any point in one of the fringe patterns is in quadrature (approximately) with the phase at the corresponding point in the other fringe pattern, irrespective of the relative lengths of the two arms of the interferometer. This can be used to provide reversible fringe counting. The laser beam is expanded to a width of 1cm so that it can propagate over distances of up to 100m without diffraction effects becoming important. The system was aligned so that the fringe patterns were as nearly uniform as possible (i.e. contained less than one fringe) and the total intensity of each pattern was focussed on a fast photodiode.

A block diagram of the electronic system used to count the fringes is shown in figure 7.3. The fringe counting logic uses the quadrature signals from the photodiodes to derive four counting pulses per fringe with the ability to discriminate between forward and reverse motion[2]. Thus the basic resolution of this system is 633nm/8 $\approx$ 80nm, which is sufficient for simple trials of the delay line servo system. The counting pulses are fed to a 24-bit counter whose output is subtracted from the output of another counter fed by a demand signal — eventually this signal will come from a microcomputer. The lowest eight bits of the result of the subtraction are fed to a digital-to-analogue converter to provide an analogue signal indicating the error between the desired and actual fringe counts. The more significant bits of the digital count are used to derive sign and overflow information so that the system can keep track of errors of up to $\pm$1.3m and maintain a sensible analogue error signal, as indicated in figure 7.4.

7.2.2 Results and Further Improvements

It was found that the high analogue bandwidth in the photodiode signal processing electronics required for slewing the delay line (3MHz for a slew rate of 1m/s) combined with the requirement of high accuracy at low rates of fringe motion (perhaps 1Hz when observing Polaris) was the most difficult specification to achieve. It was found that even with preamplifiers sited next to the photodiodes, significant phase shifts began to occur at fringe rates of 1MHz. It is proposed to replace the photodiode-preamplifier-cable combination with an optical fibre feeding a photodiode on the main circuit board and thus overcome the problem of cable capacitance reducing the bandwidth.

The other major problem is setting up the zero-levels for the digitisation of the fringe signal. This has to be made self-adjusting so that it can compensate for changes in the total intensity and contrast of the fringes due to laser power changes, beam

[2]For details of the implementation of such a logic circuit, see for example [44]

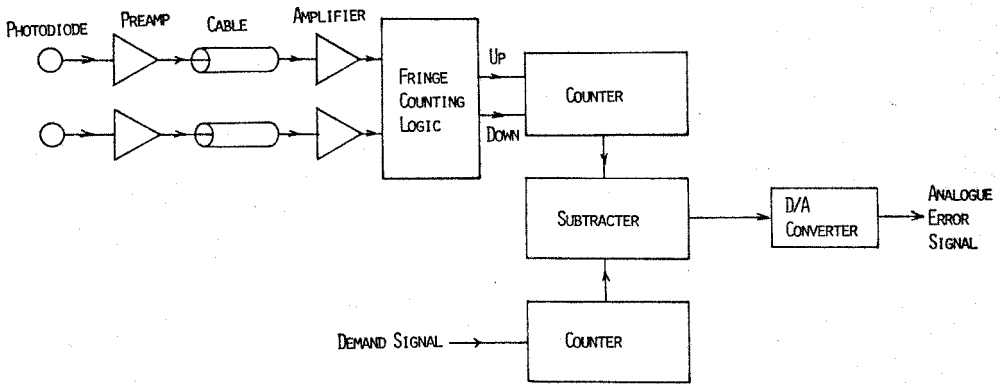

Figure 7.3: *Block diagram of the fringe counting electronics*

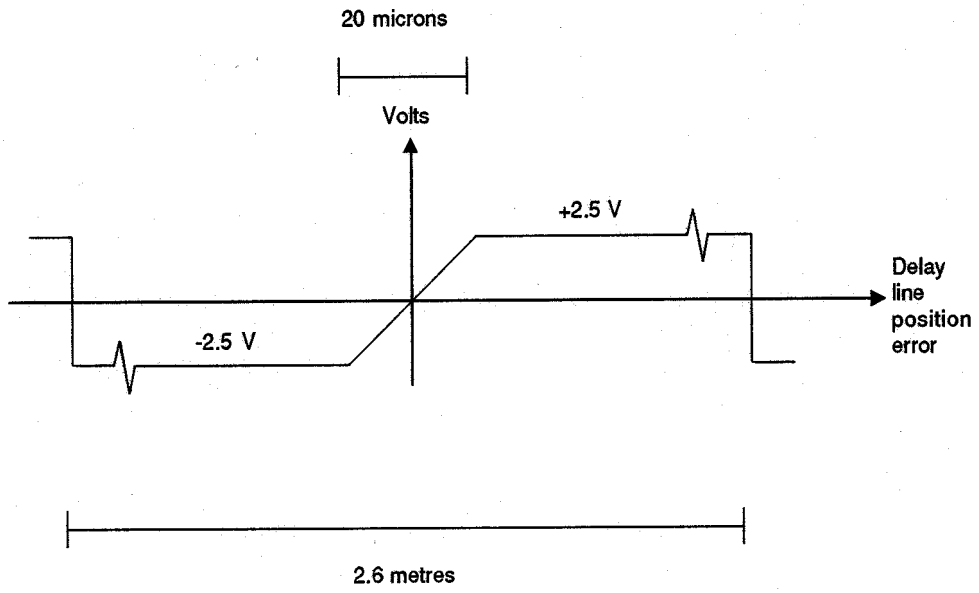

Figure 7.4: *Graph of the error signal generated by the interferometer electronics*

alignment changes and so on.

7.3 The Delay Line

The interferometer described above was incorporated into a prototype delay line. The actuating element, designed and built by Dr. Donald Wilson, consists of a corner cube retroreflector mounted on a trolley driven by electric motor along a length of aluminium track (see plate 3). The corner cube is isolated from high frequency vibrations of the trolley and the rest of the environment by a parallelogram suspension of low resonant frequency (about 5Hz). The position servo is implemented as a two-stage system: high frequency errors indicated by the laser interferometer are corrected by a moving magnet actuator attached to the corner cube and low frequency large displacements are provided by motion of the trolley, which is 'slaved' to the position of the corner cube so that the displacement of the corner cube from the equilibrium position of the suspension, as measured by a stabilised photosensor (see plate 3), is minimised.

7.3.1 Results

It was found that, after adjustment of the gain and damping parameters of the servo system, it was possible to stabilise the position of the corner cube to $\pm 1/8$th of a wavelength, i.e. the resolution of the interferometer, even in normal laboratory conditions with a workshop 5 metres away. However, when a demand signal was applied to translate the delay by about 1mm/s the position error increased to about ± 2 wavelengths due to the jerky motion of the trolley. Increasing the gain of the actuator servo in order to counteract this resulted in oscillations because of a resonance in the structure supporting the drive coil which occurred at about 70Hz. Increasing the gain of the trolley servo so as to decrease the 'jerkiness' of the motion led to an interaction of the trolley and solenoid servos resulting again in oscillations.

7.3.2 Conclusions

The above experiments show that even a crude mechanical system (built of meccano!) can be made to achieve interferometric precision in environments which are not isolated from external vibrations. To achieve the accuracy required for astronomical interferometry, the trolley will have to be redesigned with particular care being taken over structural resonances. It will also be necessary to make a more detailed analysis of the interactions of the coupled servo system, and to improve the interferometer by digitising the fringes more finely.

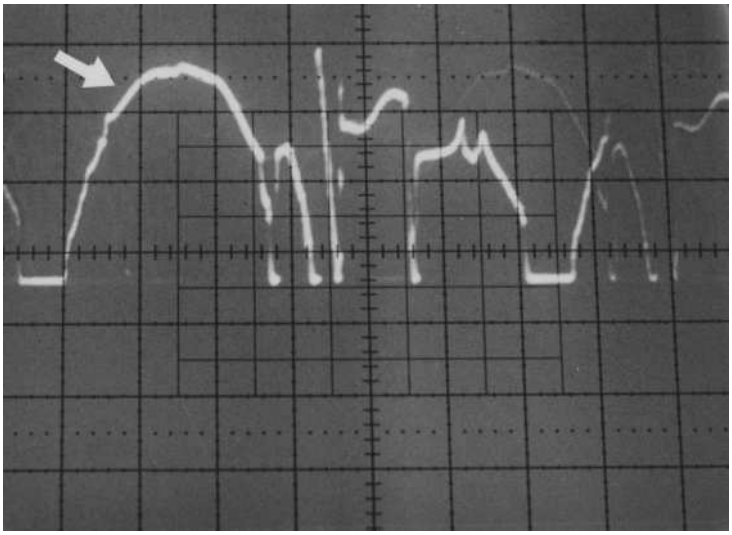

Plate 1: *The variation of beat frequency as the mode frequency is tuned across the gain profile. The vertical scale is approximately 21kHz/division and the width of the gain profile is approximately 8 horizontal divisions. The operating region of the servo is indicated by the arrow and is near the centre of the gain profile*

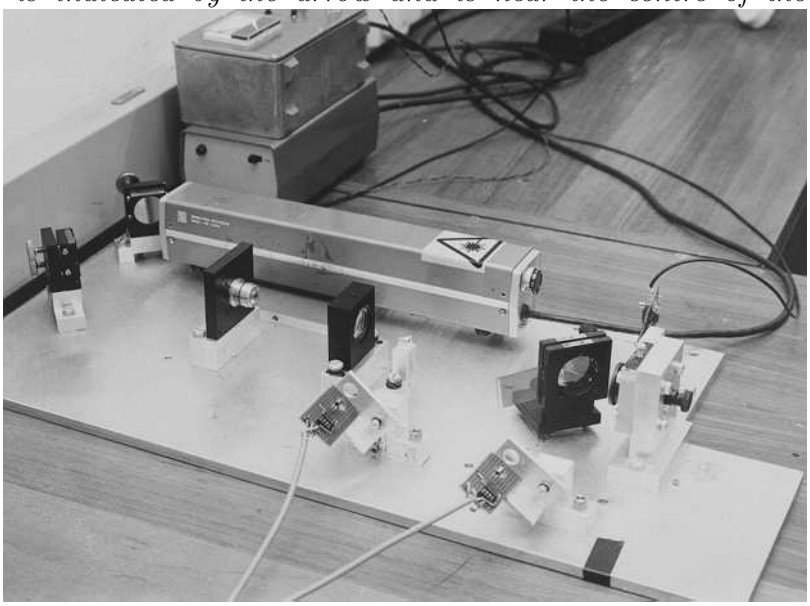

Plate 2: *Photograph of the distance-measuring interferometer. The main laser beam exits from the left of the laser. The beam exiting to the right is used for frequency stabilisation. In the foreground are the photodiodes and their associated preamplifiers. The measuring beam leaves the* 45° *beamsplitter plate towards the top right hand corner of the picture.*

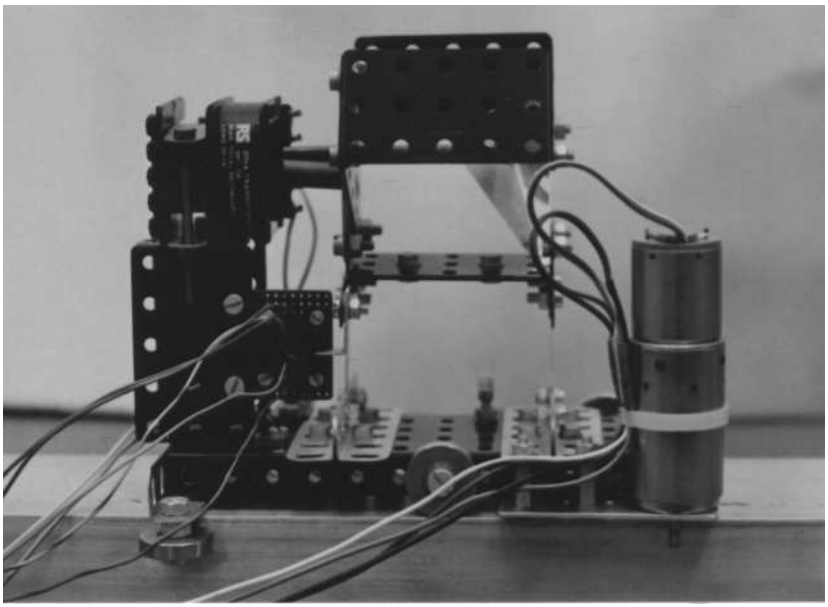

Plate 3: *The trolley. The photosensor which measures the displacement of the corner cube suspension is on the perforated circuit board at the lower left.*

Chapter 8

Aperture Synthesis Experiments on the I.N.T.

Because COAST was under construction during the period of this study, it was not possible to gain any practical experience of the problems of optical aperture synthesis from COAST itself. This chapter describes experiments that were performed in July 1987 using the Isaac Newton Telescope at La Palma as an optical testbed for astronomical interferometry.

The basic setup for these experiments was originally proposed by Fizeau: the telescope is converted into an array of small apertures by masking off the pupil apart from a few small holes. In this way we can obtain an array of apertures with baselines up to the diameter of the primary mirror — Michelson [61] went one step further than this by adding extension mirrors to the telescope structure, but for the experiments performed here the 2.5 metre primary on the INT was more than adequate.

We can describe this arrangement in terms of a separate element interferometer: path compensation is automatically achieved by moving the telescope to track the observed source in the normal way, and the beams from the sub-apertures are 'correlated' in the focal plane of the telescope to give an Airy disc crossed by fringes. It is interesting to note that, because in this case the 'input pupil' of the array and the 'output pupil' of the correlator are one and the same, the interferometric field of view of this arrangement can be as large as is permitted the size of the isoplanatic patch [7]. The disadvantage of this scheme is that the fringe frequencies are set by the length of the baselines. For long baselines this restricts the allowable bandwidth since the number of fringe cycles across the fringe pattern will be large (see later).

Many short-exposure images are recorded and the amplitudes and phases of the fringes can be measured in the same way and as for a separate-element interferometer. Coverage of the u-v plane is achieved by changing or rotating the aperture mask and then an image can be synthesised using the measured amplitudes and closure phases.

We can look upon this as a speckle interferometry experiment (see section 1.4) with the mask serving to alter the optical transfer function (OTF) of the telescope. While the instantaneous baseline coverage of an aperture masking system is much smaller than the full pupil, there are several important benefits. Firstly, we can reduce the redundancy of the beam combination. In the filled pupil, pairs of points in the aperture with the same separation all contribute to the same fringe frequency, but the contributions from different pairs will have different atmospheric phase errors at any one time, leading to 'blurring' of the fringes and an increase in the atmospheric noise. An aperture mask with non-redundant hole spacings, on the other hand, restricts the contributions to a given fringe frequency to a small set of pairs of points which are close together so that the contributions from all the pairs have almost the same phase. Aperture masking is therefore preferable at high light levels because the atmospheric noise is small.

Secondly, because of the small number of apertures, the 'dilution' of the fringes by the background flux from non-contributing apertures is small, i.e. the fringe visibilities are high. In terms of pure photon noise, this may not be the best thing to do since the increase in fringe visibility is offset by the loss in the detected photon rate, but having high fringe visibilities may be preferable if there are significant detector non-linearities present: these will usually cause errors in the measured fringe amplitudes which increase with increasing flux levels. Clearly, the higher the relative amplitude of the fringes compared with the total flux, the more resistant they will be to errors which are a fixed fraction of the flux.

At low light levels, there may be a trade-off between non-redundant and redundant arrays of apertures (the most extreme case of which is the filled pupil) and the discussion in chapter 4 will be relevant to this. However in the experiments described here only non-redundant apertures were used so as to simplify the analysis of the data.

8.1 Optical Setup

Figure 8.1 shows the optical arrangement used in the experiment. Because placing a mask directly over the primary mirror is a difficult and risky operation, the pupil is instead reimaged behind the telescope focus and the mask is placed here. This has the further advantage that the pupil is demagnified by a factor of 150 so that the masks need only be 17mm across. This demagnification is something of a mixed blessing, though, because the holes in the mask become so small ($\lesssim$ 1mm diameter) that they are difficult to drill accurately.

A filter is placed in the collimated beam to define the bandpass. The allowable bandwidth is set in these experiments by the number of fringes across the image

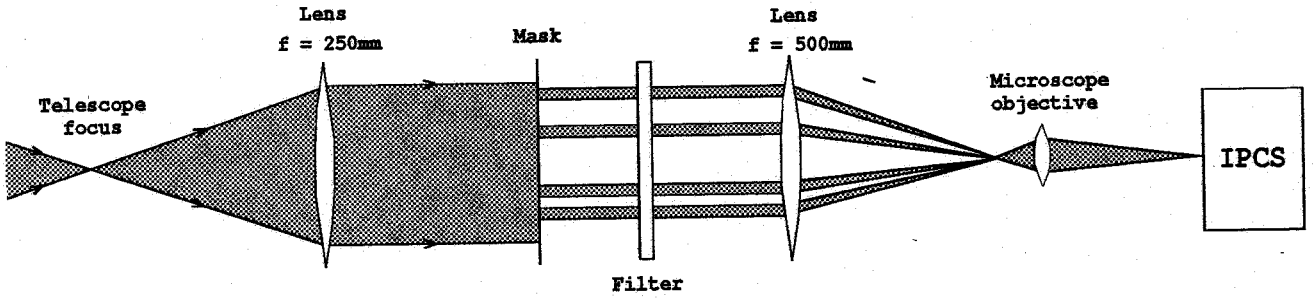

Figure 8.1: *Schematic diagram of the optical setup on the INT*

rather than the uncertainty in the optical pathlengths through the atmosphere. The latter is of order $0.4(L/r_o)^{5/6}$ wavelengths, where L is the length of the baseline being considered, which limits the fractional bandpass to

$$\Delta\lambda/\lambda \lesssim 2.5(r_o/L)^{5/6}\,, \tag{8.1}$$

whereas the former effect places the constraint

$$\Delta\lambda/\lambda \lesssim (D/L)\,, \tag{8.2}$$

where D is the diameter of the holes. This latter result can be most easily derived by considering the pairs of points in the aperture which contribute to a given fringe frequency: their separation will be directly proportional to the wavelength, which means that if a fringe at a given spatial frequency is contributed to at wavelength λ by pairs of points in the aperture which are separated by the centre-to-centre spacing L of a particular pair of holes, then at a wavelength $(1+D/L)\lambda$ only the points right at the edges of the two holes will be contributing to that fringe frequency.

The constraint in equation 8.2 is clearly more severe than that set in equation 8.1 for hole diameters $\lesssim r_o$ and even for much larger holes if the baseline is long. In these experiments the minimum hole diameter was 6cm and the maximum baseline was 2m so that a fractional bandwidth of less than 3% is adequate: the widest fractional bandwidth used here was 2.3% (a 12nm bandpass at a centre wavelength of 512nm).

The beams are brought to a focus by another lens and the resulting image is magnified by a microscope objective and detected using the IPCS (Imaging Photon Counting System). This device is basically an image intensifier which is scanned by a tv-camera-like system. The resulting electronic signals are processed in order to determine the positions of the individual photoevents, as well as to reject noise features. Its major limitations for these purposes is its relatively slow frame rate and the photon coincidence losses which are reported later, although it must be aknowledged that it was never built with interferometry in mind.

Image scale	0.541/0.271 arcsec/mm (×5/×10 microscope objective)
Pixel size	50 × 50 μm^2
Frame scan format	258 increments × 256 lines
Data window	180 × 180 pixels
Frame time	17 ms

Table 8.1: *IPCS format used for these experiments.*

The image scale on the IPCS was chosen so that the highest frequency fringes were sampled by at least four pixels per fringe cycle, thus keeping the fringe visibility loss due to finite pixel size to less than 10% (see equation 6.4). The details of the IPCS format are presented in table 8.1.

The whole optical setup was mounted at the Cassegrain focus of the INT in the RGO/RSRE imaging box. Not shown in figure 8.1 is the reflector situated in the focal plane of the telescope. This had a small hole to allow the light from the observed object to pass through to the aperture masking optics while the rest of the field of view was reflected to the standard acquisition and guiding camera of the INT. This allowed for quick acquisition despite the very small field of view at the IPCS.

8.2 Observations

The main aim of the observing run was to prove optical aperture synthesis as a working astronomical imaging technique. A number of close binary systems ($\sim$ 0.1 arcsec separation) were observed, the observations being preceded and/or followed by the observation of a point source so as to calibrate the visibility amplitudes, as explained in chapter 3. The masks used all consisted of linear arrays of 4 holes with spacings in the ratio 1:3:2 so that visibilities were measured on six equally spaced baselines. The maximum baseline was either 1 or 2 metres, depending on the desired resolution. Two-dimensional u-v coverage was achieved by rotating the Cassegrain turntable to which the imaging box was attached, thereby rotating the position angle of the mask with respect to the sky.

The hole diameters used were mostly sub-r_o (r_o was of the order of 10cm at 500nm — see section 8.5.1): this was so as to reduce the sensitivity of the calibration of the visibilities to any variations in the seeing. The photon rates were kept below about 60 photons per frame by inserting neutral density filters where necessary, in order to minimise the effects of detector coincidence losses (see section 8.4).

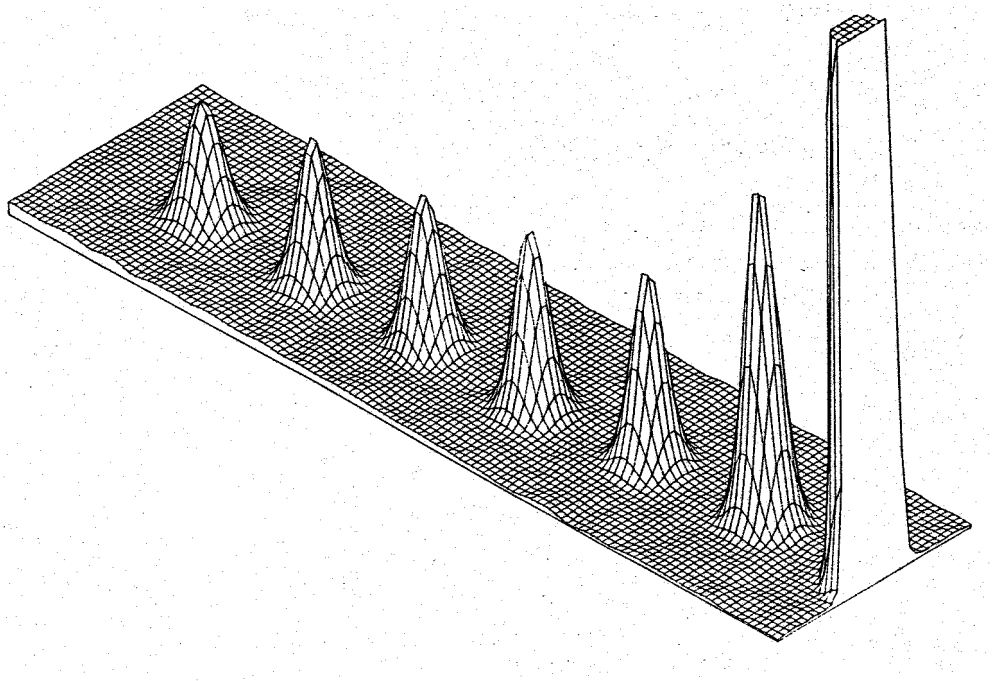

Figure 8.2: *The summed power spectrum of the images from a typical observation. The central spike has cut off above a threshold.*

8.3 Data reduction

The output of the IPCS was stored on tape as a list of the coordinates of the photo-events recorded in each frame. The first operation performed on the data was to sum the spatial power spectra of a large number of frames (∼50 000) in order to determine the fringe frequencies. In theory the latter could have been calculated by dead reckoning from the positions of the holes in the masks, but there were large uncertainties in such parameters as the relative orientation of the mask and the IPCS which made this impractical. The power spectrum was determined by summing the autocorrelation functions of the frames (this can be made very fast by using a photon-differencing algorithm — see for example [4]) and taking the Fourier Transform of the summed autocorrelation. The result then looks like figure 8.2, and the fringe frequencies can be determined by eye from the positions of the peaks.

A small amount of adjustment may be necessary in order to ensure that the 'triangles' of baselines which make up closure phases do in fact 'close', that is that the fringe frequencies corresponding to these baselines have a vector sum of zero. If a triangle does not quite close, this can lead to quite large systematic errors in the closure phases which depend on the position chosen for the phase centre of the Fourier Transform of the fringe pattern.

The closure phases were then evaluated using the triple product estimator in

equation 2.15. The individual complex fringe visibilities were calculated by taking a discrete Fourier Transform of each frame at the six fringe frequencies. Once again the process can be made very rapid by making use of the fact that the data is stored as lists of photon coordinates — evaluating a given Fourier component becomes a summation of values from a look-up table which is indexed by the coordinates of each new photon.

The fringe visibilities could in theory be evaluated in the same way, but examination of figure 8.2 shows that the background level in the power spectrum, instead of being a constant photon noise bias as predicted by equation 2.8, is in fact sloping. This is due to photon coincidence losses which are discussed in more detail in section 8.4. It is shown there that we can compensate for these losses by fitting a slowly-varying function to the background and interpolating to determine the value of the background at each fringe frequency. Having done this the mean squared fringe visibility can be determined from the height of the peak above the background, normalised by the value at the origin of the power spectrum (which has been similarly corrected for the non-uniform background). These visibilities were then calibrated by dividing them by the visibilities observed on point sources with the mask at the same position angle.

The calibrated visibilities and the closure phases (which need no calibration) were then used to construct images of the observed sources using a standard radioastronomy VLBI map reconstruction program [88]. This program takes as its input the measured amplitudes and a set of phases that are consistent with the closure phases. A starting model of the source (a point source was used here) is used to assign the unknown 'antenna' phases to the measured baselines by adjusting the antenna phases until the data fits the model as well as possible. The model is now updated so that the new model fits the 'phased up' visibilities better than its predecessor. At the same time the new model is chosen to maximise the 'entropy' of the map (for a discussion of Maximum Entropy methods see the relevant session *Indirect Imaging* ed. J. Roberts, C.U.P., Cambridge). The Maximum Entropy criterion is used to 'regularise the inverse problem' of making a map in a situation where the antenna phases are unknown and the Fourier plane is incompletely sampled. Not only does it enforce positivity in the map plane, it also tends to produce a smoother map by suppressing any features which are not justified by the signal-to-noise ratio of the data. This process of adjusting the map and the antenna phases is reiterated until the map fits the data to within the noise.

In simulations where fake data were generated corresponding to the observation of model binary sources on similar baselines to those used in the real experiment, this program proved very successful in correctly reproducing the position angle and separation of the model sources, but it was found that the flux difference between

the two components source was systematically overestimated. With the real data, therefore, the Maximum Entropy image was used to determine the source position angle and separation and these were used in the determination of the flux ratio by least-squares fitting to the observed visibilities.

Another problem encountered in the use of this program was that it used a model of the noise on the data that was inappropriate for optical aperture synthesis. In this model the noise is assumed be additive baseline noise which is small compared to the signal. In the optical case we can only get high signal-to-noise ratios if we incoherently average the amplitudes and closure phases over a large number of exposures, but then the relative magnitudes of the noises on the amplitudes and the closure phases may be incompatible with the additive baseline noise model. Thus the value of the baseline noise provided to the program must be a compromise between the actual amplitude and closure phase errors. Furthermore, the closure phase errors may not be correlated in the same way as is implicit in this model (see section 2.5.4) and so whereas in a 4-aperture system there may be 4 independent closure phase estimates to program is only able to make direct use of 3 closure phases. It is possible to provide a better estimate for these 3 closure phases using the extra closure phase as an extra constraint, but it would be better to solve this problem while including map-plane information at the same time. To get the best out of the data, therefore, programs will have to be written specifically for optical aperture synthesis which take into account a more appropriate model for the noise.

8.4 Coincidence Losses

It is clear from figure 8.2 that the power spectrum at frequencies where there is supposed to be no signal is in fact slightly sloping. Examination of the spatial auto-correlation function (see figure 8.3) immediately reveals the reason for this: there is a 'hole' around the origin where the pixels are nearly zero, whereas their neighbours have large values. Closer examination reveals that this effect is confined essentially to the pixel at the origin and the adjacent pixels (see figure 8.3(b)).

This can be explained in terms of the way the detector electronics operate. The IPCS uses a photon-centroiding algorithm to increase its resolution, that is, it locates the centre of the 'splash' of light on the output phosphor of the image intensifier caused by a photoevent at the input photocathode, thereby reducing the smearing of the image due to the finite size of the splash. This centroiding is done using a 'pattern-recognition' algorithm which recognises patterns of light and dark on the phosphor and deduces the positions of the photoevents from this [46]. However the number of different patterns which the electronics can recognise is small and if two photon splashes occur too near to each other the resulting intensity pattern is not

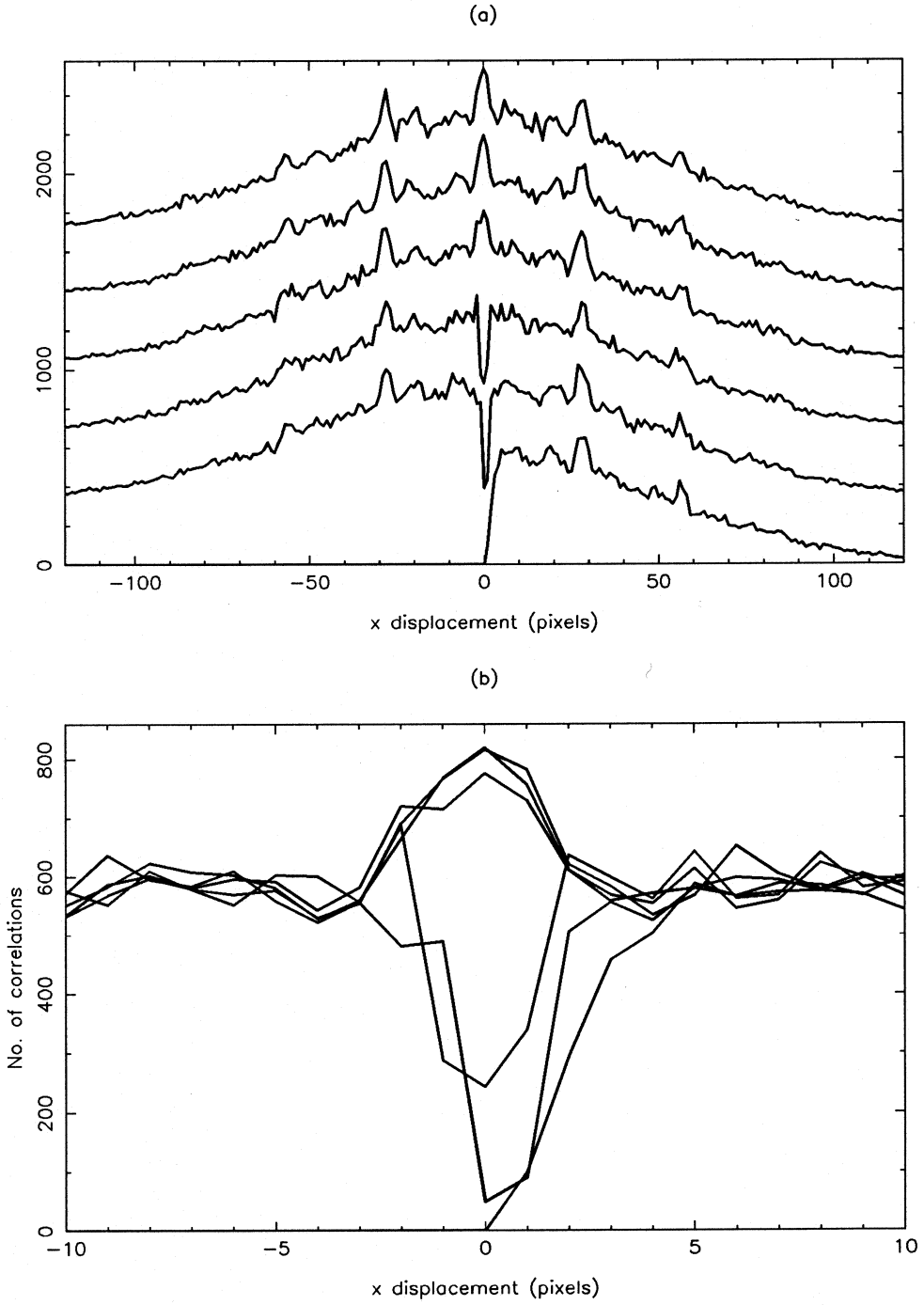

Figure 8.3: *Slices near the origin of the summed autocorrelation of a set of images recorded with the IPCS. The slice for $y = 0$ has a zero value for negative x because the information for this part of the autocorrelation can be recovered from the positive-x region. Figure(b) shows a close-up where the offset between slices has been removed so that the value of the autocorrelation can be read more easily.*

properly recognised. It is not certain whether such occurrences are treated as a single photoevent by the electronics or rejected altogether as noise, but it shall be assumed hereafter that the latter is the case. With this fault, two adjacent or coincident photoevents in one frame will never be recognised and thus the observed autocorrelation function will be zero near to the origin. The fact that in reality a small number of counts are observed near to the origin indicates that the electronics can recognise a small fraction ($\sim 1\%$) of these occurrences, but we shall neglect this in the following analysis.

Clearly, the systematic removal of close pairs of photons from the observed images will bias the fringe amplitude and closure phase measurements, and so in the next two subsections these biases and ways to compensate for them are discussed.

8.4.1 Correcting the measured visibilities

In considering the bias to the visibility amplitudes, it is helpful to consider the autocorrelation plane, i.e. the Fourier conjugate plane of the power spectrum from which the mean squared visibilities are derived. The most obvious effect of the coincidence losses is the 'hole' near the origin of the autocorrelation function. If we could 'fill in' this hole by, say, interpolation from adjacent pixels then this effect could be removed. Unfortunately there are significant small-scale features near the origin because of the presence of high-frequency fringes in the image, and this complicates the process of interpolation. It was found easier to do the interpolation in the power spectrum plane, making use of the *a priori* knowledge that the power in that plane is confined to relatively small regions.

The analysis of this process is as follows: if this first-order 'hole' effect is the only significant effect of the coincidence losses, we can write the mean measured autocorrelation function as

$$\langle \hat{a}(\mathbf{x}_i) \rangle = a(\mathbf{x}_i) - h(\mathbf{x}_i) a(\mathbf{x}_i)$$

where $a(\mathbf{x}_i)$ is the (discretised) true autocorrelation function and h is a top-hat-like function

$$h(\mathbf{x}_i) \simeq \begin{cases} 1 & |\mathbf{x}_i| \leq 1.4 \\ 0 & \text{elsewhere} \end{cases}$$

(this is only an approximate form because in reality there are some coincidence losses for photons pairs at separations of two pixel widths). In the power spectrum plane this becomes

$$\langle \hat{A}(\mathbf{u}_k) \rangle = A(\mathbf{u}_k) - H(\mathbf{u}_k) * A(\mathbf{u}_k) \tag{8.3}$$

where $A(\mathbf{u}_k)$ is the true power spectrum and $*$ denotes convolution. $H(\mathbf{u}_k)$ is the Fourier Transform of $h(\mathbf{x}_i)$ and it follows therefore that since $h(\mathbf{x}_i)$ is a narrow spike,

$H(\mathbf{u}_k)$ will be a very broad function. In comparison the spikes in the power spectrum can be approximated as a set of delta-functions; by far the largest of these will be the spike at the origin and so we have the approximate formula

$$\langle \hat{A}(\mathbf{u}_k) \rangle \simeq A(\mathbf{u}_k) - H(\mathbf{u}_k)A(\mathbf{0}).$$

This means that the background will have the shape of a shallow bowl centred at the origin, and indeed this is what we see in the measured power spectra.

In practice it is better to make use of the more exact expression in equation 8.3 when compensating for this effect: we can determine $H(\mathbf{u}_k) * A(\mathbf{u}_k)$ directly from the data, making use of our knowledge that it is slowly-varying by fitting a low-order two-dimensional polynomial to the measured background. We exclude from the fit the regions near the peaks corresponding to the fringe frequencies and then interpolate the fitted polynomial to determine the value of the background underneath these peaks. The bias in the fringe amplitudes due to the hole in the autocorrelation can then be removed by subtracting this background from the measurements of the peak heights.

If the hole in the autocorrelation were the only effect of the coincidence losses, then the removal of this background would be sufficient to compensate the data, but this is in fact only an approximation. The value of the autocorrelation function as a whole will be lowered by the removal of these photon pairs, since they could otherwise have correlated with photons much further away in the image. If this loss is a uniform fraction over the whole of the autocorrelation, the measured fringe *visibilities* will not change, since the power at the origin of the power spectrum will be affected in the same way as that at the fringe frequencies. Unfortunately, the loss of photons due to coincidences will tend to be highest near the crests of the fringes where the photon rate is highest and this will cause correlations to be preferentially lost from the spacings corresponding to the crest-to-crest distances. The fringe visibilities will therefore be systematically lowered at high photon rates.

We can see if this effect is important at the photon rates used in the experiments discussed here by setting up a crude model of this process. In this model, the fringe pattern consists of a constant illumination level superimposed by fringes of small amplitude at a single fringe frequency. The fine-grained structure of the problem is removed by making the effective pixel size large so that the coincidence loss criterion becomes simply that two photons landing in the same pixel will not be detected — adjacent photoevents *are* recognised in this model because the effective pixel size is large compared to the coincidence loss area. Furthermore, there are only two pixels per fringe cycle and so the pixel illumination will be a series of alternating values $\lambda(1 \pm V)$ where λ is the mean illumination expressed in photons per pixel per frame and V is the fringe visibility. We shall consider the situation where λ and V are

$\ll 1$. The true autocorrelation function is therefore also a square wave. The spacings 0, 2, 4 pixels etc. corresponding to crest-to-crest correlations and trough-to-trough correlations will have the value

$$\begin{aligned} a_0, a_2, \text{etc.} &= (N\lambda^2/2)([1-V]^2 + [1+V]^2) \\ &= N\lambda^2(1+V^2) \end{aligned}$$

where N is the total number of pixels (assumed large), and the spacings 1, 3, 5 etc. will have the value

$$\begin{aligned} a_1, a_3, \text{etc.} &= N\lambda^2(1-V)(1+V) \\ &= N\lambda^2(1-V^2) \end{aligned}$$

The Fourier Transform of the autocorrelation will then give us the expected value for the mean squared visibility, i.e. V^2.

We can now see what effect coincidence losses will have on this measurement. A photon will only be detected if one and only one photoevent occurs in a given pixel. The probability of this occurring is given by the Poisson distribution

$$\begin{aligned} \Pr(1) &= \lambda(1 \pm V)\exp[-\lambda(1 \pm V)] \\ &\simeq \lambda(1 \pm V)[1 - \lambda(1 \pm V)] \end{aligned}$$

where we have made use of the assumption that $\lambda \ll 1$. From this we can determine the mean of the measured autocorrelation. Clearly, if we remove the photon noise bias by not correlating photons with themselves, we have

$$\langle a_0 \rangle = 0.$$

This is the familiar 'hole' in the autocorrelation. For spacings of 2, 4, 6 pixels etc. the measured correlation will be

$$\begin{aligned} \langle a_2 \rangle, \langle a_4 \rangle, \text{etc.} &= (N/2)\lambda^2\{(1+V)^2[1-\lambda(1+V)]^2 + (1-V)^2[1-\lambda(1-V)]^2\} \\ &\simeq N\lambda^2(1+V^2)(1-2\lambda)(1-4\lambda V^2)\,, \end{aligned}$$

and the correlation for odd spacings will be

$$\begin{aligned} \langle a_1 \rangle, \langle a_3 \rangle, \text{etc.} &= N\lambda^2(1+V)[1-\lambda(1+V)](1-V)[1-\lambda(1-V)] \\ &\simeq N\lambda^2(1-V^2)(1-2\lambda). \end{aligned}$$

We can see that although the mean squared total flux is reduced by a fraction of about 2λ, the reduction in the mean squared *visibility* of the fringes is only of order $2\lambda V^2$, providing that the hole at the origin is appropriately filled in.

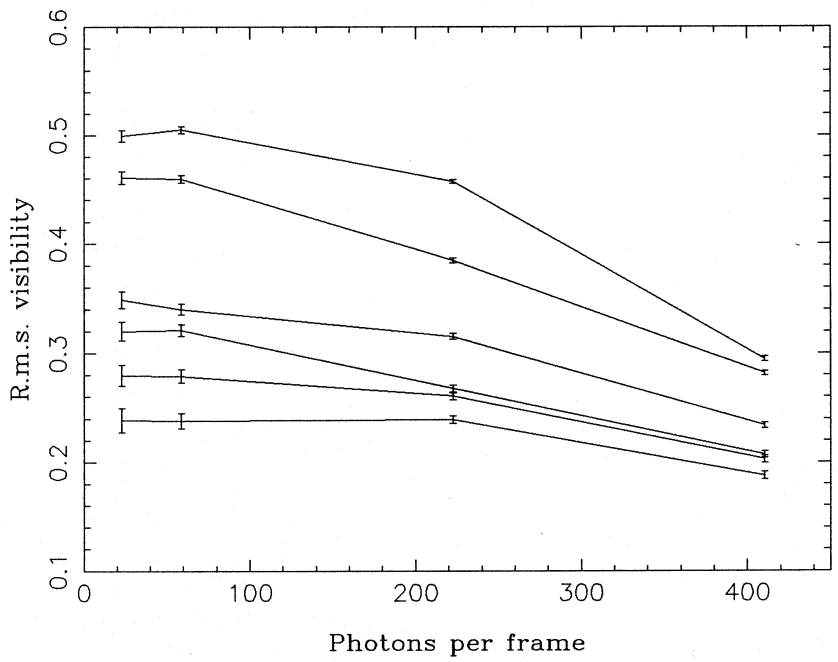

Figure 8.4: *The visibilities observed on a point source as a function of photon rate. All six baselines from a 4-hole mask are shown: the visibility decreases monotonically with baseline.*

We can apply this roughly to the IPCS data by taking the 'effective' pixel size to be the area around one photon event where the occurrence of a second photon event will cause the loss of them both, i.e. an area of about 3×3 pixels. With 100 photons per frame spread over an area of roughly 90×90 pixels, λ is then about 1/9. For a 4-hole mask, $V < 1/4$ so that the magnitude of this effect is less than 1/72. In reality the visibility will usually be less than half this value and so this 'second-order' coincidence loss will be negligible ($\ll 1\%$); compensating only for the 'first-order' effect, i.e. the hole at the autocorrelation origin, should be sufficient.

To test this prediction empirically, the variation in the observed fringe visibility as a function of photon rate was determined from a set of data taken on a point source. The visibilities were corrected for the first-order effect only and the results are displayed in figure 8.4. It can be seen that the measured visibility does not alter (to within the noise limits) between photon rates of 22 per frame and 58 per frame, but that at rates of 200 photons per frame and higher, significant losses appear. Thus for all the observations in this run, the first-order correction is adequate.

8.4.2 Closure Phase

To analyse the effect of coincidence losses on the closure phases, it is helpful to think in terms of the 'bispectrum', i.e. the triple product as a function of two of the spatial frequencies that define it

$$T(\mathbf{u}_1, \mathbf{u}_2) \equiv \langle V(\mathbf{u}_1)V(\mathbf{u}_2)V(-\mathbf{u}_1 - \mathbf{u}_2)\rangle$$

and its Fourier conjugate, the 'triple correlation' [57]

$$t(\mathbf{x}_1, \mathbf{x}_2) \equiv \langle \int i(\mathbf{x})i(\mathbf{x}+\mathbf{x}_1)i(\mathbf{x}+\mathbf{x}_1+\mathbf{x}_2)\, d^2\mathbf{x}\rangle\,,$$

where $i(\mathbf{x})$ is the intensity distribution in the image.

For lists of photon coordinates, we can estimate the triple correlation (and hence the bispectrum) from

$$\hat{t}(\mathbf{x}_1, \mathbf{x}_2) \equiv \sum_{k_1=1}^{n} \sum_{k_2\neq k_1} \sum_{k_3\neq k_2\neq k_1} \delta(\mathbf{x}_1 - \mathbf{p}_{k_1} + \mathbf{p}_{k_2}, \mathbf{x}_2 - \mathbf{p}_{k_2} + \mathbf{p}_{k_3})\,,$$

where p_k is the coordinate of the kth photon [66]. Notice that once again we avoid correlating photons with themselves in order to avoid the photon noise bias. It is clear from this formula that the triple correlation will be affected by coincidence losses in a similar way to the autocorrelation: the 'hole' in the autocorrelation becomes planes of missing correlations near to the axes of the 4-dimensional triple correlation.

To reduce this to understandable terms, we shall ignore one coordinate in the image so that the triple correlation and bispectrum are 2-dimensional. The bispectrum now looks something like the diagram in figure 8.5. Note that the 1-dimensional power spectrum appears along the axes of the bispectrum

$$T(\mathbf{u}, \mathbf{0}) = T(\mathbf{0}, \mathbf{u}) = T(\mathbf{u}, -\mathbf{u}) = \langle |V(\mathbf{u})|^2\rangle.$$

If we now consider the coincidence losses, we see that they appear as losses in correlations along lines in the triple correlation plane as shown in figure 8.6.

We can repeat the analysis of the previous subsection to show that in this case the measured triple correlation is reduced by the coincidence losses in the form

$$\langle \hat{T}(u_1, u_2)\rangle = T(u_1, u_2) - H(u_1, u_2) * T(u_1, u_2)$$

where $H(u_1, u_2)$ is the Fourier Transform of the star-shaped pattern of triple correlation losses $h(x_1, x_2)$ shown in figure 8.6(a). H is therefore itself star-shaped as shown in figure 8.6(b). We can see from this that convolution of H with the central spike in the bispectrum will not give rise to a background effect at the bispectrum points which contain the closure phase information, since these points are well away from the

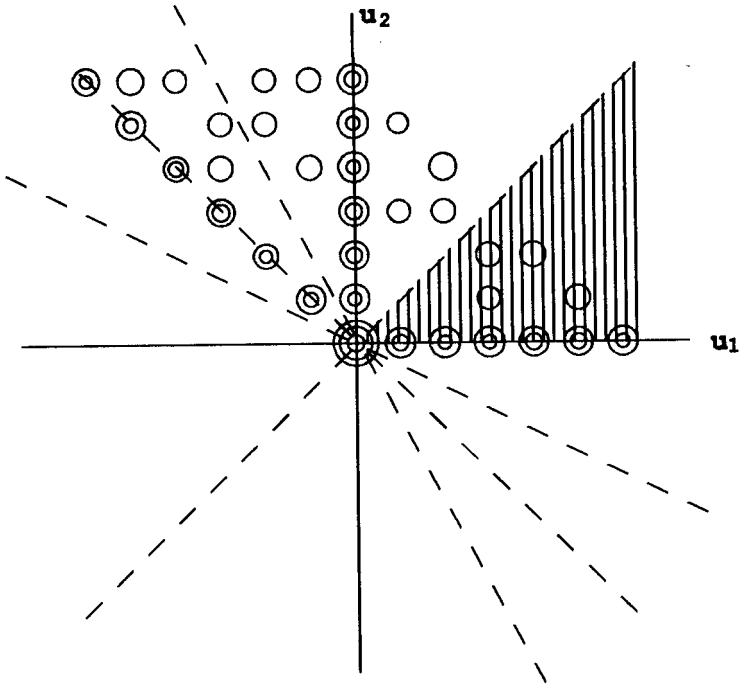

Figure 8.5: *A schematic countour plot of the bispectrum of the one-dimensional fringe pattern from a linear 4-hole mask.*
The peak at the origin has three contours, the peaks corresponding to peaks in the power spectrum are surrounded by two contours and the peaks which contain the useful closure phase information have a single contour. Because of the symmetry of the bispectrum, the information in the shaded region is repeated in the rest of the plane: the dotted lines demarcate successive regions containing the same information, but the contours have only been plotted in four of them.

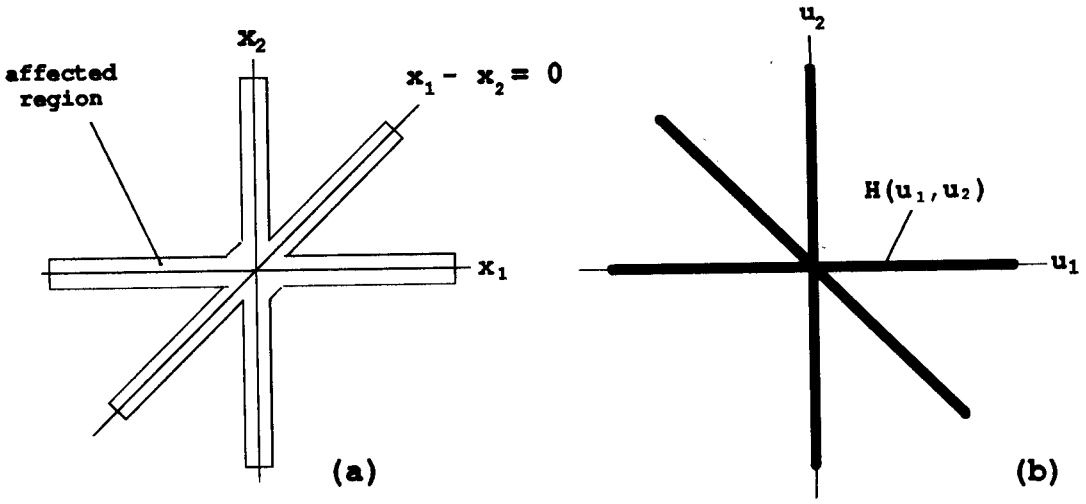

Figure 8.6: *(a) A schematic diagram of the regions in the 2-dimensional triple-correlation that are most affected by coincidence losses. (b) A schematic diagram of the function $H(\mathbf{u}_1, \mathbf{u}_2)$ defined in the text. It is the Fourier transform of the coincidence loss pattern in figure (a)*

axes. The largest effect on these bispectrum points will come from the convolution of H with the power spectrum peaks corresponding to the fringes on the baselines that make up the closure phase

$$\begin{aligned}\langle\hat{T}(u_1,u_2)\rangle &\simeq T(u_1,u_2) - T(u_1,0)H(0,u_2) - T(0,u_2)H(u_1,0) \\ &\quad - T(u_1+u_2,0)H(-u_2,u_2) + \text{etc.}\end{aligned}$$

The relative magnitude of the background bias is therefore of the order of

$$\begin{aligned}\frac{|\langle\hat{T}(u_1,u_2)\rangle - T(u_1,u_2)|}{|T(u_1,u_2)|} &\sim \frac{H(0,u_2)|V(u_1)|^2}{|\langle V(u_1)V(u_2)V(-u_1-u_2)\rangle|} \\ &\sim (H(u_1,0)|V(u_1)|/|V(u_2)||V(-u_1-u_2)|).\end{aligned}$$

Now we in fact know the magnitude of $H(0,u_2)$ empirically since it is approximately the value of the background bias to the power spectrum as a fraction of the 'd.c.' power. For these experiments it had a maximum value of typically 10^{-3}. Thus if $V(u_1)$ and $V(u_2)$ have similar magnitudes and we are prepared to tolerate closure phase errors of 5°, then we have

$$|V(-u_1-u_2)| \gtrsim 0.1$$

i.e. we can accept visibilities which are greater than about 40% of their maximum possible value in a 4-hole system without doing any background correction.

In practice no background correction was done on the bispectrum even for lower fringe visibilities than this because in such situations the closure phase errors due to photon noise were much larger than this. Furthermore this kind of closure phase error is less prone to produce large map plane effects than a constant closure phase offset; the effect of this bias is mostly to reduce the magnitude of the real part of the measured triple product, which has has little effect on closure phases close to zero and which does not bias the sign of the closure phase. The latter fact has the consequence that we can still correctly tell the 'handedness' of an asymmetric image despite this error.

8.5 Results

8.5.1 Seeing tests

Measurement of r_o

An approximate guide to the spatial scale of the atmospheric seeing was obtained at regular intervals throughout the observing period, by measuring the full width to half maximum (FWHM) of long-exposure ($\sim$ a few seconds) images observed through the full telescope aperture using the acquisition and guiding camera. The values obtained at different times were between 0.9 and 1.3 arcseconds. If we assume a Kolmogorov spectrum of atmospheric phase perturbations, this implies that r_o at 500nm wavelength was in the region 11cm to 7.6cm [75, table 2].

Measurement of t_o

This was accomplished by determining the correlation between the complex visibility measured in different frames as a function of their separation in time:

$$C(t, T_{exp}) = \langle V(t', T_{exp}) V^*(t' + t, T_{exp}) \rangle \, ,$$

where $V(t, T_{exp})$ is the visibility measured from an exposure of length T_{exp} starting at time t. This was normalised by $C(0, T_{exp})$ (i.e. the mean squared fringe visibility) so as to remove the effects of any visibility losses not caused by temporal fluctuations and the results were then compared to the values expected from the Tatarski spectrum of phase perturbations with different values of T_{exp}/t_o.

In practice only the cross-correlation for $t = T_{exp}$ was used in determining t_o because the correlation and hence the signal-to-noise ratio fell off rapidly after this; the value measured for $t = 2T_{exp}$ was used merely as a check on the theory.

Figure 8.7 shows the theoretical values for the correlation (see appendix F for the details of the calculations) for $t = T_{exp}$ and $t = 2T_{exp}$ as a function of T_{exp}/t_o. Plotted

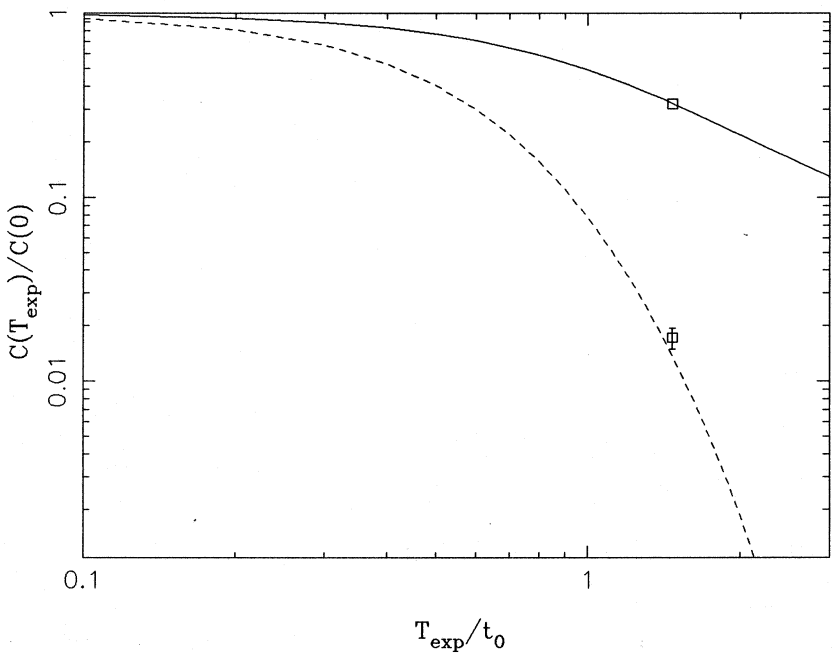

Figure 8.7: *The theoretical correlation between the visibilities in two exposures separated in time by T_{exp} (full line) and $2T_{exp}$ (dashed line), where T_{exp} is the exposure duration. Indicated also are the measured values of both parameters from a typical run.*

on this are some typical experimental values, showing that $T_{exp} = 1.47t_o$, if we believe that the spectrum of the fluctuations is truly a Tatarski spectrum. The data point for $C(2T_{exp}, T_{exp})$ lends some credence to this idea, but all the same this result must be treated with caution. Given that the IPCS frame time was 17ms, this would imply that t_o was about $12ms$ for this observation.

With this coherence time, the results of section 3.2 show there is little point in increasing the exposure time by coherently adding successive frames.

8.5.2 Point source visibilities

The r.m.s. visibilities observed on a point source in a typical observation are shown in figure 8.8. The visibilities are quoted as a fraction of their maximum possible value (i.e. 1/4 for a 4-hole mask) and this convention is adopted for the rest of this chapter. It can be seen immediately that the observed visibilities are much lower than would be expected if the spatial fluctuations of the atmospheric refractive index were the only effect causing a loss of coherence. We can list various other effects that will cause a loss in fringe visibility:

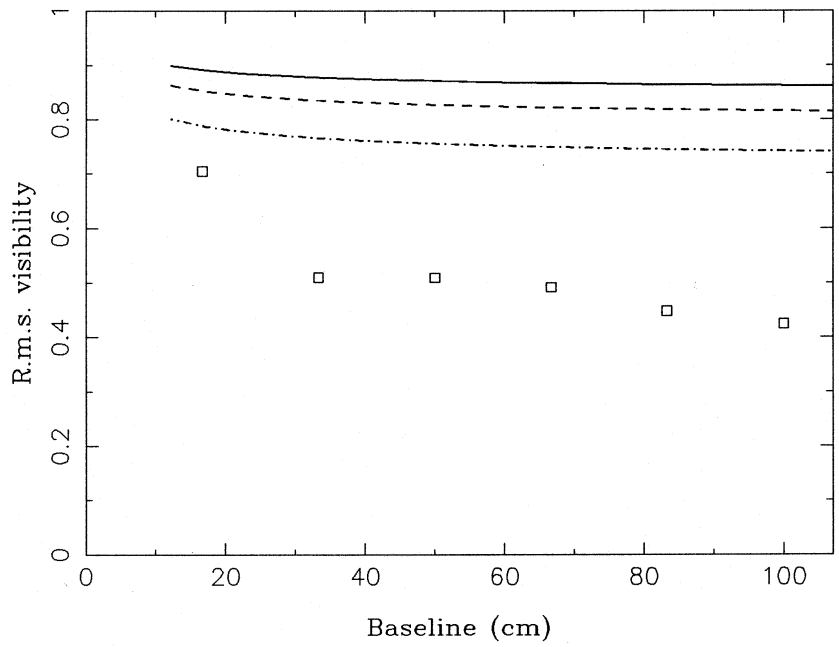

Figure 8.8: *The measured r.m.s. visibilities on the six interferometer baselines when observing on a point source (HR7948). Also shown (dotted lines) are the predicted values of the visibility for* $r_o = 10$, *13 and 16cm and the hole sizes used in the observation (6.1cm), but neglecting temporal and bandwidth effects. At the time of the observation* r_o *was estimated to be 13cm at the observing wavelength (633nm).*

Temporal fluctuations in the fringe phase. If the coherence time result from the previous subsection is to believed, this would imply a visibility loss due to the finite exposure time of about 15% (see figure 3.1).

Finite bandwidth. For this observation the filter had a centre wavelength of 633nm and a bandwidth of 11nm i.e. a fractional bandwidth of 1.7%. We can calculate the approximate loss in fringe visibility due to the finite bandwidth by calculating the fraction of the aperture area contributing to a given fringe frequency as a function of wavelength (see the discussion on the choice of filter bandwidth in section 8.1). For the fringe frequency defined by the centre wavelength and the centre-to-centre spacing of apertures, the fringes at the half-power wavelengths (i.e. $1 \pm 0.85\%$ times the central wavelength) will have a visibility a factor 0.82 lower than at the centre of the bandpass due to this effect, if the aperture spacing is 1 metre and the aperture diameters are 6cm. This means a loss factor of about 0.9 when averaged over the bandpass.

Finite pixel size. On the finest fringes there will be four pixels per fringe cycle which will attenuate the observed visibilities by a factor of 0.9 (see equation 6.4).

If we multiply together all these effects we predict a 56% r.m.s. visibility on the longest baseline compared with the observed value of 42%. A number of explanations of this extra loss may be put forward:

[1] The turbulence may not be described by a Kolmogorov spectrum and therefore the deduced values of r_o and t_o give a false indication of the real atmospheric disturbances. For instance the measurement of the spatial scale of the seeing using the full aperture is most sensitive to the small-scale phase fluctuations, whereas the fringe visibilities are most affected by relative wavefront tilts across the apertures caused by longer wavelength fluctuations. However the most obvious candidate for a departure from the Kolmogorov law is the dome seeing, which might be expected to have an excess of the small-scale fluctuations; this would mean that the seeing measurement would give a *pessimistic* prediction for the fringe visibilities. Furthermore comparative tests made in La Silla [71], where image profiles and differential tilts were measured simultaneously, have agreed with the Kolmogorov model, although the situation may be very site-dependent.

[2] There may be vibrations in the telescope structure. A motion of 0.1 arcsecond during an exposure would serve to completely wipe out the fringes on a 1 metre baseline and yet be unnoticeable in a conventional image.

[3] There may be significant imperfections in the optics. At first sight this might seem unlikely since only small patches of the optics are used ($\sim$ (6cm)^2 on the primary mirror and $\sim$ (1mm)^2 on the collimating and refocussing lenses). However

	PA	Separation	Δm
This observation	$130^\circ \pm 1^\circ$	$0.189'' \pm 0.005''$	1.0 ± 0.1
Published values [63]	135°	$0.20''$	1.0

Table 8.2: *The parameters of the binary system β Delphinius derived from this observation compared with published values. The errors quoted are the internal errors in the observation and do not include systematic effects.*

low frequency errors in the optics will translate into relative wavefront tilts across the apertures which would reduce the fringe visibilities. Spherical aberrations in the singlet lenses used might cause this sort of effect, as would telescope focussing errors.

The first two possibilities could be tested by measuring the temporal power spectrum of the fringe phase disturbances and looking for departures from the Kolmogorov law. It was hoped that this could be done using the data taken on this run but it was found that the fringe phase could not be reliably tracked from exposure to exposure because the frame time was too long compared to the atmospheric coherence time. A decision about the cause of the excess visibility loss will depend on further experiments using higher quality optics and shorter frame times.

8.5.3 Imaging

Figure 8.9 shows the image reconstructed from observations of βDel, a 0.2 arcsecond binary. The observations were made with a 4-hole mask with a maximum baseline of 1 metre, at a wavelength of 633nm and with data taken at 10 position angles of the mask. Thus the visibilities at a total of 60 independent u-v points were measured, the coverage of the Fourier plane being shown in figure 8.10.

The theoretical resolution for this observation is 0.13 arcsecond and the binary is indeed well resolved in this map — there is even some evidence that the Maximum Entropy reconstruction algorithm is perhaps giving some 'super-resolution' because of the high overall signal-to-noise ratio of the data. The dynamic range of the map appears to be greater than 50:1, although this must be treated with some caution because the Maximum Entropy algorithm will tend to artificially supress the noise. It is clear that the closure phase has resolved the 180° ambiguity in the position angle of the binary. The separation and position angle measured from in this map agree tolerably well with the published values (see table 8.2): the major uncertainties were the exact orientation of the mask and changes in the effective scale of the mask due to focussing errors. The magnitude difference obtained from a least-squares fit to the visibilities is also in good agreement with the published value.

Attempts at imaging using a 2 metre maximum baseline were less successful be-

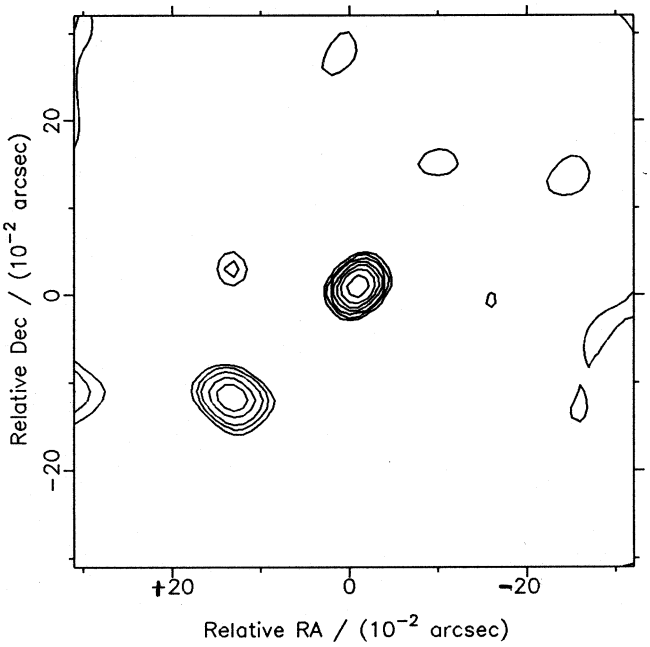

Figure 8.9: *Maximum Entropy reconstruction of the image of the binary β Delphinius. Contour levels are at 0.5%, 1%, 2%, 5%, 10%, 20% and 50% of the peak flux.*

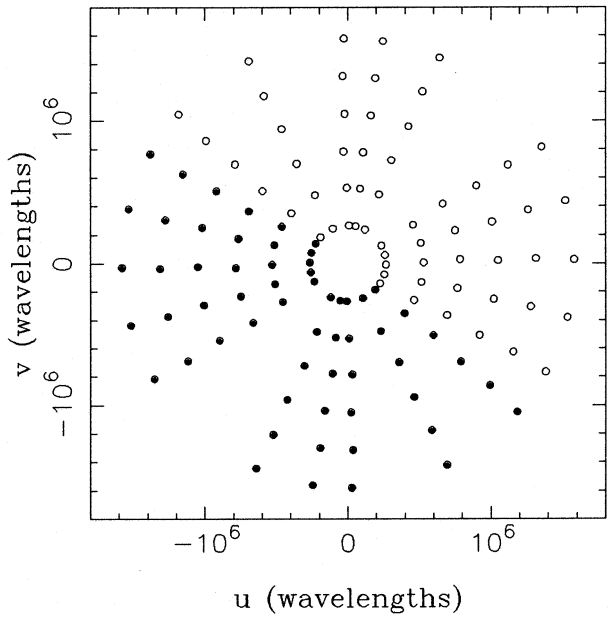

Figure 8.10: *The u-v plane coverage of the observations of β Delphinius used to construct the image in figure 8.9.*

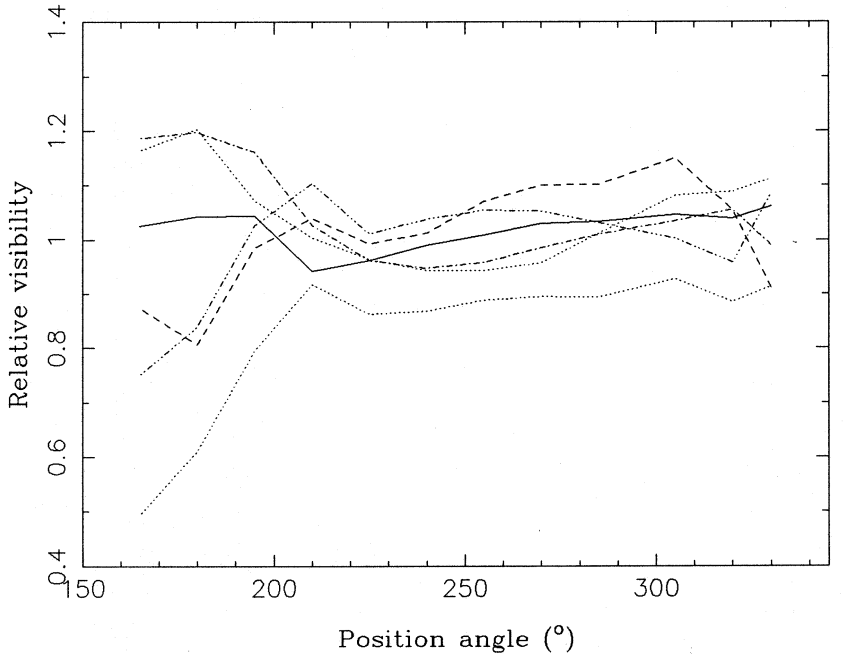

Figure 8.11: *The ratio of the visibilities measured for two different point sources (ϵ Her and epsilon Aql respectively) observed on the same night. The sources were observed at a wavelength of 512nm and a bandwidth of 11nm. Different baselines are indicated by different linestyles; in order of baseline length they are: (shortest i.e. 33cm) full, long dashes, dot-dash-dot, short dashes (longest 3 baselines: maximum 2 metres).*

cause of problems in calibrating the visibility amplitudes, which are illustrated by the example shown in figure 8.11: here the visibilities measured on one calibrator are divided by those measured on another calibrator which was observed immediately afterwards. We can see that there are large departures from the expected value of unity, especially at values of the position angle near to 165°. At this position angle, the observations of the two sources were in fact farthest apart in time, because the first source was observed starting at 165° and then at increasing values of the position angle up to 330°, while the second was observed staring at 330° and working in the reverse direction. Thus the observations at 165° were made nearly two hours apart, during which time the seeing may have changed.

There is an additional effect here, though: the effect of a seeing change would be to uniformly depress or raise the visibilities at any given position angle, but here there are baselines which have increased in visibility between the two observations and some that have decreased. This suggests that one of the holes may have been partially occulted by some obstruction in the pupil (e.g. the spider supporting the secondary) which would reduce the visibility on the baselines which included that

hole and increase the visibility on the other baselines. This occultation ought to be reproducible at the same position angle, but this has not happened here, whether because of inaccuracies in the rotation of the turntable or because of flexures or slippage somewhere in the instrument. This problem is more acute when longer baselines are present because it is harder to position the mask in the telescope pupil such that it is well away from possible sources of obscuration.

In other cases the effects are more obviously due to seeing changes: in one case there is a systematic rise in the visibility between the observations of two calibrators (indeed this rise was so fast that the binary source observed between these two calibrators sometimes had higher visibilities than the first calibrator) and this is correlated with a change in the seeing from 1.3 arcseconds at the beginning of the night to 1 arcsecond at the end of the night. It is suspected that this was mostly due to dome seeing because the dome temperature fell quite rapidly soon after the beginning of the night, indicating a large temperature difference between the inside and the outside of the building.

8.6 Conclusions

Aperture synthesis clearly works at optical wavelengths on baselines considerably larger than the scale size of the seeing, and despite quite serious deficiencies in the detector used.

The major problem to be overcome is accurate calibration of the visibilities. The effects of changes in the seeing as a function of time could be minimised by observing the source and calibrator at the same position angle in immediate succession, rather than observing the source at all position angles followed by the calibrator at all position angles as was done in these observations. This would also minimise any medium-timescale instrumental variations such as creep and thermal effects which may be leading to variable obscuration of the mask apertures.

A better understanding of the excess loss in fringe visibilities discussed in section 8.5.2 may also provide a clue as to any sources of calibration error, and any improvement in the visibilities would have the added bonus of an increased signal-to-noise ratio. To this end it would be advantageous to conduct further experiments with improved optics and higher frame rates in order to investigate the possible sources of this loss.

It would also be desirable to work with a detector which was less subject to coincidence loss problems, e.g. a CCD or the PAPA detector [68]. We have shown here that these problems can be tolerated at low photon rates, partly by virtue of the fact that only a small part of the power spectrum of the fringes contains any energy. For observations of bright objects, however, one would like to be able to observe on

many baselines simultaneously (as shown in chapter 4, this is makes the most efficient use of observing time at high light levels) and yet to work with as high a photon rate as possible so as to maximise the signal-to-noise ratio of the measurements.

Chapter 9

Conclusions

Each chapter has its own individual conclusions and so the aim here will just be to draw together the ideas that have been developed by attempting to make a rough prediction of the limiting magnitude for COAST. This sort of calculation has been made many times before, but hopefully this thesis has put at least some of the numbers onto firmer ground.

The conclusions of chapter 5 demonstrate that the magnitude limit will be reached when we can no longer reliably track the white-light fringe: while in principle a passive system (i.e. one which does not track the fringes) would be able to observe fainter objects, in practice the observing time necessary to achieve an adequate signal-to-noise ratio would prove prohibitive. The flux limit for fringe tracking is quoted in chapter 5 in terms of a 'canonical signal-to-noise ratio', defined as

$$SNR_0 = \left(|V(0)|^2 N_0\right)^{1/2} ,$$

where $V(0)$ is the high-light-level fringe visibility that would be observed if the effects of the finite exposure time and optical bandwidth could be ignored, and N_0 is the number of photons detected in the atmospheric coherence time t_o. We have shown that the fringes could be tracked for canonical signal-to-noise ratios down to about 1, and so we shall calculate here what this corresponds to in terms of the flux from a zeroth magnitude star.

Tables 9.1 and 9.2 show the estimated losses in fringe visibility and photon rate due to the atmosphere and the instrument. These are explained further below.

[1] Source spatial coherence. We shall take this as being unity, which might seem to imply that the source is unresolved and that therefore the observation being considered is not particularly interesting. However, the source might be resolved on the longer baselines but unresolved on the shorter baselines; if we can track the fringes on the shorter baselines then it might be possible to use a 'global fringe fitting' type

Cause	R.m.s. visibility attenuation
Source spatial coherence	1.00
Atmospheric spatial fluctuations	0.68
Atmospheric temporal fluctuations	—
Coherence length losses	—
Tilt errors	0.90
Delay line errors	0.90
Optical surface errors	0.87
Diffraction from aperture stop	0.94
Beamsplitter tolerances	0.97
Total	0.43

Table 9.1: *Estimated visibility losses for COAST.*

Cause	Flux attenuation
Atmospheric absorption	0.89
Diffraction from aperture stop	0.87
Reflection losses	0.56
Diffraction from surface errors	0.91
Detector quantum efficiency	0.50
Total	0.20

Table 9.2: *Estimated flux losses for COAST.*

of algorithm [87], to 'phase up' the longer baselines. Alternatively, the source might consist of a bright unresolved core and a fainter resolved component (e.g. a stellar envelope or the narrow-line region of an active galactic nucleus) in which case we would be interested in the small deviations from unity of the source visibility.

[2] Atmospheric spatial phase fluctuations. This figure is derived from figure 3.6, assuming an aperture diameter of $2r_o$ and that there is active correction of the tilt component of the perturbations. Figure 3.8 shows that the signal-to-noise ratio could be increased by going to a diameter of $3r_o$, but the gain would be small, and, as discussed in chapter 3, we are better able to calibrate the visibility amplitudes with the smaller aperture.

[3] Atmospheric temporal phase fluctuations. These effects do not need to be included here, as they are taken into account by the definition of the canonical signal-to-noise ratio.

[4] Coherence length effects. See the previous item.

[5] Tilt errors. As shown in chapter 5, it should be possible to keep the losses due to imperfect tilt correction down to less than 10%, even when the fringe tracking system is working at its limit.

[6] Delay line errors. A new delay line system is being built for COAST using a laser interferometer which has a resolution of $\lambda/16$ at 633nm, which is about $\lambda/20$ at 800nm. If we take the stability of the delay line as being roughly of this magnitude and we add together the errors of two delay lines, this implies a 10% visibility loss.

[7] Surface errors in the optics. On its passage from the sky to the correlator, each beam will be reflected off 8 mirrors and will pass through 6 other components (windows, atmospheric dispersion correctors and beamsplitters). If each component adds a distortion to the wavefront of $\lambda/20$ r.m.s. (quoted at 633nm) and the distortions added by different components are uncorrelated, then the cumulative effect of the optical train will be an r.m.s. differential phase fluctuation across two interfering wavefronts of 0.93 radian at 800nm. This could lead to a very large visibility loss, but because this distortion is constant in time, it can be removed by specially figuring one of the elements to cancel out the errors in the rest of the train. Tango [84] cites an improvement in the wavefront distortion of the Monte Porzio interferometer from 1–2 wavelengths error to around $\lambda/10$ using this technique and so we shall use this latter figure to estimate the visibility loss.

The higher spatial frequency wavefront errors (quoted as *irregularities* by optical manufacturers as opposed to lower frequency errors such as wavefront curvature or astigmatism which are called *figure errors*) will cause light to be scattered out of the main beam by diffraction and this enters into table 9.2 as the power loss due to surface errors. The magnitude of the irregularities has been taken as half that of the figure errors, i.e. about $\lambda/40$ at 633nm.

[8] Diffraction from aperture stops. The finite size of the exit pupils of the beam-reducing telescopes will give rise to Fresnel diffraction effects as the beams propagate towards the correlator. This will cause visibility losses if the paths from the telescopes to the correlator are unequal because the interfering wavefronts will have different amplitude and phase profiles. There will also be a loss in the total power because some of the light is diffracted outside stops further down the optical train. This problem is discussed in depth by Tango and Twiss [83] and the value for the visibility loss is taken from their table 1 assuming a beam diameter of 3cm and a maximum path length difference of 100m. The value for the power loss is taken from their figure 3. In both cases there is assumed to be a stop at the entrance to the correlator of the same diameter as the stop at the exit of the telescopes.

[9] Beamsplitter tolerances. We shall assume that the beamsplitter tolerances specified in chapter 6 can just be reached and that therefore there are 1% visibility losses arising from the imperfect amplitude splitting ratio and from the variation of the phase reflection coefficient with wavelength and with polarisation.

[10] Atmospheric absorption. Allen [3] gives a figure of 93% for the atmospheric transmission at 800nm, but this corresponds to very good atmospheric conditions. We would expect the atmosphere over Cambridge to contain more particulate matter and aerosols than at the best observatories, and so a the value for the losses due to dust is taken here as being roughly double that used by Allen.

[11] Reflection losses. These are calculated assuming that the reflection coefficient of the mirrors is 98% and the losses at each surface of the transmissive components is 1%. An additional loss of 1% per surface for all the components is included to take account of scattering from dust and scratches.

[12] Detector quantum efficiency. This figure is taken from the manufacturers' quoted efficiency for avalanche photodiodes.

The total visibility attenuation is estimated roughly by multiplying together the individual attenuations to give a total attenuation of 43%. Thus for a 4-telescope

system where all the beams are combined in a single fringe pattern we have

$$|V(0)| = 0.43/4 = 0.107$$

Therefore in order to be able to track the fringes the minimum number of detected photons per coherence time is

$$N_0 \approx 90.$$

The flux from a zeroth magnitude star in this time is highly dependent on the atmospheric parameters r_o and t_o. Unfortunately these are highly variable, and in the case of t_o not often measured. Visual seeing estimates on the 36 inch telescope at the Institute of Astronomy in Cambridge have a typical value of 1.5 arcseconds, but this figure must be treated with caution because visual estimates are usually optimistic [75]. Another complication is that the observed seeing will be due in part to dome seeing, which will not be present in COAST because the array elements will be free-standing, but on the other hand there will be extra perturbations due to turbulence along the air paths from the telescopes to the correlator. In the absence of better information we shall be pessimistic and assume that the 'total' seeing is 2 arcseconds. This implies that r_o is 8.8cm at 800nm. For t_o, we shall take the measurement made in La Palma and reported in chapter 8, i.e. $t_o \approx 12ms$ at 633nm wavelength under 1 arcsecond conditions, and scale it linearly with the seeing to give $t_o \simeq 7.9ms$ at 800nm in Cambridge.

Thus for a $2r_o$ diameter aperture, a 10% fractional bandwidth and a total throughput of 20% (see table 9.2), a zeroth magnitude star will give rise to 4.8×10^5 detected photons per t_o. The aforementioned flux limit of 90 photons per t_o will therefore be reached for an object with an I magnitude of 9.3.

This figure is for the total flux emitted by the source (with the proviso that in practice the emitting region must be smaller than the resolution of the smallest baseline — about 0.1 arcsecond for COAST) but the instrument will be capable of detecting features within the source which are much fainter than this. The sensitivity to such features will depend on the dynamic range of the maps produced i.e. the ratio between the brightest source and the weakest believable feature in the map. This will be a function of the signal-to-noise ratio of the data and the magnitude of any systematic errors. The experiments reported chapter 8 suggest that errors in the calibration of the visibility amplitudes will prove to be the greatest problem: if we are observing an object at the flux limit set by fringe tracking requirements, we might achieve a signal-to-noise ratio in the measurement of the amplitudes of greater than 50:1 in about 2 minutes of observation; if the total observation consists of 100 such measurements then any systematic error in the calibration greater than about 0.2% would become significant.

Given our prediction of the limiting magnitude, it will not be possible to observe one of the most interesting classes of sources, active galactic nuclei, since the brightest of these has a magnitude of 11. Even so there is a large number of sources in our own galaxy which could be observed: on average there will be more than 10 stars brighter than the limiting magnitude in a square degree of sky [3]. An excellent discussion of the range of astrophysical problems that could be tackled is given by Davis [23], and so here we shall only mention a few of these briefly.

Binary stars The overlap between the set of binary systems which are spectroscopic and those which can be resolved visually is quite small, whereas COAST will be able to resolve a large number of these systems. The combination of spectroscopic and angular information then allows the determination of the stellar masses and, if the system is a double-lined binary, a direct determine of the distance to the system can be made. This will bring invaluable information to the calibration of the stellar mass scale, and even more importantly the cosmic distance scale.

Stellar diameters COAST will be able to resolve a large number of stellar discs and if an accurate angular diameter can be measured, this can be used to determine the stars' effective temperature. If the star is the member of a double-lined spectroscopic system, the distance information available will provide a direct measurement of the linear diameter. Having simultaneous information about mass, diameter and effective temperature will be a powerful test of the predictions of theoretical models of stellar interiors.

In several classes of source, it is merely sufficient to be able to measure any asymmetry in the stellar disc, which is less demanding on the calibration of the visibilities. In the case of Mira variables, such a measurement will be able to resolve the controversy as to their pulsation mode.

Stellar surfaces For a number of giant stars, it will be possible to obtain an image with several resolution elements across the stellar disc. Indirect information suggests that it may be possible to resolve large starspots and even large chromospheric prominences on these stars. It should be pointed out that the ability to make true images using the closure phase information that COAST will provide will be especially important in such complex situations.

Stellar mass loss and mass transfer Many types of star are known or strongly suspected to have extended atmospheres and COAST will be able to resolve a large fraction of these. Images of these atmospheres will provide much-needed input to the modelling of the complex processes of stellar mass loss. Even more complex are the cases where mass transfer is occurring between the members of

a binary system. COAST will not be able to resolve the closest and therefore most extreme of these types of system, but semi-detached binaries such as Algol are within reach. In both both types of mass loss process, a high dynamic range in the map and the ability to observe in narrow spectral lines will be important, since the atmospheres will tend to be much fainter than the star.

COAST affords a large increase in angular resolution compared with previous techniques. It is therefore impossible to predict exactly what kinds of phenomena will be observed, but this is perhaps the greatest attraction of this instrument.

Appendix A

Photon noise calculations

A.1 The variance of the first order complex visibility estimator

We require to evaluate the two variances defined in equations 2.4 and 2.5 of the estimator $D(\mathbf{u})$:

$$\begin{aligned} \mathrm{var}_1[D(\mathbf{u})] &\equiv E[|D(\mathbf{u})|^2] - |E[D(\mathbf{u})]|^2 \\ \mathrm{var}_2[D(\mathbf{u})] &\equiv E[D(\mathbf{u})^2] - E[D(\mathbf{u})]^2. \end{aligned}$$

The quantity $E[D(\mathbf{u})]$ has already been derived in chapter 2 (see equation 2.3). Goodman and Belsher [33] show that

$$E[|D(\mathbf{u})|^2] = \overline{N} + \overline{N}^2\langle|V(\mathbf{u})|^2\rangle$$

and hence

$$\mathrm{var}_1[D(\mathbf{u})] = \overline{N} + \overline{N}^2\mathrm{var}_1[V(\mathbf{u})] \tag{A.1}$$

We can follow Goodman and Belsher's analysis through for the second variance (the notation is as for chapter 2):

$$\begin{aligned} E(D(\mathbf{u})^2) &= E\left[\sum_{j=1}^{N}\exp(2\pi i\mathbf{u}\cdot\mathbf{x}_j)\sum_{k=1}^{N}\exp(2\pi i\mathbf{u}\cdot\mathbf{x}_k)\right] \\ &= \left\langle\sum_{j=1}^{N}\sum_{k=1}^{N}\int\int\int\int_{-\infty}^{\infty}\exp(2\pi i\mathbf{u}\cdot(\mathbf{x}_j+\mathbf{x}_k))p(\mathbf{x}_j,\mathbf{x}_k)\,d^2\mathbf{x}_j\,d^2\mathbf{x}_k\right\rangle. \end{aligned}$$

Considering the terms of this summation separately, we can see that there are N terms for which $j = k$ and hence $p(\mathbf{x}_j, \mathbf{x}_k) = p(\mathbf{x}_j)\delta(\mathbf{x}_j - \mathbf{x}_k)$ and $N(N-1)$ cases for which $j \neq k$ giving $p(\mathbf{x}_j, \mathbf{x}_k) = p(\mathbf{x}_j)p(\mathbf{x}_j)$ Hence

$$E[D(\mathbf{u})^2] = \overline{N}\langle V(2\mathbf{u})\rangle + \overline{N(N-1)}\langle V(\mathbf{u})^2\rangle.$$

Now it can be shown that if N has a Poisson distribution, as is the case for photon statistics, then

$$\overline{N(N-1)\cdots(N-k+1)} = \overline{N}^k \,,$$

whence

$$E[D(\mathbf{u})^2] = \overline{N}\langle V(2\mathbf{u})\rangle + \overline{N}^2\langle V(\mathbf{u})^2\rangle.$$

Hence we have

$$\mathrm{var}_2[D(\mathbf{u})] = \overline{N}\langle V(2\mathbf{u})\rangle + \overline{N}^2\mathrm{var}_2[V(\mathbf{u})]. \tag{A.2}$$

Combining these results, the variance of $D(\mathbf{u})$ along a direction at an angle θ to the real axis is

$$\begin{aligned}\mathrm{var}(D(\mathbf{u}),\theta) &= \frac{1}{2}[\mathrm{var}_1(D(\mathbf{u})) + Re\{\mathrm{var}_2(D(\mathbf{u}))e^{-2i\theta}\}]. \\ &= \frac{\overline{N}}{2}(1 + Re\{\langle V(2\mathbf{u})\rangle e^{2i\theta}\}) + \overline{N}^2\mathrm{var}[V(\mathbf{u}),\theta]. \end{aligned} \tag{A.3}$$

A.2 The variance of the triple product estimator

If the visibilities are measured from three different fringe patterns, we can write the measured triple product as

$$T_{123} = \sum_{j=1}^{N_{12}}\sum_{k=1}^{N_{23}}\sum_{m=1}^{N_{31}} \exp(2\pi i\mathbf{u}_{12}\mathbf{x}_{j,12})\exp(2\pi i\mathbf{u}_{23}\mathbf{x}_{k,23})\exp(2\pi i\mathbf{u}_{31}\mathbf{x}_{m,31}) \tag{A.4}$$

where $\mathbf{u}_{12}$, $\mathbf{u}_{23}$ and $\mathbf{u}_{31}$ are the relevant fringe frequencies, $\mathbf{x}_{j,12}$ is the position of the jth photon in fringe pattern containing baseline 12 , N_{12} is the number of photo-events in that pattern and so forth. Since the photevents in different patterns are independent, we have

$$p(\mathbf{x}_{j,12},\mathbf{x}_{k,23},\mathbf{x}_{m,31}) = p(\mathbf{x}_{j,12})p(\mathbf{x}_{k,23})p(\mathbf{x}_{m,31})$$

If the visibilities are all measured from the same pattern, the same photons affect all three fringe measurements, but we remove some of the effects of this when we use Wirnitzer's estimator (equation 2.15): this latter can be shown [4] to be equivalent to

$$T_{123} = \sum_{k1=1}^{N}\sum_{k2=1,k_2\neq k_1}^{N}\sum_{k3=1,k_3\neq k_2\neq k_1}^{N} \exp(2\pi i\mathbf{u}_{12}\cdot\mathbf{x}_{k_1})\exp(2\pi i\mathbf{u}_{23}\cdot\mathbf{x}_{k_2})\exp(2\pi i\mathbf{u}_{31}\cdot\mathbf{x}_{k_3}) \tag{A.5}$$

where N is the number of detected photons in the fringe pattern. Comparison of equations A.5 and A.4 show that we expect the properties of the two estimators to be very similar: in the case where the fringe measurements all come from one pattern,

the fact that the summation explicitly avoids correlating a photon with itself will have a similar effect to having the photons coming from different patterns.

We can calculate the variances of these estimators, once again applying the analysis of Goodman and Belsher, this time with the aid of a computer program written in PASCAL to generate all the possible terms and the computer algebra program REDUCE3 to simplify the resulting expression. For the case where the measurements come from three different fringe patterns we have the isotropic component of the variance

$$\begin{aligned}
&\mathrm{var}_1[T_{123}] = \\
&\quad \overline{N}^6\ \mathrm{var}_1[V_{12}(\mathbf{u}_{12})V_{23}(\mathbf{u}_{23})V_{31}(\mathbf{u}_{31})] \\
&+\overline{N}^5(\langle|V_{12}(\mathbf{u}_{12})|^2|V_{23}(\mathbf{u}_{23})|^2\rangle \\
&\qquad + \langle|V_{12}(\mathbf{u}_{12})|^2|V_{31}(\mathbf{u}_{31})|^2\rangle \\
&\qquad + \langle|V_{23}(\mathbf{u}_{23})|^2|V_{31}(\mathbf{u}_{31})|^2)\rangle \\
&+\overline{N}^4(\langle|V_{12}(\mathbf{u}_{12})|^2\rangle \\
&\qquad + \langle|V_{23}(\mathbf{u}_{23})|^2\rangle \\
&\qquad + \langle|V_{31}(\mathbf{u}_{31})|^2)\rangle \\
&+\overline{N}^3
\end{aligned} \tag{A.6}$$

where $V_{12}(\mathbf{u}_{12})$ is the high-light-level visibility of the fringe corresponding to baseline 12 in the relevant fringe pattern and we have assumed for simplicity that the mean photon rate in all three fringe patterns is $\overline{N}$. Component of the variance which is anisotropic in the complex plane is given by

$$\begin{aligned}
&\mathrm{var}_2[T_{123}] = \\
&\quad \overline{N}^6\mathrm{var}_2[V_{12}(\mathbf{u}_{12})V_{23}(\mathbf{u}_{23})V_{31}(\mathbf{u}_{31})] \\
&+\overline{N}^5(\langle V_{12}(2\mathbf{u}_{12})V_{23}(\mathbf{u}_{23})^2V_{31}(\mathbf{u}_{31})^2\rangle \\
&\qquad + \langle V_{23}(2\mathbf{u}_{23})V_{12}(\mathbf{u}_{12})^2V_{31}(\mathbf{u}_{31})^2\rangle \\
&\qquad + \langle V_{31}(2\mathbf{u}_{31})V_{12}(\mathbf{u}_{12})^2V_{23}(\mathbf{u}_{23})^2\rangle) \\
&+\overline{N}^4\langle V_{12}(2\mathbf{u}_{12})V_{23}(2\mathbf{u}_{23})V_{31}(\mathbf{u}_{31})^2\rangle \\
&+\overline{N}^3(\langle V_{12}(2\mathbf{u}_{12})V_{23}(2\mathbf{u}_{23})V_{31}(2\mathbf{u}_{31})\rangle \\
&\qquad + \langle V_{31}(2\mathbf{u}_{31})(V_{12}(2\mathbf{u}_{12})V_{23}(\mathbf{u}_{23})^2\rangle \\
&\qquad + \langle V_{23}(2\mathbf{u}_{23})V_{12}(\mathbf{u}_{12})^2\rangle)
\end{aligned}$$

These expressions can be simplified by making the assumption that $|V(\mathbf{u})| \ll 1$ for all $\mathbf{u} \neq 0$, but that $|V(\mathbf{u})|^2\overline{N}$ has a finite value, which is $\gg 1$ at high light levels and $\ll 1$ (but still much greater than $V(\mathbf{u})$) at low light levels. The region for which $|V(\mathbf{u})|^2\overline{N} \lesssim |V(\mathbf{u})|$ is not encountered in most practical situations because the signal-to-noise ratio is then so low that there is little prospect of recovering any useful information in a finite time. With this approximation, we can eliminate any term which is smaller by a factor of order $|V(\mathbf{u})|$ than another term in the expression. In the case of the two variances of a complex number, we can also eliminate terms for

which there is an equivalent larger term in *either* of the two variances, since the two variances are added together to determine the variance along any given direction.

For the three fringe pattern case, the first part of the variance var_1 is unchanged, but in comparison the anisotropic part of the variance is very small. The only term which is not eliminated by this procedure is the high-light level anisotropic component of the variance, and this will be small in most practical cases, i.e.

$$\mathrm{var}_2[T_{123}] \simeq \overline{N}^6 \mathrm{var}_2[V_{12}(\mathbf{u}_{12})V_{23}(\mathbf{u}_{23})V_{31}(\mathbf{u}_{31})] \simeq 0. \tag{A.7}$$

The result for the case where the measurements all come from the same fringe pattern has been derived by Ayers *et al.* [4], but they are rederived here so as to present them in the form of the two variances introduced in chapter 2. Because in this formalism the part of the variance which is isotropic in the complex plane is separated from the anisotropic part, it is easier to derive those results which are independent of the value of the closure phase, and it is also easier to see how significant

the anisotropies in the noise are.

$$
\begin{aligned}
&\mathrm{var}_1[T_{123}] = \\
&\quad \overline{N}^6 \mathrm{var}_1[V(\mathbf{u}_{12})V(\mathbf{u}_{23})V(-\mathbf{u}_{12}-\mathbf{u}_{23})] \\
&+\overline{N}^5(\langle V(-\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{12})|V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2V^*(\mathbf{u}_{23})\rangle \\
&\quad + \langle V(2\mathbf{u}_{12}+\mathbf{u}_{23})|V(\mathbf{u}_{23})|^2V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{12})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23})V(\mathbf{u}_{23})|V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2V^*(\mathbf{u}_{12})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+2\mathbf{u}_{23})|V(\mathbf{u}_{12})|^2V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{12})|V(\mathbf{u}_{23})|^2V^*(2\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{23})|V(\mathbf{u}_{12})|^2V^*(\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + \langle |V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2|V(\mathbf{u}_{12})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2|V(\mathbf{u}_{23})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{12})|^2|V(\mathbf{u}_{23})|^2\rangle) \\
&+\overline{N}^4(\langle V(-\mathbf{u}_{12}+\mathbf{u}_{23})V(2\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{12}-\mathbf{u}_{23})|V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{12})V^*(\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{12})V^*(\mathbf{u}_{23})\rangle \\
&\quad + \langle V(2\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{23})V^*(\mathbf{u}_{12}+2\mathbf{u}_{23})V^*(\mathbf{u}_{12})\rangle \\
&\quad + \langle V(2\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{12})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23})V(\mathbf{u}_{12}+2\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{12})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23})V(\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{23})V^*(2\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23})V(\mathbf{u}_{23})V^*(\mathbf{u}_{12})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+2\mathbf{u}_{23})V(\mathbf{u}_{12})V^*(2\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+2\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{12})V^*(2\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{23})V^*(\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + \langle |V(2\mathbf{u}_{12}+\mathbf{u}_{23})|^2|V(\mathbf{u}_{23})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{12}+2\mathbf{u}_{23})|^2|V(\mathbf{u}_{12})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{12})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{23})|^2\rangle) \\
&+\overline{N}^3(\langle V(-\mathbf{u}_{12}+\mathbf{u}_{23})V(2\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{12}-\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23})V(\mathbf{u}_{12}+2\mathbf{u}_{23})V^*(2\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle |V(2\mathbf{u}_{12}+\mathbf{u}_{23})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{12}+2\mathbf{u}_{23})|^2\rangle \\
&\quad + 1)
\end{aligned}
$$

and

$$\begin{aligned}
&\mathrm{var}_2[T_{123}] = \\
&\quad \overline{N}^6 \mathrm{var}_2[V(\mathbf{u}_{12})V(\mathbf{u}_{23})V(-\mathbf{u}_{12}-\mathbf{u}_{23})] \\
&+\overline{N}^5(\langle V(2\mathbf{u}_{12})V(\mathbf{u}_{23})^2V^*(\mathbf{u}_{12}+\mathbf{u}_{23})^2\rangle \\
&\quad + \langle V(2\mathbf{u}_{23})V(\mathbf{u}_{12})^2V^*(\mathbf{u}_{12}+\mathbf{u}_{23})^2\rangle \\
&\quad + \langle V(\mathbf{u}_{12})^2V(\mathbf{u}_{23})^2V^*(2\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + 2\langle V(\mathbf{u}_{12})V(\mathbf{u}_{23})|V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + 2\langle V(\mathbf{u}_{12})V(\mathbf{u}_{23})|V(\mathbf{u}_{12})|^2V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + 2\langle V(\mathbf{u}_{12})V(\mathbf{u}_{23})|V(\mathbf{u}_{23})|^2V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle) \\
&\quad + \overline{N}^4(2\langle V(\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{12})V(\mathbf{u}_{23})V^*(2\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + \langle V(2\mathbf{u}_{12})V(2\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})^2\rangle \\
&\quad + \langle V(2\mathbf{u}_{12})V(\mathbf{u}_{23})^2V^*(2\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + 2\langle V(2\mathbf{u}_{12})V(\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{12})\rangle \\
&\quad + \langle V(2\mathbf{u}_{23})V(\mathbf{u}_{12})^2V^*(2\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + 2\langle V(2\mathbf{u}_{23})V(\mathbf{u}_{12})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23})\rangle \\
&\quad + \langle |V(\mathbf{u}_{12}+\mathbf{u}_{23})|^4\rangle \\
&\quad + 2\langle |V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2|V(\mathbf{u}_{12})|^2\rangle \\
&\quad + 2\langle |V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2|V(\mathbf{u}_{23})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{12})|^4\rangle \\
&\quad + 2\langle |V(\mathbf{u}_{12})|^2|V(\mathbf{u}_{23})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{23})|^4\rangle) \\
&+\overline{N}^3(\langle V(\mathbf{u}_{12}+\mathbf{u}_{23})^2V^*(2\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + 2\langle V(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{12})V^*(\mathbf{u}_{23})\rangle \\
&\quad + \langle V(2\mathbf{u}_{12})V(2\mathbf{u}_{23})V^*(2\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + \langle V(2\mathbf{u}_{12})V^*(\mathbf{u}_{12})^2\rangle \\
&\quad + \langle V(2\mathbf{u}_{23})V^*(\mathbf{u}_{23})^2)\rangle
\end{aligned}$$

Applying the approximations introduced above we have

$$\begin{aligned}
&\mathrm{var}_1[T_{123}] \simeq \\
&\quad \overline{N}^6 \mathrm{var}_1[V(\mathbf{u}_{12})V(\mathbf{u}_{23})V(-\mathbf{u}_{12}-\mathbf{u}_{23})] \\
&+\overline{N}^5(\langle |V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2|V(\mathbf{u}_{12})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2|V(\mathbf{u}_{23})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{12})|^2|V(\mathbf{u}_{23})|^2\rangle) \\
&+\overline{N}^4(\langle |V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{12})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{23})|^2\rangle) \\
&+\overline{N}^3
\end{aligned} \tag{A.8}$$

and again

$$\mathrm{var}_2[T_{123}] \simeq \overline{N}^6 \mathrm{var}_2[V(\mathbf{u}_{12})V(\mathbf{u}_{23})V(-\mathbf{u}_{12}-\mathbf{u}_{23})] \simeq 0. \tag{A.9}$$

We can see by comparing equations A.8 and A.6 that the variances of the single and multiple fringe pattern cases are very similar in form at all the light levels that would be encountered in practice.

A.3 Covariances of triple products

We can also calculate the covariances of different triple products. We shall present here only the single fringe pattern case and shall assume that the two triple products share a common baseline. Again we shall only present the results as the calculations

themselves are repetitive and long.

$$
\begin{aligned}
&\mathrm{covar}_1[T_{123}, T_{234}] = \\
&\quad \overline{N}^6 \mathrm{covar}_1[V(\mathbf{u}_{12})V(\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23}), V(\mathbf{u}_{23})V(\mathbf{u}_{34})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})] \\
&\quad \overline{N}^6 \langle V(\mathbf{u}_{12})V(\mathbf{u}_{34})|V(\mathbf{u}_{23})|^2 V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&+\overline{N}^5 (\langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12})V(\mathbf{u}_{34})|V(\mathbf{u}_{23})|^2\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12})|V(\mathbf{u}_{23})|^2 V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{34})|V(\mathbf{u}_{23})|^2 V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23})V(\mathbf{u}_{23})V(\mathbf{u}_{34})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{34})|V(\mathbf{u}_{23})|^2 V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12})V(\mathbf{u}_{23})V(\mathbf{u}_{34})V^*(\mathbf{u}_{12}+2\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12})V(\mathbf{u}_{34})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12})|V(\mathbf{u}_{23}+\mathbf{u}_{34})|^2 V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12})|V(\mathbf{u}_{34})|^2 V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23})\rangle) \\
&+\overline{N}^4 (\langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12}-\mathbf{u}_{23})V(\mathbf{u}_{23})V(\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12}+\mathbf{u}_{34})|V(\mathbf{u}_{23})|^2\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12})V^*(\mathbf{u}_{23})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12})V(\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})|V(\mathbf{u}_{23})|^2\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12}-\mathbf{u}_{23})V(\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12})V^*(\mathbf{u}_{23})V^*(\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{23}+\mathbf{u}_{34})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{23})V(\mathbf{u}_{34})V^*(\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{34})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23})|V(\mathbf{u}_{23}+\mathbf{u}_{34})|^2 V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23})|V(\mathbf{u}_{34})|^2 V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{34})V(\mathbf{u}_{23})V^*(\mathbf{u}_{12}+2\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{34})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{34})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23})V^*(\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12})|V(\mathbf{u}_{23}+\mathbf{u}_{34})|^2 V^*(\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12})|V(\mathbf{u}_{34})|^2 V^*(\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle) \\
&+\overline{N}^3 (\langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12}-\mathbf{u}_{23})V(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12}-\mathbf{u}_{23})V^*(\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{23}+\mathbf{u}_{34})V^*(\mathbf{u}_{12}+2\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{34})V^*(\mathbf{u}_{12}+2\mathbf{u}_{23})V^*(\mathbf{u}_{34})\rangle)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{covar}_2[T_{123}, T_{234}] = \\
&\quad \overline{N}^6 \mathrm{covar}_2[V(\mathbf{u}_{12})V(\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23}), V(\mathbf{u}_{23})V(\mathbf{u}_{34})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})] \\
&+\overline{N}^5(\langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12})V(\mathbf{u}_{23})^2V(\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12})V(\mathbf{u}_{23})^2V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{23})^2V(\mathbf{u}_{34})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{34})V(\mathbf{u}_{23})^2V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(2\mathbf{u}_{23})V(\mathbf{u}_{12})V(\mathbf{u}_{34})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12})V(\mathbf{u}_{23})|V(\mathbf{u}_{23}+\mathbf{u}_{34})|^2V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12})V(\mathbf{u}_{23})|V(\mathbf{u}_{34})|^2V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{23})V(\mathbf{u}_{34})|V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{23})V(\mathbf{u}_{34})|V(\mathbf{u}_{12})|^2V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle) \\
&+\overline{N}^4(\langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{23})V(\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12}+\mathbf{u}_{34})V(\mathbf{u}_{23})^2\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12})V(\mathbf{u}_{23})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(2\mathbf{u}_{23})V(\mathbf{u}_{12})V(\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{23})^2\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(2\mathbf{u}_{23})V(\mathbf{u}_{12})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12})V(\mathbf{u}_{23})V^*(\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(2\mathbf{u}_{23})V(\mathbf{u}_{34})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{23})V(\mathbf{u}_{34})V^*(\mathbf{u}_{12})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{34})V(2\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{34})V(\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{34})V(\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})V^*(\mathbf{u}_{12})\rangle \\
&\quad + \langle |V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2|V(\mathbf{u}_{23}+\mathbf{u}_{34})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{12}+\mathbf{u}_{23})|^2|V(\mathbf{u}_{34})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{23}+\mathbf{u}_{34})|^2|V(\mathbf{u}_{12})|^2\rangle \\
&\quad + \langle |V(\mathbf{u}_{12})|^2|V(\mathbf{u}_{34})|^2\rangle) \\
&+\overline{N}^3(\langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12}+\mathbf{u}_{23})V(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12}+\mathbf{u}_{34})V(2\mathbf{u}_{23})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(2\mathbf{u}_{23})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{23}+\mathbf{u}_{34})V^*(\mathbf{u}_{12})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{34})V^*(\mathbf{u}_{12})V^*(\mathbf{u}_{34})\rangle)
\end{aligned}
$$

Applying the approximations introduced in the previous section, we have

$$\begin{aligned}
&\mathrm{covar}_1[T_{123}, T_{234}] \simeq \\
&\quad \overline{N}^6 \mathrm{covar}_1[V(\mathbf{u}_{12})V(\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23}), V(\mathbf{u}_{23})V(\mathbf{u}_{34})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})] \\
&+\overline{N}^5 \langle V(\mathbf{u}_{12})V(\mathbf{u}_{34})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&+\overline{N}^4 (\langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12})V(\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{34})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})\rangle \\
&\quad + \langle V(\mathbf{u}_{12}+\mathbf{u}_{34})V^*(\mathbf{u}_{12}+\mathbf{u}_{23})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})\rangle) \\
&+\overline{N}^3 (\langle V(-\mathbf{u}_{12}-2\mathbf{u}_{23}-\mathbf{u}_{34})V(\mathbf{u}_{12}+\mathbf{u}_{34})\rangle \\
&\quad + \langle V(-\mathbf{u}_{12}-\mathbf{u}_{23}+\mathbf{u}_{34})V(\mathbf{u}_{12}-\mathbf{u}_{23}-\mathbf{u}_{34})\rangle)
\end{aligned}$$

and

$$\mathrm{covar}_2[T_{123}, T_{234}] \simeq \overline{N}^6 \mathrm{covar}_2[V(\mathbf{u}_{12})V(\mathbf{u}_{23})V^*(\mathbf{u}_{12}+\mathbf{u}_{23}), V(\mathbf{u}_{23})V(\mathbf{u}_{34})V^*(\mathbf{u}_{23}+\mathbf{u}_{34})]$$

We can see by comparing this with the equations presented in the previous section that the covariance is negligible compared with the variance at low light levels, confirming the result from the simpler model in section 2.5.4.

Appendix B

A Bayesian estimator for the fringe amplitude

The following analysis is similar to that of Jaynes [45], but adapted for the case where there are many measurements.

For this problem we set up a model of the data as being n estimates of the fringe amplitude and phase $\{d_k, \theta_k, k = 1..n\}$. These are assumed to be generated by the addition of Gaussian, circularly symmetric complex noise with variance σ^2 to the 'true' data a,$\{\phi_k, k = 1..n\}$, where the amplitude a is constant but the phase ϕ changes between frames in some unpredictable way. The assumption of Gaussian noise can be shown to be approximately correct by considering equation 2.2 as an 'almost' random walk in the complex plane with the number of steps being the number of photons in a frame, but in any case this assumption is the least committal one if we know only the mean and variance of the noise process.

Given this model, the probability density of the measured data given the model values a,$\{\phi_k\}$ is

$$\Pr(\{d_k, \theta_k\}|a, \{\phi_k\}, \sigma) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{n}\left([d_k\cos\theta_k - a\cos\phi_k]^2 + [d_k\sin\theta_k - a\sin\phi_k]^2\right)\right]$$

$$= \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{n} d_k^2 + na^2 + \sum_{k=1}^{n} 2ad_k\cos(\theta_k - \phi_k)\right)\right].$$

Now, assuming a prior probability for the model amplitude and phase as uniform in the complex plane,

$$\Pr(a, \{\phi_k\}) \propto a\,,$$

then Bayes' theorem allows us to derive the probability density for the model data as simply

$$\Pr(a, \{\phi_k\}|\{d_k\}, \{\theta_k\}, \sigma) \propto$$

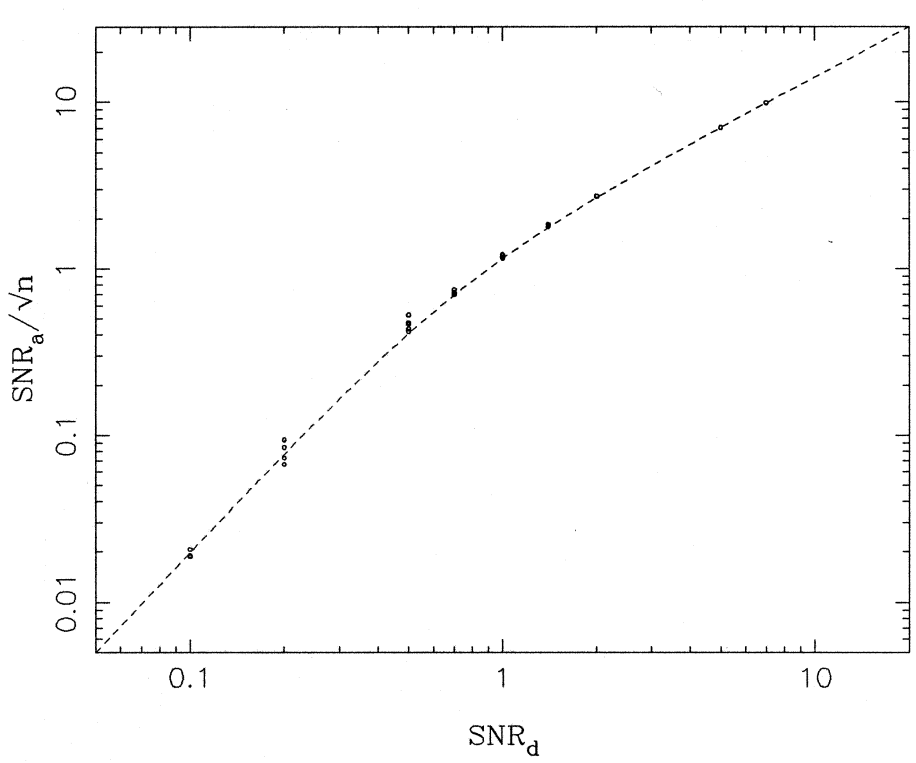

Figure B.1: *The signal-to-noise ratio of the Bayesian solution as a function of the signal-to-noise ratio of the data. The dotted line shows the perfromance of the 'frequentist' estimator A (see equation 2.8).*

$$a \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{n} d_k^2 + na^2 + \sum_{k=1}^{n} 2ad_k \cos(\theta_k - \phi_k)\right)\right].$$

Since we are only concerned with the amplitude information, we marginalise over the unknown model phases $\{\phi_k\}$

$$\begin{aligned}
\Pr(a|\{d_k\},\{\theta_k\},\sigma) \propto & \\
& a \exp\left[\sum_{k=1}^{n}(d_k^2 + na^2)2\sigma^2)\right] \times \\
& \int\int \cdots \int_0^{2\pi} \exp\left(\frac{1}{2\sigma^2}\sum_{k=1}^{n} 2ad_k \cos(\theta_k - \phi_k)\right) \prod_{k=1}^{n} d\phi_k \\
\propto \quad & a \exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(d_k^2 + na^2)\right] \prod_{k=1}^{n} I_o(ad_k/\sigma^2)\,, \qquad \text{(B.1)}
\end{aligned}$$

where I_o is a modified Bessel Function.

In common with all Bayesian methods, the performance of the estimation procedure depends on the data itself, and so this method was tested using computer-generated data with various signal-to-noise ratios. It was found that the *a posteriori*

amplitude probability distribution was approximately Gaussian for large n (i.e. n large enough that the standard deviation of the *a posteriori* distribution was smaller than the mean). Defining the signal-to-noise ratio SNR_a as the ratio of the mean to the standard deviation of this distribution, it can be seen from figure B.1 that the efficiency of this optimal estimator is very similar to that of the simpler estimator A.

Appendix C

The phase error of the sum of a set of phases

C.1 The unit vector method

In this case we are simply adding the phases and no amplitude information is included.

$$\theta_+ = \sum_{k=1}^{m} \theta_k$$

The noise can therefore be represented in terms of a pure phase error with distribution

$$p_+(\epsilon_+) = p_1 * p_2 * \cdots * p_m \,,$$

where the p_k are the error distributions on the individual phases and $*$ denotes convolution — this can be thought of either as 'circular' convolution (i.e. the convolution is done in modulo-2π space) or as a conventional convolution where some of the phase error 'spills over' outside $(-pi, +\pi]$, since the use of sinusoidal functions in our definition of the estimator automatically brings back the modulo-2π nature.

The 'phase error' α_+ can now be derived in terms of the individual phase errors $\{\alpha_k\}$ by making use of the convolution theorem of Fourier transforms. We first recall the definition of the phase error

$$\alpha = \langle s^2 \rangle / \langle c \rangle$$

where

$$\langle c \rangle = \int_{-\infty}^{\infty} p(\epsilon) \cos(\epsilon)\, d\epsilon$$

and

$$\begin{aligned} \langle s^2 \rangle &= \int_{-\infty}^{\infty} p(\epsilon) \sin^2(\epsilon)\, d\epsilon \\ &= \frac{1}{2} - \frac{1}{2} \int_{-\infty}^{\infty} p(\epsilon) \cos(2\epsilon)\, d\epsilon \end{aligned}$$

The convolution theorem therefore gives us

$$\langle c_+ \rangle = \prod_{k=1}^{m} \langle c_k \rangle$$

and

$$1 - 2\langle s_+^2 \rangle = \prod_{k=1}^{m} (1 - 2\langle s_k^2 \rangle).$$

Hence

$$\alpha_+ = \frac{1 - \prod_{k=1}^{m}(1 - 2\langle s_k^2 \rangle)}{2\prod_{k=1}^{m}\langle c_k \rangle}. \tag{C.1}$$

Now if $\alpha_k \ll 1 \forall k$ then $\langle s_k^2 \rangle \ll 1 \forall k$ since $\langle c_k \rangle \leq 1 \forall k$. This gives

$$\alpha_+ \simeq \frac{\sum_{k=1}^{m}\langle s_k^2 \rangle}{\prod_{k=1}^{m}\langle c_k \rangle}.$$

For these small phase errors, $\langle c_k \rangle \simeq 1 \forall k$ to first order, and so

$$\alpha_+^2 \simeq \sum_{k=1}^{m} \alpha_k^2 \tag{C.2}$$

For large phase errors, the p_k are nearly 'flat' distribitions so that

$$\int_{-\infty}^{\infty} p_k(\epsilon_k) cos(2\epsilon_k)\, d\epsilon_k \simeq 0$$

and hence

$$\langle s_k^2 \rangle \simeq 1/2.$$

Substituting into equation C.1 we see that

$$\alpha_+ \simeq 1/2 \prod_{k=1}^{m} 2\alpha_k^2. \tag{C.3}$$

C.2 The amplitude-weighted vector method

Here we multiply together the measured complex numbers in order to obtain an estimate for the sum of the phases

$$R_+ e^{i\theta_+} = \prod d_k e^{i\theta_k}$$

Here we model the data as the sum of the 'true' signal S_k and a noise signal n_k. For simplicity we shall assume that n_k is a circularly symmetric complex noise process with variance σ_k^2, and is uncorrelated between the different phase measurements i.e.

$$\begin{aligned} \langle n_j n_k * \rangle &= \delta_{jk}\sigma_k^2 \\ \langle n_j n_k \rangle &= 0. \end{aligned}$$

Now the phase errors on the individual phases are defined as

$$\beta_k = \sigma_k/\sqrt{2}|S_k|$$

where the factor of $\sqrt{2}$ is due to the definition of σ_k as being the total noise rather than a component in one direction in the complex plane.

To find the noise on the product of the measurements we need to evaluate

$$\begin{aligned}
\mathrm{var}_1[R_+e^{i\theta_+}] &= \left\langle \prod_{k=1}^{m}(S_k+n_k)\prod_{j=1}^{m}(S_j^*+n_k^*)\right\rangle - \prod_{k=1}^{m}|S_k|^2 \\
&= \prod_{l=1}^{m}|S_l|^2\left[\left\langle \prod_{k=1}^{m}\prod_{j=1}^{m}(1+n_k/S_k)(1+n_j^*/S_j^*)\right\rangle - 1\right] \\
&\simeq \prod_{l=1}^{m}|S_l|^2\left[\sum_{k=1}^{m}2\sigma_k^2/|S_k|^2 + \cdots + \prod_{k=1}^{m}\sigma_k^2/|S_k|^2\right].
\end{aligned}$$

where only the terms important at the high and low signal-to-noise ratio extremes have been kept.

It is easy to show that

$$\mathrm{var}_2[R_+e^{i\theta_+}] = 0\,,$$

implying that the noise process on the product is circularly symmetric and that

$$|\langle R_+e^{i\theta_+}\rangle|^2 = \prod_{l=1}^{m}|S_l|^2$$

giving

$$\beta_+^2 \simeq \sum_{k=1}^{m}\beta_k^2 \tag{C.4}$$

for $\beta_k \ll 1 \forall k$ and

$$\beta_+^2 \simeq (1/2)\prod_{k=1}^{m}2\beta_k^2 \tag{C.5}$$

for $\beta_k \gg 1 \forall k$.

Appendix D

The frequencies of the cut-offs in the power spectra of the simulated phase perturbations

The simulated phase perturbations are generated on a finite grid, and so their power spectra will be 'cut off' at some finite upper and lower frequencies. In this appendix we calculate appropriate values for these limits.

The power spectrum of a process $\Phi(f)$ is related to its structure function $D_\phi(t)$ by

$$D_\phi(t) = 2\int_{-\infty}^{\infty} [1 - \cos(2\pi f t)]\Phi(f)\, df\,,$$

so that for

$$D_\phi(t) = (t/t_o)^{5/3}$$

we have

$$\Phi(f) = 5.60 \times 10^{-3} t_o^{-5/3} |f|^{-8/3}. \tag{D.1}$$

The spectrum of the simulated perturbations will have no power at frequencies with magnitude smaller than some lower cut-off f_0 or larger than the upper cut-off at f_1. We can determine an acceptable value for f_1 by requiring that the r.m.s. value of the high frequency perturbations not included in the simulation should be less than, say, 0.1 radian. This can be written as

$$2\int_{f_1}^{\infty} \Phi(f)\, df < 0.1^2.$$

This gives, for the spectrum of equation D.1,

$$f_1 > (1.27 t_o)^{-1}\,,$$

which means that we can get away with a Nyquist sampling interval of $0.64 t_o$, but this would involve interpolating between samples because there would be significant phase

changes between successive samples. It was found simpler to decrease the sampling interval to about $0.2t_o$ and then do without interpolation.

We cannot use the same criterion for choosing the low-frequency cut-off, because the pole in the power spectrum at its origin means that the r.m.s. phase perturbation due to the very low frequency waves is infinite. However our results depend solely on phase *differences* and so we can use as our criterion that the structure function of the simulated perturbations is within some error ϵ_{max} of the 'real' structure function for timescales up to the longest timescale we are interested in, t_{max}. This implies that

$$2\int_{-f_0}^{f_0} \Phi(f)[1 - cos(2\pi f t_{max})]\, df < \epsilon_{max}^2.$$

Now if $f_0 t_{max} \ll 1$ then we can substitute $\cos(z) \simeq 1 - z^2/2$, giving

$$f_0 < (\epsilon_{max}/0.814)^6 (t_{max}/t_o)^{-5} t_{max}^{-1}.$$

The very strong dependence of f_0 on ϵ_{max} and t_{max} means that for an error of less than 0.2 radians on timescales up to $3t_0$ it would be necessary to generate realisations more than 10^6 times as large as the maximum timescale. We can reduce this requirement in two ways: firstly, we expect that it is not as important to maintain accuracy for timescales much longer than t_o because the phases become effectively decorrelated (for the purposes of interferometric systems) as the phase differences become larger than a radian; secondly, the real perturbations that we are trying to simulate will themselves have a cut-off at low frequencies, due either to the finite outer scale size of the turbulence, or, for telescope separations shorter than the outer scale, because the perturbations over the two telescopes will be correlated. We expect that for telescopes separated by a distance D with the turbulence being blown around at some characteristic speed v_{wind}, that there will be a 'break' in the spectrum of the phase differences between the telescopes at a frequency of about $(D/v_{wind})^{-1}$. Thus for $D = 100$m, $v_{wind} = 10$m/s and $t_o = 10$ms we expect the break at $(1000t_o)^{-1}$ and in practice a cut-off at about this frequency was used in the simulations.

For the simulations of the spatial fluctuations, we can apply the same arguments. A circularly symmetric spatial power spectrum $\Phi(s)$ will give a circularly symmetric structure function $D_\phi(r)$ given by

$$D_\phi(r) = 2\int_0^\infty \Phi(s)[1 - J_0(2\pi s r)]2\pi s\, ds\,,$$

so that for

$$D_\phi(r) = 6.88(r/r_o)^{5/3}$$

we have

$$\Phi(s) = 0.0229 r_o^{-5/3} s^{-11/3}. \tag{D.2}$$

Hence the high spatial frequency cut-off s_1 will satisfy

$$\int_{s_1}^{\infty} \Phi(s) 2\pi s \, ds < (0.1\text{radian})^2$$

which gives

$$s_1 > (1.08 r_o)^{-1}.$$

For the low frequency cut-off s_0,

$$2\int_{s_0}^{\infty} \Phi(s)[1 - J_0(2\pi s r_{max})] 2\pi s \, ds < \epsilon_{max}^2.$$

Substituting $J_0(z) \simeq 1 - z^2/4$, we have

$$s_0 < (\epsilon_{max}/2.92)^6 (r_{max}/r_o)^{-5} r_{max}^{-1}.$$

This shows the impracticability of achieving an adequate representation of the long-wavelength fluctuations. However, as mentioned in the main text, we expect that the fluctuations with wavelength longer than about twice the diameter of a given aperture will contribute almost exclusively to the tilts across that aperture, and therefore for simulations of tilt-corrected interferometers, the wavelength of the cut-off can be much reduced. Indeed it was found that the value for the r.m.s. residual fluctuations across tilt-corrected apertures of diameter D given by simulations with $s_0 \lesssim (6D)^{-1}$, agreed to within the statistical errors with the value analytically calculated by Noll [65].

For uncorrected apertures, it was decided to get a rough approximation to the correct fluctuations by adjusting the spectrum of the perturbations until the resulting structure function agreed with the desired structure function for scales of up to and around r_o. Examination of equation 3.9 will convince the reader that any spectrum that fulfils this requirement and furthermore has a structure function much larger than unity for separations larger than r_o will give the correct value for r.m.s. visibility, although this does not guarantee that any other parameters, e.g. the r.m.s. atmospheric noise on the visibilities, will necessarily be correct. The spectrum finally used was

$$\Phi(s) = 0.0190 r_o^{-2} s^{-4}$$

with cut-offs at $s_1 = (0.2 r_o)^{-1}$ and $s_0 = (25.6 r_0)^{-1}$. Figure D.1(b) shows the agreement of the structure function of this simulation with the theoretical structure function. We can see from figure D.1(a) that the empirical spectrum achieves this result by having an excess of longer-wavelength fluctuations so as to add the required tilt, but that it in fact underestimates the magnitude of the shorter wavelength fluctuations. Hence it is best to treat the results from the simulations of uncorrected apertures as a rough guide rather than as definitive answers.

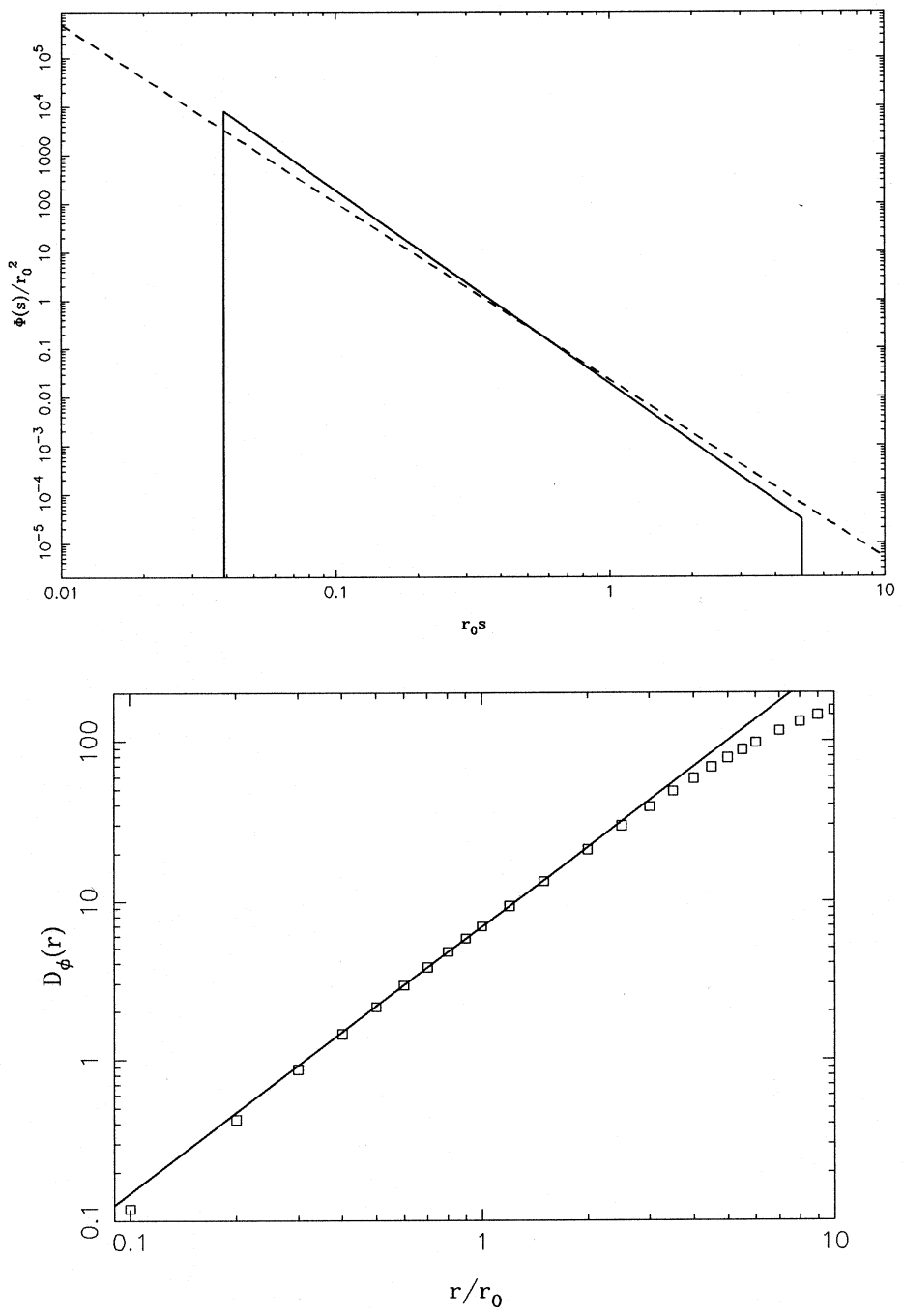

Figure D.1: *The power spectrum (a) and the resulting structure function (b) of the empirically adjusted spatial phase fluctuations (see text). In (a) the full line is the adjusted spectrum and the dotted line is the 'true' spectrum. In (b) the line represents the 'true' structure function and the squares indicate the structure function of the simulated fluctuations when the adjusted spectrum is used.*

Appendix E

The leakage coefficient for a temporally-scanned fringe pattern

This derivation is an extension of the work of Aimé [1], who considers a temproally-scanned *spatial* fringe pattern.

To recap: we are considering a temporal fringe pattern consisting of fringes which are scanned with linear phase ramps of period T, producing fringes with angular frequencies $\{k\omega\}$ where $\omega = 2\pi/T$ and k is an integer. The atmosphere will add an extra phase change $\epsilon_k(t)$ during the scan, giving an intensity pattern

$$i(t) = \sum_{k=0}^{n} \left(V_k e^{i[k\omega t + \epsilon_k(t)]} + V_k^* e^{-i[k\omega t + \epsilon_k(t)]} \right)$$

where V_k is the visibility of the fringe at frequency $k\omega$. For brevity, we shall define

$$V_k(t) \equiv V_k e^{i\epsilon_k(t)}$$

giving

$$i(t) = \sum_{k=0}^{n} \left(V_k(t) e^{ik\omega t} + V_k^*(t) e^{-ik\omega t} \right)$$

The measured visibility at fringe frequency mw will be

$$I_m = (1/T) \int_0^T i(t) e^{-im\omega t}\, dt$$

and thus the mean squared visibility can be written

$$\begin{aligned} T^2 \langle |I_m|^2 \rangle &= \left\langle \int_0^T i(t) e^{im\omega t}\, dt \int_0^T i^*(t') e^{-im\omega t'}\, dt' \right\rangle \\ &= \int_0^T \int_0^T \langle i(t) i^*(t') \rangle e^{im\omega(t'-t)}\, dt\, dt' \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{n}\sum_{j=1}^{n}\int_0^T\int_0^T \langle\Big(V_k(t)e^{ik\omega t} + V_k^*(t)e^{-ik\omega t}\Big) \\
&\quad \times \Big(V_j^*(t')e^{-ik\omega t'} + V_k(t')e^{ik\omega t'}\Big)\rangle e^{im\omega(t'-t)}\,dt\,dt' \\
&= \sum_{k=1}^{n}\sum_{j=1}^{n}\int_0^T\int_0^T \langle(V_k(t)V_j^*(t')e^{ik\omega(t-t')} + V_k^*(t)V_j(t')e^{ik\omega(t'-t)} \\
&\quad + V_k(t)V_j(t')e^{ik\omega(t+t')} + V_k^*(t)V_j^*(t')e^{-ik\omega(t+t')})\rangle e^{im\omega(t'-t)}\,dt\,dt' \qquad \text{(E.1)}
\end{aligned}
$$

Now if the total range of the fluctuations of the fringe phase is large (i.e. $\gg 1$ radian), then

$$\langle V_k(t)V_j(t')\rangle = 0\,,$$

and if furthermore the fluctuations in different fringe patterns are uncorrelated, we have

$$\langle V_k(t)V_j^*(t')\rangle = \delta_{kj}C_k(t-t')\,,$$

where C is the temporal correlation function of the fringes

$$C_k(t) = \langle V_k(t')V_k^*(t'+t)\rangle.$$

If the fluctuations of all the fringes have the same spectrum, we can write

$$C_k(t) = |V_k|^2 C(t).$$

Substituting this into equation E.1 we have

$$T^2\langle|I_m|^2\rangle = \sum_{k=1}^{n}\sum_{j=1}^{n}|V_k|^2\int_0^T\int_0^T C(t'-t)\left[e^{i(k+m)\omega(t'-t)} + e^{i(k-m)\omega(t'-t)}\right]\,dt\,dt'.$$

By transforming the integration variables we can then show that

$$\langle|I_m|^2\rangle = \sum_{k=1}^{n}\sum_{j=1}^{n}|V_k|^2(L_{|k+m|} + L_{|k-m|}) \qquad \text{(E.2)}$$

where L_j is the 'leakage coefficient' for harmonic number j

$$L_j = \frac{2}{T_{scan}}\int_0^{T_{scan}}(1 - t/T_{scan})C(t)\cos[j\omega t]\,dt. \qquad \text{(E.3)}$$

For the Tatarski spectrum of phase perturbations and widely separated telescopes the correlation function will be

$$C(t) = \exp[-(t/t_o)^{5/3}]$$

which gives equation 6.6 in the text.

The leakage coeffiecient for the two-scan integration can be derived in a similar manner, but space (and the average reader's boredom threshold) does not permit its inclusion here.

Appendix F

The correlation between exposures of finite length

We require to evaluate the correlation between the visibilities measured in two frames whose exposure time is T_{exp} and whose exposures begin at times t and t'. We shall write this as

$$C(T_{exp}, t - t') \equiv \langle V(T_{exp}, t) V^*(T_{exp}, t') \rangle$$

The problem is greatly simplified by the fact that we only want answers for the cases where $t - t'$ is a multiple of T_{exp}. We can therefore derive a result from the variation of the mean squared visibility with exposure time, which has already been computed, as follows.

Consider an exposure of length $2T_{exp}$ as being made up of two successive exposures of length T_{exp}

$$V(2T_{exp}, t) = [V(T_{exp}, t) + V(T_{exp}, t + T_{exp})]/2.$$

The mean squared visibility is therefore

$$\begin{aligned} \langle |V(2T_{exp}, t)|^2 \rangle &= \langle [V(T_{exp}, t) + V(T_{exp}, t + T_{exp})] \\ &\quad \times [V^*(T_{exp}, t) + V^*(T_{exp}, t + T_{exp})] \rangle / 4 \\ &= \langle |V(T_{exp}, t)|^2 \rangle / 4 \\ &\quad + \langle |V(T_{exp}, t + T_{exp})|^2 \rangle / 4 \\ &\quad + \langle V(T_{exp}, t) V^*(T_{exp}, t + T_{exp}) \rangle / 4 \\ &\quad + \langle V^*(T_{exp}, t) V(T_{exp}, t + T_{exp}) \rangle / 4 \end{aligned}$$

Given that the atmospheric perturbations are homogenous in time, this reduces to

$$\langle |V(2T_{exp}, t)|^2 \rangle = \langle |V(T_{exp}, t)|^2 \rangle / 2 + C(T_{exp}, T_{exp})/2.$$

Now we know $\langle |V(2T_{exp}, t)|^2 \rangle$ and $\langle |V(T_{exp}, t)|^2 \rangle$ already — they are just the mean squared visibility losses for exposure times of $2T_{exp}$ and T_{exp} respectively and have

been evaluated in section 3.2. Thus we can find the correlation between two frames separated by one exposure time from

$$C(T_{exp}, T_{exp}) = 2\langle |V(2T_{exp})|^2 \rangle - \langle |V(T_{exp})|^2 \rangle.$$

By a similar method we can determine the correlation between two frames separated by twice the exposure time

$$C(T_{exp}, 2T_{exp}) = \left[9\langle |V(3T_{exp})|^2 \rangle - 3\langle |V(T_{exp})|^2 \rangle - 4C(T_{exp}, T_{exp}) \right] /2$$

and so on.

Bibliography

[1] Aime, C., 1978. *Opt. Commun.* **26**, 139–143.

[2] Aime, C., 1986. *Opt. Commun.* **11**, 597–599.

[3] Allen, C.W., 1973. *Astrophysical Quantities*, The Althone Press, London.

[4] Ayers, G.R., Northcott, M.J. & Dainty, J.C., 1988. *J. Opt. Soc. Am. A* **5**, 963–985.

[5] Baer, T., Kowalski, F.V. & Hall, J.L., 1980. *Appl. Opt.* **19**, 3173–3177.

[6] Baldwin, J.E., Haniff, C.A., Mackay, C.D. & Warner, P.J., 1986. *Nature* **320**, 595.

[7] Beckers, J.M., 1986. *Proc. SPIE* **628**, 255–260.

[8] Bennet, S.J., Ward, R.E. & Wilson, D.C., 1973. *Appl. Opt.* **12**, 1406.

[9] Breckenridge, J.B., 1976. *J. Opt. Soc. Am.* **66**, 143.

[10] Brown, N., 1981. *Appl. Opt.* **20**, 3711–3714.

[11] Buscher, D.F., 1988. "Optimising a ground-based optical interferometer for sensitivity at low light levels", *Mon. Not. R. Ast. Soc.* In press.

[12] Caccia, J.L., Azouit, M. & Vernin, J., 1987. *Appl. Opt.* **26**, 1288.

[13] Clark, L.D., Jr., Shao, M., & Colavita, M.M., 1986. *Proc. SPIE* **627**, 838–845.

[14] Colavita, M.M. & Shao, M., 1987. In: *Proc. of the Joint NOAO-ESO Workshop on High Angular Resolution Imaging from the Ground Using Interferometric Techniques*, pp. 205–208, ed. Goad, J.W.

[15] Colavita, M.M., Shao, M. & Staelin D.H., 1988. *Appl. Opt.* **26** 4106–4112.

[16] Connes, P. & Michel, G., 1975. *Appl. Opt.* **14**, 2067–2082.

[17] Cornwell, T.J., 1981. VLA Scientific Memorandum 135.

[18] Cornwell, T.J., 1987. *Astron. Astrophys.* **180**, 269–274.

[19] Coulman, C.E., Vernin, J., Coquegniot, Y. & Caccia, J.L., 1988. *Appl. Opt.* **27**, 155-160.

[20] Dainty, J.C. & Greenaway, A.H., 1979. *J. Opt. Soc. Am.* **69**, 786.

[21] Damé, L. & Faucherre, M., 1987. In: *Proc. ESA Workshop on Optical Interferometry in Space* **ESA SP-273**, 205, eds. Longdon, N. & David, V.

[22] Danchi, W.C. 1988.In: *Proc. NOAO-ESO Conf. on High-Resolution Imaging by Interferometry*, ed. F. Merkle.

[23] Davis, J., 1979. In: *High Angular Resolution Stellar Interferometry*, Proc I.A.U. Coll. **50**, ed. J. Davis, p. 1-1.

[24] Davis, J. & Tango, W.J., 1985. *Proc. Astr. Soc. Australia* **6(1)**, 34–38.

[25] Davis, J. & Tango, W.J., 1985. *Proc. Astr. Soc. Australia* **6(1)**, 38.

[26] Davis, J. & Tango, W.J., 1986. *Nature* **323**, 234.

[27] European Southern Observatory, 1987. *Proposal for the Construction of the 16-m Very Large Telescope.*

[28] Ferguson, J.B. & Morris, R.H., 1978. *Appl. Opt.* **17**, 2924–2929.

[29] Fields, D.R., 1983. *Appl. Opt.* **22**, 645–647.

[30] Foy, R. & Labeyrie, A., 1985. *Astron. Astrophys.* **152**, L29–L31.

[31] Fried, D.L., 1966. *J. Opt. Soc. Am.* **56**, 1372.

[32] Goodman, J.W., 1984. *Statistical Optics*, Wiley, New York.

[33] Goodman, J.W. & Belsher, J.F., 1976. *S.P.I.E. Seminar Proc.* **75**, 141.

[34] Greenaway, A.H., Cheese, D.P., Bregman, J.D. & Noordam, J.E., 1987. In: *Proc. of the Joint NOAO-ESO Workshop on High Angular Resolution Imaging from the Ground Using Interferometric Techniques*, pp. 153–156, ed. Goad, J.W.

[35] Greenwood, D.P, 1977. *J. Opt. Soc. Am.* **67**, 282–290. (see also [29])

[36] Hanbury-Brown, R., 1974. *The Intensity Interferometer*, Taylor & Francis, London.

[37] Haniff, C.A., Mackay, C.D., Titterington, D.J., Sivia, D., Baldwin, J.E. & Warner, P.J. *Nature* **328**, 694.

[38] Haniff, C.A., 1988. " Diffraction limited imaging from the ground at low light levels: possibilities and limitations", In preparation.

[39] Heidbreder, G.R., 1967. *I.E.E.E. Transact. Antennas Propagation* **AP–15**, 90.

[40] Hofmann, K.-H. & Weigelt, G., 1988. *Astron. Astrophys.* **203**, L21.

[41] Hofmann, K.-H., 1988, *Appl. Opt.* **27**, 1943.

[42] Hogge, C.B. & Butts, R.R., 1982. *J. Opt. Soc. Am.* bf72, 606.

[43] Hutter, D.J., Johnston, K.J., Mozurkewich, D., Simon, R.S., Colavita, M.M., Shao, M., Hines, B.E., Staelin, D.H., Hershey, J.L., Hughes, J.A. & Kaplan, G.H., 1988. In: *Proc. NOAO-ESO Conf. on High-Resolution Imaging by Interferometry*, ed. F. Merkle.

[44] Jacobs, S.F. & Small, J.G., 1981. *Appl. Opt.* **20**, 3508–3513.

[45] Jaynes, E.T., 1987. In: *Proc. Third Workshop on Maximum-Entropy and Bayesian Methods* (1983), ed. C. Ray Smith, D. Reidel, Boston.

[46] Jenkins, C.R., 1987. *Mon. Not. R. Ast. Soc.* **226**, 341.

[47] Johnson, M.A., Betz, A.L. & Townes, C.H., 1974. *Phys. Rev. Lett.* **33**, 1617–1620.

[48] Knox, K.T. & Thompson, B.J., 1974. *Astrophys. J.* **193**, L45-L48.

[49] Koechlin, L., 1986. In: *Proc. Colloquium on Kilometric Optical Arrays in Space* **ESA SP-226**, 99, eds. Longdon, N. & Melita, O..

[50] Koechlin, L., 1988. In: *Proc. NOAO-ESO Conf. on High-Resolution Imaging by Interferometry*, ed. F. Merkle.

[51] Kolmogorov, A.N., 1941. *Compt. Rend. (Doklady) de L'Academie des Sciences de l'U.S.S.R.* **31**, 538, translated in *Classic Papers on Statistical Theory*, S.K. Friedlander and L. Topper, Eds, Interscience Publishers, Inc., New York, 1961.

[52] Korff, D., 1973. *J. Opt. Soc. Am.* **63**, 971.

[53] Labeyrie, A., 1970. *Astron. Astrophys.* **6**, 85.

[54] Labeyrie, A., 1975. *Astrophys. J.* **196**, L71–L75.

[55] Labeyrie, A., Koechlin, L., Bonneau, D., Blazit, A., & Foy, R., 1977. *Astrophys. J.* **218**, L75–L78.

[56] Lannes, A., 1987. In: *Proc. of the Joint NOAO-ESO Workshop on High Angular Resolution Imaging from the Ground Using Interferometric Techniques*, ed Goad, J.W., 187–190.

[57] Lohmann, A.W., Weigelt, G. & Wirnitzer, B., *Appl. Opt.* **22**, 4028.

[58] McAlister, H.A., Hartkopf, W.I., Gaston, B.J., Hendry, E.M. & Fekel, F.C., 1984. *Astrophys. J. Suppl. Ser.* **54**, 251–257.

[59] Melles Griot Optics Guide 4, Melles Griot, Aldershot, UK.

[60] Michelson, A.A., 1891. *Nature* **45**, 160–161.

[61] Michelson, A.A., 1921. *Astrophys. J.* **53**, 245–259.

[62] Morris R.H., Ferguson, J.B. & Warniak, J.S., 1975. *Appl. Opt.* **12**, 2808.

[63] Muller, P. & Couteau, P., 1979. *Quatrième Catalogue D'éphemerides D'étoiles Doubles*, Publications de l'observatoire de Paris.

[64] Nisenson, P., & Stachnik, R.V., 1978. *J. Opt. Soc. Am.* **68**, 169–175.

[65] Noll, R.J., 1973. *J. Opt. Soc. Am.* **66**, 207.

[66] Northcott, M.J., Ayers, G.R., & Dainty, J.C., 1988. *J. Opt. Soc. Am. A* **5**, 986–995.

[67] O'Donnell, K.A. & Dainty, J.C., 1980. *J. Opt. Soc. Am.* **70**, 1354.

[68] Papaliolios, C. & Mertz, L., 1982. *Proc SPIE* **331**, 360–365.

[69] Pearson, T.J. & Readhead, A.C.S., 1984. *Ann. Rev. Astron. Astrophys.* **22**, 97–130.

[70] Peck, E.R., 1948. *J. Opt. Soc. Am.* **38**, 66.

[71] Pederson, H., Rigaut, F. & Sarazin, M., 1988. *ESO Messenger* No.53, 8–9.

[72] Readhead, A.C.S., 1987. In: *Proc. of the Joint NOAO-ESO Workshop on High Angular Resolution Imaging from the Ground Using Interferometric Techniques*, ed Goad, J.W.

[73] Readhead, A.C.S., Nakajima, T.S., Pearson, T.J., Neugebauer, G., Oke, J.B., Sargent, W.L.W., 1988. *Astron. J.* **95**, 1278.

[74] Ribak, E., Leibowitz, E. & Hege, E.K., 1985. *Appl. Opt.* bf24, 3094–3100.

[75] Roddier, F., 1981. In: *Progress in Optics*,**19**, 283, ed. Wolf, E., North–Holland, Amsterdam.

[76] Roddier, F., 1986. *Opt. Commun.* **60** 350–352.

[77] Roddier, C. & Roddier, F., 1987. In: *Proc. of the Joint NOAO-ESO Workshop on High Angular Resolution Imaging from the Ground Using Interferometric Techniques*, pp. 25–28, ed. Goad, J.W.

[78] Roddier, F., 1987. *J. Opt. Soc. Am. A* **4**, 1396-1401.

[79] Shaklan, S., 1988. *NOAO Advanced Development Program R & D Note* **88-2**.

[80] Shaklan, S. & Roddier, F., 1988. *Appl. Opt.* **27**, 2334–2338.

[81] Shao, M. & Staelin, D.H., 1980. *Appl. Opt.*,**19**, 1519.

[82] Schneiderman, A.M. & Karo, D.P., 1978. *J. Opt. Soc. Am.* **68**, 338-347.

[83] Tango, W.J. & Twiss, R.Q., 1974. *Appl. Opt.***13**, 1814–1819.

[84] Tango, W.J., 1979. In: *High Angular Resolution Stellar Interferometry*, Proc I.A.U. Coll. **50**, ed. J. Davis, p. 13-1.

[85] Tango, W.J. & Twiss, R.Q., 1980. In: *Progress in Optics*,**17**, 239, ed. Wolf, E., North–Holland, Amsterdam.

[86] Scaddan, R.J. & Walker, J.G., 1978. *Appl. Opt.* **17**, 3379–3784.

[87] Schwab, F.R. & Cotton, W.D., 1983. *Astron. J.* **88**, 688–694.

[88] Sivia, D.S., 1987. *PhD Thesis, University of Cambridge.*

[89] Stern, J.R., Peace, M. & Dyott, R.B., 1970. *Elec. Lett.* **6**, 160–162.

[90] Storey, J.W.V., 1979. In: *High Angular Resolution Stellar Interferometry*, Proc. IAU Coll. **50**, p.20-1.

[91] Street, G., 1973. *PhD Thesis, Cambridge University.*

[92] Tatarski, V.I., 1961. *Wave Propagation in a Turbulent Medium*, Dover, New York.

[93] Umeda N., Tsukiji M. & Takasaki, H., 1980. *Appl. Opt.* **19**, 442–450.

[94] Vernin, J. & Roddier, F., 1973. *J. Opt. Soc. Am.* **63**, 270.

[95] Wilkinson, P.N., 1983. In: *Proc. Int. Conf. VLBI Techniques*, pp375–389, Cepadues-Editions, Toulouse, France.

[96] Winocur, J., 1983. *Appl. Opt.* **22**, 3711.

[97] Wirnitzer, B.,1985 *J. Opt. Soc. Am. A* **2**, 14.

[98] Woan, G. & Duffet-Smith, P.J., 1988. *Astron. Astrophys.* **198**, 375-378.

[99] York LDS Laser Delivery System brochure, York V.S.O.P., Chandler's Ford, Hampshire.