1

Making fringes

1.1 The need for angular resolution

Progress in astronomy is dependent on the development of new instrumentation that can provide data which are better in some way than the data which were available before. One improvement that has consistently led to astronomical discoveries is that of seeing finer detail in objects. In astronomy, the majority of the objects under study cannot easily be brought closer for inspection, so the typical *angular* scales subtended by objects, which depend on the ratio of the typical sizes of the objects to their typical distances from the Earth, are a more useful indicator of how easily they can be seen than their linear sizes alone. The angular separation of two features in a scene which can be just be distinguished from one another is called the *angular resolution* and the smaller this scale is, the more detail can be seen.

The impact of increased angular resolution can be appreciated from comparing the important angular scales of objects of interest with the angular resolution of the instrumentation available at different times in history. Prior to the invention of the telescope, the human eye was the premier 'instrument' in astronomy, with an angular resolution of about 1 arcminute (about 300 microradians). With the notable exceptions of the Sun, Moon and comets, most objects visible in the night sky are 'star-like': they have angular sizes smaller than 1 arcminute and so appear as point-like objects. The first telescopes improved the angular resolution of the naked eye by factors of three to six: Galileo's telescopes are thought to have had angular resolutions of about 10–20 arcseconds (Greco *et al.*, 1993; Strano, 2009) and it became possible to see that planets appear as discs or crescents and have their own moons. Subsequent improvements to telescopes have culminated in telescopes like the Hubble Space Telescope (HST) which have typical angular resolutions of around 50 milliarcseconds (about 250 nanoradians) – better by a factor of more than a thousand than the naked eye.

The increased angular resolution offered by the HST and other high-angular-resolution telescopes has transformed the study of astrophysics. Astronomers have seen many phenomena that were undreamed of even a few decades earlier: young stars surrounded by discs of material left over from their formation, hugely complex filamentary structure in the 'planetary nebulae' surrounding stars at the end of their lives and bright 'cusps' of stellar emission at the centres of galaxies indicating the presence of black holes.

Nevertheless, the angular resolution available with a conventional optical telescope is still inadequate to resolve many important astrophysical phenomena. Amongst the most obvious examples are the following:

Stars – The photospheres of the nearest stars (except for the Sun) are a few milliarcseconds across.

Planet formation – A planet in an Earth-like orbit forming around a star in the nearest star-forming region (around 150 parsecs away) will be about 6 milliarcseconds from its parent star.

Black-hole accretion – The standard model for active galactic nuclei consists of an accreting black hole surrounded by a broad-line region which reprocesses the radiaton emerging from the accretion disc, and a torus of dusty material which can block direct radiation from the accretion disk. The dust tori in the nearest active galactic nuclei have angular diameters of a few milliarcseconds and the broad-line regions are predicted to have sub-milliarcsecond angular radii. The accretion disks themselves are thought to have *micro* arcsecond-scale diameters.

Undeniably, then, there is scope for observing new phenomena if angular resolutions much greater than those available with current telescopes could be achieved.

1.2 The resolution of a single telescope

If a telescope is built so that all optical imperfections are overcome and the distorting effects of the Earth's atmosphere are removed (for example by placing the telescope in space), then the angular resolution of the telescope will be limited by diffraction. This can be understood by considering the observation of a point source of light using such an idealised telescope.

The telescope can be modelled as a perfect lens projecting an image onto a detector as shown in Figure 1.1. The finite size of the telescope is modelled as a circular aperture of diameter d placed in front of the lens. When observing

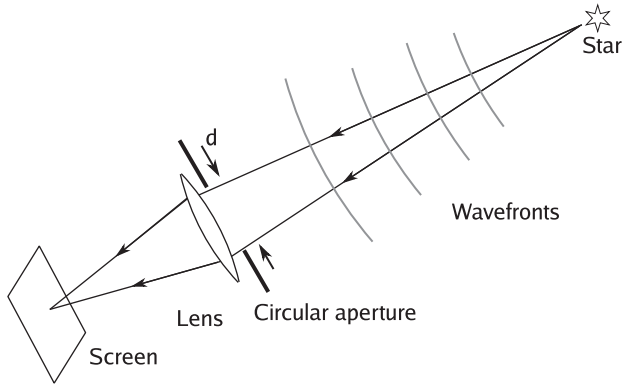

Figure 1.1 A telescope focussing the light from a point source of light (e. g. a star) onto a screen. The telescope is represented as a perfect lens with a circular aperture of diameter d in front of it.

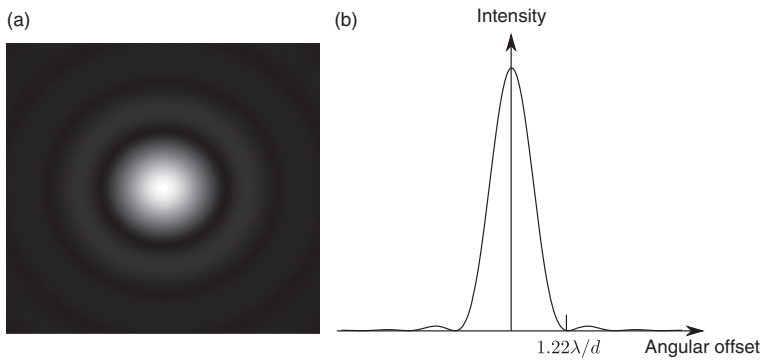

Figure 1.2 The diffraction-limited intensity pattern (known as the 'Airy disc') seen on the screen in the focal plane of the telescope in Figure 1.1 (a) and a cut through the intensity pattern (b).

a point-like object (which will be referred to as a 'star', since most stars are close enough to point-like for these purposes), this arrangement corresponds to a Fraunhofer diffraction experiment. What is seen on the screen is not an infinitely sharp point of light but rather the diffraction pattern of the circular aperture, known as an *Airy disc*, which consists of a central spot surrounded by circular rings as shown in Figure 1.2. This diffraction pattern is known as the *point-spread function* (PSF) of the telescope. Diffraction therefore introduces a finite amount of 'blurring' to the image of the point-like source, even though the lens is modelled as being free from any defects.

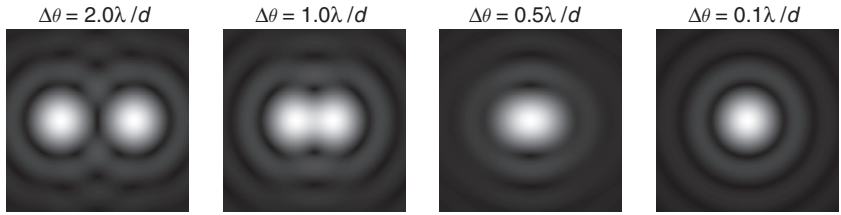

Figure 1.3 Patterns seen in the focal plane of a telescope when pairs of stars of different separations $\Delta\theta$ are observed through a telescope.

The effect of the blurring on angular resolution can be quantified by considering what happens if there is a second star close to the first one as shown in Figure 1.3. The light from the second star will produce a second identically shaped diffraction pattern on the screen, offset by an angular distance $\Delta\theta$ where $\Delta\theta$ is the angular separation of the two stars. Since the phase of the light waves from one star varies randomly and independently of the phase of the light waves from the other star, there is no interference between the two (the justification for this lack of interference is discussed further in Section 1.4.3), and so what is seen on the screen is simply the sum of the intensity patterns that would be seen with either star alone.

If the two stars are brought closer and closer to one another as shown in Figure 1.3, then at some point it becomes impossible to tell whether there is one star or two. At this point the pair of stars is said to be 'unresolved' and the separation at which this occurs is the angular resolution of the telescope. While the exact separation at which the stars become indistinguishable depends on a number of factors such as their relative brightnesses, the *Rayleigh criterion* defines the stars as being 'just resolved' when the peak of the blur pattern produced by one star overlaps with the first null of the blur pattern produced by the second. As shown in Figure 1.2, the angular distance from the peak to the first null of the Airy disc is given by $1.22\lambda/d$, where λ is the wavelength of the light being observed. Thus, the required overlap occurs when

$$\Delta\theta = 1.22\lambda/d. \tag{1.1}$$

By the Rayleigh criterion, Equation (1.1) gives value of the angular resolution of any sufficiently well-corrected telescope. As an example, we can consider the HST, which has a 2.4-m-diameter primary mirror. When observing at a visible wavelength of 500 nm, the diffraction spot from this telescope will have an angular radius of 52 milliarcseconds and so two stars closer together than this cannot be reliably distinguished.

The angular resolution can be improved by building larger telescopes. However, to achieve 1 milliarcsecond resolution at a wavelength of 500 nm would require a telescope 126 m in diameter. Even the largest optical telescopes being proposed at present will have aperture diameters of less than 40 metres, and they come with billion-dollar price tags. The cost of a telescope scales as the square or cube of the aperture diameter so building telescopes more than three times as large seems unlikely in the medium term.

What is required is a method of gaining large factors of improvement in angular resolution which do not require unfeasibly large telescopes. The only known method with this property is long-baseline interferometry. Interferometry uses the interference of light from two or more small telescopes separated by a large distance B to get images with an angular resolution of order λ/B without the mechanical and optical complexities inherent in constructing a single large telescope of diameter B. The next sections serve to show the principles of this method.

1.3 A long-baseline interferometer

An astronomical interferometer collects the light originating from a single region of the sky at two or more locations and brings the collected beams of light together to form an interference pattern. Figure 1.4 shows perhaps the simplest interferometer which could be implemented in practice. Most real interferometers include magnifying and/or demagnifying optics to make the construction of the interferometer easier and cheaper, but the design shown has the advantage that it achieves all the essential functions of an interferometer using only plane (flat) mirrors and so the optical functions of all the elements are readily understandable.

The example interferometer collects starlight using a pair of siderostats, flat mirrors which can be tilted appropriately to reflect the light from a chosen region of sky into a fixed direction. The diameter of the siderostat mirrors can be modest, perhaps only 5 cm if only bright objects are to be observed. In an interferometer used for studying faint objects, the siderostats would typically be replaced by individual telescopes acting as light collectors, each perhaps several metres in diameter.

The distance between the two collectors is typically much larger than the size of any feasible individual collector, perhaps hundreds of metres. The orientation in space of the collector separation is also important: the *baseline vector* $\boldsymbol{B}_{pq}$ between the light-collecting elements p and q of an interferometer is defined as $\boldsymbol{B}_{pq} = \boldsymbol{x}_p - \boldsymbol{x}_q$, where the elements are situated at locations $\boldsymbol{x}_p$

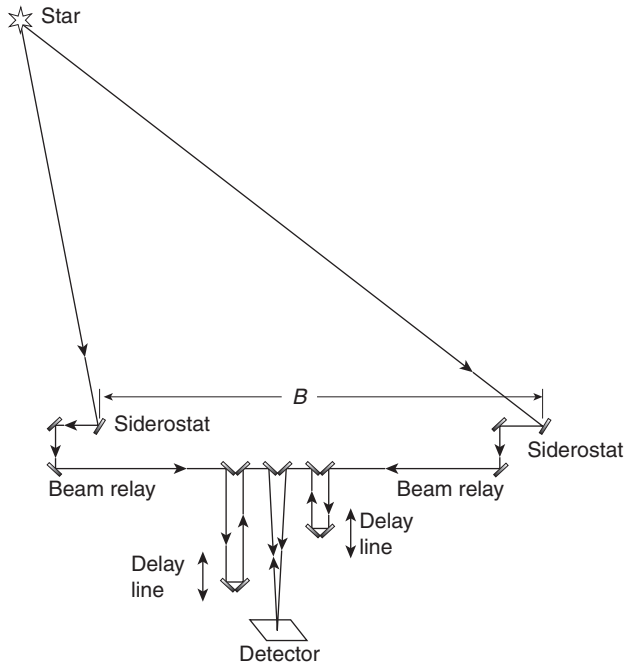

Figure 1.4 A simple long-baseline interferometer constructed out of plane (flat) mirrors. Some of the mirrors have been included to maintain certain symmetries of the optical path – these symmetries are explained in Section 4.4.2.

and $\boldsymbol{x}_q$. In much of the following the pq suffix is dropped, but it will be used later when considering multi-baseline interferometers. The length of the baseline vector is called the 'baseline length' or just the 'baseline' and is shown as B in Figure 1.4.

The collected starlight is brought to a central point using reflections off a series of 'beam-relay' mirrors. Included in the beam path are a pair of mirrors, which can be moved backwards or forwards, acting like an 'optical trombone' to vary the distance the light travels before reaching the central combination point. These 'path compensators' or 'delay lines' serve to control the relative delays between the light beams coming from different collectors: as will be discussed in Section 1.7, in practice it is necessary for the times taken for the light to travel from the object to the point of interference via the two collectors to be matched with one another to in order to see interference.

The light beams are combined in a 'beam combiner' to produce interference fringes. There are many arrangements which can be used to do this: the arrangement shown here uses a so-called 'pupil-plane' arrangement where the

two beams are simply allowed to overlap on a screen. The intensity pattern on the screen can be observed visually but it is more usual to replace the screen by an electronic detector in order to obtain more quantiative information and to observe fainter objects. The detector converts the intensity at each location on its face into an electronic signal, which is digitised, analysed and displayed on a computer.

Interference between the two beams results in a sinusoidal intensity pattern. In the next section it will be demonstrated that this 'fringe pattern' contains information about the size and shape of the object being observed.

1.4 The interferometric measurement equation

The properties of the fringe pattern seen when observing complex objects is derived in the following analysis by considering first the fringe pattern formed when observing a point source, and then how the characteristics of the fringe pattern change when the object consists of two closely spaced point sources. Finally, the properties of the fringe pattern formed when observing an arbitrary object will then be derived by considering it as a collection of closely spaced point sources.

1.4.1 The fringe pattern from a point source

The form of the fringe pattern is derived here using a model of the interferometer which is shown schematically in Figure 1.5. In this model, light arrives from a source of light that is the object of interest. The source is assumed to be a point-like 'star', which is sufficiently distant that the light can be accurately represented as a plane wave, in other words there are plane surfaces known as 'wavefronts' over which the instantaneous electromanetic field E_0 is the same at any given moment in time.

Light propagates from this wavefront along two parallel rays and arrives at the two collectors. The rays then travel via the interferometer optics to the beam combination point, at which point the light waves are superposed and then converted into an intensity $i(x)$. These rays are subject to a series of delays consisting of three different components:

1. An 'external' or 'geometric' delay τ_{ext} due to the light travel time from the wavefront to the collector.
2. An 'internal' delay τ_{int} due to the light travel time along the beam-relay and delay-line beam paths.

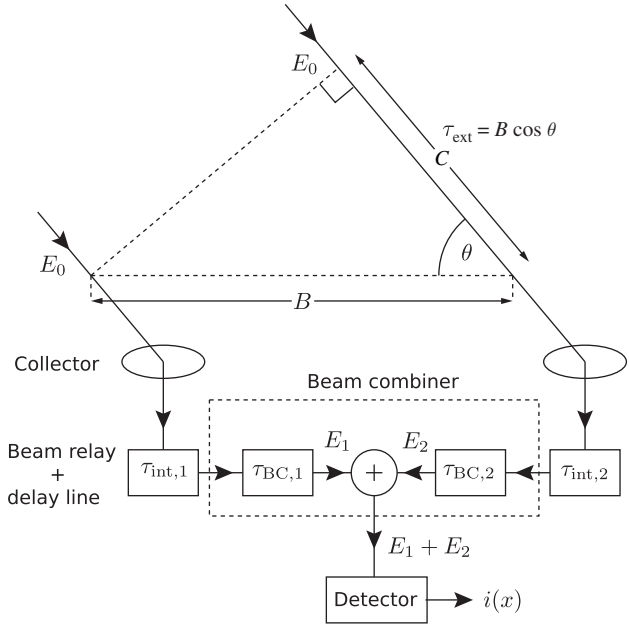

Figure 1.5 A simplified model of fringe formation in an interferometer.

3. A beam-combiner delay $\tau_{BC}(x)$, which is dependent on the location x on the detector on which the beam lands, as shown in Figure 1.6. It will be seen in the following analysis that an important function of any beam-combiner design is to allow the sampling of the interference patterns at multiple locations in 'delay space' and in this case these locations are dependent on the detector coordinate x.

This model neglects the effects of light losses in the interferometer. It also neglects other effects such as optical imperfections along the beam path because it assumes that all rays passing through one collector and arriving at the entrance of the beam combiner experience the same delay. The benefits of using this model are that the analysis is simpler and, perhaps more importantly, the results can be readily applied to interferometers of different designs. For example, the model can be straightforwardly applied to an interferometer which uses temporal coding of the fringes (see Section 4.7) instead of spatial fringes by replacing the detector coordinate x with a time coordinate t.

The analysis starts by considering the electromagnetic field incident on the interferometer. The light is assumed to be perfectly monochromatic so the light wave consists of an electromagnetic wave oscillating at frequency ν and the

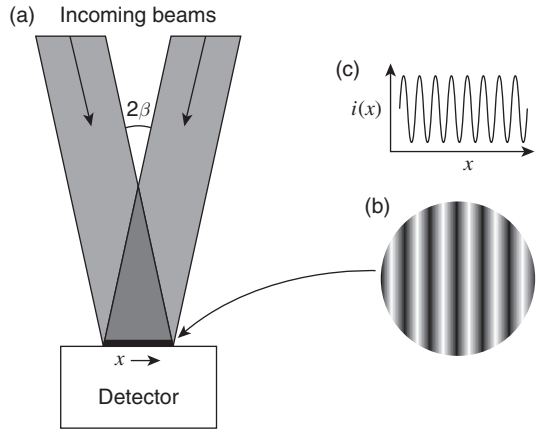

Figure 1.6 (a) Beams of starlight arriving at angles of $\pm\beta$ on a detector. (b) The fringe pattern on the detector. (c) A one-dimensional cut through the fringe pattern.

therefore instantaneous electric field at any point on the initial wavefront is given by

$$E_0 = \frac{2}{\epsilon_0}\mathrm{Re}\left[\Psi_0 e^{-2\pi i\nu t}\right], \tag{1.2}$$

where ϵ_0 is the electric permittivity of free space and Ψ_0 is a 'complex wave amplitude' given by

$$\Psi_0 = |\Psi_0|\, e^{\mathrm{i}\phi_0}, \tag{1.3}$$

where ϕ_0 is the phase of the wave. The electric field is not represented as a vector in this 'scalar wave' analysis. It is assumed that the properties of the system being analysed are the same for any polarisation, and so the vector properties of the electromagnetic field are ignored.

At optical frequencies, the oscillations of the wave are typically not directly observable: optical detectors effectively measure the accumulated light energy received over an 'exposure time' or 'integration time', which can be anywhere from several picoseconds to many minutes, whereas the oscillation period is a few femtoseconds. What is observable is the mean intensity of the wave (i. e. the mean energy crossing unit area per unit time, also known as the 'flux' from the object) given by

$$F_0 = \left\langle \epsilon_0 E_0^2 \right\rangle, \tag{1.4}$$

where $\langle\rangle$ represents averaging over the integration time of the detector. Substituting Equation (1.2) into Equation (1.4) and using the relationship

$$\mathrm{Re}\{X\} = \tfrac{1}{2}(X + X^*), \tag{1.5}$$

where X is any complex number and X^* denotes the complex conjugate of X, gives

$$F_0 = |\Psi_0|^2, \tag{1.6}$$

after dropping terms which average to zero over the exposure time. The simplicity of this expression explains the seemingly arbitrary factor of $2/\epsilon_0$ in Equation (1.2), which serves to define Ψ_0.

In an interferometer, the incident intensity F_0 is not measured directly. Instead, the wave is incident on the two collectors and travels via the beam-relay optics to the beam combiner where the beams are combined and arrive at a given location on the detector surface denoted by a coordinate x. At this location the electric field is given by the superposition of the two fields $E_1(x)$ and $E_2(x)$ that would have been observed from each collector alone. The mean intensity of the light received by the detector at a given location x is therefore given by

$$\begin{aligned} i(x) &= \epsilon_0 \left\langle (E_1(x) + E_2(x))^2 \right\rangle \\ &= \left\langle \left(\mathrm{Re}\left[\Psi_1(x) e^{-2\pi i \nu t} + \Psi_2(x) e^{-2\pi i \nu t} \right] \right)^2 \right\rangle, \end{aligned} \tag{1.7}$$

where $\Psi_1(x)$ and $\Psi_2(x)$ are defined analagously to Ψ_0.

Expanding and dropping terms which average to zero over the integration time of the detector gives

$$i(x) = |\Psi_1(x)|^2 + |\Psi_2(x)|^2 + 2\mathrm{Re}\left[\Psi_1(x)\Psi_2^*(x)\right]. \tag{1.8}$$

The intensity is therefore the sum of the intensities which would be observed on either beam alone, plus a cross term which depends on a product of the two wave amplitudes. This 'interference term' can be positive or negative, corresponding to constructive and destructive interference respectively.

Assuming that the interferometer optics introduce no losses in the light intensity, then the waves arriving at the detector are simply time-delayed versions of the incident wave $E_0(t)$, given by

$$E_1(t) = E_0(t - \tau_{\mathrm{ext},1} - \tau_{\mathrm{int},1} - \tau_{\mathrm{BC},1}(x)) \tag{1.9}$$

and

$$E_2(t) = E_0(t - \tau_{\mathrm{ext},2} - \tau_{\mathrm{int},2} - \tau_{\mathrm{BC},2}(x)). \tag{1.10}$$

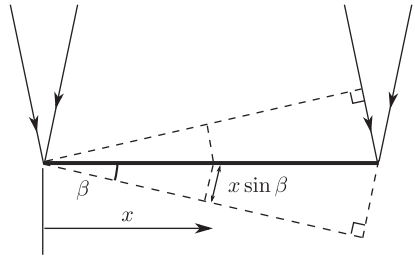

Figure 1.7 Geometry of the relative path lengths travelled by the beams arriving at the detector surface in Figure 1.6 as a function of the coordinate x.

The definition of the oscillating wave amplitude given in Equation (1.2) shows that the time delay between E_0 and E_1 and E_2 corresponds to phase shifts between the incoming wave amplitude Ψ_0 and the delayed wave amplitudes Ψ_1 and Ψ_2, so that

$$\Psi_1 = \Psi_0 e^{2\pi i\nu[\tau_{\text{ext},1}+\tau_{\text{int},1}+\tau_{\text{BC},1}(x)]} \tag{1.11}$$

and

$$\Psi_2 = \Psi_0 e^{2\pi i\nu[\tau_{\text{ext},2}+\tau_{\text{int},2}+\tau_{\text{BC},2}(x)]}. \tag{1.12}$$

Substituting Equations (1.11) and (1.12) into Equation (1.8) gives

$$\begin{aligned} i(x) &= |\Psi_0|^2 + |\Psi_0|^2 + 2|\Psi_0|^2 \text{Re}\left[e^{2\pi i\nu[\tau_{12}+\tau_{\text{BC},1}(x)-\tau_{\text{BC},2}(x)]}\right] \\ &= 2F_0\left(1 + \text{Re}\left[e^{2\pi i\nu[\tau_{12}+\tau_{\text{BC},1}(x)-\tau_{\text{BC},2}(x)]}\right]\right) \end{aligned} \tag{1.13}$$

where τ_{12} is the component of the delay difference, which is independent of the location on the detector:

$$\tau_{12} = (\tau_{\text{ext},1} - \tau_{\text{ext},2}) + (\tau_{\text{int},1} - \tau_{\text{int},2}). \tag{1.14}$$

It should be noted that the usual convention in optics is to express time-delay differences such as τ_{12} in terms of the equivalent *optical path difference* (OPD), which is the time-delay difference multiplied by the speed of light in a vacuum c. An OPD of 1 m corresponds to approximately 3.3 nanoseconds (ns) of delay difference and an OPD of 1 micron (μm) corresponds to 3.3 femtoseconds (fs). (For those who prefer to use Imperial units, 1 ns of delay corresponds quite closely to 1 ft of OPD.) In this book OPD and delay will both be used, depending on which is most convenient: in almost all cases when numerical values are quoted they will be in microns of OPD.

For the pupil-plane beam combiner given in the example interferometer, the beams arrive at the detector with angles of incidence of $\pm\beta$ at the detector, as shown in Figure 1.6(a). Figure 1.7 shows that the rays hitting the detector

surface at location x travel an extra distance $\pm x \sin\beta$ compared to the rays hitting at location $x = 0$. The OPD between the two beams therefore varies with x as

$$c\left(\tau_{\mathrm{BC},1}(x) - \tau_{\mathrm{BC},2}(x)\right) = 2x\sin\beta. \tag{1.15}$$

The intensity on the detector is therefore a sinusoidal pattern given by

$$i(x) = 2F_0\left(1 + \mathrm{Re}\left[e^{\mathrm{i}(2\pi s x + \phi_{12})}\right]\right), \tag{1.16}$$

where s is the 'fringe frequency' or the 'spatial frequency of the fringes' and is given by

$$s = 2\nu\sin\beta/c = 2\sin\beta/\lambda, \tag{1.17}$$

where $\lambda = c/\nu$ is the wavelength of the radiation, and ϕ_{12} is a phase offset given by

$$\phi_{12} = 2\pi\nu\tau_{12}. \tag{1.18}$$

This pattern is shown in Figure 1.6 and is known as a 'fringe pattern'. The alternating dark and light stripes are called 'fringes' and are the characteristic sign of interference – when light from one of the collectors is blocked off, the 'stripes' will disappear leaving a uniform illuminated disc.

The peak-to-peak spacing of the fringe pattern is given by $1/s = \lambda/(2\sin\beta)$. Since λ is on the order of a micron, quite narrow angles of incidence β are needed in order to yield macroscopic-sized fringes: for $\lambda = 0.5\,\mu\mathrm{m}$, then $\beta =$ 50 arcseconds will give a spacing between successive dark fringes of approximately 1 mm. This means that the distance between the beam-combining mirrors and the detector needs to be of order 100 m or more for 5-cm-diameter beams. In practice, different architectures of beam combiner, for example those employing beamsplitters or beam-reducing optics are used in order to allow the use of sensible-sized optics and optical paths (see Section 4.7).

1.4.2 Astrometric phase

The phase shift ϕ_{12} of the fringes depends on the external delay difference due to the difference in the optical paths from the star to the two collectors and the internal delay difference due to the difference in the optical paths from the collectors to the beam combiner. As shown in Figure 1.5, light from a star at infinity travels an additional distance external to the interferometer to get to one collector compared to the other. The delay difference due to this additional light path is given by

$$\tau_{\mathrm{ext},12} = \tau_{\mathrm{ext},1} - \tau_{\mathrm{ext},2} = B\cos\theta/c \tag{1.19}$$

where B is the baseline and θ is the angle between the direction to the star and the baseline vector.

The delay lines can be used to adjust the internal delays so that the net delay τ_{12} and hence the fringe phase shift ϕ_{12} is zero for a star in a particular direction θ_0. If the delay is now kept fixed and a star offset from this 'phase centre' is observed, the OPD will be

$$\begin{aligned} \tau_{12} c &= B\cos(\theta_0 + \Delta\theta) - B\cos\theta_0 \\ &\approx -\Delta\theta B \sin\theta_0, \end{aligned} \tag{1.20}$$

where $\Delta\theta$ is the (small) angular offset of the star from the phase centre. The phase shift of the fringes will therefore be given by

$$\phi_{12} = -2\pi u \Delta\theta, \tag{1.21}$$

where u is the length of the projection of the baseline in the direction perpendicular to the star, scaled in units of wavelengths, i. e.,

$$u = B\sin\theta_0/\lambda. \tag{1.22}$$

Importantly, then, we can see that *the phase of the fringes is sensitive to the angular position of the source and this angular sensitivity increases with the length of the projected baseline.*

To get a numerical idea of this sensitivity, we can consider an interferometer operating at a wavelength of $\lambda = 500\,\text{nm}$ and a collector spacing of $B = 50\,\text{m}$. For a phase centre which is nearly overhead so that $\sin\theta_0 \approx 1$ if the star position is shifted by $\Delta\theta \approx 1$ milliarcsecond (about 5 nanoradians) the phase of the fringes will shift by 180°, so that where there was a bright fringe there is now a dark fringe.

An interferometer can therefore act as an 'angle meter', turning small changes in position of an object into easily visible shifts in the fringe pattern. At radio wavelengths, interferometry is the premier means of precise measurement of angular positions of celestial objects (known as astrometry), achieving sub-milliarcsecond precision (Ma *et al.*, 1998). A number of instruments using interferometry for astrometry at optical wavelengths have been proposed and built (Hummel *et al.*, 1994; Armstrong *et al.*, 1998; Colavita *et al.*, 1999; Shao, 1998; Launhardt *et al.*, 2007) but for a number of instrumental reasons single-telescope methods are still predominant in astrometry.

1.4.3 The fringe pattern from two point sources

Figure 1.8 shows the geometry for the light arriving at an interferometer from two distant point-like objects or 'stars'. One star has an incident electric field

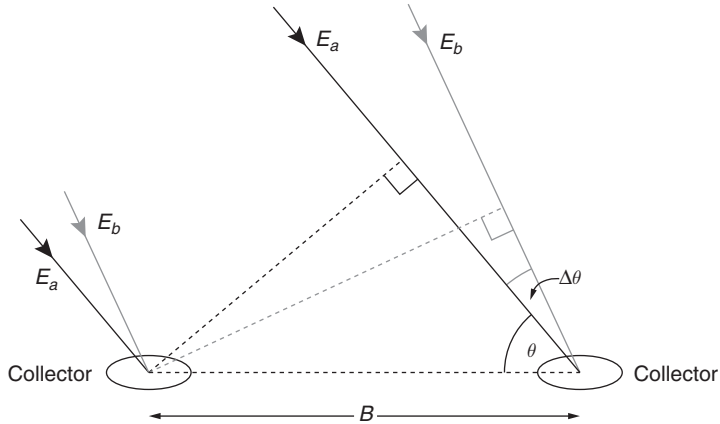

Figure 1.8 Light arriving from two stars with separation $\Delta\theta$.

E_a and is at the phase centre of the interferometer and the second star has an electric field of E_b and is at a small angular offset of $\Delta\theta$.

If $\Delta\theta$ is sufficiently small, four beams of light will arrive at the detector, one from each star via each collector. To a good approximation all the beams will fall directly on top of one another. In this case, six sets of interference effects need to be considered, as there are six possible pairwise combinations of four beams.

Superposition of intensities

In reality only the interference between pairs of beams originating from the same star will be seen. The reason for this is that the light emitted from a natural object is not a perfectly stable sinusoid as has been assumed up to this point. It is necessary instead to consider what is called a *quasi-monochromatic* light beam, where the wave amplitude Ψ is not constant but is varying randomly on a timescale which is long compared to any differential delays in the system, but short compared to the detector integration time. Any thermal source of light filtered with a narrow bandpass filter will fit this model, but so would a standard laboratory laser, which although relatively 'pure', shows random phase variations on timescales of less than a microsecond. These random variations in the wave amplitude cause the interference effects between different sources to be 'washed out' because they are uncorrelated between sources.

This can be shown by revisiting the expression for the intensity of the combination of two beams in Equation (1.7), but considering the case where Ψ_1 and Ψ_2 can have amplitudes and phases that vary randomly with time. In this case of the light pattern at the detector surface will be given by

$$i(x) = \left\langle |\Psi_1(x)|^2 \right\rangle + \left\langle |\Psi_2(x)|^2 \right\rangle + 2\left\langle \mathrm{Re}\left[\Psi_1(x)\Psi_2^*(x)\right]\right\rangle, \tag{1.23}$$

where the angle brackets denote averaging over the exposure time of the detector. Writing

$$\Psi_1 = \Psi_a e^{-2\pi i\nu\tau_a(x)} \tag{1.24}$$

and

$$\Psi_2 = \Psi_b e^{-2\pi i\nu\tau_b(x)} \tag{1.25}$$

where Ψ_a and Ψ_b are the randomly varying complex amplitudes of the incoming waves measured at a fixed point external to the interferometer and $\tau_a(x)$ and $\tau_b(x)$ represent time delays which are fixed during the exposure but can be different for different values of x. This gives an expression for the intensity

$$i(x) = \left\langle |\Psi_a|^2 \right\rangle + \left\langle |\Psi_b|^2 \right\rangle + 2\mathrm{Re}\left[\left\langle \Psi_a\Psi_b^* \right\rangle e^{2\pi i\nu[\tau_a(x)-\tau_b(x)]}\right]. \tag{1.26}$$

The first two terms are independent of x and so represent uniform illuminations corresponding to the intensity contributions from beams 1 and 2. They consist of time averages of positive quantities, and so are always finite and positive. The third term is a cross-term representing the effects of interference between the beams.

In the case where each beam arises from different stars, Ψ_a and Ψ_b will be uncorrelated in both amplitude and phase, since the variations are due to the spontaneous emission events in atoms in different stars. As a result $\Psi_a\Psi_b^*$ will have a phase which varies randomly during the exposure and so this term will average to zero. The intensity pattern will therefore simply be the sum of the intensity patterns that would be observed if each star were present individually and will contain no cross-term between stars. This can be explained in terms of constructive and destructive interference between the star beams occurring randomly and with equal probability during the exposure time: the average of this will be as if no interference occurred at all.

If instead the two beams come from the same star, then $\Psi_a = \Psi_b$ and so $\Psi_a\Psi_b^* = |\Psi_a|^2$, which is a randomly varying but always positive quantity. This does not average to zero, and gives

$$i(x) = 2\left\langle |\Psi_a|^2 \right\rangle \left[1 + \mathrm{Re}\left[e^{2\pi i\nu[\tau_a(x)-\tau_b(x)]}\right]\right], \tag{1.27}$$

which recovers the same form for the interference pattern as given in Equation (1.16) by defining $F_0 = \left\langle |\Psi_a|^2 \right\rangle$.

The combined fringe pattern

The previous subsection showed that the fringe intensity pattern seen when observing two close stars will be given by the sum of the fringe patterns from

the two stars individually. In Section 1.4.2 it was shown that the fringe pattern of the second star will have a phase shift of $-2\pi u\Delta\theta$ where u is the scaled and projected baseline defined in Equation (1.22), and so the intensity pattern on the detector will be given by

$$i(x) = F_a\left(1 + \mathrm{Re}\left\{e^{2\pi \mathrm{i} sx}\right\}\right) + F_b\left(1 + \mathrm{Re}\left\{e^{2\pi \mathrm{i}(\Delta\theta u + sx)}\right\}\right) \tag{1.28}$$

$$= F_a + F_b + \mathrm{Re}\left\{\left(F_a + F_b e^{2\pi \mathrm{i}\Delta\theta u}\right) e^{2\pi \mathrm{i} sx}\right\} \tag{1.29}$$

where F_a and F_b are the respective fluxes of the two stars. A factor of 2 in the definition of $i(x)$ has been dropped between Equation (1.16) and Equation (1.28) in order to simplify the mathematics; in any case, the absolute intensity of the fringe pattern is less important than the relative intensities of the various components of the pattern.

Equation (1.29) shows that the intensity pattern consists of a position-independent 'DC' term (in analogy to the distinction between DC and AC in electrical systems) $F_a + F_b$ and a position-dependent sinusoidal term at frequency s whose amplitude and phase depend on a complex factor $\left(F_a + F_b e^{2\pi \mathrm{i}\Delta\theta u}\right)$. The DC term therefore depends only on the fluxes of the two stars, while the sinusoidal term depends both on the fluxes of the two stars and, importantly, on their angular separation $\Delta\theta$.

This dependence on separation can be illustrated by considering the case where the fluxes of the two stars are identical, i. e. $F_a = F_b$. If the angular separation of the stars is small so that $\Delta\theta u \ll 1$ then the amplitude of the sinusoidal fringe will be equal to the DC level, and strong fringes will be seen. If instead the stars are separated by an angular separation such that $\Delta\theta = 1/(2u)$ then the fringe patterns corresponding to the individual stars will be 180° out of phase with one another and the amplitude of the summed fringe will be zero. In other words, the maxima of one fringe pattern will overlap with the minima of the other and, since the stars are of equal intensity, the fringes will vanish.

1.4.4 Relationships between the fringe pattern and source structure

The preceding analysis of the fringe pattern seen for a pair of stars illustrates a number of key facts about the interferometer:

- The appearance of the interference pattern is sensitive to the *angular structure* of the object under study: there is an observable difference in the properties of the fringe patterns seen for a single star of flux $F_a + F_b$ and a pair of stars with the same total flux, providing the stars are appropriately spaced.

- The angular scale of the structure to which the experiment is sensitive is of order λ/B radians when the source is overhead.
- Comparing this with Equation (1.1), it can be seen that the angular scales to which an interferometer is sensitive are comparable to the angular resolution of a telescope of diameter B, but B can easily be hundreds of metres, much larger than the size of any existing or planned telescope. For $B = 100\,\text{m}$ and a wavelength $\lambda = 500\,\text{nm}$, λ/B is approximately 1 milliarcsecond.

The last of these points represents the key advantage of interferometers: we can get the angular resolving power of a telescope of size B by using the interference of light from two small collectors separated by a distance B rather than by building a single large collector of size B. This means that B can be extended to sizes larger than the diameter of any feasible single telescope so that previously unattainable angular resolutions can be achieved.

The following section extends the analysis to objects of arbitrary shape, but the above results will be found to be a good guide to the basic features of interferometry.

1.4.5 The fringe pattern from an arbitrary object Vector formulation

The interferometric observation of most interest is one where the object being observed has an arbitrary angular structure. The two-dimensional angular structure of the emission from an object (this can be thought of as what the object 'looks like' from the point of view of a perfect observer) can be characterised by the *object brightness distribution* denoted as $I(\hat{\boldsymbol{S}})$, where $\hat{\boldsymbol{S}}$ is a unit vector representing a particular direction as seen from the position of an observer and where the flux coming from within a small solid angle $d\Omega$ of a direction $\hat{\boldsymbol{S}}$ is given by $I(\hat{\boldsymbol{S}})\,d\Omega$.

The external delay difference for a point source at location $\hat{\boldsymbol{S}}$ given in Equation (1.19) can be written in vector notation as

$$\tau_{\text{ext},12} = \boldsymbol{B}_{12} \cdot \hat{\boldsymbol{S}}/c, \tag{1.30}$$

where $\boldsymbol{B}_{12}$ is the baseline vector, as shown in Figure 1.9. If the delay line is adjusted to give zero OPD for a phase centre in direction $\hat{\boldsymbol{S}}_0$, then

$$\tau_{\text{int},12} = -\boldsymbol{B}_{12} \cdot \hat{\boldsymbol{S}}_0/c, \tag{1.31}$$

and so the net delay for light beams arriving at the beam combiner from a star in direction $\hat{\boldsymbol{S}}$ is given by

$$\tau_{12} = \boldsymbol{B}_{12} \cdot \boldsymbol{\sigma}/c \tag{1.32}$$

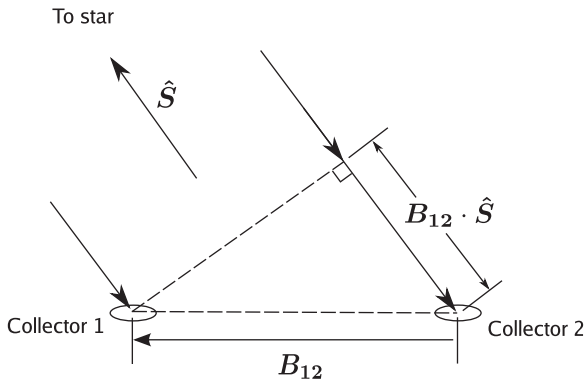

Figure 1.9 Geometry of the paths travelled by light beams from a distant star to two light collectors. The drawing is in the plane containing the vector baseline between the two collectors $\boldsymbol{B_{12}}$ and the unit vector $\hat{\boldsymbol{S}}$ pointing towards the star.

where $\boldsymbol{\sigma} = \hat{\boldsymbol{S}} - \hat{\boldsymbol{S}}_0$ is the offset between the direction to the point source and that to the phase centre. The phase shift of the fringes generated by such a source is therefore given by

$$\phi_{12} = 2\pi \boldsymbol{u} \cdot \boldsymbol{\sigma} \tag{1.33}$$

where $\boldsymbol{u} = \boldsymbol{B}_{12}/\lambda$ is the vector baseline in units of the wavelength.

The (u, v) plane

It is conventional to represent the star and baseline vectors in a right-handed coordinate system with the z axis pointing towards the phase centre, the x axis running towards the east and the y axis running towards the north as shown in Figure 1.10. Writing $\boldsymbol{u} = (u, v, w)$ and $\boldsymbol{\sigma} = (l, m, n)$ gives

$$\boldsymbol{u} \cdot \boldsymbol{\sigma} = ul + vm + nw. \tag{1.34}$$

This expression can be simplified for small fields of view such that $l, m \ll 1$. This is almost always the case in optical interferometry, where the field of view is usually a fraction of an arcsecond so $l, m \lesssim 10^{-6}$. Since $\hat{\boldsymbol{S}}$ and $\hat{\boldsymbol{S}}_0$ both lie on the surface of a unit sphere, then

$$n \approx \frac{1}{2}(l^2 + m^2) \ll |\boldsymbol{\sigma}| \tag{1.35}$$

and so

$$\boldsymbol{u} \cdot \boldsymbol{\sigma} \approx ul + vm. \tag{1.36}$$

From here on, the z coordinates of both the baseline and the source position will be dropped, writing $\boldsymbol{u} = (u, v)$ and $\boldsymbol{\sigma} = (l, m)$, so that $\boldsymbol{u}$ and $\boldsymbol{\sigma}$ represent

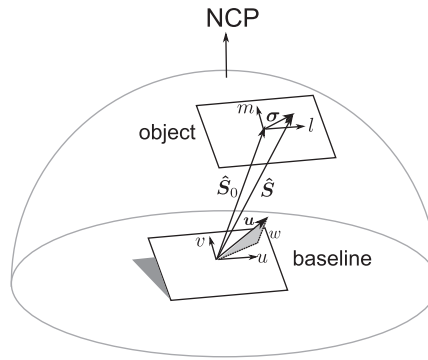

Figure 1.10 Three-dimensional geometry for an interferometric observation, showing the vector baseline $\boldsymbol{u} = \boldsymbol{B}/\lambda$, the phase centre of the image $\hat{\boldsymbol{S}}_0$ and the offset σ. The (u, v, w) coordinate system of the baseline and the (l, m, n) coordinate system for the object are shown (the n coordinate is too small to be visible). The m coordinate points towards the north celestial pole (NCP).

projections on the plane perpendicular to $\hat{\mathbf{S}}_0$ of the baseline vector and the sky coordinates respectively. These projected coordinates are said to lie in the '(u, v) plane' (sometimes known as the 'aperture plane' or the 'Fourier plane') and the 'tangent plane', respectively.

The interferometric measurement equation

An object with a brightness distribution $I(\boldsymbol{\sigma})$ can be represented by a grid of point sources of light spaced by small distances dl and dm with the flux from the point source at position σ being given by $I(\boldsymbol{\sigma})\, dl\, dm$. As in the case of the pair of sources, each point source gives rise to its own fringe pattern and so the total intensity is the sum over these fringe patterns. In the limit of an infinitely fine grid so that $dl \to 0$ and $dm \to 0$, this sum can be expressed as an integral and the fringe pattern intensity is given by

$$i(x) = \iint_{-\infty}^{\infty} I(\boldsymbol{\sigma}) \left(1 + \mathrm{Re}\left\{e^{-2\pi \mathrm{i} \boldsymbol{\sigma}\cdot\boldsymbol{u}} e^{2\pi \mathrm{i} s x}\right\}\, dl\, dm\right) \tag{1.37}$$

$$= 2 \iint_{-\infty}^{\infty} I(\boldsymbol{\sigma})\, dl\, dm + \mathrm{Re}\left\{e^{2\pi \mathrm{i} s x} \iint_{-\infty}^{\infty} I(\boldsymbol{\sigma}) e^{-2\pi \mathrm{i} \boldsymbol{\sigma}\cdot\boldsymbol{u}}\, dl\, dm\right\}, \tag{1.38}$$

where the integration limits have been taken to infinity on the assumption that $I(\boldsymbol{\sigma})$ falls to zero outside some compact region.

The integral can be simplified by defining a complex quantity called the *coherent flux* $F(\boldsymbol{u})$ given by

$$F(\boldsymbol{u}) = \iint_{-\infty}^{\infty} I(\boldsymbol{\sigma}) e^{-2\pi \mathrm{i} \boldsymbol{\sigma}\cdot\boldsymbol{u}}\, dl\, dm. \tag{1.39}$$

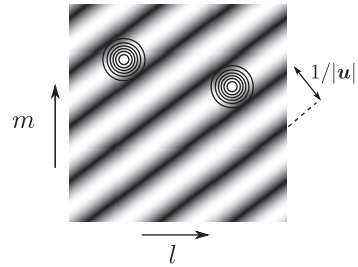

Figure 1.11 Greyscale representation of a two-dimensional sine wave at spatial frequency $\boldsymbol{u}$ overlaid with a contour map of the brightness distribution of a binary star. The coherent flux is an integral of the product of the sine wave and the brightness distribution and so will 'pick' up a binary star with this separation.

Noting that the total flux from the object can be written in terms of a coherent flux on a baseline of zero length

$$F(0) = \iint_{-\infty}^{\infty} I(\boldsymbol{\sigma})\, dl\, dm \tag{1.40}$$

(hence the term *zero-spacing flux*), then the integral can be written

$$i(x) = F(0) + \mathrm{Re}[F(\boldsymbol{u})e^{2\pi \mathrm{i} sx}]. \tag{1.41}$$

Examination of Equation (1.41) shows that the coherent flux controls three measurable properties of the fringe pattern: the zero-spacing flux $F(0)$ controls the average ('DC') level of illumination, the modulus of $F(\boldsymbol{u})$ determines the amplitude of the fringe modulation, and the argument of $F(\boldsymbol{u})$ determines the phase shift of the fringes with respect to some reference point on the detector.

At the same time, the coherent flux depends on the angular structure of the object brightness distribution $I(\boldsymbol{\sigma})$: Equation (1.39) shows that it is an integral across the field of view of the object brightness multiplied by a complex sinusoid. The real and imaginary parts of the complex sinusoid are a cosine and sine wave respectively. Each of these oscillates between positive and negative values over an angular distance on the sky of $1/|\boldsymbol{u}|$ radians as shown in Figure 1.11, and so the integral will have a large magnitude for objects which have structure on this angular scale.

Another qualitative example of how the object structure can affect the integral is shown in Figure 1.12. The integral will tend to cancel out for an object which is smooth on an angular scale of $1/|\boldsymbol{u}|$ while it will tend to be higher for an object with the same flux which is smaller than this scale. Thus the coherent flux can be thought of as the flux of the object, filtered to 'pick out' structure in the object on angular scales of order $\lambda/B_{\mathrm{projected}}$ radians.

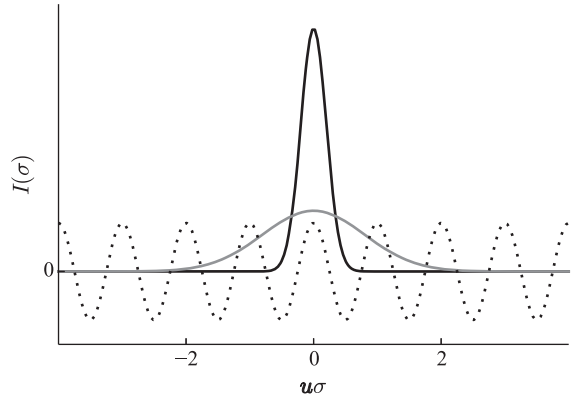

Figure 1.12 Two one-dimensional Gaussian brightness distributions plotted as a function of angular coordinate σ together with a cosine wave at frequency u (shown as a dotted line). The total flux $F(0)$ of both Gaussians is the same but the coherent flux $F(\boldsymbol{u})$ of the narrower Gaussian at frequency u is greater.

The coherent flux is the key quantity linking the fringe pattern and the object brightness distribution: by measuring the amplitude and phase of the fringes formed between two telescopes, the coherent flux can be determined, and this in turn can be related to the object structure on angular scales, which are inversely proportional to the projected baseline. This relationship, and how it can be used to reconstruct an image of the object, is explored further in Section 2.1.

1.5 Spatial coherence

Interferometry is often explained as being a means to measure the 'coherence' of a light beam. The preceeding analysis has required the use of the concept of coherence only in assuming that the light from two different stars is perfectly incoherent. A brief introduction to the idea of partial coherence is helpful for linking the concepts discussed here to the ideas of coherence, as well as explaining the origin of the term 'coherent flux'.

Given two quasi-monochromatic light beams with complex wave amplitudes Ψ_1 and Ψ_2, the *mutual intensity* of the two beams is defined as

$$M_{12} = \langle \Psi_1 \Psi_2^* \rangle , \tag{1.42}$$

where the angle brackets denote averaging over a time which is long compared to the random fluctuations of the complex wave amplitudes.

The mutual intensity is so called because it has the units of intensity. If $\Psi_1 = \Psi_2$ then $M_{12} = M_{11} = \left\langle |\Psi_1|^2 \right\rangle$, which is the intensity of a single beam. The ratio of the mutual intensity to the geometric mean of the beam intensities yields a measure of the correlation of the complex wave amplitudes. This measure is called the 'degree of coherence' (Zernike, 1938) of the beams, and is given by

$$C_{12} = \frac{M_{12}}{\sqrt{M_{11} M_{12}}}. \tag{1.43}$$

When the two beams are identical or perfectly correlated, the magnitude of the mutual degree of coherence is unity and the beams are said to be 'coherent', whereas when they are uncorrelated the degree of coherence is zero and they are said to be 'incoherent'. Intermediate values of the magnitude of the coherence correspond to 'partial coherence'. The degree of coherence and mutual intensity are complex quantities; their phases serve as measures of the mean phase difference between the two beams.

The van Cittert–Zernike theorem (van Cittert, 1934; Zernike, 1938) relates the mutual intensity to the properties of the object emitting the light. In the case of a distant object, the van Cittert–Zernike theorem gives the same result for the mutual intensity of an object with a given angular brightness distribution as is given for the coherent flux of the fringes in Equation (1.39), providing that the light emission from different parts of the object is assumed to be incoherent.

This is not unexpected. The equation for the interference pattern from a pair of quasi-monochromatic beams (Equation (1.26)) can be combined with Equations (1.15) and (1.17) to give

$$i(x) = \left\langle |\Psi_a|^2 \right\rangle + \left\langle |\Psi_b|^2 \right\rangle + 2\mathrm{Re}\left[\left\langle \Psi_a \Psi_b^* \right\rangle e^{-2\pi \mathrm{i} s x} \right]. \tag{1.44}$$

The mutual intensity of two beams appears as the complex coefficient of the interference term, i. e. the coherent flux of the fringes. An astronomical interferometer can therefore be considered as a device for measuring the coherence of light beams as a function of their transverse separation across a wavefront, otherwise known as the 'transverse coherence' or 'spatial coherence' of the light.

It should be noted that the terms 'coherent' and 'incoherent' and related terms are used in contexts other than that of the mutual degree of coherence. Coherence is more generally used to indicate phase stability in some signal, for example the 'coherence time' of the seeing is used in Section 3.1 to refer to the time over which the atmospheric perturbations can be considered to be stable, and 'coherent integration' is used in Section 8.6 to indicate integration under conditions where the phase of the fringes can be considered to be stable.

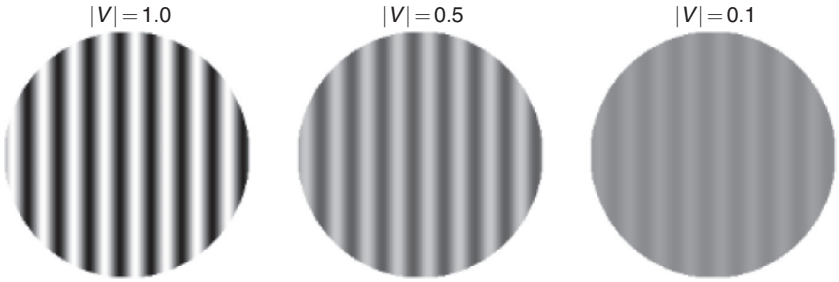

Figure 1.13 Fringe patterns with the same average brightness but different visibilities.

1.6 Nomenclature

At this juncture, a small diversion is necessary in order to establish some nomenclature, as this is somewhat confused for historical reasons. When using the human eye as a detector, the most easily observable property of interference patterns is the *contrast* of the fringes, i. e. the ratio of the intensity modulation of the pattern to the average intensity. The higher this contrast the more easy the fringes can be seen as can be appreciated from Figure 1.13. Michelson (1920) defined the *fringe visibility* as

$$V_{\text{Michelson}} = \frac{i_{\max} - i_{\min}}{i_{\max} + i_{\min}} \tag{1.45}$$

where $i_{\max}$ and $i_{\min}$ are the maximum and minimum intensities in the fringe pattern, respectively. Interestingly, the human psychovisual system (i. e. the combination of the eye and the brain's image-processing systems) is an efficient fringe-sensing tool: under optimum conditions, humans can detect the presence of fringes when $V_{\text{Michelson}} \gtrsim 0.01$.

For a fringe pattern given by Equation (1.41), the Michelson fringe visibility can be written as the modulus of a complex quantity V

$$V_{\text{Michelson}} = |V|\,, \tag{1.46}$$

where V is a normalised coherent flux

$$V = \frac{F(\boldsymbol{u})}{F(0)}, \tag{1.47}$$

known as the *complex visibility*. The complex visibility extends the Michelson visibility to include information about the phase of the fringes as well as their contrast.

In radio interferometry the normalisation factor $F(0)$ is often difficult to measure because the high sky background levels mean that determining the

total flux coming from the entire field of view is difficult. As a result, the term 'visibility' is almost always used to mean the *un-normalised* coherent flux F. It should be noted that V is always dimensionless but there is no convention for the units of F: it could be in terms of power per unit area per unit frequency, received total power or any other convenient measure.

In optical interferometry the term visibility is most frequently used to refer to the Michelson visibility $|V|$, but it is also used to refer to the complex visibility V. This book will use the term 'visibility' for V, and $|V|$ will be called the '*fringe contrast*', the '*visibility modulus*' or the '*visibility amplitude*'.

There is no universally accepted term for F. In this book, the term *coherent flux* will be used. Other terms commonly used in the literature are *correlated flux*, *mutual intensity* and *mutual coherence function*. The terms *coherent flux modulus* or *coherent flux amplitude* will be used for $|F|$.

Both F and V contain essentially the same information, but have different advantages depending on the application. The coherent flux is generally more useful mathematically as it has the important property of being a linear function of the object brightness distribution: the coherent flux for an object composed of two sub-objects is simply the sum of the coherent fluxes for the sub-objects taken individually.

In contrast(!), the visibility is a non-linear function of the object brightness distribution, but has the useful property that it allows some constants of proportionality to be discarded. For example, using Equations (1.47), (1.39) and (1.40) the visibility can be expressed in terms of the properties of the object under study as

$$V(\boldsymbol{u}) = \iint_{-\infty}^{\infty} I'(\boldsymbol{\sigma}) e^{-2\pi i \boldsymbol{\sigma}\cdot\boldsymbol{u}}\, dl\, dm, \tag{1.48}$$

where I' is a normalised brightness distribution given by

$$I'(\sigma) = \frac{I(\boldsymbol{\sigma})}{\int_{-\infty}^{\infty} I(\boldsymbol{\sigma})\, dl\, dm}. \tag{1.49}$$

Thus the visibility of objects of the same shape but different brightnesses is the same. The visibility modulus is always unity on baselines short enough that the object is unresolved.

The coherent flux and visibility can each be used to describe either an observable property of the fringes or a property of the object. For an ideal interferometer the two are identical, but for many instrumental reasons the visibility observed in the fringe pattern in a real interferometer may differ from the visibility that would be computed from Equation (1.49). The visibility in the latter role will be denoted as the '*object visibility*' and in the former role as the '*fringe visibility*'. To distinguish the two mathematically, the fringe visibility

will typically be given subscripts denoting the two telescopes being interfered, e.g. V_{12}, while the object visibility will be denoted as a function of spatial frequency $V(\boldsymbol{u})$ in analogy to the brightness distribution of the object $I(\boldsymbol{\sigma})$. The same convention will be applied to the coherent flux.

1.7 Polychromatic interferometry

Previous sections have assumed an interferometer operating at a single wavelength (or more correctly a 'quasi-monochromatic' system, where the range of wavelengths is extremely small). While this ideal can be approached by placing a narrow-band filter in front of the detector, most interferometers observe fringes using a relatively wide spectral bandpass in order to increase the number of photons collected, and this can cause the fringe pattern to deviate from the pattern for a quasi-monochromatic beam.

The fringe intensity pattern at a single frequency ν can be written as

$$i(\tau, \nu) \propto F(0, \nu) + \mathrm{Re}\left\{F(\boldsymbol{u}, \nu) e^{2\pi \mathrm{i} \nu \tau}\right\}, \tag{1.50}$$

where τ is the delay difference between the interfering beams at location x on the detector (typically $\tau \propto x$) and $F(\boldsymbol{u}, \nu)$ is the coherent flux as a function of the projected and scaled baseline $\boldsymbol{u}$ in a narrow bandpass centred at frequency ν (note that for a fixed baseline $\boldsymbol{B}$, $\boldsymbol{u}$ will be proportional to ν). The spectral coherent flux $F(\boldsymbol{u}, \nu)$ can be understood conceptually as the product of two terms:

$$F(\boldsymbol{u}, \nu) = V(\boldsymbol{u}, \nu) F(0, \nu). \tag{1.51}$$

The visibility $V(\boldsymbol{u}, \nu)$ captures the shape of the object at a given frequency, while the zero-spacing flux $F(0, \nu)$ captures the spectral distribution of the flux from the object, any wavelength-dependent absorption effects of the optics (e. g. narrow-band filters), and the variation of the sensitivity of the detector with wavelength.

To a good approximation, the fringe pattern seen when light containing a range of different wavelengths falls on a detector is the superposition of the fringe patterns seen at each of the constituent wavelengths individually – we can assume that there is no 'cross-interference' between different wavelengths, as the phase difference between light at different wavelengths changes by thousands or millions of radians during the exposure time. The pattern seen in broadband light will be therefore be given by

$$i(x) = \int_0^\infty 2F(0, \nu) + 2\mathrm{Re}\left\{F(\boldsymbol{u}, \nu) e^{2\pi \mathrm{i} \nu \tau}\right\} d\nu. \tag{1.52}$$

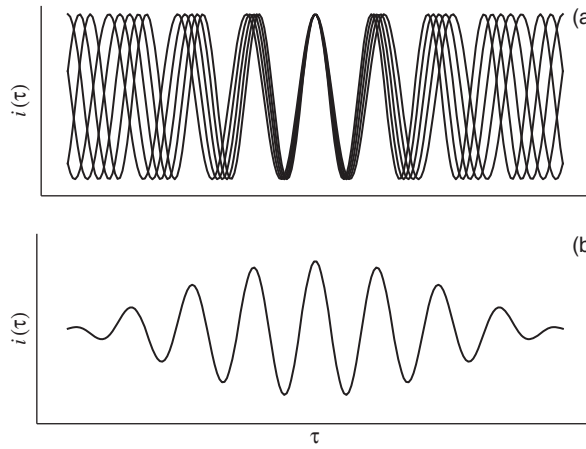

Figure 1.14 Point-source fringe patterns at multiple wavelengths (a) and their sum (b).

The effect of this superposition of fringe patterns can be appreciated by considering a number of simple cases. The first case is a single broadband point source at the phase centre so that the fringe visibility is unity at all wavelengths and baselines. The fringes at different wavelengths will all have a peak at the centre of the detector, but have different frequencies on the detector so that the fringes become more and more out of phase away from the centre. Where the fringe maximum at one wavelength overlaps a fringe minimum at another wavelength, the fringe contrast is diminished, so the fringe contrast tends to fall off away from the centre of the detector as shown in Figure 1.14.

1.7.1 Fourier relationship to the spectrum

A more quantitative insight into the shape of the fringe pattern can be obtained using a Fourier transform. The Fourier transform is discussed in more detail in Section 2.1 and Appendix A; this section can be skipped at first reading by readers less familiar with the Fourier transform.

Equation (1.52) can be rearranged to give

$$i(\tau) = 2\int_0^\infty F(0,\nu)\,d\nu + 2\int_0^\infty \mathrm{Re}\left\{F(\boldsymbol{u},\nu)e^{2\pi i\nu\tau}\right\}\,d\nu \tag{1.53}$$

$$= \int_{-\infty}^{\infty} F(0,\nu)\,d\nu + \int_{-\infty}^{\infty} F(\boldsymbol{u},\nu)\,e^{2\pi i\nu\tau}\,d\nu \tag{1.54}$$

$$= I_0 + I(\tau), \tag{1.55}$$

where $F(\boldsymbol{u}, \nu)$ has been symmetrised about $\nu = 0$ such that $F(\boldsymbol{u}, -\nu) = F^*(\boldsymbol{u}, \nu)$. The constant I_0 is an offset in intensity corresponding to the flux summed over all wavelengths,

$$I_0 = \int_{-\infty}^{\infty} F(0, \nu)\, d\nu, \tag{1.56}$$

and the fringe modulation term $I(\tau)$ is a Fourier transform of the spectral coherent flux,

$$I(\tau) = \mathcal{F}\left\{F(\boldsymbol{u}, \nu)\right\}, \tag{1.57}$$

where it should be noted that the Fourier transform is a one-dimensional transform taken over the frequency coordinate ν and not the spatial frequency $\boldsymbol{u}$.

1.7.2 Coherence length

If the zero-spacing spectrum consists of a flat-topped bandpass with band edges at $\nu_0 \pm \Delta\nu$ and the visibility is the same at all wavelengths within this bandpass, then the symmetrised coherent flux spectrum can be represented as a convolution of a 'top-hat' function and a pair of delta functions:

$$F(\boldsymbol{u}, \nu) \propto \mathrm{rect}(\nu/\Delta\nu) * [\delta(\nu - \nu_0) + \delta(\nu + \nu_0)], \tag{1.58}$$

where the rect function is defined by

$$\mathrm{rect}(t) \equiv \begin{cases} 0 & \text{if } |t| > \frac{1}{2} \\ \frac{1}{2} & \text{if } |t| = \frac{1}{2} \\ 1 & \text{if } |t| < \frac{1}{2}. \end{cases} \tag{1.59}$$

The convolution theorem can then be used to show that the modulation pattern $i(\tau)$ is the product of a sinusoidal fringe pattern with frequency ν_0 and a sinc function 'envelope':

$$I(\tau) \propto \cos(2\pi i \nu_0 \tau)\mathrm{sinc}(\pi\Delta\nu\tau), \tag{1.60}$$

where the sinc function is defined by

$$\mathrm{sinc}(x) \equiv \frac{\sin(x)}{x}. \tag{1.61}$$

The envelope function goes to zero at

$$\tau = \pm(\Delta\nu)^{-1}. \tag{1.62}$$

This scaling of the width of the 'fringe packet' as the inverse of the bandwidth is illustrated in Figure 1.15 and is consistent with the expectation that

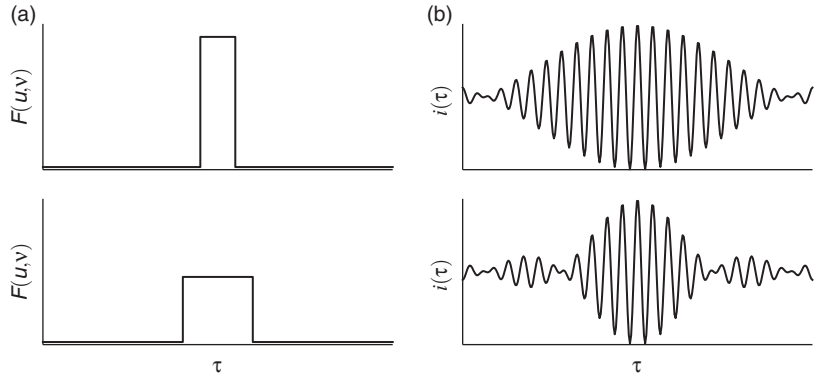

Figure 1.15 Spectral intensity patterns (a) and fringe patterns (b) for light with top-hat spectral bandpasses. The fringe envelope is a narrower, i. e. the coherence length is shorter, for the wider spectral bandpass.

the larger the range of wavenumbers, the more rapidly the fringes at different wavelengths go out of phase with increasing delay τ and so the more rapidly the contrast of the pattern made from the superposition of the fringes drops off.

The number of fringes inside the primary lobe of the fringe envelope is equal to the inverse of the fractional bandwidth, whether expressed in frequency or wavelength:

$$N_{\text{fringe}} = 2\left|\frac{\nu}{\Delta\nu}\right| \approx 2\left|\frac{\lambda}{\Delta\lambda}\right|. \tag{1.63}$$

The range of OPD $c\tau$ over which high-contrast fringes can be seen is often called the 'coherence length' of the light (this form of coherence is sometimes called 'longitudinal coherence' to distinuish it from 'lateral coherence', i. e. spatial coherence). The delay lines must equalise the pathlengths in the interferometer to an accuracy which is better than the coherence length in order to see high-contrast fringes.

If fringes are observed through a filter with a fractional bandwidth of 1% (for example a 5-nm bandpass at a wavelength of 500 nm), there will be 200 fringes inside the envelope and at a wavelength of 500 nm, this corresponds to a coherence length of order 0.1 mm. If a bandwidth of order an octave is used, for example the whole of the visible region from 400 nm to 700 nm, then the coherence envelope is of the order of one fringe in size; the position of zero OPD is then uniquely identifiable as the location of the high-contrast 'white-light fringe'.

1.7.3 Bandwidth smearing and field of view

If the object consists of a combination of a point source at the phase centre and one at an angular offset of $\boldsymbol{\sigma}$, then the observed fringe system will consist of the superposition of two sets of fringes with envelopes ('fringe packets') which are offset from one another in delay space. If the offset is large enough that the two packets do not overlap, then there will be no position in the fringe pattern where the fringes from both point sources are simultaneously present. If this occurs, then the simple relationship between the fringe visibility and the object visibility will break down – the effects of these errors in the visibility on the reconstructed image are known as 'bandwidth smearing' in radio interferometry.

This will occur when the differential delay between the stars $\boldsymbol{\sigma}\cdot\boldsymbol{u}\lambda/c$ exceeds the size of the fringe envelope as given in Equation (1.62), in other words,

$$\boldsymbol{\sigma}\cdot\boldsymbol{u} \gtrsim \frac{\nu_0}{\Delta\nu}, \tag{1.64}$$

where ν_0 is the centre frequency and $\Delta\nu$ is the bandwidth of the radiation being interfered. The two point sources can then be said to be no longer within the same 'bandwidth-smearing field of view'.

The size of this field of view along a direction parallel to the baseline is given by

$$\Delta\theta_{\mathrm{FOV}} \sim \frac{\nu_0}{\Delta\nu}\frac{1}{|\boldsymbol{u}|} \tag{1.65}$$

and since the angular resolution is given approximately by $|\boldsymbol{u}|^{-1}$ then the field of view can be written as

$$\Delta\theta_{\mathrm{FOV}} \sim \Delta\theta_{\mathrm{res}}\frac{\nu_0}{\Delta\nu}, \tag{1.66}$$

where $\Delta\theta_{\mathrm{res}}$ is the angular size of a resolution element ('resel'). Thus if fringes are observed on multiple baselines using a fractional optical bandwidth of 1% one can in principle make an image up to about 100×100 resels in size before bandwidth-smearing effects become overwhelming.

An alternative approach to making use of a wide spectral bandwidth is to observe fringes in narrow spectral channels but to observe many such channels simultaneously – so-called 'spectro-interferometry'. This approach is more complex but allows the fringe envelope to extend over wider ranges of OPD and therefore larger fields of view. More is said about spectro-interferometry in Section 4.7.4.

1.8 Chromatic dispersion and group delay

In the above analysis, the delays experienced by the light signal have all been assumed to be independent of wavelength. In a real interferometer, the light will pass through a number of optical elements such as vacuum windows and lenses whose refractive index depends on wavelength (they are said to show *chromatic dispersion*). If the differential delay between the light paths travelled by the two interefering beams depends on wavelength then the phase of the fringes will be wavelength-dependent. If the phase shifts by a large amount within the spectral bandpass used to observe the fringes, this could cause the summed fringe across the bandpass to 'wash out'.

The effects of dispersive optical elements within the interferometer can be cancelled by balancing the dispersion in both arms of the interferometer, typically by using identical optical elements in both arms, or by inserting 'compensating plates' of glass. Nevertheless, there will be some residual differential dispersion due to manufacturing tolerances. In addition, in interferometers which use delay lines in air rather than in vacuum, there will be a component of the dispersion which is due to the air in the delay line. This component is time-variable and therefore harder to compensate for, so its effects must be considered.

1.8.1 Atmospheric dispersion

The fact that air in the delay lines causes an unbalanced dispersion can be somewhat counterintuitive, as the aim of the delay lines is to provide a balancing delay to any net external delays, which in a ground-based interferometer consist in part of optical paths in air. Figure 1.16 shows that, providing that the interferometer collecting elements lie in a horizontal plane and assuming a plane-parallel atmosphere, all the optical paths in air external to the interferometer are matched between telescopes. In contrast, the geometric delay due to the phase centre not being directly overhead is a pure vacuum delay.

The beams therefore experience no differential air path due to geometric path effects if the delay line is in vacuum. If the delay line is in air, however, then tens or hundreds of metres of vacuum delay are compensated by a similar amount of air delay, and so the refractive index variations of the air become important and need to be considered. The amount of air path depends on the location of the phase centre, so the effects of dispersion are variable with time.

Light travelling through a distance l in a medium such as the air with refractive index n experiences a delay given by

$$\tau = nl/c. \tag{1.67}$$

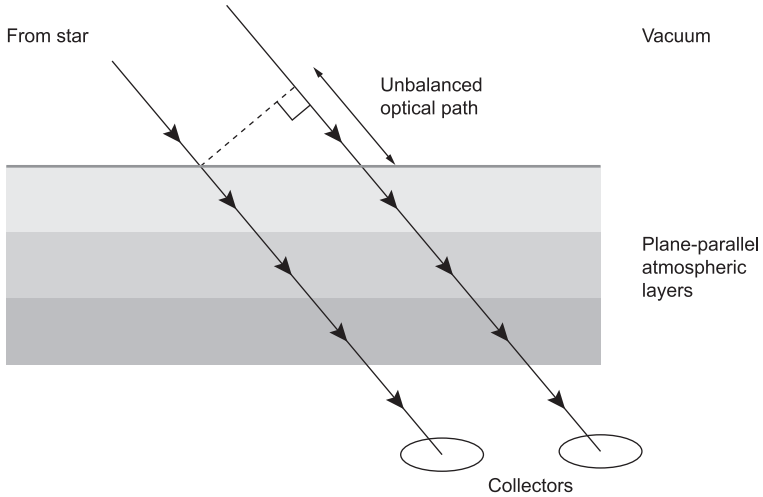

Figure 1.16 Geometry of the external light paths in a plane-parallel atmosphere and a horizontal interferometer, showing that all the optical paths in air are balanced.

If this delay is compensated by an equal path in air in the opposite arm of the interferometer, then the OPD will be given by

$$c\tau_{12} = (n - 1)l. \tag{1.68}$$

Figure 1.17 shows the refractivity $\mu = n - 1$ of dry air as a function of wavelength. It can be seen that μ changes more rapidly with wavelength at blue wavelengths compared with red and near-infrared wavelengths, and this trend continues towards mid-infrared wavelengths.

Water vapour has a refractivity curve which follows the same trend but with a different slope, so the refractivity curve of air depends on its humidity. The refractive index profiles of water vapour and of carbon dioxide have sharp features near their absorption lines at infrared wavelengths, so the detailed shape of the refractivity curve is more complex than is apparent in the figure. This can be important if narrow-band interferometry is being undertaken (Colavita *et al.*, 2004) but is ignored for the rest of this analysis.

The OPD due to differential dispersion can be compensated at a single wavelength by moving the delay line appropriately, but the differential effects with wavelength can still be quite large: the refractivity of air is measured in parts per million, but for a 100-m differential air path, the OPDs experienced at 600 nm and 800 nm will differ by 200 μm, and so the fringe shifts within a bandpass as small as a nanometre will be significant.

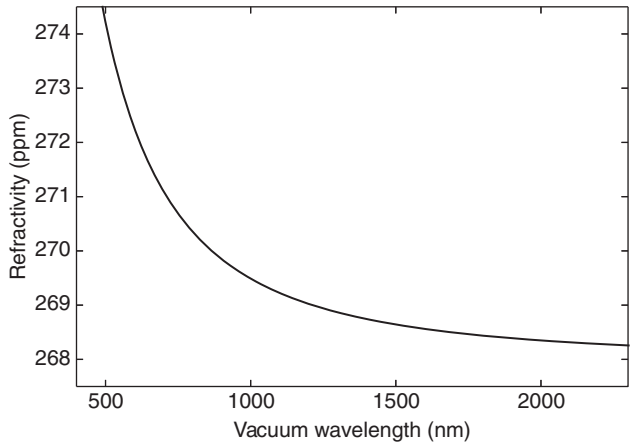

Figure 1.17 The refractivity $\mu = n - 1$ of dry air at 1 atmosphere and 20 °C at visible and near-infrared wavelengths.

This means that if fringes are observed through a filter with a moderate bandpass, they can be washed out by intra-bandpass phase shifts even if the delay lines are moved to cancel the OPD effects of refraction at a single wavelength. It turns out, however, that the fringe contrast will recover if the delay line is offset not by the refractive delay at a given wavelength (the so-called *phase delay*) but by an amount known as the *group delay*.

1.8.2 Group delay

To understand the origin of the group delay, it is helpful to consider a thought experiment in which a symmetrical interferometer consisting of a pair of collectors and a pair of variable delay lines observes a point source directly overhead. If all the optical paths are in vacuum and the delay lines are adjusted so that the net OPD is zero, the phase of the fringes at all wavelengths will be zero. Thus if the fringes are observed using a finite (but small) bandpass the resulting fringe pattern, which is the sum of the fringe patterns at each wavelength within the bandpass, will have both zero phase and high contrast as the fringe maxima will all coincide.

If one delay line is now moved to introduce a (wavelength-independent) delay τ_0 this will cause the fringes to shift in phase. The phase shift as a function of wavelength will be given by

$$\phi(\nu) = 2\pi\tau_0\nu. \tag{1.69}$$

If all the fringe patterns in a finite bandpass are now added together there will be two effects. First, the mean fringe phase will be offset by approximately $2\pi\nu_0\tau_0$, where ν_0 is the central frequency of the bandpass. A second effect is that the fringe contrast will be reduced because of the change of phase within the bandpass – in the worst cases the fringe peak at one wavelength will coincide with the trough at another wavelength within the bandpass.

This can be understood as the effect of the finite-sized fringe envelope due to the optical bandwidth as outlined in Section 1.7: moving the delay line has moved the zero-OPD point and hence the centre of the fringe envelope with respect to the centre of the fringe detector. These two effects, the phase shift of the fringes and contrast reduction, can be seen as the result of the effects of the zero-order and first-order change of phase shift with frequency respectively.

If one delay line is set to introduce a delay of τ_0 and air is then introduced into that delay line then the optical delay introduced by that delay line will be $n(\nu)\tau_0$, where $n(\nu)$ is the refractive index of air at frequency ν. The remaining (i. e. vacuum) delay line can be moved to introduce a delay of $-n(\nu_0)\tau_0$ so that the fringe phase at frequency ν_0 will be zero. The fringes at frequency ν will have a phase given by

$$\phi(\nu) = 2\pi\tau_0\nu\left[n(\nu) - n(\nu_0)\right]. \tag{1.70}$$

The phase variation with ν can be Taylor-expanded about the value at ν_0 to give

$$\phi(\nu_0 + \Delta\nu) \approx 2\pi\tau_0\left[\left.\frac{d(n\nu)}{d\nu}\right|_{\nu_0} - n(\nu_0)\right]\Delta\nu + O(\Delta\nu^2) \tag{1.71}$$

$$= 2\pi\tau_0\nu_0\left.\frac{dn}{d\nu}\right|_{\nu_0}\Delta\nu + O(\Delta\nu^2), \tag{1.72}$$

where $\nu = \nu_0 + \Delta\nu$.

For a sufficiently narrow bandpass the quadratic and higher-order terms in $\Delta\nu$ can be neglected, leaving a linear change of phase with frequency across the bandpass. The linear change of phase with frequency causes a reduction in fringe contrast in exactly the same way as if there had been an OPD error, but the fringe contrast can be restored by adjusting the vacuum delay to introduce a compensating linear phase shift. The required additional delay is given by

$$\tau_{\text{offset}} = \tau_0\nu_0\left.\frac{dn}{d\nu}\right|_{\nu_0}. \tag{1.73}$$

This means that while the compensating vacuum delay needed to make the phase of the fringes zero (known as the *phase delay*) at frequency ν_0 is

$$\tau_{\text{phase}} = \tau_0 n(\nu_0), \tag{1.74}$$

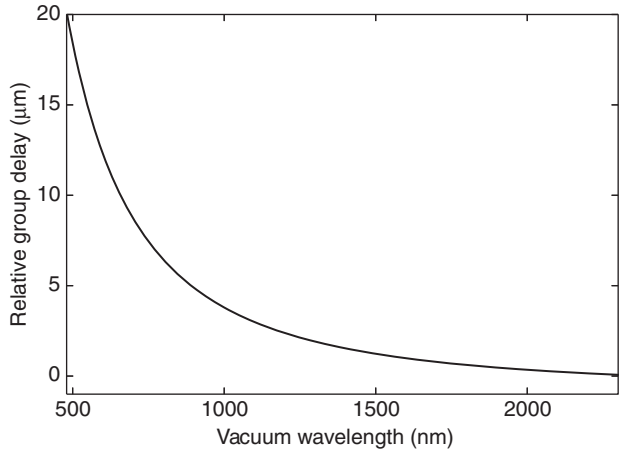

Figure 1.18 The group delay as a function of wavelength for 1 m of dry air at 1 atmosphere and 20 °C. The group delay at 2.5 μm has been subtracted to give a relative group delay.

the compensating vacuum delay for maximum fringe contrast for a bandpass centred at ν_0 (the group delay) is given by

$$\tau_{\text{group}} = \tau_{\text{phase}} + \tau_{\text{offset}} \tag{1.75}$$

$$= \tau_0 \left(n(\nu_0) + \nu_0 \left. \frac{dn}{d\nu} \right|_{\nu_0} \right). \tag{1.76}$$

Figure 1.18 shows the relative group delay introduced per metre of differential air path at different wavelengths. It can be seen that, like the refractive index, the group delay increases sharply towards the blue.

The peaks of fringe packets with different central wavelengths can be separated by significant amounts even for modest amounts of air path difference. For example, with 10 m of differential air path, the fringe envelope centres at wavelengths of 600 nm and 800 nm will be 56 μm apart in OPD. If the width of bandpasses used for forming the fringes at each wavelength are 10% of the respective central wavelength then the fringe envelopes will be 6 μm and 8 μm wide, respectively, and so fringes cannot be seen simultaneously at both wavelengths unless additional differential delay is introduced for each bandpass.

These and other atmospheric dispersion effects can be compensated for either by using narrower bandpasses or by inserting appropriate amounts of glass into the beams in the interferometer so that the dispersion of the glass cancels the dispersion of the air (Tango, 1990; ten Brummelaar, 1995; Lévêque

et al., 1996; Davis *et al.*, 1998; Thureau, 2001). A difficulty with the glass compensation approach is that the dispersion of the air is sensitive to temperature and humidity variations, so accurate knowledge of the environmental conditions along the length of the delay line is needed in order to compensate for the dispersion.