Design of an Interferometric Spectrometer for Environmental Surveillance

Kjetil Dohlen

Design of an Interferometric Spectrometer for Environmental Surveillance

Submitted by Kjetil Døhlen.

in fulfilment of the requirements for the degree of

Doctor of Philosophy

of the University of London

Submitted: October 1993
Viva: 21 January 1994

Imperial College of Science, Technology and Medicine

London SW7, Great Britain

Abstract

This thesis describes the design of a field-portable heterodyned holographic spectrometer for spectral measurements in the visible and very near infrared from natural targets such as vegetation and the atmosphere. The instrument is a non-scanning version of the Fourier transform spectrometer (FTS) where the moving mirror is replaced by either a fixed, tilted mirror (holographic FTS) or a fixed, Littrow-mounted grating (heterodyned holographic FTS). This allows for a simpler and more rugged mechanical design than that of classical FTS instruments as well as a greatly reduced power consumption, all important advantages for a portable instrument for field use.

Spectral information is produced by interfering two mutually inclined wavefronts in a Michelson interferometer. From the resulting intensity patter or *interferogram*, the spectrum of the interfering beams may be found by Fourier transformation. The interferogram is spatially localized inside the interferometer and to be measured it must be imaged onto a light sensitive surface. A purpose-built, all-reflective lens is employed to image the pattern onto a photosensitive diode array. Built-in electronic circuitry controls the detection and digitizes the measured signal, and a portable PC performs the signal processing and serves as human interface.

The thesis presents a study of the requirements for a field portable instrument and gives a review of available spectroscopic techniques. On the background of this study the choice of the holographic FTS method is justified by its combination of high optical throughput and simple and rugged design. After a review of the literature concerning this spectroscopic method and a theoretical account of its operation, the instrument's design is described. Tests of individual components as well as the entire instrument are presented. Finally, some results from field applications of the instrument are reported, accompanied by an assessment of its practical value.

To Béa, my companion

Acknowledgements

My time as a Ph.D. student has been dominated more by remarkable friendships than by remarkable advances in physics. It is never easy to start a new life in a new city, but for me the process was greatly eased thanks to the student on the desk next to mine, Juan. He knew everybody, and so, very soon, did I—at least all the Mexicans and Venezuelans around. Juan is no longer with us but his memory lives on eternally through the friendships that were formed around him. Some of us have even found our life's companion among his friends. Thank-you Juan and thank-you all my other friends who have made the London years so enjoyable.

When some physics has been done during this time it is doubtlessly due to my supervisor, Tony Cañas. He has always managed to keep the iron hot and to pull me back on the track. Thanks are also due to Lady Anne Thorne and Richard Learner for their expertise in the field of Fourier spectroscopy and their willingness to part it with me. Fred Reavell deserves a special mention for his patience during our long discussions of mechanical design details, and so do the optical and mechanical workshops for their efforts in transforming abstractions into reality.

Thanks, finally, to the Applied Optics Group for giving me the chance and to the following bodies who have given financial support during three years: SERC (Great Britain) for funding the project, the Committee of Vice-Chancellors and Principals of the United Kingdom for partially covering my student fees under the ORS scheme, and NAVF (Norway) for paying my subsistence via an educational scholarship. My fourth year has in its entirety been funded by my wife.

Marseille, 16th October, 1993

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Warned by a couple of well-publicised phenomena, the general public is getting increasingly aware of the environmental effects of excessive pollution. One phenomenon relates a 'hole' in the stratospheric layer of ozone discovered over Antarctica a few years ago to an excessive release of certain gases used in spray bottles and refrigerators. At the same time, the increasing atmospheric density of these and other pollution gases are said to affect the radiation transmission characteristics of the atmosphere, causing a general increase in world temperature—the infamous 'green house effect'. Several other phenomena are, although less vividly described in the press, also probably related to pollution, such as fish and forest deaths and over-development of certain species of sea algae.

Much research is currently being expended on understanding scientifically the various effects of pollution. This activity leads to a demand for extensive *environmental surveillance* [16] whereby pollutants may be measured and their effects quantified. Environmental surveillance may be based on direct, *in situ* analysis of objects and their processes by visual observation, chemical analysis, or other methods, but for such large scale observations as are often required in environmental research, this becomes impractical. Instead, the task is increasingly being performed by the aid of *electromagnetic remote sensing* where objects and processes are characterized by their interactions with electromagnetic radiation. This can be done from a distance, often of thousands of kilometres from a satellite in earth orbit.

Optimal collection and analysis of remotely sensed data requires a knowl-

edge of the way natural processes interact with radiation. One important tool for obtaining such knowledge is *field spectroscopy* [25], where the 'remote' sensing is performed *in situ* in parallel with direct analytical measurements. This allows building models relating directly measured, physical attributes to remotely sensed attributes. It also allows prediction of optimal measurement conditions (specifying spectral requirements, illumination and looking angles, season of the year, etc.) as well as providing calibrations by reference to well-specified ground targets.

1.1 Remote Sensing

Historically, remote sensing started in 1858 when the first aerial photograph was taken from a balloon [23]. Since then, impressive technological developments have given improvements in the design of both platforms and sensors. While the former is dominated by the advent of aeroplanes and earth orbiting satellites, the latter is characterized by a formidable increase in the amount of information collected by modern instrumentation. Instead of measurements in one single spectral band as a black and white photograph is limited to, sensors are now available with several, sometimes hundreds of bands ranging from optical radiation including the ultra violet (UV) and infrared (IR), to radar. Although radar imaging is reaching maturity, particularly by the resent launch of the ERS 1 satellite, optical remote sensing still has an important position because of the high spectral resolution attainable and the existence of many characteristic spectral features within its bounds. The comfortable existence of the sun as a powerful source of optical radiation is another of its *raisons d'être*.

Modern remote sensing started in 1972 with the launch of the Landsat 1 satellite [23]. Providing consistent and high quality images of the entire earth in four spectral bands, it had a great impact on the remote sensing community. The current 'state of the art' in space based optical remote sensing is offered by the last member of the Landsat family carrying the 'Thematic Mapper' sensor. It offers imaging in seven spectral channels ranging from the visible at 0.45 μm out to 12.5 μm in the thermal infrared. A major drawback of Landsat images

is their coarse spectral resolution, and to improve on this and other deficiencies a new series of satellites are under development, scheduled for launch in 1995 [26]. Named picturesquely *Mission to Planet Earth* it will constitute an 'Earth Observation System' (EOS). Sensors of various spectral and spatial resolutions are previewed, notably one with a continuous coverage of 10 nm wide spectral samples from 0.4 μm to 2.5 μm and a spatial picture element (pixel) size of 30 m. Clearly, this represents a quantum leap in sophistication which calls for a similar improvement in instrumentation for field spectroscopy.

It is the increased requirements for field spectroscopy that we address in this thesis. A study has been made of possibilities and limitations offered by novel spectroscopic techniques, with particular attention to miniaturization and throughput optimization. The most favourable method has been chosen and a working prototype instrument has been built. Acting as a stand-alone, local environment sensor it may provide non-destructive and unobtrusive analysis of a variety of targetsincluding agricultural crops and atmospheric pollutants, as well as immersed objects like sub-surface algae. It may also find applications in industrial process control and quality assurance [28].

1.2 Guide to the Thesis

Chapter 2 presents a radiation budget where available radiation may be sees as 'income', losses in the optical path as 'expenses', and the instrument throughput as an 'asset'. A comparison between the performance of dispersive spectrometers (using a prism or a grating) and interferometric spectrometers (using a Michelson or a Fabry-Perot interferometer) based on the radiation budget is then made which culminates in the preference for interferometric spectroscopy due to its throughput advantage. This preference is answered by a novel implementation of the Michelson interferometer in a non-scanning or *holographic* Fourier transform spectrometer (HFTS) where the interferogram is presented as a spatial rather than temporal intensity pattern. Sampled by an array detector, the interferogram is Fourier transformed to yield the spectral information. Its merits include good mechanical stability, small size, and low power consumption. Thanks to a heterodyning capability, a flexible imple-

mentation is possible, allowing freely adjustable resolving power over a large range within the entire visible and immediate near infrared spectrum (0.4 μm to 1.0 μm).

Chapter 3 presents the theoretic basis for the instrument giving an explanation of how the ideal instrument yield spectral estimates. Spectral degradation due to manufacturing inaccuracies and measurement noise is then discussed, notably treating the effects of interferogram phase and apparent errors in the sampling grid.

In **Chapter 4**, the design of the instrument is described. Particular attention has been given to critical components such as the interferometer and its centre piece, the beam splitter. Another important part of the design is the fringe imaging lens which transfers the interferogram from within the interferometer onto the detector array. Signal processing is also given a comprehensive coverage; this is the means by which an optimal spectral estimate is obtained from the measured data. Mechanical and electronic design is also described although only in the form of a summary.

Chapter 5 comments upon operating practices and presents demonstrations of the instrument including a measurement of th red edge in vegetation spectra and absorption due to atmospheric NO_2, an important air pollutant. In the red-edge measurement we observe the characteristic double inflection, and in observing scattered sunlight at a high resolution we detect the blue NO_2 absorption.

Finally, **Chapter 6** gives overall conclusions and recommendations for future work.

1.3 Units

SI units are used throughout apart from in a few cases where other units are found more illustrative. In particular, we measure distance along the interferogram in units of "photodiode elements", denoted Elements, and spatial frequencies in the interferogram in units of "cycles per photodiode element", denoted Elements^{-1}.

Optical throughput has SI unit "square metre steradian" (m^2 sr). We find

this somewhat opaque and prefer instead the unit "square centimetre-degree" (cm^2 deg^2) which is more easily envisaged.

In discussing absorption by atmospheric constituents, the product of absorption path length and the constituent's partial pressure is an important parameter. Its SI unit is "pascal meters" (Pa m), but we have used instead "atmosphere centimeters" (Atm cm). The advantage of using the atmosphere unit is that if the absolute pressure of the gas mixture is one atmosphere, then the partial pressure of a constituent measured in atmospheres equals the concentration of that gas in the mixture. A partial pressure of 10^{-3} Atm is thus equal to a concentration of 1 ppt (part per thousand). Similarly 10^{-6} Atm corresponds to 1 ppm (million), and 10^{-9} Atm corresponds to 1 ppb (billion).

Finally, we have chosen to use *wavelength* measured in microns (μm) or nanometres (nm) rather than *wavenumber* measured in cm^{-1} as the predominant spectral unit. Rooted in our own lack of spectroscopic background, we hope that this will not annoy the spectroscopist and that it will instead enhance the readability of the Thesis for the non-spectroscopist. Having said that, we must of course admit that since the instrument is interferometric and therefore has a spectral axis linear in wavenumbers rather than wavelengths, the conversion is not always practical. In those cases we will (again begging forgiveness from the spectroscopist) use the unit *reciprocal micron* (μm^{-1}) which, in the present context, appears more intuitive than the cm^{-1} unit.

Chapter 2

Concept Study

This chapter presents a study of the requirements for remote sensing spectroscopy and how these are met with different types of spectroscopic instruments. On the basis of the study a prototype instrument has been built, the design of which is the subject of the remainder of the Thesis.

A large part of the chapter is given to the construction of a *Radiation Budget* where output signal is expressed in terms of available energy, instrument throughput, and transmission losses. In analogy with an economic budget, input energy may be seen as 'income', instrument throughput is an 'asset', and transmission losses including detector efficiency as 'expenses'. Output energy in the form of an electric current thus represents the 'balance' of the budget.

One important limitation for the construction of an instrument has been given based on practical arguments: the instrument will employ silicon photodiode detectors. This is because they are easily available in many different forms and their sensitivity covers a spectral region (the visible and very near infrared, 0.4–1.0 μm) where optical glasses also have good transmission and dispersion characteristics, and where the sun provides a good source of radiation. These advantages simplify the design of a working prototype to the point where its construction is conceivable within the span of a Ph.D. project. The spectral region of silicon detectors is of importance in remote sensing, particularly for studies of vegetation where it covers the chlorophyll absorptions, and atmospheric pollution, represented in an example by the NO_2 absorption at 490 nm.

2.1 Elements of Radiometry

We will keep the radiometric discussion at a fairly basic level by avoiding the full vocabulary of terms and by making some simplifying assumptions with respect to radiation transfer and bidirectional reflectance characteristics. Hence the treatment is limited to the notions of *radiant power*, *irradiance*, and *radiance*. Radiant power equals the power contained within the radiation hitting a surface, irradiance is the amount of radiant power hitting one unit area of the surface, and radiance is the radiant power leaving the surface per unit *projected* area into a unit solid angle. Figure 2.1 illustrates these concepts,

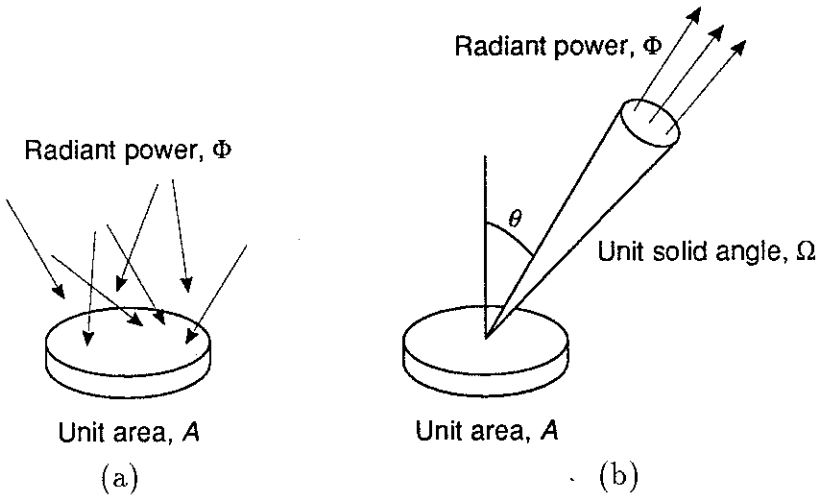


FIGURE 2.1: Illustration of the concepts of (a) irradiance, $E \equiv \dfrac{\partial \Phi}{\partial A}$ and (b) radiance, $L \equiv \dfrac{\partial^2 \Phi}{\partial \Omega \, \partial A \cos\theta}$.

and definitions of the terms and their symbols are given in Table 2.1. All the terms may be *spectral*, i.e. measured either per unit wavelength or per unit wavenumber interval, denoted by subscript λ and σ, respectively.

2.1.1 Radiation transfer

Radiation transfer calculations specify how much of the power radiated by an object (the source) is received by another object (the target). To illustrate this and present the first of our assumptions we consider the source–target geometry of Figure 2.2. By assuming both source and target to be small compared with their separation, we remove the differentials in the expressions for irradiance and radiance. The radiation transfer calculation may then proceed simply by

TABLE 2.1: Radiometric quantities, their symbols, definitions, and units[a].

Quantity	Symbol and definition	Unit
Radiant power	Φ	W
Irradiance	$E \equiv \frac{\partial \Phi}{\partial A}$	W m^{-2}
Radiance	$L \equiv \frac{\partial^2 \Phi}{\partial \Omega \, \partial A \cos\theta}$	W m^{-2} sr^{-1}
Spectral radiant power	$\Phi_\lambda \equiv \frac{\partial \Phi}{\partial \lambda}$	W μm^{-1}
or	$\Phi_\sigma \equiv \frac{\partial \Phi}{\partial \sigma}$	W μm
Spectral irradiance	$E_\lambda \equiv \frac{\partial E}{\partial \lambda}$	W m^{-2} μm^{-1}
or	$E_\sigma \equiv \frac{\partial E}{\partial \sigma}$	W m^{-2} μm
Spectral radiance	$L_\lambda \equiv \frac{\partial L}{\partial \lambda}$	W m^{-2} sr^{-1} μm^{-1}
or	$L_\sigma \equiv \frac{\partial L}{\partial \sigma}$	W m^{-2} sr^{-1} μm

[a]Adapted from references [19] and [20].

measuring the radiance (L) of the source and multiplying it by the (projected) source area ($A_S \cos\theta_S$) and the solid angle subtended by the target (Ω_T):

$$\Phi = L\Omega_T A_S \cos\theta_S. \tag{2.1}$$

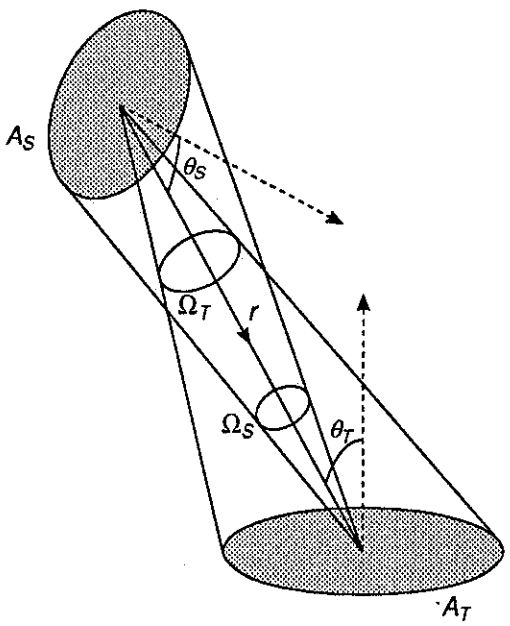


FIGURE 2.2: Geometry of radiation transfer calculations.

Seeing that $\Omega_T = A_T \cos\theta_T / r^2$ where $A_T \cos\theta_T$ is the projected area of the target and r is the distance between the objects, this may be rewritten as:

$$\begin{aligned} \Phi &= L \frac{A_T \cos\theta_T}{r^2} A_S \cos\theta_S \\ &= L\Omega_S A_T \cos\theta_T, \end{aligned} \tag{2.2}$$

where Ω_S is the solid angle subtended by the source at the target.

2.1.2 Directional reflectance

The second assumption concerns the directional distribution of the source radiance. In general a radiating object has one or more preferred directions of radiation. For objects which are not self luminous but merely reflect incident light, preferred directions are usually found due to specular reflection and/or retro-reflection (enhanced back-scatter). The radiance of a target must therefore be considered as a function of the directional coordinates θ and ϕ, see Figure 2.3, both of the observer and of the source. This gives rise to

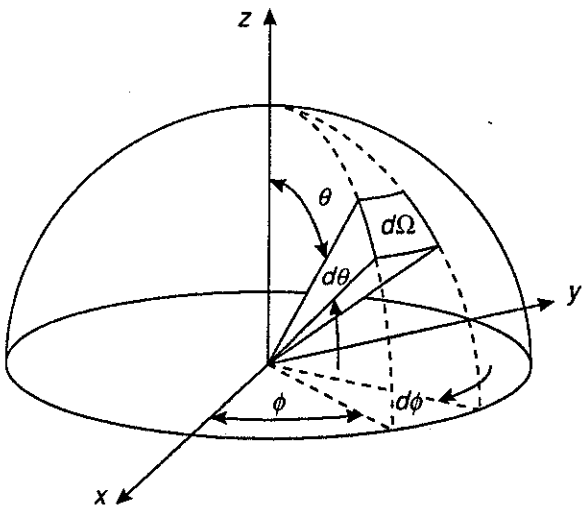


FIGURE 2.3: The spherical coordinate system used for characterizing directional radiance and irradiance. The elemental solid angle is also shown.

the *bidirectional reflectance-distribution factor*, $f_r(\theta_i, \phi_i; \theta_o, \phi_o)$ [20], obviously a difficult function to measure but of critical importance for the interpretation of remotely sensed data [25]. For the purpose of the present study we will ignore this function however, and assume all targets to act as *Lambertian* radiators, radiating equally in all directions.

Rewriting the definition of radiance from Table 2.1 in an integral form it is possible to calculate the relationship between the total power radiated from a Lambertian radiator and its radiance. Seeing from Figure 2.3 that the elemental solid angle is $d\Omega = \sin\theta\, d\theta\, d\phi$ the integral becomes:

$$\begin{aligned}
\Phi &= \int_A dA \int_{\text{Hemisphere}} L\cos\theta\; d\Omega \\
&= LA \int_{\phi=0}^{2\pi} d\phi \int_{\theta=0}^{\pi/2} \cos\theta \sin\theta\, d\theta \\
&= 2\pi LA \int_0^1 u\, du \\
&= \pi LA, \qquad (2.3)
\end{aligned}$$

with the change of variable $u = \sin\theta$. Hence:

$$L = \frac{\Phi}{\pi A}. \qquad (2.4)$$

If the target acts as a Lambertian radiator, its radiance due to an irradiance E is therefore:

$$L = \frac{RE}{\pi} \qquad (2.5)$$

where R is the dimensionless *reflectance factor* of the target. Illuminated by a source of radiance L_S subtending a solid angle Ω_S, the target irradiance is $E = L_S\Omega_S$, hence:

$$L = RL_S\frac{\Omega_S}{\pi}. \qquad (2.6)$$

As a third assumptions we will take the spectral power (Φ_λ) to be constant across a spectral channel. We thus avoid a spectral integral and are allowed to say:

$$\Phi = \Phi_\lambda\, \Delta\lambda, \qquad (2.7)$$

where $\Delta\lambda$ is the width of a spectral channel.

2.1.3 Conservation of throughput

An important feature of optical imaging systems is their conservation of radiance. To demonstrate this, consider the imaging properties of a perfect lens as illustrated in Figure 2.4. An object of projected area A is imaged onto an area A'. The lens subtends solid angle Ω seen from the object and Ω' seen

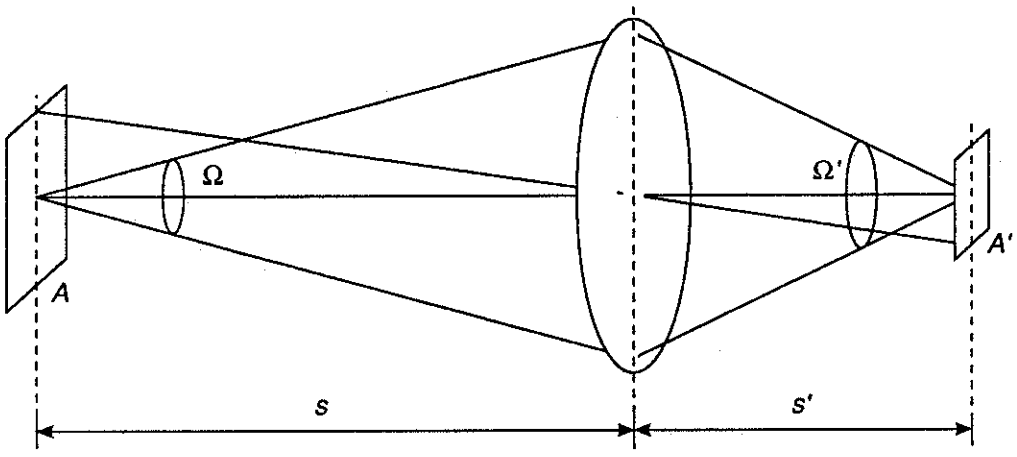


FIGURE 2.4: Construction of the imaging properties of a perfect lens.

from the image. Geometric construction then shows that $A/s^2 = A'/s'^2$ and $\Omega s^2 = \Omega' s'^2$ where s and s' are object and image distance, respectively. Hence

$$A\Omega = A'\Omega' = G, \tag{2.8}$$

an invariant of the system known as *optical throughput* and recognized as the two-dimensional equivalent of the Smith-Helmoltz invariant [6, page 165]. Invariance of radiance follows from this by conservation of energy. Let the object have radiance L and area A so that the power radiated from it into solid angle Ω is $\Phi = LA\Omega = LG$. If no losses are suffered then the power radiated into the image must also be Φ, hence the image radiance is $\Phi/(A'\Omega') = \Phi/G = L$, as required.

2.2 Radiation Budget

The Radiation Budget relates the output from a radiation collecting instrument to the level of radiation that it collects. Three basic components are involved in its calculation: the source (including losses in the radiation transfer between it and the target), the target, and the receptor. We consider these in two transfer processes: source–target and target–receptor.

2.2.1 Target to sensor transfer

Detector output current is proportional to the absorbed power with a proportionality factor called *detector responsivity*, ϱ. Ideally one electron of charge is

produced per photon incident on the detector, but practical devices only convert a fraction of the incident power. Denoting this fraction by the *quantum efficiency*, η, detector responsivity may be expressed as:

$$\varrho = \frac{e}{hc/\lambda}\,\eta = 807\,\lambda\eta \text{ [mA/W]} \tag{2.9}$$

where e is the electron charge, h is Plank's constant, c is the speed of light, and λ is measured in microns. Hence output current of a detector irradiated by spectral power Φ_λ collected within a narrow bandwidth $\Delta\lambda$ is:

$$i = \Phi_\lambda \varrho\, \Delta\lambda. \tag{2.10}$$

Figure 2.5 shows the responsivity for the detector we have used in our instrument.

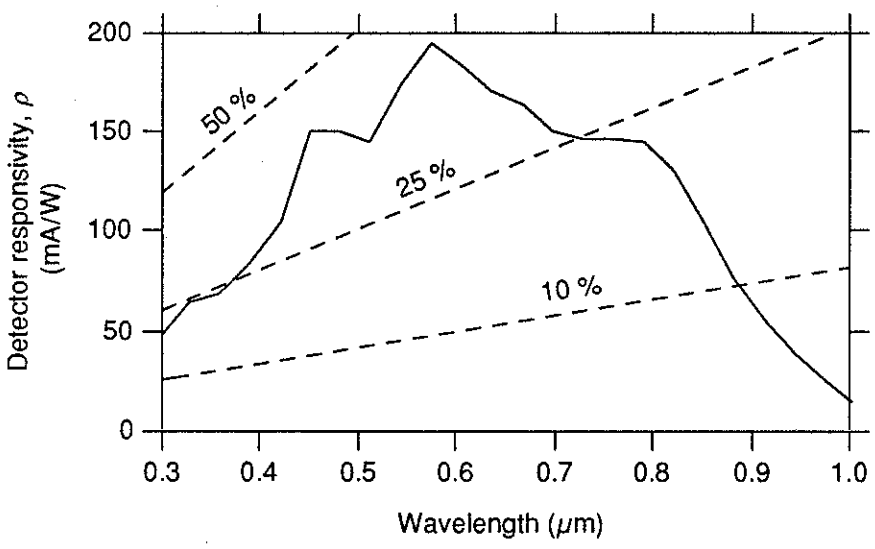


FIGURE 2.5: Responsivity of the silicon detector used in our instrument as measured by the manufacturer. Broken lines show responsivities of ideal photo electric detectors with given quantum efficiencies.

By the arguments of the preceding section, the spectral radiant power incident on the detector equals the product of target spectral radiance L_λ and throughput G of the optical system. In any real system there is a loss of power due to absorption and scattering in the optical components accounted for by the transmission factor T_O. The detector current in terms of target radiance is thus:

$$i = L_\lambda G T_O \varrho\, \Delta\lambda. \tag{2.11}$$

Before going on to consider the radiation transfer between source and target, we will have a look at a commonly used source in optical remote sensing: the sun.

2.2.2 The sun

Measured at the top of the earths atmosphere the sun has a spectrum that resembles that of a *blackbody* at an absolute temperature of 5800 K (see Figure 2.6). A blackbody is an ideal absorber [4, page 323], and its spectral

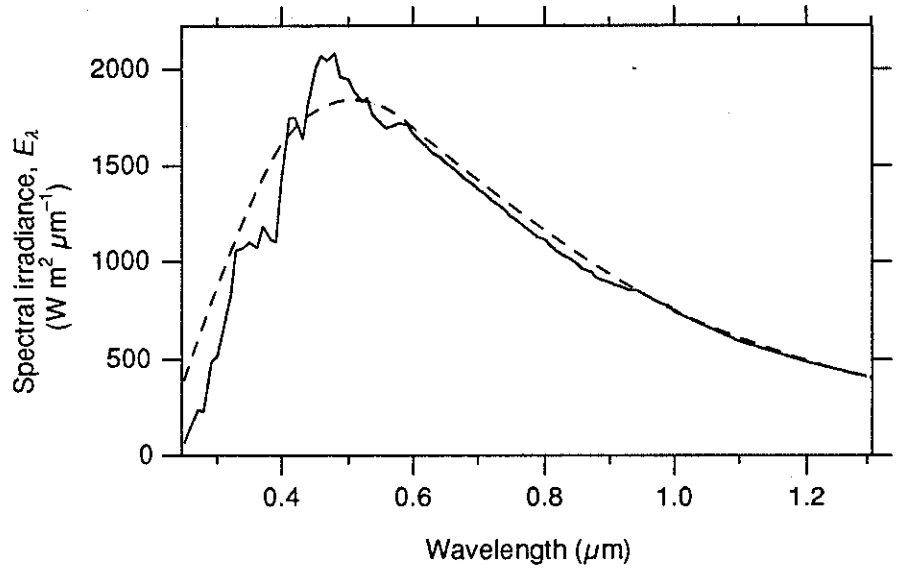


FIGURE 2.6: The solar spectral irradiance at the top of the atmosphere (full line) compared with the irradiance of a 'perfect sun' represented by a blackbody at 5800 K (broken line).
(The solar spectrum is based on reference [21].)

radiance is uniquely defined by its temperature according to Planck's law [3, page 448]:

$$L_\lambda = \frac{2hc^2}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1}, \tag{2.12}$$

where k is Bolzmann's constant, T is the absolute temperature of the body, and the other symbols are as defined earlier.

Objects which are not ideal absorbers also radiate although their radiance is always inferior to that of a blackbody. The total power radiated from an illuminated object is therefore the sum of that reflected and that emitted. Knowing the balance between solar reflection and thermal emission for a tar-

get is important in remote sensing [17]. To compare these, consider first the radiance from two different objects, one an ideal absorber at 'room temperature' ($T = 300$ K), the other a non-absorbing Lambertian reflector illuminated by the sun. The radiance of the former is given by Equation 2.12, that of the latter by Equation 2.6 by taking $\Omega_S = 6.8 \times 10^{-5}$ sr as the solid angle subtended by the sun, and L_S to be the radiance of a blackbody at 5800 K. Ignoring atmospheric absorption, the two radiances are plotted in Figure 2.7. It is the intersection point between these graphs which is important. A real, partly ab-

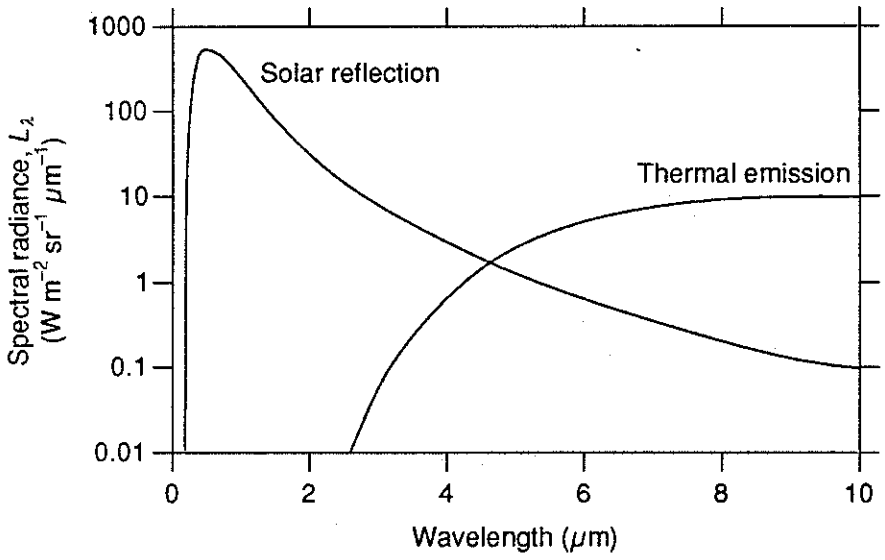


FIGURE 2.7: Comparison between the solar reflection from a perfectly diffuse, perfectly reflecting surface and the thermal emission from a perfect absorber at 300 K. The sun is assumed a 5800 K blackbody and atmospheric absorption is ignored.

sorbing, partly reflecting object has both reflection and emission curves. They are similar but inferior to those in the figure and their intersection points are at shorter wavelengths for highly absorbing objects and at longer wavelengths for highly reflecting objects. By considering the crossover points for naturally occurring reflectances two extremes materialize, marking the limits for where solar reflectance and thermal emission dominate [17]. Usually, solar reflection is said to dominate up to 2.5 μm and thermal emission above 6 μm. Since we have confined ourselves to wavelengths below 1 μm, we may therefore safely assume solar reflection to be the dominant source of radiation.

2.2.3 Source–target transfer

We distinguish between two different target types: reflective targets and transmissive targets. The **reflective target** is assumed to act as a Lambertian radiator with a reflectance factor R, illuminated by an 'ideal' sun represented by a 5800 K blackbody. Atmospheric absorption (Figure 2.8) is accounted for by the transmission factor T_A which also accounts for the effects of solar alti-

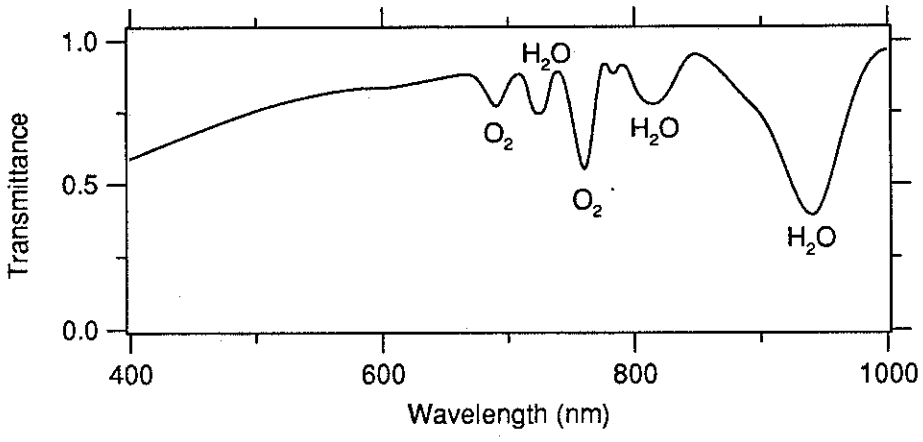


FIGURE 2.8: Atmospheric transmittance along the zenith measured at sea level under 'excellent' visibility conditions (more than 80 km).
(Adapted from reference [15].)

tude and meteorological conditions whose approximate attenuation factors are summarized in Table 2.2. Combining Equation 2.5 with Equation 2.11 now allows the Radiation Budget for reflective targets to be written in full as:

$$i = \frac{E_\lambda T_A R G T_O \varrho \, \Delta\lambda}{\pi}. \tag{2.13}$$

The **transmissive target** is one we study by looking at light transmitted through it from a source behind it, see Figure 2.9. Although such targets in general both transmit and scatter light, we will here only consider transparent gases assumed to be non-scattering fluids characterized by their transmission coefficient T. If the source or background against which the target is observed has spectral radiance L_λ and the atmosphere has transmission coefficient T_A, the Radiation Budget for such targets becomes:

$$i = L_\lambda T_A T G T_O \varrho \, \Delta\lambda. \tag{2.14}$$

Typical backgrounds include the sun, the moon, and the blue sky. Idealized radiance spectra for these are drawn in Figure 2.10 with the sun represented

TABLE 2.2: Approximate ratios between direct sunlight and other forms of natural illumination in the visible based on human eye response[a]. The column marked 'Exposure steps' shows $\log_2$ of the ratios corresponding to 'aperture steps' in photography.

Source type	Attenuation factor	Exposure steps ($\mathcal{E}$)
Direct sunlight	1	0
Full daylight (not direct sunlight)	10^{-1}	–3
Overcast day	10^{-2}	–7
Very dark day	10^{-3}	–10
Twilight	10^{-4}	–13
Deep twilight	10^{-5}	–17
Full moon	10^{-6}	–20
Quarter moon	10^{-7}	–23
Starlight	10^{-8}	–26
Overcast starlight	10^{-9}	–30

[a]Adapted from reference [15].

by a 5800 K blackbody and the moon by a Lambertian reflector illuminated by the ideal sun. Sky light is assumed to be entirely due to Rayleigh scattering of ideal sunlight [3, page 469] and its radiance curve is adopted from [15]. Note that the sun and the moon are sources of limited extent, both having an angular diameter of about 0.5 degrees. A sensor with field of view larger than this will therefore not achieve full throughput performance unless it is equipped with a telescope: by the conservation of throughput, field of view can then be reduced in return for a larger aperture, see Figure 2.11. Without such equipment, throughput G must be calculated using the solid angular extent of the source rather than the instrument's field of view. With a two degree field of view this represents a 16 times loss in throughput.

2.2.4 Signal criterion

The ultimate criterion for an instrument is the clarity of its output, often measured in terms of the *signal to noise ratio*, SNR. The various noise sources encountered in our instrument and their effects on the output spectra will be discussed in Chapter 3. Here we just mention the fundamental limit to noise

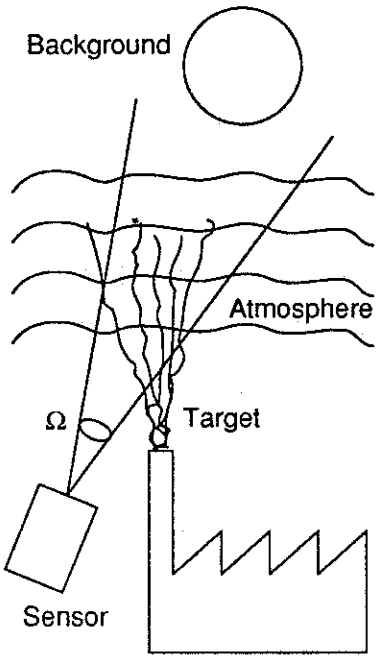


FIGURE 2.9: A typical measurement situation for transmissive targets.

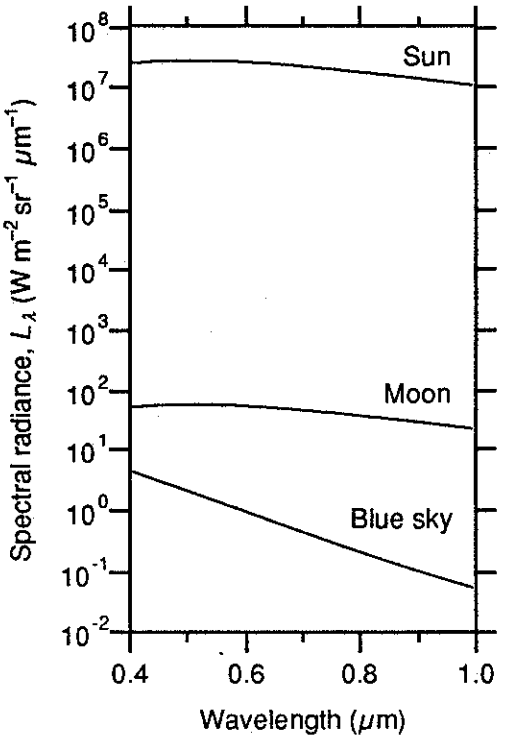


FIGURE 2.10: Spectral radiance for three background sources used for transmissive targets.

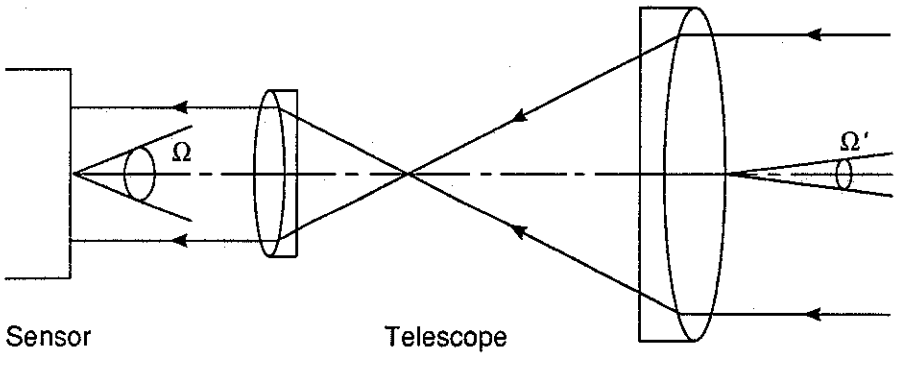


FIGURE 2.11: The telescope lens reduces field of view while conserving throughput.

performance given by the statistical uncertainty of a measurement of individual events such as photon to electron conversions. If a signal S is measured in number of electrons then its uncertainty, or noise, N equals the square root of S. Theoretically it is therefore possible to achieve any SNR by collecting sufficiently many photons.

Apart from the practical difficulties posed by target lifetimes etc. this technique is limited by detector *saturation* and *dark current*. Saturation occurs

when the number of electrons output from the detector exceeds the detector's storage capacity. In our detector the storage is limited to 20 pC or 1.24×10^8 electrons giving a theoretical maximum SNR per exposure of 11 000.

Dark current (i_D) is caused by leakage of charge across the detector diode. For very low signal levels, dark current will dominate and seriously reduce the SNR. It is therefore sensible to require

$$i \gg i_D \tag{2.15}$$

as a practical criterion for the signal current. Dark current in our detector is plotted against temperature in Figure 2.12. Note the strong temperature

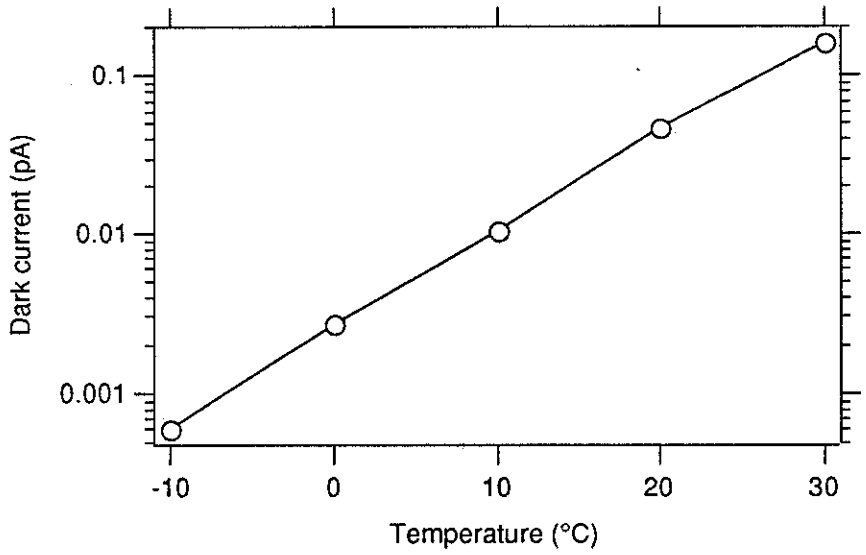


FIGURE 2.12: Dark current plotted against temperature for the detector used in our instrument as measured by the manufacturer.

dependence of this effect. At an ambient temperature of 20°C, $i_D = 0.046$ pA. We use this value to calculate instrument requirements in the examples below.

2.2.5 Examples

To illustrate the use of our Radiation Budget and set performance criteria for the subsequence analysis of instrumentation possibilities we consider two practical examples representing typical applications for spectral remote sensing in the visible: reflectance measurement of vegetation and transmission measurements of the atmosphere. In each example a minimum throughput requirement is specified and used to assess instrument performances in the next section.

A reflective target: the vegetation red edge. Several workers have measured the reflectance of vegetation in the visible, see Figure 2.13, with particular attention to the long-wavelength edge of chlorophyll absorption at about 700 nm, called the "red edge". Gates et al. [18] describes how growing

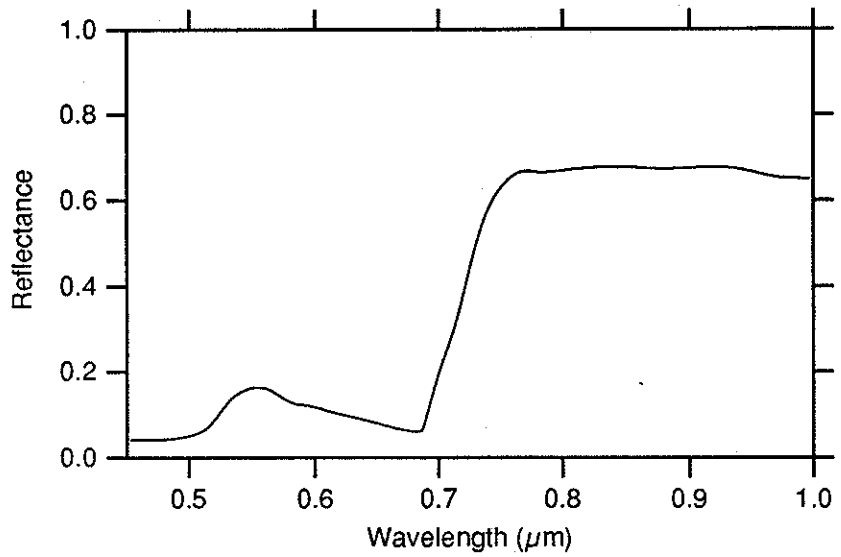


FIGURE 2.13: Reflectance spectrum of green grass as measured by the author.

oak leaves exhibit a systematic shift of this edge towards longer wavelengths and suggest its position to be a good indicator of chlorophyll content. Horler et al. [24] point out the advantages of using high resolution measurement of the red edge rather than the more commonly used ratio of two broad bands on either side of the edge for chlorophyll measurements, claiming an ability to eliminate effects of varying ground cover. More recently, Boochs et al. [27] extend the study of the red edge to more than just its position by considering first derivative spectra. They record spectra with 2 nm wide spectral channels whose first derivative displays a double rather than a single inflection peak. Having found good correlations between these features and circumstantial factors such as sowing date and nitrogen treatment, they predict a great future for plant vitality analysis by remote sensing.

Our first case study considers the Radiation Budget applied to such measurements by demanding a spectral channel width of 1 nm in the region around 700 nm. With two channels per resolution element as demanded by sampling theory, this corresponds to a resolving power of 350. The typical vegetation

reflectance spectrum of Figure 2.13 shows clearly the red edge in addition to the more familiar although rather less prominent green peak. Let us use the reflectance of the low reflectance part of the spectrum, about 10 %, as target reflectance in the budget. From curves shown earlier in the chapter we pick typical values for the other quantities required, thus:

$$\begin{aligned} E_\lambda &= 1500\,\mathrm{W/m^2/\mu m} \text{(from Figure 2.6)} \\ T_A &= 0.8 \text{ in full sunshine (from Figure 2.8)} \\ R &= 0.1 \text{ (from Figure 2.13)} \\ \varrho &= 183\,\mathrm{mA/W} \text{ (from Figure 2.5)} \\ \Delta\lambda &= 1\,\mathrm{nm} \end{aligned}$$

The 'balance' of the Radiation Budget of Equation 2.13 then becomes:

$$i = 7.0\,GT_O\ [\mathrm{mA}]. \tag{2.16}$$

Demanding $i \gg i_D = 0.046$ pA as discussed above now gives the following requirement for instrument performance:

$$GT_O \gg 6.6 \times 10^{-12}\ \mathrm{m^2\ sr}. \tag{2.17}$$

The SI unit "$\mathrm{m^2}$ sr" is somewhat opaque so let us instead introduce the more illustrative unit "square centimetre-degree": $1\,\mathrm{cm^2\ deg^2} = 3.05{\times}10^{-8}\,\mathrm{m^2}$ sr. Then:

$$GT_O \gg 2.2 \times 10^{-4}\,\mathrm{cm^2\ deg^2}. \tag{2.18}$$

It may be interesting to perform measurements under less favourable conditions than full sunshine. From Table 2.2 it is seen that a 'very dark day' requires a throughput three magnitudes higher than under optimal conditions. Increasing throughput further will allow observations even under twilight conditions. This may be found interesting for measurements performed in polar regions. Note that under such conditions the ambient temperature is probably significantly lower than 20°C. Dark current is then also reduced as shown in Figure 2.12, and hence the throughput requirement: at freezing (0°C) dark current is down by one order of magnitude.

A transmissive target: atmospheric NO_2. For our second case study, we consider the blue absorption of NO_2 centred at about 490 nm. This gas is of much concern with respect to atmospheric chemistry, vegetation damage, and respiratory problems. Its various sources and sinks are not clearly understood, and the development of better chemical understanding depends upon observational campaigns possibly based on remote sensing techniques [29].

Although the blue NO_2 absorption is a complex combination of very fine bands fully resolvable only with resolving powers of the order of 150 000, significantly lower resolving powers allow the study of its outline as seen in Figure 2.14. The absorption shows up in atmospheric transmission spectra as

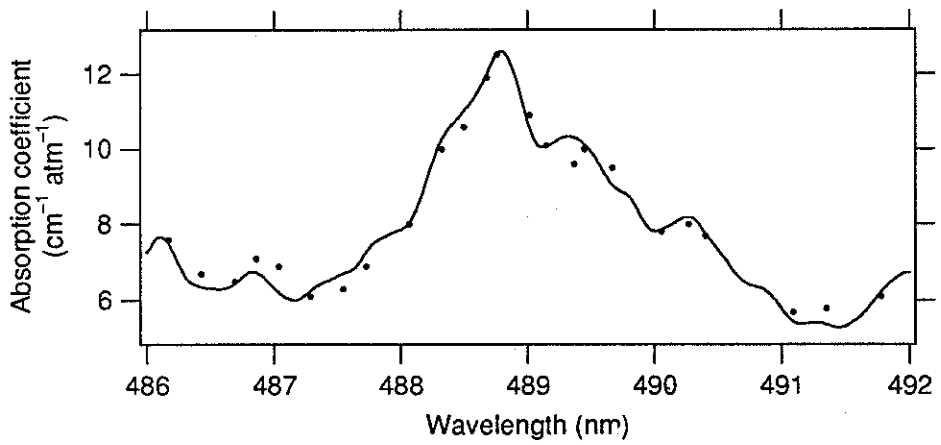


FIGURE 2.14: Absorption coefficient for NO_2 around 490 nm. The points are picked off a published curve[a] while the full line represents a measurement taken with our instrument.

[a] Reference [22].

a "hole" whose depth depends upon the concentration of the gas and the path length through it. Estimates of detection limits are given in Chapter 5.

Taking background spectral radiance values from Figure 2.10, atmospheric transmittance from Figure 2.8, detector detector responsivity from Figure 2.5, and assuming that the transmittance of the atmosphere due to its content of NO_2 and other pollutions is reduced by 0.5, we may furnish the Radiation

Budget of Equation 2.14 with the following values:

$$
\begin{aligned}
L_\lambda &= \begin{cases} 2.8 \times 10^7\,\mathrm{W/m^2/sr/\mu m} & \text{for the sun} \\ 57\,\mathrm{W/m^2/sr/\mu m} & \text{for the moon} \\ 2.2\,\mathrm{W/m^2/sr/\mu m} & \text{for the blue sky} \end{cases} \\
T_A &= 0.75 \text{ for a clear sky} \\
T &= 0.5 \\
\varrho &= 179\,\mathrm{mA/W} \\
\Delta\lambda &= 0.1\,\mathrm{nm} \\
i_D &= 0.046\,\mathrm{pA} \text{ at } 20\ ^\circ\mathrm{C}
\end{aligned}
$$

Following the same criterion as above for detector current, the instrumental demands can then be summarized as:

$$
GT_O \gg \begin{cases} 2.5{\times}10^{-16}\,\mathrm{m^2\ sr} = 8.2{\times}10^{-9}\,\mathrm{cm^2\ deg^2} & \text{for the sun,} \\ 1.2{\times}10^{-10}\,\mathrm{m^2\ sr} = 3.9{\times}10^{-3}\,\mathrm{cm^2\ deg^2} & \text{for the moon, and} \\ 3.1{\times}10^{-9}\,\mathrm{m^2\ sr} = 0.10\,\mathrm{cm^2\ deg^2} & \text{for the blue sky.} \end{cases}
$$

2.2.6 Spectroscopic techniques

The purpose of spectroscopic instruments is to disperse light along a wavelength or frequency axis so as to display its spectral content. The simplest method of dispersion is to let the light through different narrow band filters and measure the transmitted power for each filter. This is a technique which has been much employed in remote sensing, but it becomes unpractical when continuous coverage at high resolution is desired. For such work dispersion either by a prism, a grating, or an interferometer is more appropriate. Gratings outperform prisms with respect to resolving power: due to constrained availability of optical materials, it is difficult to push prisms beyond 30 000 [5, page 141]. Gratings push this limit about an order of magnitude further, and interferometers yet another. For our present purposes resolving power is not a limiting factor for any of these spectroscopic techniques however.

Prism and grating instruments have many things in common, in particular throughput. We will therefore discuss only one of them here, and since grating instruments give more freedom in the choice of dispersion and other

characteristics, they are the ones chosen. We refer to this type of instrument as *classical grating spectrometers* (CGS), the word classical is added to avoid confusion with an interferometric construction which also uses a grating.

There are two different types of interferometric spectrometers: the Fabry-Perot spectrometer and the Fourier transform spectrometer (FTS). The former is primarily used to achieve extremely high resolving powers over short spectral ranges and requires very high precision optics ($\lambda/50$) [5, page 182]. We will therefore not consider it here. The FTS principle is far more versatile and lends itself well to field remote sensing in its non-scanning form, known as holographic FTS (HFTS).

FTS has a series of advantages over CGS. Apart from the higher resolving power attainable, these include:

- Much reduced size at high resolving powers,
- Increased flexibility with respect to resolution and throughput for a given instrument,
- Noise advantage in the infrared (the 'Fellgett advantage'), and
- Considerably higher throughput (the 'Jacquinot advantage').

Since we are only interested in low resolution work in the visible (where statistical photon noise is predominant), only the latter advantage is of importance. This is also the one we are considering via the Radiation Budget.

Other factors of importance for our instrument are those related to mechanical construction and operational ease. We will therefore only consider non-scanning instruments where the spectral information is presented as a spatial intensity pattern rather than as a temporal intensity variation as in scanning instruments. This allows for much simplified mechanical constructions and avoids power consuming motors driving the scanning mechanisms because the information may be extracted by the use of an array detector. We will in the following two sections consider basic design parameters for non-scanning versions of both CGS and FTS type instruments and give examples relevant for the Radiation Budget.

2.3 Classical Grating Spectrometers

This spectroscopic technique achieves dispersion by a wavelength dependent deviation of the light. A grating deviates the light from its original direction by the phenomenon of diffraction; the simple construction based on the Fermat principle shown in Figure 2.15 uncovers the wavelength dependence of this

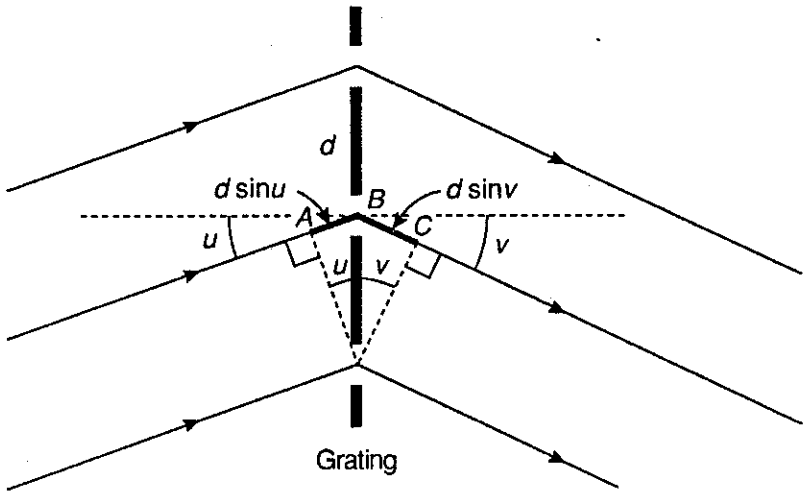


FIGURE 2.15: Construction based on the Fermat principle for diffraction by a grating: $d(\sin u + \sin v) = m\lambda$, where m is a whole number.

process. The spectral information emerging from the grating is thus angularly encoded, but for the angular code to be unambiguous, all the light 'rays' input to the grating must be parallel. This is assured by spatial filtering: the light analysed passes through an aperture slit followed by a lens (the 'collimator') before it enters the grating. In order to decode the spectrum, an 'angular decoder' in the form of a second lens (the 'camera') is placed in the output beam. The spectrum is now presented as a spatial intensity variation in the focal plane of the camera.

2.3.1 Practical instrument constructions

A schematic drawing of this system is given in Figure 2.16(a). Although perfectly feasible as a spectrometer design, more typical systems use refl ective gratings in the 'Littrow mode' where the light is retro-reflected (for one wavelength) back along its path. The collimator and camera functions are then both taken by the same lens and the spectrum is located at or close to the

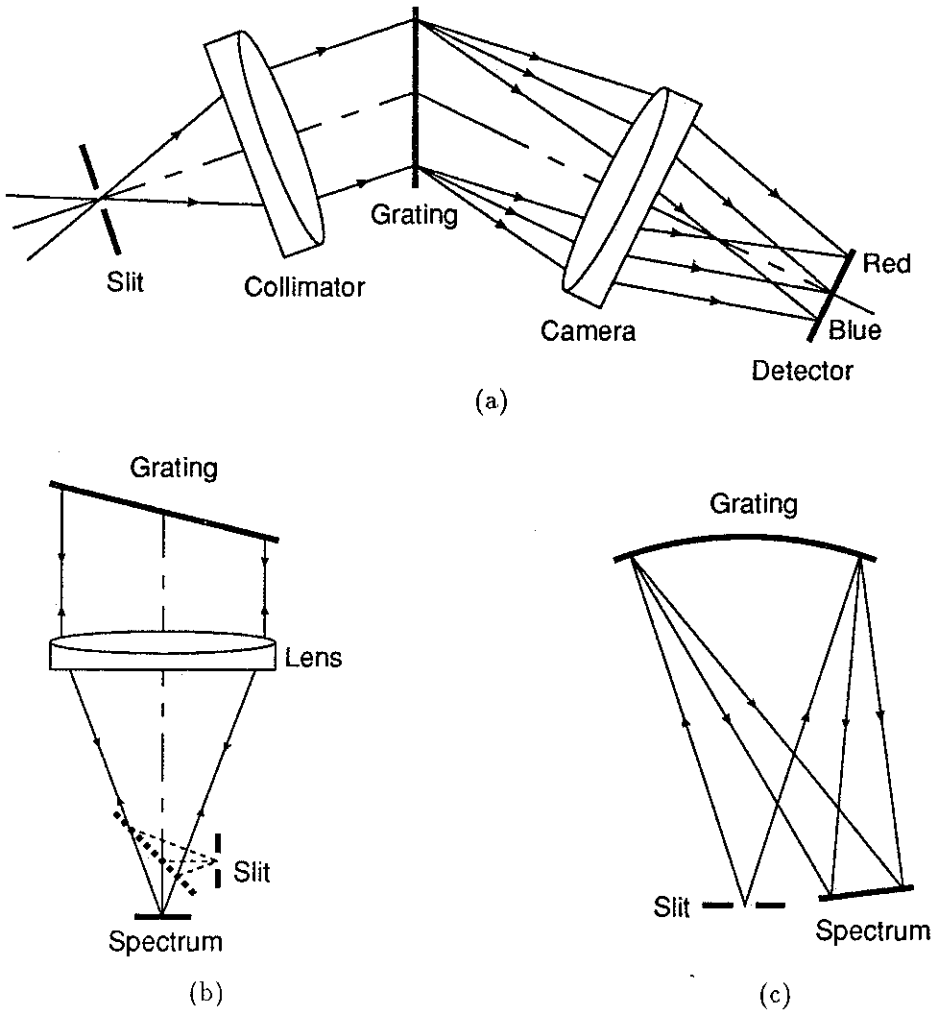


FIGURE 2.16: Principle components of a grating spectrometer (a) and two common practical implementations: the Littrow configuration (b) and the concave grating configuration (c).

entrance aperture. A flat mirror is often used to separate the two, see Figure 2.16(b). Another much used construction is the concave grating where the functions of collimator and camera are performed by the grating substrate (Figure 2.16(c)).

2.3.2 Focal ratio

A relationship between the width of the aperture slit and spectral resolution may be found by seeing that the spectral intensity distribution due to a monochromatic input beam is an image of the slit. In polychromatic illumination each wavelength produces an independent slit image slightly shifted with

respect to neighbouring wavelengths. Two spectral lines can therefore only be resolved if the aperture is sufficiently narrow that its two images do not overlap. This does not mean that infinitely fine features may be resolved just by narrowing the aperture; diffraction in the camera lens ensures the slit image never to be narrower than about $\lambda f/D = \lambda F$, where f is the focal length of the camera, D its aperture diameter, and $F = f/D$ its focal ratio.

Letting Δx denote the separation between samples, this gives rise to a focal ratio requirement:

$$F \leq \frac{\Delta x}{\lambda} \tag{2.19}$$

For visible light ($\lambda \sim 0.5$ μm) with a detector of width 25 μm, the focal ratio must therefore be kept below 100.

Reducing F improves the throughput of the instrument. To which degree F may be reduced depends upon practical considerations: production techniques limit grating widths to about 100 mm, and aberrations make focal ratios of less than unity difficult to achieve. For high resolving powers, long focal lengths are required [5, page 152] making grating size the limiting factor, but for resolving powers lower than about 10 000, focal lengths may be short enough to make aberrations the limiting factor for optical throughput.

2.3.3 Slit height

So far no mention has been made of the direction perpendicular to the dispersion. Since no spectral coding is present in this direction a much larger spread of angles can be tolerated, hence the name 'slit' for the spatial filter aperture. The slit cannot be infinitely long however: considering again as in Figure 2.15 the Fermat principle we see in Figure 2.17 that for off-axis rays the path difference between adjacent grating apertures is longer than for axial rays. The image of a long slit is therefore curved towards shorter wavelengths. Demanding that the position of a ray at the edge of the slit image should not deviate more than one resolution element from the position of a ray at the centre of the slit image, it may be shown that the angular deviation at the

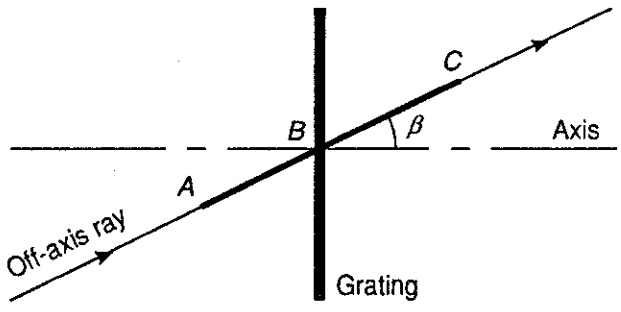


FIGURE 2.17: Top view of the grating shown in Figure 2.15 with light entering at angle β from the axis. If points A, B, and C correspond to those in the other figure, then the grating equation now becomes $d(\sin u + \sin v)/\cos\beta = m\lambda$.

grating in the non-spectral direction must be kept less than $\sqrt{2/\mathcal{R}}$ radians, where $\mathcal{R} = \Delta\lambda/\lambda$ denotes resolving power. This result is interesting because it is identical to the field of view of interferometric spectrometers, as will be seen later. For low resolution grating instruments, aberrations and detector sizes tend to limit the slit length to less than this optimum, however.

2.3.4 Instrument transmission factor

Concerning the instrumental transmission factor (T_O), diffraction efficiency of the grating is the main loss factor. Diffraction efficiency is wavelength dependent, and for low resolutions, assuming scalar blaze theory [10], the spectral transmission curve for a perfectly blazed grating is approximately as shown in Figure 2.18. Losses due to reflection and scattering at optical surfaces reduce transmission further. Although it is impossible to give a universal value for the transmission factor, an average of $T_O = 0.5$ will be used as an optimistic estimate.

2.3.5 Example: the concave holographic grating

A type of grating well suited for compact, non-scanning instruments is the concave interference or concave holographic grating. It is produced by exposing a layer of photosensitive material ('photoresist') deposited on a curved substrate to the interference between two laser beams. As pointed out earlier, the

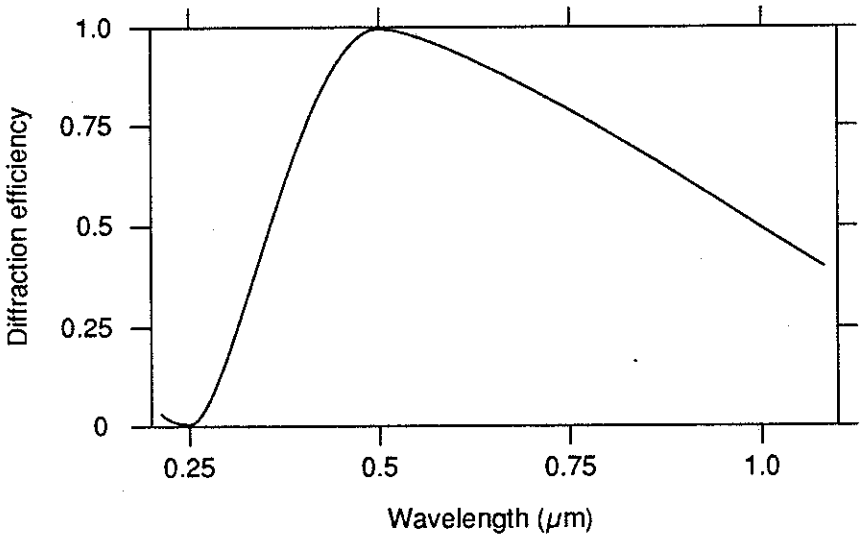


FIGURE 2.18: Idealized efficiency curve for a grating with blaze wavelength 0.5 μm.
(After reference [10].)

concave grating disposes of both collimator and camera and gives therefore a very simple optical system, see Figure 2.16(c). Producing the grating grooves by interference rather than traditional ruling allows non-straight grooves to be made by playing with the shapes of the interfering wave fronts. This allows correction of certain aberrations, exemplified by the so-called 'Type III' grating formed by two spherical wave fronts originating from certain geometrically constructed points [12]. The spectrum formed by such a grating is well focussed onto a nearly flat field and is therefore ideally suited for diode array detectors.

A typical system [11] has a focal length of 200 mm and a groove density at the centre of the grating of 300 per mm. It spreads the 400–800 nm range over 25.4 mm which, when sampled at 25 μm intervals, gives a resolution of approximately 1 nm ($\mathcal{R} \sim 600$). The maximum allowable slit height is, according to the foregoing discussion, about 20 mm but off-axis aberrations in the grating limits its useful length to about 2 mm. This corresponds well with the 2.5 mm detector length in the one-dimensional array we have used. The grating is circular with a diameter of 70 mm and a focal ratio $F = 3$. Since the solid angle of the converging beam to a good approximation is $\Omega_G = \pi/(2F)^2$,

the throughput calculated at the detector surface for a single channel is:

$$\begin{aligned} G &= A_D \Omega_G \\ &= l_D \, \Delta x \, \frac{\pi}{4F^2} \\ &= 5.5 \times 10^{-9} \text{ m}^2 \text{ sr} \end{aligned} \tag{2.20}$$

where $A_D = l_D \, \Delta x$ is area and l_D and x_D are length and width respectively of each detector element. Using the estimated instrument transmission of $T_O = 0.5$ gives a throughput-transmission figure of

$$GT_O = 2.7 \times 10^{-9} \text{ m}^2 \text{ sr} = 0.090 \text{ cm}^2 \text{ deg}^2 \tag{2.21}$$

This system can be adapted to different resolving powers by changing the grating, assuming appropriate gratings to be available. Throughput is not affected by such a change as long as the same detector is employed and the new grating has the same focal ratio.

2.4 Fourier transform spectrometers

Fourier transform spectrometers (FTS) present the spectral information in a radically different form from that of classical grating instruments. The FTS output is an *interferogram* which can only be fully interpreted after it has been mathematically Fourier transformed. This method of spectroscopic analysis was first suggested by Michelson [6, page 316] in the end of the last century, but it is only during the last few decades that it has it become useful for general spectroscopic tasks thanks mainly to developments in digital computing.

The interferogram is formed in a two-beam interferometer where the incident light is split into two parts and recombined after having travelled unequal optical paths, see Figure 2.19. On recombination, the two beams interfere, and according to their relative phase, they interfere constructively or destructively. Phase difference is proportional to the ratio between optical path difference (OPD) and wavelength, and so, if OPD and hence phase is varied linearly, the output intensity for a monochromatic input beam varies sinusoidially. The frequency of this oscillation is inversely proportional to the wavelength of the light; hence it is proportional to the *optical* frequency. When polychromatic

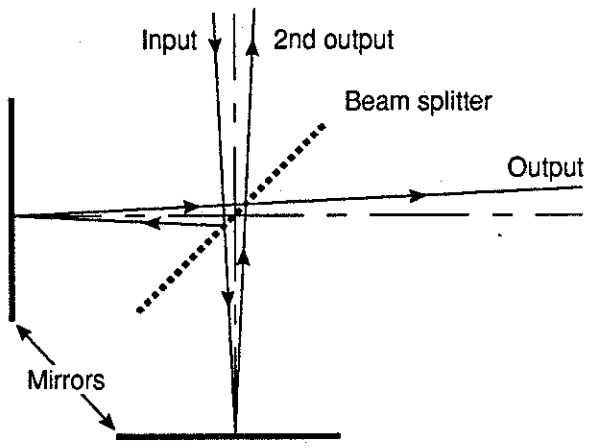


FIGURE 2.19: The Michelson interferometer.

light is fed into the interferometer, each spectral component produces its own sinusoid and the output is the incoherent sum of all these patterns. This complicated—and in white light colourful—pattern is the interferogram, see Figure 2.20.

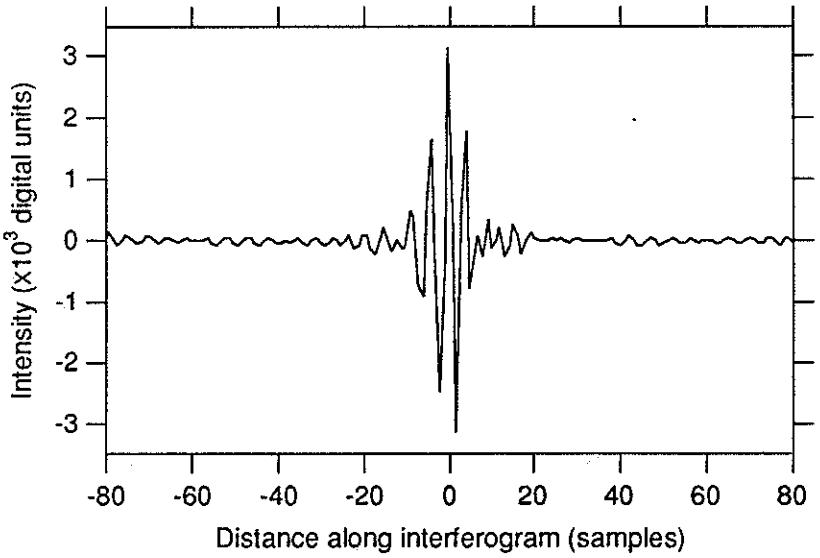


FIGURE 2.20: The central part of a typical interferogram measured with our instrument.

A signal can, according to Fourier theory, be represented uniquely as a sum of appropriately weighted sinusoids and the signal is therefore fully defined by the weighting factors (and the phases) of these sinusoids. Conversion from one of these representations to the other is performed by *Fourier transformation*. It follows from the preceding paragraph that the spectrum of the interfering light

represents the weighting factor signal for the interferogram; the spectrum is therefore the Fourier transform of the interferogram. Spectra obtained by this method are different from those obtained by classical grating spectrometers since they are linear in wavenumber (or frequency) rather than wavelength.

2.4.1 Classical and holographic FTS

Changing the OPD may be done either temporally by moving *(scanning)* one or both of the mirrors, or spatially by tilting one of the mirrors. Instruments based on the the former method (referred to as scanning or classical FTS) are usually preferred for laboratory based work since they offer both high resolution and wide spectral coverage. Those based on the latter (called non-scanning or *holographic* FTS, HFTS) are more suitable for our purpose where extreme resolving powers are not demanded, since they can be realized without moving parts. Like in the non-scanning grating instrument, an array detector collects the spatially distributed information.

Classical and holographic FTS offer identical instrumental throughputs, but while scanning systems collect all the power by a single detector, holographic systems divide it between the detectors in the array. In terms of *energy per sample*, however, the balance is regained since classical systems divides the observation *time* between all the samples while holographic systems allows all samples to be measured all the time.

The Michelson interferometer with a tilted mirror presents the interferogram as a 'fringe pattern', a series of bright and dark lines. These are fringes of equal thickness [6, page 301] parallel with the apex of the wedge formed between the mirrors. The fringes are virtual, localized near the mirror surfaces, and in order to be measured, they must be imaged onto the detector array by a 'fringe imaging lens', see Figure 2.21.

2.4.2 Spectral resolution

Spectral resolution of FTS instruments is, as shown below, equal to the reciprocal of the maximum optical path difference (OPD) between the two beams in the interferometer. This follows from the uncertainty involved in decomposing

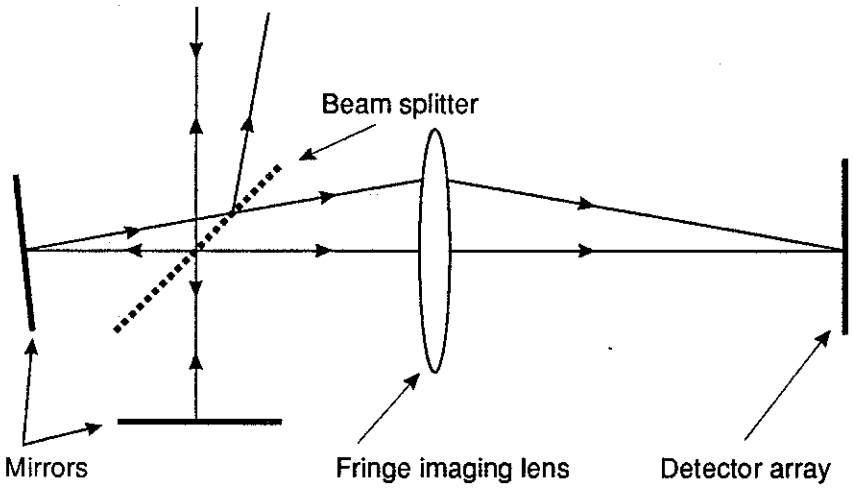


FIGURE 2.21: Sketch showing the principle of holographic FTS.

the interferogram signal into sinusoidal (Fourier) components. Like a guitar string only vibrates at frequencies corresponding to integral numbers of nodes along the string, a Fourier component can only be computed if it corresponds to an integral number of cycles, i.e. fringes, along the interferogram. Two neighbouring Fourier components σ and $\sigma - \Delta\sigma$, where $\sigma = 1/\lambda$, are therefore represented by N and $N-1$ fringes respectively where N is an integer. Since one fringe represents an OPD equal to one wavelength, the total OPD across the interferogram may be written as:

$$\xi = \frac{N}{\sigma} = \frac{N-1}{\sigma - \Delta\sigma}, \tag{2.22}$$

when N is the total number of fringes in the interferogram. Eliminating N from the equations yield the spectral resolution:

$$\Delta\sigma = \frac{1}{\xi} \tag{2.23}$$

as required. Eliminating ξ instead from Equation 2.22 gives the resolving power:

$$\mathcal{R} = \frac{\sigma}{\Delta\sigma} = N. \tag{2.24}$$

2.4.3 Heterodyned holographic FTS

The sampling theorem requires the finest fringes in the interferogram to be sampled twice per period. With a detector array of N_D detectors, the maximum number of fringes, and hence, by Equation 2.24, maximum resolving

power, is therefore $N_D/2$, i.e. typically of the order of a few hundred and about a thousand at best. This limitation may be overcome by the technique of *heterodyning*. The term is borrowed from radio theory and signifies the frequency shift seen when two signals of similar frequencies are multiplied. Such an effect is achieved in HFTS by replacing the tilted mirror with a grating, giving fringes that have spatial frequencies equal to the difference between the unheterodyned fringe frequency and the grating ruling frequency (see Section 3.2.3). Physically, this may be understood by seeing that diffracted wave fronts emerge from a grating in different directions according to their wavelength. The grating is arranged such that a wave front belonging to one end of the spectral range of interest is "retro-diffracted", i.e. it fulfils the grating's 'Littrow condition'. This wave front is therefore parallel to the reference wave front which returns from the mirror: their interference fringes are infinitely separated and have thus zero spatial frequency. At a neighbouring wavelength the diffracted wave front is no longer parallel with the reference and fringes of finite frequency are therefore produced. If the grating is well chosen, the fringes representing the other end of the spectral range equals the Nyquist frequency (half the sampling frequency) of the detector array.

There is of course an ambiguity problem involved with this method since wavelengths on either side of the Littrow wavelength may produce fringes of the same frequency. This together with the problems of aliasing posed by breaking the sampling condition, sets the following optical filtering condition: *Only light corresponding to an unambiguous spectral range must be allowed to contribute to the interferogram.* A formal expression of this condition is given in Section 3.2.4.

2.4.4 The throughput advantage

Heterodyned HFTS (usually shortened here and elsewhere HHS) can thus achieve high resolving powers at the expense of a limited spectral range. Although the maximum resolving power attainable with a given grating is the same as the diffraction limited resolving power offered by the same grating in a CGS system (see Section 3.2.7), HHS tends to be more efficient both in terms

of throughput and instrumental dimensions. HHS also offers diffraction limited resolution of the grating as a matter of course. These points are exemplified by a project proposing an HHS spectrometer of resolving power 10^5 for the 16 metres ESO telescope [62]. Due to atmospheric 'seeing', a CGS instrument utilizing the entire telescope throughput is estimated to require a grating of diameter 2.4 m while the proposed HHS system only needs 10 cm.

The main reason for this big difference is the FTS throughput advantage. In FTS the throughput is limited because of the variation in OPD with input ray angle as illustrated in Figure 2.22. For a ray inclined at an angle β to the

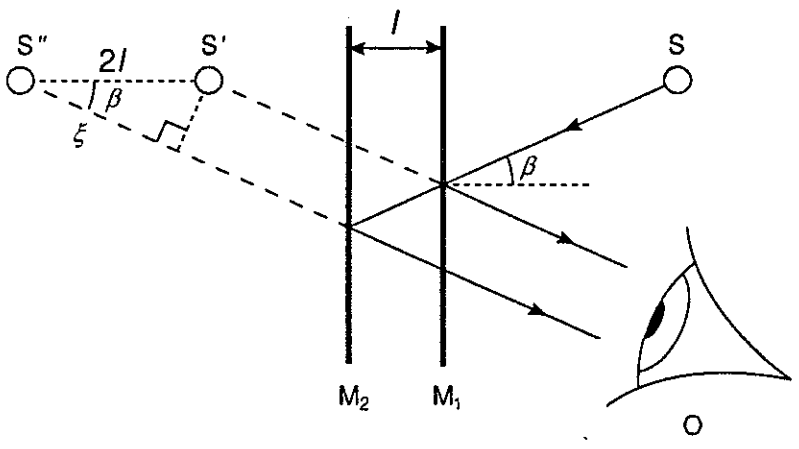


FIGURE 2.22: Construction of path difference for off-axis rays. The source S is imaged into S′ and S″ by mirrors M_1 and M_2 respectively.

axis the OPD is

$$\xi = 2l\cos\beta, \tag{2.25}$$

where l is the separation between the mirror (or mirror and grating) images. Hence, if a fringe pattern of an off-axis wave front with wavenumber σ has N fringes, it will be confused with the fringe pattern of an axial wave front of wavenumber $\sigma - \Delta\sigma$ which also has N fringes:

$$N = 2l\sigma\cos\beta = 2l(\sigma - \Delta\sigma), \tag{2.26}$$

giving:

$$\sigma\cos\beta = \sigma - \Delta\sigma. \tag{2.27}$$

Since β is always small enough that its cosine may be represented by the two first terms of its Taylor expansion, this may be approximated to:

$$\sigma(1 - \frac{\beta^2}{2}) = \sigma - \Delta\sigma, \tag{2.28}$$

and so:

$$\beta = \sqrt{\frac{2\,\Delta\sigma}{\sigma}} = \sqrt{\frac{2}{\mathcal{R}}}. \tag{2.29}$$

Note the correspondence between this result and that quoted for maximum slit height in CGS instruments. Since this limit must be respected in all directions, FTS instruments have a circular field-of-view of solid angle:

$$\Omega = \pi\beta^2 = \frac{2\pi}{\mathcal{R}} \tag{2.30}$$

2.4.5 Throughput optimization

In HFTS instruments the interferogram is presented as a pattern of straight, parallel lines. It is therefore essentially one dimensional and lends itself well to the use of a one dimensional array detector. For a large throughput, the fringes should be as long as possible, however, and this fits badly with the typical shape of array detectors. Although spectroscopic grade arrays tend to have elongated detectors (our array has elements of 25 μm by 2.5 mm), the fringes could be much longer, typically as long as the width of the fringe field. A cylindrical lens can then be used to collapse the fringes, see Figure 2.23. Optimally this lens is chosen so as to image the target onto the detector plane in the along-fringe direction.

Assuming properly collapsed fringes, it is easiest to calculate the instrument throughput at the interferometer mirrors where the fringes are located:

$$G = \Omega A_F = \frac{2\pi}{\mathcal{R}} x_F y_F, \tag{2.31}$$

where $A_F = x_F y_F$ is the area of the fringe field and x_F and y_F its 'across-fringe' and 'along-fringe' dimensions, respectively. If (as in our case) the across-fringe dimension is imaged at unit magnification onto the detectors then:

$$G = \frac{2\pi}{\mathcal{R}} N_D\, \Delta x\, y_F, \tag{2.32}$$

where Δx is the width of each detector element.

2.4.6 HFTS versus CGS

Each detector in an HFTS instrument "sees" the entire transmitted spectral range but only $1/N_D$ of the instrumental throughput (Equation 2.32). In CGS

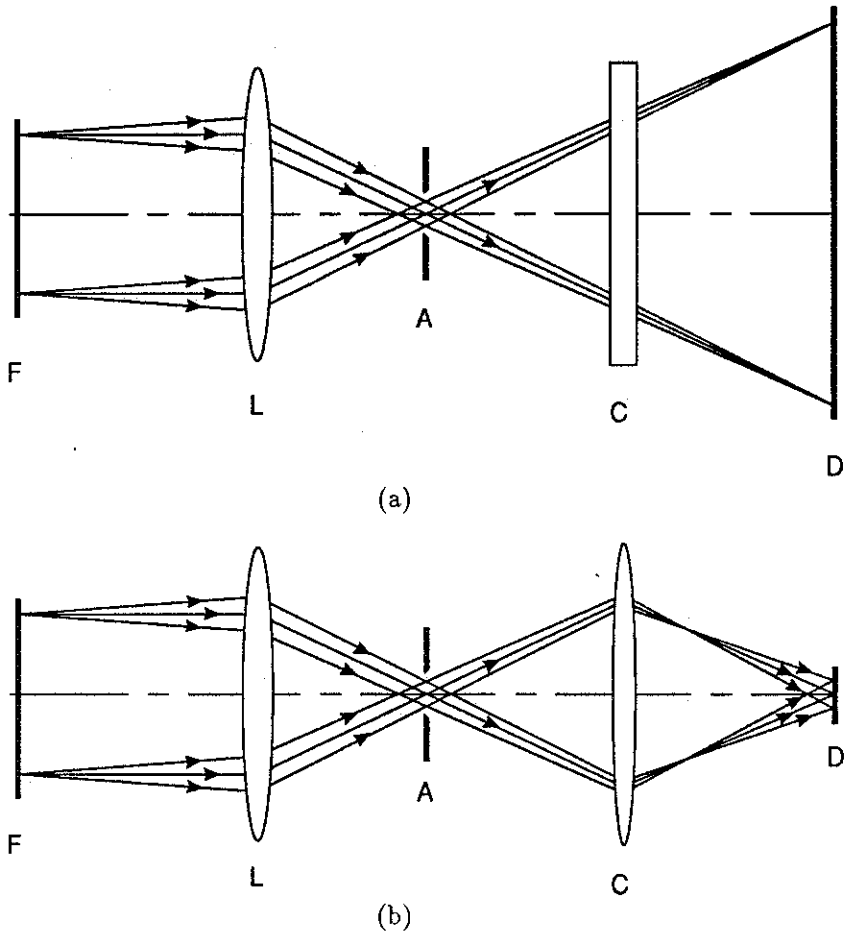


FIGURE 2.23: Two views of a simplified optical design for HFTS with a cylindrical fringe-collapsing lens. In (a) the fringes are perpendicular to the paper, in (b) they are parallel with the paper. F is the fringe field, a virtual object within the interferometer, L is the fringe imaging lens, A is the field limiting aperture, C is the cylindrical lens, and D is the detector array.

instruments, on the other hand, each detector benefits from the entire instrumental throughput (Equation 2.20), but sees only $1/N_D$ of the spectral range. To compare detector outputs, an assumption about the spectral distribution is therefore necessary: when the spectrum is quasi-continous so that all spectral channels are more or less equally filled, the two throughput expressions may be compared directly. Figure 2.24 plots throughput against resolving power for this situation assuming an HHS system using an array of 512 detectors with $\Delta x = 25\mu$m and a mirror height of $y_F = 12$ mm, and a CGS system as described in the example in Section 2.3.5. The HHS instrument is seen to have a comfortable throughput advantage over its CGS counterpart, partic-

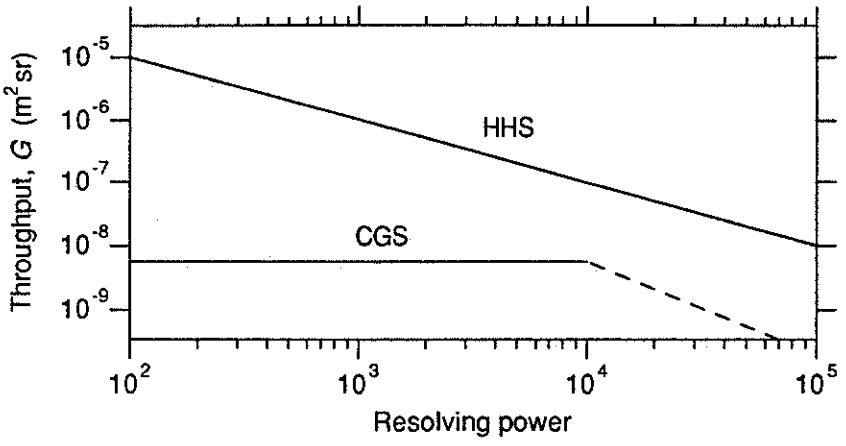


FIGURE 2.24: Comparison between optical throughput of the heterodyned holographic FTS (HHS) and the classical grating spectrometer (CGS) discussed in the text. The broken line for GCS at high resolving powers estimate the effects of grating size limitation and slit height restriction.

ularly at low resolving powers, representing a factor of 200 at $\mathcal{R} = 1000$ and of 2000 at $\mathcal{R} = 100$. [See p. 51a.]

2.4.7 Instrument transmission factor

Instrument transmission factor for HHS instruments is limited by the diffraction efficiency of the grating as in the CGS case. In addition comes a serious loss due to the beam splitter: half of the light entering the interferometer is reflected back out through the input port. Some interferometer constructions allow this light to be recuperated, but this adds a complexity to the design which has not been found worthwhile for our instrument. There are also losses due to spurious reflections at the outer surfaces of transmissive components, and if the beam splitter is metallic then there is also a loss due to absorption. Estimating the total transmission factor to $T_O = 0.2$, gives a throughput-transmission product of

$$\begin{aligned} GT_O &= \frac{1.93\times10^{-4}}{\mathcal{R}} \text{ [m}^2 \text{ sr]} \\ &= \frac{6.3\times10^{3}}{\mathcal{R}} \text{ [cm}^2 \text{ deg}^2\text{]}. \end{aligned} \tag{2.33}$$

For resolving powers of 350 and 2500 as used in our Radiation Budget examples, the proposed HHS design offers throughput-transmission figures of 18 cm^2 deg^2 and 2.5 cm^2 deg^2 respectively.

To compare the two instruments when the spectrum is not quasi-continous, we may consider the signal collected per detector element. In HHS, the measured interferogram fluctuates around a certain level I_0 which for resolved spectra equals the average interferogram signal (see Figure 3.12). Since I_0 is proportional to the total spectral signal transmitted by the instrument ($\sum B$, say), the average detector signal is:

$$\overline{I_H} = I_0 \propto \frac{G_H}{N_D} \sum B = G_H \overline{B}, \tag{2.32a}$$

where G_H is instrumental throughput (equal to G in Equation 2.32), N_D is the number of detector elements, and $\overline{B} = \sum B/N_D$ is the average spectral value in the measured (discrete) spectrum.

In CGS, the detector signal is proportional to the local spectral value B:

$$I_G \propto G_G B, \tag{2.32b}$$

where G_G is the instrumental throughput given by Equation 2.20.

Ratioing Equations 2.32a and 2.32b and substituting from Equations 2.20 and 2.32 gives:

$$\frac{\overline{I_H}}{I_G} = \frac{G_H}{G_G} \frac{\overline{B}}{B} = 8F^2 \frac{y_F}{l_D} \frac{N_D}{\mathcal{R}} \frac{\overline{B}}{B}. \tag{2.32c}$$

The two first factors are system-related: in a grating instrument signal is improved by reducing focal ratio (F) of the camera lens or by increasing slit height (i.e. detector length, l_D). In an HHS instrument signal is improved by increasing mirror height, y_F.

The third factor shows the effects of detector array size and resolving power. Note that although N_D enters the expression because of its connection with the mirror area, it may here be taken to represent the maximum *un*heterodyned resolving power (Equation 2.24). The ratio $N_D/\mathcal{R}$ may therefore be seen as a "heterodyning factor".

The ratio $\overline{B}/B = f$ is often referred to as the "spectral fill factor" and will be encountered later in noise calculations (Section 3.4.4). Quasi-contious spectra have $B \approx \overline{B}$, hence $f \approx 1$.

When $F = 3$, $y_F = 12$ mm, $l_D = 2.5$ mm, and $N_D = 512$, we find $I_H/I_G = 1.8{\times}10^5 f/\mathcal{R}$. For a spectrum with unit fill factor measured at a resolving power of 100 the ratio is about 2000 as estimated from Figure 2.24.

2.5 Conclusion

Table 2.3 summarizes the results of this concept study showing both instrumental requirements for some specific measuring tasks, and expected performances

TABLE 2.3: Summary of the radiation budget calculations showing required and estimated instrument performances as calculated in the text. (a) represents the vegetation reflectance example and (b) represents the atmospheric transmittance example. i is signal current, i_D is dark current, and τ_E is an estimate of the exposure time.

	Required	Classical grating:		Holographic FTS:	
	GT_0 product	$GT_0 = 0.090$ cm^2 deg^2		$GT_0 = 18$ cm^2 deg^2	
Illumination	(cm^2 deg^2)	i/i_D	τ_E	i/i_D	τ_E
Direct sunlight	2.2×10^{-4}	410	1 sec	8.2×10^{4}	5 ms
Overcast day	0.022	4.1	2 min	820	0.5 sec
Very dark day	.22	0.41	–	82	5 sec
Twilight	2.2	0.041	–	8.2	1 min

(a)

	Required	Classical grating:		Holographic FTS:	
	GT_0 product	$GT_0 = 0.090$ cm^2 deg^2		$GT_0 = 2.5$ cm^2 deg^2	
Illumination	(cm^2 deg^2)	i/i_D	τ_E	i/i_D	τ_E
Sun	8.2×10^{-9}	1.1×10^{7}	40 μs	3.8×10^{8}	1 μs
Moon	3.9×10^{-3}	23	20 sec	640	0.7 sec
Blue sky	.10	0.9	9 min	25	20 sec

(b)

of two different instruments, one based on classical grating spectroscopy (CGS) and the other on heterodyned holographic Fourier transform spectroscopy (HHS). A performance criterion has been obtained by requiring that the detector signal current should be greater than detector dark current. Although not very rigorous, this is a practical and operationally sound criterion.

As seen from the table, both instruments may be used with well lit targets. Using the blue sky as background for atmospheric transmission measurements pushes the CGS instrument to its limit, however, and when it comes to measuring vegetation reflectance under poor daylight conditions, this instrument

cannot cope unless some means of reducing its dark current (i.e. cooling) is employed.

The less conventional HHS design with a throughput-transmission figure almost 30 times superior to the CGS for the high resolution example, is seen to tackle the sky lit transmission measurement with a good margin. It also promises successful operation for reflectance measurements on a 'very dark day' as well as during twilight. Aided by natural cooling assuming an ambient temperature of 0°C which reduces the dark current by an order of magnitude it may even be expected to perform to specifications under 'deep twilight' condition, e.g. during polar winter.

Convinced by these possibilities and intrigued by the novelty of HHS in remote sensing, we have decided to build a spectrometer based on that principle. In the following chapters theory and design of the instrument is presented, culminating in a demonstration of its capabilities.

Chapter 3

Theory of holographic FTS

After a brief review of the literature concerning holographic and heterodyned holographic FTS, the theory for these types of instruments is presented. Basic theory is given in terms of electromagnetic interference and geometrical construction, and the Fourier transformation is introduced to explain and analyze the interference of white light. The effects of finite and sampled measurements are discussed in terms of Fourier theory.

In FTS instruments spectral information is measured in the form of an interferogram as seen in Figure 2.20. Ideally this interferogram is symmetrical, but in most practical cases it is asymmetrical due to the phenomenon of spectral phase. Correction for this effect is an important part of the signal processing of FTS data. After a presentation of the main sources of phase encountered in our instrument, we present their effects on spectral estimates and explain how to minimize them. The possibility of 'single-sided' interferograms is also discussed; such measurements offer potentially twice the spectral resolution of normal 'double sided' interferograms with the same number of samples. Such a prospect is obviously interesting for our system where the number of samples is restricted. We summarize the requirements to see under which conditions gains are to be made from this kind of operation.

The final subject to be treated is that of noise and other measurement related deficiencies. Some important sources of noise are presented together with a discussion of how to optimize the instrument's performance. Since the spectrum is obtained as a result of a mathematical transformation, the relationship between measured and spectral noise is not straight forward. We

show how the spectral estimate is affected by "white" intensity noise. Another important deficiency of the interferograms measured with our instrument is caused by aberrations in the interferometer. We show that this effect is similar to an error in the sampling grid and discuss the possibilities for correction by resampling of the interferogram.

3.1 Literature review

Holographic Fourier transform spectroscopy (HFTS) was first demonstrated in 1965 by G. W. Stroke and A. T. Funkhauser [47]. Inspired by the recent advances in holography they proposed the recording of a white light Fourier transform hologram on photographic film which, when illuminated by laser light, would yield a diffraction pattern proportional to the spectrum of the light source. One of the main advantages of the method over classical FTS at the time was that no computation was needed: the method was totally analogue.

With the revolutionary developments in computers and software algorithms the 'computational' advantage of HFTS was soon lost, leaving it with time consuming wet processing and cumbersome Fourier optical methods of spectral interrogation. The technique was to be revitalized by another technological advance however. With the advent of high quality solid state diode arrays the photographic plate could be avoided, opening up the possibility for real-time operation and compact and rugged instrument designs [54,55,58,59].

A major disadvantage of HFTS is its limited number of interferogram samples—a disadvantage that was further aggravated when diode arrays replaced high resolution photographic plates. As was demonstrated by T. Dohi and T. Suzuki in 1971 however, it is possible to *heterodyne* the interferogram by replacing one (or both) of the mirrors in the Michelson interferometer by a grating [50]. Other heterodyning schemes have since been published [51,57], but the original method is the one offering the greatest flexibility in choice of resolving power.

The first demonstration of HFTS was made using a Michelson interferometer, but other two-beam interferometers have since been used [48,49,53,63].

Of particular interest from a portability point of view is the polarization interferometer [53,60], offering very compact and rugged sensor designs. No heterodyning is available by this method however, so the resolving power is limited. For the present application where relatively high resolving powers ($\sim$ 5000) are wanted, we have chosen a standard Michelson interferometer heterodyned by a grating.

3.2 Basic theory

We start off this discussion with a presentation of the theory of interference, showing how the combined intensity of two coherent beams depends upon their relative phase. By geometric construction we then calculate the spatial variation in phase difference produced in holographic FTS and hence the intensity variation in the interference pattern produced. Resorting to the Fourier transform, we show how spectral information may be extracted from such interferograms, and by elaborating the Fourier theory we predict the shape of the spectral instrument function due to finite measurements. Finally, the effects of sampling are considered mathematically.

3.2.1 Interference

Fourier transform spectrometers use a two-beam interferometer—usually the Michelson or a variation of it—to superpose two mutually coherent light beams. It follows from electro-magnetic theory that such superposition creates *interference*: as the phase between the beams varies, their combined intensity goes through maxima which exceed the sum of their individual intensities and minima which may reach zero [6, page 256].

The process of interference may be shown in general by considering an oscillating signal described by:

$$A = a\cos(\omega t + \delta), \tag{3.1}$$

where a is its amplitude, ω its angular frequency of oscillation, δ its phase, and where t represents time. The power carried by this wave (corresponding

to the *intensity* of a light wave) is found by time averaging the square of the signal:

$$I = <A^2> = a^2 \ <\cos^2(\omega t + \delta)> = \frac{a^2}{2}. \tag{3.2}$$

When two signals of equal frequencies but different phase and amplitude are brought together, the resulting signal is $A_1 + A_2$, giving intensity:

$$\begin{aligned} I &= <(A_1 + A_2)^2> \\ &= <A_1^2 + A_2^2 + 2A_1A_2> \\ &= I_1 + I_2 + J, \end{aligned} \tag{3.3}$$

where I_1 and I_2 are the intensities of the two waves separately and J is the cross term given by:

$$\begin{aligned} J &= 2 \ <a_1 a_2 \cos(\omega t + \delta_1) \cos(\omega t + \delta_2)> \\ &= a_1 a_2 \cos(\delta_1 - \delta_2), \\ &= 2\sqrt{I_1 I_2} \ \cos \Delta\delta, \end{aligned} \tag{3.4}$$

where $\Delta\delta = \delta_1 - \delta_2$ is the phase difference between the two signals. The intensity therefore varies sinusoidally with the phase difference, a variation which is often referred to as an interference or *fringe* pattern.

In the special case when $I_1 = I_2$, the total intensity reduces to:

$$I = I_0(1 + \cos \Delta\delta), \tag{3.5}$$

where $I_0 = I_1 + I_2$ is equal to the incident intensity if the interferometer is lossless. The fringe pattern then varies between zero and twice I_0, representing optimal or 100% contrast. If instead $I_2 = c\, I_1$, the interference pattern is given by:

$$\begin{aligned} I &= I_1 + I_2 + 2I_1\sqrt{c} \cos \Delta\delta \\ &= I_0(1 + \frac{2\sqrt{c}}{1+c} \cos \Delta\delta) \\ &= I_0(1 + k \cos \Delta\delta), \end{aligned} \tag{3.6}$$

where k is the *fringe contrast*, see Figure 3.1. Contrast is an important instrument characteristic which also depends upon instrumental properties other

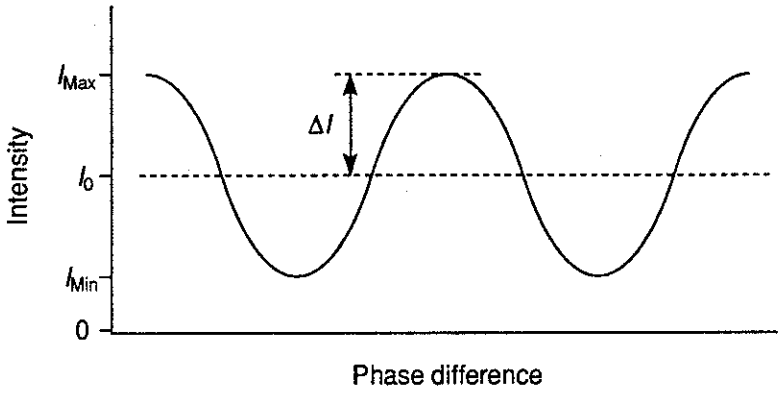


FIGURE 3.1: The contrast k of a sinusoidal intensity signal describes its depth of modulation: a fully modulated signal ($k = 1$) reaches zero at its minima. It is defined as the ratio between the sinusoidal amplitude, $\Delta I = (I_{\text{Max}} - I_{\text{Min}})/2$, and the mean signal, I_0, hence: $k = \Delta I / I_0 = (I_{\text{Max}} - I_{\text{Min}})/(I_{\text{Max}} + I_{\text{Min}})$.

than the interferometer balance alone as will be seen in Section 4.2.1.

Electro-magnetic theory models light as a transverse oscillation with two orthogonal modes or *polarizations* which do not interfere with each other. They may therefore be treated separately and the total intensity is given by their sum [6, page 259]. Polarization effects encountered in our instrument are discussed in the next chapter.

3.2.2 Fringe formation

Analogy with two classical experiments is found useful in describing the formation of fringes in holographic FTS: Young's double slit experiment provides a good qualitative explanation, and the Newton's rings experiment gives a basis for quantitative considerations.

Young's double slit. Consider, as in Figure 3.2(a), the Michelson interferometer illuminated by a point source. An observer looking into the instrument sees two coherent images of the source, slightly displaced from each other since one of the mirrors is tilted. The setup is therefore equivalent to that of Young's double slit experiment, and fringes are formed on the observers retina with fringe frequency proportional to the optical frequency of the light and the separation between the source images.

When the tilted mirror is replaced with a grating (Figure 3.2(b)), one of the

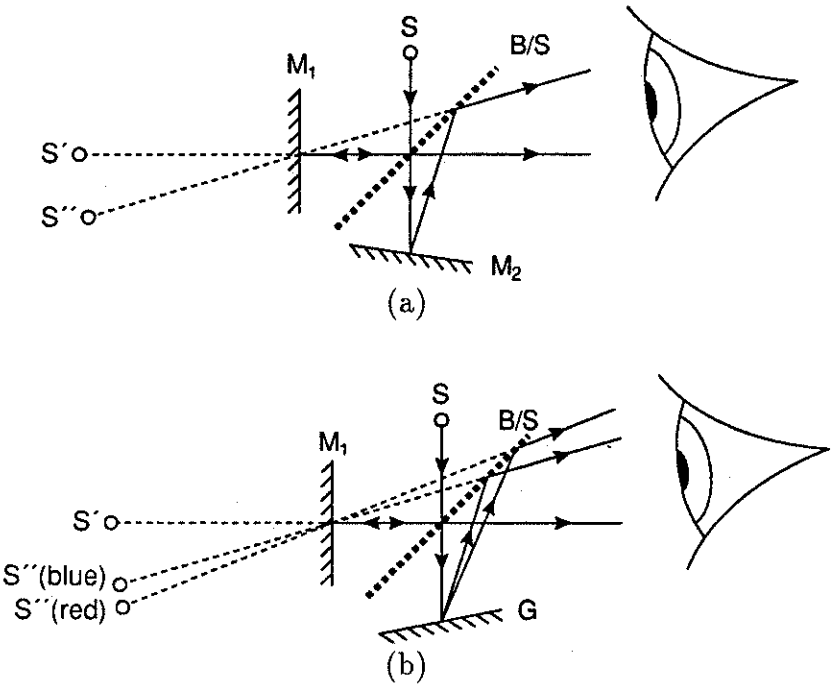


FIGURE 3.2: Qualitative understanding of the fringe formation in a Michelson interferometer with one of its mirrors (M_2) tilted may be obtained by comparison with the Young's double slit experiment (a). The analogy is also helpful to understand the effect of replacing the tilted mirror by a grating (G) as shown in (b). M_1 = fixed mirror, M_2 = tilted, adjustable mirror, G = tilted, adjustable grating, $B?S$ = beam splitter, S = source, S' = image of S in M_1, and S'' = image of S in M_2 or G.

source images takes on the shape of a spectrum. This means that each spectral component of the light from the source is represented by a separate image, and each image has a different separation from the reference image reflected from the mirror. While the fringe frequencies are still proportional to optical frequency and image separation, the relationship is now more complicated since the separation has itself become dependent upon optical frequency. As will be seen shortly, the spatial frequency of a fringe pattern is now proportional to the difference between its optical frequency and the Littrow frequency of the grating.

Newton's rings. If our observer instead of focussing on the source focuses inside the interferometer, he will notice that the two mirrors (or the mirror and the grating in the heterodyned situation) are effectively superimposed so as to

form a wedge-shaped air gap. A similar effect is seen when two irregular glass plates are brought close together, such as the cover plates of a photographic transparency. The colourful pattern often observed during slide shows known as 'Newton's rings' is exactly analogous to the interference pattern upon which our instrument is based.

Like Newton's rings, the fringes formed in our interferometer are localized near the air gap, so in order to measure them, they must be imaged onto a light sensitive surface such as a diode array. Characteristics of the fringe pattern are predicted by considering the two interfering wave fronts at the detector surface where an image of the wedge is formed as illustrated in Figure 3.3. The phase difference $\Delta\delta$ between the wave fronts is given by the wedge angle

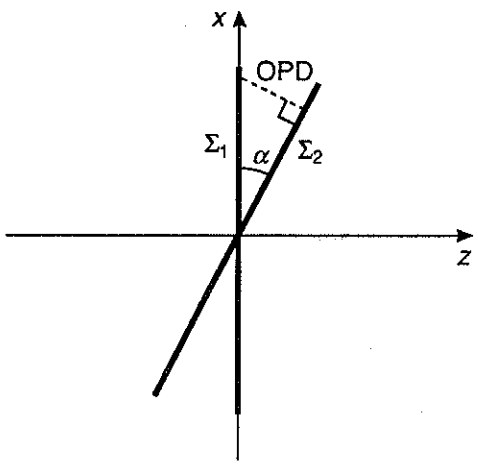


FIGURE 3.3: The wave fronts Σ_1 and Σ_2 returning from the interferometer form a wedge when superimposed. The wedge is shown here imaged onto the detector (represented by the x-axis). The dotted line joining the two wave fronts traces out the extra optical path travelled by Σ_2 compared with Σ_1.

α:

$$\Delta\delta = 2\pi\sigma x\,|\sin\alpha|, \tag{3.7}$$

where x is distance along the mirrors and the *wavenumber* $\sigma = 1/\lambda$ is proportional to the optical frequency of the light. Substituted into Equation 3.6, this gives:

$$\begin{aligned} I &= I_0[1 + k\cos(2\pi\sigma x|\sin\alpha|)] \\ &= I_0(1 + k\cos 2\pi\nu x), \end{aligned} \tag{3.8}$$

where

$$\nu = \sigma \, |\sin \alpha| \tag{3.9}$$

is the spatial frequency of the fringe pattern, referred to as the *fringe frequency.*

By the sampling theorem we require $\nu < \nu_s/2$, where $\nu_s = 1/\Delta x$ is the spatial frequency of the sampling grid and Δx is the separation between samples. Therefore, $|\sin \alpha| < \nu_s/(2\sigma) = 0.01$ at a wavelength of 0.5 μm with 25 μm sample separation, and at such small angles the approximation $\sin \alpha \approx \alpha$ is good to about 2×10^{-7}. We will hence adopt a simplified expression for fringe frequency:

$$\nu = \sigma |\alpha|. \tag{3.10}$$

We also adopt the approximations $\cos \alpha \approx 1$ and $\tan \alpha \approx \alpha$.

The dimensions of ν is m^{-1}, but since we tend to measure x in terms of a number of photodiode elements along the detector array, denoted by the unit "Elements", ν is conveniently measured in the reciprocal unit "Elements^{-1}", denoting "cycles per photodiode element" (see Section 1.3). Note that in this unit the sampling frequency becomes $\nu_s = 1.0$ Elements^{-1} and so, by the sampling theorem, the maximum allowable fringe frequency is 0.5 Elements^{-1}.

3.2.3 Expressions for fringe frequency

In the unheterodyned case the angle α between the interfering wave fronts equals twice the tilt angle of the mirror according to the law of reflection. In the heterodyned situation, α is found by the law of diffraction, described in Figure 2.15. For the present setup it may be written on the form:

$$d[\sin \theta + \sin(\theta - \alpha)] = m\lambda, \tag{3.11}$$

where θ is the angle between the grating normal and the incoming light and $\theta - \alpha$ is the diffraction angle. Manipulating this equation by well known trigonometric relations and using the above stated approximations, we find:

$$\alpha = 2 \tan \theta - \frac{m}{\sigma d \cos \theta}. \tag{3.12}$$

Noting that $1/d$ is the spatial frequency of the grating rulings, we may define the *effective grating frequency* as:

$$\nu_G = m/(d \cos \theta). \tag{3.13}$$

We define furthermore the *unheterodyned fringe frequency* as:

$$\nu_0 = 2\sigma \tan\theta, \tag{3.14}$$

which for small tilt angles becomes equal to the fringe frequency that would be observed were the grating to be replaced by a mirror. Using these definitions in Equation 3.12 and combining it with Equation 3.10 we obtain a general expression for the fringe frequency:

$$\nu = |\nu_0 - \nu_G|. \tag{3.15}$$

This expression is valid both with and without heterodyning since in the unheterodyned case $\nu_G = 0$. For the heterodyned case, it demonstrates the frequency translating or *heterodyning* property of the grating.

A relationship between fringe frequency and wavenumber for the heterodyned instrument is obtained by studying the special case when $\alpha = 0$. This is known as the *Littrow condition* and occurs at a certain wavenumber σ_L, the *Littrow wavenumber*. This condition occurs when the light of wavenumber σ_L is "retro-diffracted" and returned exactly along its incoming path, with no deviation. At this wavenumber $\nu = 0$ by Equation 3.10 and so, by Equation 3.15, $\nu_0 = \nu_G$. Hence, by Equation 3.14:

$$\nu_0 = \nu_G \frac{\sigma}{\sigma_L}. \tag{3.16}$$

Substituted into Equation 3.15 this allows the heterodyned fringe frequency to be expressed as:

$$\nu = \nu_G \frac{|\sigma - \sigma_L|}{\sigma_L}. \tag{3.17}$$

The frequency of a heterodyned fringe pattern is thus proportional to the difference between the wavenumber of the light and the Littrow wavenumber of the grating.

3.2.4 Filtering prescriptions

As pointed out in Section 2.4.3, heterodyned holographic FTS has an ambiguity problem since optical frequencies above and below the Littrow frequency produce identical fringe frequencies. This is manifested by the absolute value

in Equation 3.17. A second ambiguity problem arises due to aliasing: spatial frequencies higher than half the sampling frequency appear as frequencies lower than half the sampling frequency. To avoid ambiguous spectra, these problems must be met with an optical filter inserted into the light path with transmission T_F such that $T_F = 0$ outside the range of either

$$\sigma_L < \sigma < \sigma_L[1 + \nu_s/(2\nu_G)]$$

or

$$\sigma_L[1 - \nu_s/(2\nu_G)] < \sigma < \sigma_L.$$

The two options signify the possibility of measuring the spectrum either above or below the Littrow wavenumber. Note that for unheterodyned instruments, only the aliasing ambiguity exists, requiring therefore only that $T_F = 0$ when $\nu > 0.5$ Elements^{-1}.

3.2.5 White light: the Fourier integral

The 'magic' of Fourier transform spectroscopy appears when light of more than a single frequency is analyzed. The interferogram is then no longer described by the simple sinusoid of Equation 3.8, but as a sum of sinusoids of different frequencies, each representing a spectral component of the incident light. At the centre of the interferogram where $x = 0$, all the sinusoids are ideally in phase and add up to give a high contrast fringe. As one moves towards the edges, the sinusoids become increasingly out of phase resulting in a reduction in visibility or contrast of the fringes: for a wide-band source the contrast is lost sooner than for a narrow-band source.

Mathematically, the summation may be represented by an integral, thus for properly filtered (according to the filtering prescriptions given above) broad-band light the interferogram becomes:

$$\begin{aligned} I(x) &= \int_0^{\nu_s/2} S_\nu(\nu)[1 + k(\nu)\cos 2\pi\nu x]\, d\nu, \\ &= \int_0^{\infty} S_\nu T_F(1 + k\cos 2\pi\nu x)\, d\nu \\ &= I_0 + \int_0^{\infty} S_\nu T_F k\cos 2\pi\nu x\, d\nu, \end{aligned} \tag{3.18}$$

where $S_\nu(\nu)$ is the spectral intensity of each sinusoid, $T_F(\nu)$ represents the transmittance of an appropriate filter and $I_0 = \int_0^\infty S_\nu T_F \, d\nu$ is the total intensity collected. The remaining integral may now be recognized as a reverse Fourier cosine integral* denoted by $\mathcal{F}_{\cos}^{-1}\{\}$:

$$I(x) = I_0 + \mathcal{F}_{\cos}^{-1}\{S_\nu T_F k\}. \tag{3.19}$$

Although the ideal case is well described by the cosine transformation, it will be seen in Section 3.3 that in order to account for certain instrumental effects, it is necessary to use the complex Fourier transformation. Denoted by $\mathcal{F}\{\}$ and $\mathcal{F}^{-1}\{\}$ respectively, the forward and reverse complex Fourier transforms are written as:

$$\mathcal{F}\{\mathbf{f}(x)\} = \int_{-\infty}^{\infty} \mathbf{f}(x) e^{-i2\pi\nu x} \, dx = \mathbf{F}(\nu) \tag{3.20}$$

and

$$\mathcal{F}^{-1}\{\mathbf{F}(\nu)\} = \int_{-\infty}^{\infty} \mathbf{F}(\nu) e^{+i2\pi\nu x} \, d\nu = \mathbf{f}(x), \tag{3.21}$$

where $\mathbf{F}$ is the Fourier transform of $\mathbf{f}$ and $i = \sqrt{-1}$ is the imaginary unit. The complex transformation becomes equivalent to the cosine transformation if and only if both $\mathbf{f}$ and $\mathbf{F}$ are real and even (i.e. symmetrical about the ordinate axis).

The complex transformation uses both positive and negative frequencies. We therefore define a real and even spectral intensity B_ν such that:

$$\left.\begin{aligned} B_\nu(\nu) &= \frac{S_\nu(\nu) T_F(\nu) k(\nu)}{2} \\ B_\nu(-\nu) &= B_\nu(\nu) \end{aligned}\right\} \tag{3.22}$$

Since filter and instrument characteristics are lumped together with the source spectrum, B_ν represents the "coloured" spectrum as seen by the detectors. The interferogram may now be written on the form:

$$I(x) = I_0 + \mathcal{F}^{-1}\{B_\nu(\nu)\}, \tag{3.23}$$

*It is of course equally well recognized as the *forward* cosine transform which is exactly equivalent to the reverse transformation. We prefer the reverse notation because it reflects the usual relationship between a measured signal—here the interferogram—and its Fourier spectrum—here equal to the optical spectrum.

which is equivalent to:

$$B_\nu(\nu) = \mathcal{F}\{I(x) - I_0\} = \mathcal{F}\{\mathcal{I}(x)\}, \tag{3.24}$$

where $\mathcal{I}(x)$ is the 'active' interferogram. I_0 is a constant factor and carries no useful information; we will therefore tend to ignore it and refer to $\mathcal{I}(x)$ as 'the interferogram'.

3.2.6 Finite instrument function and apodization

Although the theoretical interferogram is infinitely long, the practically measurable interferogram has a finite length. Mathematically, this truncation may be represented by the multiplication with a rectangle or 'top hat' function, $\Pi(x/L)$, a rectangular window function of length L centered at the origin as illustrated in Figure 3.4(a):

$$\Pi(x/L) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{for } -L/2 < x \leq L/2 \\ 0 & \text{otherwise,} \end{cases}$$

Hence themeasured (active) interferogram is $\mathcal{I}(x)\Pi(x/L)$.

Modified in this way the interferogram can not yield a true representation of the optical spectrum. Instead its Fourier transform gives a spectral estimate equal to the true spectrum convolved with an *instrument function*, $A(\nu)$, equal to the Fourier transform of the truncation function. This follows from the Fourier convolution theorem:

$$\mathcal{F}\{g(x)h(x)\} = \mathcal{F}\{g(x)\} \star \mathcal{F}\{h(x)\}, \tag{3.25}$$

where $\star$ denotes convolution. The instrument function is therefore:

$$A(\nu) = \mathcal{F}\{\Pi(x/L)\} = L \operatorname{sinc} \nu L, \tag{3.26}$$

where the function $\operatorname{sinc} u \stackrel{\text{def}}{=} \sin(\pi u)/(\pi u)$, see Figure 3.4(b). The characteristic feature of this function is its undulating 'wings' with zero crossings at $\nu = n/L$ with $|n| = 1, 2, 3, \cdots$. This causes interactions between spectral components whose separation is different from n/L and so certain knowledge can only be had of components separated by

$$\Delta\nu = 1/L, \tag{3.27}$$

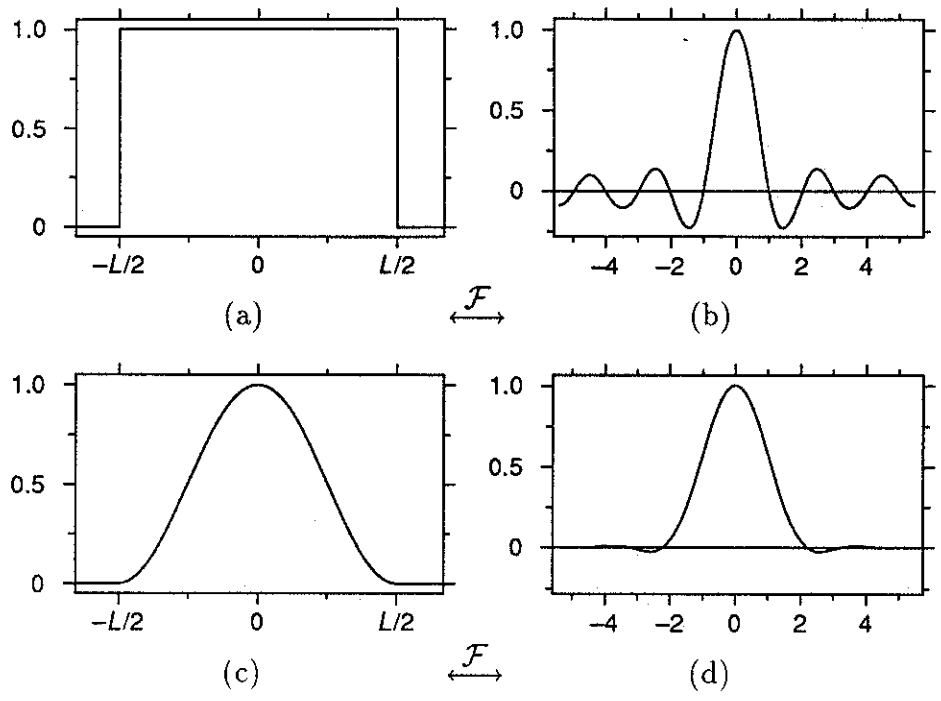


FIGURE 3.4: Normal (a) and apodized (c) truncation functions with spectral instrument functions, (b) and (c) respectively. (b) is the sinc function. The unit in the spectral domain equals the separation between independent spectral samples, $\Delta\nu = 1/L$.

the distance between independent spectral samples.

For visual inspection, FTS spectra tend to be confusing since the wings of the sinc function may appear like weak spectral features. To remedy this, interferograms are often *apodized* before transformation. This involves using a modified truncation function which smoothly tapers off the interferogram towards the edges. Note that the apodized instrument function is always broader than the unapodized one and that its zero crossings are in general no longer regularly spaced. The orthogonal properties of FTS spectra are therefore to a greater or lesser extent lost by apodization [44].

One much used apodization function is the Hann window, also called the "cosine bell", which is simply a cycle of the cosine function raised to the axis, see Figure 3.4(c). Apart from its simple functional form and high degree of "smoothness" (discontinuities occur first in the second derivative), the interest of this apodization function lies in the good orthogonal properties of the resulting instrument function.

3.2.7 Resolving power

Spectroscopic instruments are often specified in terms of their resolving power:

$$\mathcal{R} \stackrel{\text{def}}{=} \frac{\lambda}{\Delta\lambda} = \frac{\sigma}{\Delta\sigma}, \tag{3.28}$$

where $\Delta\lambda$ or $\Delta\sigma$ usually designates the spectral distance between two 'just resolved' spectral features [6, page 333] but for FTS instruments commonly is taken to represent the distance between independent spectral samples. For HFTS instruments this distance is given in terms of spatial frequencies by Equation 3.27. Differentiating Equation 3.15 gives a relationship between small increments in ν and σ, viz:

$$\frac{d\nu}{d\sigma} = \frac{d\nu_0}{d\sigma} = 2\tan\theta = \frac{\nu_0}{\sigma}, \tag{3.29}$$

and hence Equation 3.28 may be written as:

$$\mathcal{R} = \frac{\nu_0}{\Delta\nu} = \nu_0 L. \tag{3.30}$$

In the unheterodyned mode where $\nu = \nu_0$, $\mathcal{R}$ varies through the spectrum with its maximum value at the maximum fringe frequency, $\nu_{\text{Max}} = \nu_s/2 = 1/(2\Delta x)$:

$$R_{\text{Max}} = \frac{\nu_s L}{2} = \frac{N_D}{2}, \tag{3.31}$$

where N_D is the number of detector elements in the array of length L.

In the heterodyned case, by Equation 3.15:

$$R = (\nu_G \pm \nu)L \approx \nu_G L \tag{3.32}$$

when the effective grating frequency is much greater than the sampling frequency. By the definition of the effective grating frequency (Equation 3.13), letting N_G denote number of grating rulings:

$$\nu_G L = \frac{m}{d\cos\theta} L = mN_G. \tag{3.33}$$

Note that this expression equals the resolving power $\mathcal{R}_G$ of the grating when used in a classical rating spectrometer [5, page 127].

3.2.8 Sampling and discreteness

Since the interferogram is in our system measured by means of a detector array, its value is only known at discrete points corresponding to each of the detectors in the array. This discreteness is usually referred to as *sampling*, and is represented mathematically by multiplication with a *comb function*, i.e. a regular series of infinitely sharp spikes or *delta functions*. This is not the whole story, however. Each detector element has a certain width over which the interferogram is averaged as illustrated in Figure 3.5. Clearly, this has the effect of low-pass filtering the interferogram, reducing the strength of

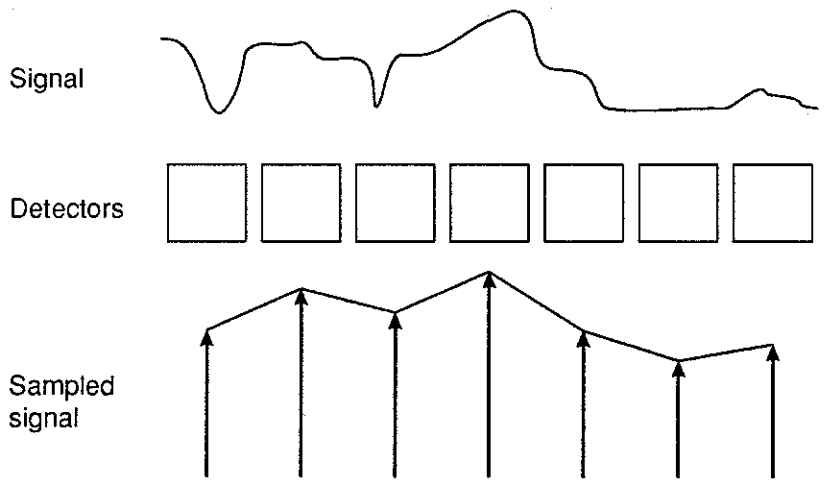


FIGURE 3.5: Sampling of a the continuous signal shown at the top of the figure by an array of detectors yields a signal which is discrete (one 'sample' per detector), but also low-pass filtered since each detector averages a certain length of the signal.

high-frequency components.

The effect of finite detector width will be discussed further in the next chapter, but for now we will ignore it and simply represent the sampling by a multiplication with a comb function, defined as:

$$\text{Ш}(u) \stackrel{\text{def}}{=} \sum_{n=-\infty}^{\infty} \delta(u-n), \tag{3.34}$$

where n is a whole number and $\delta(u)$ is the Dirac delta function, a spike located at the origin with infinitesimal width and infinite height so that $\int_{-\infty}^{\infty} \delta(u)\,du = 1$. Ignoring the finite length of the measurement, the sampling may be ex-

pressed as:

$$\sum_{n=-\infty}^{\infty} \delta(x - n\,\Delta x) = \frac{1}{\Delta x} \sum_{n=-\infty}^{\infty} \delta\left(\frac{x}{\Delta x} - n\right), \tag{3.35}$$

since[†] $\delta(ax) = \delta(x)/a$. Seeing that the reciprocal of the sample separation equals the sampling frequency ($1/\Delta x = \nu_s$) and using the definition of the comb function (Equation 3.34), Equation 3.35 may be rewritten as:

$$\nu_s \sum_{n=-\infty}^{\infty} \delta(x\nu_s - n) = \nu_s \text{Ш}(x\nu_s). \tag{3.36}$$

Fourier transforming a comb with unit spike separation gives another comb with unit spike separation: $\mathcal{F}\{\text{Ш}(x)\} = \text{Ш}(\nu)$. Applying the Fourier similarity theorem therefore gives:

$$\mathcal{F}\{\nu_s \text{Ш}(x\nu_s)\} = \text{Ш}(\nu/\nu_s), \tag{3.37}$$

a comb with spike separation equal to the sampling frequency. By the Fourier convolution theorem, the spectral estimate found from an interferogram sampled (i.e. *multiplied*) with Equation 3.36 equals the actual spectrum *convolved* with the comb of Equation 3.37. The spectral estimate is therefore infinitely repeated along the frequency axis as shown in Figure 3.6.

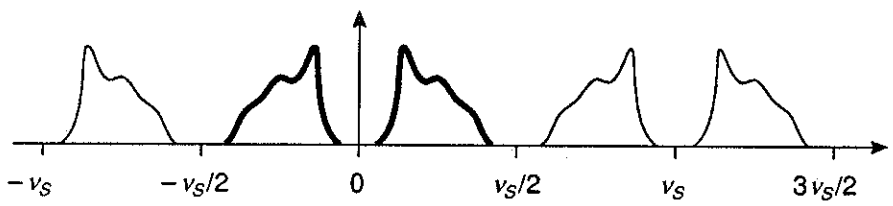


FIGURE 3.6: Sampling of the interferogram causes the spectrum to be infinitely repeated along the frequency axis. If insufficient optical filtering is employed this may cause a problem of "leakage" of spectral information into neighbouring copies or *aliases*. The leaked information adds into a different part of the spectrum producing the effect known as *aliasing*.

Sampling in the interferogram domain is seen to limit the *extent* of the spectrum. By inverting this argument we may expect a limitation of the interferogram length to correspond to a sampled spectrum. Since the Fourier transformation is defined only for infinitely extended signals (see Equation 3.20

[†] $\int \delta(ax)\,dx = \int \delta(u)\,du/a$ by the change of variables $u = ax$. Using x instead of u on the right hand side and removing the integral signs on both sides give the required relationship.

and Equation 3.21), its computational implementations of the transformation (such as the fast Fourier transform, FFT), which pretend to transform finite length signals, assume in fact that the signal is repeated *ad infinitum*. The finite interferogram is therefore convolved with a comb of period L which, when applying the convolution theorem, is equivalent to a multiplication of the spectral estimate with a comb of period $1/L$. Note that this gives a spectral sample separation equal to the distance between the zero crossings in the sinc function, hence each sample represents an independent spectral point.

3.3 Phase correction

In the ideal instrument the optical path difference (OPD) is the same for all wavelengths. Generally this is not true in real instruments, mainly due to frequency dependent refractive index variations (dispersion) of optical materials used in the interferometer. When the OPD varies from one spectral component to another, their sinusoidal interference patterns no longer have a common origin and so the total interferogram is no longer symmetric. Hence the Fourier transform is no longer real but is accompanied by a *phase*. Obtaining an acceptable, real spectral estimate from this complex function is one of the most important tasks of FTS signal processing procedures known as 'phase correction'.

Variation in OPD due to dispersion is one of several sources of spectral phase. We start off this discussion with a presentation of the main phase mechanisms encountered in our instrument. We then present the general theory of phase correction and estimate limitations in view of the present application. We also present the concept of the 'single sided' measurement and consider the extent to which it may be applicable with our instrument.

3.3.1 Dispersive phase

When light is transmitted through optically dense media, i.e. media with refractive index greater than one, the 'optical path length' (l) is no longer equal to the physical path length (s) travelled by the light. If the material traversed has a constant refractive index (n) throughout the material, then $l = ns$. Since

optical materials in general are dispersive, n, and hence ξ, are wavelength-dependent.

Figure 3.7(a) shows a simple Michelson interferometer based on a beam

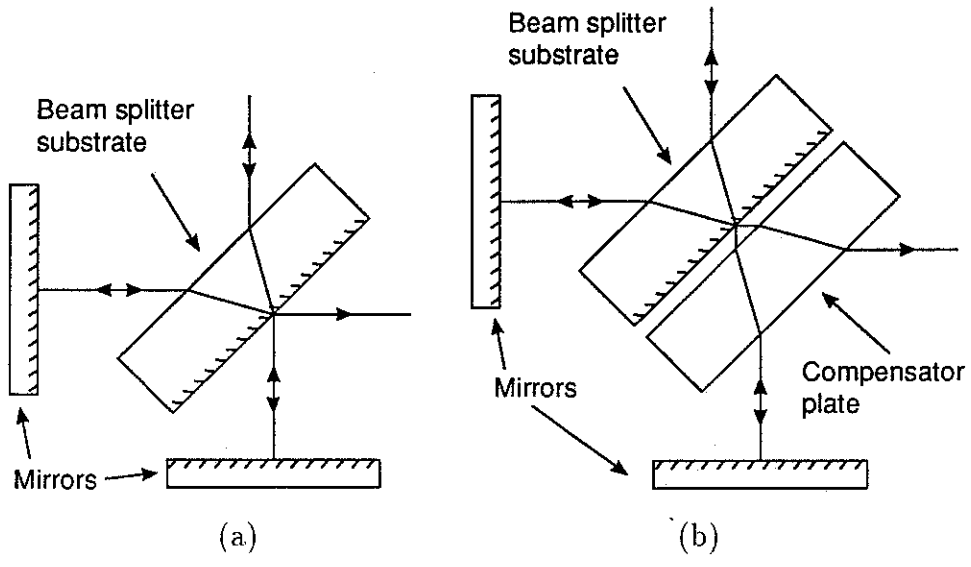


FIGURE 3.7: Uncompensated (a) and compensated (b) Michelson interferometers.

splitter supported by a substrate. The incoming light is divided in two and recombined after reflection off the mirrors in each arm. Since one of the recombined beams has passed twice through the beam splitter substrate its optical path length is different from that traversed by the other beam. By adjusting the mirrors, the OPD may be reduced to zero at one wavelength, but because of the dispersion in the substrate, other wavelengths will still have a finite OPD.

Theoretically, the effects of dispersion may be removed by ensuring that the two interferometer arms contain exactly equal amounts of dense materials as shown in Figure 3.7(b). The interferometer is then said to be *compensated.* In practice, we can only achieve a certain degree of compensation depending upon the accuracy with which the substrate and its compensator are matched. As will be described in the following chapter we have not achieved complete compensation in our instrument. This has resulted in a curved phase function which, as is shown later in the present section, tightens the tolerances for phase correction.

3.3.2 Misalignment phase

Another source of phase is the misalignment of interferogram and sampling grid. The mathematical Fourier transformation process assumes an origin which ideally should coincide with the point of zero OPD in the interferogram. When digital transformation is used, the origin must necessarily be chosen at one of the sampling points however, and a difference in the position of zero OPD and the closest point in the sampling grid shows up as spectral phase. If the difference is δx then the phase is $\phi = 2\pi\nu\,\delta x$; grid misalignment is therefore characterized by a phase function which is linear in frequency.

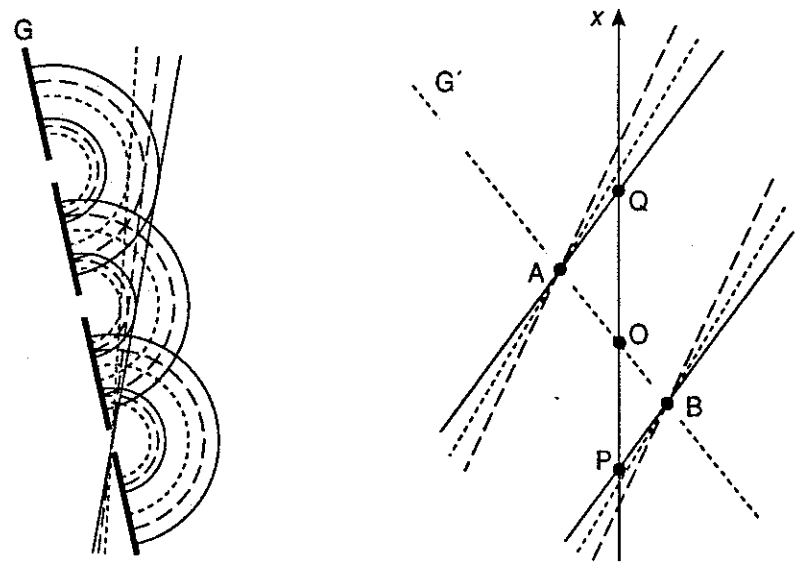
3.3.3 Grating phase

Heterodyned holographic FTS has a particular phase phenomenon associated with it due to the position of the grating rulings with respect to the interferogram. This may be understood by considering the grating as an absorbing surface with narrow reflecting lines as apertures. According to Huygens' principle, wavefronts emerging from this surface may be constructed by the summation of spherical wavelets emerging from the apertures. Figure 3.8(a) shows this construction for three different optical frequencies: clearly, the wavefronts intersect only at the apertures of the grating. The propagating wave train is therefore characterized by localized 'hubs' separated by the grating constant as seen in Figure 3.8(b). When brought to interfere with the reference wavefront, the positions of zero OPD coincide for all frequency components only when the reference wavefront coincides with a hub in the diffracted wave train.

Considering one of the frequency components, the figure denotes two neighbouring wavefront intersection points as P and Q, and the two closest hubs by A and B. The intersection between the reference wavefront and the line joining the hubs is called O. If the detecting surface coincides with the reference wavefront, we may chose O to represent the origin of the detected interferogram and the phase of the fringe pattern with respect to the origin is then:

$$\phi = 2\pi\frac{PO}{PQ}.$$

Since the hubs represent the grating apertures the line joining them is an image



(a) (b)

FIGURE 3.8: Wavefronts of three different colours (represented by continuous, broken, and dotted lines) emerging from a grating (G) are constructed by Huygens' principle, showing that they all coincide at grating apertures (a). In (b) wavefronts are imaged onto the detector array where they interfere with the reference wavefront (represented by the x-axis). The letters correspond to an argument in the text.

of the grating. We may therefore consider the grating to have a phase with respect to O given by:

$$\phi_G = 2\pi \frac{AO}{AB}.$$

Now, since triangles AOP and BOQ are similar:

$$\frac{AO}{AB} = \frac{PO}{PQ},$$

and so $\phi = \phi_G$. The phase due to this effect is therefore equal to the phase of the grating with respect to the interferogram and so constant across the spectrum.

In our system this phase phenomenon may be observed as a continuous change between symmetric and asymmetric interferogram shapes as the interferometer controls are adjusted. Although it is unique to HHS systems, it is similar to that found in deliberately aliased classical FTS systems [43]. As will become apparent in the following, constant phase terms are of no importance for the quality of phase correction; the effect is therefore not serious.

3.3.4 "Channelled phase"

An unexpected, sinusoidal variation has also been observed in the phase curve. This effect, presumably due to interference between spurious reflections at the outer surfaces of the beam splitter, will be discussed in Section 4.3.

3.3.5 The complex spectrum

Phase is introduced in the Fourier transform whenever there is a mismatch between the origin of the transformation and the position of zero OPD for the sinusoidal interferogram components. Denoting this mismatch by δx, we may rewrite Equation 3.8 on the form:

$$\begin{aligned} I_\nu &= I_0[1 + k\cos 2\pi\nu(x + \delta x)] \\ &= I_0[1 + k\cos(2\pi\nu x + \phi)], \end{aligned} \tag{3.38}$$

where $\phi = 2\pi\nu\,\delta x$ is the spectral phase of the component of spatial frequency ν. The broad-band interferogram, given by integrating Equation 3.38 over all spatial frequencies, now no longer represents an even function and its equivalence with the Fourier cosine transform is therefore lost. Use of the complex transformation is imperative for such interferograms. Letting $\phi(\nu)$ be an odd function such that $\phi(-\nu) = -\phi(\nu)$, the broad-band interferogram may be written as:

$$\begin{aligned} I(x) &= I_0 + \int_{-\infty}^{\infty} B_\nu \cos(2\pi\nu x + \phi)\, d\nu \\ &= I_0 + \int_{-\infty}^{\infty} B_\nu e^{i(2\pi\nu x + \phi)}\, d\nu \\ &= I_0 + \int_{-\infty}^{\infty} B_\nu e^{i\phi} e^{i2\pi\nu x}\, d\nu \\ &= I_0 + \mathcal{F}^{-1}\left\{B_\nu e^{i\phi}\right\} d\nu, \end{aligned} \tag{3.39}$$

By rearranging and transforming both sides we then find that:

$$B_\nu e^{i\phi} = \mathcal{F}\left\{I(x) - I_0\right\} = \mathcal{F}\left\{\mathcal{I}(x)\right\}. \tag{3.40}$$

The transformed interferogram therefore gives a complex function whose modulus is the spectrum: $B_\nu = |\mathcal{F}\left\{\mathcal{I}(x)\right\}|$. In practice, taking the modulus is not a good method by which to calculate the spectrum since when the spectrum

is noisy it causes low spectral values to fluctuate about the RMS level of the noise rather than about the actual spectral value. A better spectral estimate may be obtained if the phase is known: a *phase corrected* spectrum is then found as the real part of the complex spectrum multiplied with the conjugate of the phase function:

$$B_\nu = \Re\left[\mathcal{F}\left\{\mathcal{I}(x)\right\} e^{-i\phi}\right]. \tag{3.41}$$

The imaginary part of this complex product is zero apart from in cases when the spectrum is noisy, then it contains half the noise power, i.e. $\sqrt{2}$ of the total noise amplitude. This part of the noise is discarded and the noise left in the spectral estimate is correspondingly reduced. The advantage of phase correction over taking the modulus is therefore two-fold:

- it gives correct spectral level even at low signals, and
- it reduces spectral noise by a factor $\sqrt{2}$.

3.3.6 Phase estimation

It is tempting to take the argument of the complex spectrum directly as the phase function. Although this is indeed an estimate of the phase and the basis upon which the phase is usually determined, it will not do to use it directly since Equation 3.41 is then exactly equivalent to the modulus of the complex spectrum. Instead, a noise-less "fiducial" phase estimate must be produced. Two procedures are frequently used to extract the fiducial phase: "hard" apodization, and curve fitting. By the former method, the central part of the interferogram is isolated by the use of a narrow (hard) apodization function. Fourier transformed, this gives a low-noise spectrum with a resolution which is much lower than the original spectrum but high enough to resolve the phase which is usually a slowly varying function [37]. By the second method, the interferogram is also apodized although not necessarily as strongly as in the first method. The resulting high-resolution and noisy phase curve is then fitted either to a general function or to a specific function which is known to describe the phase variation well, the latter being preferable since it allows interpolation into spectral regions where the signal is low.

As will be seen in the following, the problem of phase correction becomes somewhat more complicated when the effect of truncation is considered. We find then that the argument of the transformed interferogram only is an approximation to the actual spectral phase, and that its error is inversely proportional to the square of the length of interferogram used in the phase estimation. This fact favours a "soft" rather than a hard apodization and hence the second rather than the first method of phase estimation.

3.3.7 The effect of truncation

The preceding phase considerations are strictly correct only if the interferogram is infinitely long. Truncating the interferogram complicates the situation because the phase shifted sinusoids no longer 'fit' within the truncation window as illustrated in Figure 3.9—it is as if a guitar string is expected to vibrate

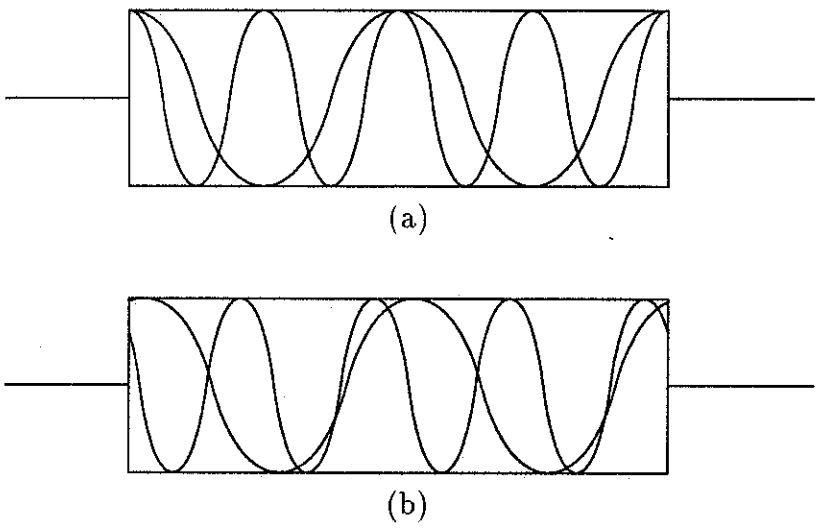


FIGURE 3.9: Two sinusoidal components of an interferogram without (a) and with (b) spectral phase, bounded by a rectangular truncation function.

with nodes which do not coincide with its fixed points. Denoting the truncation function by $\mathcal{T}(x)$, the measured (active) interferogram is:

$$\mathcal{I}(x)\mathcal{T}(x), \tag{3.42}$$

and its transform is the complex function $\mathbf{C}$. By the Fourier convolution theorem:

$$\mathbf{C} = Ce^{i\Phi} = B_\nu e^{i\phi} \star \mathbf{t}, \tag{3.43}$$

where $\mathbf{t}$ is the transform of $\mathcal{T}$. If $\mathcal{T}$ is not restricted to be symmetric about the origin of the transformation then $\mathbf{t}$ is complex and so $\mathbf{C}$ is a convolution between the complex spectrum and a complex instrument function. This is notably the case for single sided interferograms as will be discussed in Section 3.3.9 but a similar situation occurs with double sided interferograms affected by a nonlinear phase function. Our phase correction procedure must therefore not only remove the spectral phase, but also ensure the spectral estimate to be a convolution between the actual spectrum and an 'acceptable', real instrument function. For unapodized spectra the acceptability criterion requires the instrument function to be a good approximation to a sinc function, the value of 'good' depending on the instrument's performance criteria.

Writing out in full the convolution integral of Equation 3.43 for a frequency ν_0 gives:

$$\mathbf{C}(\nu_0) = \int_{-\infty}^{\infty} B_\nu(\nu_0 - \nu) e^{i\phi(\nu_0 - \nu)} \mathbf{t}(\nu)\, d\nu \tag{3.44}$$

Since the phase is slowly varying it may be written as a Taylor expansion of the form [38]:

$$\begin{aligned} \phi(\nu_0 - \nu) &= \phi(\nu_0) - \nu\phi'(\nu_0) + \frac{\nu^2\phi''(\nu_0)}{2} - \cdots \\ &= \phi(\nu_0) - 2\pi\nu\delta, \end{aligned} \tag{3.45}$$

where δ represents the variation of the phase. Note that δ has the dimensions of distance and that it in the simple case of linear phase signifies the position error δx of the interferogram on the sampling grid. In the general case of a slowly varying phase and a narrow instrument function, δ is approximately proportional to the local slope of the phase curve: $\delta \approx \phi'/(2\pi)$. The convolution integral now becomes:

$$\begin{aligned} \mathbf{C}(\nu_0) &= \int_{-\infty}^{\infty} B_\nu(\nu_0 - \nu) e^{i[\phi(\nu_0) - 2\pi\nu\delta]} \mathbf{t}(\nu)\, d\nu \\ &= e^{i\phi(\nu_0)} \int_{-\infty}^{\infty} B_\nu(\nu_0 - \nu) \mathbf{t}(\nu) e^{-i2\pi\nu\delta}\, d\nu \\ &= e^{i\phi(\nu_0)} \left[B_\nu(\nu_0) \star \mathbf{t}(\nu) e^{-i2\pi\nu\delta} \right] \\ &= e^{i\phi(\nu_0)} [B_\nu(\nu_0) \star \mathbf{A}(\nu)], \end{aligned} \tag{3.46}$$

where:

$$\mathbf{A}(\nu) = \mathbf{t}(\nu) e^{-i2\pi\nu\delta} \tag{3.47}$$

is a complex instrument function for the real spectrum. Knowing the phase function therefore allows phase correction as suggested in Equation 3.41, giving as a real spectral estimate:

$$B_R = \Re\left(\mathbf{C}e^{-i\phi}\right) = B_\nu \star A_R, \tag{3.48}$$

where $A_R = \Re(\mathbf{A})$. B_R is an acceptable estimate of B_ν as long as A_R is an acceptable instrument function.

The imaginary part of the phase corrected spectrum is:

$$B_I = \Im\left(\mathbf{C}e^{-i\phi}\right) = B_\nu \star A_I, \tag{3.49}$$

where $A_I = \Im(\mathbf{A})$. This "imaginary spectrum" becomes important when ϕ is not precisely known, in which case a certain fraction of the imaginary spectrum is added to the real spectrum. Denoting the error in ϕ by ϵ, the spectral estimate becomes:

$$\begin{aligned} B_E &= \Re\left(\mathbf{C}e^{-i(\phi+\epsilon)}\right) \\ &= B_\nu \star (A_R \cos\epsilon + A_I \sin\epsilon) \\ &\approx B_R + \epsilon B_I, \end{aligned} \tag{3.50}$$

where the approximation is valid when ϵ is small enough that its cosine approximates unity and its sine approximates itself. B_E therefore has an error ΔB approximately equal to the imaginary spectrum multiplied with the phase error:

$$\Delta B \approx \epsilon B_I. \tag{3.51}$$

3.3.8 Accuracy of phase estimation

As mentioned earlier, the basis of our phase estimate is the phase of the complex Fourier transform of the interferogram, $\mathbf{C}$. Accepting this phase function, $\Phi = \mathrm{Arg}\,(\mathbf{C})$, only to be an approximation of the actual phase function ϕ, we may write: $\Phi = \phi + \epsilon$, where ϵ is the phase error. Hence:

$$\begin{aligned} \epsilon = \Phi - \phi &= \mathrm{Arg}\left(\mathbf{C}e^{-i\phi}\right) \\ &= \tan^{-1}\left[\frac{\Im\left(\mathbf{C}e^{-i\phi}\right)}{\Re\left(\mathbf{C}e^{-i\phi}\right)}\right] \end{aligned} \tag{3.52}$$

$$= \tan^{-1}\left(\frac{B_I}{B_R}\right),$$

i.e. approximately proportional to the imaginary spectrum. Since the imaginary spectrum is minimized by using a symmetrical truncation function, such truncation should therefore allways be used for the phase estimation measurement.

If a source with a slowly varying spectrum (e.g. a blackbody) is used for the phase estimation measurement then the spectrum may be well described by a Taylor expansion similar to that of Equation 3.45. The convolution integral of Equation 3.44 may then be written as:

$$\begin{aligned}
\mathrm{C} &= \int_{-\infty}^{\infty} B(\nu_0-\nu)e^{i\phi(\nu_0-\nu)}t(\nu)\,d\nu \\
&= B_\nu e^{i\phi}\int_{-\infty}^{\infty} e^{-i(\nu\phi'-\nu^2\phi''/2+\cdots)}t(\nu)\,d\nu \\
&- B'_\nu e^{i\phi}\int_{-\infty}^{\infty} \nu e^{-i(\nu\phi'-\nu^2\phi''/2+\cdots)}t(\nu)\,d\nu \\
&+ \frac{B''_\nu e^{i\phi}}{2}\int_{-\infty}^{\infty} \nu^2 e^{-i(\nu\phi'-\nu^2\phi''/2+\cdots)}t(\nu)\,d\nu \\
&- \cdots
\end{aligned}$$

where $t(\nu)$ is the (real) Fourier transform of the symmetrical truncation function. If $t(\nu)$ is sufficiently narrow, then ν is small enough within the nonzero range of $t(\nu)$ that orders in ν higher than the second may be ignored. With this approximation the expression may be written as a sum of a series of integrals many of which disappear because their integrands are odd—remembering that $t(\nu)$ is even. Inserting what is left into Equation 3.52 it may be written as [45]:

$$\epsilon \approx \left(\frac{B'}{B}\,\phi' + \frac{\phi''}{2}\right)\frac{\int_{-\infty}^{\infty}\nu^2 t(\nu)\,d\nu}{\int_{-\infty}^{\infty} t(\nu)\,d\nu}. \tag{3.53}$$

The first term of this expression shows how the phase error depends upon the *shape* of the phase curve. In parts of the spectrum where spectral intensity changes rapidly—typically at optical filter edges—the spectral derivative is considerable and so the first derivative of the phase dominates the error. Where the spectral value is essentially constant, it is the second phase derivative which has the greatest effect.

The second term may be recognized as the normalized second moment (variance) of the instrument function. For phase estimation a quickly diminishing instrument function is therefore important; this is achieved by the use

of a smooth apodization function. Note also that by the change of variable $u = \nu/a$, where a is the full-with at half maximum (FWHM) of the instrument function, the integrals may may be written as:

$$\frac{\int_{-\infty}^{\infty} \nu^2 t(\nu)\, d\nu}{\int_{-\infty}^{\infty} t(\nu)\, d\nu} = a^2 \frac{\int_{-\infty}^{\infty} u^2 t_N(u)\, du}{\int_{-\infty}^{\infty} t_N(u)\, du}, \tag{3.54}$$

where t_N is a 'normalized width' version of t: $t_N(u) = t(\nu)$. The variance is therefore proportional to the square width of the instrument function.

To estimate the variance consider the particularly smooth apodization provided by a Gaussian function, see Figure 3.10. Since the Fourier transform of

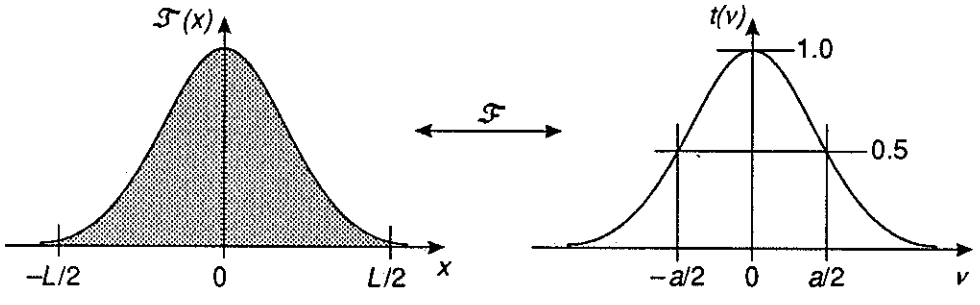


FIGURE 3.10: The Fourier transform of a Gauss function is itself a Gaussian function. If 94% of the area under the curve in the spatial domain is contained within a length L as shown, then the reciprocal of $L/2$ equals the FWHM, a, of the curve in the frequency domain: $a = 2/L$.

a Gaussian is another Gaussian, this is also the functional shape of the instrument function:[‡]

$$\mathcal{T}(x) = e^{-\pi x^2} \Leftrightarrow t(\nu) = e^{-\pi \nu^2},$$

so that the FWHM of $t(\nu)$ is $a = 2\sqrt{(\ln 2)/\pi}$. It is useful to note that $\mathcal{T}(1/a) = 0.028$ and that the interval $-1/a < x < 1/a$ therefore contains almost all (94%) of the area under $\mathcal{T}$. Letting $L = 2/a$ represent the length of interferogram used in the phase estimation and performing the integrals of Equation 3.53, we find the variance of our Gaussian instrument function to be $1/(2L^2 \ln 2) \approx 0.72/L^2$.

Calculation of variance for other instrument functions is less straight forward but we will assume that for a smooth apodization function (such as the

[‡]We ignore here for simplicity that the apodizing Gaussian necessarily is truncated causing the instrument function to be somewhat modified.

Hann window) the variance is of the order of $1/L^2$. Hence we estimate the error of the phase estimate to:

$$\epsilon \sim \left(\frac{B'}{B} \phi' + \frac{\phi''}{2} \right) \frac{1}{L^2} \tag{3.55}$$

3.3.9 Single or double sided measurements

An ideal interferogram is, as we have seen, perfectly symmetrical about a strong central peak. There is therefore a redundancy of information and it is sufficient for a complete reconstitution of the spectral information to measure only one of the two sides. Such a measurement is called "single sided" as opposed to a symmetrical measurement where both sides are included, called "double sided". The interest of single sided measurements lie in the prospect of a more efficient use of a given instrument: for classical FTS instruments it represents a halving of the scan length required for a given resolving power, in holographic FTS it represents a potential doubling of the resolving power in a given instrumental configuration.

When the interferogram is not symmetrical, the idea of single sided measurements is no longer evident. It is still possible but, as will be shown in the following, it requires much improved accuracy of the phase estimate for proper phase correction. Also, in order to measure the phase, a certain length of interferogram is needed from the 'other side' of the origin. This reduces the potential gain in resolving power.

Double sided. Double sided interferogram measurements use a symmetrical truncation function which—in the unapodized case—is a top hat function centered at the origin. Its Fourier transform is the real and symmetrical sinc function, see Figure 3.4(b):

$$\mathrm{t} = \mathcal{F}\{\mathcal{T}_{\mathrm{Sym}}\} = \operatorname{sinc} L\nu. \tag{3.56}$$

Due to spectral phase, the instrument function given by Equation 3.47 is still complex however:

$$\mathbf{A}_\nu = e^{-i2\pi\delta\nu} \operatorname{sinc} L\nu. \tag{3.57}$$

Taking real and imaginary parts and using well known trigonometric relations gives the following real and imaginary instrument functions respectively:

$$A_R = \frac{\sin(\pi L\nu - 2\pi\delta\nu) + \sin(\pi L\nu + 2\pi\delta\nu)}{2\pi L\nu} \tag{3.58}$$

and

$$A_I = \frac{\cos(\pi L\nu + 2\pi\delta\nu) - \cos(\pi L\nu - 2\pi\delta\nu)}{2\pi L\nu}. \tag{3.59}$$

When δ is small compared with L, these relationships may to good approximations be written as:

$$A_R \approx \operatorname{sinc} L\nu \tag{3.60}$$

and

$$A_I \approx \frac{2\delta}{L} \operatorname{cosc} L\nu, \tag{3.61}$$

where $\operatorname{cosc} x \stackrel{\text{def}}{=} (1 - \cos \pi x)/(\pi x)$ is the antisymmetric function shown in Figure 3.11.

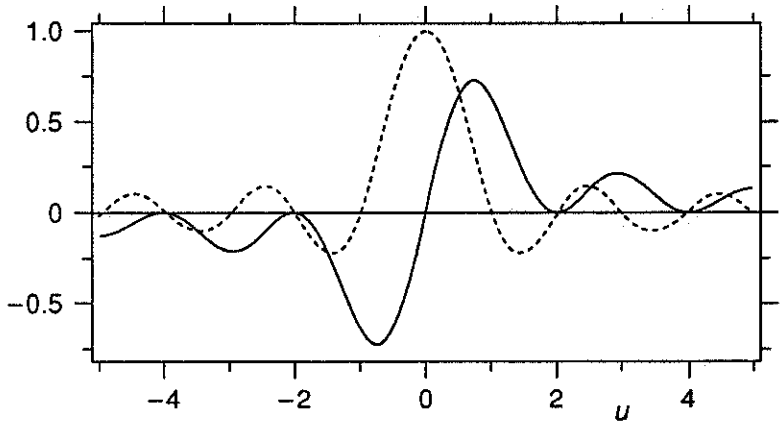


FIGURE 3.11: The cosc function, $\operatorname{cosc} u = (1 - \cos \pi u)/(\pi u)$, plotted as a solid line compared with the sinc function (dotted).

As long as both ϵ and δ are small, the error ΔB in the spectral estimate of a double sided interferogram may now be calculated by Equation 3.51:

$$\Delta B \approx \frac{2\epsilon\delta}{L} (B_\nu \star \operatorname{cosc} L\nu). \tag{3.62}$$

Since the cosc function is asymmetric, its effect is to asymmetrize and shift sharp spectral features. For it to be negligible, ΔB should be smaller than the spectral noise, hence requiring:

$$\frac{2\epsilon\delta}{L} \lesssim \frac{1}{\mathrm{SNR}_\nu} \tag{3.63}$$

where SNR_ν is the spectral signal to noise ratio. Hence the phase error tolerance is:

$$\epsilon \lesssim \frac{L}{2\delta\,\mathrm{SNR}_\nu} \tag{3.64}$$

Combining this phase error tolerance with the estimated phase error of Equation 3.55, using the approximation that $\delta \approx \phi'/(2\pi)$ gives:

$$\left(\frac{B'}{B}\,\phi' + \frac{\phi''}{2}\right)\frac{1}{L^2} \lesssim \frac{L\pi}{\phi'\mathrm{SNR}_\nu} \tag{3.65}$$

which may be expressed as a requirement for interferogram length:

$$L \gtrsim \left[\mathrm{SNR}_\nu\phi'\left(\frac{B'}{B}\,\phi' + \frac{\phi''}{2}\right)\frac{1}{\pi}\right]^{1/3} \tag{3.66}$$

As shown in Figure 4.9 the instrument suffers from a parabolic phase function due to poor dispersion compensation. For an unheterodyned spectrum well centered in the instrument's spectral range the phase has a curvature of $\phi'' \approx 200$ rad Elements^2. If we assume that the spectrum used for phase estimation has a relatively flat central portion where $B'/B \approx 0$ within which $\phi' \lesssim 10$ rad Elements then, for an $\mathrm{SNR}_\nu \approx 1000$, the phase estimation interferogram must at least be of length $L \gtrsim 68$ Elements. Clearly, with a detector array of length 512 Elements this requirement is easily met.

Single sided. For single sided interferograms, the truncation function is no longer symmetric but shifted so as to start at or near the origin. Its transform is found by the Fourier shift theorem, yielding in the unapodized case:

$$\mathrm{t} = \mathcal{F}\left\{\mathcal{T}_{\mathrm{sym}}(x - L/2)\right\} = e^{-i2\pi\nu L/2}\,\mathrm{sinc}\,L\nu. \tag{3.67}$$

By the use of well known trigonometric relations this complex function may be rewritten as:

$$\begin{aligned}
\mathrm{t} &= \frac{\sin\pi L\nu\cos\pi L\nu}{\pi L\nu} - i\,\frac{\sin^2\pi L\nu}{\pi L\nu} \\
&= \frac{\sin 2\pi L\nu}{2\pi L\nu} - i\,\frac{1-\cos 2\pi L\nu}{2\pi L\nu} \\
&= \mathrm{sinc}\,2L\nu - i\,\mathrm{cosc}\,2L\nu.
\end{aligned} \tag{3.68}$$

Its real part is thus a sinc function of half the with of the double sided instrument function, and its complex part is the asymmetric cosc function. The

complex instrument function defined in Equation 3.47 may therefore be expressed in terms of its real and imaginary parts as:

$$\begin{aligned} A_R &= \operatorname{sinc} 2L\nu \cos 2\pi\delta\nu - \operatorname{cosc} 2L\nu \sin 2\pi\delta\nu \\ &\approx \operatorname{sinc} 2L\nu + \frac{\delta}{L} \operatorname{sinc} 2\delta\nu \, (1 - \cos 2\pi L\nu) \end{aligned} \tag{3.69}$$

and:

$$\begin{aligned} A_I &= -\operatorname{sinc} 2L\nu \sin 2\pi\delta\nu - \operatorname{cosc} 2L\nu \cos 2\pi\delta\nu \\ &\approx -\operatorname{cosc} 2L\nu - \frac{\delta}{L} \operatorname{sinc} 2\delta\nu \, \sin 2\pi L\nu \end{aligned} \tag{3.70}$$

respectively, the approximations being valid when $\delta \ll L/2$. Ignoring the additional terms for the moment, these functions are similar to their double sided equivalents given by Equation 3.60 and Equation 3.61. Their widths are halved, however, and the imaginary instrument function is much larger since it has lost its $2\delta/L$ factor. While the former effect causes the desired increase in resolving power, the latter effect causes a greatly increased sensitivity to errors in the phase estimate. Still ignoring the additional terms, the error in the spectral estimate given by Equation 3.51 now becomes:

$$\Delta B \approx \epsilon(B_\nu \star \operatorname{cosc} 2L\nu). \tag{3.71}$$

Hence, by the same criterion as for double sided measurements, the single sided tolerance on ϵ becomes:

$$\epsilon \lesssim \frac{1}{\mathrm{SNR}_\nu} \tag{3.72}$$

For an SNR of 1000, the phase must therefore be determined with an accuracy exceeding 1 milliradian.

Combining this tolerance with the phase error estimate of Equation 3.55 gives the following length tolerance for the double sided section of a single sided measurement:

$$L \gtrsim \left[\mathrm{SNR}_\nu \left(\frac{B'}{B}\phi' + \frac{\phi''}{2}\right)\right]^{1/2} \tag{3.73}$$

which, with the same assumptions as in the double sided case described above gives $L \gtrsim 320$ Elements. It is therefore necessary to include about 160 Elements on the "short" side of the central peak in single sided measurements,

leaving about 350 Elements on the long side. Instead of the ideal doubling of resolving power this only offers an improvement of factor 1.4 compared with double sided measurements. We note however that in the heterodyned modes the phase curvature is smaller, possibly allowing a higher gain in resolving power.

We return now to the additional terms in Equation 3.69 and Equation 3.70: these are found to be significant because of their shapes rather than their size. The extra sinc appearing in the expression for A_R has its first zero crossing at $\nu = 1/(2\delta)$. In the case of a strictly linear phase error, where $\delta \leq 0.5$ Elements always may be ensured by an appropriate choice of origin, this function is therefore wider than the spectral range allowed by the sampling theorem. Convolved with the spectrum, it therefore spreads spectral signal outside the region bounded by the sampling theorem, and, by the phenomenon of aliasing, this signal is folded back into the spectrum and appears as a spurious background.

[See p.85a] A different remedy which is interesting in our case because we only have very short interferograms is to perform the phase correction

as a *convolution* in the interferogram plane rather than as a *multiplication* in the spectral plane [38] by rewriting Equation 3.41 on the form:

$$B_\nu = \Re\left[\mathcal{F}\left\{\mathcal{I}(x) \star \mathcal{F}^{-1}\left(e^{-i\phi}\right)\right\}\right] \tag{3.74}$$

By this method, demonstrated in Figure 4.38, all the sinusoidal interferogram components are brought into phase *before* the Fourier transformation thus avoiding the awkward problem depicted in Figure 3.9. Note that the problem of overlapping aliases does not occur for double sided interferograms; such measurements may therefore still enjoy the simplicity of multiplicative phase correction.

3.4 Noise

In addition to the fundamental limitations to the FTS technique in general and the HFTS technique in particular discussed until now there is a range of other, more or less system dependent effects which limit the quality of spec-

To demonstrate the signifficance of these terms we consider the trivial case of a quasi-continous spectrum where all spectral elements have value B. Including the effects of overlapping aliases, the contributions from the first and second terms of Equation 3.69 to a spectral point are then proportional to the area under their functions. Remembering that the area under the 'unit width' sinc function is $\int_{\infty}^{\infty} \mathrm{sinc}\,(u)du = 1$, this equals for the first term $1/(2L)$. Ignoring the quickly varying cosine factor in the second term, the area under this curve is also $1/(2L)$. Although less populated spectra have a smaller contribution from the error term, this calculation does demonstrate its importance.

Efficient reduction of this effect is achieved by using a trapezeium-shaped truncation function. The short part of the interferogram measured on the 'other' side of the central peak is then taken into account by multiplying the central region with a slope passing through 0.5 at the peak [38]. Further improvement by an additive correction method has also been demonstrated [40].

tral estimates. Some of these such as thermodynamic and photon statistical fluctuations appear as random *noise* in the measurements and set limitations for optimal use of the instrument. Other effects including spatial throughput variations and dark current nonuniformity are fixed from one measurement to another and may be corrected for provided appropriate additional measurements are taken, see Section 4.6. Aberrations in the interferometer causes a different kind of measurement error, affecting the *phase* of the interferogram rather than its *intensity*. They therefore affect spectral estimates in a rather different manner, as will be sen in Section 3.5. We presently consider theoretically the consequences of random noise.

After a presentation of the most prominent sources of noise encountered in our instrument, their balance is discussed in the view of achieving optimal noise performance. The effect of measurement noise on the spectral estimate is then considered, leading to a comparison between the noise performance of FTS instruments and classical grating instruments.

3.4.1 Sources of random noise

In holographic FTS the main sources of random noise are:

1. *Shot noise* (ϵ_S): statistical uncertainty in the signal level due to discreteness of photo-electric events,

2. *Resetting noise* (ϵ_R): thermodynamic uncertainty in resetting a detector cell after readout,

3. *Digital noise* (ϵ_D): error in the digitized signal due to the discrete nature of digital numbers,

4. *Amplifier noise*: random signal fluctuations produced in the analogue circuitry,

5. *Flatness error*: spatial variations in optical and electronic gain, and

6. *Dark noise*: spatial and temporal variations in the level of dark current,

where ϵ_i represent the RMS noise of a system dominated by noise source i. As will be demonstrated in Section 4.6, flatness error and dark noise are under

most circumstances well eliminated by the use of auxiliary measurements. Assuming also that the amplifier noise is sufficiently reduced by good electronic design, the system is dominated by either of the three first noise sources. All of these are assumed to be 'white', i.e. equally affecting all frequency components of the measured signal.

Shot noise is inherent to the signal, representing the minimum amount of noise associated with a certain signal. In optimizing noise performance we therefore want to make Shot noise the limiting noise source. It shares with resetting noise the property of being indeterminate, i.e. to vary randomly from one measurement to another even if the signal to be measured stays constant. Averaging many measurements therefore reduces the noise by a factor equal to the square root of the number of measurements. Digital noise on the other hand is determinate: a certain signal level produces a certain, calculable error in the digitized signal. Averaging has no effect on a signal dominated by digital noise.

3.4.2 Quantification of noise

In the evaluation of the noise from these three sources, we will assume an ideally symmetric interferogram with optimal contrast, see Figure 3.12. This facilitates the treatment because the signal level at the well-defined, centrally located peak of the interferogram then equals twice the background level: $I_P = I(0) = 2I_0$. Hence the peak of the *active* interferogram is $\mathcal{I}(0) = I(0) - I_0 = I_0$. Letting ϵ_x be the total RMS noise of the interferogram, we define the signal-to-noise ratio of the interferogram as $\mathrm{SNR}_x = \mathcal{I}(0)/\epsilon_x = I_P/(2\epsilon_x)$. Similarly, we define the *component* signal-to-noise ratio as $\mathrm{SNR}_i = \mathcal{I}(0)/\epsilon_i$ which is used to describe the noise contributions from the individual noise sources listed above. Since these sources are independent from each other, their contributions to the total noise adds in a root-square fashion, i.e. $\epsilon_x = \sqrt{\sum_i \epsilon_i^2}$, and so it follows that

$$\mathrm{SNR}_x = \frac{1}{\sqrt{\sum_i \mathrm{SNR}_i^{-2}}}. \tag{3.75}$$

Shot noise measured in photo-electric events equals the square root of the signal when this is also measured in photo-electric events. Since the average

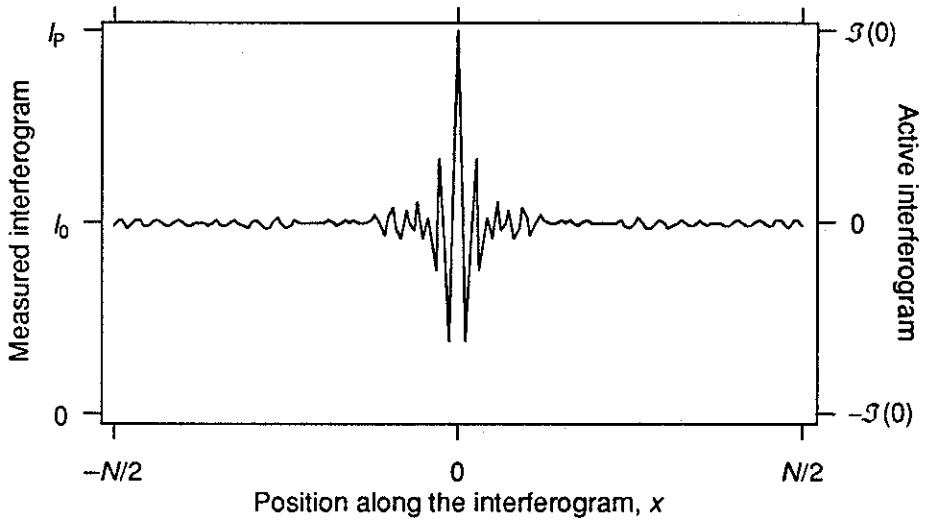


FIGURE 3.12: The ideal interferogram assumed in the noise calculations: perfectly symmetric and with optimal contrast. Intensity of the *active* interferogram (right hand axis) equals intensity of the *measured* interferogram (left hand axis) minus the background level: $\mathcal{I}(x) = I(x) - I_0$.

signal level of the interferogram equals (or at least approximates) I_0, its average Shot noise equals:

$$\epsilon_S = \sqrt{I_0} = \sqrt{\frac{I_P}{2}}. \tag{3.76}$$

Hence the Shot noise component of the SNR is:

$$\mathrm{SNR}_S = \frac{\mathcal{I}(0)}{\epsilon_S} = \sqrt{\frac{I_P}{2}}. \tag{3.77}$$

Resetting noise is independent of signal level and measured by the manufacturer to $\epsilon_R \approx 2000$ photo-electric events. The component SNR is therefore:

$$\mathrm{SNR}_R \approx \frac{\mathcal{I}(0)}{2000} = \frac{I_P}{4000}. \tag{3.78}$$

Digital noise depends upon the number of bits (N_B) used in the digital representation of the measurement: the value of each interferogram sample is digitized with a certain *dynamic range* given by: $D = 2^{N_B}$. Since the digital signal only contains integers, the smallest variation in its size is unity, corresponding to a signal variation of ΔI. The discrete nature of the digital signal gives it a 'staircase' appearance; the RMS noise implied by this misrepresentation of the measured signal may be shown to be:

$$\epsilon_D = \Delta I/(\sqrt{12}) \approx 0.3\,\Delta I. \tag{3.79}$$

If the peak of the interferogram exactly fills the dynamic range of the ADC, then $\Delta I = I_P/D$, and so the component SNR is given by:

$$\mathrm{SNR}_D \approx \frac{I_P/2}{0.3\,\Delta I} \approx 1.7D. \tag{3.80}$$

3.4.3 Optimal operating conditions

Figure 3.13 shows SNRs plotted against peak interferogram intensity. Thin,

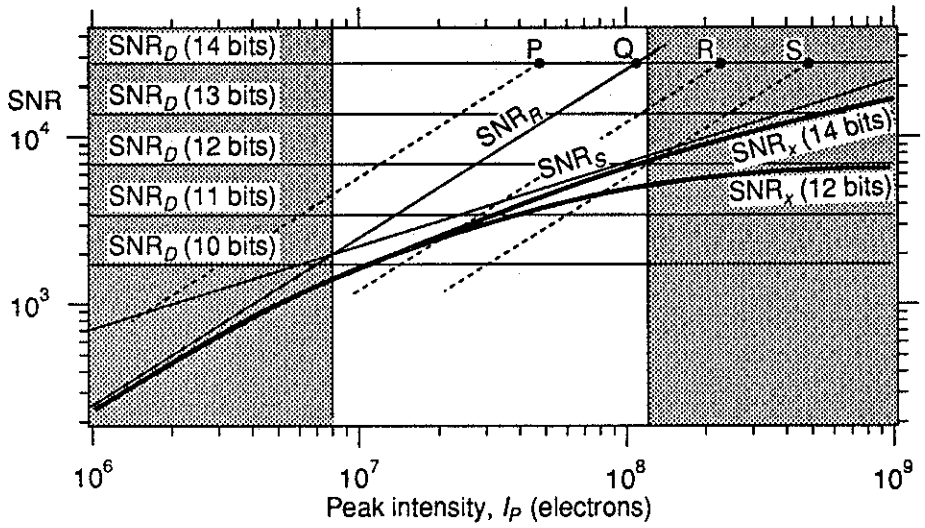


FIGURE 3.13: Signal-to-noise ratio (SNR) plotted against peak signal level in electrons. Thin, solid lines represent 'component' SNRs and thick lines show total SNR_x for 12 and 14 bit digitization. Operating points are shown as discussed in the text.

solid lines show the component ratios, and total ratios, calculated for 12 and 14 bit digitization, is represented by thick lines. The latter are seen to increase steeply at low intensities where the resetting noise is dominant, then rising less steeply as Shot noise takes over, and at high intensities they flatten out as digital noise becomes dominant. The unshaded region represents the useful range of the detector: below it resetting noise dominates, and above it the detector saturates.

Maximum interferogram SNR for a single exposure is $\sqrt{I_{\mathrm{Sat}}/2} \approx 8\times10^3$ electrons, where $I_{\mathrm{Sat}} = 1.2\times10^8$ is the saturation signal of the detector. This may only be achieved with at least 13 bit digitization and by letting a 'just unsaturated' signal completely fill the digital dynamic range. Shot noise limited

performance may also be achieved with 11 and 12 bit digitization, but then only at signal levels well below detector saturation.

In practice, 'just unsaturated' operation should be avoided because of the danger of nonlinearities close to saturation, and since the signal may accidentally exceed the saturation level due to fluctuations. To investigate such non-optimal operation, consider the concept of a "digital operating point". Such a point may be plotted in Figure 3.13 on the curve for the digital SNR corresponding to the digitization used, at the ordinate corresponding to a completely filled digital dynamic range. The horizontal position of the operating point may thus be varied by adjusting the gain of the analogue amplification.

Four such points (P, Q, R, and S) are plotted in the figure for the case of 14 bit digitization. From these points a dotted line is drawn showing the reduction in digital SNR as the collected signal is reduced: as long as the signal level stays within the unshaded region and the dotted line does not cross the Shot noise curve, the instrument is still Shot noise limited. 'Safe' operation, for which digital noise never dominates, clearly requires that the operating point (e.g. P) be situated at a signal level lower than the intersection between the resetting and digital noise curves (Q). If the operating point is set at a higher level (e.g. at R), the range of Shot noise limited operation reduces quickly with the danger of being rendered impossible (as in case S). Optimally, the operating point should be placed in the vicinity of point Q.

3.4.4 Spectral effects of white noise

White noise is equally distributed over all the frequency components of the measured signal. Our spectral estimate—which is the Fourier transform of the measured signal—is therefore assumed to have a uniform level of noise *unaffected by local spectral value.* This is different from the case of Shot noise limited spectra produced by classical grating spectrometers where the noise varies with the square root of the local spectral value. In the following we will find the relationship between interferogram noise and spectral noise in an HFTS system and compare its performance with that of a corresponding grating system.

Two important Fourier theorems are useful in these calculations: the Parseval identity known as the 'Power Theorem' relates the noise in the two domains, and the 'Area Theorem' relates the signals. The former states that the *power* contained in a function equals the power contained in its Fourier transform. Hence since the spectral noise with RMS value ϵ_ν is the transform of the interferogram noise with average RMS value ϵ_x, the theorem may be expressed as:

$$L\,\epsilon_x^2 = \nu_s\,\epsilon_{T,\nu}^2, \tag{3.81}$$

where L is the interferogram length, ν_s is the sampling frequency (i.e. the length of the spectrum including negative frequencies), and $\epsilon_{T,\nu}$ is the total spectral RMS noise. Half of the spectral noise power is in the imaginary part however, so for a phase corrected spectrum the RMS noise in the real spectral estimate is:

$$\epsilon_\nu = \frac{\epsilon_x}{\sqrt{2}}\sqrt{\frac{L}{\nu_S}} = \epsilon_x\sqrt{\frac{L\,\Delta x}{2}} = \epsilon_x \Delta x\sqrt{\frac{N}{2}}, \tag{3.82}$$

where $\Delta x = 1/\nu_s$ is the sample separation and $N = L/\Delta x$ is the number of interferogram samples.

The Fourier Area Theorem states that the value of a function at its origin equals the area under its Fourier transform. Assuming the active interferogram to be perfectly symmetrical with its origin chosen at the central peak (see Figure 3.12), its transform is real and equals the spectrum B_ν. The Area theorem may then be expressed as:

$$\mathcal{I}(0) = \delta\nu \sum_{-\nu_s/2}^{\nu_s/2} B_\nu(\nu) = \overline{B}\,\nu_s = \frac{\overline{B}}{\Delta x}, \tag{3.83}$$

where the summation represents the sum of all the spectral samples and $\overline{B} = \sum B_\nu/N$ is their average. Since interferogram signal-to-noise ratio is $\mathrm{SNR}_x = \mathcal{I}(0)/\epsilon_x$, the interferogram noise may now be written as:

$$\epsilon_x = \frac{\overline{B}}{\mathrm{SNR}_x\,\Delta x}, \tag{3.84}$$

and so, by the power theorem, the spectral noise becomes:

$$\epsilon_\nu = \frac{\overline{B}}{\mathrm{SNR}_x}\sqrt{\frac{N}{2}}. \tag{3.85}$$

Local spectral signal-to-noise ratio may therefore be expressed in terms of interferogram signal-to-noise ratio as:

$$\mathrm{SNR}_\nu = \frac{B_\nu}{\epsilon_\nu} = \mathrm{SNR}_x \frac{B_\nu}{\overline{B}} \frac{1}{\sqrt{N/2}}. \tag{3.86}$$

The ratio $\overline{B}/B_\nu$ is often called "spectral fill factor". For quasi-continous spectra where $B_\nu \approx \overline{B}$, the fill factor is unity. We find then for our instrument where $N = 512$ that $\mathrm{SNR}_\nu \approx \mathrm{SNR}_x/16$, i.e. giving a reduction with respect to interferogram's SNR. For single line emission spectra where almost all the signal is concentrated in one single spectral element, the fill factor is very low. For $B_\nu \approx \overline{B}\, N/2$ the spectral SNR at the line is $\mathrm{SNR}_\nu \approx 16\ \mathrm{SNR}_x$, i.e. greater than the interferogram's SNR.

As noted earlier, a 'just unsaturated' interferogram could reach a maximum Shot noise limited SNR of $\sqrt{I_{\mathrm{Sat}}/2}$, i.e. approximately $7.7{\times}10^3$. For a quasi-continous spectrum, the optimal SNR is therefore about 480. If only half the spectral range is "filled" however, the fill factor is a half and so the SNR increases to about 1000. For single line emission spectra the SNR soars to 10^5, but this value will probably never be reached due to other effects such as sampling grid errors as will be discussed shortly.

3.4.5 Comparison with CGS instruments

Due to the fundamental difference between the way FTS and CGS type instruments produce spectral estimates, measurement noise affects the estimates in different ways. For the FTS pioneers, working in the infrared where detector noise was predominant, this difference turned out as an advantage, called the multiplex or Fellgett advantage. Modern detectors in the visible are, as shown above for our detector, usually Shot noise limited; the situation is then different and the multiplex advantage —more appropriately called the multiplex *effect*—turns out to be a mixed blessing.

For a comparison with classical grating spectrometers we consider the non-scanning detector array based type discussed in Chapter 2. Here the spectral signal-to-noise ratio is simply equal to the square root of the number of photo-electric events contained in the signal. Assuming a spectrum whose peak value

'just fails to saturate' the detector, the local SNR in the CGS instrument may then be expressed as:

$$\mathrm{SNR}_G = \sqrt{\frac{B_\nu}{B_{\mathrm{Peak}}} I_{\mathrm{Sat}}}. \tag{3.87}$$

Writing out the spectral SNR for the HFTS instrument under similar conditions, i.e. for a 'just unsaturated' interferogram, by combining Equation 3.77 and Equation 3.86 gives:

$$\mathrm{SNR}_\nu = \frac{B_\nu}{\overline{B}} \sqrt{\frac{I_{\mathrm{Sat}}}{N}}. \tag{3.88}$$

The ratio between these two expressions may be written:

$$\frac{\mathrm{SNR}_\nu}{\mathrm{SNR}_G} = \sqrt{\frac{B_\nu}{\overline{B}}} \sqrt{\frac{B_{\mathrm{Peak}}}{N\overline{B}}}. \tag{3.89}$$

For a quasi-continous spectrum where $B_\nu \approx \overline{B} \approx B_{\mathrm{Peak}}$, grating instruments therefore have an advantage of factor $\sqrt{N}$, while at the peak of a single line emission spectrum where $B_\nu = B_{\mathrm{Peak}}$ and $\overline{B} \approx 2B_{\mathrm{Peak}}/N$, the HFTS instrument has an advantage of factor $\sqrt{N}/2$.
[See p. 93a.]

3.4.6 Discussion

The foregoing analysis is obviously stylised, notably in its assumption of optimal interferogram contrast. The contrast k of a sinusoidal interferogram component enters as a multiplying factor in the Fourier integral, giving a reduction of the spectral signal. It does not enter the integral for the background intensity, however, and so it does not affect the noise. The effect of reduced contrast is therefore a proportional reduction in spectral SNR.

Another simplification is the disregard of dark current in the noise calculations. Dark current produces noise in two different ways: its level varies randomly from element to element, and it contributes to the Shot noise. Fortunately the spatial variations are virtually constant in time and may therefore be reduced greatly by measuring the dark signal separately and subtracting it from signal measurements as will be demonstrated in Section 4.6.3. The Shot noise contribution can not be removed, but as long as the dark current is small compared to the signal, its Shot noise contribution is insignificant to

In the above comparison we have allowed the exposure time to vary, choosing optimal, just unsaturated exposure for both instruments. This is a necessary condition for Shot-limited operation when charge-integrating detectors are used (see Section 3.4.3). A more complete comparison between the two types of instrument should also take into account the throughput advantage (Section 3.4.6), however.

For the grating instrument, detector signal I_G may be expressed as in Equation 2.32b, giving an SNR of:

$$\mathrm{SNR}_G = \sqrt{I_G} \propto \sqrt{G_G\, B} \tag{3.89a}$$

(where $B \equiv B_\nu$). For HHS instruments, the local SNR may be found from Equation 3.88 by replacing I_{Sat} with $2I_0 = 2\overline{I_H}$, with $\overline{I_H}$ given by Equation 2.32a:

$$\mathrm{SNR}_H = \frac{B}{\overline{B}}\sqrt{\frac{2\,\overline{I_H}}{N_D}} \propto B\sqrt{\frac{2G_H}{\overline{B}N_D}} \tag{3.89b}$$

(where $\mathrm{SNR}_H \equiv \mathrm{SNR}_\nu$ and $N_D \equiv N$).

Ratioing the two SNRs now give:

$$\frac{\mathrm{SNR}_H}{\mathrm{SNR}_G} = \sqrt{\frac{2G_H}{N_D G_G}}\sqrt{\frac{B}{\overline{B}}}. \tag{3.89c}$$

Note the similarity between this ratio and that of detector signals in Equation 2.32c. As might be expected, the SNR ratio depends upon the square root of the throughput ratio. Less obvious is its dependence upon the *inverse* of the fill factor ($f = \overline{B}/B$) which gives a net improvement in favour of the HHS method as f decrases.

Substituting for throughputs from Equations 2.20 and 2.32 and using the same instrumental parameters as in the throughput comparison of Section 3.4.6, Equation 3.89c evaluates to:

$$\frac{\mathrm{SNR}_H}{\mathrm{SNR}_G} = \sqrt{\frac{16\,F^2\,y_F}{f\,l_D\mathcal{R}}} = \sqrt{\frac{690}{f\mathcal{R}}}. \tag{3.89d}$$

Solved for the condition $\mathrm{SNR}_H > \mathrm{SNR}_G$ this equation gives $\mathcal{R} < 690/f$. For quasi-continous spectra ($f \approx 1$) HHS instruments therefore have a net advantage over CGS instruments when resolving power is less than 690.

the total noise. At long exposures however, when dark current starts to dominate over signal current, its effect becomes serious. As seen in Chapter 2, the throughput advantage of FTS instruments over CGS instruments allows shorter exposure times and hence less dark current. For broad-band spectra it is therefore mainly under poor light conditions that our instrument may be expected to gain a noise advantage over its CGS equivalent.

3.5 Interferometer aberrations

Some of the most serious deficiencies of our prototype instrument are caused by aberrations in the interferometer due to poor mounting of the reflectors. This causes the interfering wavefronts to be curved rather than plane, giving reductions in spectral quality in two ways:

- Curved fringes which, when imaged onto a one-dimensional detector array, causes a "leakage" of information between the detectors, and hence a reduction in contrast, and
- Non-linear relationship between phase difference and distance across the interferogram causing the period of the fringes to vary with position.

The former effect is blamed for the instrument's poor fringe contrast or modulation transfer (see Section 4.3.6), resulting in a reduction in signal-to-noise ratio. The latter effect modifies the spectrum as will be seen in the following.

Poorly mounted reflectors is not the only cause of interferometer aberrations. Grating ruling errors also give rise to such errors, as do inhomogeneities in optical components. When considering the effects of interferometer aberrations we distinguish between three different types:

Random, caused chiefly by small-scale inhomogeneities and surface defects of optical components in the interferometer,

Periodic, caused typically by grating ruling errors, and

Monotonic, caused by poor reflector mounting.

3.5.1 Apparent sampling grid errors

A monochromatic interferogram produced by an aberrated interferometer appears as if it had been sampled on an irregular sampling grid. Hence the error may be counteracted by *resampling*. For unheterodyned measurements the apparent sampling error is the same for all spectral components; resampling therefore also improves broad band spectra in this mode. In heterodyned modes, however, the apparent sampling error is frequency dependent, leaving resampling of broad-band measurements ineffective.

If $w_\varepsilon(x)$ is the wavefront aberration in the interferometer, then the distance between the wedged, interfering wavefronts is given by $x \sin\alpha + w_\varepsilon(x)$. Equation 3.7, describing the phase difference between the wavefronts, must hence be rewritten as:

$$\begin{aligned} \Delta\delta &= 2\pi\sigma[x\sin\alpha + w_\varepsilon(x)] \\ &= 2\pi\nu\,(x+\varepsilon), \end{aligned} \tag{3.90}$$

where $\varepsilon = w_\varepsilon(x)\sigma/\nu$ is the apparent sampling error and ν is the fringe spatial frequency as defined in Equation 3.10. In unheterodyned operation, ε is frequency independent since spatial and optical frequencies are proportional. For heterodyned systems this proportionality is lost and the apparent sampling error varies from one spectral component to another, leaving resampling of broad-band measurements useless.

We think that it is still possible to correct heterodyned interferograms, but that this involves a much increased computational load. The "original", unheterodyned interferogram may be recreated by shifting the spectrum onto a frequency axis which is proportional to wavenumbers before transforming it back into the interferogram domain. This original interferogram would have a number of points of the order of the resolving power of the spectrum. Presumably, resampling applied at this point should give the required spectral improvement.

A better way to go about things is of course to remove the interferometer aberrations in the first place. Avoiding the monotonic error should be achieved relatively easily by improving the way the reflectors are supported, see Sec-

tion 6.4. Periodic errors, assumed to be caused by grating ruling deficiencies, are also seen to be significant however (Figure 4.35), and so, to remove such errors without placing too high demands upon grating manufacture, a computational correction method is still desirable.

3.5.2 Spectral effects of interferometer aberrations

We have modelled theoretically the effects of interferometer aberrations on spectral estimates by assuming the sampling grid error equivalence.

Let $\mathcal{I}(x)$ represent an ideally symmetric interferogram and $\mathcal{I}_\varepsilon(x) = \mathcal{I}(x+\varepsilon)$ an erroneously sampled version of it. They may be related in terms of a Taylor series, viz:

$$\begin{aligned} \mathcal{I}_\varepsilon(x) &= \mathcal{I}(x+\varepsilon) \\ &= \mathcal{I}(x) + \varepsilon\mathcal{I}^{(1)}(x) + \frac{\varepsilon^2}{2}\mathcal{I}^{(2)}(x) \\ &\quad + \cdots + \frac{\varepsilon^n}{n!}\mathcal{I}^{(n)}(x) + \cdots, \end{aligned} \tag{3.91}$$

where $\mathcal{I}^{(n)}(x)$ is the nth derivative of the interferogram. Fourier transformation of this series according to the Fourier Derivative Theorem [31] gives the spectral estimate:

$$\begin{aligned} B_E &= B_\nu + \Re\left(\mathbf{e}_1 \star i2\pi\nu B_\nu\right) - \Re\left(\mathbf{e}_2 \star (2\pi\nu)^2 B_\nu\right) \\ &\quad + \cdots + \Re\left(\mathbf{e}_n \star (i2\pi\nu)^n B_\nu\right) + \cdots, \end{aligned} \tag{3.92}$$

where $\mathbf{e}_n = \mathcal{F}\{\varepsilon^n/n!\}$ is an "error instrument function of the nth order". The spectral estimate is affected by a series of error terms which may be found by convolving the error instrument functions with spectra modified by the multiplication with $(i2\pi\nu)^n$. Note that smooth spectral features act as filters for the error instrument function so that 'smooth' spectra are less affected than 'sharp' spectra.

As long as the sampling error is small, the size of the error instrument function decreases with the order, and a good estimation of the error may then be obtained by considering only the first order term in Equation 3.92. For random and periodic errors this condition must necessarily be fulfilled, otherwise the interferogram should be modified beyond recognition. In the

case of monotonic errors however, the error may well become considerable towards the edges of the interferogram. For such errors we have therefore found it necessary to include also the second order.

When ε is **random**, with RMS value $\epsilon_{\varepsilon x}$ the first order error instrument function is also random with RMS value in each of its real and imaginary parts as given by the Fourier Power Theorem (Equation 3.81):

$$\epsilon_{\varepsilon\nu} = \epsilon_{\varepsilon x}\,\Delta x\sqrt{\frac{N}{2}}. \tag{3.93}$$

Because of the filtering effect noted above, broad spectra appear less noisy than sharp spectra: the effect of random sampling error is therefore opposite to the effect of intensity noise for which sharp (emission-type) spectra have a noise advantage.

A **periodic** sampling error has $\varepsilon = \varepsilon_0 \cos(2\pi\omega x + \theta)$ where ε_0 is its amplitude, ω its spatial frequency, and θ its phase. It is useful to decompose the error into even and odd in-phase contributions: $\varepsilon = \varepsilon_0(\cos\theta\cos 2\pi\omega x - \sin\theta\sin 2\pi\omega x)$. The first order error instrument function, $\mathbf{e}_1$, is given by the Fourier transform of ε and has therefore a real part with two positive peaks of size $(\varepsilon_0\cos\theta)/2$ at $\pm\omega$ and an imaginary part with two peaks of size $(\varepsilon_0\sin\theta)/2$, negative at $+\omega$ and positive at $-\omega$. By Equation 3.92, only the imaginary part of $\mathbf{e}_1$ contributes to the real spectral estimate, and the resulting spectral error is therefore a pair of oppositely signed *ghosts* of relative magnitude $\pi\nu_0\varepsilon_0\sin\theta$ around sharp spectral features at $\nu_0 \pm \omega$, where ν_0 is the frequency of the mother feature.

For sufficiently small errors, only the first order ghosts will be visible over the noise, but as the error increases, higher order ghosts increase in importance and will eventually dominate over both first order ghosts and the mother feature. The effect is similar to diffraction at a sinusoidal phase grating where the intensity of the various diffraction orders are described by Bessel functions with respect to the modulation depth of the grating [1].

As will be argued in Section 4.6.5, periodic sampling errors are encountered in heterodyned holographic FTS as the result of a periodic ruling error in the grating. The resulting ghosts are different from those occurring if the grating

was used in a classical grating spectrometer in that they are one positive and one negative. They are also larger since they are directly proportional to the amplitude of the ruling error rather than to the amplitude square as in the classical case.

Monotonic errors may have many functional shapes. The simplest is linear, but this only corresponds to an extra wedge between the interfering wavefronts. Of more serious consequence is the quadratic error ($\varepsilon \propto x^2$) caused by astigmatism in the interferometer. Letting $\varepsilon = kx^2/\Delta x$, the nth term of the interferogram expansion (Equation 3.91) may be written as:

$$\frac{k^n}{(\Delta x)^n n!}\, x^{2n} \mathcal{I}^{(n)}(x). \tag{3.94}$$

The Fourier transform of this expression yields the spectral error term of the nth order and may, by the Fourier Derivative Theorem, be seen to equal:

$$\Delta \mathbf{B}_n(\nu) = \left(\frac{-ik}{2\pi\,\Delta x}\right)^n \frac{D^{2n}\left[\nu^n B_\nu(\nu)\right]}{n!}, \tag{3.95}$$

where D^p stands for the derivation operator $d^p/d\nu^p$. Hence each order of the spectral error consists itself of a series of terms, and these may be found by the Leibnitz rule for higher derivatives of products.

Note that, since $D^p(\nu^n) = 0$ when p is greater than or equal to n, ${B^{(n)}}_\nu(\nu)$ is the lowest derivative of $B_\nu(\nu)$ present in the nth order error term. Since only even order terms are real, the lowest spectral derrivative affecting the real estimate is therefore the second. For smooth spectra whose derivatives are small the effect is therefore also small. Sharp features such as unresolved lines have large derivatives however, and, assuming that each order of the spectral error is dominated by the term containing the highest spectral derivative, the following approximation may be justified:

$$\Delta \mathbf{B}_n(\nu) \approx \left(\frac{-i\,k}{2\pi\,\Delta x}\right)^n \frac{\nu^n {B_\nu}^{(2n)}(\nu)}{n!}. \tag{3.96}$$

Hence the nth order spectral error term is proportional to the $2n$th spectral derivative. A line feature is therefore al ways affected by a symmetric error, both in its real and imaginary parts. Note also that the error depends upon the nth power of frequency and is hence highly variable across the spectrum.

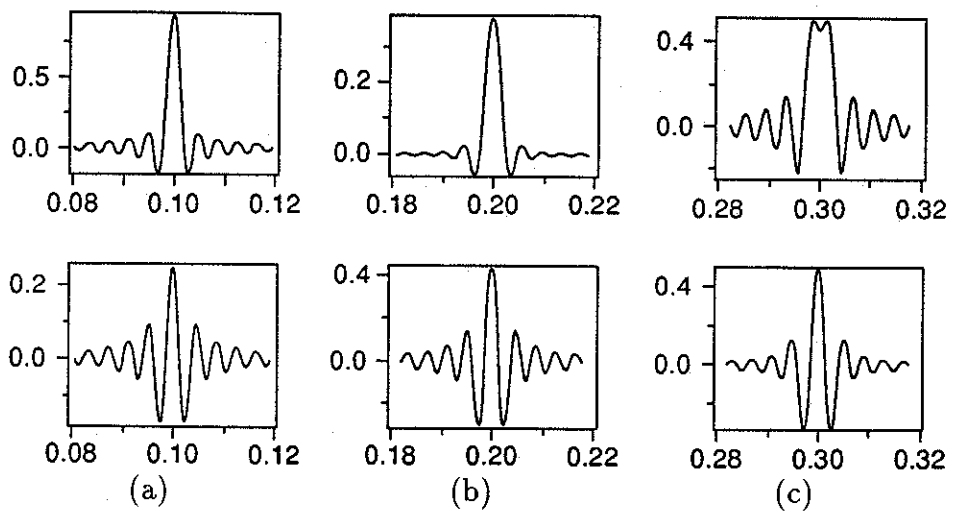


FIGURE 3.14: The effect of a monotonic sampling error of $k = 2 \times 10^{-5}$ at different points in the spectrum. Top graphs show real instrument functions, bottom ones show imaginary instrument functions. The horizontal (frequency) axis has units of fringe cycles per detector element (Elements^{-1}).

Figure 3.14 shows the real and imaginary parts of an unresolved spectral line calculated by this approximation at three different frequencies by using error terms of the first and second orders. Comparisons with measured line spectra (see Section 4.6.5) verifies the validity of these calculations. In (a) the error is apprechiable only in the imaginary part because the second order term is not yet significant. In (b) and (c) the real spectral estimate is also affected: first as a broadening and a reduction of the wings, later a double peak appears with large and slowly decreasing wings.

3.6 Conclusion

After a brief review of the literature and historical development of holographic Fourier transform spectroscopy, we have in this chapter shown how the interference pattern from a two-beam interferometer yields spectral information via the Fourier transformation. By the aid of familiar physical phenomena, the formation of interference in a non scanning Michelson interferometer has been explained, and the effects of finite and sampled measurements have been presented.

An important part of the signal processing required for obtaining good

spectral estimates is phase correction. We give a theoretical treatment of the problem with particular attention to effects relevant to our instrument. At the end of the section we give numerical examples where operational tolerances for our instrument are calculated. Of some interest are the results regarding the technique of single sided measurements by which ideally a factor of two in resolving power may be gained. According to our calculations we may expect a gain of at least 1.4 by this technique.

Causes and effects of random intensity noise are presented, and a method for choosing the optimal operating point is described. A comparison of the noise performance of our instrument with that of a non scanning grating spectrometer is given, showing that the HFTS technique suffers a *disadvantage* with respect to CGS instruments for quasi-continuous spectra under Shot noise limited conditions. Signal-to-noise ratios in HFTS spectra of the order of 1000 are expected however, even under very poor light conditions.

Finally, we discuss the problem of apparaent sampling errors caused by interferometer aberrations. The aberrations are classified in three groups: random, periodic and monotonic. Periodic aberrations are encountered in heterodyned holographic FTS when the grating has a periodic ruling error and produces ghosts as in classical grating spectrometers. Monotonic aberrations occur in holographic FTS when the interferometer is affected by astigmatism. The interfering wavefronts are then curved instead of plane, causing a variation in fringe frequency across the interferogram. As shown by simulations, the effect of this error is primarily to broaden the instrument function. Interferometer aberrations may, at least in some cases, be cancelled by resampling of the interferogram.

Chapter 4

Instrument Design

Designing and constructing the prototype instrument has been the major part of the work for this thesis and it is the purpose of this chapter to describe and characterize the working instrument. Far from trying to describe the entire design process with all its stumbling and 'cut-and-try,' we will concentrate on a description of the main components of the design. These are presented against the background of the design considerations upon which they are based and their measured performance.

Two components of the design stand out as key elements: the beam splitter and the fringe imaging lens; a large part of the chapter is consecrated to describing these two components. Mechanical and electronic designs are also described but only in broad lines. Signal processing is of course an important part of the design of an FTS-type instrument, without which its output would be practically meaningless. Although the scope of our work has not included the development of a purpose built signal processing system, we have implemented all its basic features in an experimental system based on a general purpose mathematics programme with hardware interface capability. We describe the features of this implementation and discuss points of experience. First, however, we offer a brief overview of the system.

4.1 Design overview

Apart from the ability to gather spectral information of a certain resolution with a certain sensitivity, three criteria have been particularly important for

the design of this instrument:

- Ruggedness,
- Compactness, and
- Low power consumption.

In order to fulfil these criteria, we have kept the number of moving parts at a minimum, and motorized drives and detector cooling have been avoided.

Optical and electronic systems are both contained within a single unit. Supplied with power from a small external battery pack, the instrument is capable of stand-alone, automatic data logging controlled by its own on-board micro-processor. Manual operation from a portable computer is ensured via a standard RS-232 serial data link, see Figure 4.1. Processing of data is performed on the portable computer.

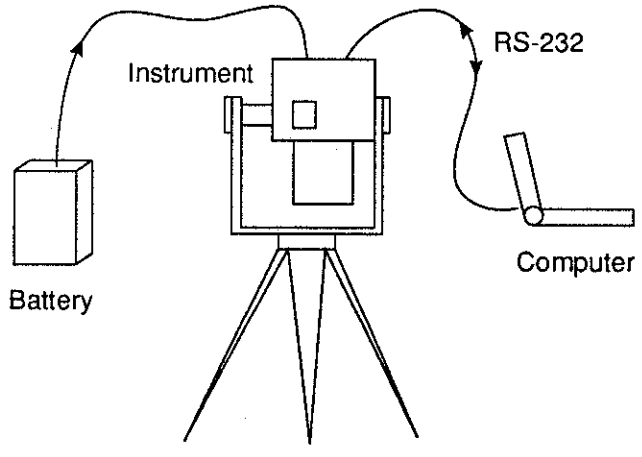


FIGURE 4.1: Schematic view of the instrument system. Two-way communication with a portable computer is ensured via a standard serial link.

Looking into the instrument itself reveals—apart from an impressive presence of electronics—a construction based on four optical units: a filter compartment, the interferometer, the fringe imaging lens, and the detector housing. Figure 4.2 shows a perspective view of the optical components stripped of their mechanical fixtures. It is clearly a rather simplistic optical design, and we have tried to reflect its simplicity in the mechanical design as will be described in Section 4.4.

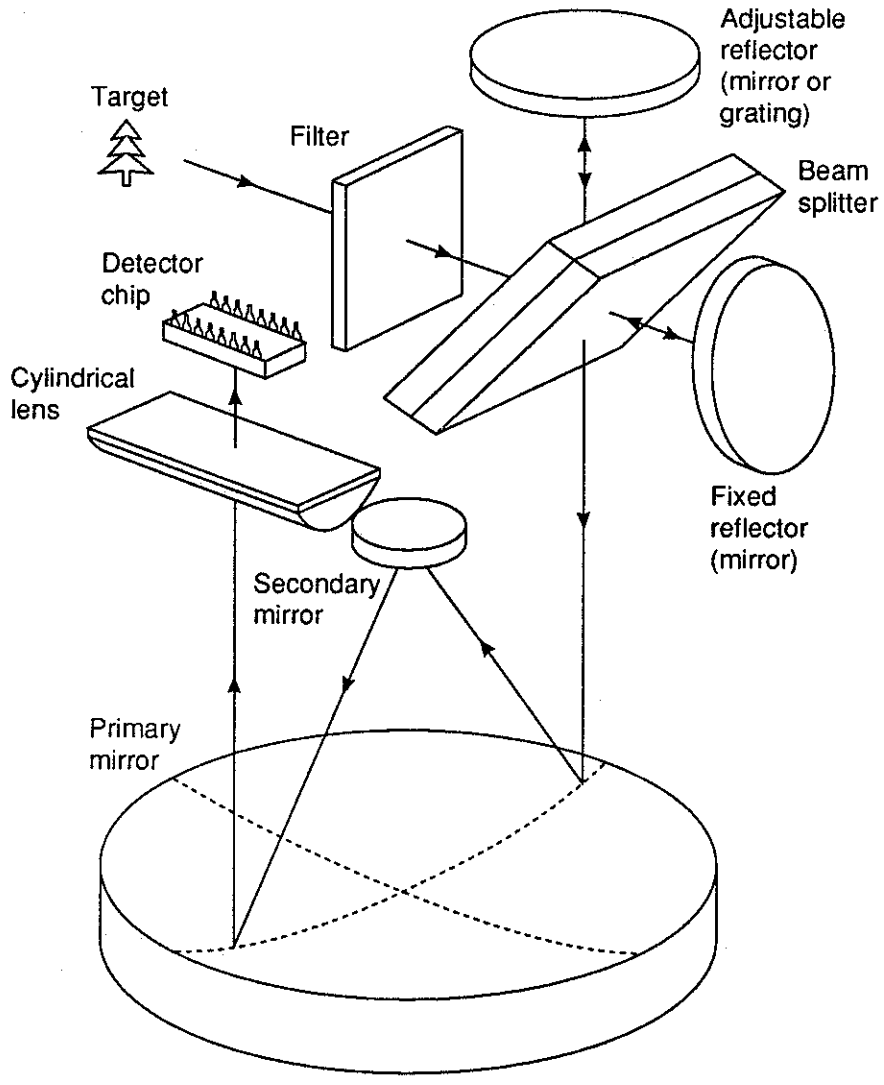


FIGURE 4.2: Perspective view of the optical system, stripped of its support mechanisms.

We have seen in the preceding chapter that the HFTS method of spectroscopy consists of forming, by interference, an intensity pattern known as the interferogram. Imaged onto a detector array, this pattern is measured and digitized and yields spectral information after some signal processing. Design criteria for the optical part of the system are based mainly on a desire to optimize the contrast of the interferogram. The notion of fringe contrast or modulation depth has been introduced in the context of interference but, as will be seen in the following, it is also affected by stray light and the quality of the imaging system.

The instrument may operate with a resolving power of up to about 10 000 and cover a spectral window of 256 independent spectral samples anywhere within the range 0.4–1.0 μm. Changing resolving power is done by a change of grating, while a shift of the spectral window is achieved by adjusting the interferometer and changing the filters. The instrument is optimized for operation at a resolving power of 1000 with its window centred at about 700 nm; this mode has been specifically conceived for studies of the vegetation red-edge and will be referred to as the **medium resolution** mode. Unheterodyned operation with its maximum resolving power of 256 covers the entire spectral range of the instrument and is referred to as the **low resolution** mode. A third mode, the **high resolution mode**, with resolving power 5000 has also been implemented to demonstrate the instrument's high resolution capability and with the specific aim of studying atmospheric NO_2 absorption. At this resolution we resolve well the sodium doublet and observe Fraunhofer lines in the solar spectrum.

4.2 Beam splitter assembly

The beam splitter assembly consists of two optical components: the substrate onto which a semi-transparent film is deposited, and a dispersion compensation plate. In order to save space and gain ruggedness we have chosen to glue the two components together, forming a sandwich with the semi-transparent film sealed between them, see Figure 4.3. The successes and failures of this choice will be discussed, together with considerations with respect to materials and coatings, but in order to gain objective criteria for the specification of the beam splitter, we first take a closer look at how it affects fringe contrast and instrument transmission factor.

4.2.1 Fringe contrast

Equation 3.6 defines fringe contrast in terms of the ratio between intensities of the recombining beams. We assumed then a loss-less interferometer, but in harsh reality this assumption is not always valid and we must therefore take into account absorptions as well as outer surface reflections. We introduce for

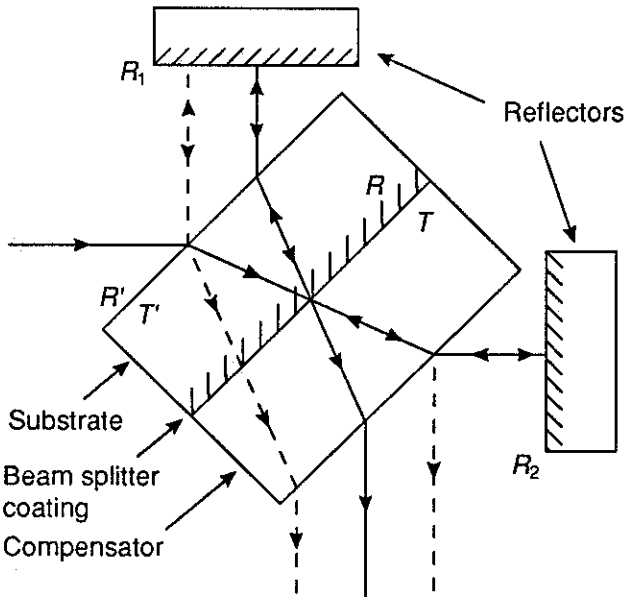


FIGURE 4.3: Cross section of the interferometer showing the sandwiched beam splitter design. The figure also shows the notation used in the text for reflection and transmission coefficients. Full lines represent the desired light paths and broken lines the main spurious light paths.

this some new quantities (see Figure 4.3):

- R and T: Reflection and transmission coefficient of the beam splitter whose sum, when the beam splitter is absorbing, does not reach unity. Since the assembly is essentially symmetric, the coefficients are assumed to be equal for light incident from either side. This is not strictly true since the layer of glue makes the sandwich asymmetrical, but the refractive index of the glue is close to that of the substrate and its asymmetrizing effect is therefore assumed to be negligible.

- R' and T': Reflection and transmission coefficients respectively of the outer surfaces of the beam splitter assembly. There is no absorption since it is a boundary between two dielectrics, and their sum is therefore equal to one.

- R_1 and R_2: Reflection coefficients of the two interferometer reflectors. In the heterodyned mode, where a grating replaces the mirror in one arm, R_2 denotes the diffraction efficiency of the grating.

With reference to Figure 4.3, we may now express the intensities of the two interfering beams as:

$$I_1 = I_0 T' R T' R_1 T' T T' = I_0 T'^4 R_1 R T \tag{4.1}$$

and:

$$I_2 = I_0 T' T T' R_2 T' R T' = I_0 T'^4 R_2 R T. \tag{4.2}$$

The beam splitter also produces stray light in the form of spurious reflections off its outer surfaces. In general these spurious reflections are out of phase with each other and do not create interference; they contribute therefore only to the background level, thus reducing the contrast. We will see in Section 4.2.6 that some spurious reflections do interfere, but since these have seen several passes through the beam splitter they are very weak. Here we only consider spurious reflections of the first order: those which suffer a single pass through the beam splitter. There are two such paths, drawn in dotted lines in Figure 4.3, and their intensities are:

$$I_1' = I_0 R' R_1 T' T T' = I_0 R' T'^2 T R_1 \tag{4.3}$$

and

$$I_2' = I_0 T' T T' R_2 R' = I_0 R' T'^2 T R_2, \tag{4.4}$$

respectively.

With reference to Equation 3.3 and Equation 3.4 and assuming I_1' and I_2' to be incoherent, the interference pattern may then be written as:

$$\begin{aligned} I &= I_1 + I_2 + I_1' + I_2' + 2\sqrt{I_1 I_2}\cos\Delta\delta \\ &= \left(1 + \frac{2\sqrt{I_1 I_2}}{\sum I}\cos\Delta\delta\right)\sum I, \end{aligned} \tag{4.5}$$

where $\sum I$ is the sum of the individual intensities, equal to the background level of the interferogram. It may be written as:

$$\sum I = I_0 T'^4 T (T'^2 R + R')(R_1 + R_2), \tag{4.6}$$

and so the interferogram contrast defined in Equation 3.6 may be expressed as:

$$k = \frac{2\sqrt{I_1 I_2}}{\sum I}$$

$$= \frac{T'^2 R}{(T'^2 R + R')} \frac{2\sqrt{R_1 R_2}}{(R_1 + R_2)}$$
$$= k_B k_R, \tag{4.7}$$

where k_B and k_R are contrast factors related to the beam splitter and the reflectors, respectively. These factors become more accessible by assuming R' to be much less than R and the difference between R_1 and R_2 (δ_R, say) much less than their average (R_R, say). We then find the following approximations:

$$k_B \approx 1 - \frac{R'}{R} \tag{4.8}$$

and

$$k_R \approx 1 - \frac{\delta_R^2}{8R_R^2}. \tag{4.9}$$

Equation 4.8 shows that interferogram contrast is proportional to the beam splitter's outer surface reflection, but not at all affected by the balance between reflectance and transmittance in the beam splitter. By Equation 4.9, the balance between *reflector* reflectances is important only in the second order; this is of great comfort for heterodyned operation where the diffraction efficiency of the grating may be expected to be lower than the reflectance of the mirror and to vary greatly across the spectrum.

Figure 4.4 shows plots of k_B against R for different values of R', and k_R

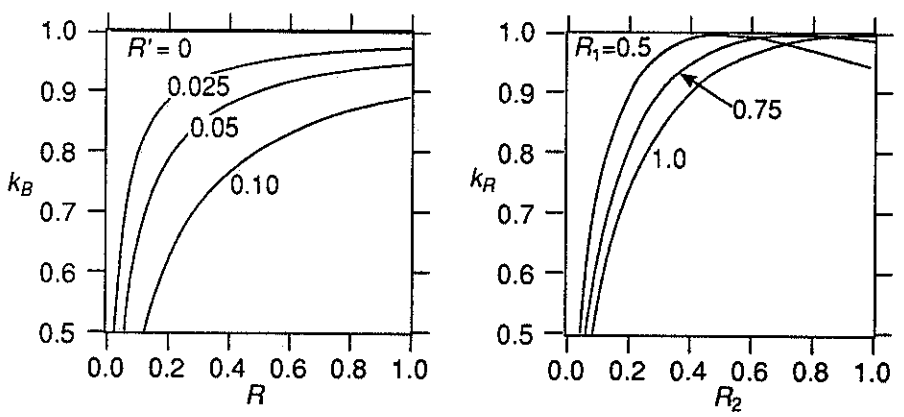


FIGURE 4.4: Plots of the contrast factors k_B (a) and k_R (b). The former is plotted against beam splitter reflectance (R) and outer surface reflectance (R'), the latter against reflector reflectances (R_1 and R_2).

versus R_2 for a range of R_1 as calculated from their exact definitions in Equation 4.7. Useful in setting manufacturing specifications, these curves reiterate

the conclusions drawn from the approximate relations. For the interferometer to produce high-contrast fringes it is clearly of great importance to control the outer surface reflection at the beam splitter but, as is further discussed in Section 4.2.5, this is difficult due to the strong polarization effect at 45° incidence.

4.2.2 Transmission factor

High contrast is not the only criterion for the interferometer, however; it should also have a high transmission factor so that as much as possible of the incoming light is passed on to the detectors. Since the total transmitted light is $\sum I$ the transmission factor may be found by deviding Equation 4.6 by I_0, the incident light. Assuming an ideal situation where R_1 and R_2 approximate unity and outer surface reflections are small, then:

$$\frac{\sum I}{I_0} \approx 2RT = \frac{(1-A)^2 - \delta_{RT}^2}{2}, \tag{4.10}$$

where A is the beam splitter's absorption coefficient and $\delta_{RT} = R - T$ is the difference between the beam splitter's reflection and transmission coefficients, the "beam splitter balance". The approximation is plotted against δ_{RT} for some values of A in Figure 4.5. Its maximum value of 0.5 reflects the fact that half the collected radiation returns out through the input.

Estimates of contrast and transmission factors for the constructed interferometer will be given a bit further on (Figure 4.7), and a measurement of the actual contrast in the form of the *modulation transfer function* is shown in Figure 4.21.

4.2.3 Beam splitter design

The semi-transparent film of the beam splitter is supported by a substrate of fused silica ('synthetic quartz'). In order to compensate for the dispersion of the substrate, a *compensator plate* of identical material and thickness is situated on the other side of the film. For compactness and ruggedness, the two pieces are in our design cemented together. To avoid that the layer of cement ruins the dispersion compensation, the compensator is made slightly

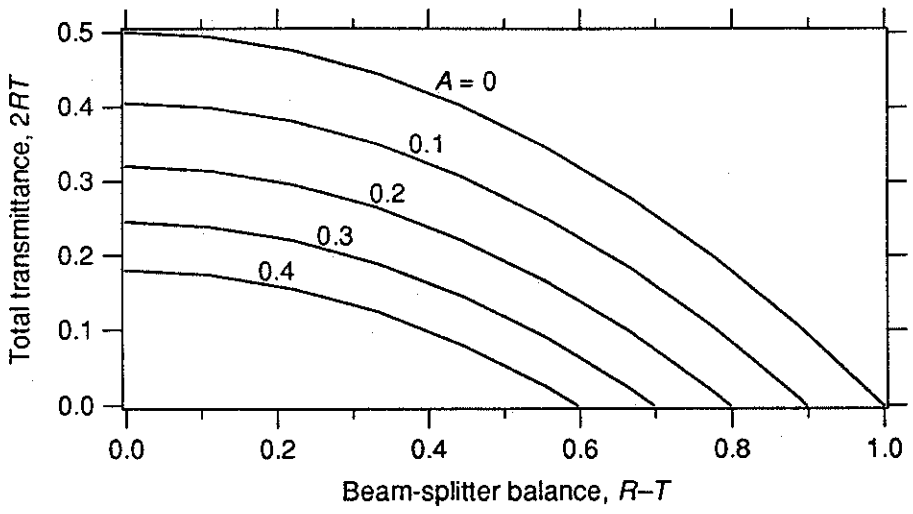


FIGURE 4.5: Beam splitter transmittance-reflectance product representing the transmission factor of an idealized interferometer (see text). The product is plotted against the difference δ_{RT} for various levels of absorptance.

(9 μm) thinner than the substrate by cutting the two plates from a single, wedge-shaped work-piece.

Fused silica has been chosen as substrate material mainly because of its low coefficient of thermal expansion which allows substrate and compensator to be cemented together immediately after handling. With optical glass the two pieces would have to rest for about an hour after touch to regain thermal equilibrium and this would require more complicated and expensive rigging. Otherwise the materials are equivalent since the good transmission characteristics of silica in the UV are lost to a strong absorption in the cement just short of 400 nm [67].

The beam splitter assembly is placed at 45° to the direction of the incoming light, but thanks to refraction at the outer surfaces the angle of incidence at the semi reflecting surface is only 29.5°. This reduction in angle is beneficial since it weakens the effect of polarization.

4.2.4 Beam splitter coating

For the semi-reflecting surface we have considered three different thin films: a "standard" metallic film, a silver film, and an all-dielectric multi-layer film. The standard metallic film has a very good reflection-transmission balance

($R = T = 0.3 \pm 0.05$), but it was rejected because of its very high absorption ($A = 0.4$). Both of the remaining films promised far better performances with absorptions less than 0.1 for the silver and essentially zero for the dielectric film. When the dielectric option eventually was dropped, this was chiefly because its manufacturer was uncooperative with respect to detailed specifications. For the silver film, which was produced in a local facility under our control, properties and characteristics could be studied in the literature.

The optical performance of thin silver films is strongly affected by the formation of 'isolated islands' during the deposition process which causes greater absorption than predicted from bulk properties [64]. For films produced by thermal evaporation the problem may be considerably reduced by fast evaporation (10 Å/s) giving films with virtually bulk characteristics for thicknesses down to about 150 Å [66]. Since calculations based on bulk properties showed that a film of thickness 160 Å would perform well in our instrument the problem should therefore be negligible.

Figure 4.6(a) shows a comparison between calculated and measured per-

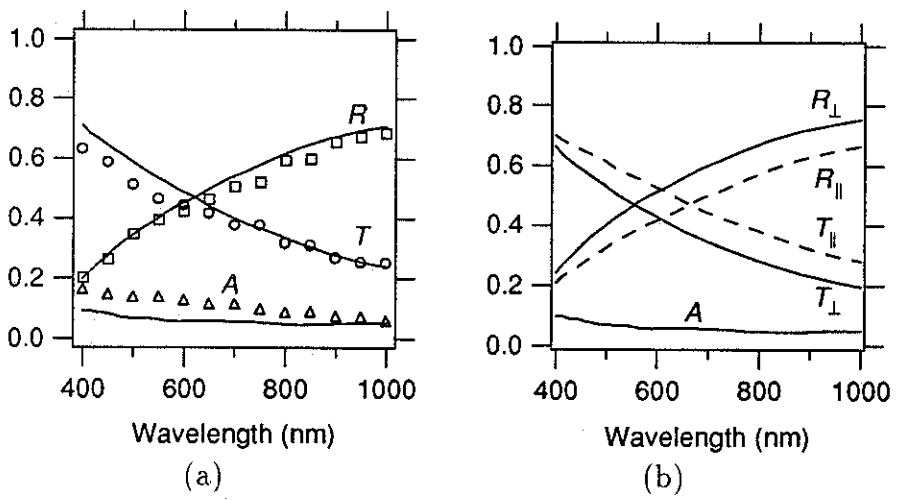


FIGURE 4.6: Comparison between simulated and measured beam splitter characteristics at normal incidence (a), and simulated characteristics at 45° incidence of the manufactured sandwich with a silver semi-reflective coating (b). In (b) solid lines represent the perpendicular polarization and broken lines parallel polarization.

formance of the finished beam splitter assembly at normal incidence; apart from a somewhat higher absorption, the film performs well compared with

theory. Figure 4.6(b) shows calculated performance for the assembly at 45 degrees to the incident radiation—direct measurement of its performance at this angle has not been possible. The effect of this change of angle is seen to give a difference of about 0.1 in both R and T between the two polarization components. Apart from a slight polarization dependence for the transmission factor, this is of little consequence for the instrument performance.

There are two important classical problems with silver films: poor adherence to the substrate and a tendency to oxidize. These are both solved in our design by hermetically sealing the film between substrate and compensator. A test of the gluing process was made with a film deposited on a microscope slide to ensure that it was not ruined by the sandwiching. Because of these problems however, silver films are little used nowadays; alternative materials with better mechanical and chemical properties are used instead. As a consequence, practical experience is rare, and we found it necessary to run an extensive series of tests in order to establish appropriate routines and a sufficiently accurate thickness calibration. This took a lot of time and may be a good reason for looking into alternative coatings for future work. As will be seen further on, other arguments indicate the same conclusion.

4.2.5 Anti-reflection coating

Equation 4.8 suggests a strong dependence of interferogram contrast on outer surface reflection, promising high returns from a reduction in its value, i.e. by applying an anti-reflective (AR) coating. We find however that for an angle of incidence of 45°, the polarization dependence of AR coatings is very strong, making it difficult to control both polarizations at once.

An uncoated surface at 45° incidence reflects 8.0% of the perpendicular component and 0.64% of the parallel component, the good performance of the latter is due to the Brewster effect [6, page 43]. Its average of 4.3% gives according to Equation 4.8 a contrast of 0.91 in an otherwise ideal interferometer ($R = 0.5$ and $R_1 = R_2 = 1.0$). With a single, quarter-wave layer of magnesium fluoride (MgF_2), reflectance for the perpendicular component may be kept below 5% within the interval 450 to 900 nm with a minimum of 3.7% while

that for the parallel component is reduced to below 0.3%. Contrast is then increased to 0.96. Of more sophisticated coatings, a three layer structure specifically designed for a 45 degree incidence has been described [65], but its effectiveness is still limited to only one of the polarizations and its total performance is hardly any better than that of the single layer.

Figure 4.7(a) shows the expected contrast and throughput for an uncoated

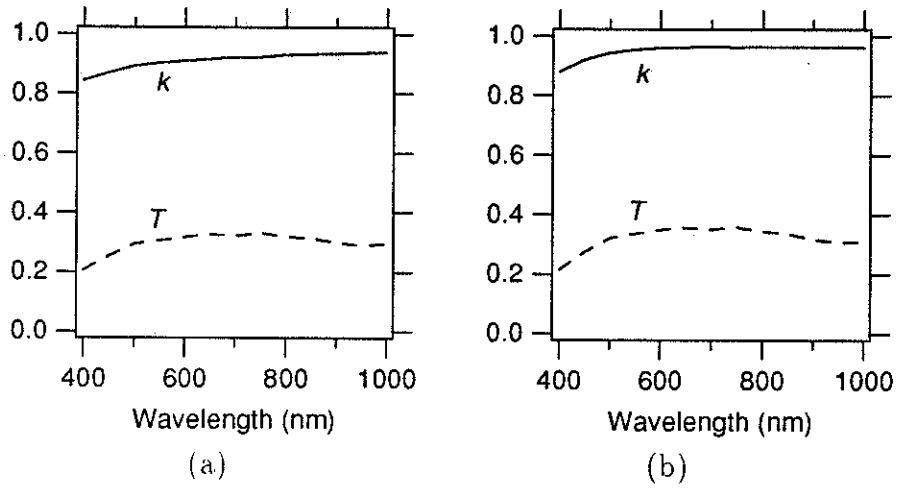


FIGURE 4.7: Expected contrast (solid lines) and transmission factors (broken lines) for the unheterodyned interferometer without (a) and with (b) anti-reflection coating.

beam splitter, and (b) shows its performance when its outer surfaces have been coated with a single layer of MgF_2. Although clearly present, the improvement is limited and has not been considered worth the investment. The beam splitter's outer surfaces have therefore been left uncoated.

4.2.6 Dispersion compensation

The purpose of the compensator plate is to make sure that the two interfering beams travel equal optical paths regardless of their frequency, usually ensured by making the compensator plate identical to the beam splitter substrate. Our glued sandwich structure renders such a simplification impossible since the layer of glue—necessarily restricted to one side of the semi-transparent film—has itself a dispersion. In order to achieve optimal compensation, the compensator must instead be made slightly thinner than the substrate.

A small difference in plate thickness is accurately achieved by cutting both components from a single work-piece into which a wedge has been polished, taking the compensator from the thinner part. Of much greater difficulty is the task of getting the glue thickness right, and as will be seen shortly we have not been very successful in this respect. Using shims was found unattractive because the thinnest shims available of 12.7 μm would dictate an unpractically large wedge.* Instead we chose to rely upon an estimated thickness of between 4 and 5 μm for an optimally squeezed-out layer of cement. A more precise method which might be looked into is to thermally evaporate 'shims', whose thickness may be controlled very accurately, directly onto the substrate.

Figure 4.8 shows the refractive indexes of fused silica and optical cement

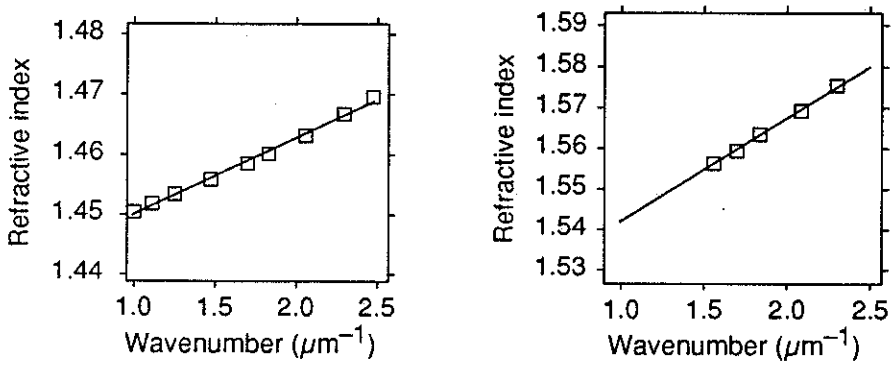


FIGURE 4.8: Refractive index of fused silica (a) and optical glue (b) plotted against wavenumber. The data fit well to the straight line model discussed in the text.

plotted versus wavenumber. Since our method of compensation can only hope to achieve a first order dispersion compensation, we have fitted these data to straight lines. Fortunately the fits are good, at least within the ranges were data has been available. These straight lines may be expressed mathematically as:

$$n_i \approx n_{i0} + a_i \sigma, \tag{4.11}$$

where the subscript i is replaced with Q for fused silica and C for the cement. The coefficients for the two materials are listed in Table 4.1.

*The wedge is controlled by observing the fringes formed by interference between the surfaces of the work-piece. Too dense a fringe pattern renders the method unpractical.

TABLE 4.1: Refractive indices of fused silica and optical cement used are found to fit well to the equation $n_i = n_{i0} + a_i\sigma$. In addition to refractive indices at $\lambda = 0.5$ μm, the table shows the results of such fits.

Material	n_i at 0.5 μm	n_{0i}	a_i (μm)
Fused silica (Q)	1.46	1.437	1.278×10^{-2}
Cement (C)	1.57	1.516	2.573×10^{-2}

Optical path difference (OPD) between the two interfering beams due to their passage through the beam splitter assembly is:

$$\begin{aligned} \Delta l &= |l_1 - l_2| \\ &= \pm[2n_Q s_{Q1} - 2(n_Q s_{Q2} + n_C s_C)] \\ &= \pm 2(n_Q\,\Delta s_Q - n_C s_C), \end{aligned} \tag{4.12}$$

where s_i represents physical path lengths and $l_i = n_i s_i$ represents optical path lengths. Using our straight line index model the OPD may be approximated by:

$$\Delta l \approx \pm 2[n_{Q0}\,\Delta s_Q - n_{C0} s_C + \sigma(a_Q\,\Delta s_Q - a_C s_C)]. \tag{4.13}$$

Hence, for it to be independent of σ, we must require:

$$a_Q\,\Delta s_Q - a_C s_C = 0, \tag{4.14}$$

and so, using the coefficients from our linear fits (Table 4.1):

$$\frac{\Delta s_Q}{s_C} = \frac{a_C}{a_Q} = 2.0. \tag{4.15}$$

Physical path length through a plate of thickness t_i depends upon the angle α_i at which it is traversed: $s_i = t_i/\cos\alpha_i$. When a ray of light changes from one material to another, its angle changes by the law of refraction, but since the indexes of silica and glue are similar (see Table 4.1), the angular change is very small (less than 10%), and the ratio of the cosines is within 3% of unity. We adopt therefore the compensation requirement:

$$\Delta t_Q = 2.0\,t_C, \tag{4.16}$$

where Δt_Q is the difference in thickness between substrate and compensator and t_C is the glue thickness.

Since it was estimated that the glue thickness would be between 4 and 5 μm, we specified the work-piece wedge to give a thickness difference of 9.0 μm between substrate and compensator (55 arc seconds, corresponding to 12 fringes of helium-neon laser light per cm or 32 fringes per inch).

That our compensation scheme has been less than successful is clear from the phase curve displayed in Figure 4.9. This curve is parabolic with a second

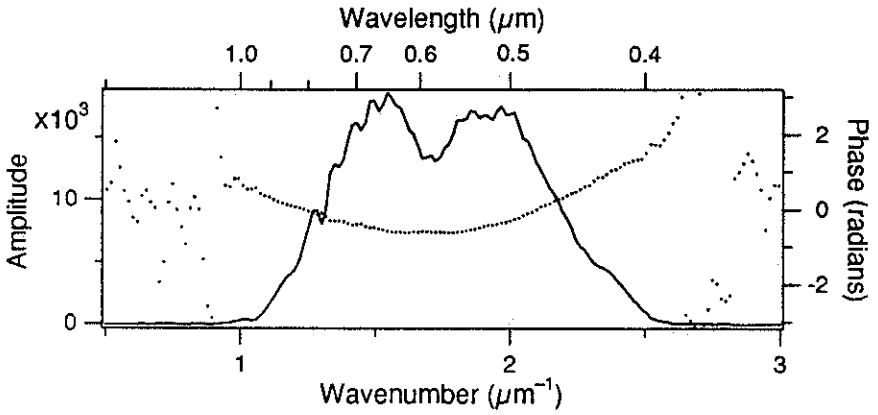


FIGURE 4.9: The spectrum of a blue sky as measured with our instrument in the low resolution mode ($\mathcal{R} = 256$). Superposed on it, plotted as dots, is the corresponding phase curve with its characteristic parabolic shape.

order coefficient of 3.37 μm^2. Optical path difference (OPD) is related to the phase by $\phi = 2\pi\sigma\Delta l$ so this phase curve signifies a linear variation of OPD with wavenumber at a slope of 0.536 μm^2. The compensation condition of Equation 4.14 is therefore violated, and by differentiating Equation 4.13, using thicknesses instead of path lengths:

$$\frac{d\,\Delta l}{d\sigma} = 0.536\ \mu\text{m}^2 = \pm\frac{2(a_Q\,\Delta t_Q - a_C t_C)}{\cos\alpha}. \qquad (4.17)$$

Solving with respect to t_C with $\alpha = 29.5°$and $\Delta t_Q = 9.0$ μmand taking a_Q and a_C from Table 4.1, we find $t_C = 13.5$ μm as an estimate for actual glue thickness.

It must be mentioned that the parabolic coefficient of phase curves from different spectra show a considerable variation, and that estimated glue thicknesses are generally greater than that found in the above calculation. An average over seven measurements thus gives $t_C = 15$ μm with a standard error of ± 10 %. We offer no rigorous explanation for this, but it may be due to

the fact that most spectra measured in the low resolution mode tend to have a Gaussian shape. Combined with the parabolic phase function, this spectral shape produces a curvature error in the phase according to Equation 3.53. The chosen spectrum (Figure 4.9) is that of the blue sky which, through the response of our instrument, is exceptionally flat over a considerable range. Our confidence in this explanation is strengthened by the considerations on channelling described in the following section.

4.2.7 Channelling

Subtracting the best fitting parabola from the phase curve of Figure 4.9 reveals the existence of a sinusoidal ripple, see Figure 4.10. The same pattern is also found in the spectrum, there even more clearly. The ripple period,

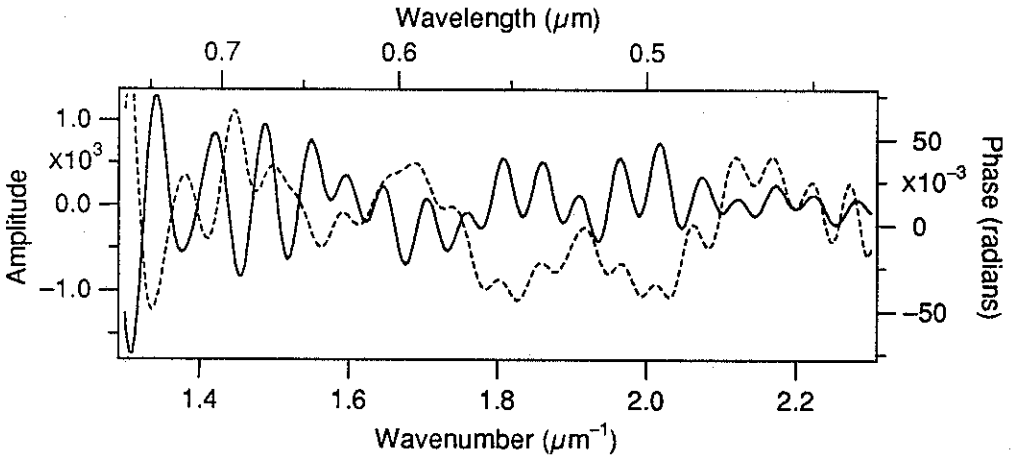


FIGURE 4.10: Spectrum (solid line) and phase (dotted line) of the same blue sky as in the previous figure but now with slowly varying features removed. Channelling is clearly present in both curves.

$\Delta\sigma_R = 0.053\ \mu\text{m}^{-1}$, is highly repeatable from measurement to measurement and between different resolutions with an error of less than three per cent. In the medium resolution mode the ripple represents only about six periods across the spectral range. It is then no longer apprec iable in the spectral amplitude, but clearly present in the phase. At the highest resolution the phase is still affected although only about one period is present.

These ripple effects are most certainly due to "channelling", a common problem in spectroscopic instruments. Due to interference between spurious

reflections, usually from the surfaces of plane parallel plates, it tends to occur where it is least expected. Figure 4.11 shows how channelling is usually

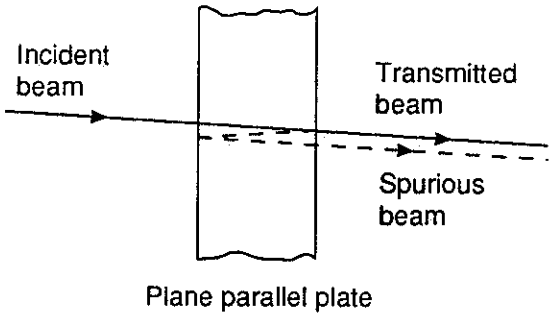


FIGURE 4.11: Spectral channelling due to spurious reflections in a plane parallel plate.

produced, giving the transmittance of a glass plate maxima when the spurious beam is in phase with the transmitted beam. If the extra optical path represented by two crossings of the plate is Δl then the period of the spectral ripple is $\Delta\sigma_R = 1/\Delta l$.

In our system, the spectral resolution is sufficiently small that we had expected all channelling to be negligible simply by using sufficiently thick optical components so that $\Delta\sigma_R$ would be less than the resolution element. Clearly, we have not succeeded, but we must seek other explanations than that given by the classical "plane parallel plate" model since the observed ripple suggests an OPD of only 19 μm, i.e. a plate thickness of 9.5 μm. Note also that the plane parallel plate explanation does not suggest a ripple on the phase curve.

The observed ripple period leads the attention to the beam splitter. Consider the spurious reflections marked 'A' in Figure 4.12. They interfere for a second time after each beam has travelled twice through its respective half of the beam splitter assembly. Since the two halves are not equal—their difference is aggravated by the poor adjustment of the glue thickness—channelling occurs and the amount of light which is transmitted through to the detector varies sinusoidally with wavenumber.

Optical path difference (Δl) between the spuriously reflected beams is given by Equation 4.12. Solving for the cement thickness gives:

$$t_C = \frac{2n_Q\Delta t_Q \mp \Delta l\cos\alpha}{2n_C}. \tag{4.18}$$

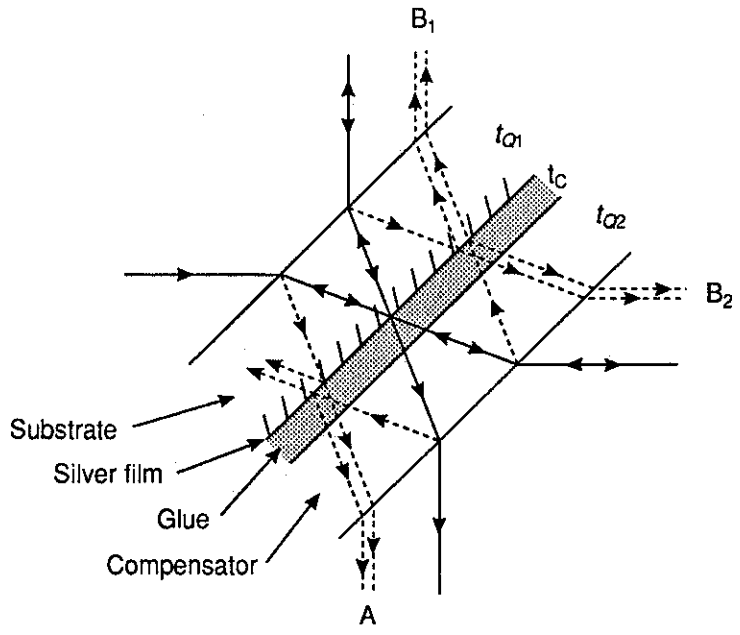


FIGURE 4.12: Suggested explanation of channelling due to the cemented beam splitter. Solid lines represent desired light beams, dotted lines represent interfering spurious beams. Those marked 'A' are supposed linked with channelled spectra, those marked 'B_1' and 'B_2' are supposedly the culprits of channelled phase.

Using the reciprocal ripple period for Δl we find two positive solutions according to the choice of sign: 3.16 μm and 13.58 μm. Arguing that the first solution is wrong since it would have given a much better dispersion compensation, we chose the second solution which, lo and behold, corresponds well with the thickness calculated previously.

Since we also see channelling in the phase curve, a kind of oscillation with wavenumber must occur in the optical path difference between the two interferometer arms. It seems that the spurious reflections marked 'B_1' and 'B_2' in Figure 4.12 may produce this effect. Here two beams are brought together to interfere *before* they are launched into the interferometer arms and we imagine that this creates a 'switching' of the light between the arms.

We note with some alarm that the observed channelling effects cannot be entirely removed in a glued beam splitter even if the glue thickness is well adjusted. This is because it is impossible to make both Equation 4.12 and Equation 4.14 vanish simultaneously. Unless some acceptable compromise can

be found, the only way to remove it (and at the same time facilitate phase compensation) is to discard the layer of cement altogether and hold the plates together by other means (optical contacting or mechanical pressure). This may render silver impractical as film material since it would no longer be hermetically sealed. Even better from a channelling point of view would be to introduce an air gap between the substrate and compensator, thus ensuring the spurious reflections to be sufficiently out of phase not to create problems. This solution definitely renders silver films impractical and it is less advantageous from a miniaturization point of view.

We must search for the bright sides of life, though! Spectral channelling is efficiently suppressed when two spectral measurements are divided (the channelled signal being proportional to the spectral signal), a process which is usually carried out anyway in order to remove the transmission coefficient of the instrument itself. It may also be used as a means of spectral calibration: in situations where a calibration source is not available the period (and phase in the heterodyned modes) of the channelled signal reveals the relationship between spatial frequency and wavenumber to an accuracy of 3%.

4.3 Fringe imaging lens

The purpose of fringe imaging is to transfer the interference pattern, localized at the interferometer mirrors, onto the detector surface. One important prerequisite for the design of a lens to perform this imaging is that it should only accept light with a direction close to parallel with the axis; i.e. it must have a *telecentric* entrance aperture. This requirement follows from the cosine dependence of optical path difference with ray angle shown in Equation 2.25: unless telecentricity is ensured, the variations in OPD becomes unnecessarily large and the symmetry of the instrument function disappears. Telecentricity is achieved by locating the aperture mask (or its image) at 'infinity', i.e. in the focus of a lens, see Figure 4.13.

Physically, the aperture may be situated either before or after the interferometer. The former is more 'correct' since it ensures identical masking of both the interfering beams. It requires an additional optical system ('fore optics')

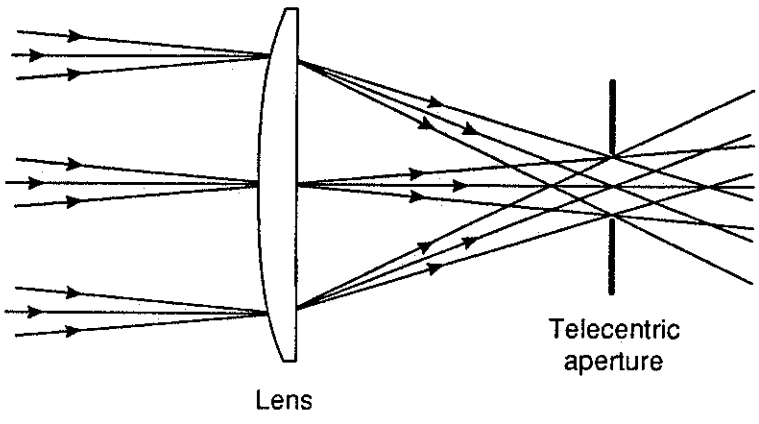


FIGURE 4.13: Illumination at near-normal incidence is ensured by the telecentric aperture, placed in the focal plane of the imaging lens.

however, and is therefore less desirable than the latter option where the aperture may be located within the fringe imaging lens. By this latter scheme the two interfering beams are masked differently resulting in reduced contrast and an asymmetry in the instrument function, but for resolving powers less than 10 000 these effects are very small.

Although the fringe pattern is two-dimensional, the along-fringe dimension does not carry any information. To optimize optical throughput the fringes should therefore be as long as possible, but this in turn requires long detector elements. A more efficient way of achieving high throughput is to collapse the fringes with the aid of a cylindrical lens. While improving throughput (as long as the numerical aperture of the cylindrical lens is larger than that of the fringe imaging lens), this also offers the opportunity for spatial resolution by using a two-dimensional array detector. This option is not exploited because of the increase in complexity required: a high-quality, possibly custom designed cylindrical lens would then be needed. Instead we have found that for the present instrument without imaging a plano-convex cylindrical singlet with a focal ratio of 2 offers sufficient quality, see Section 4.3.5.

4.3.1 Alternative fringe imaging lenses

The simplest alternative for fringe imaging is a doublet with an aperture placed in its focal plane, see Figure 4.14(a). It may be optimized for zero spherical aberration and coma, and well corrected for chromatic aberrations, but it

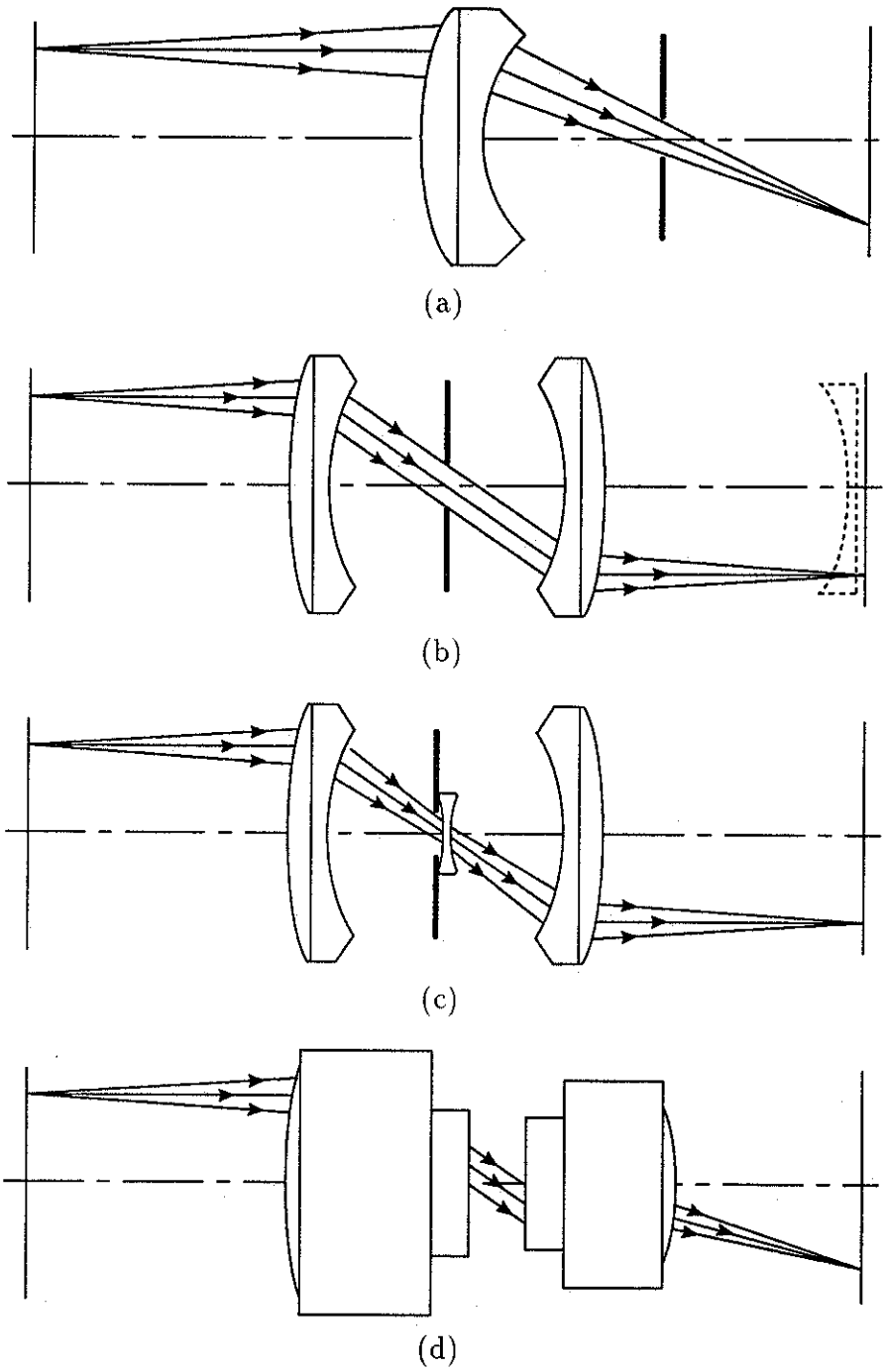


FIGURE 4.14: Alternative designs for the fringe imaging lens using specially designed doublets (a, b, and c) and off-the-shelf profile-projection and camera lenses (d).

suffers from curved image fields and distortion. Symmetrizing the design by using two doublets, see Figure 4.14(b), removes the distortion, but not the field curvature. For this to be eliminated, a negative element must be included, either in the form of a field flattener as shown in dotted line in the figure, or as a negative lens at the aperture, see Figure 4.14(c). Attempts have been made at designing such lenses, but no successful version with acceptable dimensions has been found.

Alternatively, off the shelf units such as 'profile projection lenses' may be used. Two such lenses placed back to back, or one followed by a photographic lens, see Figure 4.14(d), would probably offer more than adequate image quality. They would contain many surfaces however, causing a reduction in transmittance, and they would be very expensive.

4.3.2 The Offner lens

The lens we have found most satisfactory is an all-reflecting design consisting of two concentric, spherical mirrors one of which has a radius of curvature twice as long as that of the other [8]. As illustrated by the meridional[†] section through the lens shown in Figure 4.15, it is used off axis, the light being reflected twice off the large primary mirror (M1) and once off the secondary (M2). Since M2 is located in the focal plane of M1, an aperture mask mounted in front of M2 ensures that the lens is telecentric.

Free from all third order ray aberrations[‡], the lens is dominated by fifth order astigmatism causing the tangential focal surface[§] to have a shape given

[†]The meridional plane of an axially symmetric optical system is that containing the axis and the object point. By symmetry, the image is also contained in this plane [6, page 151].

[‡]To simplify algebraic analysis of optical systems the aberration function is usually expressed as an expansion. Ray aberration terms of first order in the object's position coordinates correspond then to paraxial optics. Even orders do not appear in axially symmetric systems, but third order terms are of great importance, often referred to as 'Seidel' or 'primary' aberrations. Fifth order terms are the most important of the 'higher order' aberrations.

[§]An astigmatic system images a point object in two *focal lines*, the *sagittal* and the *tangential*. The sagittal focal line lies in the meridional plane, the tangential focal line is normal to this plane. Their positions in the image space for varying object positions give

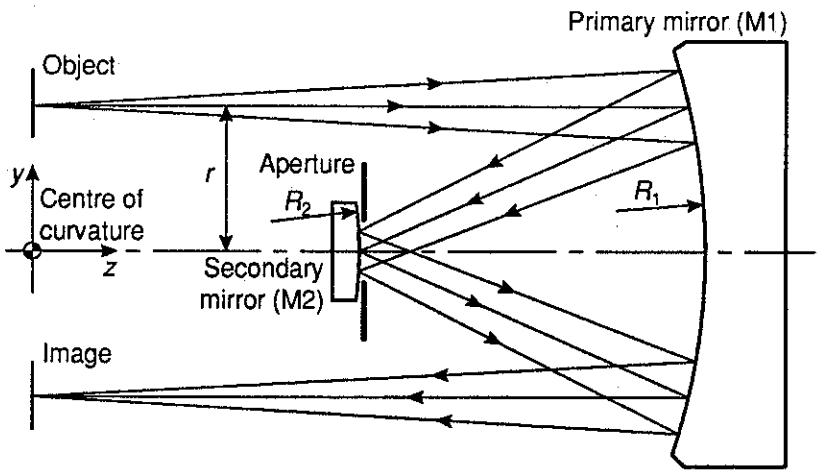


FIGURE 4.15: Meridional section of the Offner lens. With two concentric, spherical mirrors and the object (and hence image) placed in the plane containing the centre of curvature, the lens offers excellent unit magnification imaging.

by:

$$\Delta z = \frac{a_5}{R_2^3} r^4 \tag{4.19}$$

where Δz is distance from the paraxial focal plane, R_2 is radius of curvature of M2, r is the height of the object above the axis, and a_5 is a dimensionless coefficient which, by ray tracing, is found to be 0.322. The sagittal focal surface is plane, the properly focussed lens images therefore a point source as a radially directed line of length:

$$l = 2\beta\,\Delta z, \tag{4.20}$$

where β is the numerical aperture identical to the interferometer's field of view (see Equation 2.29).

Because of the one-dimensional nature of the fringe pattern the system is tolerant to curvature of the tangential focal surface as long as the fringes are parallel with the focal line. This condition is violated for fringes outside the meridional plane as shown in Figure 4.16, and a tolerance limit is therefore necessary. It seems reasonable to demand that the component l_x of the focal line in the across-fringe direction (the x-direction) should not exceed the spacing between detector elements,[¶] $\Delta x = 25\ \mu$m. From Figure 4.16 we see

rise to two surfaces, the *tangential* and *sagittal focal surfaces* [6, page 215].

[¶]H. H. Hopkins [7] has calculated the amount of astigmatism which produces a 20% loss

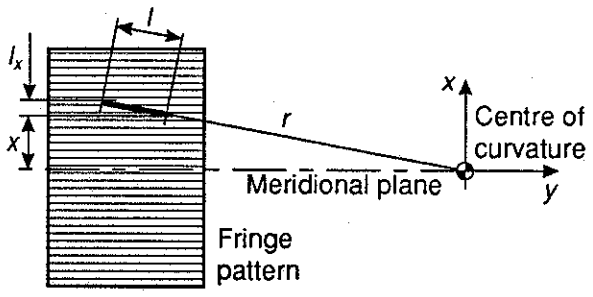


FIGURE 4.16: The astigmatic focal line is shown for a point in the fringe pattern image. l_x is its component in the across-fringe direction

that:

$$l_x = l\,\frac{x}{r} \tag{4.21}$$

hence, by Equation 4.19 and Equation 4.20, $l_x < \Delta x$ is equivalent to:

$$2\beta x a_5 \left(\frac{r}{R_2}\right)^3 < \Delta x. \tag{4.22}$$

Evaluating this inequality in the worst case we choose a point at the edge of the fringe field where $x = \Delta x\, N_D/2$. The aperture is at its greatest in the unheterodyned mode, where, by Equation 3.31 substituted into Equation 2.29: $\beta = 2/\sqrt{N_D}$. The condition of Equation 4.22 may then be rewritten as:

$$\frac{R_2}{r} > \left(2a_5\sqrt{N_D}\right)^{1/3} \tag{4.23}$$

The physical dimensions of the lens are here related to the number of detector elements so in our case, where $N_D = 512$, the R_2/r-ratio should exceed 2.4. Since our interferometer design requires a maximum value of r equal to 29 mm, the radius of curvature of the secondary mirror in the Offner lens is designed to be $R_2 = 70$ mm.

4.3.3 Manufacturing tolerances

The Offner lens is well behaved with respect to manufacturing tolerances. Simulations show that even quite large perturbations of design parameters add no other aberrations of significance than third order astigmatism and field

of contrast at a given spatial frequency. Adapted to our system with light of wavelength 0.5 μm, his calculation shows that this loss of contrast occurs for a spatial frequency equal to half the sampling frequency when the astigmatic focal line has a length of 22.4 μm.

curvature, causing changes in the curvatures of the two focal surfaces. Mathematically, we may define the shapes of the focal surfaces by the expansion:

$$\Delta z_i = {}_2C_i\, y^2 + {}_4C_i\, y^4, \tag{4.24}$$

where Δz_i is the distance between the focal surface and the paraxial image plane, and the subscript i is replaced by S for the sagittal focal surface and T for the tangential. The ${}_2C_i$ coefficients correspond to third order aberrations and denote spherical curvature of the focal surfaces; the ${}_4C_i$ coefficients correspond to fifth order aberrations and describe parabolic surface shapes. Thus the ideal Offner lens has ${}_2C_T = {}_2C_S = {}_4C_S = 0$, while, according to Equation 4.19, ${}_4C_T = a_5/R_2^3$. We may furthermore define an *astigmatic defocus* denoting the distance between the two focal surfaces:

$$\Delta z_A = \Delta z_S - \Delta z_T. \tag{4.25}$$

It may also be written as an expansion, and its coefficients, ${}_2C_A$ and ${}_4C_A$, are found by subtracting the appropriate S and T coefficients.

We have simulated the effects of errors in object distance, mirror separation, and mirror curvatures; Table 4.2 shows the observed changes in third

TABLE 4.2: Astigmatic effects of manufacturing errors. Δ_2C_S and ${}_2C_T$ denote changes in third order coefficient for the sagittal and tangential focal surfaces respectively, and $\Delta_2C_A = \Delta_2C_T - \Delta_2C_S$ is the change in astigmatic defocus coefficient. The manufacturing tolerances are set as discussed in the text.

Error (mm)		Aberrations ($\times 10^{-4}$ mm^{-1})			Manufacturing
Type	Amount	Δ_2C_S	Δ_2C_T	Δ_2C_A	tolerance (mm)
Object position	10.0	-0.646	-1.852	1.206	10
Mirror separation	3.0	6.178	17.87	-11.69	0.5
M1 radius of curv.	3.0	-6.179	-12.84	6.664	0.5
M2 radius of curv.	3.0	5.717	5.248	0.469	0.5

order coefficients for the two image surfaces and for the astigmatic defocus. Note that when a combination of errors occur, the coefficients add up. It is therefore possible to add, say, tangential field curvature while keeping the

sagittal curvature zero by increasing R_2 and decreasing the mirror separation. This allows for a balance between third order and fifth order astigmatism [8].

To set manufacturing tolerances we must consider the case when the sagittal focal surface is curved, see Figure 4.17. Following the imaging criterion used

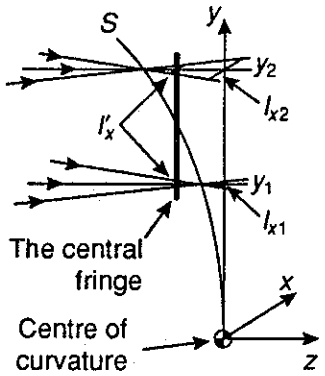


FIGURE 4.17: Optimal focussing of the interferogram's central fringe giving a maximum blur in the x-direction of l'_x at the extremes of the fringe. The y-axis coincides with the paraxial image plane and the sagittal focal surface is indicated by the line S.

in the previous section, we demand that the width of the focal spot in the x-direction never exceeds the detector spacing (Δx). If the detector is placed in the paraxial image plane (represented by the y-axis in Figure 4.17), then the width l_x of the spot at $x = 0$ varies with position as:

$$l_x = C_S \beta y^2. \tag{4.26}$$

A more optimal focus may be found where one point on the central fringe coincides with the sagittal focal surface and the two extremes of the fringe are equally far away from it (as illustrated). At the extremes of the fringe (at $y = y_1$ and y_2, say), the spot then has a width of:

$$l'_x = \frac{l_{x2} - l_{x1}}{2} = {}_2C_S \beta \, \frac{y_2^2 - y_1^2}{2}. \tag{4.27}$$

Hence, for the width to be less than the detector spacing we require $l'_x < \Delta x$, i.e.:

$${}_2C_S < \frac{2\,\Delta x}{\beta(y_2^2 - y_1^2)} = 1.30 \times 10^{-3} \text{ mm}^{-1}, \tag{4.28}$$

when $\Delta x = 25$ μm, $\beta = 0.071$ radians, $y_1 = 16.5$ mm, and $y_2 = 28.5$ mm. To allow for the effects of $x \neq 0$ and several errors added together, we have

divided this criterion by 10; the resulting manufacturing tolerances are given in Table 4.2.

4.3.4 Interferometric test

The finished lens has been tested in a Fizeau interferometer for optical quality and aberrations. Figure 4.18 shows a photo of the measurement setup and Figure 4.19 shows a typical interferogram. Apart from the slight wedge between the fringes due to astigmatism, the fringes are nice and straight, proving the high optical quality of the lens.

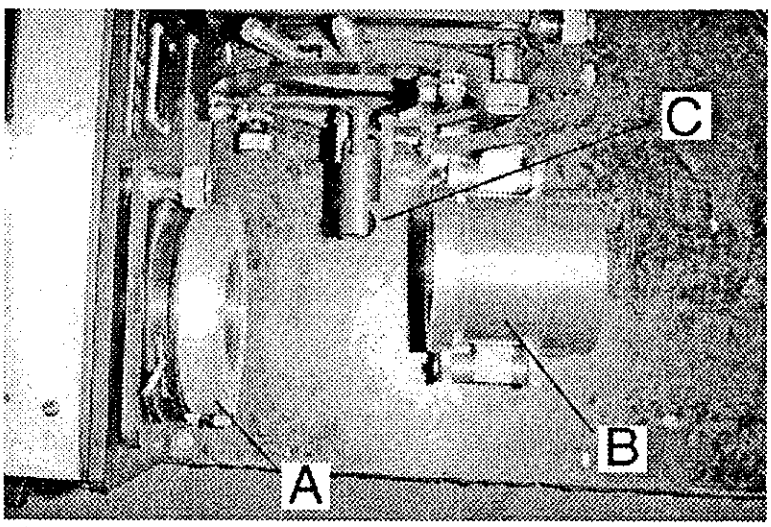


FIGURE 4.18: The interferometric test setup. A converging laser beam emerges from the Fizeau interferometer (A) and enters the Offner lens (B). In the image plane a spherical test surface (C) intercepts the beam and returns it back along its path.

Interferometric measurements do not measure the absolute position of the focal surfaces but the distance between them, i.e. the astigmatic defocus. It is therefore impossible from this test to predict the actual performance of the lens. Still, astigmatic defocus has been measured for two different object positions and a fourth order polynomial has been fitted to the values as shown in Figure 4.20. Also shown is the astigmatism expected for the ideal lens. It is clear from comparing the two curves that the manufactured lens suffers from third order astigmatism in addition to the expected fifth order astigmatism. Partly, this is due to an error of 10 mm in the axial object distance during

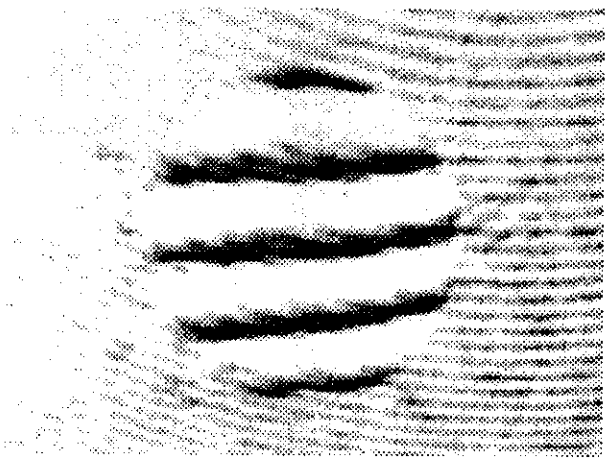


FIGURE 4.19: Straight fringes produced at an object height of 22.5 mm. The straightness of the fringes prove the good quality of the optical surfaces. Looking closely reveals a slight wedge between the fringes due to astigmatism.

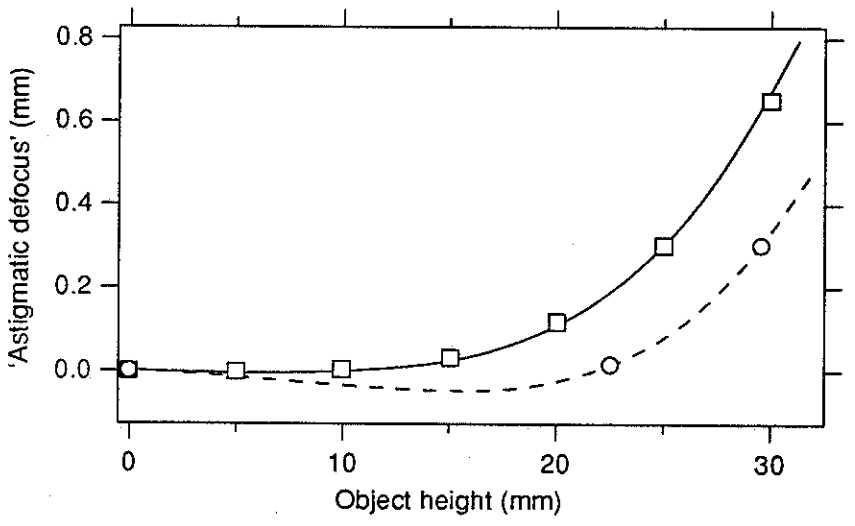


FIGURE 4.20: Comparison between simulated (solid line) and measured (broken line) astigmatism in the Offner lens. The third order shape of the measured curve is partly due to the measurement setup, partly manufacturing errors.

the test due to bulky components. Subtracting the astigmatism predicted for this error leaves a residue which is probably due to manufacturing errors. Although it is impossible to predict how the error is distributed among the design parameters, we may make an "informed guess" by assuming that mirror separation is more likely to be erroneous than mirror curvatures. We then find a separation error of 0.5 mm, which is within the specified tolerance. The lens should therefore perform according to our requirements.

4.3.5 Other optical components

Fringe imaging is also affected by the other optical components in the system, i.e. the beam splitter and the cylindrical lens. These contribute with spherical aberration, but as one is in a diverging beam and the other is in a converging beam, their contributions have opposite signs and cancel to some extent, and we find that the net effect is negligible. Aberrations introduced by the tilt of the beam splitter are also found to be small. The focussing performance of the cylindrical lens in the along-fringe direction is limited by spherical aberration, but with a blur spot of length just under 1 mm its performance is sufficient since the height of the detector elements is 2.5 mm.

The focal plane of the primary mirror, located at the secondary mirror, is also affected by spherical aberration with a blur spot of 0.1 mm diameter. This affects the accuracy with which the field limiting aperture defines the field-of-view of the instrument, but since the aperture diameter is 2 mm at a resolving power of 10 000, this blur represents at most only 5% of the field.

4.3.6 Modulation transfer function

As an ultimate test for the optical system as a whole (beam splitter and fringe imaging lens together) we have measured the system's *modulation transfer function* (MTF). This function describes the depth of modulation or contrast with which a sinusoidal intensity pattern is imaged through the system. As has been pointed out earlier, this is particularly important in our system where each spectral component is represented by a sinusoidal intensity (fringe) pattern. A reduction in contrast is equivalent to a reduction in apparent spectral intensity.

We have measured the system MTF by recording interferograms from a monochromatic source with different 'fluff' adjustments, i.e. at different spatial frequencies. From each interferogram, contrast is estimated as shown in Figure 3.1, resulting in the curve show in Figure 4.21. Clearly, the performance is rather poor with a modulation of about 0.4 throughout most of the range. The dotted line shows the MTF curve of the detector alone as measured by the manufacturer: ideally, this should be the main factor in the MTF

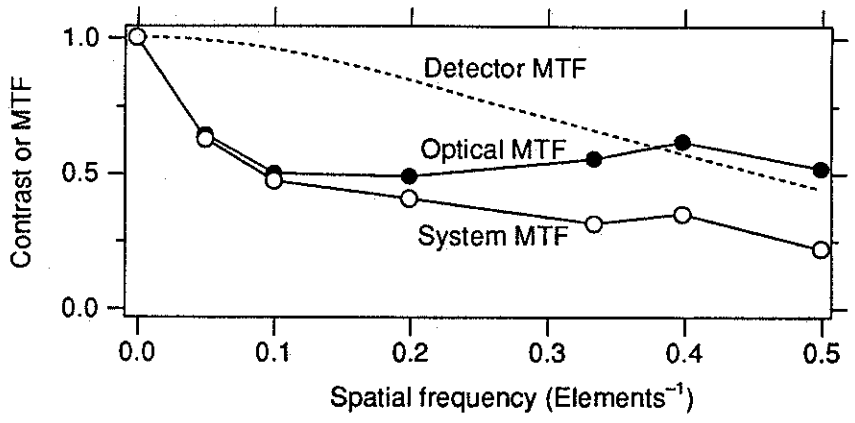


FIGURE 4.21: Measured fringe contrast plotted against fringe spatial frequency (○). The curve corresponds to the system MTF. Shown in dotted line is the detector MTF as measured by the manufacturer. Optical MTF (●) is found as the ratio between system MTF and detector MTF.

budget. Taking the ratio of these two curves[‖] gives the optical MTF due to interferometer and fringe imaging lens.

Although we expect some modulation loss due to stray light from outer surface reflections at the beam splitter (Figure 4.4) and small-scale, random homogeneity and surface errors in optical components [9], we believe that most of the loss is due to the astigmatic deformation of the interferometer reflectors. We have already discussed one serious effect of this deficiency: the spatial variation in fringe frequency appearing as a monotonic sampling error. Another effect is that the fringes to become curved. Since the pattern is then no longer one-dimensional, a one-dimensional detector array is not capable of measuring it properly. Each fringe is effectively averaged over several detectors, resulting in the observed loss of contrast.

A reduction in MTF causes a proportional reduction in the spectral signal. It does not reduce the spectral noise however, since the noise relates to an interferogram's *background* level. The SNR is therefore proportional to the MTF as will be verified in an example in Section 4.6.

[‖] An important feature of the MTF concept is that the total system MTF equals the product of the MTFs of each component of the system [9].

4.4 Mechanical design

Many details of the mechanical design have cost considerable care and consideration. This fact is of course partly—if not wholly—due to our inexperience in the work involved, a discrepancy which to some degree has been corrected. To avoid this Thesis becoming a construction manual however, the mechanical design will only be presented in the form of an overview.

4.4.1 General description

The instrument, see Figure 4.22, is contained in a box of 158 mm by 188 mm by 125 mm onto which a cylinder of diameter 110 mm and length 100 mm containing the fringe imaging lens is mounted. Although care has been taken

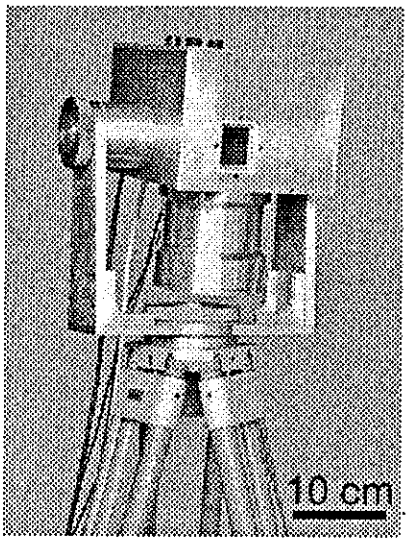


FIGURE 4.22: External view of the instrument mounted in its gimbal.

to avoid leakage of light through the cover, no particular precautions have been taken to make it moisture proof. This might be necessary in later versions, but the extra work involved has not been granted for the prototype. The assembly is suspended in a gimbal which, attached to a tripod as shown in Figure 4.22, allows the instrument to be pointed and held stable in any direction. A central hole in the tripod head allows unobstructed view straight down.

Light enters the instrument through a window in the front panel. Before entering the interferometer, it passes through a **filter compartment** with room for three 50 mm by 50 mm filters of 5–6 mm thickness. This allows for

a high-pass, a low-pass, and a band-pass filter to be employed simultaneously.

4.4.2 Interferometer unit

The interferometer unit, see Figure 4.23, is as far as possible constructed in one piece for thermal and mechanical stability. This piece has the shape of a

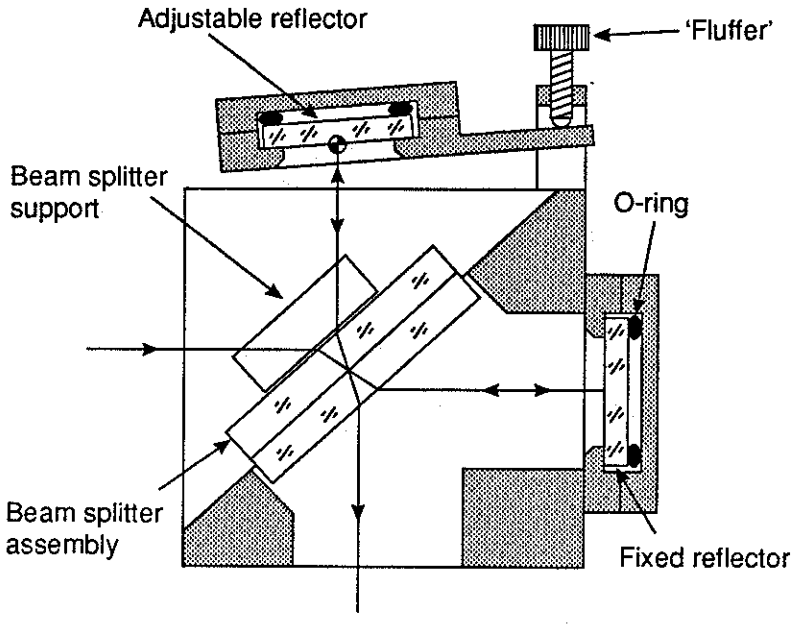


FIGURE 4.23: Cross section of the interferometer unit. It consists of a solid main block with holes machined for beam splitter support and light paths, and two reflector cells, one fixed and one adjustable.

cube in which suitable holes for light paths and beam splitter supports have been machined. The beam splitter is supported at 45 degrees to the optical axis on three points, each of which consists of a ball-bearing fixed to a shelf in the cube with a spring loaded ball-bearing directly opposite as shown in Figure 4.24(a).

A reflector cell blocks each of the interferometer arms. One of them, containing a mirror, is completely fixed and bolted onto the cube; the other, containing a mirror in the unheterodyned mode and a grating in the heterodyned modes, is adjustable and supported in a 'hole-slot-plane' (HSP) fashion. This involves three spherical support points (i.e. ball-bearings) one of which rests in a cone-shaped hole and has therefore no possibility for lateral movement, the second one rests in a V-shaped slot and is therefore allowed movement

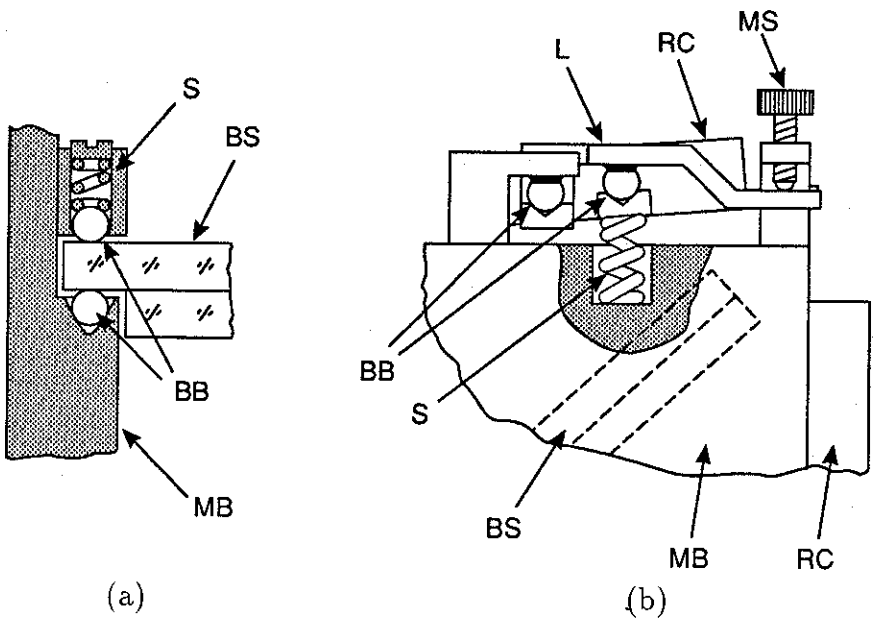


FIGURE 4.24: Details of the mechanical construction: the support system for the beam splitter (a), and the lever providing fine adjustment for fringe pattern rotation (b). S = spring, BS = beam splitter, BB = ball bearings, MB = main block, L = lever, RC = reflector cell, and MS = micrometer screw.

in one direction, and the third rests on a flat surface, having freedom in two directions. The cell is thus fixed in space at an angle to the optical axis freely adjustable by vertical movement of the support points.

Two of the support points are situated on opposite sides of the reflector such as to produce an axis of rotation located in the plane of its reflecting surface and in a direction parallel with the fringes. Adjusting either of these support points produces a rotation of the fringe pattern. Since such adjustment is generally required to be very fine, these points are connected to micrometer screws via levers, see Figure 4.24(b). HSP type mounting ensures stability and unrestrained movements of the levers as well.

Adjustment of the third support point changes tilt of the reflector and hence the spatial frequency or "fluff" of the fringes; the micrometer screw controlling this point is therefore called the "fluffer". Its movements are usually quite coarse so no lever action is required. When the instrument is used in a heterodyned mode with a grating as reflector the fluffer is adjusted to provide the required grating tilt.

4.4.3 Reflector supports

Both reflectors are supported on flat, turned surfaces and secured by the pressure of an O-ring on their back surface. This method of support was chosen because it allows the position of the mirror surfaces to be known accurately with respect to a reference surface. The choice proved to be a poor one however, since due to it the interfering wave fronts are contaminated by astigmatism which causes an apparent sampling error (Section 3.5) as well as a reduction of the fringe contrast (Section 4.3.6). Pressed against a supporting surface, the reflector substrates are forced to take on the shape of this surface; any shape error is therefore transmitted to the mirror. Shape errors in turned surfaces often result from vibrations in the lathe, or from 'warping' due to a release of stresses in the material when the work piece is machined after turning. The deficiency is serious and should be removed in future redesigns, probably by using three-point supports. This is further discussed in Section 6.4.

Aperture masks are mounted in front of the mirrors to give a rectangular fringe field, and slots for insertion of shutters are provided. Shutting off each interferometer arm in turn provides one method for measuring the background illumination, see Section 4.6.1.

4.4.4 Fringe imaging lens

In the fringe imaging lens, see Figure 4.25 the light undergoes three reflections off the two spherical, concentrically mounted mirrors. The mirrors are mounted according to a commonly used principle: a spherical surface resting on a sharp circular edge has its centre of curvature fixed with respect to the edge. The mirror substrate may thus "wander about" on the edge without affecting the direction of the optical axis and sufficient criterion for the two mirrors to be co-axial is that the support edges are co-axial. Since the mirrors face each other it is difficult to turn the edges in the same piece without changing the lathe setup. Two pieces are therefore made, fitting into each other with a 'sliding fit' which ensures the axial symmetry to within ± 0.1 mm.

O-rings press the mirrors against their support edges, following the same principle as for the interferometer reflectors. Again this might provoke some

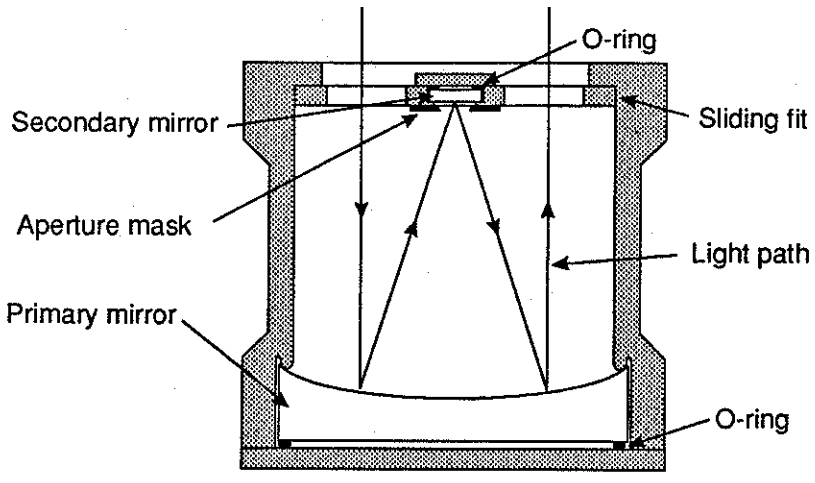


FIGURE 4.25: Mechanical construction of the Offner fringe imaging lens.

astigmatic error in the mirror surfaces, but it is not as critical here as in the interferometer since it would not affect the *shape* of the interference pattern. Instead it would modify the (already astigmatic) aberration function of the lens with an essentially field independent astigmatic component. Such an effect was not seen during the interferometric test of the lens (Section 4.3.4) and we assume it therefore to be negligible.

The telecentric aperture mask (see Section 4.3) is mounted in front of the secondary mirror. It is accessible only by removing the primary mirror from its support cell, a process which is less than optimal because it carries the risk of damaging the mirror. A re-design should take care of this problem by changing the way in which the two main pieces of the lens are joined.

4.4.5 Detector housing

Both interferometer and lens are bolted onto the 'floor' of the instrument case; interferometer on the upper side, lens on the lower side. Light enters the lens through a hole in the floor and exits through another hole displaced by 45 mm. It then enters the detector housing, a little box mounted via a slide to the side of the interferometer unit. The slide allows focussing of the interferogram on the detector array, but this is considered only to be necessary as a one-off adjustment to take up manufacturing errors. The cylindrical fringe-collapsing lens is mounted at the entrance of the detector house. Since the detector

elements are 2.5 mm tall, the focussing tolerance of this lens is ±1 mm which is assumed achievable without the need for adjustments. In the other end of the detector house a printed circuit board (PCB) is mounted carrying the detector chip itself.

4.5 Electronic design

The electronic design for the instrument has been left in safe and far more experienced hands than ours. Although it is therefore not strictly part of the work done for this thesis, completeness requires an overview of it as well.

4.5.1 General description

A purpose-built, microprocessor-based electronic system for control and data-logging has been constructed. Apart from its power source, the system including memory to hold about 100 interferograms is fully contained within the instrument. Powered by a small battery pack, it is capable of automatic stand-alone operation. Interrogation and programming requires a suitable terminal such as a portable personal computer with which communication is ensured through a standard serial port.

The electronics is distributed on a number of printed circuit boards. Care has been taken to screen all analogue signal paths and to separate physically analogue and digital circuitry in order to control noise due to interference from high frequency digital signals.

4.5.2 Detector control

The detector chip contains a one-dimensional array of photo diodes. It is a 'switched access array' and should not be confused with the more familiar CCD (charge coupled device) array where the charge packets collected by each diode are read out strictly serially in a conveyer belt fashion. Instead, all the cells are connected via switches to a common 'video' line and may be interrogated randomly, although in practice the cells are accessed one after the other in a well defined, regular sequence called the *readout cycle*.

Each photodiode consists of a "bar" of p-type silicon embedded in an n-type substrate as shown in Figure 4.26. The equivalent electronic circuit for this

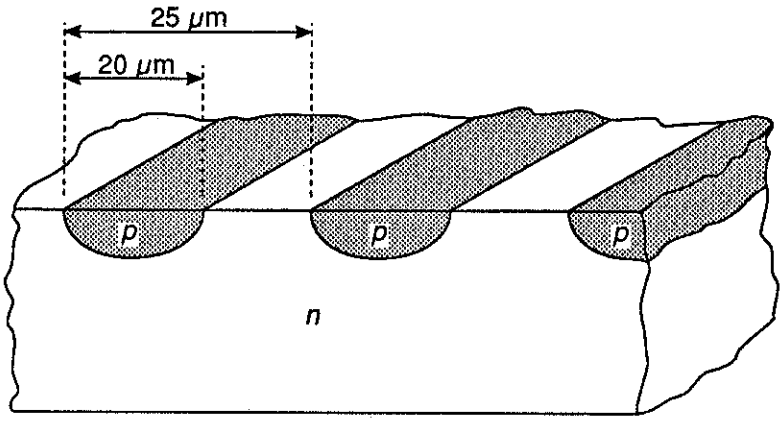


FIGURE 4.26: Cut through the diode array chip showing the bars on p-type silicon embedded in an n-type substrate.

structure is a parallel connection of a diode and a capacitor, see Figure 4.27(a). The capacitance is inherent in the diode construction, but contrary to usual

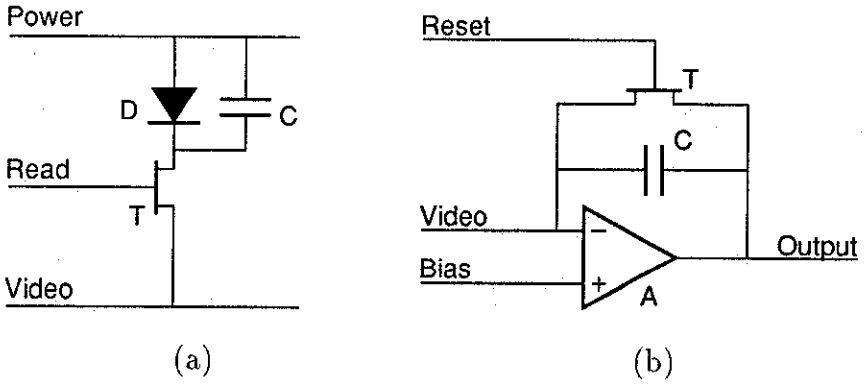


FIGURE 4.27: Circuit diagrams for two key components in the electronic design: the detector diode cell (a), and the charge integrator (b). D = diode, C = capacitor, T = transistor switch, and A = operational amplifier.

practice where it is minimized, it is instead optimized and serves as charge integrator. When a photon is absorbed by the diode the capacitor increases its charge by one electron.

Interrogation of a diode in the array after a suitable exposure proceeds by closing the "read" switch connecting that diode to the video line. Its contained charge is thus transferred to a charge-to-voltage converter consisting of an operational amplifier with a capacitor connected in negative feed-back,

see Figure 4.27(b). This configuration translates the charge very accurately into a proportional voltage suitable for input to an analogue to digital converter (ADC). After conversion the integrator is reset by shorting its feedback capacitor. This is achieved under logic control by closing a transistor switch connected in parallel with the capacitor.

4.5.3 Timing

The process of opening and closing the switches required for measuring the charge collected by each diode cell is controlled by a logic sequence called the *diode cycle* whose timing diagram is shown in Figure 4.28. A diode cycle lasts six clock cycles, so with a clock frequency of 154 kHz its duration is 39 μs.

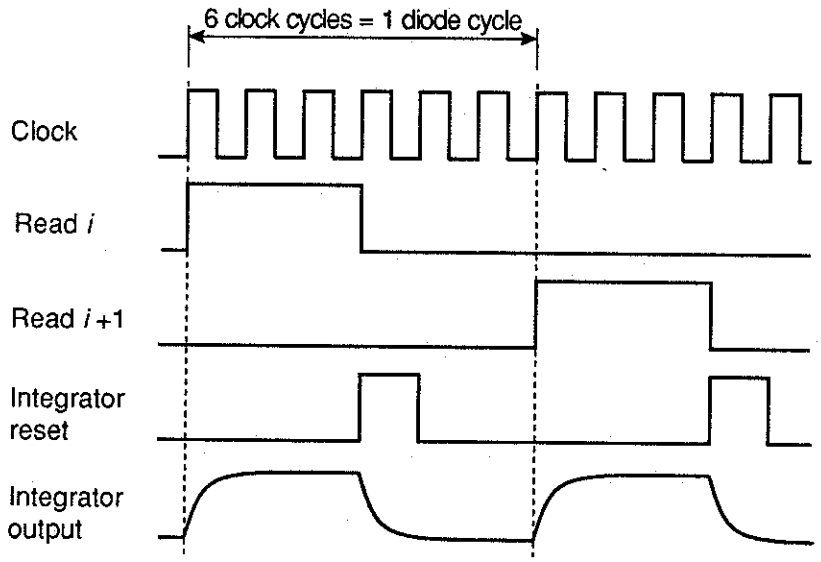


FIGURE 4.28: Timing diagram for the interrogation of diode cells. 'Read i' is the read pulse for the ith detector element. It lasts three clock cycles and causes all the charge collected by that detector to be transferred to the voltage convertor. This also serves to reset the charge to zero thus preparing for a new exposure. The output of the convertor is sampled during the third clock cycle and reset by the 'Reset' pulse during the fourth clock cycle. The system is then allowed two clock cycles to stabilize before the next diode is interrogated.

The readout cycle for the 512 element array lasts 515 diode cycles including starting up and closing down procedures, i.e. 20 ms.

An *exposure cycle* contains two readout cycles separated by a time delay, and the *exposure time* (τ)—the time allowed for each diode to collect photons—

is equal to the time between the start of these readout cycles. Hence the minimum exposure, called an *exposure unit* (τ_0), has the length of a readout cycle. It is in practice convenient to measure the exposure in logarithmic units; we introduce therefore the *exposure step*, $\mathcal{E}$, defined by:

$$\mathcal{E} = \log_2 \frac{\tau}{\tau_0}. \tag{4.29}$$

A change of one exposure step doubles the collected radiation and is equivalent to 'stepping the aperture one stop' in traditional photography. The system offers exposure times in multiples of the exposure unit up to a maximum of 256 units (about 5 seconds) covering a range of 8 steps. Longer exposures are possible by using the real-time clock included in the system. In practice, the exposure time is limited by the dark current in the detector diodes which at room temperature saturates the capacitors after 3 minutes ($\mathcal{E} = 13$). The dark current is highly temperature sensitive however, and at 0°C the dark saturation time is increased to almost 3 hours representing a range of 29 exposure steps. Operation of the instrument could then be pushed into conditions of deep twilight (Section 2.5). With the current design, such operation depends upon natural cooling from ambient temperature. Artificial cooling is possible, but it has not been implemented because of its high power consumption.

4.5.4 Data logging and power

The analogue to digital converter has a word length of 14 bits giving it a dynamic range of 16384 steps. Its linearity is guaranteed to better than half a least significant bit thanks to a self-calibration facility, and its noise is dominated by the digitizing noise.

Superior control, storage of data, and communication with the user (represented by the personal computer) is provided by an on-board micro processor system. Among its features are:

- Standard serial communication via RS-232;
- Capability of simultaneous collection of up to eight scans;
- Arithmetic unit for simple mathematical operations on these scans, e.g. adding them all up, thus increasing the precision to 16 bits;

- Storage space for about 100 scans including a header with circumstantial information such as date, time, and exposure; and
- Real time clock for timed operation and referencing of scans.

Power is supplied from two 12 V rechargeable lead acid batteries, each of 1.2 Ah capacity. They are connected to give $\pm$ 12 V as required by the operational amplifiers, and the +5 V required for digital circuits is derived from +12 V. Since the detector timing circuit consumes considerably more power than the rest of the system, a relay has been incorporated to put the detector to 'sleep' when not in use. The lifetime of fully recharged batteries is thus increased from 7.5 hours under full operation towards 40 hours under minimal operation. A future redesign of the timing circuit is expected to reduce its consumption considerably, thus allowing a corresponding increase in battery lifetime during full operation.

4.6 Signal processing

It has been within the scope of our work to investigate signal processing methods for the instrument. An experimental processing environment has been created with the aid of a general purpose mathematics programme in which macros and functions are written in a high level 'language'. We present here the various processing elements involved and discusses alternative implementations.

Signal processing for Fourier transform spectroscopy is naturally divided into two parts separated by the Fourier transformation: *space domain* (or time domain in classical FTS) and *frequency domain* calculations. Some calculations may equally well be performed in either domain, however. Apodization is an example of this where the 'tailoring' of the instrument function achieved by multiplying the interferogram with a smoothly tapered truncation function may just as well be done in the frequency domain by a convolution between the unapodized spectrum and the appropriate instrument function. Phase correction of double sided interferograms is another example although here the situation is inverted: the simple multiplicative correction performed in the fre-

quency plane may be substituted by a convolution in the interferogram plane. Some processes also cross the boundary between the domains within themselves. During phase correction, for example, the apodized spectrum which is required for determination of the phase function is obtained by going back to the interferogram and performing an additional, apodized, transformation.

The processing is generally done in two main steps: background correction, and transformation with phase correction. Interferograms taken at long exposures are additionally corrected by dark current subtraction, and, in appropriate cases, the interferogram is resampled before transformation. We describe the operations performed in each step and illustrate them with examples. At the end of the section a description of the present, experimental operator interface is given including remarks on practical aspects and recommendations for an operative version.

4.6.1 Interferogram correction

Raw interferograms measured with our instrument show some distinct features, see Figure 4.29(a):

- Offset from zero,
- "Waviness", and
- High frequency ripple.

Apart from the offset which is a natural attribute of the interference pattern, these features are caused by instrumental deficiencies such as vignetting, dust, variations in detector sensitivity, etc., and they have seriously detrimental effects upon the spectral estimate as seen in Figure 4.29(b). The tall spike close to the origin is due to offset and waviness, the poor signal to noise ratio (SNR) is caused by the ripple.

Although the waviness varies somewhat from target to target—probably due to variations in the illumination geometry affecting stray light and vignetting—we find that the ripple is virtually constant from one interferogram to another, andt is therefore possible to remove it. The literature suggests two methods for measuring the background alone by removing the interferometric

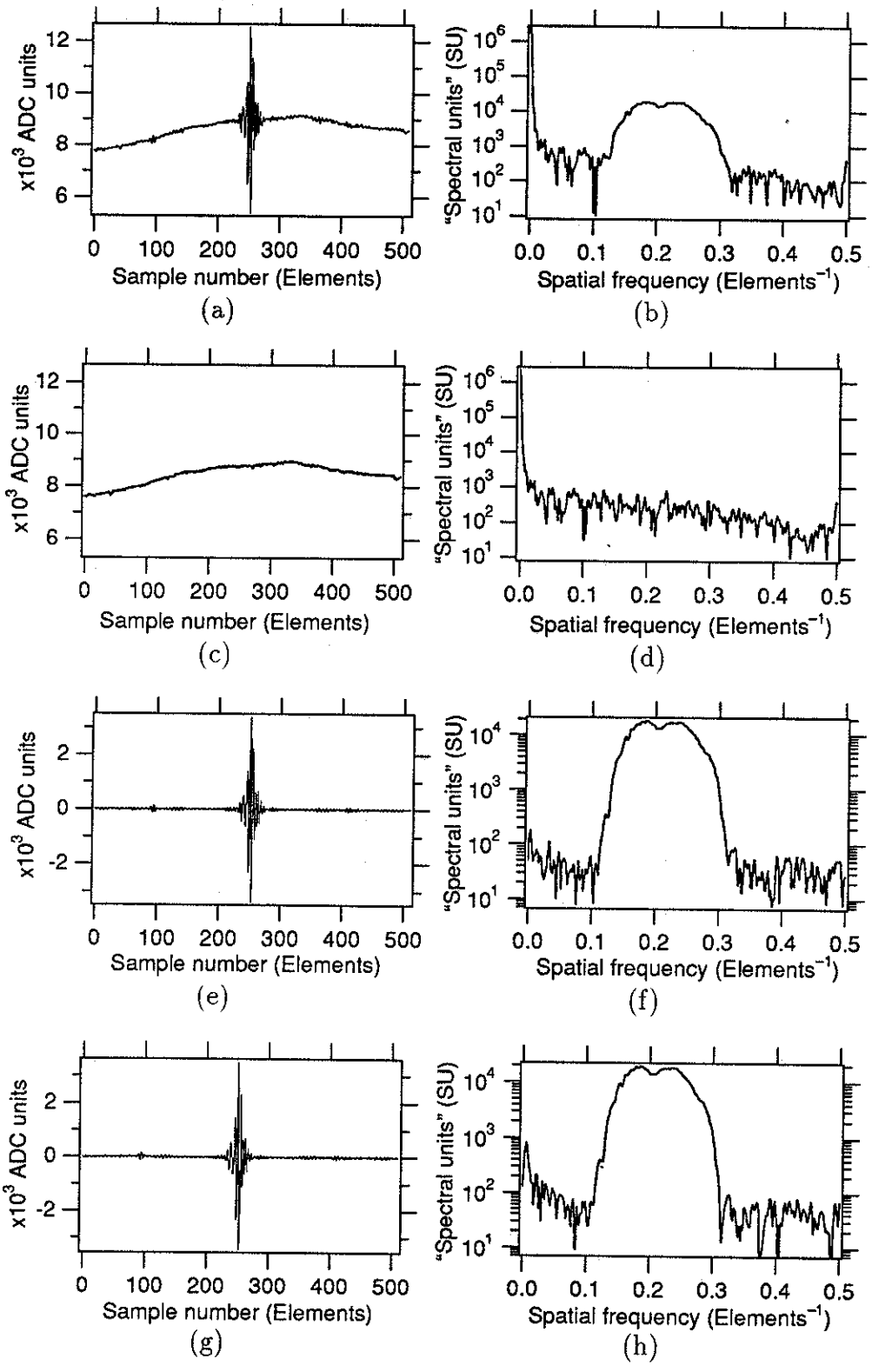


FIGURE 4.29: Demonstration of the interferogram correction procedure. The left hand column show interferograms in various stages of the correction and the right hand column shows the complex amplitude of their corresponding Fourier transforms (not phase corrected) plotted on a logarithmic scale. See text for explanation.

information. One method involves the use of two measurements where each of the interferometer arms has been blocked off in turn [61]. By the other method the interferometer is adjusted so as to give the fringes a tilt with respect to the detector array: each detector then sees an average over several fringes, leaving them essentially invisible [56].

Mechanically, we have allowed for both methods, but since the former seems likely to introduce more uncertainties than it removes (not only will the shutters never be completely black, but each time they are inserted or removed they will probably disturb particles of dust inside the interferometer and may even introduce new ones) this method remains untested. Instead we use to great satisfaction the second method by which the background shown in Figure 4.29(c) has been measured. Its Fourier transform (d) displays the same features as the uncorrected spectrum, apart, luckily, from any of the spectral signal. Correction proceeds now in principle by dividing the measured interferogram (I_M) by the background (I_B) and subtracting unity to give an improved estimate of the active interferogram:

$$\mathcal{I}_E = \frac{I_M}{I_B} - \mathcal{I}_0, \tag{4.30}$$

where $\mathcal{I}_0 = 1$. In practice the process is somewhat more involved: in order to take into account differences in illumination between the two measurements, it is the *average* of their ratio which is subtracted instead of unity: $\mathcal{I}_0 = \overline{I_M/I_B}$. The result is afterwards multiplied with the average of I_B in order to give a scaling of the corrected interferogram similar to that of the measured interferogram. Figure 4.29(e) shows the result of these operations and (f) shows the spectral estimate produced from it.

When interferogram and background have been measured with different targets, the corrected interferogram tends to be affected by a slope and some residual waviness. We reduce these deficiencies quite efficiently by letting $\mathcal{I}_0$ be a polynomial curve fitted to the I_M/I_B-ratio. The interferogram in Figure 4.29(g) is the same as in (e) but corrected with a different background. Again the spectrum has a good reduction of the high frequency ripple while leaving a somewhat higher noise level in the low frequency range, see Figure 4.29(h). Considering the simplifications in measuring technique conferred

by the use of a standard background (no need for readjustment of the interferometer nor recalibration for every new target) this loss seems fully aceptable.

4.6.2 Noise evaluation

Background correction has reduced the RMS noise level in this example by a factor of about 20 in the low frequencies, and about 2 in the high frequencies. The RMS noise level in Figure 4.29(f) is thus brought down to about 30 SU (spectral units**) across the range from about 0.05 Elements^{-1} to 0.5 Elements^{-1}. According to the Fourier power theorem (Equation 3.81) this corresponds to an average RMS noise level in the interferogram of $\epsilon_x = \epsilon_\nu/(\Delta x\sqrt{N}) \approx 1$ ADC unit with 512 interferogram samples.

Signal-to-noise ratio (SNR) reaches in this example a peak value of about 650 at the highest spectral point of 20 000. In Equation 3.88 the spectral SNR of HFTS instruments was estimated to $\mathrm{SNR}_\nu = \sqrt{(I_{\mathrm{Sat}}/N)}/f$, where the square root equals $\sqrt{(1.2\times10^8/512)} \approx 500$ and $f = \overline{B}/B_\nu$ is the spectral fill factor. With a fill factor of 0.2 in the current example, the ideal SNR is thus 2500, i.e. roughly four times the measured SNR. This comparison is incorrect because of three factors: (1) the measured spectrum has not been phase corrected, (2) it is the result of two independent measurements (interferogram and background), and (3) it has been apodized. While the two first factors cause a reduction in the measured SNR of $1/\sqrt{2}$ each, the third factor improves it somewhat. Assuming that improvement to represent about a factor $\sqrt{2}$ as well, the ratio between measured and ideal SNR should be corrected by the remaining $\sqrt{2}$, giving approximately 0.4.

Note that this factor corresponds to the system MTF presented in Figure 4.21; the stated relationship between modulation transfer and noise performance is hence verified. Note also that the measured interferogram itself bears witness of the poor modulation: the ratio between its peak value and its average background level is $3500/8500 \approx 0.4$.

**The spectral unit denotes power per spectral interval. Calibrated, this should be given in watts per micron (W $\mu\mathrm{m}^{-1}$) or watts per inverse micron (W $\mu\mathrm{m}$). Uncalibrated however, the unit of power is the ADC unit and the spectral unit is the inverse diode array element separation. Hence: 1 SU = 1 ADC-unit Elements.

4.6.3 Dark signal subtraction

For long exposures, the dark signal becomes significant. Dark signal contributes to the noise by two different mechanisms: Shot noise (equal to the square-root of the number of dark electrons) and diode-to-diode variations. The former contribution is indeterminate and can not be removed, but the latter, referred to as "dark noise", is constant (or almost constant: a slow, '$1/f$'-type, variation is expected [52]) and is efficiently removed by subtracting a "dark scan" (a scan taken with the aperture covered) from both interferogram and background measurements. The RMS level of the dark noise, denoted ϵ_d, is proportional to the dark current I_d:

$$\epsilon_d = pI_d, \tag{4.31}$$

where $p \approx 6\times10^{-3}$ according to our measurements, is the proportionality factor.

To ensure that dark noise never affects our measurements we demand that a dark signal be subtracted whenever the dark noise exceeds the digitizing noise, ϵ_D. Hence: $\epsilon_d > \epsilon_D$, which, by Equation 3.79 becomes:

$$pI_d > \frac{0.3I_P}{D} \tag{4.32}$$

where I_P is the peak signal and D is the digital dynamic range. Dark signal measured in number of electrons may be expressed in terms of the dark current i_d as: $I_d = i_d\tau/e$, where τ is exposure time and e is the electron charge, and so the dark correction condition may be rewritten as:

$$\tau > \frac{0.3I_P}{D}\frac{e}{i_d p}. \tag{4.33}$$

At 20°C ambient temperature, $i_d = 0.045$ pA according to Figure 2.12, so for an operating point (see Section 3.4.3) adjusted to $I_P = 10^8$ and 14 bit digitization, dark correction is required for measurements with exposures longer than 1.2 seconds ($\mathcal{E} = 5.8$). Note that at this exposure the dark signal is less than 1% of the peak signal.

Figure 4.30 demonstrates the effect of dark correction showing a 'raw' dark scan measured at an exposure of $\mathcal{E} = 8$ (a) and the difference between two

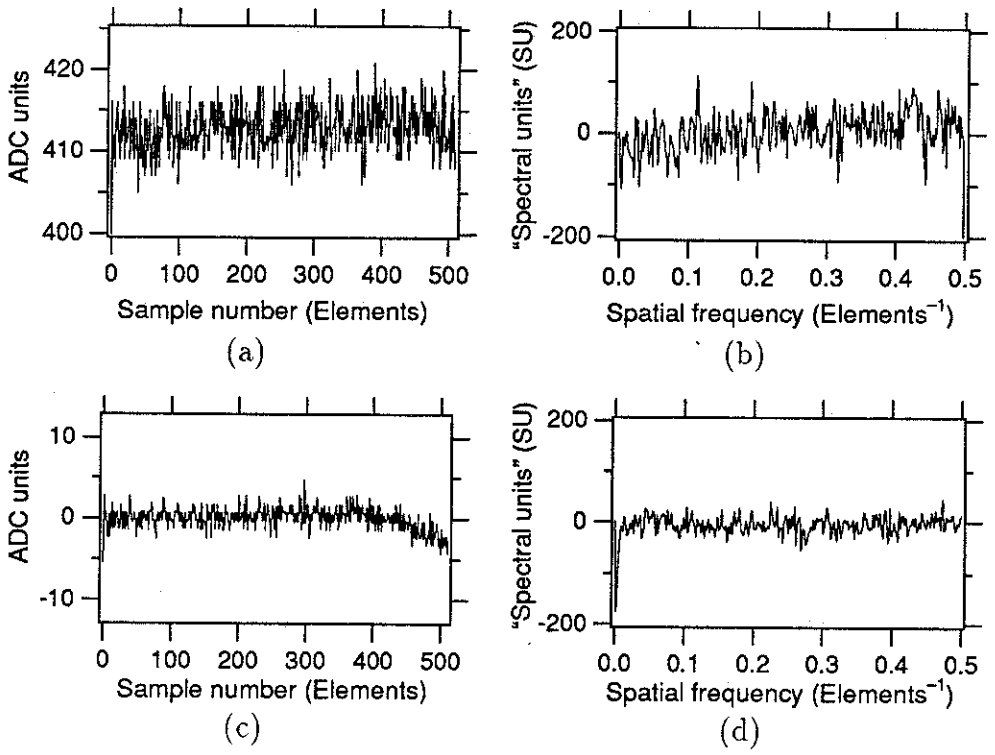


FIGURE 4.30: A "raw" dark scan (a) and its Fourier transform (b) compared with the difference between two dark scans (c) and its transform (d).

dark scans (c) together with the real part of their Fourier transforms (b and d, respectively). Note that for the raw scan the mean has been subtracted before transformation in order to avoid the spike at zero frequency. The noise reduction is dramatic, leaving a RMS noise in the difference scan of about one ADC unit. The RMS noise in the corresponding spectrum is about 16 SU, as predicted by Equation 3.82.

The remaining noise compares well with what we predict from the treatment of noise in Section 3.4.1. We see from Figure 3.13 that with 14 bit digitization and an operating point at $I_P = 10^8$, the resetting noise and digital noise are approximately equal: $\epsilon_R \approx \epsilon_D \approx 0.3$ ADC units. In addition there is the shot noise of the dark current, $\epsilon_S = \sqrt{I_d}$, equivalent to 0.24 ADC units for the present measurement. These noise components add together in a root-sum-square fashion, not forgetting to count each component twice since we have combined two measurements, to give an expected noise level of 0.7 ADC units. When the measured noise is slightly higher, it may signify that the detector resetting noise has been underestimated.

Note the sharp spike (negative in this plot) at the highest frequency ($\nu =$ 0.5 Elements^{-1}) in Figure 4.30(b). It is due to a ripple in the detected signal of peak-to-peak amplitude 1.2 ADC units with a very accurate period of two samples. Present in all our measurements, it is probably caused by 'leakage' from a digital signal. It is harmless, however, since it is confined to one single frequency component situated at the extremity of the spectrum, and it disappears entirely when the difference between two scans is taken.

4.6.4 Interferogram resampling

An astigmatic aberration in the interferometer causes an apparent error in the sampling of interferograms, as described in Section 3.5. The error is particularly troublesome in unresolved emission spectra, but it is also noticed in unresolved absorption spectra, e.g. the solar spectrum measured in the instrument's high resolution mode. We describe a method by which it may be measured and corrected for in software. Two examples, the sodium doublet and the spectrum of a fluorescent "daylight" tube, are given.

Measuring the sampling error. Measurement of the error is possible by comparing the measured interferogram for a known spectral distribution with its theoretically predicted interferogram. This is particularly easy when the source has a single, unresolved spectral line. Then the active interferogram is given by:

$$\mathcal{I}(x) = \cos 2\pi\nu_1 x, \tag{4.34}$$

where ν_1 is the spatial frequency representing the optical frequency of the line. When a sampling error is present, the sampled version of the interferogram must, according to Equation 3.91, instead be written:

$$\begin{aligned} \mathcal{I}_\varepsilon(x) = \mathcal{I}(x+\varepsilon) &= \cos 2\pi\nu_1[x+\varepsilon(x)] \\ &= \cos[2\pi\nu_1 x + \theta_\varepsilon(x)], \end{aligned} \tag{4.35}$$

where:

$$\theta_\varepsilon(x) = 2\pi\nu_1\varepsilon(x) \tag{4.36}$$

is the phase of the sampling error. Note that this phase is *spatially* varying (across the interferogram) rather than *spectrally* varying as the dispersive phase described in Section 3.3.

With analogy to radio theory we may regard the erroneous interferogram as a carrier signal of frequency ν_1 phase modulated by the function $\theta_\varepsilon(x)$ [2, page 577]. Fourier transforming the interferogram thus produces the transform of the modulation function convolved with a pair of δ-functions at $\pm\nu_1$. Since the modulation function is slowly varying, its transform is quite narrow; the resulting spectrum therefore consists of two peaks, identical apart from a difference in sign in the imaginary part. Demodulation may now proceed by isolating one of the peaks and shifting it to the position of zero frequency. Transformed back into the spatial (interferogram) domain, this produces a complex function whose argument is θ_ε.

This may be proved mathematically by rewriting Equation 4.35 on the form:

$$\mathcal{I}_\varepsilon(x) = \cos 2\pi\nu_1 x \cos\theta_\varepsilon - \sin 2\pi\nu_1 x \sin\theta_\varepsilon \tag{4.37}$$

whose Fourier transform is:

$$\begin{aligned}\mathcal{F}\{\mathcal{I}_\varepsilon(x)\} &= [\delta(\nu-\nu_1)+\delta(\nu+\nu_1)] \star \mathcal{F}\{\cos\theta\} \\ &\quad -i[\delta(\nu-\nu_1)-\delta(\nu+\nu_1)] \star \mathcal{F}\{\sin\theta\},\end{aligned} \tag{4.38}$$

where $\delta(\nu)$ denotes the Dirac delta function. Rearranging and using exponential notation this expression may be written as:

$$\begin{aligned}\mathcal{F}\{\mathcal{I}_\varepsilon(x)\} &= \delta(\nu-\nu_1) \star \mathcal{F}\left\{e^{-i\theta_\varepsilon}\right\} \\ &\quad +\delta(\nu+\nu_1) \star \mathcal{F}\left\{e^{+i\theta_\varepsilon}\right\},\end{aligned} \tag{4.39}$$

as required.

In practice the isolation of one of the peaks (the positive) is done by multiplication with a bell-shaped function to avoid ripple in the estimate of θ_ε. The technique works very well and without intervention as long as θ_ε never exceeds $\pm\pi$, in which case a discontinuity occurs. Algorithms which can handle such occurrences may be devised but none has as yet been implemented. If discontinuities do occur, they must be corrected manually.

Correcting the sampling error. Correct sample values may now be found by interpolation between existing samples. Theoretically this is permissible as long as the interferogram is sampled according to the sampling theorem. Since interferograms usually are sampled close to critically (two samples per period), however, simple linear interpolation is unreliable. Again a recourse is made to the Fourier transform: the transformed interferogram is extended (or 'padded') with zeros to, usually, four times its length before it is transformed back into interferogram space. This produces three new samples between each original pair of samples, and further interpolation may now be simply linear. Nonlinear interpolation would probably improve the accuracy of the resampling even further, but since the linear method gives satisfactory results for demonstration purposes we have not implemented such refinements at this stage.

As pointed out in Section 3.5 this straightforward resampling process is strictly only valid for unheterodyned spectra. It does improve narrow-band heterodyned measurements however, as demonstrated by the second example below. We think it is possible to achieve resampling of broad-band heterodyned measurements as well by a different method, but this has not yet been tried.

4.6.5 Resampling examples

A fluorescent "daylight" spectrum Improvements in an unheterodyned broad-band spectrum by resampling its interferogram is well demonstrated by considering the spectrum of a fluorescent "daylight" tube (standard Blackett Laboratory illumination). As seen in Figure 4.31 the spectrum consists of a 100 nm wide continuous band centred at 600 nm with two strong peaks on its blue flank. This version of the spectrum is calculated by apodizing the interferogram and the effect of resampling is therefore not very great. For the unapodized version however, the effect is evident as seen in Figure 4.32. In the original version (a), both peaks appear as doublets, and the imaginary part contains a considerable signal. Note the similarity between these peaks and those shown in Figure 3.14(c), found from the simulation model. After resampling, Figure 4.32(b), the real peaks have refound their ideal sinc-shape and the imaginary parts are significantly reduced.

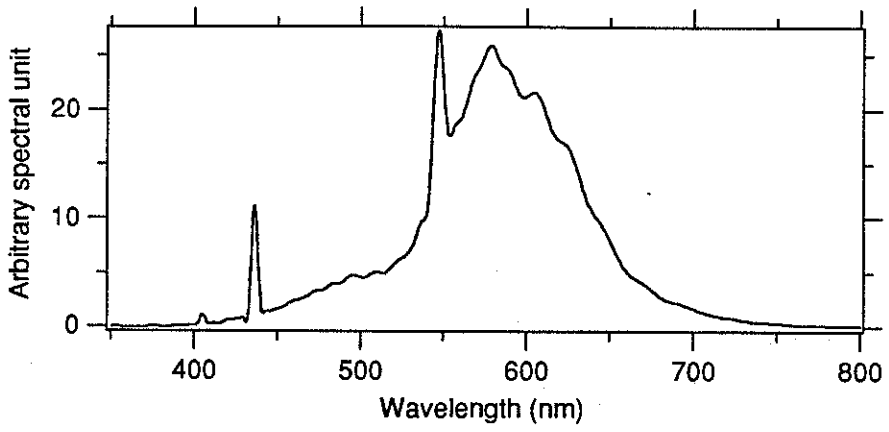


FIGURE 4.31: Apodized version of the fluorescent lamp spectrum.

The apparent sampling error has been measured from the interferogram of a 670 nm red diode laser which at this resolving power appears strictly monochromatic. As seen in Figure 4.33 the apparent sampling error function is smooth and follows a parabolic shape as expected. At one edge of the interferogram the error reaches 1.5 samples, representing a considerable distortion of the measurement.

Example 2: The sodium doublet. Good improvement of the spectral estimate has also been achieved by resampling the interferogram of narrow-band, heterodyned spectra. This is clearly demonstrated on a high-resolution sodium spectrum, see Figure 4.34. The dotted and solid traces represent the spectrum before and after resampling, respectively. No apodization was applied so the ideal instrument function is the sinc function (Figure 3.4) with a distance between the peak and the first zero crossing equal to 1/(512 Elements) $= 1.95\times10^{-3}$ Elements^{-1}. The corrected version comes close to this ideal with a peak-to-first zero distance equal to 2.4×10^{-3} Elements^{-1}. In contrast, the uncorrected spectral instrument function is twice as wide and has a marked double peak as predicted by theory (Figure 3.14).

Once again a parabolic shape due to astigmatism in the interferometer dominates the apparent sampling error, see Figure 4.35. Now a ripple is also clearly present however, and isolating the ripple by subtracting a best fitting paraboloid reveals a quasi-sinusoidal function with a period of about 38 Elements, corresponding to about 1 mm. It is unlikely that an aberration in

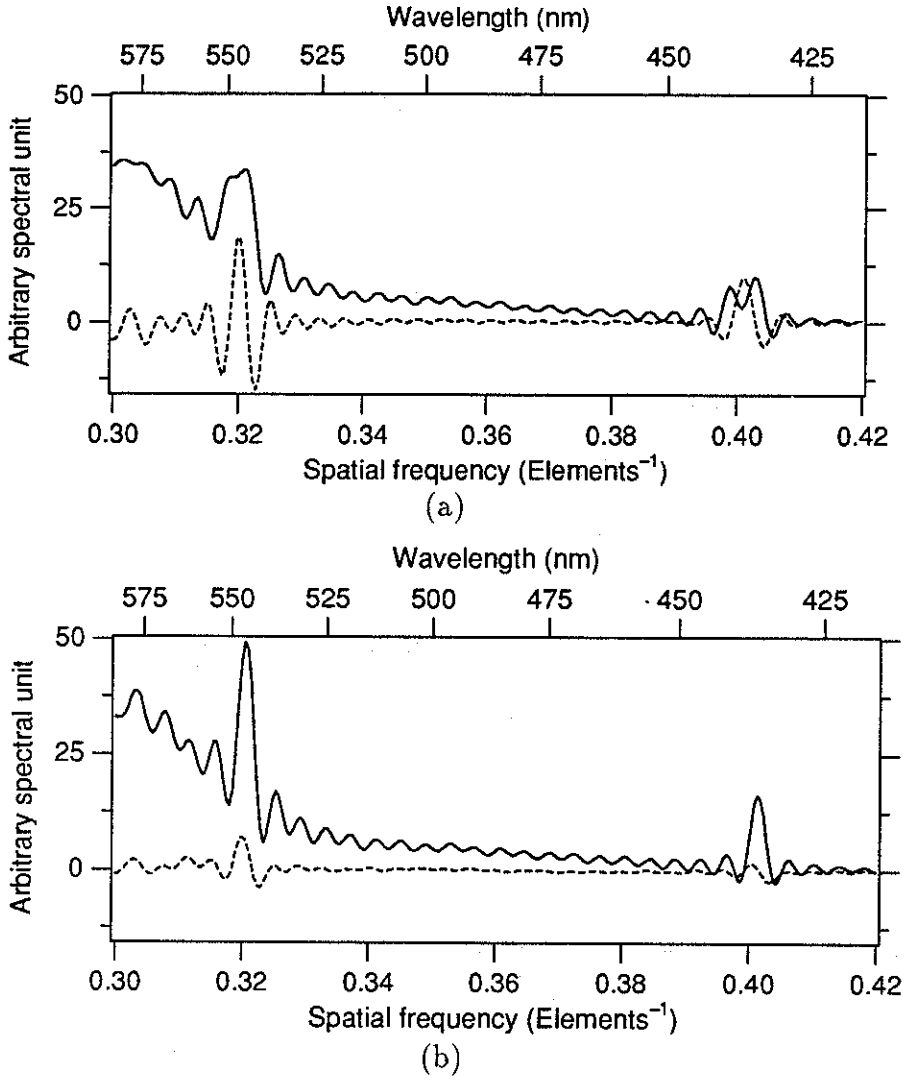


FIGURE 4.32: Unapodized versions of a part of the fluorescent lamp spectrum before (a) and after (b) resampling of the interferogram. Solid lines show real parts, dotted lines show imaginary parts.

the interferometer should take on such a fast variation; we propose instead that the error reflects a ruling error in the grating. By the treatment of periodic sampling errors in Section 3.5, we predict this error to cause spectral ghosts. As will be seen shortly this is indeed the case, and the magnitude of the ghosts correspond well with those predicted by the grating's specification sheet.

Figure 4.36 shows logarithmic plots of the uncorrected (a) and corrected (b) spectra. This time apodization has been applied in order to reduce the wings of the instrument functions. Note that the dotted traces show negative intensities. Apart from a reduction of about half a decade of the 'grass' around the double peak, the most striking improvement in the corrected spectrum

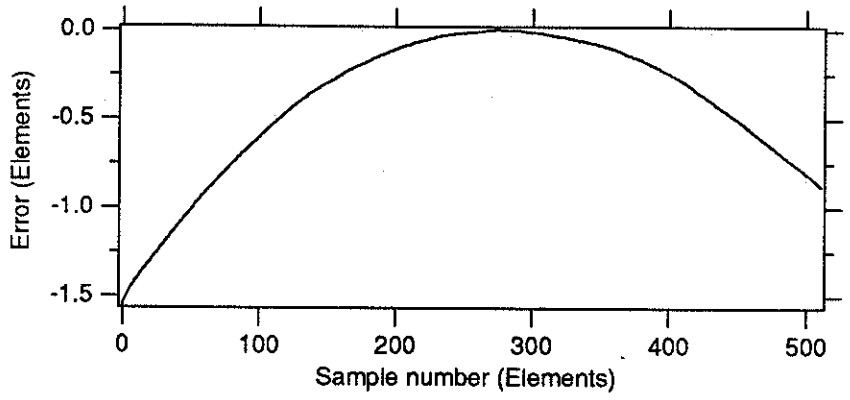


FIGURE 4.33: Estimate of the apparent sampling error in the fluorescent lamp interferogram. Its smooth, parabolic shape is due to the astigmatic aberration in the interferometer.

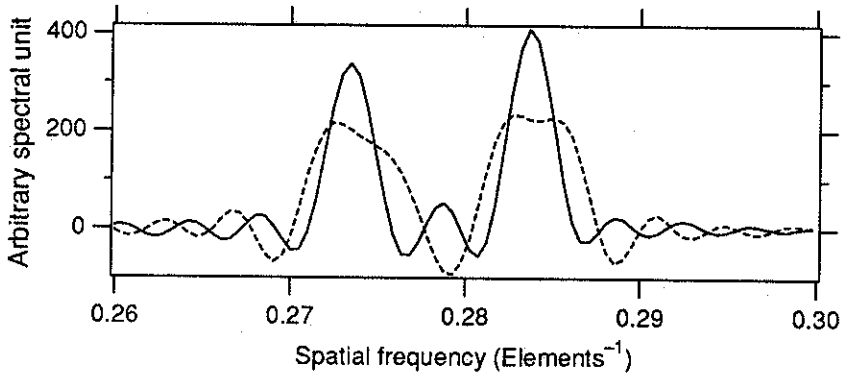


FIGURE 4.34: Comparison between the Sodium D-line spectrum before (dotted line) and after (solid line) resampling of the interferogram.

is the disappearance of two pairs of 'humps' seen only in the uncorrected spectrum (marked with arrows). They resemble the sodium lines in that they have the same inter-peak separation, but they have only a fraction of the height (1.5%). Note that one pair is positive while the other is negative, an unmistakable sign of ghosts due to a periodic sampling error (Section 3.5). The ghosts are separated from their "mother" feature by 0.026 Elements^{-1}, whose reciprocal, 38 Elements, is equal to the period of the ripple seen on the sampling error.

The specification sheet for the grating warns about ghosts with maximum intensity 0.030% of the mother feature. The specification is made for the classical grating user, however, and he is bothered not by the amplitude of the ruling error but the amplitude squared. We must therefore take the square

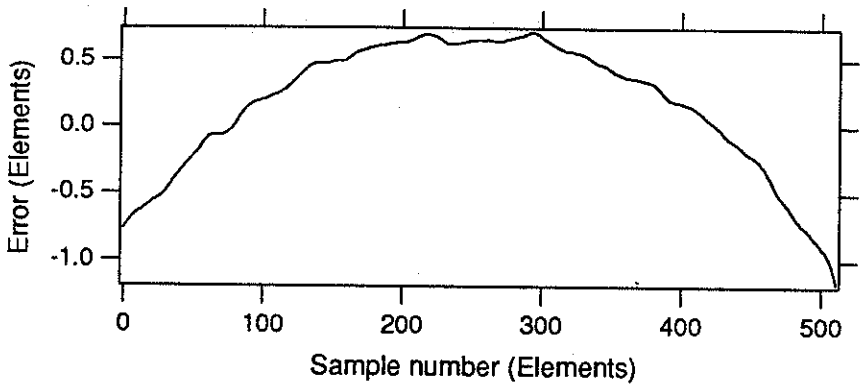


FIGURE 4.35: Estimate of the apparent sampling error for the sodium interferogram. Again the dominating shape is parabolic but an undulating structure is also seen, supposedly due to quasi-periodic ruling errors in the grating.

root of the specified ghost intensity to make it applicable to our situation, arriving at 1.7%, almost exactly equal to the observed ghost intensity. This result is somewhat disconcerting since it seriously toughens the tolerances for grating ruling errors. Seeing that we are in fact able to correct for it (at least in narrow-band spectra) is therefore a great comfort.

A new, unwanted feature has appeared in the corrected spectrum of Figure 4.36(b): two "blobs" (marked with arrows), one on either side of the doublet. Like ghosts they are of opposite signs, but they are introduced by the correction process itself and witness the fact that the reference source was not strictly monochromatic. Again the red diode laser was used as reference, but as shown in Figure 4.37 it displays a great mode-hopping activity and is never quite monochromatic. The measurement used as reference in the present example (plotted in solid line in the figure) was the best one of the lot, but it has a marked contribution from secondary modes.

4.6.6 Transformation and phase correction

Fourier transformation of the interferogram is performed by the fast Fourier transformation (FFT) algorithm [34]. For an N-points, complex input signal the FFT routine produces at its output another N-points, complex signal. If, as in our case, the input is real, the output becomes symmetrical about its origin, leaving $N/2 + 1$ (including the origin) independent complex points in the output.

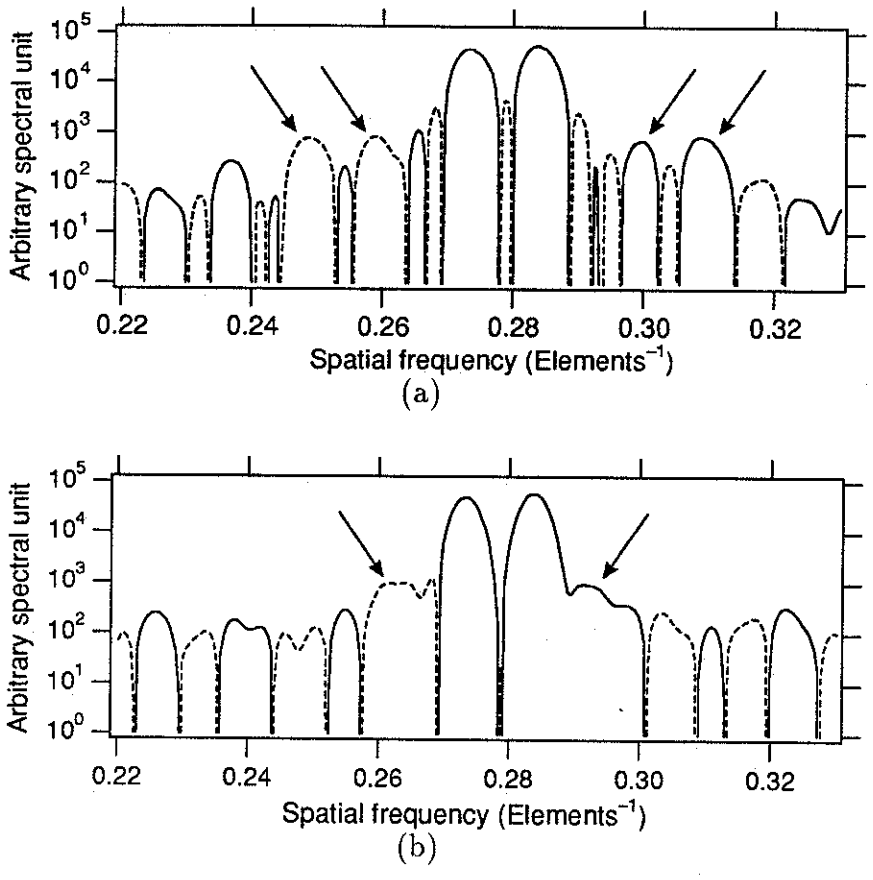


FIGURE 4.36: Logarithmic plot of the apodized sodium spectrum before (a) and after (b) resampling of the interferogram. Dotted lines show negative spectral values.

It is possible to use the magnitude of the complex FFT output as spectral estimate, but as seen in Section 3.3 a better estimate may be found by applying *phase correction*. This process requires knowledge of the phase introduced by the interferometer and as a basis for its estimation the phase of the complex spectrum itself is used. As seen by Equation 3.53, the latter is affected by an error depending upon both slope and curvature of the phase. Given a certain curvature, we must use the freedom of choice of interferogram origin to reduce the slope as much as possible, i.e. to obtain a position of *stationary phase* [32, page 21]. Often this is achieved by placing the interferogram origin at the sample of greatest amplitude, but not always, as in the case of an antisymmetric interferogram whose optimal origin is at the zero crossing between the two main peaks.

We are currently using the simple option of placing the origin at the in-

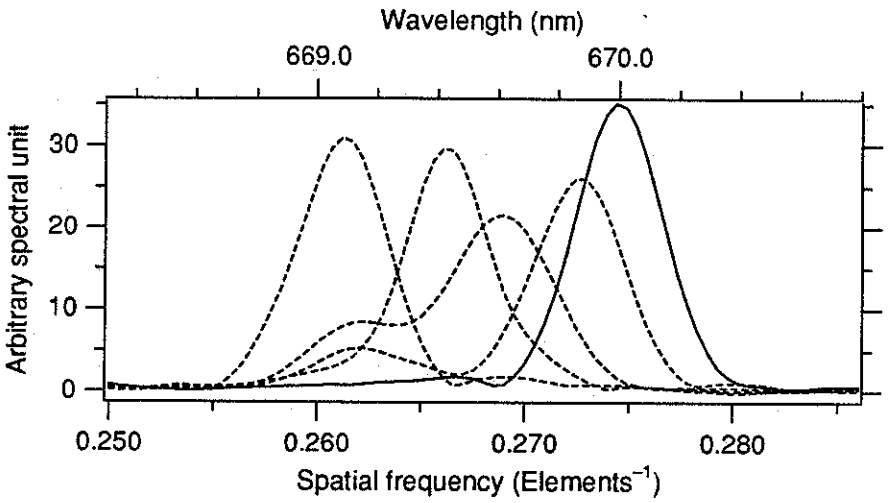


FIGURE 4.37: Sucsessively measured high resolution spectra of the diode laser used as reference for resampling. The unstable line structure and position is attributed to the phenomenon of "mode hopping".

terferogram sample of highest amplitude, but an intelligent routine searching for a more optimal choice should not be too difficult to implement. We have found, however, that if the peak value is negative it is beneficial to multiply the entire interferogram by -1 causing the phase curve to be located in the vicinity of zero rather than $\pm\pi$. Since the argument of a complex function is determined within the range $-\pi$ to $+\pi$, this avoids unnecessary discontinuities.

Having found the optimally flat phase curve, we choose a range within which the spectral amplitude is also reasonably flat and strong enough to ensure a well defined phase. A parametric function is then fitted to the phase curve within this range to give a 'fiducial phase function' which may be extrapolated into the less well defined regions. For this technique to be efficient, a function which describes the phase curve well with a minimum of coefficients should be used. Alternatively, a numerical function could be determined once and for all and the fit involving simply a change of its slope and zero offset. Presently we have not fully implemented either of these techniques; instead we use simply a polynomial fit of variable order. This gives good results in the unheterodyned case where the phase is essentially parabolic, but for high resolution measurements where it is the sinusoidal phase shape due to channelling which dominates, the method offers very poor extrapolation.

From the fiducial phase ϕ_f a conjugated complex phase function is calcu-

lated:

$$e^{-i\phi_f} = \cos\phi_f - i\sin\phi_f. \tag{4.40}$$

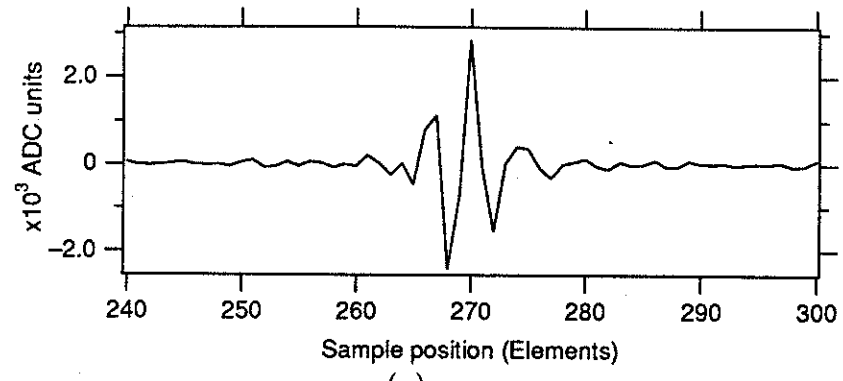
Calculation of the phase-corrected spectral estimate now proceeds according to Equation 3.41. Note that all complex arithmetic is performed on complex numbers in their rectangular rather than polar form to avoid nonlinear operations such as tangents and square roots.

Single sided interferograms. Although we have not made an in-depth performance study of the instrument's capability to measure single sided interferograms, we have implemented the basic algorithms necessary for treating such measurements. The major difference compared with double sided interferograms is that phase correction according to Equation 3.41 is no longer desirable because of the problem of overlapping aliases (Section 3.3.9). Instead, correction must be performed in the interferogram space according to Equation 3.74.

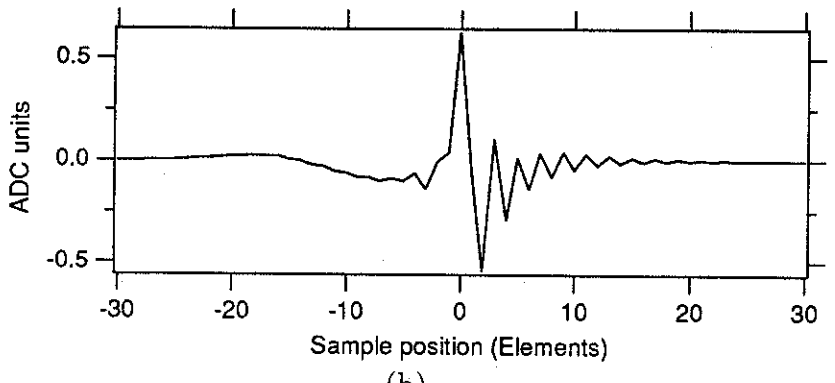
As seen in Section 3.3.9 the "single sided" interferogram should never be entirely single sided. A double sided interferogram of a certain length is necessary (140 samples was calculated as a minimum for the unheterodyned interferogram due to its strongly nonlinear phase curve) in order to calculate a fiducial phase function. A fraction of the total interferogram length must therefore be sacrificed to include the required number of samples on the "other" side of the central peak.

Again a conjugated complex phase function is produced, but this time it is reversely Fourier transformed into a "phase interferogram." Since ϕ is slowly varying, the phase interferogram has only a very short non-zero range. This is convenient since a good approximation of the convolution in Equation 3.74 is obtained by using only a few points on either side of its central peak. The result of the convolution is impressive, the corrected interferogram is (or appears, at least) completely symmetrical as illustrated in Figure 4.38.

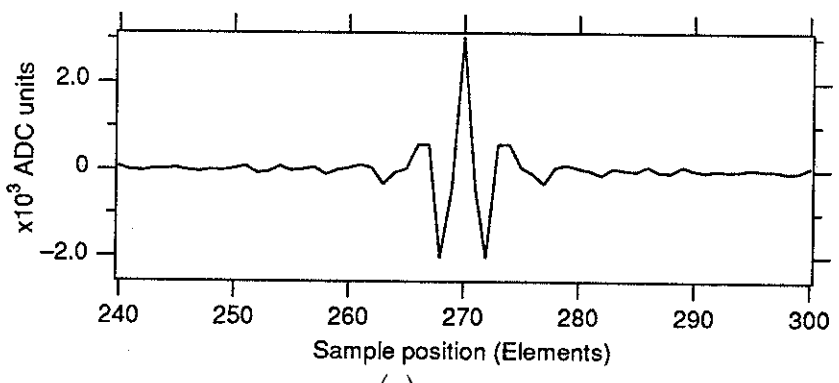
Fourier transformed after having discarded its short part and arranged its peak to be at the origin, the corrected interferogram produces a complex spectrum whose real part is the desired spectral estimate. Our phase model



(a)



(b)



(c)

FIGURE 4.38: Demonstration of the convolution method for phase correction. The original interferogram (a) is convolved with the 'phase interferogram' (b) to yield the symmetrized interferogram (c).

is at present not good enough to produce reliable spectral estimates however, and it remains therefore to be seen how well actual single-sided measurements may perform and which gains in resolving power may be achieved.

4.6.7 Calibration

The final signal processing operation necessary before displaying the spectral information is calibration of the spectral unit. Spectral calibration of our instrument requires in general three pieces of information, although when used in its unheterodyned mode, only one of them (the first) is required:

- One known frequency component within the spectral range,
- The resolving power of the grating, and
- The *direction* of the spectrum.

The calibration formula may be deduced from Equation 3.17 by solving it with respect to the Littrow frequency:

$$\sigma_L = \frac{\nu_G \sigma}{\pm\nu + \nu_G}. \tag{4.41}$$

If a known frequency component σ_1 is represented by spatial frequency ν_1, then, since σ_L is constant:

$$\frac{\nu_G \sigma_1}{\pm\nu_1 + \nu_G} = \frac{\nu_G \sigma}{\pm\nu + \nu_G} \tag{4.42}$$

which, when solved with respect to σ, gives:

$$\sigma = \sigma_1 \frac{\pm\nu + \nu_G}{\pm\nu_1 + \nu_G}. \tag{4.43}$$

By Equation 3.33 the effective grating frequency is given by $\nu_G = mN_G/L = \mathcal{R}_G/N_D$, where $\mathcal{R}_G = mN_G$ is the grating's resolving power, L is the length of the interferogram, and N_D is the number of detector elements; the equality between L and N_D is valid when the sample separation is used as unit of length along the interferogram. Hence the relationship between optical and spatial frequencies may be written as:

$$\sigma = \sigma_1 \frac{N_D\nu \pm \mathcal{R}_G}{N_D\nu_1 \pm \mathcal{R}_G} \tag{4.44}$$

where top and bottom have been multiplied with ± 1.

The signs in Equation 4.44 are chosen according to the direction of the spectrum: $+$ for positive direction, $-$ for negative direction, and depends

upon which of the two filtering conditions of Section 3.2.4 has been chosen, i.e. whether the Littrow wavenumber is above or below the measured spectral range. It may be deduced from observing the interferogram when the grating angle (and hence the Littrow wavenumber) is varied: if the fringes widen (i.e. their spatial frequency decreases) as the grating angle increases then the direction is *negative*; if the fringe frequency increases with grating angle then the direction is *positive*.

Calibration is most easily performed by the use of a monochromatic reference source within the spectral range. Its peak position may then be determined automatically and calibration proceed without intervention. A known feature in a broad-band spectrum may also be used, but manual determination of the spatial frequency of the reference point is then required.

If no reference feature is known for a spectrum then the channelling effect described in Section 4.2.7 may be used to determine the calibration, either from the period of the ripple (unheterodyned mode) or from the phase of the ripple (heterodyned modes). We have not considered this possibility in any detail, however.

In FTS spectra measured with a photon-counting device such as the photodiode, spectral intensity is obtained in photons per unit wavenumber, B_σ. Spectral intensity may be converted to photons per unit wavelength by the following relationship:

$$B_\lambda = \sigma^2 B_\sigma. \tag{4.45a}$$

For conversion to *power* per unit wavelength, we note that the energy contained in a photon equals hc/λ. For exposure time τ, the (average) spectral power is therefore given by:

$$\Phi_\lambda = \sigma^3 B_\sigma \frac{hc}{\tau}. \tag{4.45b}$$

4.6.8 User interface

Thanks to the hardware interface of our general purpose signal processing software, we have incorporated a simple user interface for operation of the instrument. Although slow, it provides a good test bed for the various features of such an interface.

In a test mode, the instrument is instructed to scan the diode array continuously and to send to the computer the highest and lowest interferogram values. From these a contrast figure is calculated and displayed, continuously

updated, on the computer screen. This feature allows readjustment of the interferometer, e.g. after a background scan has been taken. For adjustment in connection with a change of grating or spectral range, the full interferogram is required. This is currently available only by connecting the analogue video line directly to an oscilloscope (the instrument is equipped with a co-axial connector for the purpose), but a specially designed control software should be fast enough to down-load the central section of the interferogram to the computer and display it in real time on the screen.

To read scans, the programme asks for a name for the session, the type of scan (background, target, or reference), and a label (in case many scans of the same type are taken). It then acquires a pre-scan from which the optimal exposure level is determined, sets the exposure level accordingly and acquires the measurement scan. Down-loaded to the computer this scan is stored together with a header (containing exposure level, date, and time) in a file named according to session, type, and label. The data is also displayed on the screen and stored in programme memory in one of three 'accumulators' containing the most recent scan of each type. Data may similarly be read in from a file and stored in the appropriate accumulator.

Data analysis may be performed in two simple steps: calibration of the spectral axis according to the current reference scan, and full transformation including background correction of the interferogram and phase correction of the spectrum. Optionally, the process may be performed in several separate steps; this is at present necessary for the process of resampling. There are also several possibilities for tailoring of the process, allowing additional interferogram correction by polynomial fitting, alternative apodization functions and widths, manual choice of interferogram origin, etc.

4.7 Conclusion

We have in this chapter described the design of a prototype holographic FTS instrument. Small, rugged, and with a low power consumption the instrument is capable of self-contained operation in the field, controlled either automatically by its own, built-in micro-processor, or manually via a portable computer.

Its resolving power is adjustable by the change of a grating, and the spectral range of 256 independent spectral samples may be placed anywhere between 0.4 and 1.0 μm.

Much attention has been given to the optical designs of beam splitter and fringe imaging lens. The former is particularly critical and appears to be at the root of some undesired deficiencies in the instrument: poor dispersion compensation and a "channelling" effect both in the spectrum and in the phase. Apart from a sinusoidal ripple on the spectral estiamte which disappears when two spectra are ratioed, these effects are found not to impede seriously on normal operation of the instrument.

For fringe imaging we have chosen an all reflective, two-mirror lens whose advantages over transmission optical designs include simplicity and compactness. It is also tolerant to manufacturing errors and we have devised simple design criteria for the lens in the current application. Interferometric measurements show that its performance is according to its specifications.

A second weak point in the instrument design is the method chosen for supporting the interferometer reflectors. The deficiency manifests itself in two ways: as a reduction in the system's modulation transfer function and hence the signal-to-noise ratio, and as a broadening and sometimes splitting of the spectral instrument function. While nothing can be done to recover from the reduction in MTF, correction of the deformed instrument function by resampling of the interferogram has been demonstrated. As a by-product of resampling we find that ghosts, presumably due to periodic ruling errors in the grating, are also efficiently suppressed.

While mechanical and electronic designs are presented only in terms of their main features, the signal processing is described in somewhat more detail. The 'cleaning up' procedures for the interferograms are described and illustrated with examples, and noise calculations which verify well the theoretical predictions are shown. Phase correction and spectral calibration are described and the user interface of the experimental control system is portrayed.

The prototype instrument is currently in working order and produces good quality spectra. Some examples of its performance will be shown in Chapter 5,

notably giving a demonstration of the high resolution mode where an apodized resolving power of 2100 is achieved. This allows for identification of Fraunhofer lines in the solar spectrum and detection of the atmospheric pollutant NO_2 at an estimated concentration of 7 parts per billion.

Chapter 5

Practical Operation

We describe in this chapter the instrument in practical operation, starting off with an overview of the three resolution settings or 'modes of observation' that we have implemented: the low, medium, and high resolution modes. The sodium doublet spectrum is used to illustrate each mode. We then give a description of the procedures for initial adjustment and mode changing. Explained in some detail, this description will no doubt interest the user more than the examiner, but it is included here to give an idea of the operational complexity of the instrument. On this basis we discuss the possibilities for untrained operation.

In the following sections, each mode is described separately in terms of practical examples. In the low resolution mode we have measured the blue sky spectrum in which we observe the major atmospheric absorptions due to water and oxygen. We use this spectrum also to estimate the spectral response of the instrument.

To present the two heterodyned modes we return to the examples used to illustrate our radiation budget in Chapter 2. A plant reflectance spectrum has been measured in the medium resolution mode; at this resolution the 'red edge' is isolated within a window covering the range 650 to 800 nm and analyzed with a resolution of about one nanometer. The idea is to study fine features of the edge which are brought to light by taking the first derivative of the spectrum. High resolution operation is presented via a study of absorption by the atmospheric pollutant NO_2. Wavelength calibration is here achieved by using strong Fraunhofer lines. Two different 'targets' are measured for

the absorption: a glass tube containing concentrated NO_2, and the London atmosphere. By applying the Beer-Lambert law of absorption in a gas together with—in the case of the London atmosphere measurement—a crude guess of the absorption path length, gas concentrations are estimated in the two cases.

5.1 Modes and adjustments

Although our instrument may be adjusted to any resolving power up to about 10 000 and operated within any spectral window between 400 and 1000 nm, we have implemented—by the purchase of appropriate gratings and filters—two heterodyned modes in addition to the unheterodyned mode. We have therefore been able to demonstrate the following:

- A **low resolution mode** with a maximum resolving power of 256, covering the entire spectral range visible to the instrument (400 to 1000 nm). This is the unheterodyned mode in which both interferometer reflectors are plane mirrors and no band limiting filters are required.

- A **medium resolution mode** with a maximum resolving power of 1024. For this mode we are equipped with filters to study the range between 650 and 800 nm, chosen to cover the red edge of plant reflectance spectra.

- A **high resolution mode** with a maximum resolving power of 5120. The mode is applied to the study of an NO_2 absorption centred at 489 nm. At this wavelength the mode covers a range of 25 nm out of which we isolate 10 nm by the use of a standard interference filter.

Maximum resolving powers are here quoted for unapodized spectra according to the definitions given in Equation 3.31 and Equation 3.32. In practice apodization is often used to reduce sidelobes around sharp spectral features and—in the present prototype—the effect of interferometer aberrations. Resolving power of apodized spectra may be estimated as the ratio between the wavelength of an unresolved line and its full-width at half the maximum (FWHM). When the Hann window is used as apodization function the thus

estimated resolving power is down by a factor of about 2.4 with respect to the unapodized resolving power.

While in heterodyned modes the resolution is very nearly constant within the spectral range, in the unheterodyned mode it varies proportionally with wavenumber and the maximum resolving power is therefore attained only at the highest wavenumber measured. This mode also has a considerable 'dead' capacity wasted on frequencies outside the detectable range. In practice the useful range tends to be centred about half-way up the spatial frequency axis, and the resolving power is therefore usually between 50 and 100 for apodized spectra in this mode.

5.1.1 Spectral analysis of the sodium doublet

Figure 5.1 shows interferograms and spectra measured off a sodium discharge lamp in each of the modes. The interferograms have been resampled as described in Section 4.6 and left unapodized in order to show their optimal resolution. At low resolution the doublet lines centred at 589.3 nm are seen as essentially monochromatic. At medium resolution the doublet is still not resolved, although the spectrum is seen to have a finite bandwidth. Note the automatic suppression of sidelobes ("self apodization") in this mode caused by the width of the unapodized resolution element (0.6 nm) being about equal to the separation between the doublet lines (0.597 nm). Sidelobes of opposite signs are then superposed and all but cancel. Measured in the high resolution mode the doublet is fully resolved and a difference in intensity between them is observed as expected. Their separation is here measured as 0.593 nm, i.e. with an error of 0.004 nm with respect to the published value, corresponding to 0.7%.

Note the correspondence between these observations and the features seen in the interferograms: while the unresolved, low resolution measurement is featureless, the medium resolution one fits in exactly one 'hump' of the characteristic sodium interferogram. It appears therefore as if it has been apodized, hence the suppression of spectral wings. The high resolution interferogram includes several of the humps and is therefore fully resolved.

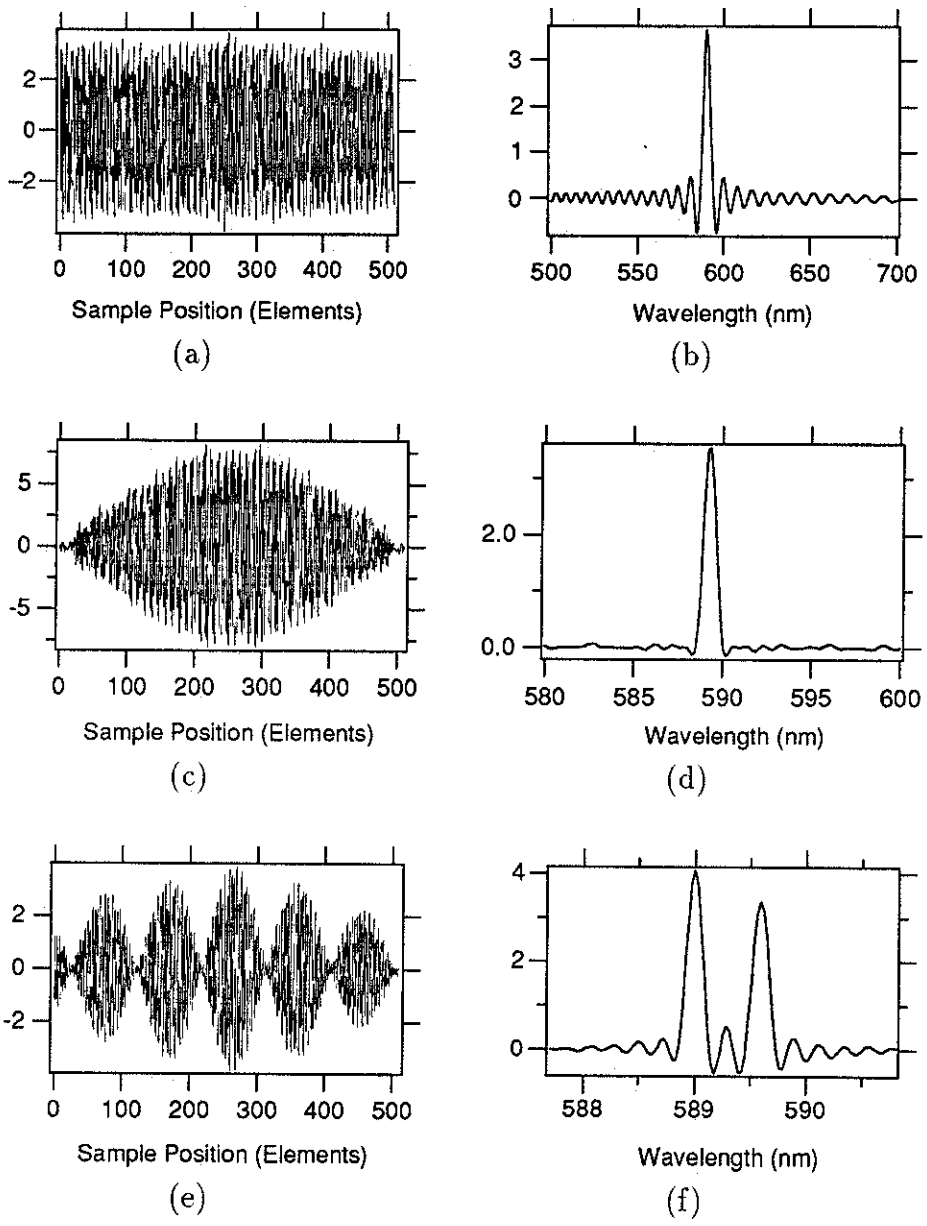


FIGURE 5.1: Interferograms and spectra of a sodium lamp measured at different resolutions. (a) and (b): 'low' resolution; (c) and (d): 'medium' resolution; (e) and (f): 'high' resolution.

5.1.2 Initial adjustment

The interferometer is—not unlike most—quite 'fiddly' to adjust, particularly at assembly and for an inexperienced user. This situation is aggravated by the lack of a single adjustment for lateral movement of the adjustable reflector: for adjustment of path length, all three control screws must be adjusted in turn. Once zero path difference has been found, however, such movement is

no longer necessary: from then on one of the side screws may (and should!) be left untouched, leaving only the centre screw (the 'fluffer') for adjustment of fringe frequency, and one side screw for adjustment of fringe rotation.

One half of the interfering light escapes the interferometer through its input aperture. Although a nuisance in radiation budget calculations, this 'deficiency' is of great help in the adjustment procedure since it allows visual observation of the interferogram.

Initial adjustment should be done with mirrors in both interferometer arms. Looking into the input aperture at the reflection of a point object allows alignment of the mirrors by ensuring the two images to be superposed. Switching to diffuse illumination from a sodium lamp should now give fringes; further adjust appropriate screws to give horizontal fringes of a comfortably visible spatial frequency.

Now switch to diffuse illumination from a mixture of white and monochromatic light; fluorescent light tubes provide an excellent mixture (see Figure 4.31). Weak fringes should be visible due to the monochromatic components, and as the path length is varied (by turning all three screws one after the other in the same direction) the contrast should either increase or decrease according to whether or not zero path difference is approached. Choose the direction of increasing contrast; this should eventually lead to the position of zero path difference where the full splendour of white light fringes is displayed. Ensuring the fringes to be horizontal, the electronics may be switched on and fringes should be visible as a sinusoidal pattern on the oscilloscope screen. Fine adjustment of a side screw allows optimization of fringe contrast.

The fringes are modulated by a bell-shaped envelope whose peak represents the position of zero path difference. Adjusting the fluffer at this point will probably cause this position to move across the screen, but some additional fine tuning allows an adjustment to be found where the position of zero path difference is almost stationary, located at the centre of the fringe field. This corresponds to the position of the axis between the support points of the side screws (see Section 4.4). When this position has been found, one of the side screws should be left strictly untouched during all subsequent adjustments.

5.1.3 Change of resolving power

Increasing the resolving power of the instrument is done by replacing the mirror in the adjustable reflector cell with a grating. The most critical point in this process is to get the grating rulings well aligned with the direction of the fringes: unless this is achieved, fringes of different frequencies will be orientated at different angles, resulting in reduced—and variable—fringe contrast. A well oriented grating will also greatly facilitate adjustment between different wavelength bands. The crux of the method for achieving good alignment of the gratings is to start off from a well adjusted interferometer and to leave both side screws untouched until the new grating is well secured in its cell.

Starting off from a well-adjusted, unheterodyned mode, open the reflector cell and remove the mirror. Insert the grating roughly oriented in the right direction (rulings running perpendicular to the instrument's line of sight). Without closing the cell, look into the input at the reflection of a point source and adjust the fluffer until a spectral image of the source produced by the grating becomes visible in addition to the white image produced by the stationary mirror. Now rotate the grating in its cell *(don't touch the side screws!)* until the spectral image is in line with the white image. Further adjust the fluffer to make the yellow spectral band coincide with the white spot; switching to diffuse sodium illumination should now again give fringes.

Looking at the sodium fringes, rotate the grating to make them horizontal; switching on the electronics, the sinusoidal pattern should again be visible on the screen. Replace the cover on the reflector cell without tightening its screws. Slots in the cover allow it to be rotated, and by pushing it lightly down while rotating, friction allows fine adjustment of the grating orientation. Thus tune the fringe contrast to its maximum before tightening the screws. Following this procedure ensures the grating rulings to be well aligned with the fringe pattern. Changing directly from one grating to another follows the same procedure, but if good adjustment is accidentally lost it is recommended to restart with the mirror.

5.1.4 Change of wave band

Looking again at a white point source, the heterodyned interferometer may be adjusted to any desired wave band simply by adjusting the fluffer to make the white image of the source coincide with the appropriate colour in its spectral image. Inserting filters and switching on the electronics should produce fringes on the screen. Adjust the fluffer to produce the desired fringe frequency and one of the side screws (always the same one) to optimize contrast. Note that the fringes may be difficult to see, representing a mere 'twiddle' on the oscilloscope screen. A facility for displaying fringes on the computer screen in real-time should improve the situation greatly.

Adjustment is greatly facilitated if a monochromatic source is available within the desired band. The instrument may then be fine tuned using this source before switching onto white light.

Operation beyond the visible range requires of course a different approach to the adjustment. Lack of experience deprives us of the ability to describe it, but we think that given a well adjusted grating, 'blind' adjustment should be possible without too much difficulty.

When the desired wave band is found, the fluffer must be adjusted to position the band within the appropriate range of spatial frequencies. Care must be taken to avoid information leakage below zero or above half the sampling frequency, as well as to choose the 'direction' of the spectrum correctly (see Section 4.6). If a monochromatic source is available, the spatial frequency that it should occupy is calculated from Equation 3.17 and the fluffer adjusted to produce fringes of that frequency. If no monochromatic or narrow band source is available, some trial and error may be required. Note that a broad-band source whose spectral intensity is approximately uniform across the band gives a fringe pattern whose mean spatial frequency corresponds to the optical frequency in the middle of the band. An interferogram whose fringes occupy about four detectors per cycle (i.e. $\nu = 0.25$ Elements^{-1}) therefore gives a nicely centred spectrum.

5.1.5 Background measurements

After the optimal adjustment has been found, one additional operation is required before starting to take spectral data: measurement of the background. As described in Section 4.6.1 this is done by adjusting one of the side screws to rotate the fringes with respect to the detector elements. When the fringes are well removed (requiring about half a turn of the screw) a scan is taken and the interferometer readjusted by turning *the same screw* back to its original position. This is a critical operation which—at least with the present display facilities—requires some practice if it is to be performed 'in the field'. It is possible, however, and it should be made easier by improved display software.

Note that the slightest adjustment of the interferometer changes the wavelength calibration of the instrument. Spectra taken before and after a background measurement or even only a readjustment of fringe contrast must therefore be calibrated separately. If a reference spectrum is required from a special purpose source, a new measurement of this source must be made after each readjustment.

5.1.6 The possibility of untrained operation

While the full adjustment procedure probably requires a certain "feel" for optical instruments in general and interferometers in particular, we think that operation of the preadjusted instrument may be done by an "untrained" operator. The only manual intervention required at this stage is that of background measurements and this only involves adjustments of a single screw. With the current display facility the operation is a bit critical, but the situation should greatly improve with more sophisticated software.

Looking a step further, towards a redesign of the instrument, full automation of the instrument may be envisaged by the use of motorized or piezoelectric interferometer adjustments. Contrast optimization and background measurement may then be performed automatically at the push of a button.

5.2 Low resolution: Unheterodyned operation

When the instrument is operated without heterodyning, spatial frequencies in the fringe pattern are directly proportional to optical frequencies (wavenumbers). Zero spatial frequency represents zero wavenumber ($\lambda = \infty$), and so, when any radiation visible to the instrument is measured in this mode, all (invisible) radiation beyond $\lambda = 1$ μm is also 'measured', although with very poor resolution.

5.2.1 The blue sky spectrum

To demonstrate this mode we take a closer look at the blue sky spectrum of Figure 4.9 where it was shown there on a linear plotted on a wavenumber scale with wavelengths indicated on the top axis. Note that, on the wavelength scale, the blue end of the spectrum is favoured with a considerably better resolution than the red end.

Assuming that the blue tint of the clear sky is produced purely by Rayleigh scattered sunlight, the solar spectrum may be deduced from this measurement by multiplying it with λ^4 [3, page 514]. Dividing the result by the standard solar spectrum at the top of the earth's atmosphere (see Figure 2.6) gives the transmission spectrum of the entire path from the top of the atmosphere to the detector, including the response of the detector itself. Figure 5.2 shows the result of these operations, plotted on a wavelength scale.

5.2.2 Spectral instrument response

As seen from the published atmospheric transmission spectrum shown in Figure 2.8, atmospheric absorption features are fairly narrow compared with the instrumental bandwidth (400 to 1000 nm). In contrast, both the instrumental transmission spectrum as estimated in Figure 4.7 and the published spectral sensitivity of the detectors (Figure 2.5) have very broad features. We assume therefore that the general shape of Figure 5.2 represents the spectral instrument response, and that its high frequency features represent the atmospheric

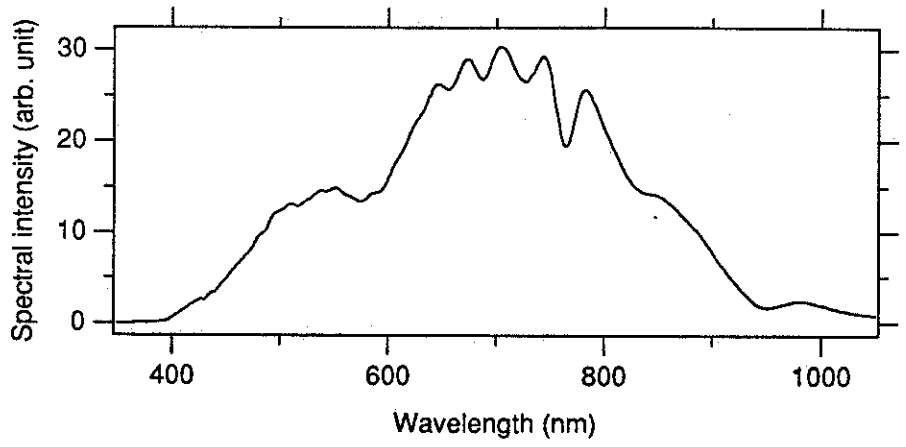


FIGURE 5.2: Combined atmospheric transmittance and instrumental response spectrum obtained from a blue sky spectrum by "correcting" it for Rayleigh scattering and dividing it by the solar spectral irradiance.

transmission.

We obtain the general shape of the measured spectrum by low-pass filtering it. The result, shown in Figure 5.3, is not quite as expected, however. Given

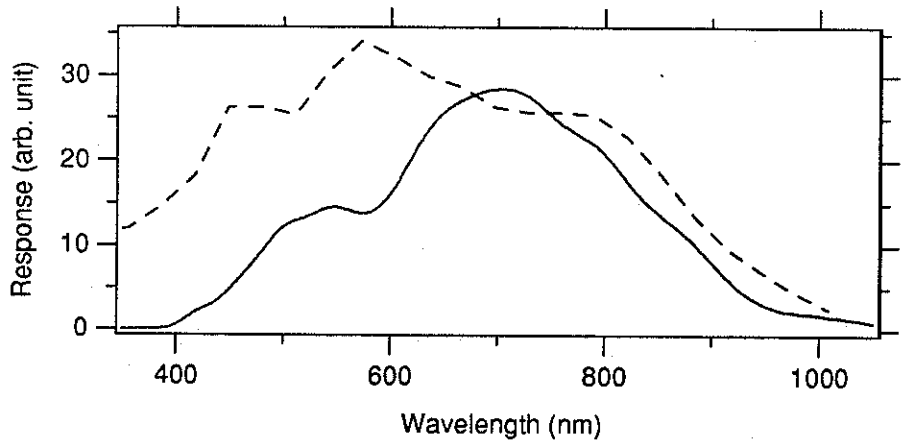


FIGURE 5.3: Estimated instrumental response spectrum (solid line) plotted together with the detector response as measured by the manufacturer (broken line).

by Figure 4.7 that the predicted instrumental transmission spectrum is quite flat, we would expect the response curve to look very much like the detector sensitivity curve, shown as a broken line in Figure 5.3. The correspondence is good at the red end of the spectrum with a ten percent cut-off at about 1000 nm. At the blue end, however, the response has its 10% cut-off at 420 nm instead of carrying on well into the ultra violet. This deficiency may be

due to instrumental absorptions which have not been accounted for, e.g. an ageing of the beam splitter film or a deficiency in the detector array. We have not had the opportunity to study the problem, but this may be done by using an appropriate source (e.g. a high temperature tungsten-halide filament lamp) together with a 'short pass' filter to attenuate the longer wavelengths.

5.2.3 Atmospheric absorptions

High frequency features in the blue sky spectrum are isolated by dividing it with the estimated instrument response, see Figure 5.4. Comparing this

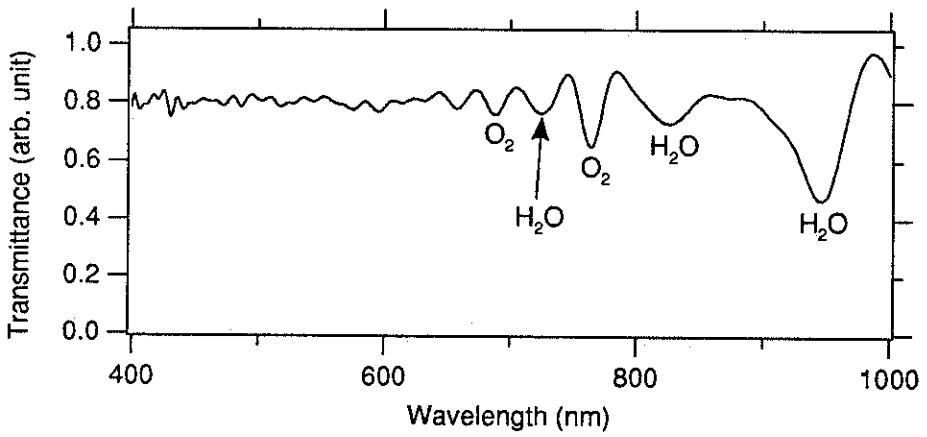


FIGURE 5.4: Atmospheric transmission spectrum obtained as described in the text.

spectrum with the published atmospheric transmission spectrum of Figure 2.8 allows all the major absorption bands due to water and oxygen to be recognized and identified. We notice, however, that short of 650 nm where atmospheric absorption shows little detail, our measurement has a ripple of about 10% peak-to-peak amplitude. This is due to the channelling effect shown in Figure 4.10.

5.3 Medium resolution: The vegetation red edge

In this mode, the interferometer is heterodyned with a grating of 80 grooves per millimeter. In the region of the vegetation red edge (see Figure 2.13) at about 700 nm, this gives a spectral resolution of down to 0.7 nm and a free spectral range of about 170 nm. To isolate the required band, a 'long pass' filter cutting

off at 600 nm and a 'short pass' filter cutting off at 850 nm are combined as shown in Figure 5.5. Note that choice of spectral direction (Section 4.6) is

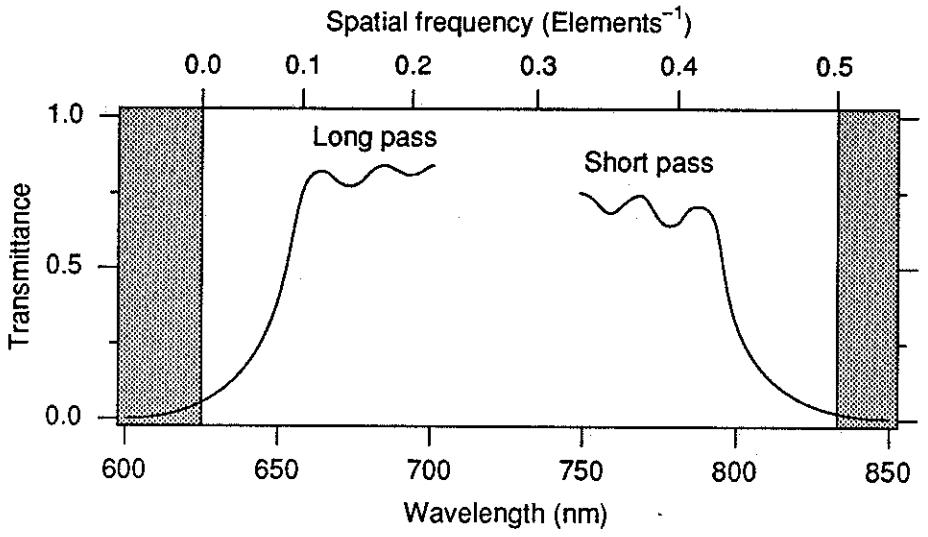


FIGURE 5.5: Transmittance spectra of the band pass limiting filters estimated from the manufacturer's specifications. The top axis shows spatial frequencies for a well adjusted instrument, and the shaded areas are outside the filtering conditions.

critical in this mode; the positive direction giving a considerably shorter free spectral range than does the negative. With our choice of filters it is important that the negative direction be chosen.

Calibration is provided by a diode laser emitting at 670 nm, conveniently within the spectral range of interest. The laser has the size of a small torch and, powered by a little 9 volt battery, is therefore easily carried out in the field.

5.3.1 Grey card measurement

Reflectance spectra are measured by taking the ratio between reflected and incident spectral intensities. A practical method of measuring the incident intensity is to measure the reflection off a surface with a known, preferably flat, reflectance spectrum, a so called 'grey card'. Our grey card has a reflectance equal to 0.526 with a standard deviation of only 0.002 within the range 400–1000 nm. We will therefore assume it to be ideally flat.

Our demonstration spectrum of the vegetation red edge was taken under less than optimal conditions. A green plant was brought into the laboratory and illuminated by the ambient light, a combination of scattered sunlight and artificial light from a fluorescent "daylight" tube whose spectrum is shown in Figure 4.31. Dominated by the fluorescent light the spectral irradiance fell off steeply with wavelength, resulting in a rather coloured grey-card measurement as shown in Figure 5.6. This situation provoked a reduction in signal to noise ratio in the red end of the spectrum.

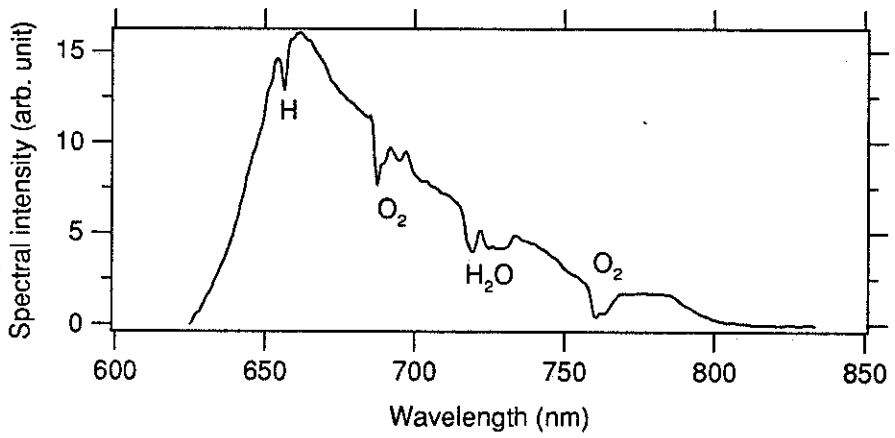


FIGURE 5.6: The grey-card spectrum. Its sharp decrease across the spectrum is caused by an illumination dominated by fluorescent light. Atmospheric absorptions are identified as labelled.

We recognize the presence of scattered sunlight from the features of atmospheric absorptions observed in the spectrum. They are now better resolved than in the low-resolution spectrum presented above, and a new absorption has also appeared: the unresolved c-line due to hydrogen at 656.2816 nm [3, page 453]. This feature may be useful as an alternative calibration reference.

5.3.2 Plant measurement

Figure 5.7 shows the measured green-plant spectrum. Here the monotonic decline in illumination with increasing wavelength is counteracted by the sharply rising edge at 700 nm. Dividing this spectrum with that of the grey card clearly shows this feature, see Figure 5.8. The absorptions identified in the grey card spectrum have all but disappeared in the ratioed spectrum, apart from some traces of the oxygen line at 760 nm which is probably an artifact of excessive

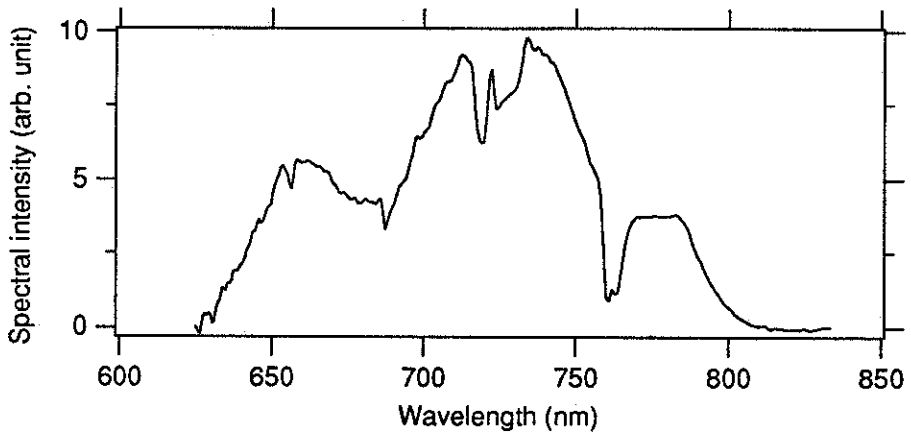


FIGURE 5.7: Spectral radiance from the green plant.

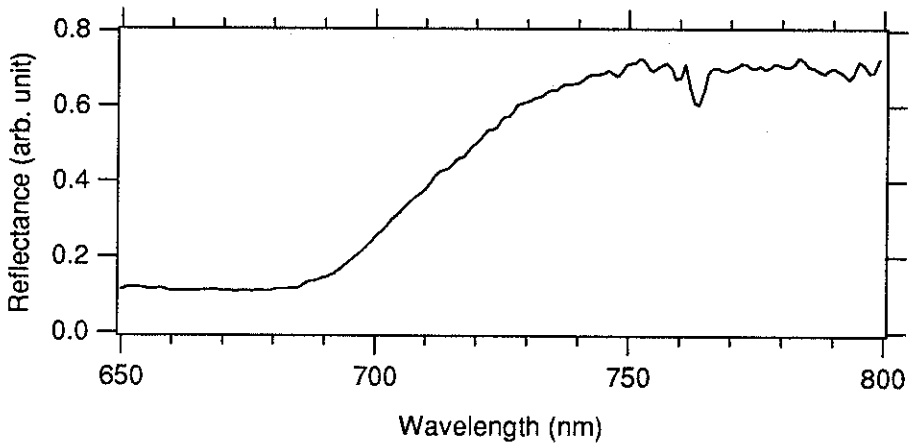


FIGURE 5.8: Reflectance spectrum of the green plant obtained by dividing the plant's spectral radiance with that of the grey card.

noise. The poor illumination has clearly influenced the noise level by giving a noticeable increase in spectral fluctuations towards the red end.

5.3.3 Analysis

According to Horler et al. [24] and Boochs et al. [27] the diagnostic features of the vegetation red edge are brought out by differentiating the reflectance spectrum. Differentiation amplifies the noise, however, so we have found it necessary to reduce the noise by low-pass filtering the spectrum before differentiation. This has been done by convolution: each spectral value in the filtered spectrum is found as the weighted sum of neighbouring spectral samples in the original spectrum. Note that this process is equivalent to apodization, the apodization window and the weighting function being each other's Fourier

transform. We find the convolution method more practical to implement in the present context, however.

Differentiation is done numerically by calculating the difference between neighbouring samples. It is therefore important that the spectrum is smooth with high-frequency noise well suppressed, i.e. that the transform of the weighting function falls off quickly. Using the rectangular (top hat) function as weighting function, representing simply an averaging of a certain number of neighbouring samples, is sub-optimal because of the sidelobes in its Fourier transform cause "leakage" of high-frequency noise. Instead the weighting function should be bell-shaped, falling off smoothly towards the edges. We have found that the raised cosine function (the Hann window) works very well as weighting function, efficiently reducing the noise. Figure 5.9 compares the ef-

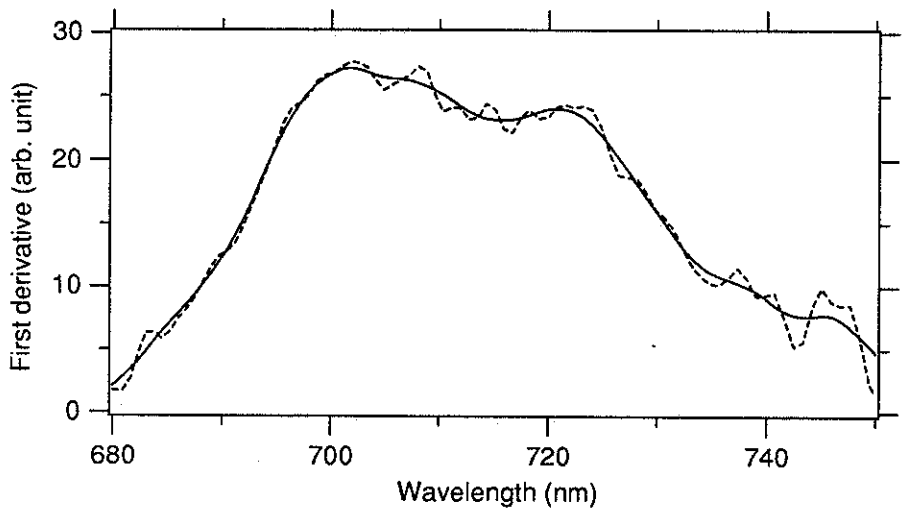


FIGURE 5.9: First derivatives obtained after convolving the red-edge spectrum with a rectangular function (dotted line) and a raised cosine function (solid line). Both functions had FWHM = 8 nm.

fect on the red edge derivative of using a rectangular filter (dotted line) and a Hann filter (solid line). Both filters had a width (FWHM) of 8 nm. Note that the rectangularly filtered version has details of width down to 2 nm.

Figure 5.10 compares the effect of using different filter widths: 4, 8, and 12 nm. While the highest resolution version has large fluctuations not accounted for in the literature, the medium resolution has lost these and is left with two clear peaks, one at 705 nm and one at 721 nm. Similar peaks are described

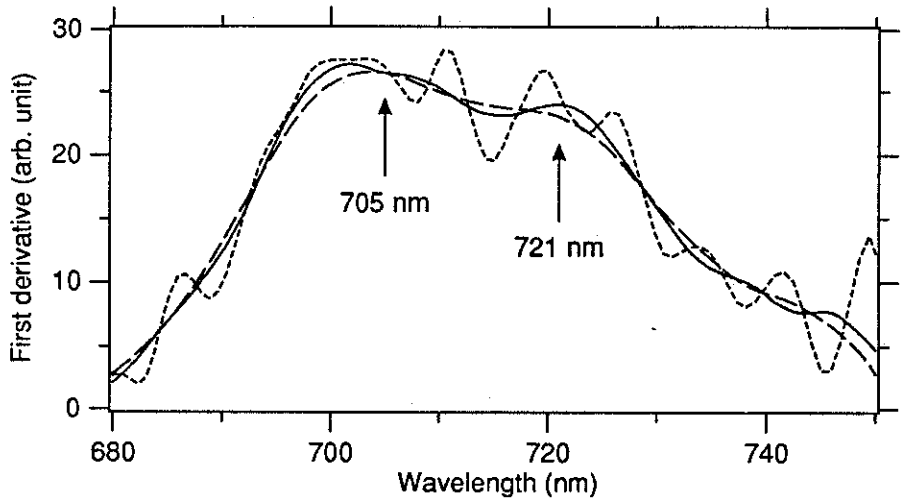


FIGURE 5.10: First derivatives obtained after convolving the red-edge spectrum with different raised cosine functions: FWHM = 4 nm (dotted line), 8 nm (solid line), 12 nm (broken line).

in the literature and is one of the main topics discussed by Horler et al. [24]. They argue that the first peak (705 nm) represents the edge of the chlorophyll absorption band while the second peak (721 nm) is caused by a sharp increase with wavelength in scattering within the leaf. A shift in the position of the red edge may therefore be accounted for by a variation in importance of these mechanisms. Note that in the lowest resolution spectrum in Figure 5.10 the peaks are hardly visible, leaving instead just a linear section. According to an experienced worker in the field of "botanic spectroscopy" the study of the red edge is often hampered by such a lack of clear peaks. We suggest that this may be due to a lack of resolution.

5.3.4 Requirement study

To evaluate the requirements for use of our instrument to obtain reliable red-edge spectra we have added noise to a synthetic spectrum and estimated the SNR of its derivative after applying filters of different widths. Figure 5.11 shows the results of this experiment, revealing an essentially linear relationship between the SNR in the spectrum before filtering, $\mathrm{SNR}_{\mathrm{Edge}}$, and that of the differentiated, filtered spectrum, $\mathrm{SNR}_{\mathrm{Diff}}$:

$$\mathrm{SNR}_{\mathrm{Edge}} \approx \frac{w}{w_0} \mathrm{SNR}_{\mathrm{Diff}}, \tag{5.1}$$

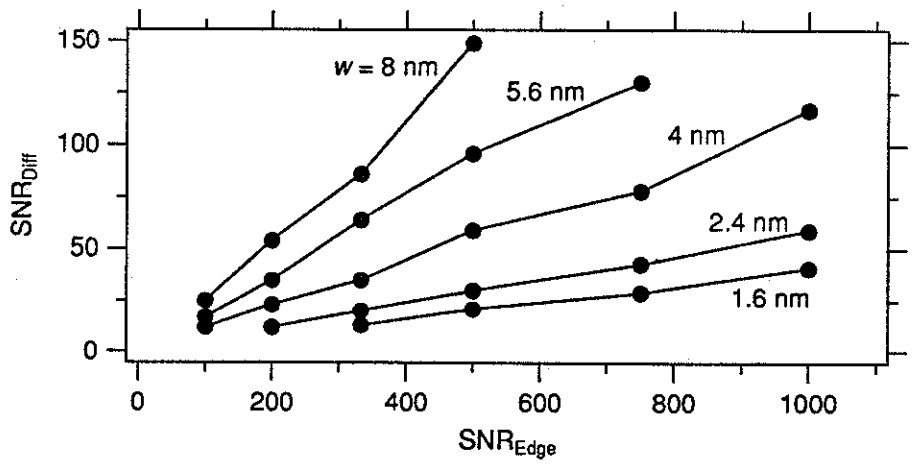


FIGURE 5.11: Estimated signal-to-noise ratio (SNR) in the first derivative spectrum plotted against the SNR of the measured red-edge spectrum for a range of filter widths (w).

where w is the FWHM of the filter and w_0 is a constant, evaluated to 34 ± 7 nm. Requiring a resolution of 5 nm and an $\mathrm{SNR_{Diff}}$ of 100, we must therefore require an $\mathrm{SNR_{Edge}}$ of about 700.

[See page 179a.] Since a spectral estimation usually requires two measurements (an interferogram and a background), its actual SNR is reduced by $1/\sqrt{2}$. Given a more even illumination than that used in our demonstration measurement, the grey card spectrum has a fill factor of about 0.7 while the plant spectrum has a fill factor of 0.5. Their respective SNRs are therefore about 200 and 270. Combined according to Equation 3.75 this gives an estimated SNR of the red edge spectrum of $\mathrm{SNR_{Edge}} \approx 160$, i.e. 4.4 times less than required by the above study.

Improving $\mathrm{SNR_{Edge}}$ may be achieved by averaging several measurements, thus reducing the noise by the square root of the number of measurements. This requires 20 measurements to be taken. With reference to Table 2.3(a) a well illuminated vegetative target requires an exposure of 5 ms. The minimum exposure allowed by the instrument is 20 ms however, so attenuation by the use of neutral density filters is in this case necessary. For 20 scans the total exposure time is then 0.4 seconds. Note that with an improvement in instru-

Equation 3.88 gives an expression for the Shot-limited signal-to-noise ratio in HFTS instruments in the case of optimal fringe contrast: $\mathrm{SNR}_\nu = \sqrt{(I_{\mathrm{Sat}}/N)/f}$ where $I_{\mathrm{Sat}} = 1.2{\times}10^8$ electrons is the detector saturation signal, $N = 512$ is the number of detectors in the array, and $f = \overline{B}/B_\nu$ is the spectral fill factor. According to the discussion in Section 3.4.6, a reduction in fringe contrast may be taken into account by multiplying with a factor k, measured to about 0.4 (see Figure 4.21). Hence, the estimated spectral SNR is $190/f$.

ment contrast the red edge SNR may be expected to reach 400, in which case the requirement is reached with the averaging of only 3 scans.

5.3.5 Discussion

We have studied the possibilities for reliable red edge measurements with our instrument. For a spectral resolution of 5 nm it has been found necessary to produce spectral estimates with signal-to-noise ratios of about 700. With the instrument in its present condition this requires averaging of many (~ 20) independent measurements, but with an improvement in contrast performance the number is considerably reduced.

Although it seems possible to obtain satisfactory results by this method, we put a question mark at its efficiency. Is it really necessary to use such high resolution (the medium resolution mode has an optimal resolution of 0.7 nm in this wavelength range) when the resolution anyway is reduced to 5 nm? If we instead use the low resolution (unheterodyned) mode to measure the spectral range from 650 to 800 nm, a resolution of 5 nm may be achieved. More importantly, we would benefit from a very small fill factor of about 1/7.
[See p. 1802.]

This calculation demonstrates a key feature of the instrument: its great flexibility allows it to be employed in different ways in different situations, allowing an optimization of its performance according to the requirements at hand.

5.4 High resolution: NO_2 absorption

The atmospheric pollutant NO_2 absorbs strongly in a 2 nm wide band centred at 489 nm, and we have implemented a 'high resolution mode' to study this absorption with an apodized resolution of 0.2 nm. The free spectral range is about 25 nm, but we use a standard 10 nm wide interference filter for band-pass limitation. This reduction in band width is beneficial from a signal-to-noise point of view since it reduces the spectral fill factor. It also

The effect of this change may be seen by considering Equation 3.88. Assume the spectrum to consist of $N_S = N_D/2$ spectral elements (where $N_D \equiv N$ is the number of detectors), with N_F elements of value B and the rest of value zero. The average spectral value is then $\overline{B} = 2BN_F/N_D$ and so Equation 3.88 may be rewritten:

$$\mathrm{SNR}_H = \frac{\sqrt{N_D I_{\mathrm{Sat}}}}{2N_F} \tag{5.1a}$$

(where $\mathrm{SNR}_H \equiv \mathrm{SNR}_\nu$). Whether spectral resolution is reduced by filtering or by a change of mode, the number of filled spectral elements in the resulting (low-resolution) spectrum remains constant. Filtering is equivalent to shortening the interferogram, i.e. wasting detector elements. N_D is therefore not constant in the two cases, the low-resolution measurement benefiting from seven times more detectors than the filtered case. According to Equation 5.1a this gives a spectral SNR advantage of $\sqrt{7} \approx 2.6$ in the low-resolution measurement.

For optimally measured data with the instrument in its present state, we estimated in the above that an improvement in SNR of factor 4.4 was required for proper red-edge measurements. This was proposed achieved by averaging $(4.4)^2 \approx 20$ measurements. Opting for the low-resolution mode brings the improvement factor down by $\sqrt{7}$ to 1.7, thus the number of required measurements to three.

gives a good tolerance to instrument adjustment by making it easier to attain the filtering condition.

5.4.1 Calibration

We do not possess a portable calibration source for this wave band. Instead we find that Fraunhofer lines seen in the measured spectra themselves provide excellent calibration signals. The line structure fits well with that measured in 1940 by Minnaert et al. [14], and in particular two strong lines are recognized as the F line due to hydrogen (486.1327 nm) and the c line due to iron (495.7609 nm) [3, page 453].

Figure 5.12 shows the spectrum measured off the blue sky and calibrated with respect to the F line. The spectrum is apodized, and it has been normal-

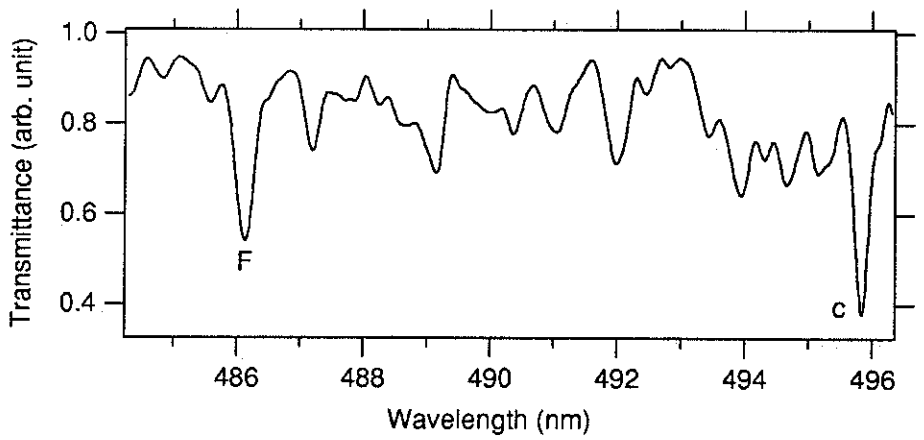


FIGURE 5.12: Fraunhofer lines measured with our instrument at a resolving power of about 2100. The two strong lines are the F-line due to hydrogen and the c-line due to iron.

ized by dividing it with a low-pass filtered version of the same spectrum. The c line now falls at 495.836 nm, having an error of 0.075 nm with respect to its published position. This error represents 0.8% of the separation between the peaks, i.e. essentially the same as the error found in the high resolution sodium spectrum (see Section 5.1.1).

Note that the FWHM (full width at half the maximum) of the narrowest spectral feature (the c-line) is about 0.24 nm. This suggests an actual resolving power of 2100 which, despite of the interferometer aberrations, compares well with the expected (apodized) resolving power of this mode of about 2200.

5.4.2 Concentrated NO_2

We have been provided with concentrated NO_2 in a glass tube of length 100 mm and diameter 22 mm. The tube bears evidence of its content by displaying a light, yellowish-brown tint. We have measured its spectral transmittance by placing it in front of the instrument aperture while pointing at the blue sky and dividing the thus obtained spectrum with that of the blue sky alone, see Figure 5.13.

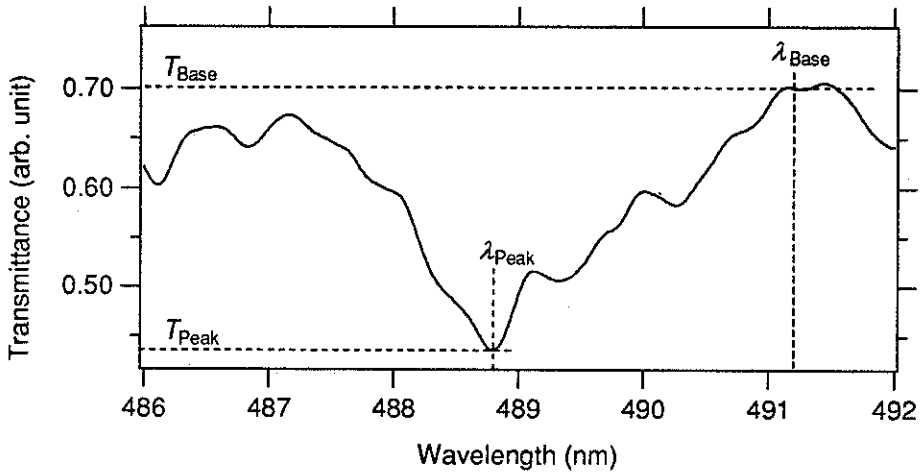


FIGURE 5.13: The absorption spectrum of an NO_2-filled glass tube as measured with our instrument.

Absorption in a gas is described by the Beer-Lambert law which may be expressed as:

$$T = \frac{S}{S_0} = e^{-pL\alpha(\lambda)} \tag{5.2}$$

where T is transmittance, S is transmitted spectral power, S_0 is the spectral power measured off the background, p is partial pressure of the gas, L is absorption path length, and $\alpha(\lambda)$ is absorption coefficient of the gas. When the gas is contained in a mixture with other gases, the ratio between its partial pressure and the total pressure of the mixture gives the *concentration* of the gas in number of molecules per total number of molecules.

The absorption coefficient for NO_2 has been measured by Woods and Jolliffe [22], see Figure 2.14. The good fit between our curve and the published results indicates that our measurement is of a good quality.

We may estimate the concentration of NO_2 in the tube by comparing the two curves quantitatively. In order to avoid errors due to ambiguities in abso-

lute levels of the measured spectra, it is necessary to compare the transmittance at two points. We chose two extreme points on the absorption coefficient curve: $\lambda_{\text{Peak}} = 488.8$ nm and $\lambda_{\text{Base}} = 491.2$ nm, at which points the transmittance is T_{Peak} and T_{Base} and the absorption coefficient is α_{Peak} and α_{Base}, respectively. The product between partial pressure and path length, refered to as the "absorption depth" $\mathcal{D}$, may then be found as:

$$pl = \mathcal{D} = \frac{-\ln(T_{\text{Peak}}/T_{\text{Base}})}{\alpha_{\text{Peak}} - \alpha_{\text{Base}}}. \tag{5.3}$$

Picking the required values of α from Figure 2.14 ($\alpha_{\text{Peak}} = 12.6$ cm^{-1} atm^{-1} and $\alpha_{\text{Base}} = 5.7$ cm^{-1} atm^{-1}) and those of T from Figure 5.13 ($T_{\text{Peak}} = 0.437$ and $T_{\text{Base}} = 0.701$) gives an absorption depth for the NO_2 trapped in the glass tube of 6.86×10^{-2} cm atm. Hence, since the absorption path is 10 cm, the partial NO_2 pressure is 6.86×10^{-3} atm. Assuming the total gas pressure in the tube to equal one atmosphere, the concentration of NO_2 is therefore about seven parts per thousand (ppt).

5.4.3 Atmospheric NO_2

We have also attempted to measure absorption due to the atmospheric content of NO_2. Installed on the roof of the Blackett laboratory, we have pointed the instrument towards the east; the city of London. The measurements were taken in the afternoon, with the sun at our back, on a clear spring day with good visibility and little wind. We took four measurements of the blue sky spectrum, the first pointing towards the horizon, with a sightline about 5° over the horizon in order to avoid obstacles in the city, the second pointing 30° over the horizon, the third at 60°, and the last towards the zenith.

The light measured in this setup is predominantly sunlight scattered off air molecules and dust particles. Assuming a simple first order scattering model as illustrated in Figure 5.14 where each photon is scattered only once, it is clear that light measured close to the horizon has suffered a considerably longer absorption path length than light measured close to the zenith. We are also tempted to believe that the concentration of NO_2 decreases exponentially with altitude so that the horizontal path suffers more frequent encounters with NO_2

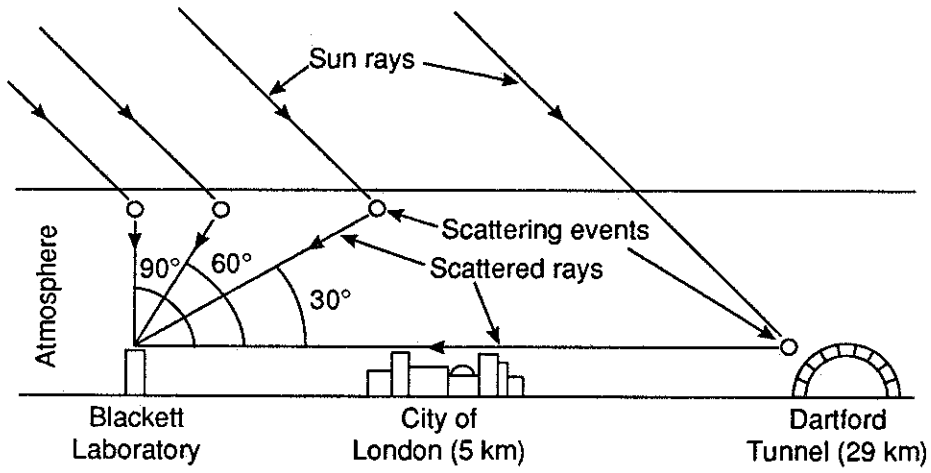


FIGURE 5.14: In our simplified, first order model for atmospheric scattering we assume each observed light ray to have undergone only one scattering event. The model predicts a considerably longer absorption path for horizontal than for zenithal rays.

molecules than the vertical path. We expect therefore to see a difference in absorption depth between the four measurements, a difference which should be brought out by ratioing the spectra with each other. Using the zenith measurement as a reference, we have hence divided all the others by it. These ratios are dominated by slow variations, typically spectral tilt and curvature, probably due to other absorption and scattering mechanisms. We remove these variations by dividing by a best fitting polynomial curve to obtain flat transmittance curves.

The 30° and 60° measurements show no visible trace of NO_2 absorption; this is no surprise considering the weak increase in path length predicted by the model of Figure 5.14. For the horizontal measurement a clear "hole" in transmittance curve is seen, however, see Figure 5.15. The curve is markedly noisy, but its shape is unmistakably similar to that of the curve obtained from measuring concentrated NO_2. Estimating 'Peak' and 'Base' transmittances we find that this measurement represents an absorption depth of 1.64×10^{-2} cm atm, i.e. about one quarter of that found in the glass tube. Although this is not strictly a measure of absolute absorption depth for the horizontal path, the large difference between horizontal and vertical path lengths should ensure it to be a good estimate.

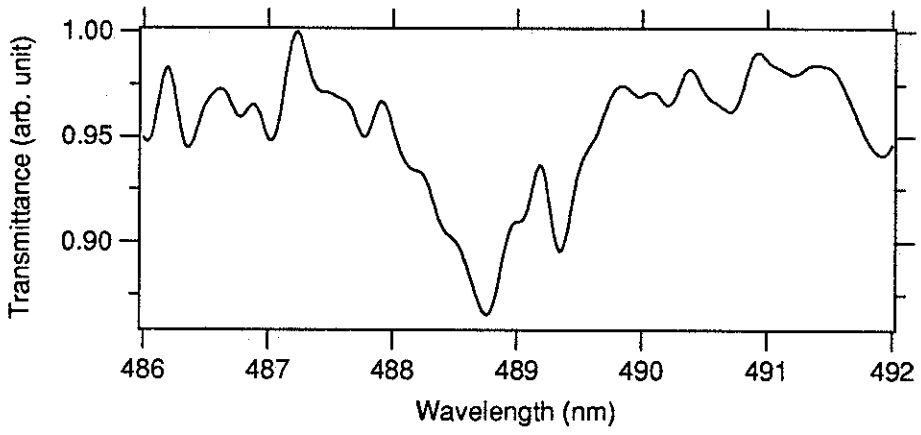


FIGURE 5.15: The NO_2 absorption is clearly recognized in this ratio between a horizontal and a zenithal blue sky spectrum.

It is difficult to take this measurement any further since the absorption path is unknown. We note however that the 'visibility' or 'meteorological range' for a 'standard clear atmosphere' is 23.5 km [15]. Looking east from the roof of the Blackett laboratory this brings us almost to the Dartford tunnel. Assuming this range to represent the order of magnitude for the horizontal absorption path we find that the partial pressure of NO_2 over London is about 7×10^{-9} atm, representing a concentration of seven ppb (parts per billion).

5.4.4 Discussion

While a concentration of 7 ppb appears to be of the right order, it is possible that a more sophisticated scattering model together with a method for measuring the meteorological range would yield a more precise estimate of the concentration. Another, more commonly used method of measuring atmospheric gas concentrations, is to compare a transmission spectrum measured towards the zenith with a synthetic spectrum generated from atmospheric models. Since, according to our scattering model, the absorption depth is much smaller towards the zenith than towards the horizon, this method puts considerably higher demands on instrumental performance.

The spectrum of Figure 5.15 with its signal-to-noise ratio of $\sim$ 100 does not represent the optimal performance of our instrument. Apart from the reduction in SNR caused by the aberrant interferogram, a further reduction of about a factor of three is due to the omission of dark-signal correction: the sky

light is sufficiently weak to require 3 second exposures, i.e. about three times longer than the limit for dark current correction estimated in Section 4.6.3. Furthermore, only single exposures were taken, thus missing the opportunity to improve the SNR by averaging. Properly dark-signal corrected, and averaged over eight measurements, spectra should attain an SNR of about 800.

5.4.5 Detection limit

If we demand the absorption "hole" to be three times as deep as the RMS noise level (ϵ_ν) for it to be detectable, we may calculate a minimum detectable absorption depth, $\mathcal{D}_{\text{Min}}$. Since the SNR outside the absorption feature is :

$$\text{SNR}_\nu = \frac{T_{\text{Base}}}{\epsilon_\nu}\,, \tag{5.4}$$

the detection requirement dictates:

$$T_{\text{Base}} - T_{\text{Peak}} > 3\epsilon_\nu = 3\,\frac{T_{\text{Base}}}{\text{SNR}_\nu}\,, \tag{5.5}$$

equivalent to:

$$\frac{T_{\text{Peak}}}{T_{\text{Base}}} < 1 - \frac{3}{\text{SNR}_\nu}\,. \tag{5.6}$$

Substituted into Equation 5.3 this gives the minimum detectable absorption depth:

$$\mathcal{D}_{\text{Min}} = \frac{-\ln(1 - 3/\text{SNR}_\nu)}{\alpha_{\text{Peak}} - \alpha_{\text{Base}}}\,. \tag{5.7}$$

which, with an SNR of 800, gives $\mathcal{D}_{\text{Min}} = 2{\times}10^{-4}$ cm atm, or 2 ppb km. Assuming an atmospheric NO_2 concentration of 7 ppb as estimated above this detection limit allows absorption path lengths down to a minimum of 300 meters.

5.5 Conclusion

We have in this chapter presented the instrument in practical operation. The three modes of operation which have been implemented are first described and illustrated by their rendering of the sodium doublet. In the low resolution mode which results from using a mirror as reflector in both interferometer arms, the doublet is completely unresolved and appears as monochromatic.

In the medium resolution mode, where an 80 grooves-per-millimeter grating replaces one of the mirrors, the doublet is still unresolved but the spectrum no longer appears truly monochromatic. We note that in this mode the width of the resolution element equals the separation of the doublet giving the effect of auto apodization. In the high resolution mode, using a grating of 400 grooves per millimeter, the sodium doublet is fully resolved.

A description of the adjustment procedures for the instrument is included to give a picture of the complexity involved in practical operation of the instrument and to evaluate the possibilities for untrained operation. We think that for routine situations where changes of mode and spectral range are not required, such operation is fully possible.

Each mode is then presented more thoroughly through their application in case studies. The low resolution mode is used for an analysis of the solar spectrum within the instrument's entire spectral range in which we recognize the major atmospheric absorption bands due to water and oxygen. We also estimate the spectral response of the instrument from this measurement. At long wavelengths it corresponds well with what has been expected, but at short wavelengths the instrument cuts off earlier than predicted. The reason for this discrepancy remains unknown.

For the presentation of the medium resolution mode we study the vegetation red edge in a window from 650 to 800 nm. Despite of a poor choice of illumination a good rendering of the edge is obtained. The diagnostic features of the red edge are most clearly displayed by taking the first differential of the spectrum. For this our demonstration spectrum is too noisy however, so we have reduced the noise by filtering, thus reducing resolving power.

It has been outside the scope of the present work to contest the field of botanic spectroscopy, we have therefore not had the opportunity to use the instrument in a *quantitative* study of the vegetation red edge. Instead we have considered the qualitative aspect through a study of noise performance requirements. The study shows that for a signal-to-noise ratio of 100 in the differentiated spectrum, a ratio of 700 between signal and noise in the original red-edge spectrum is needed. With the instrument in its present state this

requires averaging of 20 independent measurements together with a strong reduction in resolution when the medium resolution mode is used. Calculations suggests, however, that performing the same measurements with the low resolution mode may yield the required performance with only one scan.

In our demonstration of the high resolution mode we look at the absorption band of NO_2 in a 10 nm wide window at 490 nm. Calibration is in this mode achieved by the aid of strong Fraunhofer lines observed in the blue sky spectrum. We have successfully reproduced the transmittance spectrum of NO_2 by measuring light transmitted through a concentration of the gas contained in a glass tube. Comparing our measurement with a published NO_2 absorption curve, we have estimated the concentration to seven parts per thousand. By ratioing two blue-sky spectra, one measured towards the horizon, the other towards the zenith, we have also detected NO_2 absorption in the atmosphere, here at an estimated concentration of seven parts per billion. With the instrument in its present state it should be possible to detect atmospheric NO_2 at this concentration with absorption path lengths down to about 300 meters.

Chapter 6

Conclusion

The purpose of the work presented in this thesis has been to design an instrument for field spectroscopy which answers demands for increased spectral resolution and throughput. Different spectroscopic techniques have been studied and the most promising one chosen on the basis of radiation budget calculations.

A prototype instrument has been built and tested. It has been found to perform satisfactorily, although some poor design solutions prevent optimal performance. These faults have been pointed out and improvements have been suggested. Demonstration projects, involving different resolving powers and wave bands, have been implemented to show the flexibility offered by the instrument.

We present in this chapter the main conclusions drawn in the thesis and sum up recommendations for future work.

6.1 Choice of concept

A new family of satellite-based optical remote sensors with much higher spectral resolution than its predecessors set new standards for ground based reference measurements. To meet the new requirements, but also as an analytic tool in its own right, we have set out to design a new, high resolution field spectrometer.

Field operation sets strong mechanical constraints for the instrument: it must be light, rugged, and consume little power. Translated into practical

terms these constraints dictate a minimum of moving parts, no motorized scanning, and no artificial cooling.

We have studied various solutions to the optical problem of spectral analysis and found that the most appropriate candidates for our purpose are the concave grating spectrometer (CGS) and the holographic Fourier transform spectrometer (HFTS). Both are based on a linear array detector as the means of collecting spectral information without mechanical scanning.

To compare the two techniques we have chosen two practical remote sensing applications for which an estimate of available optical power has been made. This 'radiation budget' has allowed an assessment of the two spectroscopic techniques by showing under which conditions they may be used. Thanks to the throughput advantage of Fourier transform spectrometers, the HFTS method allows spectroscopic measurements to be made under far less favourable conditions than does the CGS method. While the latter reaches its limit for medium resolution vegetation reflectance measurements under conditions of 'overcast daylight', the former pushes operation into conditions of twilight, or even further by assuming natural cooling from the ambient temperature. This may be interesting for operation during the dark season in northern areas. High resolution measurements of atmospheric absorptions using the blue sky as background similarly pushes the CGS instrument to its limit while being well within reach for the HFTS construction.

In comparing noise performances of the two instrument types, we have assumed optimal exposure in both cases in order to ensure Shot-limited noise performance. Under these conditions, for quasi-continous spectra, CGS instruments have a noise advantage due to the multiplex effect. Considering equal exposure conditions however, the HFTS method is found to be superior. We note that under normal conditions (i.e. optimal fringe contrast) HFTS spectra with fill factor $\sim 1/2$ attain signal-to-noise ratios of about 1000.

6.2 Theory of operation

Having pinned down the concept, we proceed with a study of the theory behind it. Fourier transform spectrometers work on the principle of interference, and

by analogy with classical physics experiments we explain the particular way in which the holographic method produces its 'interferogram.' We see thus that the spectral information is obtained as the Fourier transform of the intensity distribution in this interferogram. An important feature of the HFTS method is that by using a grating instead of a mirror in one of the interferometer arms, the interferogram may be frequency shifted or 'heterodyned.' This allows for the implementation of high resolution measurements.

One of the major challenges in the design of FTS-type instruments is the minimization of phase difference between the sinusoidal interferogram components and to correct for the residual phase differences by signal processing. In studying this problem we have found that heterodyned HFTS instruments suffer from a particular type of phase related to the position of the grating rulings with respect to the position of the zero path difference. The phase caused by this effect is constant throughout the spectrum however, and is therefore not considered to be of importance for the functioning of the instrument. More serious phase effects are due to improper compensation of the beam-splitter substrate's dispersion which causes a nonlinear phase curve and toughens the tolerances for the phase correction procedure. We find an unexpectedly large amount of such error in our instrument due to a poor choice of beam-splitter construction, but the resulting phase-correction tolerances are well within reach for normal operation of the instrument.

Noise performance of HFTS instruments has been estimated by assuming an optimal balance between the most important noise sources. We find that Shot noise limited performance is attainable, but that the SNR of the spectral estimate is strongly dependent upon the shape of the spectrum. A single line emission spectrum may thus (theoretically) achieve an SNR of 10^5, while a broad-band spectrum filling half the available spectral range has a maximum SNR of 1000.

6.3 The prototype instrument

The performance of our prototype instrument suffers from the effects of two poor design solutions: a cemented beam splitter assembly and the use of turned

surfaces as the supports for the interferometer reflectors. One effect of the former is poor dispersion compensation causing a non-linear phase curve. This effect is particularly important in the low resolution, unheterodyned mode of operation since the bandwidth is then large. In heterodyned modes where the bandwidth is much smaller, the dispersion effect is considerably reduced. Another effect of the cemented beam splitter is channelling, giving a sinusoidal modulation of the spectral estimate as well as the phase curve. The spectral modulation is found to be efficiently suppressed by ratioing two measurements.

Probably due to a release of tension in the metal after turning, the surfaces upon which the interferometer reflectors rest have lost their flat shape. Since the reflectors are pushed against these surfaces by the pressure of an O-ring, they are forced to take on the shape error, and the interfering wave fronts are therefore deformed. This causes the effects of non-straight and non-equidistant interferogram fringes. Non-straight fringes cause a loss of contrast and hence a reduction in SNR, estimated to a factor 0.4. Variation in fringe separation causes an effect similar to that of a sampling error and has been found to corrupt the shape of spectral lines. Counteracting the error by a resampling of the interferogram has been demonstrated, but while this method works well for unheterodyned spectra, its effectiveness is restricted to very narrow-band spectra when heterodyning is used.

The other components of the design have been found to function well. Particular attention has been given to the lens which images the interferogram onto the detector array. An all-reflective, two-mirror design, the Offner lens, has been chosen for this purpose. Optimal design criteria are calculated and the satisfactory results of an interferometric test is presented. Mechanical and electronic designs for the instrument are likewise found to fulfil their roles with merit.

Instrument control and signal processing, implemented in a rudimentary form on a portable computer, have been described. Although some less than optimal shortcuts have been taken at this stage, the system demonstrates well the capabilities of the instrument. In particular, the interferogram correction procedures allow noise to be reduced to the expected level, and the phase

correction routine takes good care of the phase problem.

The instrument has been applied to the measurement of the vegetation red edge at 700 nm and the atmospheric NO_2 absorption at 489 nm. We observe the double inflection of the red edge and succeed in detecting atmospheric NO_2 at an estimated concentration of seven parts per billion.

6.4 Recommendations

Apart from the development of a purpose designed control and signal processing system, an improvement of the two noted instrumental deficiencies has the highest priority on the list of recommendations for future work. Such improvements may be undertaken without a full redesign of the instrument.

The problems related to the cemented beam splitter may be partly or fully removed by making a new beam splitter assembly where the compensator plate is fixed to the substrate by some other means than cementing. The two pieces should then be made identically thick for optimal dispersion compensation. They may be mounted in optical contact, but it is probably safer with respect to interference between spurious reflections to introduce a wedged air gap between them.

Rectification of the faulty reflector supports may be done by letting the reflectors rest on three ball bearings instead of the turned surface. The pressure necessary to keep the reflectors in position may be supplied from three rubber ball bearings placed directly opposite, see Figure 6.1. As a preliminary solution one may try to rectify the shape error by reinstalling the reflector cells in the turning lathe and remove a few microns from the support surfaces.

Other improvements of the instrument hardware which might be tried include the use of a stray-light baffle at the input. Stray-light performance has not been tested directly, but it seems to be insufficient particularly when a strong point source (e.g. the sun) is approached [30]. It is also possible that differences in the shape of the interferogram background which have been observed as the target geometry changes are caused by stray light. Adding baffling may therefore improve background correction and hence the instrument's low frequency performance.

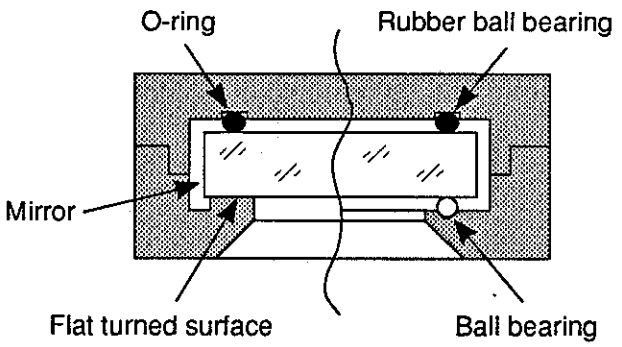


FIGURE 6.1: Cross section of the interferometer reflector cell showing the suggested new support system using three ball bearings (right hand side) compared with the existing system (left hand side).

In the original design it was thought that the filter compartment placed in front of the interferometer would provide sufficient baffling. Further protection may be added by mounting a tube in front of the input aperture with an internal structure similar to that shown in Figure 6.2.

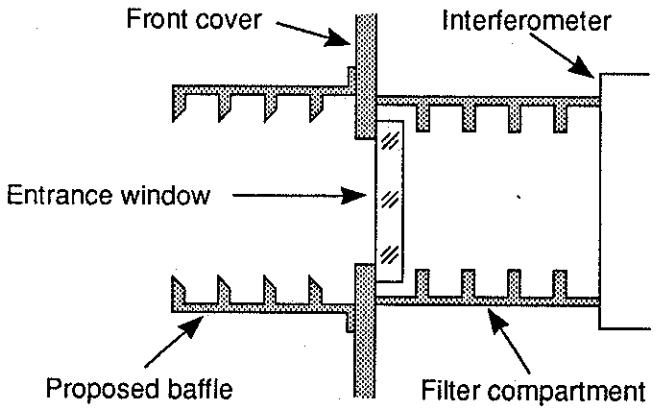


FIGURE 6.2: A typical baffle design added onto the existing instrument.

The instrument may be made more accessible to the untrained user by the inclusion of automatic fringe-contrast control. The simplest way to implement this would be to replace one of the interferometer adjustment screws with a motorized micrometer screw. Controlled from software, this may allow automatic background measurements, thus removing the need to 'fiddle' with instrument adjustments in routine measurement situations.

Finally, we recommend that a test is made of the measurement of single sided interferograms, a technique which theoretically promises a doubling of the resolving power. Although we have predicted a somewhat lower gain in

our instrument due to the excessive phase curvature, we still believe that it is worth the effort to test its possibilities, particularly if an improved beam splitter design results in an flatter phase function.

References

The reference list is organized in thematic groups. Books and papers are listed separately in each group. The entries are sorted chronologically according to their publication date.

PHYSICS AND OPTICS IN GENERAL

Books

[1] Joseph W. Goodman, *Introduction to Fourier optics* (McGraw-Hill, San Francisco 1968).

[2] Paul Horowitz and Winfield Hill, *The art of electronics* (Cambridge University Press, Cambridge, 1980).

[3] Francis A. Jenkins and Harvey E. White, *Fundamentals of Optics* (fourth edition, McGraw-Hill International Book Company, Tokyo, 1981).

[4] Francis W. Sears, Mark W. Zemansky, and Hugh D. Young, *University Physics* (sixth edition, Addison-Wesley Publishing Company, Reading, Massachusets, 1982).

[5] Anne P. Thorne, *Spectrophysics* (second edition, Chapman and Hall, London, 1988).

[6] Max Born and Emil Wolf, *Principles of Optics* (sixth (corrected) edition, Pergamon Press, Oxford, 1989).

Papers

[7] H. H. Hopkins, "The aberration permissible in optical systems," Proceedings of the Physical Society **70**, 449 (1957).

[8] Abe Offner, "New concepts in projection mask aligners," Optical Engineering **14**, 130 (1975).

[9] William B. Wetherell, "The calculation of image quality." In *Applied Optics and Optcal Engineering volume VIII*, eds. Robert R. Shannon and James C. Wyant (Academic Press, London, 1980), p. 171.

[10] M. C. Hutley and W. R. Hunter, "Variation of blaze of concave diffraction gratings," Applied Optics **20**, 245 (1981).

[11] J. M. Lerner, R. J. Chambers, and G. Passereau, "Flat field imaging spectroscopy using aberration corrected holographic gratings," Proc. SPIE **268**, 122 (1981).

[12] M. C. Hutley, *Diffraction gratings* (Academic Press, London, 1982).

[13] R. N. Wilson, F. Franza, and L. Noethe, "Active optics I. A system for optimizing the optical quality and reducing the costs of large telescopes," Journal of Modern Optics **34**, 485 (1987).

REMOTE SENSING

Books

[14] M. Minnaert, G. F. W. Mulders, and J. Houtgast, *Photometric atlas of the solar spectrum from λ3612 to λ8771* (D. Schnabel, Amsterdam, 1940).

[15] RCA, *Electro-optic handbook* (RCA Commercial Engineering, Harrison, NJ 07029, 1968).

[16] Dag T. Gjessing, *Remote surveillance by electromagnetic waves for air – water – land* (Ann Arbor Science publishers inc., Michigan, 1980).

[17] Philip Slater, *Remote sensing optics and optical systems* (Addison-Wesley Publishing Company, Reading, Massachusets, 1980).

[18] David M. Gates, Harry J. Keegan, Jhon C. Schleter, and Victor R. Weider, "Spectral properties of plants," Applied Optics **4**, 11 (1965).

[19] Fred E. Nicodemus, "Directional reflectance and emissivity of an opaque surface," Applied Optics **4**, 767 (1965).

[20] Fred. E. Nicodemus, "Reflectance nomenclature and directional reflectance and emissivity," Applied Optics **9**, 1474 (1970).

[21] Matthew P. Thekaekara, "Extraterrestrial solar spectrum, 3000–6100 Å at 1-Å intervals," Applied Optics **13**, 518 (1974).

[22] P. T. Woods and B. W. Jolliffe, "Experimental and theoretical studies related to a dye laser differential lidar system for the determination of atmospheric SO_2 and NO_2 concentrations," Optics and Laser Technology, February 1978, p. 25.

[23] Alexander F. H. Goetz, Barrett N. Rock, and Lawrence C. Rowan, "Remote sensing for exploration: an overview," Economic Geology **78**, 573 (1983).

[24] D. N. H. Horler, M. Dockray, and J. Barber, "The red edge of plant leaf reflectance," Int. J. Remote Sensing **4**, 273 (1983).

[25] E. J. Milton, "Principles of field spectroscopy," Int. J. Remote Sensing **8**, 1807 (1987).

[26] Deborah Vane, "Earth Observing System: a platform for imaging spectrometers," Proc. SPIE **834**, 176 (1987).

[27] F. Boochs, G. Kupfer, K. Dockter, and W. Kübauch, "Shape of the red edge as a vitality indicator for plants," Int. J. Remote Sensing **11**, 1741 (1990).

[28] U. Eschenauer, O. Henck, M. Hühne, P. Wu, I. Zebger, and H. W. Siesler, "Near-infrared spectroscopy in chemical research, quality assurance, and

process control." In *Near Infra-red spectroscopy*, eds. K. I. Hildrum, T. Isaksson, T. Naes, and A. Tandberg (Ellis Horwood, London, 1992), p. 11.

[29] A. Cañas and J. D. Haigh, *Ultraviolet Fourier transform spectrometer for atmospheric sensing* (SERC research grant application, Imperial College, London, 1993).

[30] R. Eiesland, *Mise au point expérimentale d'un spectromètre de terrain* (Rapport de Travail de Fin d'Etudes, Ecole des Mines de Paris, 1993).

GENERAL FTS

Books

[31] R. N. Bracewell, *The Fourier Transform and Its Applications* (McGraw-Hill, New York, 1965).

[32] L. Mertz, *Transformations in Optics* (John Wiley & Sons, Inc., New York, 1965).

Papers

[33] Janine Connes, "Recherches sur la spectroscopie par transformation de Fourier," Revue d'Optique **40**, 45, 116, 171, 231 (1961).

[34] James W. Cooley and John W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series," Mathematics of Computation **19**, 297 (1965).

[35] Michael L. Forman, W, Howard Steel, and George A. Vanasse, "Correction of Asymmetric Interferograms Obtained in Fourier Spectroscopy," J. Opt. Soc. Am. **56**, 59 (1966).

[36] Hajime Sakai and George A. Vanasse, "Hilbert Transform in Fourier Spectroscopy," J. Opt. Soc. Am. **56**, 131 (1966).

[37] L. Mertz, "Auxiliary computations for Fourier spectrometry," Infrared Physics **7**, 17 (1967).

[38] H. Sakai, G. A. Vanasse, and M. L. Forman, "Spectral recovery in Fourier Spectroscopy," J. Opt. Soc. Am. **58**, 84 (1968).

[39] D. A. Walmsley, T. A. Clark, and R. E. Jennings, "Correction of Off-Center Sampled Interferograms by a Change of Origin in the Fourier Transform; the Important Effect of Overlapping Aliases," Applied Optics **11**, 1148 (1972).

[40] R. B. Sanderson and E. E. Bell, "Multiplicative Correction of Phase Errors in Fourier Spectroscopy," Applied Optics **12**, 266 (1973).

[41] Thomas P. Sheahen, "Chirped Fourier Spectroscopy. 1: Dynamic Range Improvement and Phase Correction," Applied Optics **13**, 2907 (1974).

[42] Thomas P. Sheahen, "Chirped Fourier Spectroscopy. 2: Theory of Resolution and Contrast," Applied Optics **14**, 1004 (1975).

[43] Jyrki Kauppinen, "Correction of the linear phase errors of one-sided interferograms," Infrared Physics **16**, 359 (1976).

[44] Robert H. Norton and Reinhard Beer, "New apodizing functions for Fourier spectrometry," J. Opt. Soc. Am. **66**, 259 (1976).

[45] James W. Brault, "High Precision Fourier Transform Spectrometry: The Critical Role of Phase Corrections," Mikrochimica Acta [Wien] **III**, 215 (1987).

[46] James W. Brault, *Fourier Transform Spectrometry* (National Solar Observatory, Tucson, Arizona 85726).

HOLOGRAPHIC FTS

Papers

[47] G. W. Stroke and A. T. Funkhauser, "Fourier-transform spectroscopy using holographic imaging without computing and with stationary interferometers," Physics Letters **16**, 272 (1965).

[48] Kunio Yoshihara and Atsuo Kitade, "Holographic spectra using a triangle path interferometer," Japan J. Appl. Phys. **6**, 116 (1967).

[49] Kôgo Kamiya, Kunio Yoshihara, and Katsuhiko Okada, "Holographic spectra obtained with Lloyd's mirror," Japan J. Appl. Phys. **7**, 1129 (1968).

[50] T. Dohi and T. Suzuki, "Attainment of high resolution holographic Fourier transform spectroscopy," Applied Optics **10**, 1137 (1971).

[51] F. Lanzl, B. Reuter, and W. Waidelich, "Moiré technique in high resolution holographic Fourier transform spectroscopy," Optics Communications **5**, 354 (1972).

[52] Yair Talmi and R. W. Simpson, "Self-scanned photodiode array: a multichannel spectrometric detector," Applied Optics **19**, 1401 (1980).

[53] H. Barnils and J. M. Simon, "Spectral analysis by polarization interferographs," Optik **68**, 209 (1984).

[54] Takayuki Okamoto, Satoshi Kawata, and Shiego Minami, "Fourier transform spectrometer with a self-scanning photodiode array," Applied Optics **23**, 269 (1984).

[55] Henry Ayamanya-Mugisha and Ronald R. Williams, "A Fourier transform diode array spectrometer for the UV, visible and near-IR," Applied Spectroscopy **39**, 693 (1985).

[56] T. H. Barnes, "Photodiode array Fourier transform spectrometer with improved dynamic range," Applied Optics **24**, 3702 (1985).

[57] Takayuki Okamoto, Satoshi Kawata, and Shiego Minami, "Optical method for resolution enhancement in photodiode array Fourier transform spectroscopy," Applied Optics **24**, 4221 (1985).

[58] T. H. Barnes, T. Eiju, and K. Matsuda, "Heterodyned photodiode array Fourier transform spectrometer," Applied Optics **25**, 1864 (1986).

[59] Takayuki Okamoto, Satoshi Kawata, and Shiego Minami, "A photodiode array Fourier transform spectrometer based on a birefringent interferometer," Applied Spectroscopy **40**, 691 (1986).

[60] Shigeo Minami, "Fourier transform spectroscopy using image sensors," Mikrochimica Acta [Wien] **III**, 309 (1987).

[61] Nigel Douglas and Harvey Butcher, "Heterodyned holographic spectroscopy and the ESO VLT." In *ESO conference on very large telescopes and their instrumentation*, eds. M.-H. Ulrich (Garching, March 1988), p. 1223.

[62] N. Douglas, F. Maaswinkel, H. Butcher, and S. Frandsen, *A study of the potential of heterodyned holographic spectrometry for application in astronomy* (Technical report No. 15, European Southern Observatories, 1991).

[63] J. Harlander and F. L. Roesler, *Spatial heterodyne spectroscopy: a novel interferometric technique for ground-based and space astronomy* (Dept. of Physics, University of Wisconsin, 1150 University Avenue, Madison, Wisconsin 53706).

THIN FILMS

Papers

[64] R. S. Sennett and G. D. Scott, "The structure of evaporated metal films and their optical properties," J. Opt. Soc. Am. **40**, 203 (1950).

[65] T. Turbadar, "Equi-reflectance contours of triple-layer anti-reflection coatings," Optica Acta **11**, 195 (1964).

[66] Jing-Jiang Xu and Jin-Fa Tang, "Optical properties of extremely thin films: studies using ATR techniques," Applied Optics **28**, 2925 (1989).

[67] Norland optical adhesive 61 data sheet, Tech Optics Ltd., Tonbridge, Kent.