SURFACE PLASMON APPLICATIONS -

MICROSCOPY AND SPATIAL LIGHT MODULATION

by

Eric Morgan Yeatman

A Thesis Submitted for the Degree of
Doctor of Philosophy
of the
University of London

Department of Electrical Engineering
Imperial College of Science, Technology and Medicine
London SW7

February 1989

ABSTRACT

This thesis is concerned with the study of the excitation of surface plasmons on laterally non-uniform structures, and with two practical applications of surface plasmons involving such structures - microscopy and spatial light modulation. The previous literature on surface plasmons is reviewed and discussed, and the theory of optical excitation of plasmons on uniform surfaces is presented. This section includes detailed original derivations of important known results as well as some new results. The generalization of this analysis to the case of focussed incident light is presented, and the results discussed. A new theoretical model is then advanced for the approximate analysis of finite beam interaction with non-uniform plasmon supporting structures.

Experimental results from a new type of microscopy using surface plasmons are presented. Images of periodic oxide structures on silver films have been obtained by both scanned and broad beam techniques; these demonstrate lateral resolution of about 20 μm combined with vertical sensitivity on the order of an angstrom. The use of surface plasmons for spatial light modulation is proposed, and the potential for this technique in relation to previously reported devices is discussed. Liquid crystal devices have been constructed; these are described, and measurements of their performance are presented. These devices demonstrate resolution comparable with conventional liquid crystal spatial light modulators, along with speeds improved by about one order of magnitude. Other configurations are proposed for future development. The thesis concludes with a discussion of the merits and restrictions of the use of surface plasmons in these two application areas, and recommendations for future work.

Acknowledgements

I am happy to take this opportunity to thank my supervisor, Professor Eric Ash, for his guidance and support, and for making the last two and a half years both enriching and enjoyable. My special thanks also go to Professor Mino Green, for his unfailing enthusiasm, encouragement and support. I would also like to thank Professor Hugh Jones, who set me on the track that led to Imperial.

I gratefully acknowledge the support of the Association of Commonwealth Universities and the British Council, who made it possible for me to study in Britain through a Commonwealth Scholarship. I would also like to acknowledge the financial support of this work by the Science and Engineering Research Council.

For their helpful advice and discussion, I would like to thank John Cozens, Keith Leaver, Richard Clarke, and especially Richard Syms, whose suggestion sparked the SLM part of this project. For their assistance in practical aspects of the work, I am grateful to Bernard Adler, Martin Caldwell and Annette Glaeser. Special thanks to my fellow students Andy Smith, Ian Healy, Colin Yeoh and Dave Atkinson, for their companionship and support. Thanks also to my colleagues at Linstead Hall, Jan Bradley, Alice Vrielink and Paul Dwyer, for helping to make hall life enjoyable and rewarding. Finally, I would like to thank my family, who despite the distance have been an encouragement and an inspiration.

My ideas for the analysis of plasmons on non-uniform surfaces finally came together on the train from Salzburg to Vienna, so I also gratefully acknowledge the unknowing aid of the Austrian National Railway.

This thesis is dedicated to the people and spirit of Pepperell, 1983-86.

and to the memory of Sylvia Marguerite Giraud Yeatman.

Table of Contents

List of Figures

CHAPTER 1

INTRODUCTION AND BACKGROUND

Surface plasmons are electromagnetic waves of transverse magnetic (TM) polarization, which propagate along interfaces between conducting and non-conducting media. They are associated with longitudinal oscillations of the free electrons at the conducting surface. The existence of these waves was first indicated by the so-called Wood's anomalies: at about the turn of the century, R.W. Wood observed that the reflectance of metallic gratings showed strong minima at certain angles of incidence [1.1]. He had in effect produced a grating coupler for plasmons - the radiative and surface wave fields were coupled when the spatial frequency of the grating equalled the difference between the parallel component of the incident wavevector and the (higher) surface wave propagation vector.

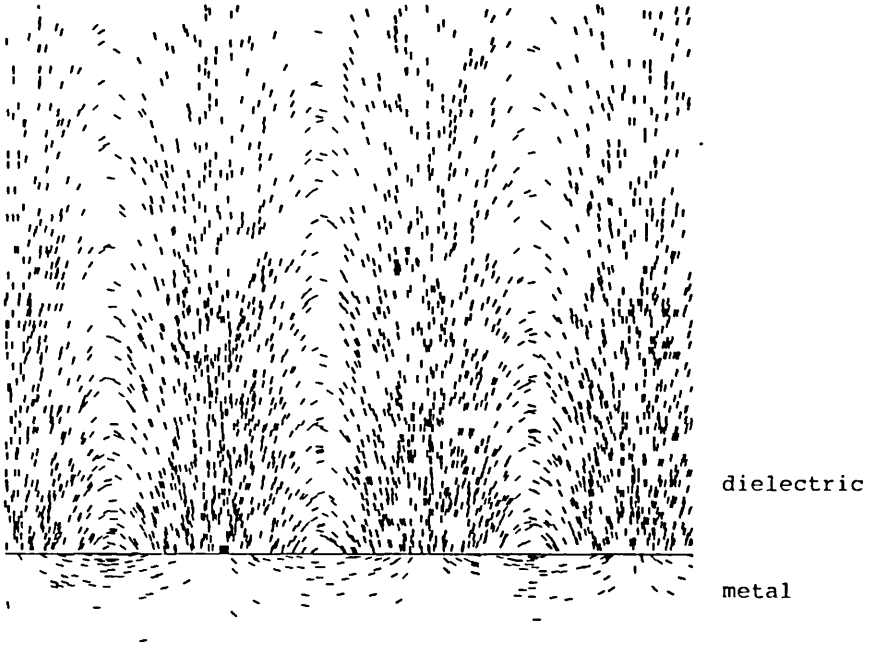


Figure 1.1 Instantaneous electric field pattern for a surface plasmon at a metal/dielectric interface. The free space wavelength is 633 nm, plotted area is 1μm x 1μm, and density of the director lines is proportional to field amplitude.

Electromagnetic surface waves were given a theoretical basis by Sommerfeld [1.2] and Zenneck [1.3], as solutions of Maxwell's equations at planar boundaries. The name 'Zenneck wave' has been given to a certain type of radio-frequency surface wave, while surface waves at visible and infra-red wavelengths are usually called plasmons. A theoretical treatment of surface plasmons in terms of the complex dielectric constant was reported in 1962 by Ritchie and Eldridge [1.4]; this has since become the standard form of analysis. For a non-radiative plasmon, the field is evanescent in the direction of the surface normal, on both sides of the surface. Consequently, power flows only along the surface, and not into the bulk

media. The extent of the field, and consequently the fraction of field energy, is much less in the conductor than in the dielectric; this results in a propagation length for surface plasmons that is substantially greater than the absorption length for plane waves incident on the conducting surface. Figure 1.1 illustrates the distribution of the electric field amplitude and direction, for a 633 nm wavelength plasmon at a silver/air interface.

Interest in surface plasmons increased considerably after the reporting by Otto of a method of optical excitation by prism coupling [1.5]. In Otto's method, the exciting beam is reflected at the base of a prism, at an angle of incidence greater than the critical angle for total internal reflection, and the evanescent field at the point of reflection extends across an air gap to a metallic surface. Coupling to the surface waves occurs when the phase matching condition $\beta_r = n_p k_0 \sin\theta$ is satisfied, where β_r is the surface plasmon propagation vector, $k_0 = \omega/c$, and n_p and θ are the prism refractive index and incident angle from the surface normal respectively. A modified configuration was introduced by Kretschmann [1.6], in which the evanescent field extends through a metal film in contact with the prism base, and excites surface waves on the far side of the metal. Both configurations are illustrated in figure 1.2.

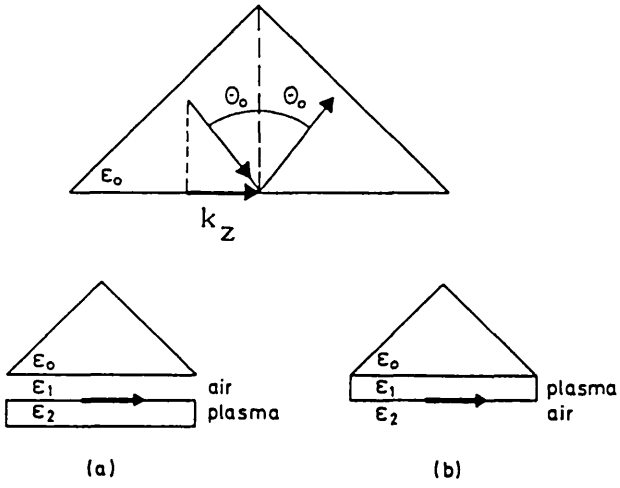


Figure 1.2 Two methods of generating surface plasmons using prism couplers: a) Otto configuration; b) Kretschmann configuration

The prism coupling method is a form of resonant excitation; when the coupling is optimized, the field intensity obtained at the plasmon-supporting surface can be 100 times higher than that of the incident beam. Optical excitation of

plasmons is thus often called surface plasmon resonance, or SPR. Figure 1.3 shows an example of the field intensity profile for plasmon excitation on a silver film in the Kretschmann configuration, illustrating the intensity enhancement. When the surface wave is generated, energy is absorbed due to collision losses in the metal. This results in attenuation of the reflected beam in the prism. For this reason, prism-coupled SPR measurements are also called attenuated total reflection, or ATR, measurements.

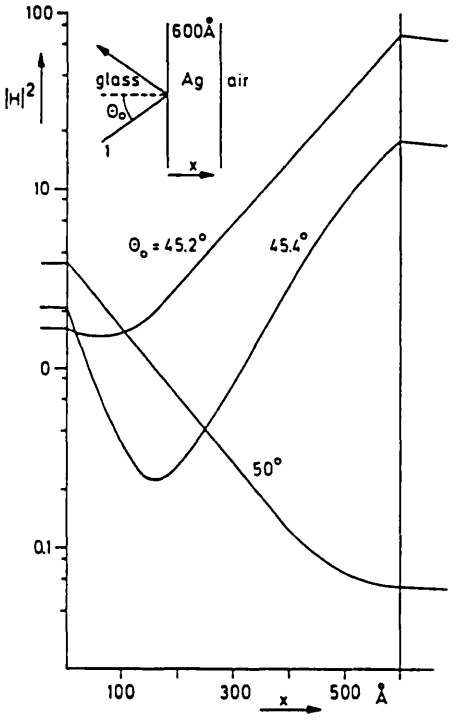


Figure 1.3 Field intensity profile in a silver film in the Kretschmann configuration, where θ_0 is the surface plasmon coupling angle, showing the intensity enhancement at resonance.

If the reflectivity is measured as a function of angle of incidence in one of the geometries described above, using a wide collimated beam, a dip is seen at the angle corresponding to phase-matching. The width of this dip is related to the propagation length of the plasmons, the depth is determined by the relationship between the absorption rate and the coupling strength, and the position of the minimum is determined by the dispersion relation for the plasmons and the phase-matching condition given above. These relations will be derived in chapter two. Figure 1.4 shows a series of examples of SPR curves for films of different metals and thicknesses. The practical interest in SPR results from the very high sensitivity of this angular response to conditions at the plasmon-supporting surface.

The deposition of organic monolayers on a metal film, for instance, can be easily detected by measuring the shift in the SPR response. This makes the use of surface plasmons for sensing an attractive possibility, and it is consequently in this area that most of the practical work on surface plasmons has been done.

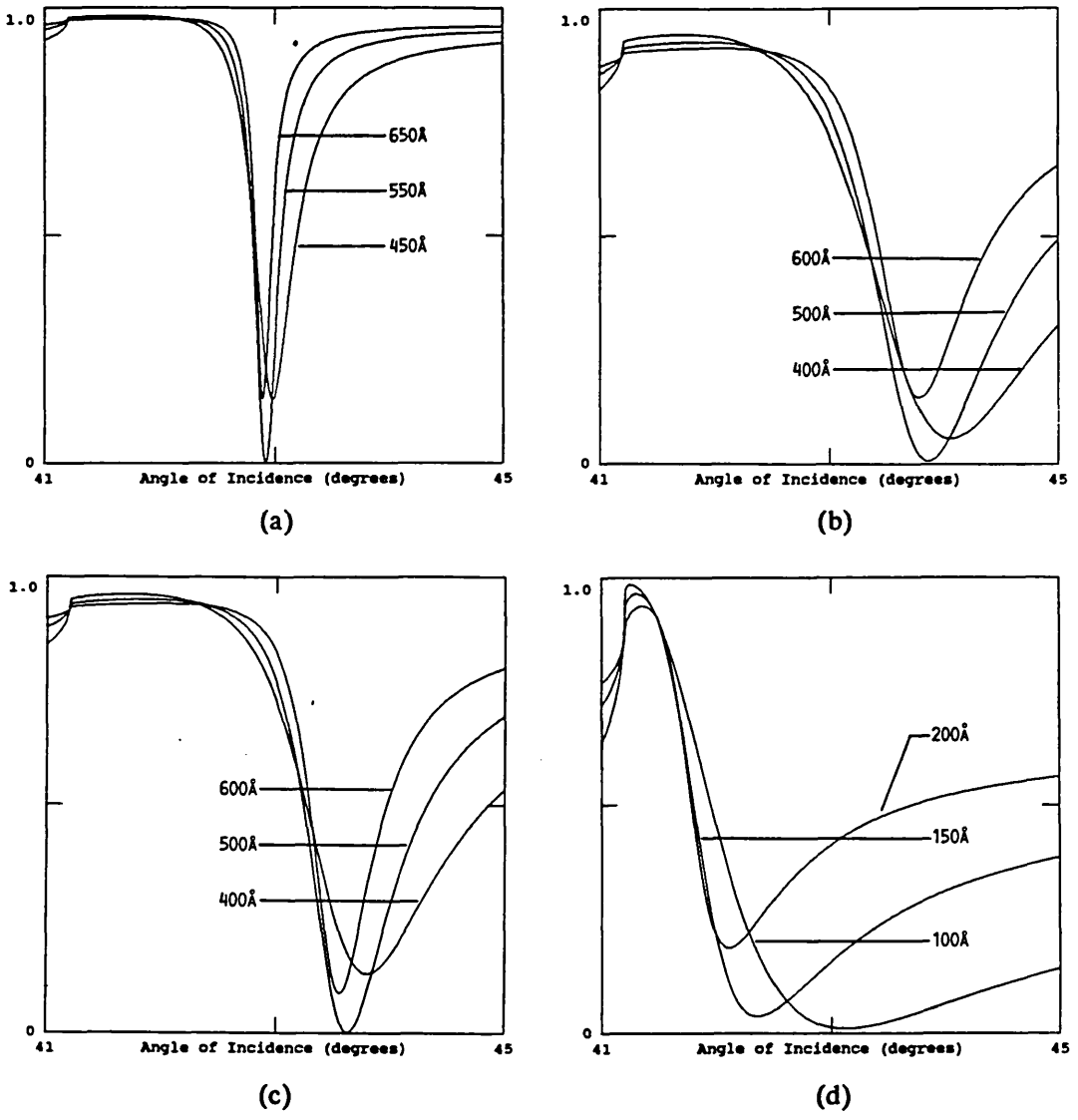


Figure 1.4 Calculated plasmon resonance curves (reflection coefficient vs. incident angle) for the Kretschmann configuration at 633 nm wavelength, with film thicknesses as indicated of: a) silver; b) gold; c) copper; d) aluminium.

A number of substantial reviews have been published on surface plasmons, for example by Ritchie [1.7], Abelès [1.8] and Raether [1.9]. The latter of these is often cited in the literature; it is quite comprehensive, and includes a very extensive bibliography. However, it suffers from a drawback that is shared to some extent by most of the literature on this topic: the basic theoretical analysis is

presented in a sketchy, confusing and sometimes misleading fashion, and many important results are not derived. For this reason, I have presented in chapter two a detailed analysis of the basic theory, starting essentially from first principles, in order to provide a clear basis for the new theoretical results.

The investigation of organic structures using surface plasmons was reported in 1977 [1.10]. This has led to the specific application of immunosensing [1.11], where the SPR response is used to test for the presence of specific organic species in solution, using surface coatings that bond to these specific molecules. Other applications of surface plasmons that have been investigated include gas detection [1.12], measurement of refractive index and absorption of thin dielectric films deposited on the metal surface [1.13], and measurement of the optical constants of the metal films themselves [1.14].

While a great deal of theoretical and experimental work on SPR has been published, the majority of it deals with laterally uniform structures, with the exception of some work on scattering from rough surfaces [1.15]. The main goal of the work described in this thesis was initially to investigate the interaction of focussed and unfocussed beams with surface plasmons on laterally non-uniform structures, and in particular, to develop a microscopy technique based on this principle. A principal part of this task was to develop a theoretical model which would allow prediction of the resolution attainable with such a technique, and analysis of the interaction of sensitivity ('vertical' resolution) with lateral resolution and surface structure.

Microscopy using evanescent fields to obtain sub-wavelength resolution has been investigated for some time, and is generally called near-field optical microscopy [1.16],[1.17]. Figure 1.5 illustrates a typical configuration. A glass slide is covered on one face by a 20 nm aluminium film in which there is a small (typically 50 nm) pinhole. A HeNe laser beam is coupled into the slide, and a conventional microscope objective focussed on the pinhole captures the scattered light as shown. The sample is scanned laterally just below the slide, so that it interacts with the evanescent field extending from the pinhole. The variation of the scattered signal as a function of sample position is then used to create a scanned image. Lateral resolution is in theory limited only by the pinhole diameter - however, the signal drops rapidly as this size is decreased, so that signal-to-noise ratio will be the practical limiting factor. In addition, the metal film must be thick enough to be highly reflecting, and a pinhole smaller than the film thickness will

produce minimal field in the region of the sample (i.e. beyond the metal surface). Scanned images have, however, demonstrated resolution better than 20 nm [1.16].

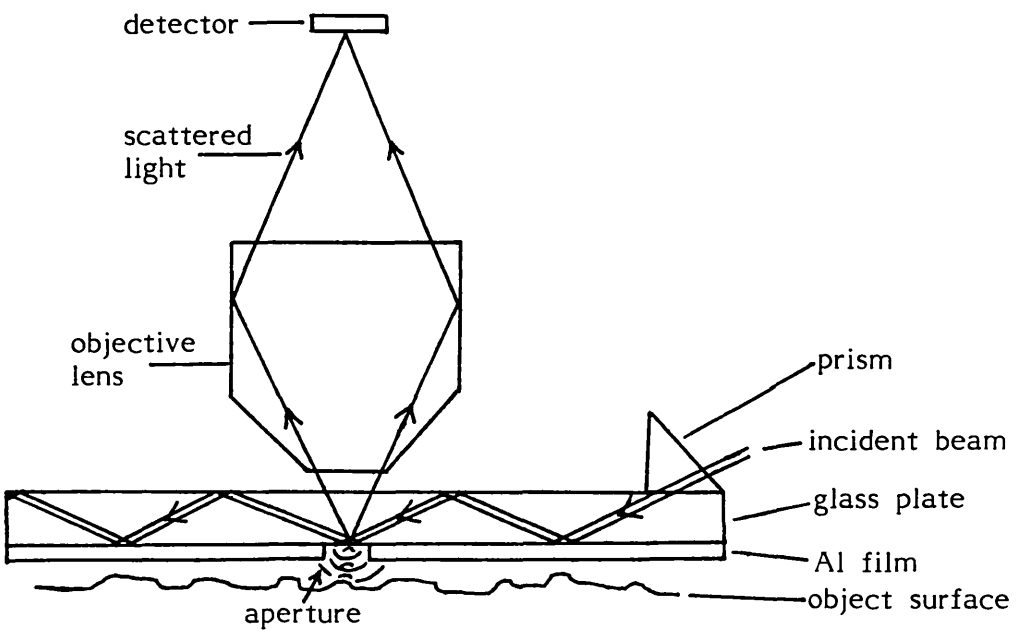


Figure 1.5 A scanning near-field optical microscope, in which the object is probed using the evanescent field extending from a sub-wavelength aperture in a conducting film (from [1.16]).

More recently, a super-resolving probe microscope using surface plasmons has been proposed by Wessel [1.18]. In this instrument, the probe would be a metal ellipsoid of sub-wavelength dimensions, mounted on a transparent base. Light focussed onto the base would create strong fields around the particle due to resonant oscillations of its surface electrons. The interaction of these fields with a sample in proximity with the particle would modify the light scattered from it, and this signal could be used to create a scanned image. The efficiency of coupling into the plasmon modes would be very low, however, since the incident field shape would be poorly matched to these modes. Reasonable contrast would therefore be difficult to achieve. To compensate, Wessel proposes using this method to measure highly non-linear phenomena, such as Raman scattering. Other than these near-field techniques, the type of microscopy most similar to that described in this thesis is perhaps acoustic microscopy, in which the generation of surface waves can play an important role in generating contrast [1.19].

In this thesis, I will describe a method of imaging features on planar structures, where surface plasmons can be generated with high efficiency using the

prism coupling methods described above. This method is of interest as a new form of microscopy because it has a different contrast mechanism from that in conventional optical microscopy, and will therefore give new information. There is no suggestion that surface plasmon microscopy will offer improvements in lateral resolution over conventional methods; in fact, the lateral resolution will tend to somewhat worse, as we shall see. However, the high sensitivity of the measurement indicates that substantial improvements in vertical resolution should be possible, thus allowing the detection of features which do not otherwise produce contrast. This should be possible with low optical power, and without vacuum or sophisticated instrumentation. The initial experimental results reported in chapter four have been published [1.20], and shortly afterward, similar results were reported by B. Rothenhausler and co-workers [1.21]. Otherwise, I know of no other work being done or having been done on this type of microscopy. The use of SPR for spatial light modulation, which developed from the work on microscopy, is treated separately in chapter five.

CHAPTER 2

PLASMONS ON UNIFORM SURFACES

2.1 Introduction

Plasmons on uniform surfaces have been extensively studied, and many of the important theoretical results are well known. The derivations of these results, however, are in most cases not reported, and those that have been published are generally sketchy and difficult to follow, particularly where approximations have been made. For this reason, this chapter will include derivations of important known results relevant to this project, as well as some new analysis.

Surface plasmons are a type of guided electromagnetic wave; therefore the starting point for our analysis is given by Maxwell's equations. These we can write in their general form as :

$$\nabla \cdot D = \rho_f \tag{2.1.1}$$

$$\nabla \cdot B = 0 \tag{2.1.2}$$

$$\nabla \times E + \frac{\partial B}{\partial t} = 0 \tag{2.1.3}$$

$$\nabla \times H - \frac{\partial D}{\partial t} = J_f \tag{2.1.4}$$

Here E and H are the electric and magnetic field vectors, and D and B the electric displacement and the magnetic induction, respectively. The quantities ρ_f and J_f are respectively the free charge density and the current density due to free charges.

Let us now introduce the definitions $D=\epsilon E$ and $B=\mu H$, where ϵ and μ are the permittivity ($\epsilon \equiv 1+\chi_e$) and permeability respectively. In addition, we will restrict ourselves to fields having a sinusoidal variation in time. To write this time dependence in exponential notation, we must choose a sign convention: we will set the time factor to be $e^{-j\omega t}$, so that a forward-going wave will have the form $e^{j(kz-\omega t)}$. We can now rewrite the latter two of Maxwell's equations as:

$$\nabla \times E - j\omega\mu H = 0 \tag{2.1.5}$$

$$\nabla \times H + j\omega\epsilon E = J_f \tag{2.1.6}$$

The current is related to the electric field according to $J_f = \sigma E$, σ being the conductivity. In the linear regime, the conductivity will be a constant, but it need not necessarily be real, as we will see later. Then (2.1.6) becomes:

$$\nabla \times H + (j\omega\epsilon - \sigma)E = 0 \tag{2.1.7}$$

It is now useful to introduce the complex permittivity, whch is defined as:

$$\underline{\varepsilon} \equiv \varepsilon + j\sigma/\omega = 1 + \chi_e + j\sigma/\omega \qquad (2.1.8)$$

This simplifies the analysis by allowing the conductivity and permittivity of a material to be replaced by a single (complex) parameter. It will prove to be convenient in later analysis if we take as our single parameter the relative complex permittivity, or dielectric constant, which we will label ϵ, and define as follows:

$$\epsilon \equiv \underline{\varepsilon}/\varepsilon_0 \qquad (2.1.9)$$

where ε_0 is the permittivity of free space. (Note the difference in symbols: ε for absolute and ϵ for relative permittivity). We will also restrict ourselves to non-ferromagnetic materials, where the permeability is approximately that of free space (μ_0). Thus the relative complex permittivity ϵ is the single parameter we will use to characterize materials in the analysis that follows. That being the case, the latter two of Maxwell's equations can be written in the form:

$$\nabla xE - j\omega\mu_0 H = 0 \qquad (2.1.10)$$

$$\nabla xH + j\omega\epsilon\varepsilon_0 E = 0 \qquad (2.1.11)$$

Combining these gives:

$$\frac{\nabla x\nabla xH}{j\omega\epsilon\varepsilon_0} + j\omega\mu_0 H = 0 \qquad (2.1.12)$$

Since $\nabla\cdot H=0$, we can substitute $\nabla x\nabla xH=-\nabla^2 H$. Then:

$$\nabla^2 H + \omega^2\epsilon\varepsilon_0\mu_0 H = 0 \qquad (2.1.13)$$

This can be simplified using the equation for the speed of light $c^2 = 1/(\varepsilon_0\mu_0)$, and by defining the free space wave vector, k_0:

$$k_0 = \omega/c \qquad (2.1.14)$$

Equation (2.1.13) can then be written as:

$$\nabla^2 H + \epsilon k_0^2 H = 0 \qquad (2.1.15)$$

This is one form of the electromagnetic wave equation for homogeneous media.

We will be concerned with structures consisting of one or more parallel planar interfaces. We can use these interfaces as a reference by which to define our coordinate axes and our principal polarizations. Any arbitrary field can be considered as a linear superposition of two fields having the two orthogonal principal polarizations. These we define as follows:

i) transverse electric (TE): E is parallel to the interface planes

ii) transverse magnetic (TM): H is parallel to the interface planes

Then we define our three axial directions:

i) $\hat{x}$ is normal to the interface planes
ii) $\hat{y}$ is in the direction of the field vector parallel to the interfaces (E for TE mode, H for TM mode)
iii) $\hat{z}$ is parallel to the interfaces and perpendicular to $\hat{y}$

Here, $\hat{x}$,$\hat{y}$ and $\hat{z}$ are unit vectors along the x,y and z axes.

2.2 Surface Waves on a Single Interface

We can define a surface wave as one that propagates along a surface or interface, without radiating energy into the two bulk media. The confinement of such a wave to the interface implies that the field intensity should approach zero as one moves away from the interface on either side. Our goal is to look for solutions to Maxwell's equations which satisfy these criteria.

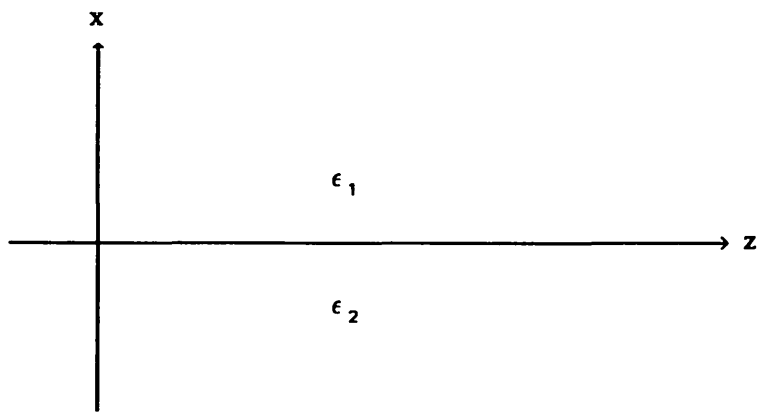


Figure 2.1 Geometry for single interface structure.

The figure above shows the coordinate system for a structure consisting of two homogeneous half-spaces which meet at the x=0 plane. The two media are characterized by complex relative permittivities ϵ_1 and ϵ_2 as shown. We will consider the two principal polarizations in turn.

First let us consider TM polarization. The H vector will be in the $\hat{y}$ direction, and the E vector will lie in the $\hat{x}$-$\hat{z}$ plane. The vector equations (2.1.10) and (2.1.11) now yield three scalar equations:

$$\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} - j\omega\mu_0 H = 0 \tag{2.2.1}$$

$$\frac{\partial H}{\partial x} = -j\omega\epsilon\epsilon_0 E_z \tag{2.2.2}$$

$$\frac{\partial H}{\partial z} = j\omega\epsilon\epsilon_0 E_x \tag{2.2.3}$$

Plane waves are solutions to the wave equation, and in TM polarization have the form:

$$H = H_0 \exp[j(\gamma x + \beta z)] \tag{2.2.4}$$

(In general, we will omit the time dependence for the sake of brevity.) Such a wave will satisfy (2.1.15), when:

$$\gamma^2+\beta^2 = \epsilon k_0^2 \tag{2.2.5}$$

We will look for a surface wave field solution which has this form in both media:

$$H = \begin{cases} H_{01}\exp[j(\gamma_1 x + \beta_1 z)] & x>0 \\ H_{02}\exp[j(\gamma_2 x + \beta_2 z)] & x<0 \end{cases} \tag{2.2.6}$$

The relationships between the various parameters are determined using the two boundary conditions, which arise directly from Maxwell's equations: the parallel components of E and H must be continuous at the interface. The latter condition is satisfied in (2.2.5) if we set $H_{01}=H_{02}$ (henceforth H_0), and $\beta_1=\beta_2$ (henceforth β). To satisfy the former condition we first obtain an expression for E_z, using (2.2.2). This yields:

$$E_z = \frac{-j}{\omega\epsilon\epsilon_0}\frac{\partial H}{\partial x} \tag{2.2.7}$$

$$E_z = \frac{\gamma H_0}{\omega\epsilon\epsilon_0} \exp[\, j(\gamma x+\beta z)] \tag{2.2.8}$$

with γ and ϵ having the appropriate subscripts in the upper and lower media. Continuity of E_z now gives us the result:

$$\frac{\gamma_1}{\epsilon_1} = \frac{\gamma_2}{\epsilon_2} \tag{2.2.9}$$

Expressions for γ_1 and γ_2 are obtained from (2.2.5), giving:

$$\frac{(\epsilon_1 k_0^2 - \beta^2)^{\frac{1}{2}}}{\epsilon_1} = \frac{(\epsilon_2 k_0^2 - \beta^2)^{\frac{1}{2}}}{\epsilon_2} \tag{2.2.10}$$

Manipulation of this equation gives us an expression for the propagation vector β:

$$\frac{\beta}{k_0} = \left[\frac{\epsilon_1 \epsilon_2}{\epsilon_1 + \epsilon_2}\right]^{\frac{1}{2}} \tag{2.2.11}$$

In order for equation (2.2.6) to describe a non-radiative surface wave, the field amplitudes must drop to zero for large positive or negative values of x. This means that the imaginary components of γ_1 and γ_2 must be positive and negative respectively. In order for this to be true, β/k_0 must be greater than both ϵ_1 and ϵ_2. However, (2.2.11) indicates that β/k_0 will be less than both values if both are positive, and will be imaginary if both are negative. The condition for existence of a non-radiative wave propagating along the surface, therefore, is that one of the media must have a negative dielectric constant; in addition, the magnitude of this constant must be greater than that of the positive ϵ of the other medium. Taking medium 1 to have a positive ϵ, we can write this condition as:

$$\epsilon_2 < -\epsilon_1 \tag{2.2.12}$$

Let us now consider TE polarization. In this case, E will be in the y direction and H in the x-z plane. Then equations (2.1.10) and (2.1.11) can be written as:

$$\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} + j\omega\epsilon\epsilon_0 E = 0 \tag{2.2.13}$$

$$\frac{\partial E}{\partial x} = j\omega\mu_0 H_z \tag{2.2.14}$$

$$\frac{\partial E}{\partial z} = -j\omega\mu_0 H_x \tag{2.2.15}$$

These will give a wave equation for E in the same form as (2.1.15), and so we can take for the form of E:

$$E = \begin{cases} E_0 \exp[j(\gamma_1 x + \beta z)] & x>0 \\ E_0 \exp[j(\gamma_2 x + \beta z)] & x<0 \end{cases} \tag{2.2.16}$$

Here we have included the condition of continuity of E. To satisfy the continuity of H_z we use (2.2.14), from which we see that H_z is proportional to γE. So we can write:

$$\gamma_1 = \gamma_2 \tag{2.2.17}$$

From (2.2.5) we can see that equation (2.2.17) cannot be satisfied unless $\epsilon_1 = \epsilon_2$, which would mean the interface would not exist physically. Therefore we can

conclude that electromagnetic surface waves can only exist in TM mode, on the surface of a material having a negative dielectric constant.

2.3 Waves in Plasmas

The waves we are concerned with propagate on the surfaces of metals. At low frequencies, metals do not have negative dielectric constants, but at high frequencies, they can behave as plasmas, which do have negative dielectric constants; thus the name surface plasmon. We shall now consider why this is so.

For our purposes we will consider a plasma to be a dilute gas of ionized particles, having an equal concentration of positive and negative charges. Let us call this concentration N_e, the number of charges, in units of the electron charge 'e', per unit volume. Let us first examine the properties of bulk oscillations of these charges.

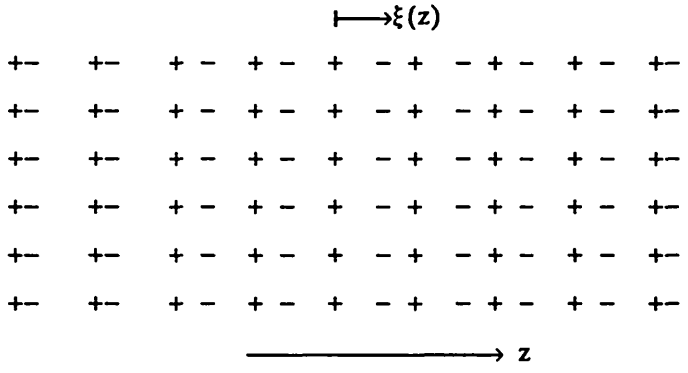


Figure 2.2 Oscillations of free electrons in a plasma. The function $\xi(z)$ gives the displacement of electrons from their equilibrium position as a function of that position (z).

As the masses of the positively charged ions are several orders of magnitude greater than the electron mass, motion of the ions will be negligible compared to that of the electrons. We will neglect both thermal motion and collisions. Let us consider an oscillation in the z direction, as indicated in figure 2.2. The relative displacement of the electrons as a function of position is given by $\xi(z)$. If we approximate our distribution of discrete electrons as a continuum, then the local charge density, ρ, will be given by:

$$\rho = eN_e \partial\xi/\partial z \qquad (2.3.1)$$

We can relate this to the electric field using the first of Maxwell's equations (2.1.1); since our system is uniform in the x and y directions, the divergence has only the z component, so we write:

$$\epsilon \partial E/\partial z = \rho = eN_e \partial \xi/\partial z \qquad (2.3.2)$$

We can also relate the field strength to the electron acceleration in a straightforward way:

$$-eE_z = m_e \partial^2 \xi/\partial t^2 \qquad (2.3.3)$$

Here, m_e is the electron mass. In this microscopic theory, we are treating the metal as a collection of point charges in a vacuum, so the permittivity ϵ will be that of free space. We now combine (2.3.2) and (2.3.3) to get:

$$\frac{e^2 N_e}{m_e \epsilon_0} \frac{\partial \xi}{\partial z} = - \frac{\partial}{\partial z} \frac{\partial^2 \xi}{\partial t^2} \qquad (2.3.4)$$

The solution to this equation is a temporally periodic oscillation, of the form $\xi=\xi(z)\exp(-j\omega_p t)$, where the frequency ω_p is given by:

$$\omega_p^2 = \frac{e^2 N_e}{m_e \epsilon_0} \qquad (2.3.5)$$

This is called the plasma frequency. Equation (2.3.4) gives no restriction on the form of $\xi(z)$; the dispersion relation is just the line $\omega=\omega_p$, so the group velocity is zero, and no information or power can be transported by bulk plasma waves. The value of ω_p depends only on the charge density. If we think of a metal as a plasma of free electrons in a fixed matrix of positive ions, we can calculate its plasma frequency accordingly; for most metals the value of ω_p will be in the near ultra-violet.

In order to determine the complex permittivity of a plasma, we need to calculate the conductivity. Again, we will consider the regime where collisions do not play an important role. The acceleration of the electrons will then be proportional to the field amplitude, according to (2.33), and the current density J will be proportional to the velocity of the electrons:

$$J = -eN_e \partial \xi/\partial t \qquad (2.3.6)$$

The conductivity is defined by $\sigma \equiv J/E$, so we get:

$$\sigma = \frac{-eN_e(-j\omega\xi)}{-m_e(-\omega^2\xi)/e} \qquad (2.3.7)$$

This we can simplify, using (2.3.5), to:

$$\sigma = j\epsilon_0\omega_p^2/\omega \qquad (2.3.8)$$

We can see, therefore, that in the absence of collisions, the current and field are $\pi/2$ out of phase, so that the high frequency conductivity of a plasma is imaginary. Combining (2.3.8) with (2.1.8) and (2.1.9), we can obtain the complex dielectric constant of a plasma:

$$\epsilon = 1 - \omega_p^2/\omega^2 \qquad (2.3.9)$$

We can see from (2.3.9) that for frequencies below the plasma frequency, the dielectric constant is a negative real. This is what we require for a material that will support a surface wave.

Collisions play an increasingly important role in the conductivity of metals as the frequency is reduced; in the low-frequency ohmic regime the conductivity is real, and the dielectric constant is consequently imaginary.

2.4 The Surface Plasmon

We have now seen that non-radiative surface waves can exist, with TM polarization, at the surface of a material having a negative dielectric constant, and that plasmas satisfy this requirement for frequencies below but near the plasma frequency as given by (2.3.5). As metals have plasma frequencies typically in the near ultra-violet, they will behave as plasmas at optical frequencies. In this work, as in most of the published literature on surface plasma waves, we are concerned with waves at the interfaces between metals and dielectric materials, excited at visible or near-visible wavelengths. The quanta of these waves are known as surface plasmons. Labelling the dielectric constants of the metal and dielectric as ϵ_m and ϵ_d respectively, we can rewrite (2.2.11) as:

$$\beta = k_0 \left[\frac{\epsilon_m\epsilon_d}{\epsilon_m + \epsilon_d}\right]^{\frac{1}{2}} \qquad (2.4.1)$$

For the dielectric, the index of refraction n_d will be given by:

$$n_d = \epsilon_d^{\frac{1}{2}} \qquad (2.4.2)$$

If ϵ_m is given by (2.3.9) then we can write (2.4.1) as:

$$\beta = n_d k_0 \left[\frac{\omega^2 - \omega_p^2}{\omega^2 - \omega_p^2}\right]^{\frac{1}{2}} \qquad (2.4.3)$$

We have now obtained an explicit dispersion relation for surface plasmons, which is

shown in the figure below.

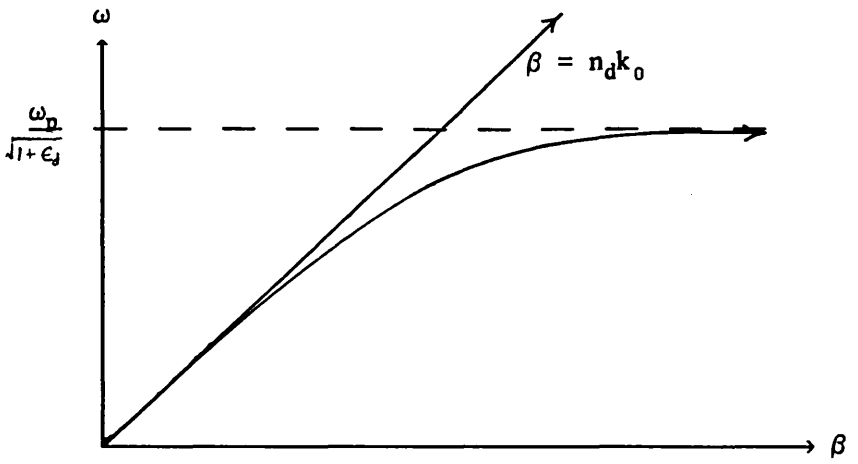


Figure 2.3 Dispersion relation for surface plasmons. The straight line gives the dispersion relation for light, in the dielectric, propagating in the z direction.

Since the plasmon propagation vector is always greater than that of light in the dielectric, the plasmon does not radiate into the bulk, and equivalently cannot be generated by light incident on the surface, since it is not possible to phase match along the interface.

In practice, collisions do play some role in the conductivity of metals at visible wavelengths, so σ does have a small real part, and consequently the complex permittivity has a small imaginary part. It is this characteristic that makes the propagation of plasmons lossy, which means that β is complex. Let us write the metal dielectric constant as:

$$\epsilon_m = -\epsilon_r + j\epsilon_x \qquad (2.4.4)$$

where ϵ_r and ϵ_x are both positive reals. Assuming that $\epsilon_x \ll \epsilon_r$, we can obtain approximate expressions for the real and imaginary parts of (2.4.1):

$$\frac{\beta}{k_0} \cong \left[\frac{\epsilon_r\,\epsilon_d}{\epsilon_r - \epsilon_d}\right]^{\frac{1}{2}} + j\frac{\epsilon_x}{2\epsilon_r^2}\left[\frac{\epsilon_r\,\epsilon_d}{\epsilon_r - \epsilon_d}\right]^{3/2} \qquad (2.4.5)$$

Here the real part gives the spatial frequency of the plasmon in the direction of propagation, while the imaginary part is a decay coefficient, indicating the absorption of energy due to collision losses in the metal.

2.5 Surface Waves on a Layer of Finite Thickness

In order to generate surface plasmons optically using the ATR method, one of the two media must be of finite thickness, so that a third medium (the coupling prism) can be introduced. We therefore need to determine how the propagation vector β differs in a three-phase structure from that determined previously (equation 2.4.5).

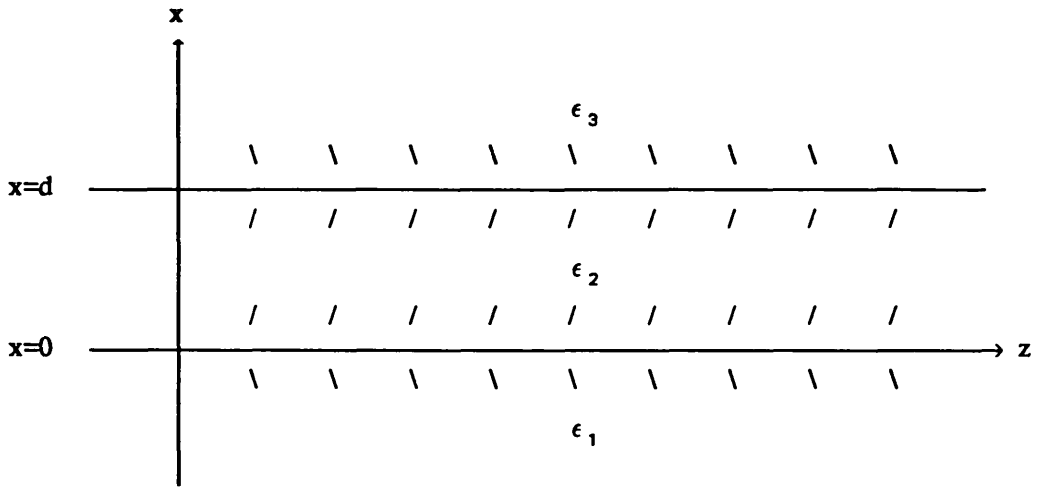


Figure 2.4 Plane boundaries separate medium 2, of finite thickness d, from semi-infinite media 1 and 3.

We begin the solution by assuming that the fields in each of the three media have the form given by (2.2.4) and (2.2.5). Because medium 2 has finite thickness, its field does not have to go to zero for large x. Therefore it can have field components with both positive and negative γ values. We can thus write the total field as:

$$H = \begin{cases} A \exp[j(\beta z + \gamma_1 x)] & x<0 \\ B \exp[j(\beta z + \gamma_2 x)] + C \exp[j(\beta z - \gamma_2 (x-d))] & 0<x<d \\ D \exp[j(\beta z + \gamma_3 (x-d))] & x>d \end{cases} \qquad (2.5.1)$$

Continuity of H at the two boundaries gives us:

$$A = B + C \exp(j\gamma_2 d) \qquad (2.5.2)$$

$$B \exp(j\gamma_2 d) + C = D \qquad (2.5.3)$$

We know from (2.2.7) that ϵE_z is proportional to $\partial H/\partial x$. Therefore, continuity of E_z gives us the following conditions:

$$\frac{\gamma_1 A}{\epsilon_1} = \frac{\gamma_2}{\epsilon_2} (B - C \exp(j\gamma_2 d)) \qquad (2.5.4)$$

$$\frac{\gamma_3 D}{\epsilon_3} = \frac{\gamma_2}{\epsilon_2}(B\exp(j\gamma_2 d) - C) \tag{2.5.5}$$

Let us also introduce the quantities:

$$q = \exp(j\gamma_2 d) \tag{2.5.6}$$

$$J = \gamma_1\epsilon_2/\gamma_2\epsilon_1 \tag{2.5.7}$$

$$K = \gamma_3\epsilon_2/\gamma_2\epsilon_3 \tag{2.5.8}$$

We can now write our four boundary conditions as:

$$-A + B + qC = 0 \tag{2.5.9}$$

$$qB + C - D = 0 \tag{2.5.10}$$

$$JA - B + qC = 0 \tag{2.5.11}$$

$$-qB + C + KD = 0 \tag{2.5.12}$$

The solution of this set of homogeneous linear equations can be found by setting the determinant of the associated matrix equal to zero. Thus:

$$\begin{vmatrix} -1 & 1 & q & 0 \\ 0 & q & 1 & -1 \\ J & -1 & q & 0 \\ 0 & -q & 1 & K \end{vmatrix} = 0 \tag{2.5.13}$$

The solution to the equation can then be written as:

$$q^2(J+1)(K-1) - (J-1)(K+1) = 0 \tag{2.5.14}$$

$$q^2 = \left[\frac{J-1}{J+1}\right]\left[\frac{K+1}{K-1}\right] \tag{2.5.15}$$

Using our definitions for q, J and K, we can expand this equation to give:

$$\exp(j2\gamma_2 d) = \left[\frac{\gamma_1\epsilon_2 - \gamma_2\epsilon_1}{\gamma_1\epsilon_2 - \gamma_2\epsilon_1}\right]\left[\frac{\gamma_3\epsilon_2 + \gamma_2\epsilon_3}{\gamma_3\epsilon_2 + \gamma_2\epsilon_3}\right] \tag{2.5.16}$$

As the γ values are all functions of β according to (2.2.5), (2.5.16) is a transcendental equation for β. We shall now be concerned with looking for its solutions.

Because surface plasmons are studied using reflectivity measurements, it is

useful to consider the relationship between the modal solution for the double-interface system, as given above, and the reflection equations for the same system. This we do by noting that the two factors on the right side of (2.5.16) have the same form as the Fresnel reflection coefficients for plane waves obliquely incident on boundaries between two half spaces. These can be written as [2.1] :

$$r_{ij} = \frac{\gamma_i \epsilon_j - \gamma_j \epsilon_i}{\gamma_i \epsilon_j + \gamma_j \epsilon_i} \qquad (2.5.17)$$

Here we use the labelling convention that r_{ij} is the amplitude reflection coefficient for a wave incident in medium *i* on the border with medium *j* (if both media are semi-infinite). To rewrite (2.5.16) in terms of *r* values, we must consider the signs of the γ values. In the r_{ij} equations, these would all be taken as positive for waves incident in the $+\hat{x}$ direction (up). The signs in (2.5.16) are for decay away from the boundaries, so γ_2 and γ_3 are positive and γ_1 is negative. This means we can rewrite (2.5.16) as:

$$\exp(j2\gamma_2 d) = (r_{12} r_{32})^{-1} \qquad (2.5.18)$$

For a very thick middle layer, q^2 will be approximately equal to zero, and therefore the two solutions to (2.5.16) will be given by $\gamma_1 \epsilon_2 = \gamma_2 \epsilon_1$ and $\gamma_2 \epsilon_3 = -\gamma_3 \epsilon_2$. These are just the single boundary solutions, as in (2.2.9), for each of the two boundaries, noting that γ_2 will have opposite signs in the two cases. If we begin to decrease the thickness d, the two solutions will be perturbed from their single interface values. We shall look for an approximate expression for this perturbation, in cases where it is small. We shall also assume that ϵ_1 and ϵ_3 are substantially different so that the two solutions are well removed from one another. Let us consider the solution at the 1–2 boundary, which we will call β_0, and label as $\Delta\beta$ its deviation from the d=∞ solution β_∞, so:

$$\beta_0 \equiv \beta_\infty + \Delta\beta \qquad \text{where} \qquad (\beta_\infty/k_0)^2 = \epsilon_1 \epsilon_2/(\epsilon_1+\epsilon_2) \qquad (2.5.19)$$

The approximation we will make is that in the vicinity of the 1–2 solution, the function r_{12} is rapidly varying, whereas r_{23} and q are much more slowly varying; therefore we can take the values of the latter two to be the same at β_0 as they are at β_∞. Then we write:

$$r_{12}^{-1} = r_{23}(\beta_\infty)\, q^2(\beta_\infty) \qquad (2.5.20)$$

Assuming the perturbation of β is small, we can use the approximation $\beta_0^2 = \beta_\infty^2 + 2\beta_\infty\Delta\beta$. Then since $\gamma^2 = \epsilon k_0^2 - \beta^2$, we can write:

$$\gamma_1^2 = \epsilon_1 k_0^2 - \epsilon_1 \epsilon_2/(\epsilon_1+\epsilon_2)\, k_0^2 - 2\beta_\infty\Delta\beta \qquad (2.5.21)$$

$$\gamma_1^2 = \epsilon_1^2/(\epsilon_1+\epsilon_2)\, k_0^2 - 2\beta_\infty\Delta\beta \tag{2.5.22}$$

$$\gamma_1^2 = \left((\epsilon_1/\epsilon_2) - 2\Delta\beta/\beta_\infty \right) \beta_\infty^2 \tag{2.5.23}$$

In evaluating the γ expressions, care must be taken in choosing the signs of the roots. We are using the convention that γ_1 is negative imaginary, and both terms in the brackets in (2.5.23) are negative. Taking the negative root, and using $\Delta\beta \ll \beta_\infty$ we get:

$$\gamma_1 = - (\epsilon_1/\epsilon_2)^{\frac{1}{2}}\beta_\infty - (\epsilon_2/\epsilon_1)^{\frac{1}{2}}\Delta\beta \tag{2.5.24}$$

Similarly:

$$\gamma_2 = (\epsilon_2/\epsilon_1)^{\frac{1}{2}}\beta_\infty + (\epsilon_1/\epsilon_2)^{\frac{1}{2}}\Delta\beta \tag{2.5.25}$$

In (2.5.24) and (2.5.25) the correct signs are indicated outside the roots. Inserting these expressions in the reflection coefficient gives:

$$r_{12}^{-1} = \frac{(\epsilon_1\epsilon_2)^{\frac{1}{2}}\beta_\infty + \epsilon_1(\epsilon_1/\epsilon_2)^{\frac{1}{2}}\Delta\beta - (\epsilon_1\epsilon_2)^{\frac{1}{2}}\beta_\infty - \epsilon_2(\epsilon_2/\epsilon_1)^{\frac{1}{2}}\Delta\beta}{(\epsilon_1\epsilon_2)^{\frac{1}{2}}\beta_\infty + \epsilon_1(\epsilon_1/\epsilon_2)^{\frac{1}{2}}\Delta\beta + (\epsilon_1\epsilon_2)^{\frac{1}{2}}\beta_\infty + \epsilon_2(\epsilon_2/\epsilon_1)^{\frac{1}{2}}\Delta\beta} \tag{2.5.26}$$

Again using $\Delta\beta \ll \beta_\infty$, we obtain:

$$r_{12}^{-1} = \frac{\Delta\beta}{2\beta_\infty} \frac{\epsilon_1(\epsilon_1/\epsilon_2)^{\frac{1}{2}} - \epsilon_2(\epsilon_2/\epsilon_1)^{\frac{1}{2}}}{(\epsilon_1\epsilon_2)^{\frac{1}{2}}} \tag{2.5.27}$$

$$r_{12}^{-1} = \frac{\Delta\beta}{2\beta_\infty} \frac{\epsilon_1^2 - \epsilon_2^2}{\epsilon_1\epsilon_2} \tag{2.5.28}$$

$$r_{12}^{-1} = - \Delta\beta\, (\epsilon_2-\epsilon_1)\, k_0^2/2\beta_\infty^3 \tag{2.5.29}$$

Using (2.5.20) and (2.5.29), we can now write an expression for the perturbation:

$$\Delta\beta = - \frac{2k_0}{\epsilon_2-\epsilon_1}\left[\frac{\epsilon_1\epsilon_2}{\epsilon_1+\epsilon_2}\right]^{3/2} r_{23}\, \exp(j2\gamma_2 d) \tag{2.5.30}$$

As we are mainly interested in the Kretschmann configuration, we will take the middle medium as the metallic one, so that ϵ_2 is given by (2.4.4). For the purposes of the perturbation analysis, the imaginary part can be neglected, so that $\epsilon_2 = -\epsilon_r$. The lower medium will be the dielectric $\epsilon_1 = \epsilon_d = n_d^2$. We can also approximate γ_2 (as given by 2.2.5) as $j\sqrt{\epsilon_r}k_0$. This gives:

$$\Delta\beta = \frac{2k_0}{\epsilon_r+\epsilon_d}\left[\frac{\epsilon_d\epsilon_r}{\epsilon_r-\epsilon_d}\right]^{3/2} \exp(-2\sqrt{\epsilon_r}k_0 d)\, r_{32}(\beta_\infty) \tag{2.5.31}$$

All the quantities in this expression are real except r_{32}, which in the Kretschmann configuration is the reflection coefficient at the prism/metal boundary for an infinite

metal thickness. If γ_3 is real, which it necessarily will be in a coupling prism, then the magnitude of r_{32}, as given by (2.5.17), is unity, and consequently r_{32} is only a phase factor. The real part of (2.5.31) gives the increase in the real part of β, which will correspond to an increased coupling angle for optical excitation. The imaginary part gives the loss due to radiation into medium 3 (the coupling prism). This result is given without derivation in the original publication by Kretschmann [2.2]; it is also found in the frequently referenced review paper by Raether [2.3], but again it is not derived, and in fact contains an error in the argument of the exponent (possibly typographical).

2.6 Optical Coupling Characteristics for the Kretschmann Configuration

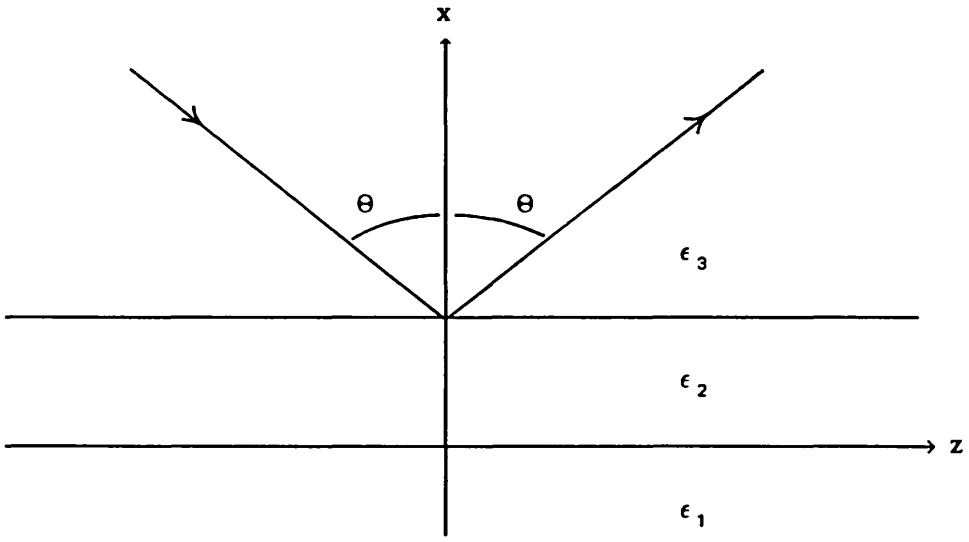


Figure 2.5 A plane wave, propagating at an angle θ to the surface normal, is incident on the boundary between media 3 and 2.

In the experiments conducted in this work, and in most of those reported in the literature, plasmons are generated using light incident in a coupling prism on a metallic layer of finite thickness (the Kretschmann configuration), as described in chapter 1. The plasmon properties are determined by measuring the reflectivity at the prism base as a function of incident angle. This is depicted in figure 2.5; a plane wave in medium 3 (the prism) is incident on the metal film at an angle θ to the surface normal. If this incident angle is greater than the critical angle for total internal reflection ($\sin\theta_c=n_1/n_3$), there will be no power propagated into medium 1; consequently, any power not absorbed by the metal will be reflected back into the prism, at an equal angle to the normal (Snell's law). The magnitude

of the propagation vector of both the incident and reflected waves is $n_p k_0$, where n_p is the refractive index of the prism ($\epsilon_3 \equiv n_p^2$), and the parallel component of this wavevector is given by $\beta = n_p k_0 \sin\theta$.

Using the labelling conventions defined previously in this chapter, we can write the amplitude reflectivity of a three phase structure as [2.1] :

$$r = \frac{r_{32} + r_{21}\, q^2}{1 + r_{32} r_{21}\, q^2} \tag{2.6.1}$$

where the single interface reflection coefficients r_{ij} are as defined in (2.5.17). This equation will allow us to plot exactly the form of the reflectivity for any specific combination of physical parameters. It does not, however, give us much insight on the physics involved. What we will seek is an approximate equation in the vicinity of resonance which will explicitly give the form of the coupling curve, and the dependence of this form on the parameters of the system.

We will approximate r above by assuming, as we did in section 2.5, that q^2 and r_{32} are both very slowly varying compared to r_{21}, and so approximate them as constants. Let us also introduce the notation $r_{ij} = z_{ij}/n_{ij}$, where:

$$z_{ij} = \epsilon_j \gamma_i - \epsilon_i \gamma_j \tag{2.6.2}$$

$$n_{ij} = \epsilon_j \gamma_i + \epsilon_i \gamma_j \tag{2.6.3}$$

We can now rewrite (2.6.1) as:

$$r = \frac{n_{21} + z_{21} r_{32}^{*}\, q^2}{n_{21} + z_{21} r_{32}\, q^2}\; r_{32} \tag{2.6.4}$$

Here we have used the fact that because the magnitude of r_{32} is unity, $r_{32}^{-1} = r_{32}^{*}$.

It is n_{21} that goes to zero when $\beta = \beta_\infty$, so we will write n_{21} as an expansion about β_∞. But first we must introduce some new notation. We have seen that in the finite layer solution there are two decay mechanisms, absorbtion loss due to collisions in the metal, and radiation loss into the prism. These are given, respectively, by the imaginary parts of β_∞ and $\Delta\beta$. We will label these two quantities Γ_i and Γ_r respectively, while the real parts of β_∞, $\Delta\beta$ and β_0 we will call $\beta_{\infty r}$, $\Delta\beta_r$ and β_{0r}. So we can write:

$$\beta_0 \equiv \beta_\infty + \Delta\beta = \beta_{0r} + j(\Gamma_i + \Gamma_r) \tag{2.6.5}$$

$$\beta_\infty \equiv \beta_{\infty r} + j\Gamma_i \tag{2.6.6}$$

$$\Delta\beta \equiv \Delta\beta_r + j\Gamma_r \tag{2.6.7}$$

For plane wave reflection, all amplitudes will be uniform in the $\hat{z}$ direction, so β will be real. We will write it as an expansion about $\beta_{\infty r}$:

$$\beta = \beta_{\infty r} + \rho = \beta_\infty + \rho - j\Gamma_i \tag{2.6.8}$$

Let us now evaluate n_{21}. We begin by finding expressions for γ_1 and γ_2. Here we will use the approximation that both ρ and Γ_i are small compared to β, so that:

$$\beta^2 \cong \beta_\infty^2 + 2\beta_\infty (\rho - j\Gamma_i) \tag{2.6.9}$$

Then we can write:

$$\gamma_1 = k_0 \left[\epsilon_1 - \frac{\epsilon_1\epsilon_2}{\epsilon_1+\epsilon_2} - \frac{2\beta_\infty(\rho-j\Gamma_i)}{k_0^2} \right]^{\frac{1}{2}} \tag{2.6.10}$$

We combine the first two terms in the brackets, and approximate the square root, again using $\rho \ll \beta_\infty$ and $\Gamma_i \ll \beta_\infty$:

$$\gamma_1 = k_0 \frac{\epsilon_1}{(\epsilon_1+\epsilon_2)^{\frac{1}{2}}} \left[1 - (\rho-j\Gamma_i) \frac{\beta_\infty}{k_0^2} \frac{\epsilon_1+\epsilon_2}{\epsilon_1^2} \right] \tag{2.6.11}$$

And similarly:

$$\gamma_2 = -k_0 \frac{\epsilon_2}{(\epsilon_1+\epsilon_2)^{\frac{1}{2}}} \left[1 - (\rho-j\Gamma_i) \frac{\beta_\infty}{k_0^2} \frac{\epsilon_1+\epsilon_2}{\epsilon_2^2} \right] \tag{2.6.12}$$

The minus sign appears in γ_2 because in the reflection coefficient we have to take the same sign for both γ values, and ϵ_2 is negative. We can now write an expression for n_{21} :

$$n_{21} = - \frac{\epsilon_1\epsilon_2 k_0}{(\epsilon_1+\epsilon_2)^{\frac{1}{2}}} (\rho-j\Gamma_i) \frac{\beta_\infty}{k_0^2} \frac{(\epsilon_1+\epsilon_2)(\epsilon_1^2-\epsilon_2^2)}{\epsilon_1^2\,\epsilon_2^2} \tag{2.6.13}$$

$$n_{21} = G\,(\rho - j\Gamma_i) \tag{2.6.14}$$

where

$$G = - \frac{\epsilon_2^2 - \epsilon_1^2}{(\epsilon_1\,\epsilon_2)^{\frac{1}{2}}} \tag{2.6.15}$$

For z_{21} we can neglect the ρ and Γ_i terms, giving:

$$z_{21} = - \frac{2\epsilon_1\epsilon_2 k_0}{(\epsilon_1+\epsilon_2)^{\frac{1}{2}}} \tag{2.6.16}$$

$$z_{21} = \left[\frac{\epsilon_1\epsilon_2}{\epsilon_1+\epsilon_2}\right]^{3/2} G \frac{2k_0}{\epsilon_2 - \epsilon_1} \tag{2.6.17}$$

This we can further approximate as:

$$z_{21} = -\left[\frac{\epsilon_d \epsilon_r}{\epsilon_d + \epsilon_r}\right]^{3/2} G \frac{2k_0}{\epsilon_r + \epsilon_d} \tag{2.6.18}$$

From this expression, and using (2.5.31), we obtain:

$$z_{21}\, q^2\, r_{32} = -G\,(\Delta\beta_r + j\Gamma_r) \tag{2.6.19}$$

$$z_{21}\, q^2\, r_{32}^{*} = -G\,(\Delta\beta_r - j\Gamma_r) \tag{2.6.20}$$

We can now rewrite the reflectivity equation (2.6.1) as:

$$r = \frac{\rho - j\Gamma_i - \Delta\beta_r + j\Gamma_r}{\rho - j\Gamma_i - \Delta\beta_r - j\Gamma_r}\, r_{32} \tag{2.6.21}$$

Since $\rho = \beta - \beta_{or}$, and $\beta_{or} + \Delta\beta_r = \beta_r$, we can rewrite this as:

$$r = \frac{(\beta - \beta_r) - j(\Gamma_i - \Gamma_r)}{(\beta - \beta_r) - j(\Gamma_i + \Gamma_r)}\, r_{32} \tag{2.6.22}$$

We have now obtained a useful approximate equation for the reflectivity in the vicinity of maximum coupling, from which the amplitude and phase behaviour are directly expressed in terms of the incident parallel wavevector β and the two loss terms Γ_i and Γ_r. As most experiments measure the intensity of the reflected beam as a function of angle, we need to obtain the intensity reflection coefficient $R = r^{*}r$. As the magnitude of r_{32} is one, we get:

$$R = \frac{(\beta - \beta_r)^2 + (\Gamma_i - \Gamma_r)^2}{(\beta - \beta_r)^2 + (\Gamma_i + \Gamma_r)^2} \tag{2.6.23}$$

$$R = 1 - \frac{4\,\Gamma_i\,\Gamma_r}{(\beta - \beta_r)^2 + (\Gamma_i + \Gamma_r)^2} \tag{2.6.24}$$

The second term in (2.6.24) gives the dip in reflectivity due to surface plasmon coupling, and indicates that its shape is Lorentzian. This is an approximation, as we have seen, which will be most accurate when the resonance is sharp, i.e. when the loss terms Γ_i and Γ_r are small. As the losses increase, the shape of the reflectivity dip will become more and more a symmetrical. We can rewrite the reflectivity in terms of the angle of incidence θ, using $\beta = n_p k_0 \sin\theta$, and defining the resonance angle θ_r as the incident angle giving phase matching with the surface plasmon ($\beta_r = n_p k_0 \sin\theta_r$). Then we get:

$$R = 1 - \frac{4\,\Gamma_i\,\Gamma_r}{n_p^2 k_0^2 (\sin\theta - \sin\theta_r)^2 + (\Gamma_i + \Gamma_r)^2} \tag{2.6.25}$$

Although the reflected amplitude is probably the most important function for most practical applications, it is also useful to know the behaviour of the other fields, particularly the evanescent field in the dielectric medium, which we shall henceforth call the plasmon field. We will now show how these functions can be obtained. Let us begin by considering the structure of figure 2.4; this time we will add an extra component to the fields, to represent the incident wave. The fields are now given by:

$$H = \begin{cases} A\ \exp[j(\beta z + \gamma_1 x)] & x<0 \\ B\ \exp[j(\beta z + \gamma_2 x)] + C\ \exp[j(\beta z - \gamma_2 (x-d))] & 0<x<d \\ D\ \exp[j(\beta z + \gamma_3 (x-d))] + E\ \exp[j(\beta z - \gamma_3 (x-d))] & x>d \end{cases} \qquad (2.6.26)$$

Using the boundary conditions as in section 2.4, we obtain a matrix equation:

$$\begin{bmatrix} -1 & 1 & q & 0 & 0 \\ 0 & q & 1 & -1 & -1 \\ J & -1 & q & 0 & 0 \\ 0 & -q & 1 & K & -K \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ D \\ E \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad (2.6.27)$$

From this equation we can obtain relations between any two of the field amplitudes by elimination; for instance, we can solve for D/E, which is equal to the reflectivity r, and obtain an expression exactly equivalent to (2.6.1). We are interested in the plasmon field amplitude as a function of incident amplitude, so we solve for A/E. First, we add and subtract lines 2 and 4 of (2.6.27), giving equations for B and C:

$$C = -\frac{K-1}{2}D + \frac{K+1}{2}E \qquad (2.6.28)$$

$$B = \frac{K+1}{2q}D - \frac{K-1}{2q}E \qquad (2.6.29)$$

Line 3 gives us JA=B−qC; inserting (2.6.28) and (2.6.29) gives:

$$A = \{\ [(K+1) + q^2(K-1)]D - [(K-1) + q^2(K+1)]E\ \}\ /2qJ \qquad (2.6.30)$$

If the coupling is not too strong, then q^2 will be very small and we can neglect the q^2 terms in (2.6.30). Making the substitutions r=D/E and $r_{32}=(K-1)/(K+1)$, we obtain:

$$\frac{A}{E} = \frac{K-1}{2qJ}\left[\frac{r}{r_{32}} - 1\right] \qquad (2.6.31)$$

Near $\beta=\beta_{\infty}$, $J\cong 1$. In addition, we can manipulate our definition of r_{32} to obtain:

$$K-1 = 2r_{32}/(1+r_{32}) \qquad (2.6.32)$$

From (2.6.22) we can obtain an expression for $(r/r_{32} - 1)$. Inserting this, along with (2.6.32) and J=1, into (2.6.31) gives:

$$\frac{A}{E} = \frac{r_{32}q^{-1}}{1+ r_{32}} \quad \frac{j2\Gamma_r}{(\beta-\beta_r) - j(\Gamma_i+\Gamma_r)} \qquad (2.6.33)$$

This gives us the form of the rise in amplitude of the plasmon field at resonance, corresponding to the drop in reflected amplitude. It is also interesting to consider the intensity; we simply take the magnitude squared of (2.6.33) to obtain:

$$\left|\frac{A}{E}\right|^2 = \frac{q^{-2}}{|\,1+ r_{32}\,|} \quad \frac{4\Gamma_r^2}{(\beta-\beta_r)^2 + (\Gamma_i+\Gamma_r)^2} \qquad (2.6.34)$$

This equation indicates the field intensity enhancement that can be obtained at the surface of the metal using resonant optical excitation.

2.7 Sensitivity to Surface Parameters

As we have seen, the SPR curve takes the form of a sharp dip in the angular reflectivity response, the shape and position of which depend on the optical properties of the metal film and the dielectric material in contact with it. The sensitivity of the SPR curve to these parameters can be calculated for specific cases using the analysis presented in previous sections; however, it is useful to consider whether there are any general reasons why SPR should be a particularly useful sensing method. There are such reasons, and they can be relatively simply stated. The plasmon mode is a type of guided wave, and as such has many properties in common with guided modes in dielectric planar structures, which can also be excited using prism couplers. An important difference, however, is that the plasmon mode is guided by a single interface, and that most of the field energy is in what can be considered the cladding of the waveguide. This means the field is accessible for sensing purposes, much more than in most dielectric waveguides. The propagation characteristics of the plasmon are thus strongly influenced by the optical properties of this cladding region.

The prism coupled SPR measurement results in a strong enhancement of the

field intensity, and as this enhanced field is evanescent, it can be confined to a layer of thickness considerably less than a wavelength. Intensity gain could be obtained with a Fabry–Perot type resonant structure, but the confinement of the field would be less; consequently, a greater volume would have to be perturbed in some way in order to get a measurable change in response. The use of dielectric waveguides as sensors has been investigated in this research group [2.4], with promising results. Because these guides have very low propagation losses, the intensity enhancement can be much greater than with SPR, and the angular width of the (phase) response can be made almost arbitrarily small by reducing the coupling strength. The extent of the evanescent field in the cladding can also be less than a wavelength, although more than half the field energy will be in the core and the buffer region between the core and the coupling prism, and thus unavailable for sensing. In the silver film SPR measurement (at 633 nm), more than 99% of the guided field energy is in the accessible dielectric region. The practical limitation on obtaining sharp resonances and intense fields with the dielectric structure is this: the required lateral extent of both the coherence of the incident beam and the uniformity of the structure, to excite low-loss modes, increases with the modal propagation distance. As the guiding characteristics are highly dependent on structural properties, this places stringent requirements on the manufacturing precision for the guides. The plasmon structure is easier to construct; it consists of a single evaporated layer, of a thickness to which the guiding characteristics are not highly sensitive. Finally, SPR is inherently an amplitude modulating technique, which is often more convenient for experimental purposes than the phase modulation obtained with these other two methods.

If we examine equation (2.6.24), we see that the reflection coefficient will only fall to zero at phase-matching ($\beta=\beta_r$) if we satisfy:

$$\Gamma_r = \Gamma_i \qquad (2.7.1)$$

When this is true, the rates of energy dissipation by the propagating plasmon due to intrinsic damping and reradiation are equal. I have found it useful to consider the one-dimensional analogy of a transmission line coupled at its termination to a damped resonator. In this case, the incident energy is completely absorbed if the coupling strength is such as to double the loss rate of the oscillator (or halve its Q). This means that, as above, the contributions to energy dissipation of absorption and reradiation are equal. Thus we can think of (2.7.1) as an impedance matching condition. It is this condition that is usually used to determine the metal film thickness for an SPR measurement. For silver, with HeNe laser light incident at

633 nm, the thickness giving zero reflection at resonance is about 560 angstroms, a figure often repeated in the literature.

We will now examine some examples of the influence of various parameters on the SPR curve. In the course of this work, a computer program was written to calculate the reflectivity of a multi-layered system. This was done on an IBM PC using the Pascal language, and the calculations were performed using the method described by Hansen [2.1] and Abelès [2.5]. This is a matrix method which is commonly used, for instance, to design anti-reflection coatings. It assumes that the media are homogeneous, linear and isotropic, and that the interfaces between them are planar, parallel, and infinite in extent. The program is called OPTO; it calculates amplitude, phase and power reflectivity, and plots field profiles, and was used to generate many of the figures in this thesis, including all those in this section.

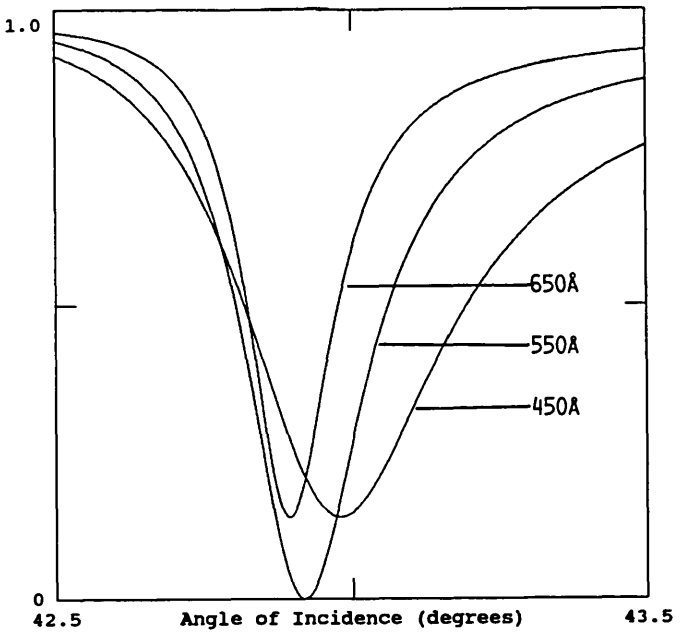


Figure 2.6 SPR curves for an Ag film, with various thicknesses as indicated, between a prism of index 1.52 and air. ($2\pi/k_0$ = 633 nm, ϵ_m = -16+j0.53)

Figure 2.6 shows the SPR curves for several thicknesses of silver film in the Kretschmann configuration. As the thickness is reduced, the coupling gets stronger, and consequently the resonance broadens. At one particular thickness, as predicted by the equations, the reflectivity minimum falls to zero. As the curves were calculated without making the approximations needed to obtain (2.6.24), they are not exactly symmetrical. Part of this asymmetry, however, is simply due to the fact

that the curves are plotted as functions of angle, while our equations have been in terms of β, which is proportional to $\sin\theta$.

Figure 2.7 shows the very high sensitivity of the SPR response to surface properties, in particular to adlayers (thin layers of dielectric deposited on the Ag surface). For a small perturbation, the best indication of sensitivity to some surface parameter is the angular shift in the resonance caused, compared to the resonance curve width. Even with high noise levels and simple instrumentation, it should be possible to measure quite easily a resonance shift of one tenth of the width at half-maximum. The curves shown indicate that such a shift would be produced by about an angstrom adlayer thickness. With a more precise measurement, it should be possible to improve this figure by at least an order of magnitude.

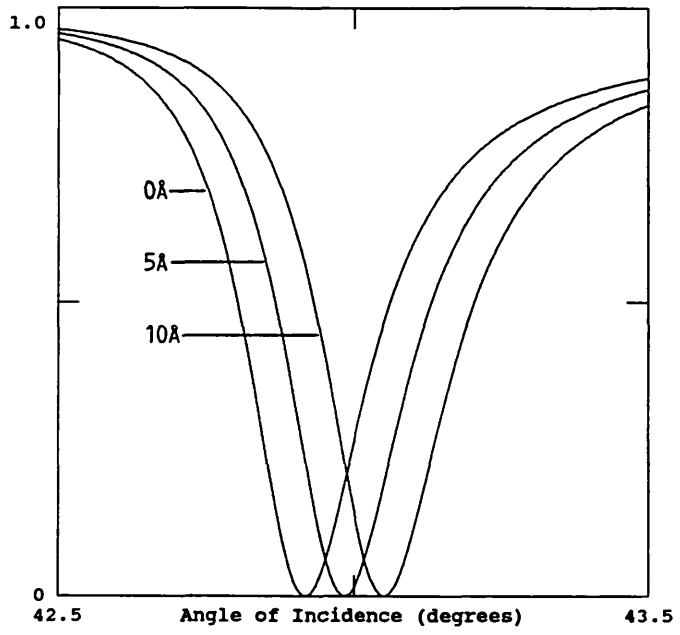


Figure 2.7 SPR curves for a 550 A Ag film, with various thicknesses as indicated of a dielectric coating of index 2.0. Other parameters as for figure 2.6.

The criterion of zero reflectivity is almost always used, in experimental work reported in the literature, to choose the metal film thickness. This allows, of course, the highest possible contrast in an intensity measurement, but it is not obvious that this should also give the highest possible sensitivity. For a very small perturbation of the system, the most sensitive detection method would be to have light incident at the angle of highest slope on the resonance curve, so that a small shift in resonance position gives the maximum signal change. Therefore I have calculated the effect of coupling strength (metal thickness) on the slope of the

resonance curve, and used this calculation to derive a new condition giving maximum sensitivity.

For convenience we will determine the slope as a function of β, rather than angle. Then we can use (2.6.24) directly, and write:

$$\partial R/\partial\beta = \frac{8\,\Gamma_i\,\Gamma_r\,(\beta - \beta_r)}{[(\beta - \beta_r)^2 + (\Gamma_i + \Gamma_r)^2]^2} \qquad (2.7.2)$$

To determine when this slope is maximum, we take the second derivative and set it equal to zero. This gives:

$$\frac{\partial^2 R}{\partial\beta^2} = 8\,\Gamma_i\,\Gamma_r \frac{(\beta - \beta_r)^2 + (\Gamma_i + \Gamma_r)^2 - 4(\beta - \beta_r)^2}{[(\beta - \beta_r)^2 + (\Gamma_i + \Gamma_r)^2]^3} = 0 \qquad (2.7.3)$$

This equation is satisfied if:

$$(\Gamma_i + \Gamma_r)^2 = 3(\beta - \beta_r)^2 \qquad (2.7.4)$$

Let us label the slope R'; the maximum, which we will call R'_m, is obtained by combining (2.7.2) and (2.7.4). This gives:

$$R'_m = \frac{3\sqrt{3}\,\Gamma_i\,\Gamma_r}{2(\Gamma_i + \Gamma_r)^3} \qquad (2.7.5)$$

If we now take the derivative of this function with respect to Γ_r, we will obtain the relationship between Γ_i and Γ_r that maximizes the slope.

$$\frac{\partial R'_m}{\partial\Gamma_r} = \frac{3\sqrt{3}}{2} \frac{\Gamma_i(\Gamma_i + \Gamma_r) - 3\Gamma_i\,\Gamma_r}{(\Gamma_i + \Gamma_r)^4} = 0 \qquad (2.7.6)$$

The optimum coupling to maximize sensitivity, therefore is obtained by satisfying the criterion:

$$\Gamma_r = \Gamma_i/2 \qquad (2.7.7)$$

Expressions for Γ_i and Γ_r are given by the imaginary parts of (2.4.5) and (2.5.31) respectively. Since Γ_r falls with thickness as e^{-d}, the difference in the thicknesses needed to satisfy each of the two criteria is not great; the increase in slope obtained in going from the zero reflectivity to the maximum slope condition is on the order of 5 to 10%.

The table below lists optical constants for several metals, at the frequency of the 633 nm HeNe line, obtained from reference [2.6]. Also listed are the

thicknesses needed to satisfy the criteria of zero reflectivity and maximum slope, given by d_1 and d_2 respectively. In practice, the optical constants of thin metal films will vary somewhat in manufacture, due to grain size variations, impurities introduced during evaporation, and possibly other effects; they are also temperature dependent. Remembering that $\epsilon_m = -\epsilon_r + j\epsilon_x$, we have:

Metal	ϵ_r	ϵ_x	d_1	d_2
Ag	16.9	0.55	560	640
Au	11.0	1.35	505	610
Cu	11.9	0.99	540	640
Al	39.6	14.7	145	200

Table 2.1 Complex relative permittivities for various metals, at $2\pi/k_0$ = 633 nm, and optimum thicknesses (in angstroms) for zero reflectivity (d_1) and maximum SPR slope (d_2).

The validity of the d_1 values is indicated by figure 1.4. Figure 2.8 shows SPR slope curves, also calculated using OPTO, for several thicknesses of each of the four metals above.

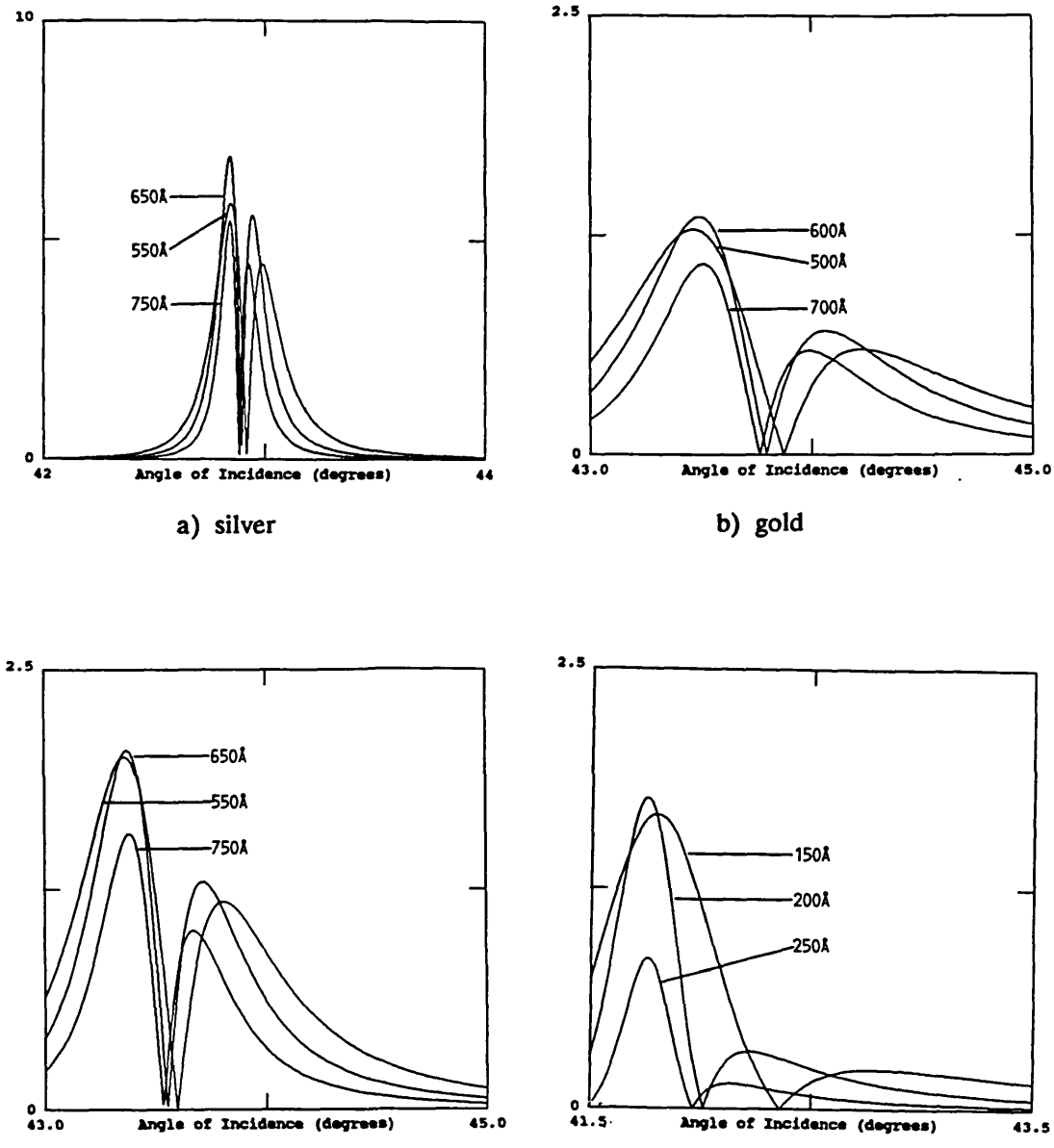


Figure 2.8 Calculated SPR resonance slopes ($dR/d\theta$), in units of degrees^{-1}, for the metals described in table 2.1, with thicknesses as indicated. Other parameters as in figure 2.6. Note the different vertical scale for silver.

2.8 Phase Considerations

The approximation we have derived for the amplitude reflectivity gives us the phase change on reflection as well as the amplitude. If we manipulate equation (2.6.22) we can obtain the real and imaginary parts as follows:

$$\mathrm{Re}\{r/r_{32}\} = \frac{(\beta - \beta_r)^2 + (\Gamma_i^2 - \Gamma_r^2)}{(\beta - \beta_r)^2 + (\Gamma_i + \Gamma_r)^2} \tag{2.8.1}$$

$$\mathrm{Im}\{r/r_{32}\} = \frac{2\Gamma_r (\beta - \beta_r)}{(\beta - \beta_r)^2 + (\Gamma_i + \Gamma_r)^2} \tag{2.8.2}$$

Calling the phase change on reflection φ, we obtain:

$$\varphi = \tan^{-1}\frac{2\Gamma_r (\beta - \beta_r)}{(\beta - \beta_r)^2 + (\Gamma_i^2 - \Gamma_r^2)} + \varphi_{32} \tag{2.8.3}$$

where φ_{32} is the phase change on reflection from an infinite metal thickness, i.e.:

$$r_{32} = \exp(j\varphi_{32}) \tag{2.8.4}$$

We can see from the equations above that $\mathrm{Im}\{r/r_{32}\}$ will always cross through zero at $\beta=\beta_r$; $\mathrm{Re}\{r/r_{32}\}$, however, will only be zero at this point if $\Gamma_i=\Gamma_r$, giving zero reflectivity.

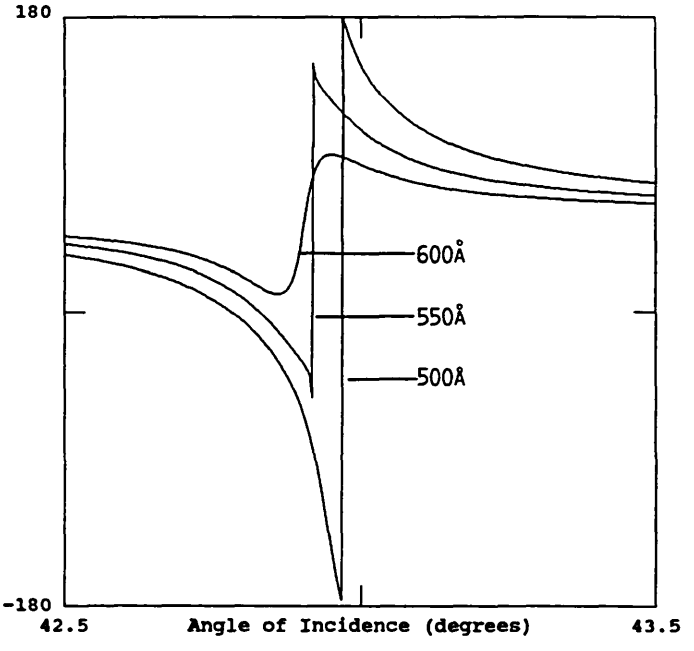


Figure 2.9 Phase change on reflection from a silver film, with thicknesses as indicated (values of parameters as in figure 2.4).

Figure 2.9 above shows examples of the phase response of a silver resonance, for various film thicknesses, calculated using OPTO. The vertical offset is due to φ_{32}.

It is instructive to consider the derivative of (2.8.3); this will give us an indication of the sensitivity of a phase-based measurement, and will also allow us to calculate the Goos-Haenschen shift. If the system is undercoupled ($\Gamma_r<\Gamma_i$), so that $\mathrm{Re}\{r/r_{32}\}$ remains positive, then the function (2.8.3) is reasonably well behaved, in that the phase remains within the range $(-\pi/2,\pi/2)$, and we can use the usual form for the derivative of $\tan^{-1}x$, which is $1/(1+x^2)$. Let us introduce two new variables for convenience:

$$\Lambda = \Gamma_i^2 - \Gamma_r^2 \tag{2.8.5}$$

$$\nu = \beta - \beta_r \tag{2.8.6}$$

Then we can obtain, for $\Lambda>0$:

$$\frac{\partial\varphi}{\partial\nu} = \frac{2\Gamma_r\,(\Lambda - \nu^2)}{(\nu^2+\Lambda)^2 + (2\Gamma_r\nu)^2} \tag{2.8.7}$$

The maximum slope will be at the resonance angle $\nu=0$, where:

$$(\partial\varphi/\partial\nu)_0 = 2\Gamma_r/\Lambda \tag{2.8.8}$$

This equation indicates that for exact matching ($\Lambda=0$), the slope of the phase change at resonance goes to infinity. This is because the imaginary part changes sign while the real part is zero. One might conclude that this offers a chance for an experimental measurement of unlimited sensitivity; however, there are two factors which limit this possibility. The first is that the high slope is associated with a signal of vanishing amplitude, so in practice the sensitivity of the apparatus, and the signal to noise ratio, will limit what can be achieved. Secondly, the high slope will only occur over a very limited angular range, and so to achieve it one would need a beam of equally narrow angular width. In practice, this would mean both coherence of the beam and uniformity of the film over large lateral distances.

For the undercoupled case above, the phase shift reverses direction twice, finally returning to its original value. For the overcoupled case where $\mathrm{Re}\{r/r_{32}\}$ goes below zero, the phase change is monotonic, with a 2π change over the whole resonance. In figure 2.10 below, the same phase scale of $-\pi$ to $+\pi$ is used for all plots, so that an artificial discontinuity appears in the overcoupled case. The three cases plotted correspond to slightly undercoupled, well matched, and slightly

overcoupled. Both the phase change and the real and imaginary parts of the reflectivity r are shown. In the well matched case there is an abrupt phase change of π.

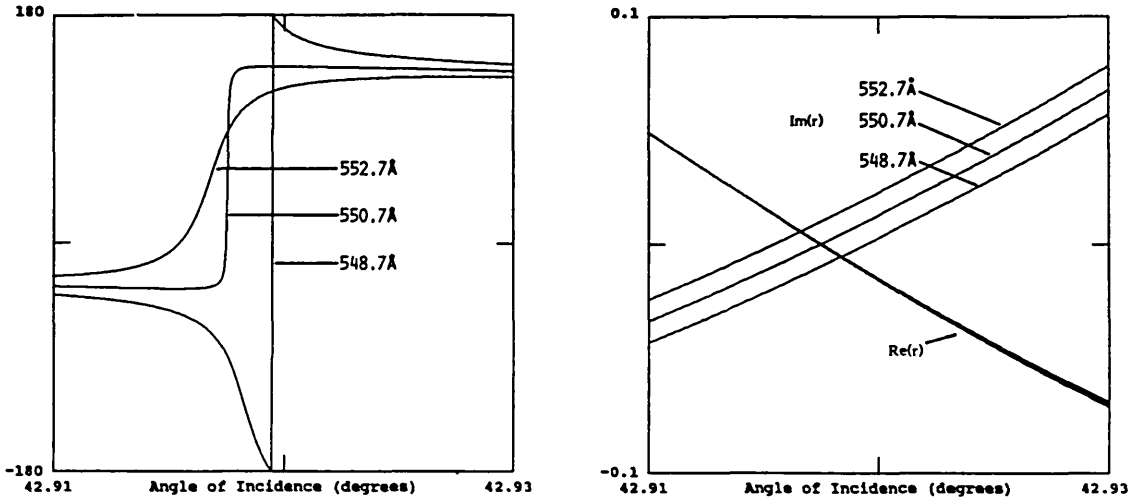


Figure 2.10 Detail of (a) phase change φ and (b) real and imaginary parts of r, for silver thicknesses as indicated.

The Goos-Haenschen shift describes the apparent lateral displacement of a beam of finite width, upon reflection from some structure. If we call the lateral shift (along the surface) z_0, then this is given by [2.7] :

$$z_0 = (\partial\varphi/\partial\theta)/(n_p k_0 \cos\theta) \qquad (2.8.9)$$

The sign is opposite from that in the reference because of our different sign convention for the exponential notation (see section 2.1). Since $\nu = n_p k_0 \sin\theta - \beta_r$, then $\partial\nu/\partial\theta = n_p k_0 \cos\theta$, and using $\partial\varphi/\partial\theta = (\partial\varphi/\partial\nu)(\partial\nu/\partial\theta)$, we get:

$$z_0 = 2\Gamma_r/\Lambda \qquad (2.8.10)$$

Using the approximation $\Lambda = \Gamma_i^2 - \Gamma_r^2 = 2\Gamma_r(\Gamma_i - \Gamma_r)$, we get:

$$z_0 \cong 1/(\Gamma_i - \Gamma_r) \qquad (2.8.11)$$

(Note that as these expressions use (2.8.8), they are also only valid for $\Lambda>0$; the corresponding equations for $\Lambda<0$ are similar but with the sign reversed so that they also give a positive shift).

2.9 Multi-Modal Systems

The Kretschmann configuration will, as we have seen in section 2.5, support two guided modes; if one medium is a coupling prism, the only mode that can be excited is the one on the low-index side of the metal, as the propagation vector of the plasmon mode on the prism side will be too high. Thus for the purposes of the prism coupled experiment, we are effectively dealing with a single mode waveguide. More complex structures, however, can support multiple modes with different characteristics; in particular, we can generate plasmon-like modes which have much lower loss (and thus greater propagation lengths) than the basic surface plasmon. These low-loss modes are of interest because of the potential increase in sensitivity, due to the sharpness of their resonance curves.

There are two main categories of plasmon-like modes, which we shall consider in turn. The first is produced by the coupling between modes on similar surfaces. In our treatment of the finite layer problem in section 2.5, we assumed that the indices on the two sides of the metal were sufficiently different to ensure that the two surface modes acted quite independently. If we considered instead a symmetrical system, such as a finite metal layer bounded on both sides by air, then this would no longer be the case. The solution would now consist of two guided modes both having an equal amount of field energy in the two dielectric media, but with the phase on the two sides being equal in one mode and opposite in the other. These modes are called symmetric and antisymmetric. One will have a greater proportion of field energy in the metal, and thus will have both a higher value of β and a shorter decay length. The low loss mode can have a propagation distance many times greater than that of the single-interface plasmon, although that of the high loss mode will be correspondingly reduced. For this reason they are frequently referred to in the literature as long-range and short-range surface plasmons [2.8],[2.9].

In order to excite these modes, we can introduce a coupling prism at some distance from the metal film on one side, as depicted in figure 2.11(a). Figure 2.11(b) shows an example of the sort of resonance curve that results from this configuration, with one sharp and one broad resonance, calculated using OPTO. Figure 2.12 shows calculated field intensity profiles for the two modes. In the higher loss mode, the intensity is zero at the center of the metal, indicating that the fields on opposite sides are out of phase.

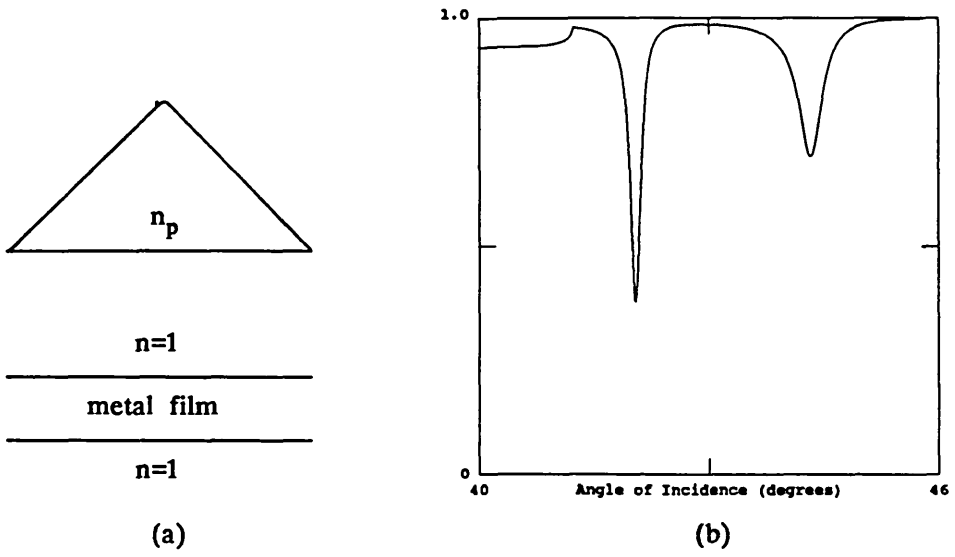


Figure 2.11 (a) Configuration for optical excitation of long-range and short-range surface plasmons; (b) angular resonance curve for a 450 angstrom silver film in such a configuration, with an air gap of 0.85 microns. ($2\pi/k_0$=633 nm)

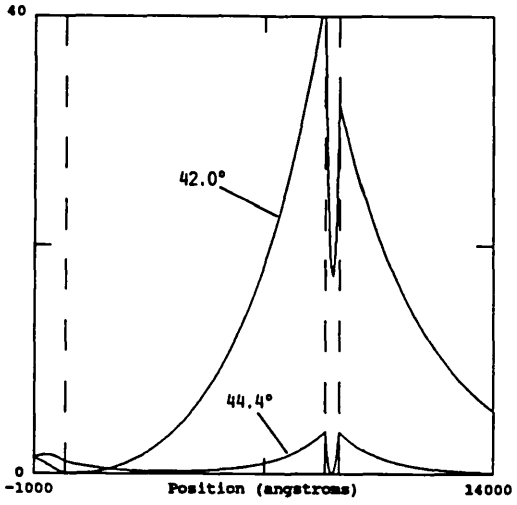


Figure 2.12 Field intensity profiles for the two resonances indicated in figure 2.11(b). The value plotted is the magnitude squared of H, normalized to an incident amplitude of H=1. Position is measured relative to the prism base, with dashed lines indicating boundaries between media.

The other method of obtaining plasmon-like modes with long propagation distances is by having a structure like a dielectric waveguide, but with one or both of the guiding boundaries being a metal surface. Thus we obtain a guided mode where part of the field is not evanescent. The simplest way to produce these modes is by placing a dielectric layer of sufficient thickness between the metal and air in the Kretschmann configuration. Such a system is illustrated in figure 2.13(a), the resulting resonance curve (calculated) is shown in figure 2.13(b), and the field

profiles for two of the modes are shown in figure 2.14. The effective index (β/k_0) of the plasmon mode is greater than the index of the dielectric layer, since almost all of the field will be in this medium. Between the critical angle and the plasmon resonance angle are seen a number of very sharp modes. These modes are evanescent in the silver and air, but periodic within the dielectric. The thicker this dielectric layer, the more such modes will be supported, just as in a conventional slab waveguide which is bounded by lower index dielectric materials on both sides. The propagation lengths of these modes can be orders of magnitude greater than that of the simple plasmon mode.

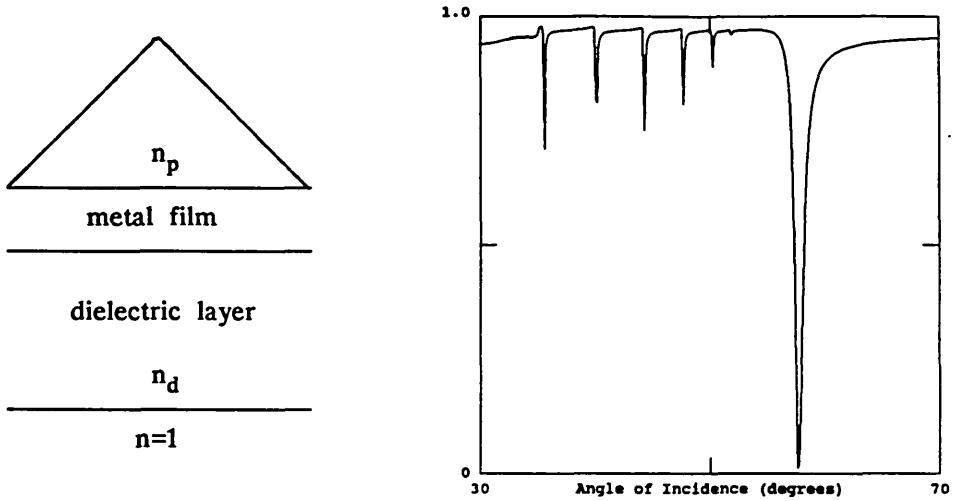


Figure 2.13 (a) Configuration for plasmon-like modes in a dielectric slab waveguide; (b) angular resonance curve for a 550 angstrom silver film in such a configuration, with a 2 μm dielectric layer of index 1.38. ($2\pi/k_0$=633 nm)

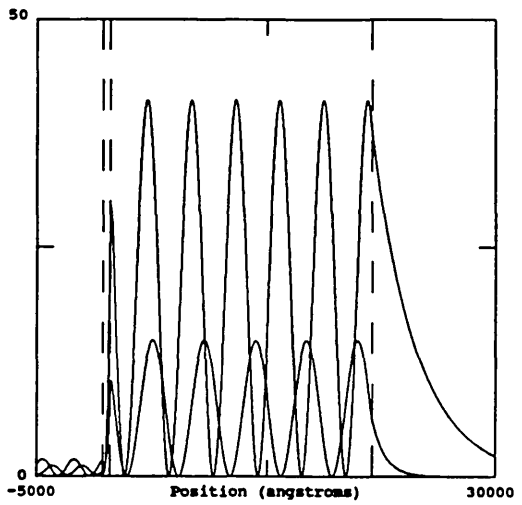


Figure 2.14 Field intensity profiles for two of the modes indicated in figure 2.13(b) (position measured relative to the prism base).

While both these mode types suggest possibilities for tremendous sensitivity due to the very sharp angular response, there are several limitations. Firstly, as in the case of the sharp phase response discussed in section 2.8, a narrow angular resonance requires high uniformity of the beam and structure over a large lateral extent; manufacturing and experimental considerations will thus place a limit on what can be achieved in this way. Even without this requirement for lateral uniformity, all these structures are more difficult to fabricate than the basic Kretschmann stucture. Also, in this particular project, we are concerned with measuring spatial variation, and thus we must consider the effect on lateral resolution of long propagation distances. This topic will be discussed in detail in chapter 4.

From the point of view of sensing in general, these special modes have an additional drawback, which is that less of the field is in the region where sensing occurs, and thus the modes are likely to be less affected by changes within this region. In the slab waveguide modes, most of the field is within the dielectric layer, as we have seen, while in the long-range surface plasmon about half the field is in the air gap between the prism and silver. However, it may be that a low-loss mode offers higher sensitivity despite this consideration, and indeed that the most sensitive method might be a phase measurement using a dielectric waveguide, which has effectively no absorption losses [2.4].

CHAPTER 3

PLASMONS ON NON-UNIFORM SURFACES

3.1 SPR Measurements Using Focussed Beams

The SPR reflectivity equations we derived in section 2.6 are for plane waves, which cannot in practice be individually obtained. However, any field within a uniform medium can be described as a unique angular spectrum of plane waves. We shall only be concerned with monochromatic fields, for which the angular spectra will in general be complex functions of two parameters (the three angular components of the wave vector being related by the wave equation).

Let us consider the incident beam shown in figure 2.5. In the prism (medium 3), plane waves will all have propagation vectors of magnitude $n_p k_0$. If we restrict ourselves to a two dimensional approximation, where both the system and the waves are constant in the $\hat{y}$ direction, then the direction of the plane wave propagation vectors can be described by a single parameter; this could be the angle θ, or equivalently the $\hat{z}$ component β. Let us take as our phase reference the point (z=0, x=d), and call the angular spectra of the incident and reflected beams $A_i(\theta)$ and $A_r(\theta)$ respectively. Then we can write:

$$A_r(-\theta) = r(\theta)\ A_i(\theta) \tag{3.1.1}$$

where $r(\theta)$ is the angular reflectivity function, given by (2.6.22) with $\beta=n_p k_0 \sin\theta$.

We can now calculate the effect of measuring the SPR reflectivity curve with a spatially confined beam, i.e. one with an angular spectrum of non-zero width. We will start with a beam having a fixed angular spectrum about a direction θ_c, which will be the angle of the axis of the beam to the surface normal. If we then rotate the prism relative to the beam, we will vary θ_c while keeping the relative angular spectrum $A_i(\theta-\theta_c)$ constant. If we measure the resonance curve by collecting all the reflected power in a detector, we will obtain the power reflectivity as a function of the axial incident angle θ_c. We will call this total reflectivity $R_t(\theta_c)$. Since power is proportional to the modulus squared of the amplitude, we can write:

$$R_t(\theta_c) = \frac{\int_0^{\pi/2} R(\theta)\ A_i(\theta-\theta_c)\ A_i^*(\theta-\theta_c)\ d\theta}{\int_0^{\pi/2} A_i(\theta-\theta_c)\ A_i^*(\theta-\theta_c)\ d\theta} \tag{3.1.2}$$

where $R(\theta)$ is the modulus squared of the amplitude reflectivity function, and is given by (2.6.25).

Equation (3.1.2) tells us that the total power reflectivity is given by the convolution of the plane wave reflectivity with the modulus squared of the incident angular spectrum. The most important consequence is that the resonance curve is broadened, which results in loss of sensitivity, and that this broadening increases with increased angular width of the incident beam. This also gives a condition on the extent of the illuminated area. The amplitude distribution at the metal surface will be given by the Fourier transform of the incident angular spectrum; thus the angular width increases as the size of the spot illuminated decreases. If we write the spectrum in terms of $(\beta-\beta_c)$ rather than $(\theta-\theta_c)$, and label as $a_i(z)$ the amplitude of the incident beam at the prism/metal interface, then

$$a_i(z) = \exp(j\beta_c z) \int_0^{n_p k_0} A_i(\beta-\beta_c) \exp[j(\beta-\beta_c)z] \, d\beta \qquad (3.1.3)$$

The amplitude reflectivity function, as given by equation (2.6.22), also is associated with a characteristic length; its Fourier transform is a decaying exponential of the form $\exp[-(\Gamma_i+\Gamma_r)z]$, as we shall show in section 3.3. Thus the total reflectivity (3.1.2) is the convolution of the magnitude squared of two functions, which are related to the illuminated area and the decay length $1/(\Gamma_i+\Gamma_r)$ respectively. The implication is that to keep the angular beam width low enough to avoid broadening the resonance significantly, the illuminated area must be greater than the characteristic plasmon decay length.

We can think of this requirement on illuminated area in terms of interaction length. The basis of sensing measurements using guided waves lies in determining the propagation vector of the guided mode with the maximum possible precision; this precision is proportional to the length over which the exciting field and the guided mode interact coherently. In the case of SPR, the maximum coherence length of the plasmon mode is its decay length; this gives a maximum precision in the measurement of β which is proportional to the resonance width. If both the exciting beam and the structure are uniform over at least this lateral distance, maximum sensitivity can be achieved.

We will now consider the resonance broadening effects on incident beams with specific angular spectra. By assuming that the plane wave resonance is very sharp, we can make the approximation:

$$\sin\theta - \sin\theta_r \cong (\theta - \theta_r) \cos\theta \qquad (3.1.4)$$

If we define a as the angular half-width of the plane wave resonance, then we

can write:

$$R(\theta) = 1 - \frac{\alpha^2}{x^2 + a^2} \tag{3.1.5}$$

where $x = \theta - \theta_r$, and the constant α is given by:

$$\alpha^2 \equiv \frac{4\Gamma_i\Gamma_r}{(n_p k_0 \cos\theta_r)^2} \tag{3.1.6}$$

We now introduce two new variables:

$$u \equiv \theta_c - \theta_r \tag{3.1.7}$$

$$y \equiv \theta - \theta_c = x - u \tag{3.1.8}$$

remembering that θ_c and θ_r are the axial incident angle of the beam and the resonant coupling angle respectively. We will take as our first example a rectangular incident angular spectrum of half-width b :

$$A_i(y) = \begin{cases} A_0 & |y| \leq b \\ 0 & |y| > b \end{cases} \tag{3.1.9}$$

Inserting (3.1.9) and (3.1.5) into (3.1.2) gives:

$$R_t(u) = \frac{A_0 \int_{u-b}^{u+b} \left[1 - \frac{\alpha^2}{x^2 + a^2}\right] dx}{A_0 \int_{-b}^{b} dx} \tag{3.1.10}$$

from which we obtain:

$$R_t(u) = 1 - \frac{\alpha^2}{2b}\left[\tan^{-1}\frac{u+b}{a} - \tan^{-1}\frac{u-b}{a} \right] \tag{3.1.11}$$

The angular half-width of this ~~measured~~ finite beam resonance, which we will label u_0, will be the value of u for which the bracketed expression in (3.1.11) has half the value it has at u=0. This we can write as:

$$\tan^{-1}\frac{u_0+b}{a} - \tan^{-1}\frac{u_0-b}{a} = \tan^{-1}(b/a) \tag{3.1.12}$$

Introducing the normalized variables $\gamma \equiv b/a$ and $v \equiv u_0/a$, we have:

$$\tan^{-1}(v+\gamma) - \tan^{-1}(v-\gamma) = \tan^{-1}(\gamma) \tag{3.1.13}$$

This is a transcendental equation for the normalized broadened resonance width v. In the limits of very narrow and very wide incident beams, we can use Taylor expansions to obtain the following approximations:

$$v = 1 + \gamma^2/3 \qquad \text{where } \gamma \ll 1 \tag{3.1.14}$$

$$v = \gamma + 1/2\gamma \qquad \text{where } \gamma \gg 1 \tag{3.1.15}$$

These expressions agree with the expected result; for a very narrow incident angular spectrum, the resonance width is only slightly greater than the plane wave ideal, while for very broad incident angular spectra, the resonance is slightly greater than the beam width. Another approximation,

$$u_0 = (a^2 + b^2)^{\frac{1}{2}} \qquad \text{or } v = (\gamma^2 + 1)^{\frac{1}{2}} \tag{3.1.16}$$

was found empirically which fits (3.1.14) and (3.1.15) quite well, and can be used over the entire range of γ. This function was found to fit the numerical solution of (3.1.13) to a maximum error of 1.5%. Figure 3.1 gives a comparison of the three approximations to the numerical solution of (3.1.13).

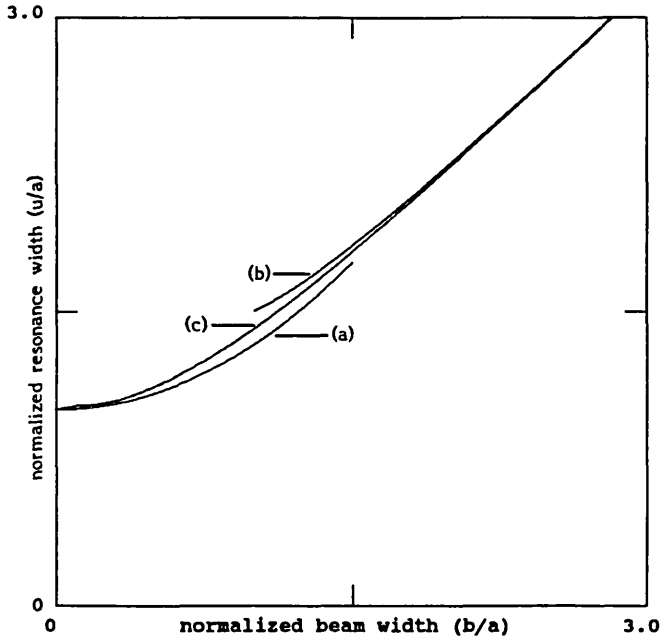


Figure 3.1 Normalized resonance half-width, as a function of normalized beam half-width, for a rectangular incident angular spectrum, using the approximations: a) $v = 1 + \gamma^2$; b) $v = \gamma + 1/2\gamma$; c) $v^2 = \gamma^2 + 1$; Exact numerical solution is also plotted, but is indistinguishable from (c).

We can relate the resonance broadening to the illuminated area in the rectangular beam case using equation (3.1.3). If the angular half-width of the beam is b, then using (3.1.4) we obtain a width in terms of β of $(n_p k_0 \cos\theta)b$. The amplitude at the prism base will be the Fourier transform of a rectangular function, which is a sinc function:

$$a_i(z) = a_0 \exp(j\beta_c z)\ \sin(n_p k_0 \cos\theta\ bz)/z \tag{3.1.17}$$

Let us characterize the illuminated area by a spot radius z_0 within which half the incident power falls:

$$\int_0^{z_0} \frac{\sin^2(\beta_0 z)}{z^2}\, dz = \frac{1}{2} \int_0^{\infty} \frac{\sin^2(\beta_0 z)}{z^2}\, dz \qquad (3.1.18)$$

where $\beta_0 = (n_p k_0 \cos\theta)b$. The second definite integral can be found in tables [3.1] to have a value of $\pi\beta_0/2$. By setting the first integral equal to $\pi\beta_0/4$, we can obtain a numerical solution for z_0, which is:

$$z_0 \cong 0.85/\beta_0 \qquad (3.1.19)$$

To obtain a relation between z_0 and the angular width b, we will assume typical values of $n_p = 1.52$, $k_0 = 2\pi/0.633\ \mu m^{-1}$, and $\cos\theta \cong 0.7$; then we get:

$$z_0 \cong 0.08/b \quad \mu m/rad \quad = 4.56/b\ \mu m/degree \qquad (3.1.20)$$

Combining (3.1.20) with (3.1.16) gives a relation between focussed spot size and resonance width:

$$u_0 = (a^2 + (4.56/z_0\ deg/\mu m)^2)^{\frac{1}{2}} \qquad (3.1.21)$$

In figure 3.2, this is plotted for several values of the plane wave resonance width a.

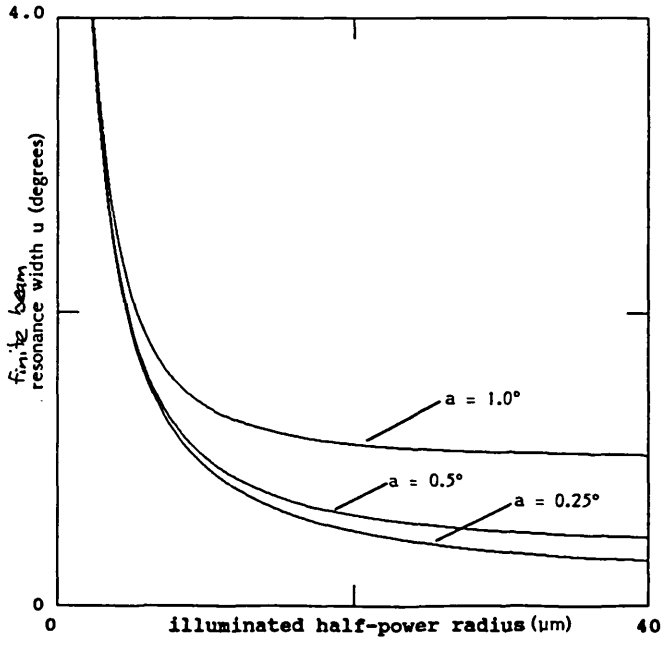


Figure 3.2 Effective resonance half-width u_0, as a function of focussed spot radius z_0, for a rectangular incident angular spectrum, plotted for several values of plane wave half-width a as indicated.

Another important example is that of a Gaussian incident angular spectrum. This we will write as:

$$A_i(y) = A_0 \exp(-y^2/b^2) \tag{3.1.22}$$

The total reflectivity will now be given by:

$$R_t(u) = \frac{A_0 \int_{-\theta_r}^{\pi/2-\theta_r} \left[1 - \frac{\alpha^2}{x^2 + a^2}\right] \exp(-2(x-u)^2/b^2)\, dx}{A_0 \int_{-\pi/4}^{\pi/4} \exp(-2y^2/b^2)\, dx} \tag{3.1.23}$$

In this case the half-power half-width u_0 of the resonance will be given by:

$$\int_{-\theta_r}^{\pi/2-\theta_r} \frac{\exp(-2(x-u_0)^2/b^2)}{x^2 + a^2}\, dx = \frac{1}{2} \int_{-\theta_r}^{\pi/2-\theta_r} \frac{\exp(-2x^2/b^2)}{x^2 + a^2}\, dx \tag{3.1.24}$$

The illumination amplitude is obtained using (3.1.3), and will also be Gaussian:

$$a_i(z) = a_0 \exp(-z^2\beta_0^2/4) \tag{3.1.25}$$

with, again, $\beta_0 = (n_p k_0 \cos\theta)b$. Consequently, the half-power spot radius can be obtained from:

$$\int_0^{z_0} \exp(-z_0^2\beta_0^2/2)\, dz = 1/2 \int_0^{\infty} \exp(-z_0^2\beta_0^2/2)\, dz \tag{3.1.26}$$

This was also solved numerically to give:

$$z_0 = 0.675/\beta_0 \tag{3.1.27}$$

Using the same assumptions as for (3.1.20), this gives:

$$z_0 \cong 0.064/b \quad \mu m/rad \quad = 3.66/b \ \mu m/degree \tag{3.1.28}$$

Equation (3.1.23) cannot be solved explicitly, but a numerical solution was performed. This was found to agree to within about 5% with the approximation given by (3.1.16), which combined with (3.1.27) can be written as:

$$u_0 = (a^2 + (3.66/z_0 \ deg/\mu m)^2)^{\frac{1}{2}} \tag{3.1.29}$$

To quantify the resonance broadening effect, we will examine the half-power spot radii for which the resonance width is double its plane wave value, i.e. $u_0 = 2a$. We can obtain expressions for this value, which we will call ρ, directly from (3.1.21) and (3.1.28):

$$\rho \cong 2.63/a \ (\mu m/deg) \qquad \text{(rectangular incident spectrum)} \tag{3.1.30}$$

$\rho \cong 2.11/a$ (μm/deg) (Gaussian incident spectrum) (3.1.31)

These values give a rough indication of the lateral resolution that can be obtained in an SPR measurement without too much loss of sensitivity. A table of examples is given below.

a (degrees)	ρ (μm) rectangular	Gaussian
0.1	26	21
0.2	13	10
0.5	5	4

Table 3.1 Half-power illuminated spot radii for which the effective resonance half-width u_0 is twice the plane wave resonance half-width a, for several values of a, and for both rectangular and Gaussian incident angular spectra.

3.2 AC Circuit Analogy

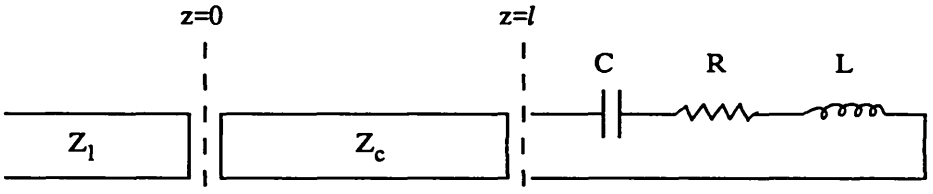


Figure 3.3 A transmission line of impedance Z_l is coupled to an L-R-C oscillator via a length l of coupling line with imaginary impedance Z_c.

In our analysis of surface plasmons, we have assumed the system is uniform in time and in the $\hat{y}$ dimension, as is usual in the literature, and thus the problem is two dimensional. It has proved very useful, in gaining physical insight to this problem, to compare it with the one-dimensional, time-dependent analogy of an electrical transmission line coupled to a damped oscillator, as depicted in figure 3.3. We can show that the equations governing the behaviour of such a circuit are very similar in form to those derived in chapter 2, and consequently the transient behaviour of the circuit can provide inspiration for treating the spatially non-uniform surface plasmon problem.

To obtain evanescent coupling, we use a length l of transmission line having an imaginary impedance Z_c, and an imaginary wavevector k_c. The line in which

the signal is incident has a real impedance Z_l, and a real wavevector k_l. Both the line and the coupling line will support forward- and backward-going waves; if we leave out the time factor $\exp(-j\omega t)$, we can write the voltages as:

$$V_l = A \exp(jk_l z) + B \exp(-jk_l z) \tag{3.2.1}$$

$$V_c = C \exp(jk_c z) + D \exp(-jk_c z) \tag{3.2.2}$$

and the currents as:

$$Z_l I_l = A \exp(jk_l z) - B \exp(-jk_l z) \tag{3.2.3}$$

$$Z_c I_c = C \exp(jk_c z) - D \exp(-jk_c z) \tag{3.2.4}$$

Continuity of I and V at z=0 gives the conditions:

$$A + B = C + D \tag{3.2.5}$$

$$Z_c(A - B) = Z_l(C - D) \tag{3.2.6}$$

The boundary condition at $Z=l$ is:

$$V_c(l)/I_c(l) = R - j\omega L + j/\omega C \tag{3.2.7}$$

We now define two new quantities:

$$\omega_0 \equiv 1/(LC)^{\frac{1}{2}} \tag{3.2.8}$$

$$q \equiv \exp(jk_c l) \tag{3.2.9}$$

Then we can write (3.2.7) as:

$$Z_c \frac{Cq^2 + D}{Cq^2 - D} = - (jL/\omega) [\omega^2 - \omega_0{}^2 + j\omega R/L] \tag{3.2.10}$$

We are interested initially in two aspects of the circuit response: the resonant frequency and damping rate for the undriven case, and the frequency response of the reflectivity for the driven case. Let us first consider the undriven case, where A=0. Equations (3.2.5) and (3.2.6) now give:

$$B = C + D \tag{3.2.11}$$

$$B = Z_l/Z_c \, (D - C) \tag{3.2.12}$$

Which we can combine to give:

$$\frac{C}{D} = \frac{Z_l - Z_c}{Z_l + Z_c} = - r_{lc} \tag{3.2.13}$$

where r_{lc} is the reflection coefficient at the Z=0 junction for an infinite coupling length l. We have stated that Z_c is imaginary; we will call its magnitude ρ, so that:

$$Z_c \equiv j\rho \tag{3.2.14}$$

where ρ is real. We will also introduce a new variable ν, where:

$$\nu \equiv \omega - \omega_0 \tag{3.2.15}$$

and make our first approximation: that we are concerned with a small range of frequencies, such that:

$$\omega^2 - \omega_0^2 \cong 2\omega_0\nu \tag{3.2.16}$$

We can now write (3.2.10) as:

$$\frac{\rho}{2L}\left[\frac{1 - r_{lc}q^2}{1 + r_{lc}q^2}\right] = \nu + jR/2L \tag{3.2.17}$$

For weak coupling, $q^2 \ll 1$, so we can make the approximation:

$$\nu \cong (\rho/2L)\,[1 - 2r_{lc}q^2] - jR/2L \tag{3.2.18}$$

Equation (3.2.18) gives the devation of the oscillation frequency from that of an undamped, uncoupled oscillator; using the notation of chapter 2, we will have three terms to this deviation: a decay constant Γ_i due to internal losses, a decay constant Γ_r due to coupling losses, and a real frequency shift δ due to coupling. These three terms will be given by:

$$\Gamma_i = R/2L \tag{3.2.19}$$

$$\Gamma_r = (\rho/2L)q^2\mathrm{Im}\{r_{lc}\} \tag{3.2.20}$$

$$\delta = (\rho/2L)\,[1 - 2q^2\mathrm{Re}\{r_{lc}\}] \tag{3.2.21}$$

where:

$$\omega = \omega_r - j(\Gamma_i + \Gamma_r) \tag{3.2.22}$$

$$\omega_r \equiv \omega_0 + \delta \tag{3.2.23}$$

The next step is to determine the reflection coefficient as a function of driving frequency, which will be defined as:

$$r(\omega) \equiv B(\omega)/A(\omega) \tag{3.2.24}$$

From (3.2.5) and (3.2.6) we can obtain:

$$2C = A(1 + Z_c/Z_l) + B(1 - Z_c/Z_l) \tag{3.2.25}$$

$$2D = A(1 - Z_c/Z_l) + B(1 + Z_c/Z_l) \tag{3.2.26}$$

Using $B/A \equiv r$, we can combine (3.2.24) and (3.2.26) to get:

$$\frac{C}{D} = \frac{1 - r_{lc}r}{r - r_{lc}} \tag{3.2.27}$$

By inserting (3.2.27) into the final boundary condition (3.2.10), we can obtain:

$$\frac{\rho}{2L}\left[\frac{r - r_{lc} + q^2(1 - r_{lc}r)}{r - r_{lc} - q^2(1 - r_{lc}r)}\right] = \nu + jR/2L \tag{3.2.28}$$

Because r_{lc} will have unit amplitude, $r_{lc}^* = r_{lc}^{-1}$. Using this result, we can obtain from (3.2.28) an expression for the reflectivity; after some quite involved algebra, which we shall omit, we can obtain:

$$\frac{r}{r_{lc}} \cong \frac{(\omega - \omega_r) + j(\Gamma_i - \Gamma_r)}{(\omega - \omega_r) + j(\Gamma_i + \Gamma_r)} \tag{3.2.29}$$

This has the same form as the surface plasmon reflectivity function given by equation (2.6.22). The power reflectivity will also have the analogous form:

$$R = 1 - \frac{4\,\Gamma_i\,\Gamma_r}{(\omega - \omega_r)^2 + (\Gamma_i + \Gamma_r)^2} \tag{3.2.30}$$

The condition for zero reflection is the same as in the plasmon case: $\Gamma_i = \Gamma_r$.

It is also useful to derive, as we did in chapter 2, an expression for the resonant amplitude as a function of driving frequency. This we will call t, so:

$$t \equiv \frac{V(z=l)}{A} = Dq^{-1}/A \tag{3.2.31}$$

The derivation of t is similar to that described above for the reflectivity, and the result is:

$$t = -r_{lc}\left[\frac{Z_l + Z_c}{q\,Z_l}\right] \frac{j\,\Gamma_r}{(\omega - \omega_r) + j(\Gamma_i + \Gamma_r)} \tag{3.2.32}$$

This, again, is has the same form as the surface plasmon equivalent, equation (2.6.33).

3.3 A Diffraction Theory for Surface Plasmon Microscopy

This section will present a theoretical model for the analysis of the interaction of an incident beam of finite extent with surface plasmons on a non-uniform surface, in the Kretschmann prism-coupled geometry. The basis of this analysis is that we assume that both excitation and re-radiation can be considered perturbations on the guided surface plasmon mode. By splitting the reflected wave into two components, one due to direct reflection at the prism/metal interface and the other due to re-radiation from the plasmon field, we will be able to derive expressions for the interaction of arbitrary input beams with surfaces having arbitrary lateral variation, within limitations that will be discussed later in this chapter.

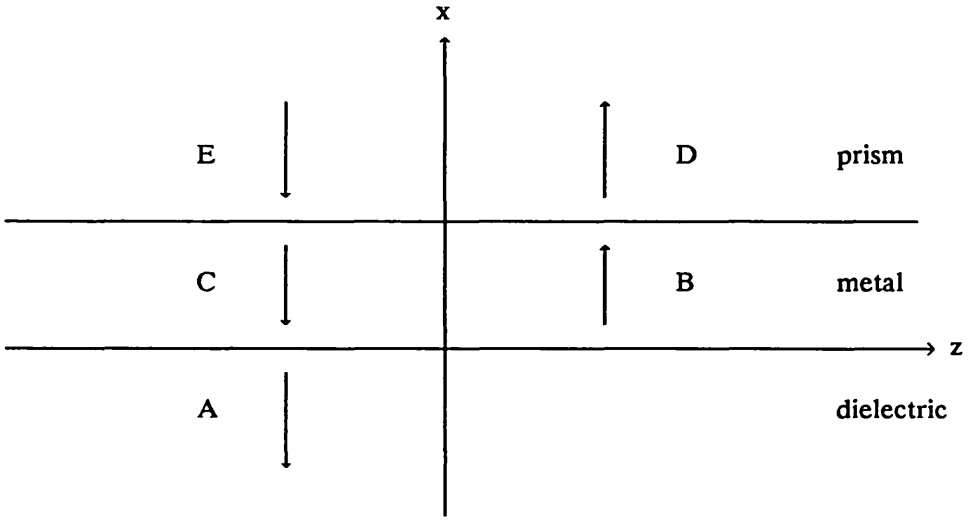


Figure 3.4 Kretschmann prism coupler geometry, indicating the labelling of the various field components.

In the notation of (2.6.26), the amplitude coefficients of the incident, reflected and plasmon fields are E, D and A respectively. This is illustrated in figure 3.4 above. We can rewrite the plane wave reflectivity (2.6.22) as:

$$\frac{D}{Er_{pm}} = 1 + \frac{2j\Gamma_r}{(\beta - \beta_r) - j(\Gamma_i + \Gamma_r)} \qquad (3.3.1)$$

Here we have relabelled r_{32}, the reflectivity between prism and metal half-spaces, as r_{pm}. We will consider re-radiation to be a distinct and separable contribution to the reflected signal. An expression for the re-radiated amplitude can be obtained using the analysis in section 2.5, where the incident amplitude is zero. We will call the amplitude of this re-radiated signal D^+. Using equations (2.5.9) to (2.5.12) we

can derive an expression for D^+ as a function of the plasmon field amplitude A:

$$\frac{D^+}{A} = \frac{J-1}{q(K-1)} \tag{3.3.2}$$

We have derived in section 2.6 expressions for the plasmon field amplitude in a uniform system; we can rewrite (2.6.31) as:

$$\frac{A}{E} = \frac{K-1}{2qJ} \left[\frac{r}{r_{pm}} - 1 \right] \tag{3.3.3}$$

Multiplying (3.3.2) and (3.3.3) gives an expression for the re-radiation amplitude of the uniformly excited plasmon field:

$$\frac{D^+}{E} = \frac{J-1}{2q^2J} \left[\frac{r}{r_{pm}} - 1 \right] \tag{3.3.4}$$

Assuming we are near resonance, equation (2.5.15) for q^2 will be approximately correct, so we can combine it with (3.3.4), giving:

$$\frac{D^+}{E} = \frac{J+1}{2J} (r - r_{Pm}) \tag{3.3.5}$$

This can be further simplified, since near resonance, $J \cong 1$, giving:

$$\frac{D^+}{E} = r - r_{Pm} \tag{3.3.6}$$

Finally, using r=D/E, we can write:

$$D = Er_{pm} + D^+ \tag{3.3.7}$$

In this approximation we can consider the reflected signal to be composed of two separate components: the first is the signal that would be reflected at the interface between the prism and a metal half-space, and the second is the re-radiated signal from the plasmon field. It is when these two components are equal and have opposite phase that the reflectivity is zero.

We can generalize the approximation of (3.3.7) to the case of arbitrary incident beams by integrating over the incident angular spectrum of plane waves. We will no longer be able to quantify the reflected signal by a single amplitude value; the amplitude will have to be written as a function of one parameter. We therefore introduce a function D(z) which describes the amplitude envelope of the reflected beam at the prism/metal boundary, where the spatial 'carrier frequency' is given by the center β of the incident angular spectrum (labelled β_c in section 3.1); the amplitude of the reflected beam at the prism/metal interface x=d is thus given by $D(z)\exp(j\beta_c z)$. In the limit where the incident field is a single plane wave, D(z)

will have the constant value given by (3.3.1). The amplitude of the reflected beam at any other point in the prism can be obtained from this function by performing the appropriate diffraction integral. D(z) can be considered as the amplitude distribution of a one-dimensional source, with the $\exp(j\beta_c z)$ factor having the effect of shifting the direction of output of this source. Similarly, we will introduce two other functions, E(z) and $D^+(z)$, which give the amplitude 'envelope' of the incident beam, and the re-radiation component of the reflected beam, respectively.

We will write the incident angular spectrum as a function $\tilde{E}(\beta)$ of the lateral wavevector β. The reflected amplitude D(z) is given by integrating (3.3.1) over the angular spectrum:

$$\frac{D(z)}{r_{pm}}\exp(j\beta_c) = \int_{-\infty}^{\infty}\tilde{E}(\beta)\exp(j\beta z)d\beta + \int_{-\infty}^{\infty}\frac{\tilde{E}(\beta)j2\Gamma_r\exp(j\beta z)}{(\beta-\beta_r) - j\Gamma_t}\,d\beta \qquad (3.3.8)$$

where $\Gamma_t = \Gamma_i + \Gamma_r$. Both of the integrated terms in (3.3.8) have the form of Fourier transforms. The first, the transform of the incident β spectrum, gives the incident amplitude at the prism base, $E(z)\exp(j\beta_c z)$. The second is a transform of the product of two functions, and therefore can be written as the convolution of the transforms of each. The first function is again $E(\beta)$, and the second has a transform given by:

$$\int_{-\infty}^{\infty}\frac{j2\Gamma_r\exp(j\beta z)}{(\beta-\beta_r) - j\Gamma_t}\,d\beta = \begin{cases} 2\Gamma_r\exp(j\beta_r z - \Gamma_t z) & z\geq 0 \\ 0 & z<0 \end{cases} \qquad (3.3.9)$$

If we rewrite (3.3.8) in the form of (3.3.7), we obtain:

$$D(z) = E(z)r_{pm} + D^+(z) \qquad (3.3.10)$$

then $D^+(z)/r_{pm}$ is the convolution of the incident amplitude E(z) with the decaying exponential of the plasmon mode; we can write this as:

$$D^+(z)/r_{Pm} = 2\Gamma_r\exp(-j\beta_c z)\int_{-\infty}^{z}\exp[(j\beta_r-\Gamma_t)(z-\sigma)]\,E(\sigma)\exp(j\beta_c\sigma)d\sigma \qquad (3.3.11)$$

We have derived equation (3.3.11) from an approximate treatment of the interaction of an arbitrary incident beam with a uniform plasmon-supporting structure. The limits of integration indicate that in this approximation, the reflected amplitude at some point z_1 is due only to the incident amplitude in the region

$z<z_1$; this implies that there is no plasmon field propagating in the $-\hat{z}$ direction, which is reasonable if the incident angular spectrum has only positive β values. Equation (3.3.11) will prove more useful in differential than integral form; this can be obtained as follows:

$$D^+(z)/r_{Pm} = 2\Gamma_r \exp[-(\Gamma_t + j(\beta_c - \beta_r)z] \int_{-\infty}^{z} \exp[-(j\beta_r - \Gamma_t)\sigma] E(\sigma) \exp(j\beta_c \sigma) d\sigma \qquad (3.3.12)$$

$$\partial[D^+(z)/r_{Pm}]/\partial z = [\Gamma_t + j(\beta_c - \beta_r)]\ D^+(z)$$
$$+ 2\Gamma_r \exp[-(\Gamma_t + j(\beta_c - \beta_r)z] \left[\exp[-(j\beta_r - \Gamma_t)\sigma] E(\sigma) \exp(j\beta_c \sigma) \right]_{\sigma=z} \qquad (3.3.13)$$

$$dD^+(z)/dz = [\Gamma_t + j(\beta_c - \beta_r)]D^+(z) + 2\Gamma_r E(z) r_{pm} \qquad (3.3.14)$$

This gives the local rate of change of the re-radiation amplitude, which is proportional to the rate of change of the plasmon field amplitude; the three contributing components are damping and re-radiation loss, phase mis-match, and excitation by the incident beam.

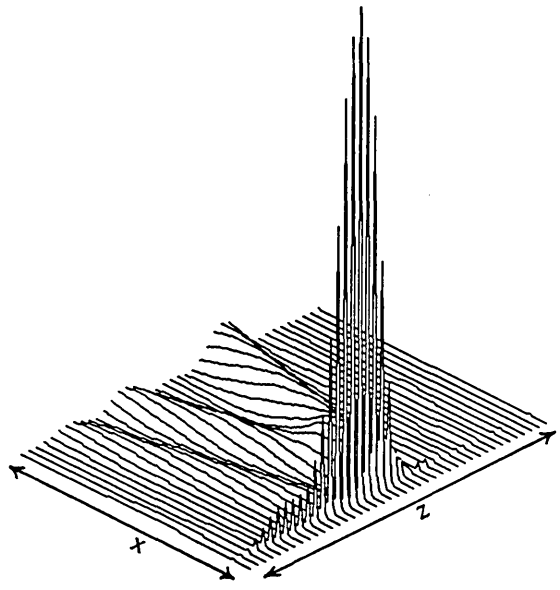


Figure 3.5 Field intensity profile, in the plane of incidence, for the resonant excitation of surface plasmons on a silver film by a focussed Gaussian beam of 2° angular width. Plotted area is 55 μm ($\hat{x}$) by 260 μm ($\hat{z}$).

Equation (3.3.14) allows the calculation of the plasmon field created by a focussed incident beam. This can also be obtained numerically, by calculating the Fresnel transmission coefficients for plane waves incident on a multi-layered

structure, and integrating over the angular spectrum. Figure 3.5 gives an example of a plasmon field profile calculated in this way, showing the intensity rapidly increasing in the region of excitation and then decaying in the direction of forward propagation with the characteristic decay length $1/\Gamma_t$. This figure helps to illustrate why the resolution of the plasmon measurement cannot be increased simply by strongly focussing the incident light. The contribution the plasmon resonance makes to the reflected signal arises from the re-radiation of the plasmon field as it propagates forward, and this clearly can be over a region extending significantly beyond the illuminated portion of the object.

Other than the assumptions and approximations made in obtaining the reflection equations of section 2.6, the treatment presented above is rigorous for arbitrary incident fields. The basis of the new model for non-uniform fields is to postulate that equations (3.3.10) and (3.3.14) are also valid if the surface parameters β_r, Γ_i and Γ_r are themselves functions of z, so long as they are slowly varying compared to the spatial 'carrier' frequency β_c, so that β_r has a definable local value. The basis of this postulate is the assumption that changes in the surface do not directly influence the plasmon field amplitude, but only the rate of excitation and re-radiation, both of which are components of (3.3.14). Thus we can use (3.3.10) and (3.3.14) to analyze the interaction of arbitrary incident beams with arbitrarily varying surfaces.

There are two important limitations to this generalization. The first is that as the model assumes a single forward propagating wave, reflections of the plasmon field are not included. In practice, first reflections can be neglected, since they will not be phase-matched with the forward-going wave and will re-radiate towards the incident beam in the prism, rather than in the direction of the reflected beam. Secondary, and higher even-order reflections, will alter the plasmon and reflected fields; however, such waves will be of very low amplitude, except for those due to highly reflecting multiple line features which are closer together than the plasmon decay length.

The second limitation to equation (3.3.14) is that it does not include the loss of energy from the plasmon field due to scattering into bulk radiation modes in the dielectric medium, which will occur for features having spatial frequency components greater than $\beta_r - n_d k_0$. We can add this contribution in a straightforward way in the case of line features (section 3.4), but not so easily for periodic structures (section 3.5).

Neglecting multiple reflections and scattering losses for the moment, equations (3.3.14) and (3.3.10) taken together comprise a general formulation for treating the interaction of arbitrary incident beams with plasmons on surfaces having lateral variations. This will allow the calculation of the expected reflection images, in surface plasmon microscopy, from arbitrary objects, and prediction of the spatial resolution attainable in various circumstances. It should also be valid for prism coupled imaging using other types of guided modes, although for multimode structures, mode coupling would have to be included in the model. In the case of a non-lossy guide (Γ_i=0), these equations are equivalent to those given by Ulrich for a tapered coupler [3.2],[3.3], although his formulation was developed for a quite different application: that of maximizing the coupling efficiency from a truncated prism into a lossless slab waveguide.

This model can be in some ways easier understood with reference to the AC circuit analogy of section 3.2. The equations derived for reflectivity of the circuit are equivalent to the SPR equations of chapter 2, with temporal frequency ω taking the place of the lateral spatial frequency β. We can consider the dimension of time in the AC model to be analogous to the lateral dimension $\hat{z}$ in the two-dimensional plasmon model. We can therefore compare an incident optical beam of finite lateral extent to an AC input pulse of finite duration, and lateral variations on the plasmon-supporting surface can be considered analogous to temporal variation of the AC circuit parameters. This comparison will be particularly useful in the treatment of line features.

3.4 Line Features

By setting the derivative in (3.3.14) equal to zero, we obtain the equilibrium value of D^+ which is that derived earlier for a uniform system. If we call this value D_0^+, then:

$$D_0^+ = \frac{-2\Gamma_r E(z) r_{pm}}{\Gamma_t + j(\beta_c - \beta_r)} \tag{3.4.1}$$

We can now rewrite (3.3.11) as:

$$dD^+(z)/dz = -[\Gamma_t + j(\beta_c - B_r)](D^+(z) - D_0^+) \tag{3.4.2}$$

The solution to this equation is given by:

$$D^+(z) = D_0^+ + C\exp[-(\Gamma_t + j(\beta_C - \beta_r))z] \tag{3.4.3}$$

where C is a complex constant, the value of which will be determined by boundary conditions. Let us now consider the system of figure 3.6, where we have two adjacent uniform surface regions, characterized by propagation vectors $\beta_1 + j(\Gamma_{i1} + \Gamma_{r1})$ and $\beta_2 + j(\Gamma_{i2} + \Gamma_{r2})$ respectively. The solution within each of these regions will have the form given by (3.4.2), assuming that the incident illumination A(z) is very slowly varying. In region 1 the incident and plasmon field amplitudes will increase slowly from zero with increasing z, so here the value of C will be zero. In the second medium, the initial amplitude of D^+ will be D_0^+ in the first medium, times the transmission coefficient for the plasmon field as it crosses the line boundary. This takes into account the energy lost at the boundary due to scattering into bulk modes in the dielectric. The value of C for medium two can be calculated from this initial value for D^+.

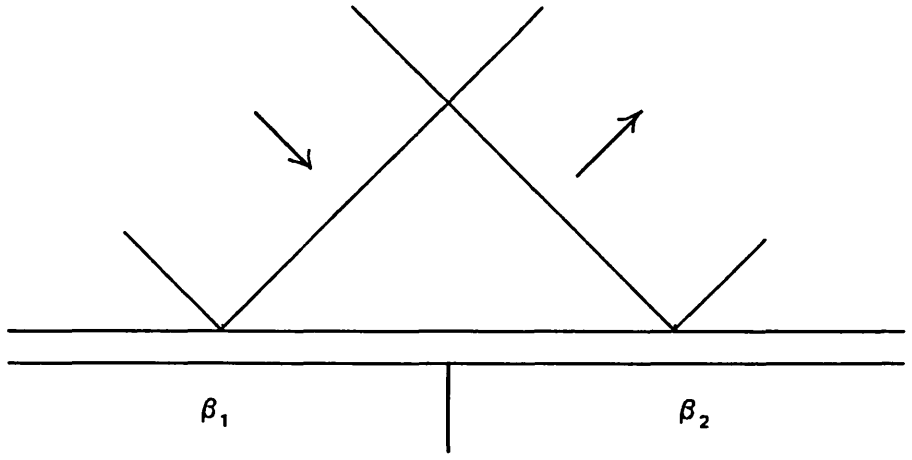


Figure 3.6 A beam of finite width, simultaneously incident on two uniform regions with different SPR properties.

Let us label the transmission coefficient for plasmons crossing the boundary as t_{sp}. This will in general be a complex quantity giving the amplitude and phase change of the plasmon field, and also of D^+, which is proportional to the plasmon amplitude A. If we label D_0^+ in the first and second media as D_{01}^+ and D_{02}^+ respectively, then the complete solution for the re-radiated field after the boundary will be:

$$D^+(z>0) = D_{02}^+ + (t_{sp}D_{01}^+ - D_{02}^+)\exp[-(\Gamma_t + j(\beta_C-\beta_r))z] \qquad (3.4.4)$$

All that remains is to calculate t_{sp}. This we can do by making the approximation that the transmission coefficient is the same in the prism coupled case as it would be for a plasmon travelling on an interface between metal and dielectric half-spaces. This problem has been considered in the literature, both for plasmons at optical wavelengths and for radio-frequency Zenneck waves.

A method for calculating the behaviour of a surface plasmon at a line or strip feature has been reported by T.A. Leskova [3.4]. She describes a complex integral equation method for calculating the radiated field pattern, and the reflected and transmitted surface wave amplitudes. In the case where the strip is wide enough that the radiation field at the first boundary does re-couple into the surface mode at the second boundary, an explicit equation is given for the strip transmission and reflection coefficients. We are only interested in the transmission coefficient at a single boundary, which Leskova gives as (variable labels adjusted to my conventions):

$$t_{sp} = \frac{\gamma_2(\gamma_2-\gamma_1)}{\beta_2(\beta_2-\beta_1)} \frac{\Psi_+(\beta_2)}{\Psi_+(\beta_1)} \qquad (3.4.5)$$

To evaluate the function Ψ_+, Leskova gives a related function:

$$\Psi(\beta) = 1 - \frac{\gamma_2 - \gamma_1}{(k_0^2 - \beta^2)^{\frac{1}{2}} - \gamma_1} \qquad (3.4.6)$$

which then has to be factored into two functions, Ψ_+ and Ψ_-, which are analytic and non-zero in the upper and lower half-planes, respectively, of the complex variable β. This can be done using Cauchy's theorem, as follows [3.5] :

$$\Psi = \Psi_+\Psi_- \qquad (3.4.7)$$

$$ln\Psi = ln\Psi_+ + ln\Psi_- \qquad (3.4.8)$$

$$ln\Psi_+(\beta) = \frac{1}{2\pi j}\int_{-\infty-i\epsilon}^{\infty-i\epsilon}\frac{ln\Psi(z)}{z-\beta}\,dz \qquad (3.4.9)$$

$$ln\Psi_-(\beta) = \frac{-1}{2\pi j}\int_{-\infty-i\epsilon}^{\infty+i\epsilon}\frac{ln\Psi(z)}{z-\beta}\,dz \qquad (3.4.10)$$

Here ϵ is an infinitesimal constant, included to indicate that the two integrations are performed just below and above the real line respectively, and thus indicating the sign that should be chosen for the square root in $\Psi(z)$, in the case where $|z| < |\omega/c|$. In Ψ_+, the imaginary part must be negative, because Leskova gives the condition that the real part is always positive. By evaluating the complex integral and taking the exponential of the result of (3.4.9), we obtain the values needed in (3.4.5).

The problem of surface wave reflection at a single line discontinuity has also been analyzed by Barlow and Brown [3.6], in the context of Zenneck waves. As in Leskova's paper, they consider only the fields in the dielectric medium, the effects of the 'lower' medium being taken account of by the appropriate boundary conditions. Then they evaluate the transmitted, radiated and reflected field components, obtaining explicit solutions using a Wiener–Hopf method. If we combine Barlow & Brown's equations 12.11 and A.3.82, we obtain the magnitude squared of the amplitude transmission coefficient:

$$|t_{sp}|^2 = \frac{4\beta_1\gamma_2{}^2}{\beta_2(\gamma_1+\gamma_2)^2} \qquad (3.4.11)$$

Taking the square root of this expression, we obtain:

$$t_{sp} = (\beta_1/\beta_2)^{1/2}\,\frac{2\gamma_2}{\gamma_1+\gamma_2} \qquad (3.4.12)$$

Figure 3.7 shows a comparison of t_{sp} calculated using (3.4.12) with a numerical solution of the Leskova method, as a function of the ratio β_2/β_1.

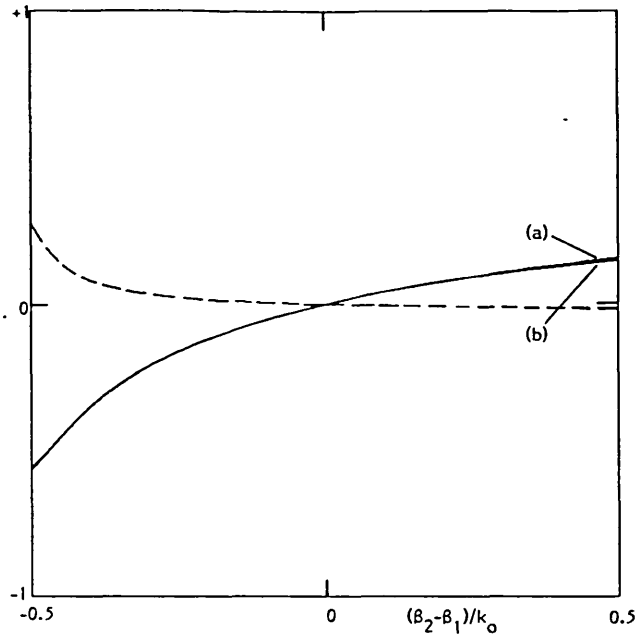


Figure 3.7 Transmission coefficient t_{sp} of a plasmon crossing a line boundary between two surface regions having different values of β_r. The solid lines give $Re\{t_{sp}\}-1$, calculated by the methods of Leskova (a) and Barlow and Brown (b), while the dashed line gives $Im\{t_{sp}\}$, the results of the two calculations being indistinguishable.

Figure 3.8 shows the reflected intensity distributions at the prism base, for a series of angles of incidence of a broad Gaussian beam, for four specific combinations of surface properties, all calculated using the method described above. The most dramatic results are the fringes seen for example in 3.8(a), where the incident beam is wholly absorbed in the first medium (giving a strong plasmon field) but is significantly phase-mismatched in the second. These fringes have been measured experimentally at a number of wavelengths by Rothenhausler [3.7], and they agree in form with the predictions of the model described above. In figure 3.8(a), the plasmon vectors in the two regions are separated by four resonance widths, which gives strong fringes at the boundary. In 3.8(b), this separation is reduced to one resonance width, with a consequent reduction in the amplitude and pitch of the fringes. In 3.8(c), the fringes are damped compared to 3.8(a) because the second medium has a stronger re-radiation coefficient. Finally, 3.8(d) illustrates the loss of resolution of the feature when it is illuminated with more strongly focussed light.

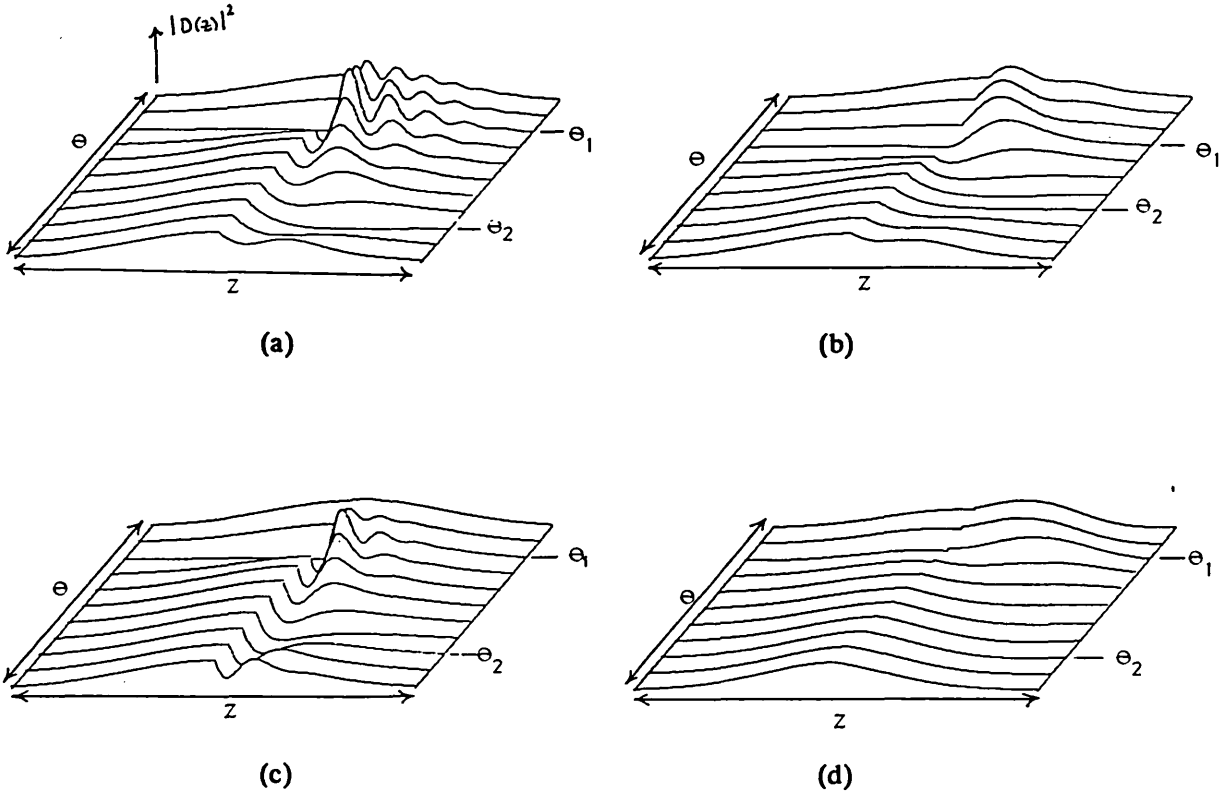


Figure 3.8 Reflected intensity profiles for a Gaussian beam incident on a line feature as illustrated in figure 3.6, plotted for various angles of incidence θ with respect to the resonant coupling angles θ_1 and θ_2 in the left and right media respectively. Parameters for (a) are:

$\beta_{r1} = 1.07k_0$; $\beta_{r2} = 1.03k_0$;

$\Gamma_{i1} = \Gamma_{r1} = \Gamma_{i2} = \Gamma_{r2} = 0.0025k_0$;

incident beam width = $1400/k_0$;

plotted Z range = $1600/k_0$;

Parameters for (b),(c) and (d) are as for (a) except for the following alterations:

(b) $\beta_{r1} = 1.04k_0$

(c) $\Gamma_{r2} = 0.0075k_0$

(d) beam width = $200/k_0$; plotted Z range = $260/k_0$

3.5 Periodic Features

The theory presented in section 3.3 can be used to analyze the case of periodic features, the simplest being sinusoidal variation of the plasmon propagation vector β_r. Let us consider the case of a small sinusoidal perturbation of β_r, of amplitude M and spatial frequency ρ, about the incident center frequency β_c. Then:

$$\beta_r = \beta_c + M\cos(\rho z) \tag{3.5.1}$$

We will define the equilibrium value of D^+ as that in the absence of the perturbation, so that:

$$D_0^+ = -2\Gamma_r A(z) r_{pm}/\Gamma_t \tag{3.5.2}$$

We will also introduce a variable $\Lambda \equiv D^+(z) - D_0^+$. Then the differential equation for D^+ (3.4.2) can be written as:

$$d\Lambda(z)/dz = -\Gamma_t \Lambda(z) - jM\cos\rho z\ D^+(z) \tag{3.5.3}$$

If the perturbation is small, then we can make the approximation that $D^+(z) \cong D_0^+$ in the last term of (3.5.3), which gives:

$$d\Lambda(z)/dz = -\Gamma_t \Lambda(z) - jMD_0^+ \cos\rho z \tag{3.5.4}$$

Taking the derivative of (3.5.4) gives:

$$d^2\Lambda(z)/dz^2 = \Gamma_t^2 \Lambda(z) + jM\Gamma_t D_0^+ \cos\rho z - jM\rho D_0^+ \sin\rho z \tag{3.5.5}$$

We can find a solution to this equation of the form:

$$\Lambda(z) = P\cos(\rho z + \varphi) \tag{3.5.6}$$

This gives:

$$d^2\Lambda(z)/dz^2 = -\rho^2 \Lambda(z) \tag{3.5.7}$$

Inserting (3.5.7) into (3.5.5) gives:

$$(\Gamma_t^2 + \rho^2)P\cos(\rho z + \varphi) = -jD_0^+ M(\Gamma_t \cos\rho z + \rho\cos\rho z) \tag{3.5.8}$$

This gives us solutions for the amplitude and phase of the perturbation on the reflected signal:

$$P = \frac{-jMD_0^+}{(\Gamma_t^2 + \rho^2)^{1/2}} \tag{3.5.9}$$

$$\tan\varphi = -\rho/\Gamma_t \tag{3.5.10}$$

Figure 3.9 shows the variation of the reflected signal with the spatial frequency of

the perturbation, for four values of the perturbation amplitude M, calculated using (3.3.10) and (3.3.11). As predicted by (3.5.9), the amplitude of the sinusoidal variation on the reflected signal drops as ρ increases past Γ_t. As the perturbation amplitude increases, the approximation made in (3.5.4) becomes less valid, and the signal variation becomes less sinusoidal, showing a strong non-linear response. In figure 3.9(c), the low spatial-frequency response curves show a ringing effect similar to that seen for the line features. It should be borne in mind that these calculations do not include the effect of radiation into bulk modes in the dielectric, which become significant for high spatial frequency features.

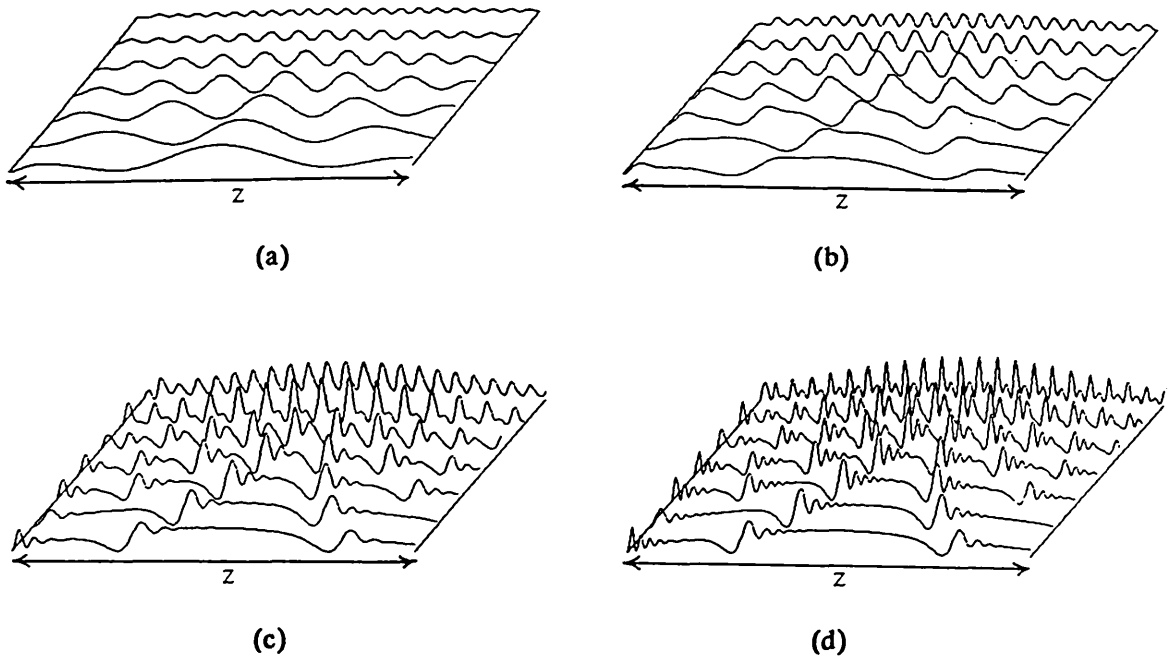


Figure 3.9 Reflected intensity profiles for a broad Gaussian beam incident on a film on which the plasmon coupling angle has a spatially periodic variation about the incident angle. In each figure, this is plotted for the following set of values of the spatial period of perturbation (in terms of $1/\Gamma_t$): 5, 3.3, 2.2, 1.5, 1.0, 0.67, 0.44. The perturbation amplitudes are as follows:

(a) $M=\Gamma_t/2$; (b) $M=2\Gamma_t$; (c) $M=6\Gamma_t$; (d) $M=10\Gamma_t$;

CHAPTER 4

SURFACE PLASMON MICROSCOPY - EXPERIMENTAL RESULTS

4.1 Resonance Curve Measurements

The first experimental task performed was the measurement of surface plasmon resonance curves on uniform films, using previous literature as a guide, in order to get experience of the relevant procedures. The metal chosen for the experiments was silver, which is used in most SPR work due to its high conductivity at optical frequencies (and consequently sharp resonance), its ease of deposition and its relatively high chemical stability. Gold was also used for a few experiments. The silver films were prepared by evaporation in the standard way, using pieces of pure silver wire in a molybdenum evaporation dish through which a high current was passed. The source to sample separation was about 20 cm, the effective source size about 1 cm^2, and the evaporation pressure typically between $2x10^{-7}$ and 10^{-6} torr.. Monitoring of the deposition rate and thickness was done using a quartz oscillator, which was calibrated for each material deposited by measuring film thicknesses on a talystep machine, which has a resolution of about 10 angstroms. A deposition rate of about 10 angstroms/second was found to give good Ag films, while allowing precise control over the thickness. Much slower depositions produced films with poor conductivity, as evidenced by broad plasmon resonance curves. This was probably due to excessive heating of the substrate, which resulted in oxidation or other contamination of the silver as it was being deposited. Samples were given at least 20 minutes to cool after deposition before being exposed to room pressure. The films were deposited on glass microscope slides which were first cleaned with acetone, then microsoap, then de-ionized water, with about three minutes in an ultrasonic bath at each step.

Figure 4.1 shows the apparatus used for the original resonance curve measurements. All measurements were made using a one milliwatt helium-neon laser. The corresponding silver thickness for maximum contrast, as discussed in chapter two, is 560 angstroms. The glass slides on which the films have been deposited are attached to the coupling prism using a drop of index-matched oil. The easiest way to see the resonance is to look for strong scattering at the silver/air surface. If the film were perfectly uniform, there would be no such scattering, but the inevitable surface roughness causes some coupling between the evanescent field and the radiation modes. At the plasmon resonance angle, the scattering intensity increases due to the strong enhancement of the field intensity at the silver surface. This appears as a red glow which is often visible even if the drop in reflectivity is too small to be noticeable to the eye. The scattered light was only used to locate the resonance, not for any quantitative measurements.

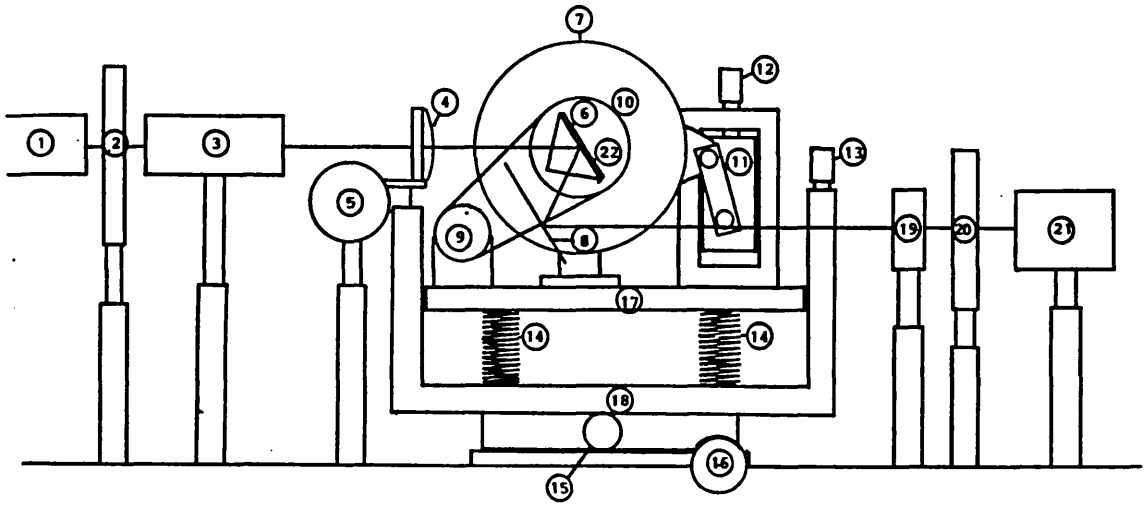


Figure 4.1 Apparatus for plasmon resonance curve measurements, and for scanned plasmon images. Components are:

1. HeNe laser (1 mw)
2. polarizer
3. beam expander
4. lens
5. focus adjustment
6. prism
7. rotating platform
8. mirror
9. servo potentiometer
10. drum for potentiometer belt
11. rotation control linkage
12. rotation control micrometer
13. vertical translation micrometer
14. support springs
15. horizontal translation micrometer
16. servo potentiometer
17. vertically translating platform
18. horizontally translating cradle
19. lens
20. polarizer
21. photodetector
22. slide with metal film

The resonance curves were measured by the following technique (where numbers refer to those of figure 4.1): the beam from the laser (1) is TM polarized (2), expanded (3), and in some instances focussed (4), before entering the coupling prism (6) onto which the sample (22) is attached. The platform (7) on which the prism is mounted is rotated manually via a linear micrometer stage (12), connected to the platform with a pinned rod arrangement (11). A mirror parallel to the prism base (8) restores the beam to its original direction; the beam is then focussed (19) onto a photodetector (21), through a polarizer (20) which adjusts the intensity to suit the sensitivity range of the detector. The output of the photodetector controls one axis of a chart recorder, while the other axis is controlled by the voltage on a servo potentiometer (9), which is connected by a belt to a drum (10) that rotates with the platform (7).

The first curves were measured with an unexpanded, unfocussed beam of about 1 mm diameter. Next, some slight focussing was added to reduce the spot size. In general, this broadened the resonance curve somewhat, as can be seen in figure 4.2. However, in some cases the resonance was narrower in the focussed case. This could be simply due to non-uniformity of the film, so that the response with a larger illuminated area was averaging a greater variation of SPR response. This effect was usually obtained with poor quality films: it was discovered that silver films left exposed to the air gave SPR readings that degenerated significantly after 2-3 days, due to oxidation or reaction with atmospheric contaminants (probably sulfur). Having produced SPR curves with some focussing (spot size about 100 microns), it was now possible to look for variation of this response with position. Figure 4.3 shows two measurements made at one position on the sample, to indicate repeatability, and a third reading at a different position, with a marked change in response. These experiments indicated that a scanned surface plasmon image would be possible; contrast was present, despite the fact that the films appeared completely uniform in a conventional microscope, although the mechanism for this contrast could not be determined.

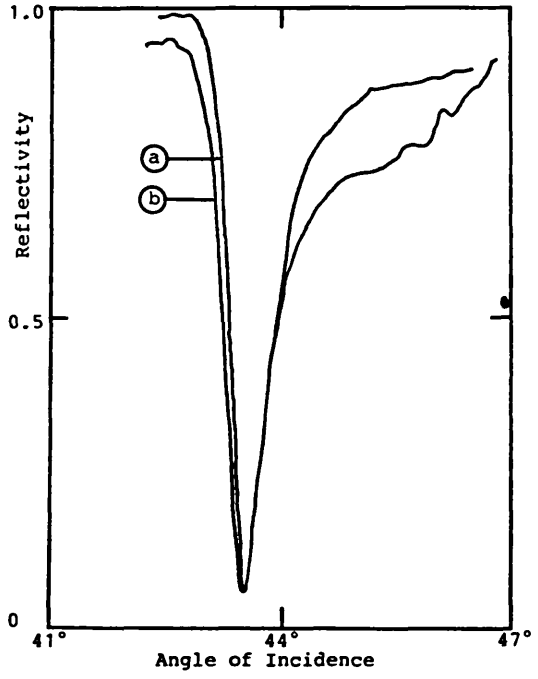


Figure 4.2 Measured SPR curves for a uniform 560 Å Ag film, using:
a) an unfocussed beam
b) a beam of ≈1° angular width

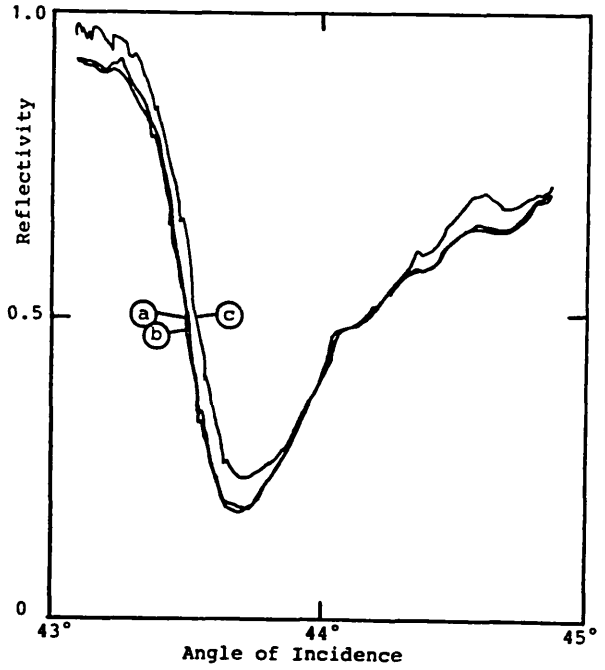


Figure 4.3 Measured SPR curves as in fig. 4.2 b), where a) and b) are measured at the same position, and c) at a second position.

4.2 Scanned Plasmon Images

Figure 4.3 indicates that if the beam is incident at an angle where the slope of the resonance curve is maximum, the reflected intensity will vary with lateral position as the resonance curve changes in position and/or shape. By plotting this varying intensity as a function of position, a single line scan is obtained, and by plotting a series of parallel line scans measured along adjacent paths on the sample, a two-dimensional image of the surface can be constructed. The apparatus used was that of figure 4.1, with the horizontal translation micrometer (15) controlling one axis of the chart recorder via a servo potentiometer (16) as shown. After each line was plotted, the plotter pen and sample were both shifted by fixed increments perpendicularly to their direction of motion in each scan line.

The first surface plasmon images made in this way were of nominally uniform silver films; figure 4.4 is an example. The features obtained correspond to changes in reflected intensity of as much as 40 to 50%. The scanned area is about 600x600 μm, and features of about 50 μm lateral extent can be clearly seen. The mechanism responsible for these features could not be determined; it could be surface corrosion or physical damage. The image did, however, indicate that a surface plasmon scan could expose features not visible using a conventional microscope.

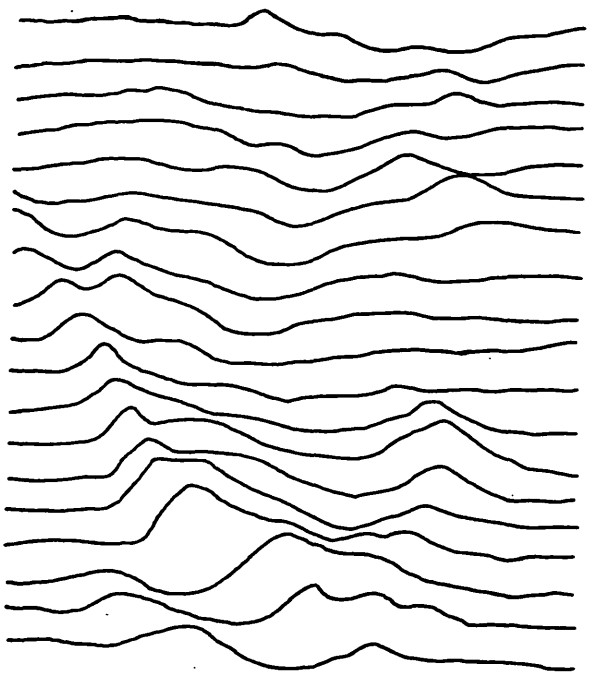


Figure 4.4 Surface plasmon line image of a nominally uniform evaporated Ag film. Scanned area is about 600x600 μm.

4.3 Imaging of Known Features

The first known feature to be imaged was a line boundary between two thicknesses of Ag film. This was made by direct contact masking. The line scan image, figure 4.5, was made using the same method as for figure 4.4. Despite the large step height of 500 angstroms, the line could not be detected with the instrumentation used if the angle of incidence was significantly removed from the plasmon resonance angle, but on the slope of the resonance it was very clearly visible.

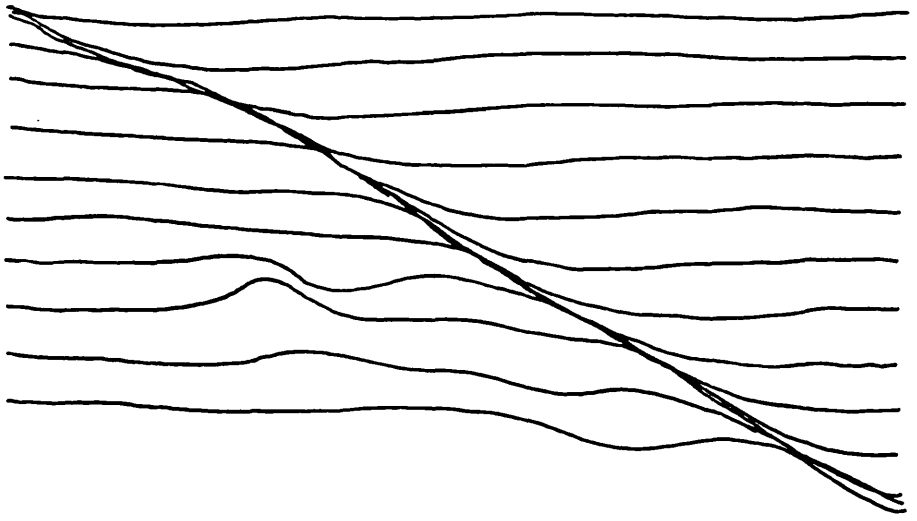


Figure 4.5 Surface plasmon line scan image of Ag film, showing step between film thicknesses of 600 and 1100 angstroms. Scanned area is about 500x500 μm.

The theoretical analysis indicates that the SPR effect is much more sensitive to conditions at the metal surface than to the metal thickness. Therefore, the next test samples were prepared by evaporating very thin oxide layers over Ag films, using a fine wire mesh as a mask. With this method, care had to be taken to ensure close and uniform contact between the mesh and the sample; otherwise, the finite size of the source would cause blurring of the deposited pattern, due to the umbra/penumbra effect (partial shadowing). Since the lack of flatness of the mesh sets a minimum average separation between it and the sample, the edges of the deposited pattern would have a minimum transition width of about one wire radius (about 5 microns). This is consequently also a limit on the spatial resolution that can be measured with these samples.

The oxide layers were impossible to detect with the unaided eye, but were easily detected using SPR. Figure 4.6 shows single line scans across two samples.

Sensitivity is clearly better than one angstrom, and spatial resolution is at least 20 microns. The lack of uniformity and poor edge definition could both result from the manufacturing process; they are not necessarily indicative of limitations of the imaging technique. Layers of 50 angstroms and less were deposited, and such thin depositions could not be monitored precisely using the crystal monitor in the vacuum chamber; measuring the shift in SPR response was in fact the most precise way of determining the thickness. In all cases the resonance shifted substantially compared to its width, so that the optical signal strength was a large fraction of the incident beam intensity. The thinnest sample had a feature height of about two atomic radii, and while clearly the limit of sensitivity of the imaging technique had not been reached, it was not possible to produce substantially thinner features with the method used, since they would not have been detected by the crystal monitor.

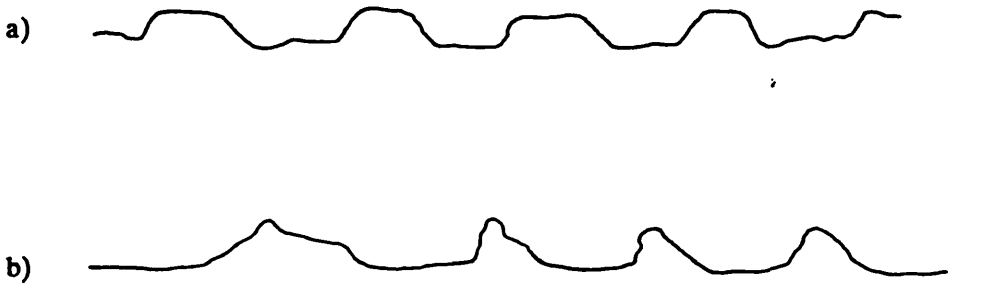


Figure 4.6 Surface plasmon line scans of WO_3 grids on uniform silver films, with grid parameters:
a) 120 micron periodicity, 30±3 angstrom thickness
b) 60 micron periodicity, 7±3 angstrom thickness

4.4 Full-frame Surface Plasmon Microscopy

When the incident beam used for the scanned measurements was not properly focussed on the object, features could be seen (in particular the oxide grids) in both the reflected and the scattered light. This indicated that the SPR effect could be used to produce images of very weak features without scanning, by simply illuminating the entire object or a desired portion with a uniform collimated coherent beam, at the appropriate angle of incidence. This makes the instrumentation even simpler, and presents the image directly to the viewer optically as in a conventional microscope. It has the additional advantage of being a real-time technique, so it could be used for observing temporally varying features without any limitations due to scan rate.

Figure 4.8 shows an example of a full-frame plasmon image of a 25 angstrom thick oxide grid. The instrumentation used is shown in figure 4.7; the print was made by photographing the image screen (8). This is a practical demonstration of a method which produces a high contrast image of a dielectric material less than 0.005 wavelengths thick, without any signal processing or electronics of any kind, and with low optical power (1 mw). Unlike interference techniques, it is not particularly sensitive to alignment, requiring only control of the incident angle and beam divergence to about 0.1°. The analysis presented in chapter three indicates that the images produced in this way are easier to interpret and analyze in general than scanned plasmon images.

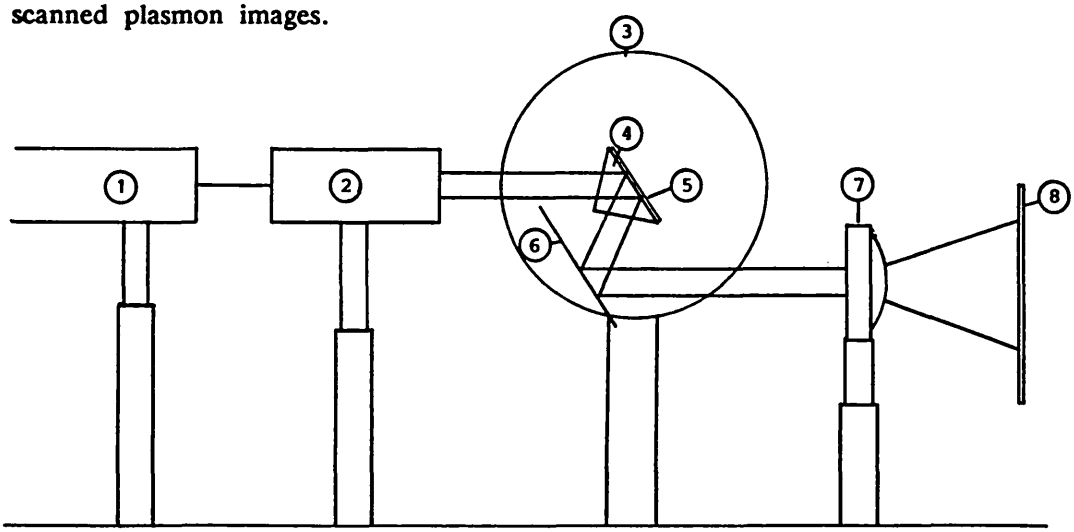


Figure 4.7 Apparatus for full-frame surface plasmon imaging. Components are:

1) 1 mw HeNe polarized laser
2) beam expander
3) rotating platform
4) prism
5) sample slide
6) mirror
7) lens
8) image screen

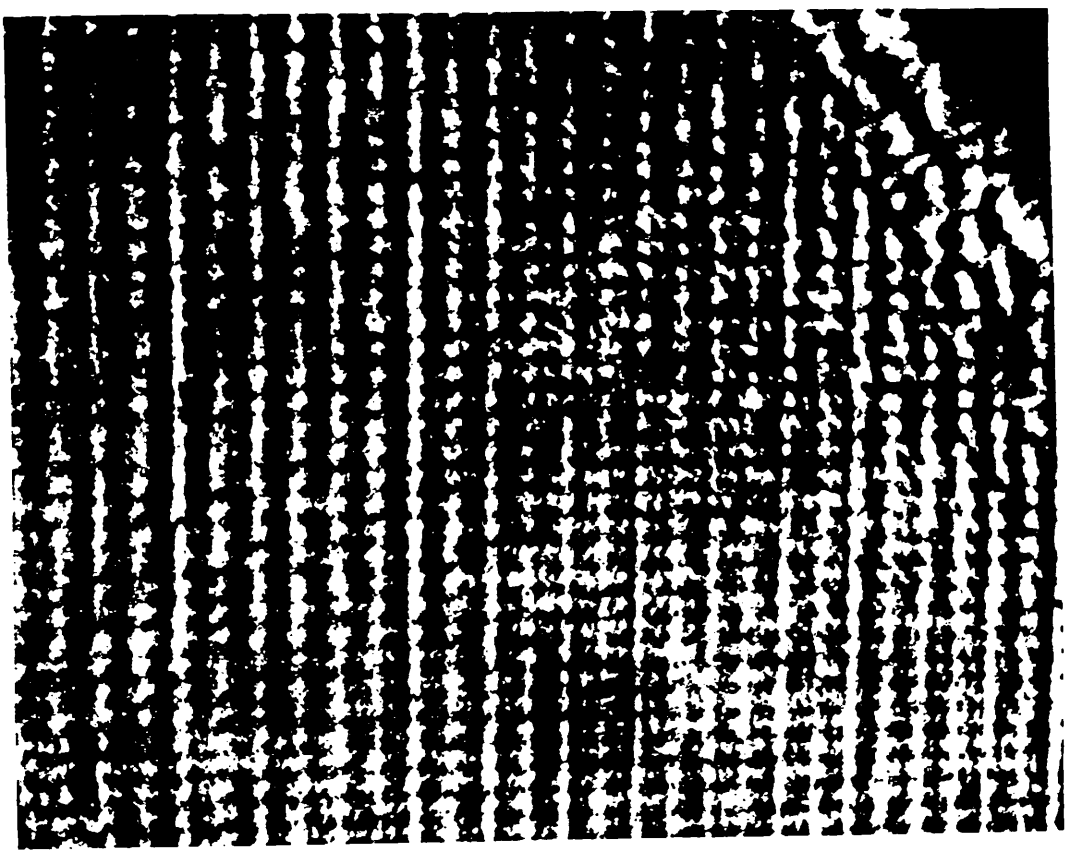


Figure 4.8 Full-frame surface plasmon image of WO_3 grid on a uniform silver film. Grid thickness is 25±3 angstroms, and periodicity is 60 μm.

CHAPTER 5

SPATIAL LIGHT MODULATION

5.1 Introduction

Spatial light modulators, or SLM's, are devices that generate a temporal and spatial variation in the amplitude or phase of a light beam in response to some input signal. The spatial variation is in one or two dimensions perpendicular to the propagation direction of the beam, and may be a continuous image or an array of discrete pixels. The input may be a serial electrical signal, or an optical signal; in the latter case, the purpose of the device is typically to transfer the image from an incoherent to a coherent beam, or between beams of different spectral composition.

Spatial light modulators act as the input devices for optical signal processing systems. These systems take advantage of the inherent parallelism of light, by performing operations simultaneously on a two-dimensional set of data. This is an area of research which is still at an early stage of development. A variety of optical signal processing systems, both analog and digital, have been proposed and investigated; all rely on some form of SLM as their input ports, although for static demonstration a photographic transparency is usually sufficient. While a number of types have been developed, all suffer from significant weaknesses; the lack of adequate SLM's is currently one of the main impediments to the development of optical processing systems of useful complexity and performance.

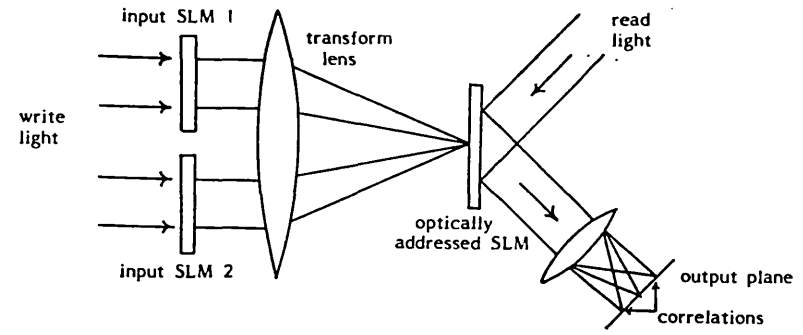


Figure 5.1 A dual inline optical correlator.

One of the most important applications of SLM's is in real-time optical pattern recognition. Figure 5.1 illustrates an example; such systems have application potential in military and security surveillance, robotic assembly, and text recognition. Another major application area is analog and digital optical computing, including optical implementation of neural networks.

The performance and utility of SLM's can be characterized by a number of parameters. These are: speed, lateral resolution, phase uniformity, contrast, grey–scale, effective area, linearity, storage (memory) ability, input sensitivity, efficiency, and spectral response. In addition, the factors of cost, reliability and lifetime will naturally be important. The relative importance of the performance parameters is very much dependent on the specific application. In the case of analog image processing, broadcast video provides a convenient benchmark for speed, contrast and grey scale, as well as the information content (resolution x active area). For optical computing, much higher speeds will be required. In digital applications, grey–scale is unnecessary and bistable operation may be desirable. In any coherent signal processing system, phase uniformity will be a critical parameter, although this may depend more on the manufacturing process than on inherent properties of the device. Input sensitivity is especially important for optically addressed SLM's. Efficiency refers to the fraction of the optical power passed by the device (in the unmodulated parts of the beam); this will become an increasingly important parameter as systems become more complex.

5.2 Current SLM Technology

Most of the important types of SLM that have been reported are described in a 1977 review paper by Casasent [5.1]. Since then several of them have been substantially improved and a number of new variations have been reported, some of which are described in a recent review by Collings [5.2]. However, no important new technology has emerged, with the exception of the greatly increased use of active semiconductor structures as part of the implementation of earlier techniques. The major classes of SLM are described below.

Liquid Crystal SLM's

Devices using liquid crystals (LC's) make up the most important class of SLM's. This includes the Hughes liquid crystal light valve (LCLV), which is depicted in figure 5.2. A twisted nematic cell is employed, as in display devices, to rotate the polarization of the 'read' beam as it passes through (once for transmission and twice for reflection devices). The liquid crystals, which are highly birefringent, are aligned with their high refractive index axes parallel to the cell faces, but gradually rotating through 90° within this plane as they pass between the two faces, which are coated with perpendicularly oriented aligning layers. This alignment is altered by applying an electric field across the cell; the high index

axes of the LC's tend to align in the direction of the field. A polarizer at the front face sets the initial polarization and converts the rotation of the exiting beam to an amplitude modulation. The difference between an SLM and a normal LC display device is that in the SLM the applied field, and thus the resulting modulation, is spatially controlled at a small lateral scale. In the LCLV depicted, this is done by illuminating a CdS photoconducting (PC) layer with an input image, and applying a bias voltage across the LC/mirror/PC sandwich. In parts of the image with higher intensity, the increased conductivity will increase the fraction of the bias voltage across the LC, thus reducing the polarization twist and the resulting amplitude reduction. In this way, the image is transferred from the input to the (usually coherent) write beam. The input may be from a conventional display, or light focussed from a physical object.

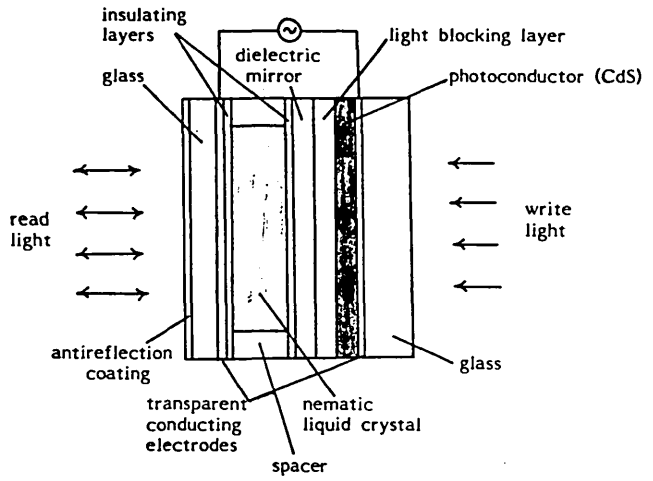


Figure 5.2 The Hughes liquid crystal light valve (LCLV).

The LCLV gives good grey-scale and reasonable contrast, operates at low bias voltages (≈ 5V), and exhibits resolution of about 50 linepairs/mm [5.1]. Its main weakness is speed, which is limited both by the mechanical response of the LC's, and by charge storage time in the photoconductor. Typically these devices have rise and fall times from 10 to 100 milliseconds. A number of elaborations on the basic LCLV have been reported. Hughes have developed an SLM in which the twisted nematic LC cell is switched by an optically activated MOS silicon structure [5.3]. At GEC, another Si/LC device has been developed in which a photodiode array replaces the PC layer [5.4]. Other PC materials have been investigated for use in SLM's, including amorphous silicon [5.5], and GaAs [5.6].

A number of electrically addressed LC SLM's have been reported. Hughes

have produced a variation on the light valve in which a 256 x 256 CCD (charge-coupled device) array provides a matrix of serially addressed analog memory cells which switch corresponding portions of the liquid crystal [5.7]. At STC, devices have been made with electrode array back-planes containing active addressing circuitry [5.2]; in this case, ferroelectric liquid crystals are used, which exhibit much faster switching than the conventional nematic cells. These ferroelectric LC cells are orientationally bistable; therefore they have an inherent memory capability, and are most suited for binary operation, which makes the active addressing circuitry similar to that for semiconductor memory.

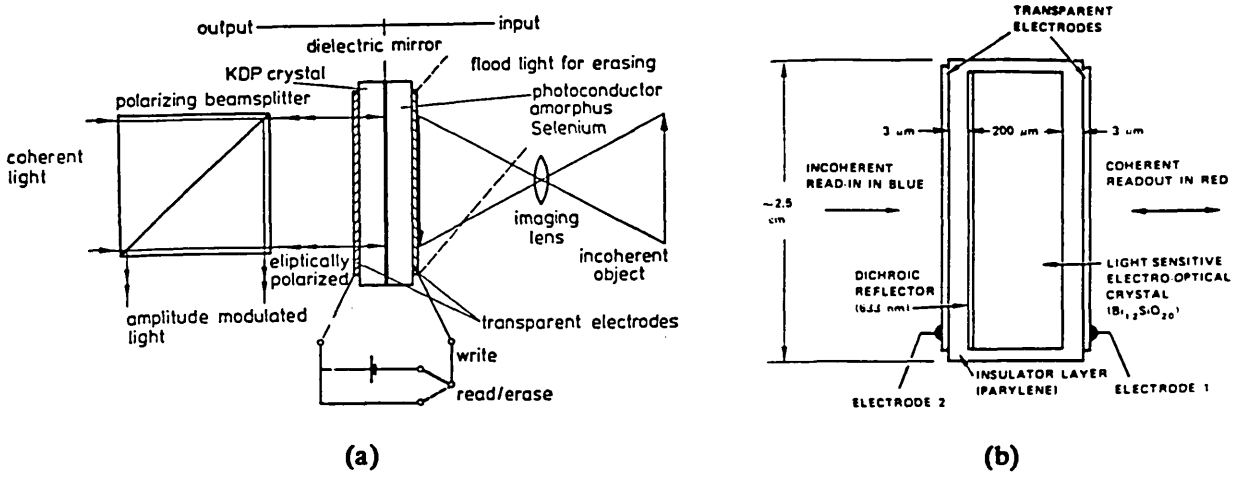
In the last few years, a number of authors have reported the use of LC televisions as spatial light modulators [5.8],[5.9]. The primary motivation is that because they are mass produced, LC TV's constitute a very inexpensive electrically driven LC matrix. The optical quality is very poor, but this can be greatly improved by replacing the low-quality polarizers and immersing the display screen in index-matched oil between optically flat glass plates. They are usually used in transmission mode, and are conveniently addressed by a standard video signal.

Pockel's Effect SLM's

These devices are similar to liquid crystal SLM's, except that the polarization rotation of the 'read' beam is performed by a solid crystal exhibiting the first order electrooptic (Pockel's) effect. In such crystals, when an electric field E is applied in the direction of one of the crystal axes, the refractive indices n_1 and n_2 along the other two axes shift linearly according to:

$$n_{1,2} = n_0 \pm \alpha E \qquad (5.2.1)$$

Light initially polarized at 45° to these two axes will have its polarization rotated as it propagates through the crystal, and this polarization change is converted to an amplitude modulation by a polarizer at the exit plane, as in the LC devices. To obtain maximum modulation the polarization must be rotated through 90°, and as the rotation is proportional to the product of the path length and the field E, it is simply a function of the applied voltage and the Pockel's coefficient α. Therefore we can define the half-wave voltage as that which produces a half wavelength difference in optical path length between the two polarization components, and thus a 90° rotation. One of the disadvantages of these devices is that the half-wave voltages are usually in the kV range.



(a) (b)

Figure 5.3 Two Pockel's effect SLM's, using (a) DKDP and (b) BSO. From [5.1].

A number of Pockel's effect SLM's have been reported [5.10],[5.11],[5.12]. Figure 5.3 illustrates two of these. The first uses a KDP crystal which is switched using an amorphous selenium photoconductor. Write time is low (≈30 μs), but the image must be erased by reversing the bias potential and flooding with blue or UV light [5.1]; therefore this is not really a real-time device. Resolution is 75 linepairs/mm. The device is operated at -50°C to reduce its half-wave voltage to 80V. The device in figure 5.3(b) uses a BSO ($Bi_{12}SiO_{20}$) crystal, which is both electrooptic and photoconducting. In order to reduce the influence of the read beam on the conductivity, its intensity must be low, and its wavelength must lie in a weaker part of the spectral response of the photoconductive effect than that of the write beam. A bias voltage of 2 kV is used. Spatial resolution is about 50 linepairs/mm for 50% modulation, although this varies with write intensity. Very high contrast (5000:1) is possible at lower resolution, and phase flatness better than λ/5 can be achieved.

Microchannel SLM's

These are actually Pockel's effect devices as well, but they use a microchannel plate as an electron amplifier between the photoconductive layer and the electrooptic crystal [5.13],[5.14]. Figure 5.4 shows a typical construction. The read and write beams are isolated, so dual wavelength operation is not needed. Various electrooptic materials have been used in these devices, including DKDP, BSO and lithium niobate ($LiNbO_3$), and various photoconductors can be used as well, providing possible write wavelengths ranging from near IR to soft x-ray regions. Microchannel SLM's are complex to construct, and require high voltage power

supplies, but they provide very high sensitivity to the write beam (2.2 nJ/cm^2). They are also reasonably fast (40 frames/sec at full modulation), but the resolution is poor (less than 10 linepairs/mm at 50% modulation). Research into these devices seems to have decreased, as solid state techniques such as those referred to previously are developed which also provide a high input sensitivity.

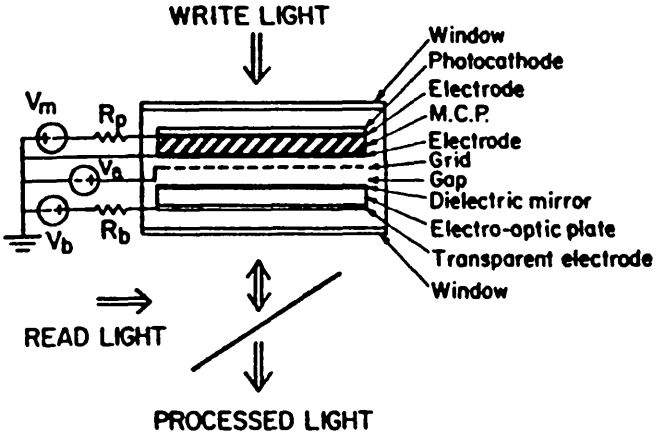


Figure 5.4 Construction of a microchannel SLM (from [5.13])

Deformable SLM's

There are a variety of devices, exhibiting a variety of qualities, which modulate incident light by the local mechanical deformation of a surface. Thermoplastics are one category of these; they are deformed at high temperature by a charge variation produced by shining the image onto an adjacent photoconductive layer. This method can be used to generate a holographic matched filter; resolution is high (2000 linepairs/mm) but writing is slow and complex. SLM's have been developed which use a scanning electron beam in a vacuum tube to write patterns on oil films; these devices have good speed and resolution, and have been extensively used in projection television. Optically addressed deformable oil film SLM's have also been reported [5.15]. Recently, another optically addressed device using a deformable gel has been developed [5.16]. A number of SLM techniques have been investigated which use the local deformation of a membrane mirror to produce phase variation in the reflected beam. One recently reported electrically addressed device has a reflecting membrane deposited over a silicon VLSI structure, in which varying voltages on pixel electrodes deflect the membrane into etched cavities [5.17]. Present manufacturing technology does not allow pixels smaller than about 75 μm.

Magnetic SLM's

Electrically addressed magneto-optic SLM's can be manufactured by epitaxially depositing a magnetic garnet film onto a non-magnetic garnet substrate [5.18]. This film is then etched into an array of elements with associated drive lines, and the magnetic polarity of the elements is controlled by row-addressed current switching. This is intrinsically a binary technique, so it is well suited to digital processing applications [5.19], but not very well to analog applications. The incident light undergoes a polarization rotation due to the Faraday effect, the rotation direction depending on the magnetic polarity. This system has some inherent storage ability, reasonable speed, and elements can be as small as 10 μm.

Other SLM Types

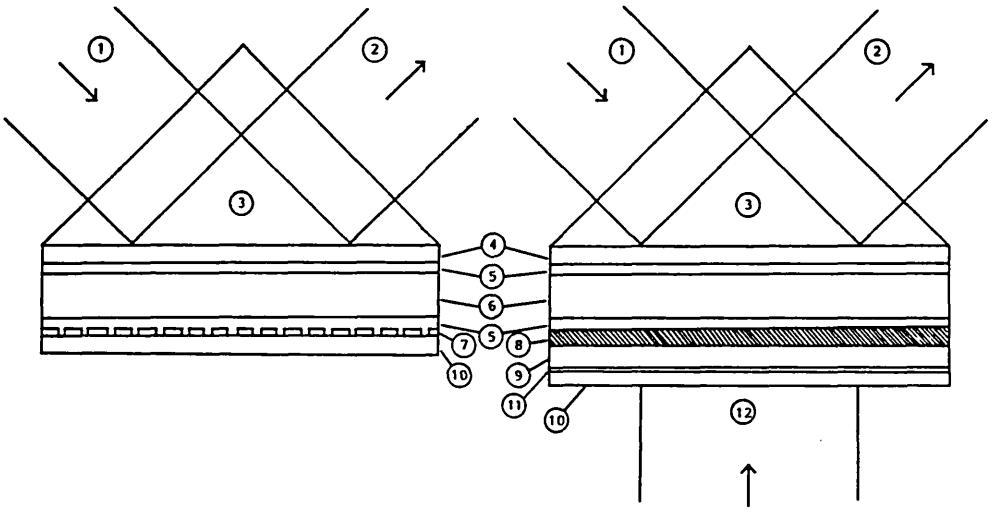
A few types of SLM have been reported which do not fit into the categories described above. One is the photodichroic SLM [5.20]. These devices use colour center defects in alkali-halide host materials. The defects cause spectral absorbtion bands; by illumination at some wavelength, electron transfer can produce variation in the absorbtion at other wavelengths. Because the mechanism is on the lattice scale, the resolution is very high (diffraction limited), but the sensitivity is very low (150 mJ/cm^2). Write and read bands vary depending on the material used; some are written by visible light and read in the infrared, others written in UV and read at visible wavelengths.

In addition to the Pockel's effect devices, there are also those that use the second order electrooptic effect, in which the refractive index varies as the square of the field. A recently reported device uses PLZT as the electrooptic layer deposited on an active silicon substrate [5.21].

5.3 A Surface Plasmon Spatial Light Modulator

Experimental results on surface plasmon microscopy indicated that the amplitude of a collimated coherent beam could be spatially modulated, with high contrast and reasonable resolution, by small variations in the optical properties of a metal/dielectric interface. It was suggested by Richard Syms [5.22] that this might form the basis of an optical image convertor, and it was from this suggestion that the work on surface plasmon SLM's began. The basic principle can be stated in simple and general terms: plasmons are generated on some planar structure by optical excitation, using prism or grating coupling, and the optical properties of the

structure are spatially varied so as to vary the SPR response. The specific configuration that was investigated was that of a Kretschmann geometry with a metallic film bounded on one side by the coupling prism, and on the other by a dielectric layer, the refractive index of which can be locally varied by an applied electric field. The surface plasmon SLM is not just a new type of device to be added to the present list; it introduces a new physical mechanism by which the goal of spatial modulation of a light beam can be accomplished. For this reason, many of the structures described in section 5.2, in particular the liquid crystal and electrooptic devices, could be modified to operate by surface plasmon resonance.



(a) (b)

Figure 5.5 Possible configurations for liquid crystal SLM's operating by surface plasmon resonance; a) electrically adressed, and b) optically adressed. Components are as follows:

1) incident coherent 'read' beam
2) reflected modulated beam
3) coupling prism
4) silver film
5) aligning layers
6) nematic liquid crystal cell
7) electrode matrix
8) light blocking layer
9) photoconducting layer
10) glass slide
11) transparent electrode
12) incident 'write' beam

Figure 5.5 illustrates possible configurations for electrically and optically addressed SPR liquid crystal SLM's. A nematic liquid crystal layer is aligned parallel to a silver film, deposited on high index glass (not shown), by a thin aligning layer. Surface plasmons are generated at this interface using a coupling prism. As most of the plasmon field has its E vector normal to the surface, the effective index of the LC layer is the ordinary index (normal to the long crystal axis). If a bias potential is applied across the cell, the crystals will rotate towards

the normal direction, increasing the index in that direction towards the higher extraordinary index (parallel to the long axis). This will increase the SPR coupling angle, causing modulation of the ouput. In the first configuration this is accomplished by an array of electrodes on the back plate, while in the optically addressed configuration, a bias potential is split between the LC cell and a photoconducting layer, as in the Hughes light valve, so that an image in the 'write' beam is reproduced in the LC orientation pattern, and consequently in the reflected intensity of the 'read' beam.

While spatial light modulation using surface plasmons has not been previously reported, some work has been published on single-cell electrically controlled SPR modulators. Pockel's effect devices have been proposed, in which the SPR response for a prism/Ag/electrooptic crystal structure is modulated by an applied potential across the crystal [5.23],[5.24]. One device has been constructed in which a piezoelectric stack is used to vary the thickness of an air gap between a prism and a silver layer (Otto configuration) [5.25]. This primarily varies the coupling between the incident beam and the plasmon, and consequently the depth of the resonant reflectivity dip. One could conceive of a spatial light modulator as simply being an array of these individual modulators; however, most current or envisaged SLM applications require spatial variation of the modulation on a very small scale, preferably approaching diffraction limits. To effectively investigate the use of plasmons for such applications requires an understanding of the physics involved in the excitation of plasmons on non-uniform planar structures, and a theoretical model by which the response of specific configurations can be predicted. This I believe is now provided by the analysis described in chapter 3. The model developed allows the prediction of the distortion of image features as a function of size, orientation and intensity. In the case of pixellated SLM's, where the device is divided into a matrix of individually addressed elements, the model can be used to investigate the performance of the elements as a function of size, and the cross-talk between them.

There are several potential advantages to using SPR for spatial light modulation. Let us first consider sensitivity. In the liquid crystal and Pockel's effect devices described in section 5.2, a half wavelength difference between the optical path lengths of two polarization components is needed to achieve full modulation. Therefore, if the refractive indices along two axes are altered by $\pm\delta n$, the product of the index change and physical path length L will be given by:

$$(\delta n/n)L = \lambda/4 \tag{5.3.1}$$

We can calculate an equivalent parameter for SPR modulation. We have seen in chapter 2 that the plasmon propagation vector β for a high conductivity metal film is slightly greater than nk_0, where n is the index of a uniform dielectric layer in contact with the metal, and that the half-width of the resonance is given by Γ_t. To modulate the reflected beam by 50% will require shifting the resonance by its half-width, so that:

$$\delta n/n = \Gamma_t/\beta \tag{5.3.2}$$

$$\delta n \cong \Gamma_t/k_0 \tag{5.3.3}$$

Since the field decays exponentially, the effect of index variations on the SPR response will be greater close to the interface. Let us assume that if the index is not uniform in the $\hat{x}$ direction, where x is the distance from the metal surface, β will be determined by an average index $\bar{n}$, weighted according to the field amplitude, i.e.:

$$\bar{n} = \gamma\int_0^\infty n(x)e^{-\gamma x}dx \tag{5.3.4}$$

Let us now uniformly modulate a layer of thickness L adjacent to the metal, such that:

$$n(x) = \begin{array}{ll} n_0+\epsilon & x \leqslant L \\ n_0 & x > L \end{array} \tag{5.3.5}$$

Inserting this distribution in (5.3.4) gives:

$$\bar{n} = n_0 + \epsilon(1-e^{-\gamma L}) \tag{5.3.6}$$

To get 50% modulation we now need:

$$\delta\bar{n} = \epsilon(1-e^{-\gamma L}) = \Gamma_t/k_0 \tag{5.3.7}$$

If the active layer is thin compared to the field decay length, i.e. $L \ll 1/\gamma$, then $e^{-\gamma L} \cong 1-\gamma L$. This gives us an index change - path length product of:

$$\epsilon L = \Gamma_t/\gamma k_0 \tag{5.3.8}$$

For a well matched system, we have seen that $\Gamma_t \cong 2\Gamma_i$, where Γ_i is given by the imaginary part of (2.4.5), so we get:

$$\Gamma_t/k_0 \cong \epsilon_x/\epsilon_r^2 \tag{5.3.9}$$

We can also approximate γ, for $\epsilon_x \ll \epsilon_r$, as:

$$\gamma \cong k_0/\epsilon_r^{\frac{1}{2}} \tag{5.3.10}$$

We can now obtain an expression for ϵL in terms of the material parameters:

$$\epsilon L \cong \frac{\lambda \epsilon_x}{2\pi \epsilon_r^{3/2}} \qquad (5.3.11)$$

For Ag at 633 nm we can insert values of ϵ_x and ϵ_r from table 2.1, and obtain:

$$\epsilon L \cong \lambda/800 \qquad (5.3.12)$$

For the polarization rotation devices, the corresponding figure for 50% modulation will be $\lambda/8$, so this indicates a potential improvement in sensitivity of two orders of magnitude for SPR devices. As discussed in chapter 3, this enhanced sensitivity is due to the resonantly amplified field intensity, and a similar improvement could be obtained using a Fabry–Perot resonant cavity. In the plasmon case, however, the active region can be considerably less than a wavelength thick.

Another advantage of the SPR SLM is that because the 'read' light does not propagate through the active layer, there is much less phase distortion. In the bulk LC and Pockel's devices, all interfaces encountered by the read beam must be flat and parallel compared to the wavelength if coherence is to be maintained. The layers must also be homogeneous so that the optical path lengths have no random variation. In the SPR device, the active layer does not need to be flat; only the two surfaces of the silver need to be phase flat. If the active layer (e.g. the liquid crystal) is not uniform and homogeneous, this will only affect the strength of the signal, not the absolute phase. No liquid crystal SLM's yet reported have low enough phase distortion to be of much use in coherent signal processing systems; with an SPR device, this should be much easier to achieve.

In the specific case of liquid crystal SLM's, the surface plasmon technique offers a considerable speed enhancement. Previous work in this research group has demonstrated that switching times for nematic liquid crystal cells are much lower near the surface regions than in the bulk of the cell [5.26]. This is because the recovery of the cell when the bias is reduced is due to the surface aligning forces, and the realigning force only extends gradually into the bulk as the crystals rotate, with a speed limited by their viscosity. The results described in section 5.4 indicate an order of magnitude increase in switching speed for the SPR devices.

Perhaps the most exciting advantage of SPR devices is that because they are both highly sensitive and ideally suited to the use of very thin active layers, they

may open up the possibility of using new types of materials and effects for modulation that would not be practical with the bulk devices. One possibility is the use of reversible electrochemical or photochemical phenomena, which would tend to be too slow, due to diffusion times, in anything but very thin layers. Another possibility is the use of optically active monolayers deposited by the Langmuir-Blodgett technique.

5.4 Experimental Results

It was decided to use a liquid crystal structure to investigate and demonstrate the use of SPR for spatial light modulation. This was primarily because the facilities and experience relevant to LC fabrication were available in the research group, due to ongoing work on LC display devices. The configuration constructed is depicted in figure 5.6. While LC modulators using SPR have not been previously reported, some work has been done at the University of Exeter where surface plasmon excitation was used to investigate the orientational distribution in liquid crystal cells as a function of applied voltage, from which surface binding forces could be deduced [5.27]. Some of the devices described in this section were constructed and tested by Martin Caldwell.

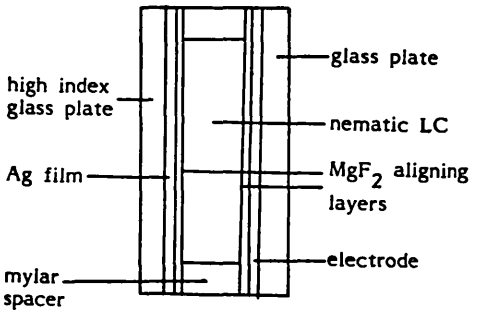


Figure 5.6 Schematic of SLM's constructed.

Nematic liquid crystals of the types K15, E7 and E47 were used, all supplied by BDH Chemicals. These have high birefringence and a relatively low ordinary index (≈1.5). The glass slide on which the silver film is evaporated has to have an index significantly higher than this ordinary index, so that the coupling angle will not be too close to grazing in the unswitched state. Glass of index 1.805 was used for these front plates and for the coupling prism. The completed cell was attached to the prism using a drop of high index coupling fluid.

The aligning layers used were MgF_2 deposited by evaporation at an angle of about 60° to the normal direction. This is a well known technique for parallel alignment [5.28]; essentially, the MgF_2 forms a series of parallel hills, due to shadowing produced by the initial irregularities in the coating thickness. Total thicknesses of about 200 angstroms are used; this appears to provide adequate aligning force and uniformity, without taking up too much of the field energy that would otherwise be in the active region. SPR measurements were used to calibrate the thickness of the MgF_2 layers simply by measuring the resonance curve on coated and uncoated parts of the Ag film, and comparing the angular shift with that predicted by OPTO, using a refractive index for MgF_2 films of 1.38. Figure 5.7 shows an example of these two measured curves; addition of the MgF_2 causes some resonance broadening due to increased surface roughness.

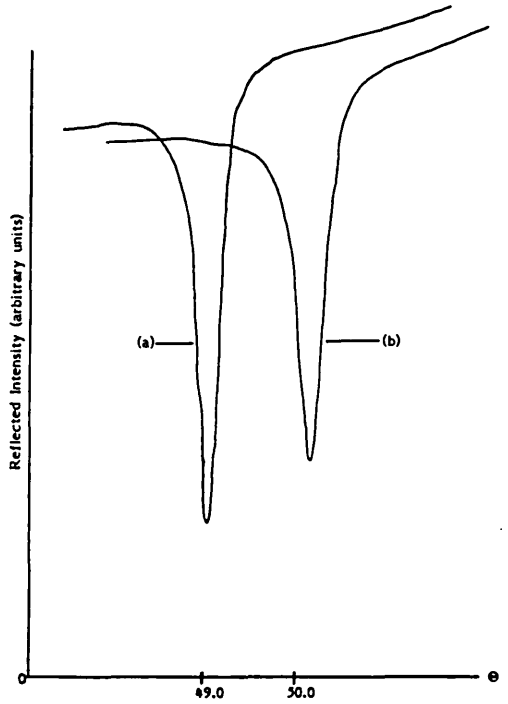


Figure 5.7 SPR resonance curves for portions of a Ag film (a) uncoated and (b) coated with an MgF_2 aligning layer.

Once the glass plates have been cleaned (as in section 4.1) and coated with their metal and aligning layers, they are glued together with epoxy using strips of mylar as spacers. The mylar used was 6 μm or 7.5 μm thick. Gaps are left at opposite ends of the cell through which the liquid crystal can be inserted by capillary action, before the cell is completely sealed. All these steps were carried out in a clean room of approximately class 1000. One problem encountered was

that the high index matching fluid tended to slowly seep through the epoxy, and then enter and corrupt the cell. Less corrosive fluids could not be obtained of sufficiently high index, so this problem was reduced simply by using the minimum possible amount of fluid and taking extra care in handling and applying it, to keep it away from the epoxy bonds.

The first devices made had uniform back electrodes, in order to demonstrate modulation without spatial variation. In general, it was found that the SPR curves for the finished devices were poorer than those of the bare Ag films – they were broader, shallower, and less smooth. This was probably due to lack of uniformity in the LC surface alignment, which caused scattering losses. With more careful control of the deposition and LC filling procedures, this problem was reduced, and devices were manufactured with better than 99% extinction at resonance. Figure 5.8 shows the response of one of these devices, where the angle of incidence is set for minimum reflectance at zero bias voltage; a contrast ratio of about 120:1 was obtained. This device shows a voltage threshold for switching of about 5 V p–p, which corresponds to a bias amplitude of 2.5 V. This is higher than those of typical conventional LC cells (about 1 V); the cause of this difference is unknown. The signal applied was a high frequency square wave (about 10 kHz) rather than DC, since DC currents tend to reduce the lifetime of LC devices. Figure 5.9 shows SPR curves for the same device, measured at several values of bias voltage. This indicates that the main effect is not a shift in the resonance, but a reduction in its depth. This is probably due to the complicated director profile taken up by the LC cell in the applied field, but a detailed analysis of this point has not been carried out.

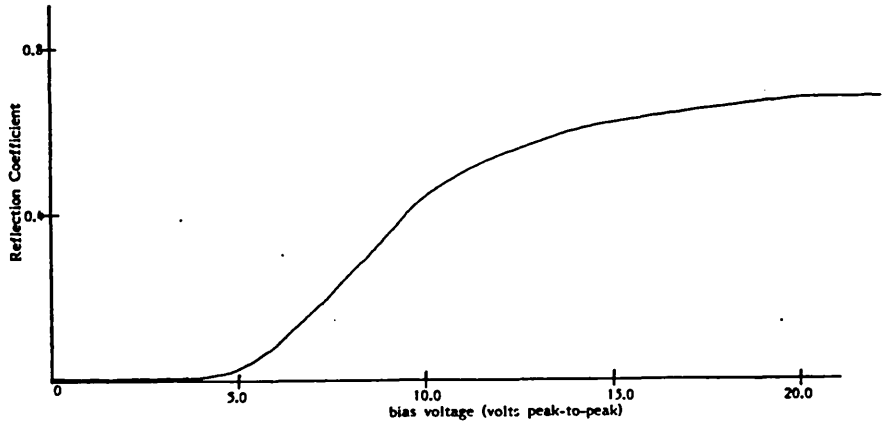


Figure 5.8 Reflected intensity for a surface plasmon LC light modulator, as a function of bias voltage. The reflection coefficient at zero bias is 0.005.

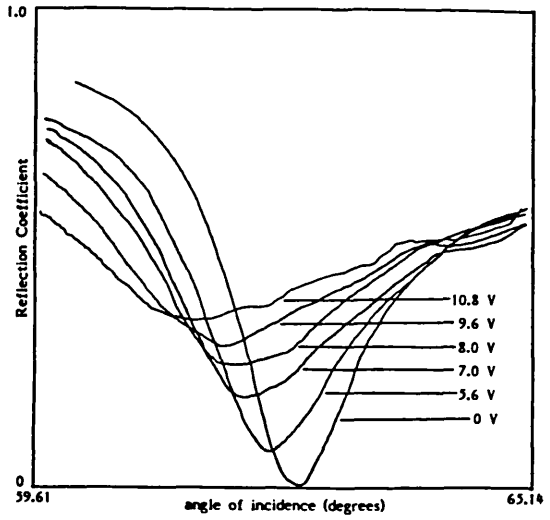


Figure 5.9 SPR curves for the device of figure 5.8, at several values of bias potential as indicated.

The frequency response of one of the devices made is shown in figure 5.10; this was obtained by amplitude modulating the square wave bias with a sinusoidal signal, and measuring the modulation amplitude of the output light intensity as a function of modulation frequency. The frequency for 50% reduction in response is about 700 Hz, which is about ten times higher than for a typical nematic LC cell used in a conventional mode.

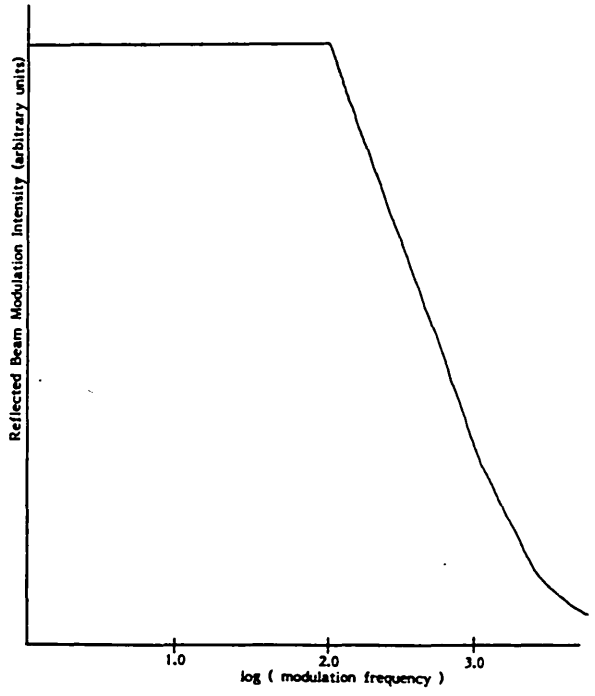


Figure 5.10 Variation of reflectance modulation with bias modulation frequency, for a fixed bias amplitude.

In order to show spatial variation of the modulation, devices were made in which the back plane electrodes were in a pattern of narrow lines, 4 μm wide and 40 μm apart. It proved very difficult to obtain uniform alignment of the liquid crystals in these devices, so that only patches would show a reasonable plasmon resonance and respond to the bias voltage. Figure 5.11 shows a spatial profile of the output intensity across two fingers in one of these patches, with the bias potential on and off. The spatial variation of the field is clearly resolved. This device exhibited on and off switching times of 0.5 ms and 2 ms respectively.

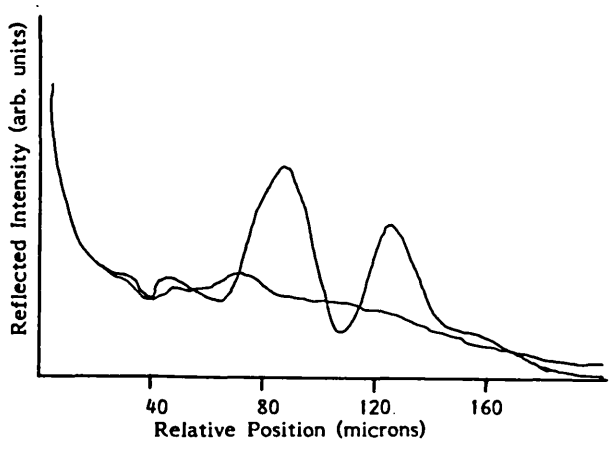


Figure 5.11 Spatial variation of reflected intensity over a portion of a plasmon SLM with digitated back electrodes, for bias on and off; in the 'on' line two digit electrodes are clearly resolved.

CHAPTER 6

CONCLUSION

The feasibility of using prism coupled surface plasmon measurements both for microscopy and for spatial light modulation has been demonstrated, and a theoretical model has been developed by which the excitation of surface plasmons on non-uniform structures can be analyzed. From the results presented, a number of conclusions can be drawn, and a number of recommendations made for future work in this area.

The model presented in section 3.3 has proved very useful in allowing both numerical and analytical prediction of the interaction of surface plasmons with various types of features. Some experimental evidence exists for its validity, as discussed in section 3.4, but it is very limited. It would be of considerable value to test the accuracy of the model in a precise and rigorous way over a wide range of feature types and values of the relevant parameters. In particular, this would give a better idea of the circumstances in which scattering into bulk modes in the dielectric, which is not included in the model, causes significant error in the results. Extension of the analysis to include this scattering would of course be of value.

The analysis presented has all been based on a two-dimensional approximation, the direction perpendicular to the plane of incidence being assumed uniform in all respects. This means that the features analyzed have been one-dimensional, when in most practical situations they will be two-dimensional. To some extent this approximation is legitimate because the important dimension for the interaction of plasmons with the structure is in their direction of propagation; it seems reasonable to assume that interaction along the surface in the orthogonal direction will be mainly determined by conventional diffraction considerations. That is to say, because the resolution limit imposed by the plasmon propagation length is about an order of magnitude greater than the wavelength, in a first order approximation we can neglect conventional diffraction phenomena and assume that the feature will only be distorted in the direction of propagation. However, rigorous analysis of the point would be of considerable benefit.

The results of the theoretical model essentially confirm the initial prediction that resolution would be limited by the surface plasmon decay length. They also show the nature of the image distortions caused by this limitation, and how these distortions depend on the amplitude (contrast) of the features. This also means that one could now imagine developing image processing routines to remove or decrease the image degradation, for instance by some form of deconvolution. In cases where

there is some *a priori* knowledge of the nature of the objects being examined, image restoring algorithms might prove highly effective. Another very attractive possibility for improving the lateral resolution is to combine images of a structure taken at several directions of plasmon propagation along the same surface, to take advantage of the directional nature of the measurement. For any work that may be attempted in developing image processing routines for surface plasmon microscopy, the analysis presented indicates that plane wave images will be easier to analyze and interpret than scanned images, and that a series of plane wave (actually broad-beam) images is preferable to a complete set of resonance curves at each scan position for this reason.

The potential application of surface plasmon microscopy will be primarily limited by its structural requirements. This is essentially a method of looking at variations of the optical properties of a planar interface between a metal and a dielectric, where one of these materials must be of an appropriate thickness to allow excitation by prism coupling. This means it is well suited to analysis of the thin films themselves, and could be very useful for examining the structure of other thin films which can be deposited on metals. The contrast mechanism is different from other types of microscopy, which means that new information will be obtained from any system examined. Also, the technique is non-destructive, highly sensitive, and can produce images essentially instantaneously, which suggests its use for dynamic examination of chemical and biological thin film processes.

The technique presented will be restricted to systems which include a metal layer, and if this layer is not part of the object it must be part of the instrument, and brought into very close proximity to the object. One could, for instance, conceive of an instrument for examining the uniformity of silicon wafers, in which a silver film on a microscope head is brought to within a half micron or less of the Si surface and then the plasmon image obtained. However, the examination of metal film surfaces is certainly of interest in itself; one possible immediate application of surface plasmon microscopy is the investigation of oxidation and anodization processes on aluminium films.

The use of surface plasmons for spatial light modulation is very promising. Immediate improvements in sensitivity, speed and phase uniformity are indicated, and exciting prospects exist for the use of entirely new active thin film techniques to which the use of plasmons is ideally suited. The resolution limit imposed by the decay length is acceptable even for silver, and could be improved considerably by

use of a lossier metal, albeit with a corresponding reduction in sensitivity. The use of grating rather than prism coupling in a plasmon SLM also warrants investigation, this would lead to a planar geometry which could be more convenient for some implementations. While only liquid crystal devices have been constructed to date, implementations using active semiconductor structures are currently being considered. If plasmons are generated with part of the field in a semiconductor in contact with the metal, then flooding this junction region with carriers could alter the optical properties of the semiconductor sufficiently to modulate the plasmon response. This flooding could be controlled in an optically addressed configuration by photon generation of carriers in another layer. Designs along these lines are currently being investigated.

BIBLIOGRAPHY

Chapter 1

1.1 R.W. Wood, Physical Optics, New York: MacMillan, 1934.

1.2 A. Sommerfeld, Ann. Physik 28, 665 (1909).

1.3 J. Zenneck, Ann. Physik 23, 846 (1907).

1.4 R.H. Ritchie and H.B. Eldridge, Phys. Rev. 126, 1935 (1962).

1.5 A. Otto, Z. Physik 216, 398 (1968).

1.6 E. Kretschmann, Z. Physik 241, 313 (1971).

1.7 R.H. Ritchie, Surface Science 34, 1 (1973).

1.8 F. Abelès, Surf. Sci. 56, 237 (1976).

1.9 H. Raether, Phys. Thin Films 9, 145 (1977).

1.10 I. Pockrand, J.D. Swalen, J.G. Gordon II and M.R. Philpott, Surf. Sci. 74, 237 (1977).

1.11 M.T. Flanagan and R.H. Pantell, Electronics Letters 20, 968 (1984).

1.12 C. Nylander, B. Liedberg and T. Lind, Sensors & Actuators 3, 79 (1982/83).

1.13 I. Pockrand, Surf. Sci. 72, 577 (1978).

1.14 J. Bosenberg, Z. Phys B 22, 267 (1975).

1.15 S.O. Sari, D.K. Cohen and K.D. Scherkoske, Phys. Rev. B 21, 2162 (1980).

1.16 U.C. Fischer, U.T. Durig and D.W. Pohl, App. Phys. Lett. 52, 249 (1988).

1.17 U.C. Fischer, D.W. Pohl and F. Rohmer, J. App. Phys. 59, 3318 (1986).

1.18 J. Wessel, J. Optical Soc. Amer. B 2, 1538 (1985).

1.19 M.G. Somekh, H.L. Bertoni, G.A.D. Briggs and N.J. Burton, Proc. Roy. Soc. London A 401, 29 (1985).

1.20 E.M. Yeatman and E.A. Ash, Elect. Lett. 23, 1091 (1987).

1.21 B. Rothenhausler and W. Knoll, App. Phys. Lett. 52, 1554 (1988).

Chapter 2

2.1 W. Hansen, J. Opt. Soc. Am. 58, 380 (1968).

2.2 Kretschmann (ref. 1.6).

2.3 Raether (ref. 1.9), p. 155.

2.4 W.D. Mason, Prism Coupler Surface Sensing, Ph.D. thesis, Dept. of Electrical Engineering, Imperial College, U. of London, 1988.

2.5 F. Abelès, Ann. Phys. (Paris) 5, 596 (1950).

2.6 L.G. Schultz, J. Opt. Soc. Am. 44, 357 (1954).

2.7 J.E. Midwinter, Optical Fibres for Transmission, New York: Wiley and Sons, 1979, pp. 33–39.

2.8 F.Y. Kou and T. Tamir, Optics Lett. 12, 367 (1987).

2.9 K. Welford, in Surface Plasmon Polaritons, (IoP Short Meetings Series No. 9), London: Institute of Physics, 1988, pp. 25–78.

Chapter 3

3.1 W.H. Beyer, Ed., CRC Standard Mathematical Tables, Boca Raton, Florida: CRC Press, 1978.

3.2 R. Ulrich, J. Opt. Soc. Am. 61, 1467 (1971).

3.3 P.K. Tien and R. Ulrich, J. Opt. Soc. Am. 60, 1325 (1970).

3.4 T.A. Leskova, Solid State Comm. 50, 869 (1984).

3.5 Prof. C. Atkinson, Imperial College, private communication.

3.6 H.E.M. Barlow and J. Brown, Radio Surface Waves, Oxford: Clarendon Press, 1962, pp. 137–142.

3.7 Rothenhausler (ref. 1.21).

Chapter 5

5.1 D. Casasent, Proc. IEEE 65, 143 (1977).

5.2 N. Collings, in IEE electronics colloquium on "Optical Techniques for Information Processing", digest no. 1987/105.

5.3 U. Efron, J. Grinberg, P.O. Braatz, M.J. Little, P.G. Reif and R.N. Schwartz, J. App. Phys. 57, 1356 (1985).

5.4 S.G. Latham and M.P. Owen, GEC Journal of Research 4, 219 (1986).

5.5 P.R. Ashley and J.H. Davis, App. Optics 26, 241 (1987).

5.6 M.C. Hebbron and S.S. Makh, Proc. SPIE vol. 825 (1987).

5.7 M.S. Welkowsky, U. Efron, W. Byles and N.W. Goodwin, Optical Engineering 26, 414 (1987).

5.8 H.K. Liu, J.A. Davis and R.A. Lilly, Opt. Lett. 10, 635 (1985).

5.9 D.A. Gregory, App. Optics 25, 467 (1986).

5.10 R.A. Sprague, J. App. Phys. 46, 1673 (1975).

5.11 B.A. Horwitz and F.J. Corbett, Opt. Eng. 17, 353 (1978).

5.12 M.P. Petrov, A.V. Khomenko, M.V. Krasinkova, V.I. Marakhonov and M.G. Shlyagin, Sov. Phys. Tech. Phys. 26, 816 (1981).

5.13 C. Warde, A.M. Weiss, A.D. Fisher and J.I. Thackara, App. Optics 20, 2066 (1981).

5.14 C. Warde and J.I. Thackara, Opt. Eng. 22, 695 (1983).

5.15 B. Schneeberger, F. Laeri and T. Tschudi, Opt. Comm. 31, 13 (1979).

5.16 K. Hess, R. Dandliker and R. Thalmann, Opt. Eng. 26, 418 (1987).

5.17 D.G. Vass, R.M. Sillitto, I. Underwood, D.J. McKnight, P.H. Willson and M.J. Ranshaw, in IEE electronics colloquium on "Optical Techniques for Information Processing", digest no. 1987/105.

5.18 W.E. Ross, D. Psaltis and R.H. Anderson, Opt. Eng. 22, 485 (1983).

5.19 D.L. Flannery, A.M. Biernacki, J.S. Loomis and S.L. Cartwright, App. Optics 25, 466 (1986).

5.20 D. Casasent, F. Caimi, App. Optics 11, 1426 (1972).

5.21 S.C. Esener, J.H. Wang, T.J. Drabik, M.A. Title and S.H. Lee, Opt. Eng. 26, 406 (1987).

5.22 R.R.A. Syms, Imperial College, private communication.

5.23 R. Reinisch, P. Vincent, M. Neviere and E. Pic, App. Optics 24, 2001 (1985).

5.24 C. Plumereau, A. Bouchoux and A. Cachard, SPIE vol. 8000, 79 (1987).

5.25 G.T. Sincerbox and J.C. Gordon II, App. Optics 20, 1491 (1981).

5.26 M. Green and J.S. Menown, Proc. SID 26, 163 (1985).

5.27 K.R. Welford and J.R. Sambles, App. Phys Lett. 50, 871 (1987).

5.28 J. Cognard, Mol. Cryst. Liq. Cryst. supp. 1, (1982).