Wavelength-Modulated Surface Plasmon Resonance

by

Jeremy R. Cooper

Submitted in partial fulfillment of the requirements of the University of Washington Evening M.S. Degree Program in Physics

August 21, 2002

Project Advisor: Prof. Larry Sorensen

Table of Contents

Section 1: Introduction

My interests lie at least as much in the teaching of science as in my own particular scientific research. Thus, my goals in this paper are two-fold: First, I would like to develop the theory of surface plasmon resonance (SPR) in a clear and detailed manner. Second, I would like to present my own work in the development of a variety of techniques for wavelength-modulated SPR.

A very large portion of this paper (all of Section 2) is spent developing the basic theory of surface plasmons. Understanding this theory was a big challenge for me and I found myself frequently discouraged that none of the papers and books I reviewed on the subject provided a thorough enough analysis for me to grasp. Thus, in Section 2, I've tried to present a development of surface plasmon theory that I might have wished for as a beginner to the subject. My hope is that someone like myself (who may not be able to blindly whip through applications of Maxwell's equations, but with a little help is more than capable of following the theory) will be able to learn the physical principles of SPR with minimal frustration. Section 2 may no doubt be tedious to a reader well versed in surface plasmon theory or electromagnetic theory in general, so feel free to skip to later sections at your own discretion.

Section 3 offers a brief description of the classic techniques for studying SPR (utilizing the Kretschmann geometry). Both angle sweep and wavelength-sweep techniques are covered, including data from both of these techniques.

Section 4 discusses the modulation of input variables as a method of improving upon the sensitivity of the classic techniques for studying SPR. First, I present a generic analysis of the theory behind modulating input variables. Then, I briefly discuss the work of a research group that applied a technique of prism incidence angle modulation to SPR.

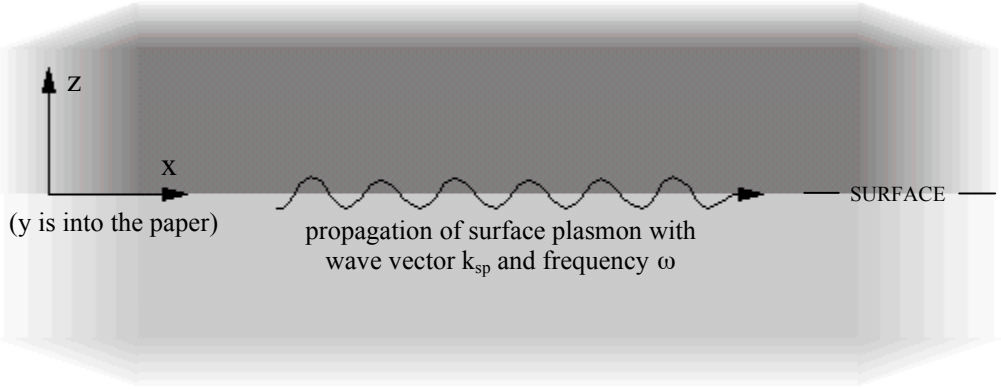
In Section 5, I present my own research which involves three separate techniques of wavelength modulation as applied to SPR. This was the main thrust of my research and it is subsequently the most significant section of this paper. The data from two of these methods demonstrates good sensitivity to very small shifts in the prism geometry, while the third shows moderate potential. Sections 4 and 5 will likely be the most interesting to a reader that is already an expert in the field of surface plasmon resonance.

Section 6 is a brief survey of the practical applications of SPR as a detection device for the characterization of biomolecules (mainly proteins).

Section 2: Surface Plasmon Theory

In this section, I would like to derive the equations that describe surface plasmon waves, starting with a minimum of assumptions: only that the electric and magnetic fields obey the standard solutions to the wave equations for electromagnetic radiation (equations (1) and (2)), and that the direction of propagation is parallel to the surface along what is defined to be the x-axis (see Figure 1).

Figure 1: Surface plasmon propagating along interface in x-direction



By defining the x-axis in this manner, all dependence upon y is eliminated, and thus the mathematics is simplified considerably. Note, however, that the amplitude vectors B_0 and E_0 are in fact allowed a dependence upon z, which will become significant as the derivation progresses. The standard solutions to the wave equations for electromagnetic radiation propagating along the x-axis are as follows:

$$\vec{E} = \vec{E}_0 \left(e^{i(k_{sp} \cdot x - \omega t)} \right) \tag{1}$$

$$\vec{B} = \vec{B}_0 \left(e^{i(k_{sp} \cdot x - \omega t)} \right). \tag{2}$$

I'll begin with Maxwell's Equations (in SI units) for linear media. Note that ε and μ are the *relative* electric permittivity (aka. the dielectric constant) and the *relative* magnetic permeability, respectively, of the material(s) in question. Also note that the **J** term is absent from equation (4) due to the fact that we are considering the case in which there is no free current.

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{3}$$

$$\vec{\nabla} \times \vec{B} = \frac{\mu\varepsilon}{c^2} \cdot \frac{\partial \vec{E}}{\partial t}. \tag{4}$$

Expanding equations (3) and (4) into component form yields:

$$\left(\frac{\partial E_z}{\partial y}-\frac{\partial E_y}{\partial z}\right)\hat{x}+\left(\frac{\partial E_x}{\partial z}-\frac{\partial E_z}{\partial x}\right)\hat{y}+\left(\frac{\partial E_y}{\partial x}-\frac{\partial E_x}{\partial y}\right)\hat{z}=\left(-\frac{\partial B_x}{\partial t}\right)\hat{x}+\left(-\frac{\partial B_y}{\partial t}\right)\hat{y}+\left(-\frac{\partial B_z}{\partial t}\right)\hat{z} \tag{5}$$

$$\left(\frac{\partial B_z}{\partial y}-\frac{\partial B_y}{\partial z}\right)\hat{x}+\left(\frac{\partial B_x}{\partial z}-\frac{\partial B_z}{\partial x}\right)\hat{y}+\left(\frac{\partial B_y}{\partial x}-\frac{\partial B_x}{\partial y}\right)\hat{z}=\left(\frac{\mu\varepsilon}{c^2}\cdot\frac{\partial E_x}{\partial t}\right)\hat{x}+\left(\frac{\mu\varepsilon}{c^2}\cdot\frac{\partial E_y}{\partial t}\right)\hat{y}+\left(\frac{\mu\varepsilon}{c^2}\cdot\frac{\partial E_z}{\partial t}\right)\hat{z}\,. \tag{6}$$

Based upon the previously stated definition that the electric and magnetic fields do not depend upon y, this simplifies to:

$$\left(-\frac{\partial E_y}{\partial z}\right)\hat{x}+\left(\frac{\partial E_x}{\partial z}-\frac{\partial E_z}{\partial x}\right)\hat{y}+\left(\frac{\partial E_y}{\partial x}\right)\hat{z}=\left(-\frac{\partial B_x}{\partial t}\right)\hat{x}+\left(-\frac{\partial B_y}{\partial t}\right)\hat{y}+\left(-\frac{\partial B_z}{\partial t}\right)\hat{z} \tag{7}$$

$$\left(-\frac{\partial B_y}{\partial z}\right)\hat{x}+\left(\frac{\partial B_x}{\partial z}-\frac{\partial B_z}{\partial x}\right)\hat{y}+\left(\frac{\partial B_y}{\partial x}\right)\hat{z}=\left(\frac{\mu\varepsilon}{c^2}\cdot\frac{\partial E_x}{\partial t}\right)\hat{x}+\left(\frac{\mu\varepsilon}{c^2}\cdot\frac{\partial E_y}{\partial t}\right)\hat{y}+\left(\frac{\mu\varepsilon}{c^2}\cdot\frac{\partial E_z}{\partial t}\right)\hat{z}\,. \tag{8}$$

Before proceeding, it is prudent to make an assumption about the polarization of the electromagnetic waves. I shall assume that the waves are polarized such that the magnetic field lies parallel to the surface of propagation (the y-axis). This forces E_y, B_x, and B_z to be zero. (Note: the validity of this assumption will be discussed a bit later). Thus, equations (7) and (8) simplify further to:

$$\left(\frac{\partial E_x}{\partial z}-\frac{\partial E_z}{\partial x}\right)\hat{y}=\left(-\frac{\partial B_y}{\partial t}\right)\hat{y} \tag{9}$$

$$\left(-\frac{\partial B_y}{\partial z}\right)\hat{x}+\left(\frac{\partial B_y}{\partial x}\right)\hat{z}=\left(\frac{\mu\varepsilon}{c^2}\cdot\frac{\partial E_x}{\partial t}\right)\hat{x}+\left(\frac{\mu\varepsilon}{c^2}\cdot\frac{\partial E_z}{\partial t}\right)\hat{z}\,. \tag{10}$$

Splitting these equations into components and executing the noted derivatives upon equations (1) and (2), leaves the following three equations:

$$\frac{\partial E_{0x}}{\partial z}\left(e^{i(k_{sp}\cdot x-\omega t)}\right)-ik_{sp}E_{0z}\left(e^{i(k_{sp}\cdot x-\omega t)}\right)=i\omega B_{0y}\left(e^{i(k_{sp}\cdot x-\omega t)}\right) \tag{11}$$

$$-\frac{\partial B_{0y}}{\partial z}\left(e^{i(k_{sp}\cdot x-\omega t)}\right)=-\frac{i\omega\mu\varepsilon}{c^2}E_{0x}\left(e^{i(k_{sp}\cdot x-\omega t)}\right) \tag{12}$$

$$ik_{sp}B_{0y}\left(e^{i(k_{sp}\cdot x-\omega t)}\right)=-\frac{i\omega\mu\varepsilon}{c^2}E_{0z}\left(e^{i(k_{sp}\cdot x-\omega t)}\right). \tag{13}$$

The oscillating terms of equations (11), (12), and (13) cancel due to the fact that both the wave vector (k_{sp}) and the frequency (ω) remain constant throughout the surface plasmon wave, yielding the following:

$$\frac{\partial E_{0x}}{\partial z} - ik_{sp}E_{0z} = i\omega B_{0y} \tag{14}$$

$$\frac{\partial B_{0y}}{\partial z} = \frac{i\omega\mu\varepsilon}{c^2}E_{0x} \tag{15}$$

$$E_{0z} = -\frac{k_{sp}c^2}{\omega\mu\varepsilon}B_{0y}\,. \tag{16}$$

Substituting equation (16) into equation (14) and simplifying, leaves the following expression:

$$\frac{\partial E_{0x}}{\partial z} = i\omega\left(1 - \frac{k_{sp}^2 c^2}{\omega^2\mu\varepsilon}\right)B_{0y}\,. \tag{17}$$

Taking the partial derivative of both sides of equation (17) with respect to *z* and then plugging in equation (15) results in the following differential equation:

$$\frac{\partial^2 E_{0x}}{\partial z^2} = \left(k_{sp}^2 - \frac{\omega^2\mu\varepsilon}{c^2}\right)E_{0x}\,. \tag{18}$$

The solution of this differential equation is given below, with A being the amplitude of the electromagnetic wave:

$$E_{0x} = Ae^{-\left(\sqrt{k_{sp}^2 - \frac{\omega^2\mu\varepsilon}{c^2}}\right)z}\,. \tag{19}$$

This result can then be applied to equation (17) in order to determine B_{0y}:

$$B_{0y} = \left(-\frac{i\omega\mu\varepsilon}{c^2}\right)\left(k_{sp}^2 - \frac{\omega^2\mu\varepsilon}{c^2}\right)^{-\frac{1}{2}} Ae^{-\left(\sqrt{k_{sp}^2 - \frac{\omega^2\mu\varepsilon}{c^2}}\right)z}\,. \tag{20}$$

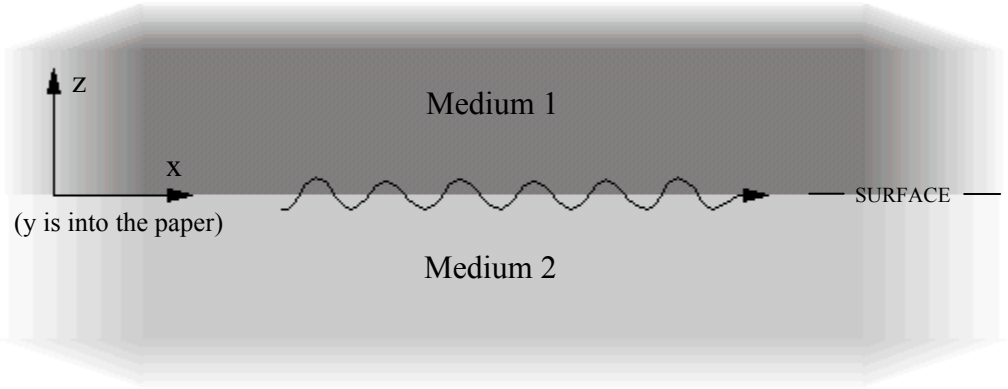
Which can subsequently be applied to equation (16) to determine E_{0z}:

$$E_{0z} = \left(ik_{sp}\right)\left(k_{sp}^2 - \frac{\omega^2\mu\varepsilon}{c^2}\right)^{-\frac{1}{2}} Ae^{-\left(\sqrt{k_{sp}^2 - \frac{\omega^2\mu\varepsilon}{c^2}}\right)z}\,. \tag{21}$$

We can apply these solutions to any sharp interface between two different isotropic and homogenous media. For now, the physical characteristics of the media in question are not restricted in any way, and thus I will refer to them simply as medium 1 and medium 2. However, in the following pages I will develop restrictions upon the physical properties of the involved media that must be met in order for surface plasmon waves to exist.

The general arrangement for the interface between the two media is shown in Figure 2, with the axes defined such that medium 1 is on the positive z side of the origin and medium 2 is on the negative z side of the origin.

Figure 2: Surface plasmon propagates along interface between medium 1 and medium 2



Substituting equations (19), (20), and (21) into equations (1) and (2) gives the electric and magnetic fields at all points near the surface, as follows:

In Medium 1

$$E_{1x} = A_1 \left(e^{-\left(\sqrt{k_{sp}^2 - \frac{\omega^2 \varepsilon_1}{c^2}} \right) z} \right) \left(e^{i(k_{sp} \cdot x - \omega t)} \right) \tag{22}$$

$$E_{1z} = (ik_{sp}) \left(k_{sp}^2 - \frac{\omega^2 \varepsilon_1}{c^2} \right)^{-\frac{1}{2}} A_1 \left(e^{-\left(\sqrt{k_{sp}^2 - \frac{\omega^2 \varepsilon_1}{c^2}} \right) z} \right) \left(e^{i(k_{sp} \cdot x - \omega t)} \right) \tag{23}$$

$$B_{1y} = \left(-\frac{i\omega \varepsilon_1}{c^2} \right) \left(k_{sp}^2 - \frac{\omega^2 \varepsilon_1}{c^2} \right)^{-\frac{1}{2}} A_1 \left(e^{-\left(\sqrt{k_{sp}^2 - \frac{\omega^2 \varepsilon_1}{c^2}} \right) z} \right) \left(e^{i(k_{sp} \cdot x - \omega t)} \right) \tag{24}$$

$$E_{1y} = B_{1x} = B_{1z} = 0 \quad \text{(by definition).}$$

In Medium 2

$$E_{2x} = A_2\left(e^{\left(\sqrt{k_{sp}^2-\frac{\omega^2\varepsilon_2}{c^2}}\right)z}\right)\left(e^{i\left(k_{sp}\cdot x-\omega t\right)}\right) \tag{25}$$

$$E_{2z} = \left(-ik_{sp}\right)\left(k_{sp}^2-\frac{\omega^2\varepsilon_2}{c^2}\right)^{-\frac{1}{2}}A_2\left(e^{\left(\sqrt{k_{sp}^2-\frac{\omega^2\varepsilon_2}{c^2}}\right)z}\right)\left(e^{i\left(k_{sp}\cdot x-\omega t\right)}\right) \tag{26}$$

$$B_{2y} = \left(\frac{i\omega\varepsilon_2}{c^2}\right)\left(k_{sp}^2-\frac{\omega^2\varepsilon_2}{c^2}\right)^{-\frac{1}{2}}A_2\left(e^{\left(\sqrt{k_{sp}^2-\frac{\omega^2\varepsilon_2}{c^2}}\right)z}\right)\left(e^{i\left(k_{sp}\cdot x-\omega t\right)}\right) \tag{27}$$

$E_{2y} = B_{2x} = B_{2z} = 0$ (by definition).

At this point, there are three things to note:

1. The magnetic permeability (μ) is omitted (and will be from here on out) because we will assume that all of the involved media are not magnetic (and thus have a relative magnetic permeability sufficiently close to 1).
2. The dielectric constants (ε) for the two media are distinguished by subscripts 1 and 2.
3. For the set of equations defining the fields in medium 2, all appearances of z merit an extra negative sign (since medium 2 lies below the origin). This also causes the overall signs of E_{2z} and B_{2y} to be reversed from that of E_{1z} and B_{1y}, respectively, due to the partial derivative with respect to z taken when applying equation (17).

Now we take the boundary condition for electric fields *parallel* to a surface, $E_{1\|} = E_{2\|}$ (at z = 0), and apply it to equations (22) and (25) to determine that

$$A_1 = A_2 . \tag{28}$$

The boundary condition for electric fields perpendicular to a surface, $\varepsilon_1 E_{1\perp} = \varepsilon_2 E_{2\perp}$ (at z = 0), can then be applied to equations (23) and (26), which gives

$$\varepsilon_1\left(k_{sp}^2-\frac{\omega^2\varepsilon_1}{c^2}\right)^{-\frac{1}{2}} = -\varepsilon_2\left(k_{sp}^2-\frac{\omega^2\varepsilon_2}{c^2}\right)^{-\frac{1}{2}} . \tag{29}$$

We know that the $\left(k_{sp}^2 - \frac{\omega^2 \varepsilon_1}{c^2}\right)^{-\frac{1}{2}}$ term and the $\left(k_{sp}^2 - \frac{\omega^2 \varepsilon_2}{c^2}\right)^{-\frac{1}{2}}$ term must both be positive, otherwise equations (22) through (27) would diverge as z approached either negative or positive infinity. Therefore, in order for equation (29) to be true, either ε_1 or ε_2 (but not both) must be negative. For the sake of definition, I will choose that ε_2 is positive which forces ε_1 to be negative.

The next step is to derive a dispersion relationship from equation (29). Squaring both sides and rearranging gives the following:

$$\varepsilon_1^2 k_{sp}^2 - \frac{\omega^2 \varepsilon_2 \varepsilon_1^2}{c^2} = \varepsilon_2^2 k_{sp}^2 - \frac{\omega^2 \varepsilon_1 \varepsilon_2^2}{c^2}. \tag{30}$$

Now, solving equation (30) for k_{sp} and simplifying the result produces the well known dispersion relationship for surface plasmons,

$$k_{sp} = \frac{\omega}{c}\left(\frac{\varepsilon_1 \varepsilon_2}{\varepsilon_1 + \varepsilon_2}\right)^{\frac{1}{2}}. \tag{31}$$

Note that this result reveals another constraint upon the relative dielectric constants of the two media: namely that $-\varepsilon_1 > \varepsilon_2$. If this were not the case, k_{sp} would be imaginary (since we have already shown that ε_1 must be negative) and thus a propagating, oscillating wave could not exist.

With these constraints in mind, I'll go ahead and add a bit more definition to the emerging physical arrangement that is necessary for surface plasmons to exist. First, since ε_1 is negative, we can say that medium 1 must be either a metal or a semiconductor (in practice, silver and gold are commonly used, but other metals, such as aluminum, copper, nickel, and platinum also have some functionality[1]). Second, ε_2 should be small and positive, which is easily fulfilled by a variety materials, the most practical of which being air. Air is not only relatively abundant, but with a dielectric constant nearly equal to one, it makes the rest of the analysis a bit easier.

An important characteristic of metals is that their dielectric constants are not in fact constant. Instead, they depend significantly upon the frequency of electromagnetic waves traveling through the material. Luckily, the dielectric constant can be approximated by the following relationship[2]:

$$\varepsilon(\omega) = 1 - \left(\frac{\omega_p}{\omega}\right)^2. \tag{32}$$

The relationship stated in equation (32) is substituted for ε_1 in equation (31), while setting $\varepsilon_2 = 1$ (for air). Squaring both sides then gives the following:

$$k_{sp}^2 = \frac{\omega^2}{c^2}\left(\frac{1-\left(\frac{\omega_p}{\omega}\right)^2}{2-\left(\frac{\omega_p}{\omega}\right)^2}\right). \tag{33}$$

If both sides are multiplied by $\left(\omega^2\left(2-\left(\frac{\omega_p}{\omega}\right)^2\right)\right)$, equation (33) simplifies to

$$2k_{sp}^2\omega^2 - k_{sp}^2\omega_p^2 = \frac{\omega^4}{c^2} - \frac{\omega_p^2\omega^2}{c^2} \tag{34}$$

and a bit more rearrangement results in the equation

$$\omega^4 - \left(2c^2k_{sp}^2 + \omega_p^2\right)\omega^2 + c^2k_{sp}^2\omega_p^2 = 0\,. \tag{35}$$

Using the quadratic formula to solve for ω^2 gives the following result:

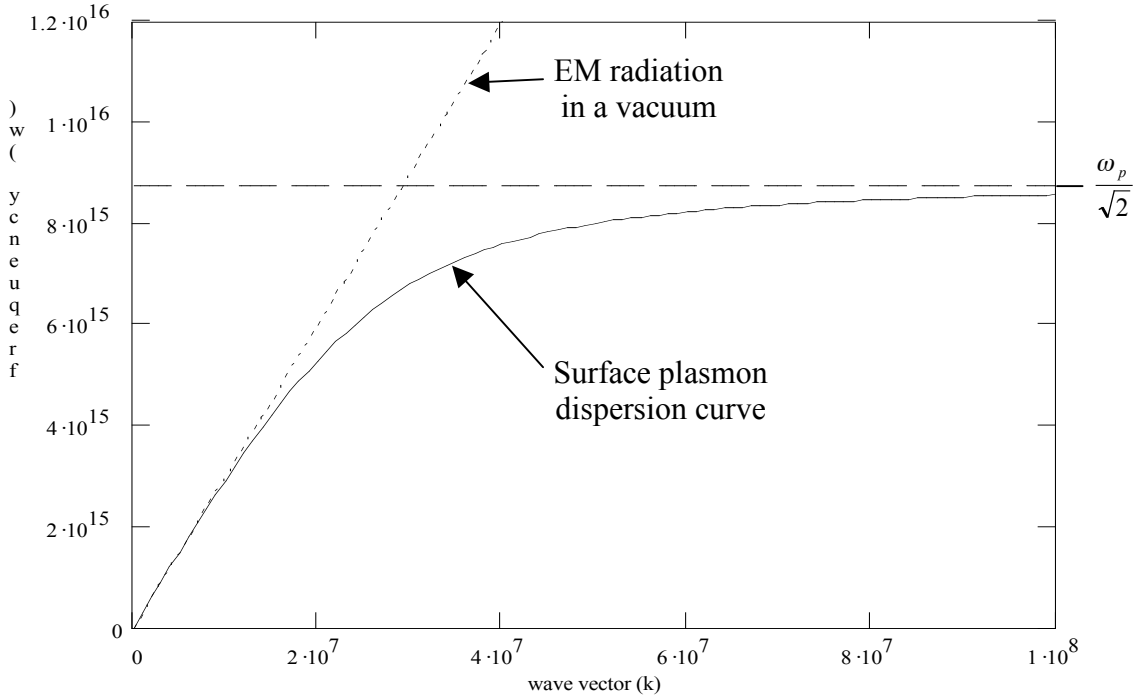
$$\omega^2 = \frac{\omega_p^2}{2} + \left(ck_{sp}\right)^2 \pm \sqrt{\frac{\omega_p^2}{4} + \left(ck_{sp}\right)^4}\,. \tag{36}$$

In equation (36), the solution with the positive square root term requires that $\omega > ck_{sp}$, which is impossible because it would mean that the wave had a velocity greater than the speed of light. Thus, the only physically relevant solution is the one with the negative term. Keeping this term and solving for ω results in a more complete dispersion relationship for surface plasmon waves:

$$\omega = \sqrt{\frac{\omega_p^2}{2} + \left(ck_{sp}\right)^2 - \sqrt{\frac{\omega_p^2}{4} + \left(ck_{sp}\right)^4}} \tag{37}$$

An informative illustration can be made by plotting this dispersion relationship (See Figure 3 on the next page).

Figure 3: Plot of surface plasmon dispersion relationship



In this plot, the solid line is a plot of the dispersion relationship given by equation (37) and the dashed line corresponds to the value $\omega = \frac{\omega_p}{\sqrt{2}}$, which the surface plasmon frequency tends towards for large wave vectors (k_{sp}). Of great importance is the dotted line, which represents the relationship $\omega = ck$ (i.e. the dispersion relationship for a photon in a vacuum). Although the surface plasmon dispersion relationship tends towards this line for small wave vectors, the two never in fact intersect (except at the origin, which is trivial). This creates a problem when attempting to initiate surface plasmon waves. In order to drive the initiation of surface plasmons from photons, we must match both their frequencies and wave vectors at the same time. For a single photon being converted into a single plasmon, this is equivalent to matching the both the energy and momentum of the two particles, since $E = \hbar\omega$ and $p = \hbar k$.

It is, however, possible to match both the frequency and wave vector if the electromagnetic radiation (which is driving the surface plasmon waves) is traveling through a medium with a large index of refraction ($n = \sqrt{\varepsilon}$). The wave vector for such a wave is increased by a factor of the index of refraction according to the following:

$$k = \frac{\omega}{c} n \tag{38}$$

A plot of this dispersion relationship is similar to that of the straight line from standard electromagnetic radiation, except now the slope is decreased by a factor of $1/n$. This allows a single point of intersection between the respective dispersion relationships of the surface plasmons and the electromagnetic radiation. Thus, if the incident electromagnetic radiation is

traveling through a medium with an index of refraction significantly greater than one (such as glass), it is possible to initiate surface plasmons. A plot of this can be seen in Figure 7.

The only problem now is designing a physical arrangement for which this kind of coupling is possible. Neither the metal nor the air have the optical properties required to be the medium through which the incident radiation travels (this is because the metal has a negative dielectric constant and the air has a dielectric constant which is nearly equal to one).

The solution to this problem is subtle, but not too complicated. We can place a third medium (usually glass) on the opposing side of either the air or metal layer. These two possible arrangements are shown in Figures 4 and 5 below:

Figure 4: Kretschmann Geometry

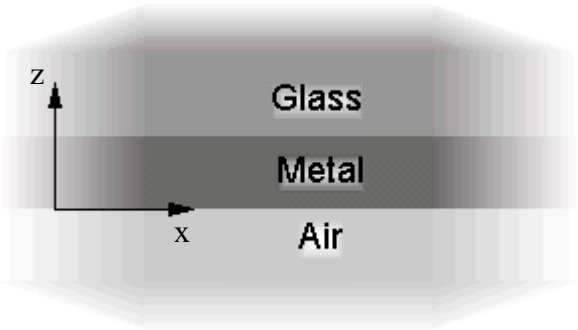
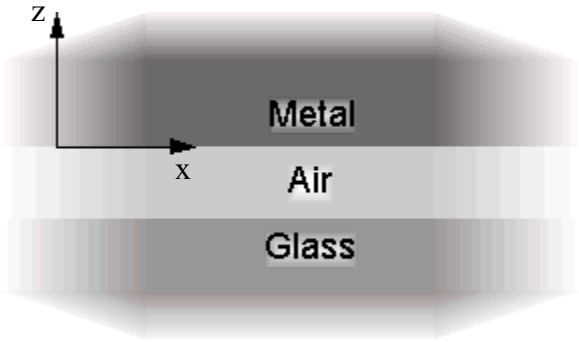


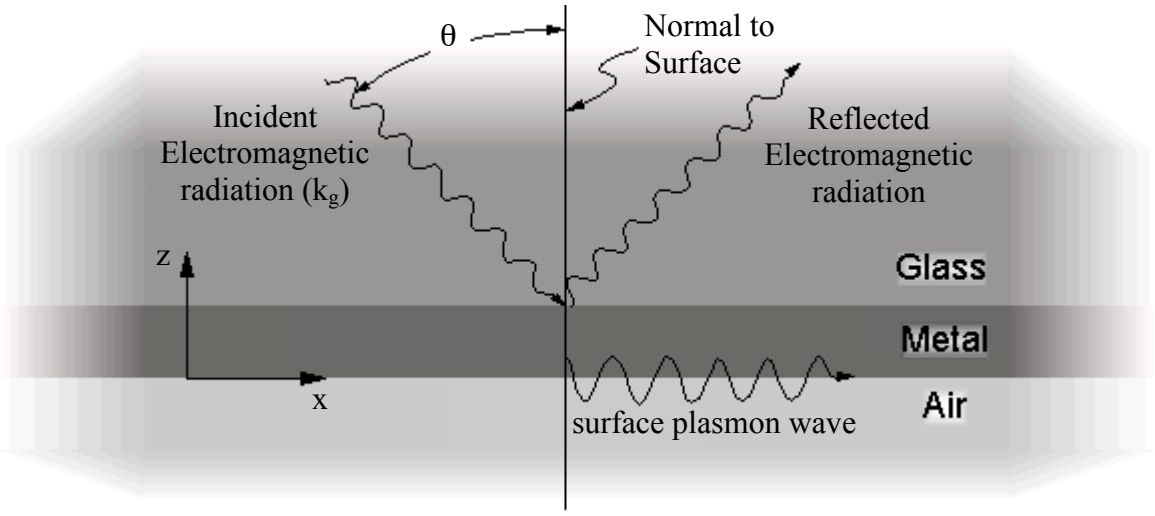
Figure 5: Otto Geometry



The arrangement shown in Figure 4 is commonly referred to as the Kretschmann geometry after the work by E. Kretschmann and H. Raether in 1968[3]. The second arrangement (shown in Figure 5) is referred to as the Otto geometry after a similar experiment performed by A. Otto[4]. Although Otto's work preceded that of Kretschmann and Raether, the Kretschmann geometry is an overwhelmingly more common method of creating surface plasmons in current research. The major reason for this is that the Kretschmann geometry is much more practical to assemble. As I will discuss later, the metal layer (or the air layer for the Otto geometry) must be exceedingly thin (less than 100 nm). For the Kretschmann geometry, a thin and consistent metal layer can be produced without too much trouble by vapor deposition onto the glass surface. However, with the Otto geometry, the air layer is created by moving a glass and metal block to within 50 nm or so of each other. Controlling this gap width can be a bit of challenge, which is why the Otto geometry is not commonly used. However, once a finely tuned apparatus is in place to control the gap width, the Otto geometry shows considerable merit due to the fact that the gap width can be adjusted to the desired value. An additional bonus is that the Otto geometry is a bit easier to analyze from a reflectivity standpoint, however, that is beyond the scope of this paper (see Sprokel and Swalen[5] for a simple derivation of the reflectivity for both Otto and Kretschmann geometries. See Hansen[6] for a more complete derivation of the reflectivity for the Kretschmann geometry with multiple layers).

In the Kretschmann geometry, we can think of electromagnetic radiation that travels through the glass medium and encounters the metal layer at some angle of incidence (θ) as shown in figure 6. In figure 6, the wave vector of the incident electromagnetic radiation (k_g) has x and z components but no y component (which is consistent with our original definition about the direction of surface plasmon propagation). This implies that the plane of incidence (defined by the lines representing the incident and reflected wave vectors) is parallel to the x-z plane.

Figure 6: Electromagnetic radiation (photons) converted into surface plasmons



As stated previously, the wave vectors must match in order for coupling to take place. Thus, since the wave vector of the surface plasmons (k_{sp}) lies in the x-direction, the x-component of k_g must be equal to k_{sp}, as follows:

$$k_g \sin(\theta_r) = k_{sp}. \tag{39}$$

Here, θ_r is the angle of incidence that is required for coupling to occur. Equations (31) and (38) can then be substituted into equation (39) to find

$$\frac{\omega}{c} n_g \sin(\theta_r) = \frac{\omega}{c}\left(\frac{\varepsilon_1 \varepsilon_2}{\varepsilon_1 + \varepsilon_2}\right)^{\frac{1}{2}}. \tag{40}$$

Here, n_g represents the index of refraction of the glass coupling medium. Simplifying equation (40) and inserting one for the dielectric constant of air (ε_2) gives the following result:

$$n_g \sin(\theta_r) = \left(\frac{\varepsilon_m}{\varepsilon_m + 1}\right)^{\frac{1}{2}}. \tag{41}$$

In equation (41), ε_1 has been more explicitly labeled as ε_m, which is the dielectric constant of the metal layer. To fulfill the constraint that $-\varepsilon_1 > \varepsilon_2$ (which is discussed immediately following equation (31)), ε_m must be less than negative one. Subsequently, the right side of equation (41) must be slightly greater than one. Thus, we can also state that

$$n_g \sin(\theta_r) > 1. \tag{42}$$

We know from optics that:

$$n_g \sin(\theta_c) = 1 \tag{43}$$

where θ_c is the critical angle, for incidence angles above which total internal reflection occurs. Thus, the following must be true:

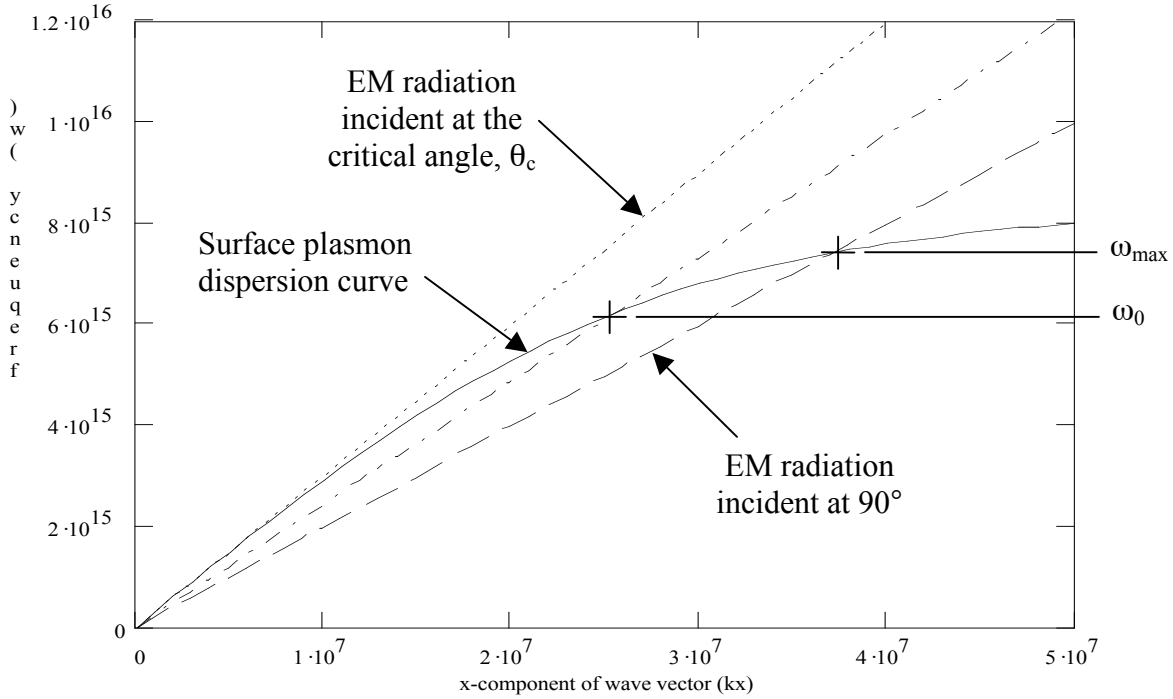
$$\theta_r > \theta_c \tag{44}$$

This is an interesting result because it means that surface plasmons can be generated only when the incident electromagnetic radiation is experiencing total internal reflection. So, in a case such as this, the wave that transfers energy across the metal layer (from the point of incidence on the glass/metal boundary to the surface where plasmons can exist on the metal/air boundary) must be evanescent. An evanescent wave is one that decays to zero at only a very short distance away (about a wavelength or so) from its point of origin[7]. The neat thing about an evanescent wave such as this is that it maintains the same wave vector (k_g) as it had in the glass, even though it is actually traversing the metal layer. Thus, it can be used quite effectively to initiate surface plasmons. The only restriction is that the metal layer must be quite thin (usually less than 100 nm) in order for the evanescent wave to have maintained a significant portion of its amplitude by the time it reaches the metal/air boundary. This same concept applies to the Otto geometry, except that in the case of the Otto geometry it is the *air* layer that must remain very thin.

Now, in order to initiate surface plasmons, we must tune either the incidence angle or the frequency of incident radiation until the condition of equation (41) is met (keeping in mind that ε_m is dependant on the frequency of electromagnetic radiation). The most common method is to keep the frequency fixed at a value ω_o and tune the angle of incidence. In Figure 7 below, the surface plasmon dispersion curve (solid line) is plotted against the dispersion relationships of the incident radiation for a variety of incidence angles (note that only the x-component of the wave vectors are accounted for in this plot, which is why the incidence angle makes a difference). The dashed line represents electromagnetic radiation with an incidence angle of 90°. The dotted line is the dispersion relationship for radiation incident at the critical angle, θ_c (which does not intersect the surface plasmon dispersion curve except for at the origin). Finally, the dot-dash line represents electromagnetic radiation with a tunable angle of incidence. Adjustment of the incidence angle causes a shift in the slope of the dot-dash line, which is dependant on θ_r by a factor of $1/\sin(\theta_r)$. In this way it is possible to match both the frequency and wave vector for any predetermined value of ω_0. The only constraint is that the incidence angle must be maintained within a range of $\theta_c < \theta_r < 90°$. Thus, ω_o cannot exceed ω_{max} (the point at which the dashed line intersects the surface plasmon dispersion curve).

It is also possible to hold the angle of incidence fixed at a value θ_o and tune the frequency of incident electromagnetic radiation. Graphically speaking (in terms of figure 7), this is equivalent to fixing the slope of the dot-dash line and then sliding up or down along this line to reach the point of intersection with the surface plasmon dispersion curve. This method is much

Figure 7: Surface plasmon dispersion relationship for a variety of incidence angles



less common than the method of variable incidence angle because it requires a light source with an accurately tunable frequency (whereas adjusting the angle of incidence only requires some accurate means of rotating the sample surface).

Surface plasmons only exist with a polarization in which their magnetic field vectors lie parallel to the surface. I "blindly" made this assumption near the beginning of the preceding derivation (see equations 9 and 10), however it turns out that surface plasmons of the alternate polarization (ie. electric field vector parallel to the surface) are not physically possible. A proof of this is offered in the appendix at the end of this paper. Because the electric and magnetic fields must match in order for energy transfer to take place, the incident electromagnetic radiation which initiates these plasmons must have a similar orientation. Thus, one final and important requirement must be placed upon the polarization of the incident radiation, namely that its oscillating magnetic field vector be perpendicular to the plane of incidence (which is referred to as TM polarization from "transverse magnetic"). Incident at an angle θ (as shown in figure 6), TM light has components of the electric field in both the x and z-directions, but only a y-component of the magnetic field. Notice that this corresponds nicely with our derived fields for surface plasmons (equations (22) through (27)).

To summarize this section, here are the requirements that must be fulfilled in order to generate surface plasmons:

a) One of the two media that defines the surface must have a negative dielectric constant ($\varepsilon_1 < 0$). Metals fulfill this requirement nicely.
b) The second medium must have a small, positive dielectric constant. Air works well for this ($\varepsilon_2 \cong 1$).

c) Of the two dielectric constants, the one belonging to the metal must have the greater magnitude: $|\varepsilon_1| > |\varepsilon_2|$, Which means that $-\varepsilon_1 > \varepsilon_2$.
d) A third medium (with $\varepsilon_3 > 1$) must be used to couple the energy and momentum of incident photons to that of surface plasmons. Optical glass is the standard material for accomplishing this task.
e) For the Kretschmann geometry (see Figure 4), the thickness of the metal layer must be significantly less than the wavelength of the incident electromagnetic radiation. With visible light, a metal layer thickness between 50 nm and 100 nm is commonly used.
f) The electromagnetic radiation must be at an angle of incidence which fulfills the dispersion relationship in equation (41). This requires that the incidence angle be greater than the critical angle for total internal reflection.
g) The incident radiation must have a transverse magnetic (TM) polarization (magnetic field vector lies perpendicular to the plane of incidence).

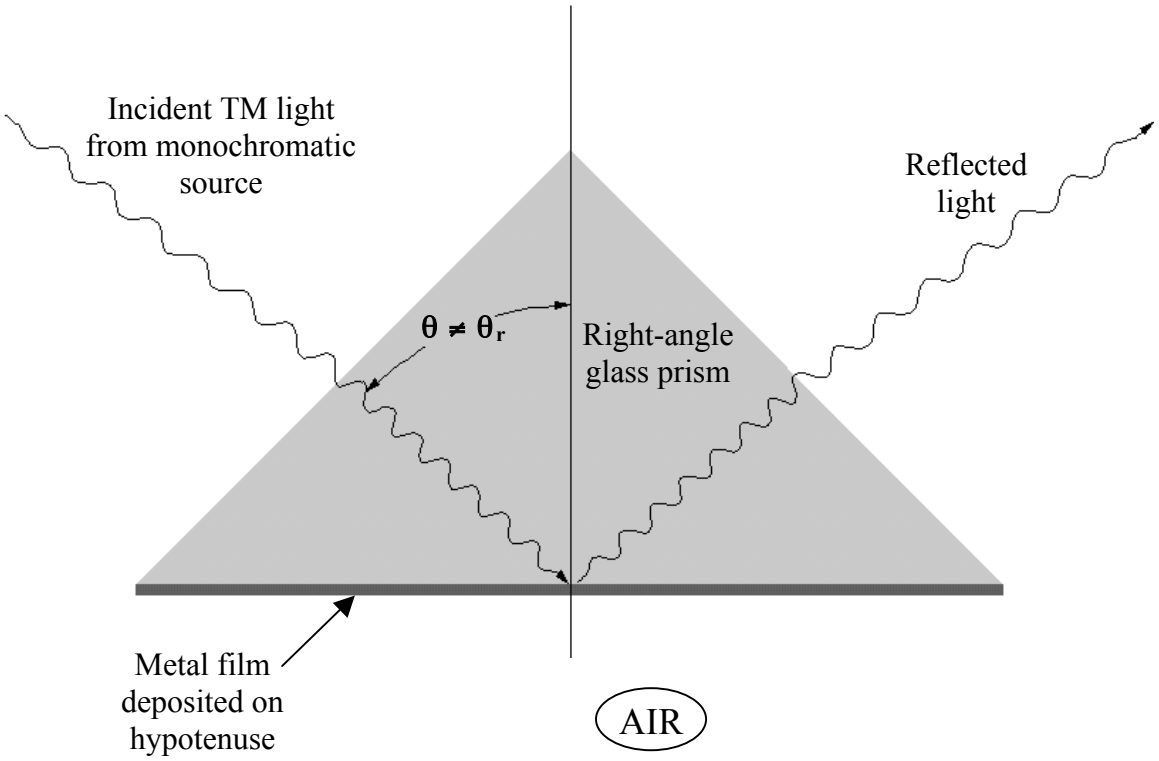
With these requirements in mind, the following section will discuss the phenomenon of surface plasmon resonance in practice.

Section 3: Practical Surface Plasmon Resonance

The preceding section discussed surface plasmon theory and described the conditions necessary to generate surface plasmons. One of the most important requirements is stated in equation (41), which describes the relationship between the incidence angle and frequency in order for coupling to occur (note that ε_m is dependant upon the frequency via equation (32)). If we take the common technique of holding the frequency fixed (at ω_o) while varying the incidence angle, this implies that the transfer of energy from photons to surface plasmons will only occur at a sharply defined region around the angle θ_r. Thus, θ_r is referred to as the *resonance angle* and the phenomenon as a whole is described as surface plasmon resonance (SPR). Alternatively, surface plasmon resonance can also be detected when the incidence angle is fixed at a predetermined value θ_o while the frequency is swept through a range of values. In this case, energy transfer occurs at a frequency ω_r which is referred to as the *resonance frequency*.

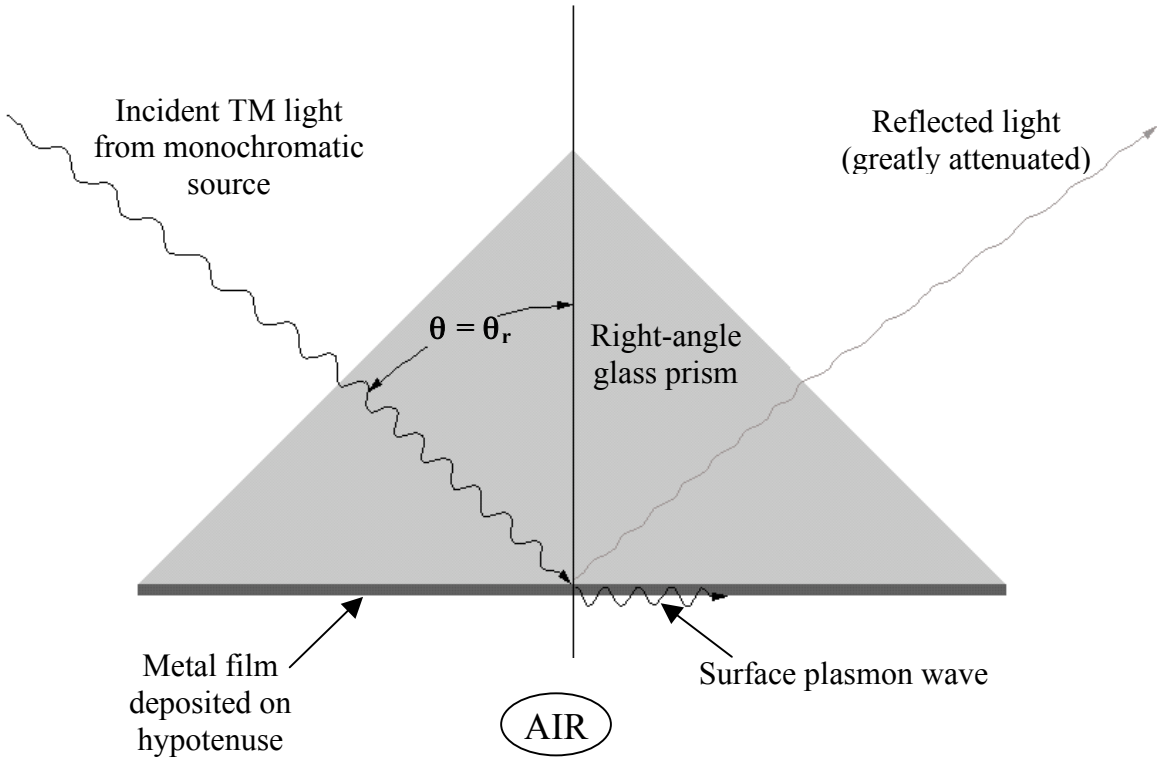
In practice, surface plasmon resonance is commonly achieved by using a right-angle glass prism. A thin layer of metal (usually gold or silver) is deposited onto the hypotenuse face of this prism. Then, monochromatic TM light is directed at one of the open faces such that it travels through the prism and strikes the hypotenuse surface at an angle greater than the critical angle. Since this is a total internal reflection scenario, all of the incident light should be reflected off the back surface unless the incidence angle is sufficiently close to the surface plasmon resonance angle (θ_r). See Figure 8 below:

Figure 8: Kretschmann geometry – not at resonance



However, when the resonance angle is reached, the majority of the incident light will be absorbed by the surface in the form of surface plasmons (See Figure 9 below). Surface plasmons themselves are difficult to detect because they cannot decay radiatively. However, it is quite easy to detect the amount of light that is absorbed by measuring the reflected light intensity. At incidence angles 1°-2° above or below the resonance angle, the reflected light will have an intensity near 100%. However, at the resonance angle, the reflected light intensity can drop almost to zero as the incident light energy is converted almost entirely to surface plasmons. See Figure 11 for an example of this abrupt dip in reflectivity.

Figure 9: Kretschmann geometry – at resonance



An experiment to detect the phenomenon of surface plasmon resonance can be carried out as follows.

First, the prism shown in Figures (8) and (9) is mounted onto a rotating stage with an angular precision of about 0.1° or better. Next, a monochromatic light source with a relatively stable intensity is required. Lasers are commonly used as a light source because of the collimated, monochromatic beam that they provide (the high intensity of lasers is also a bonus). However, there is no need for the incident beam to be coherent, and thus any light source that is properly filtered to a narrow bandwidth and focused into a collimated (parallel) beam will suffice.

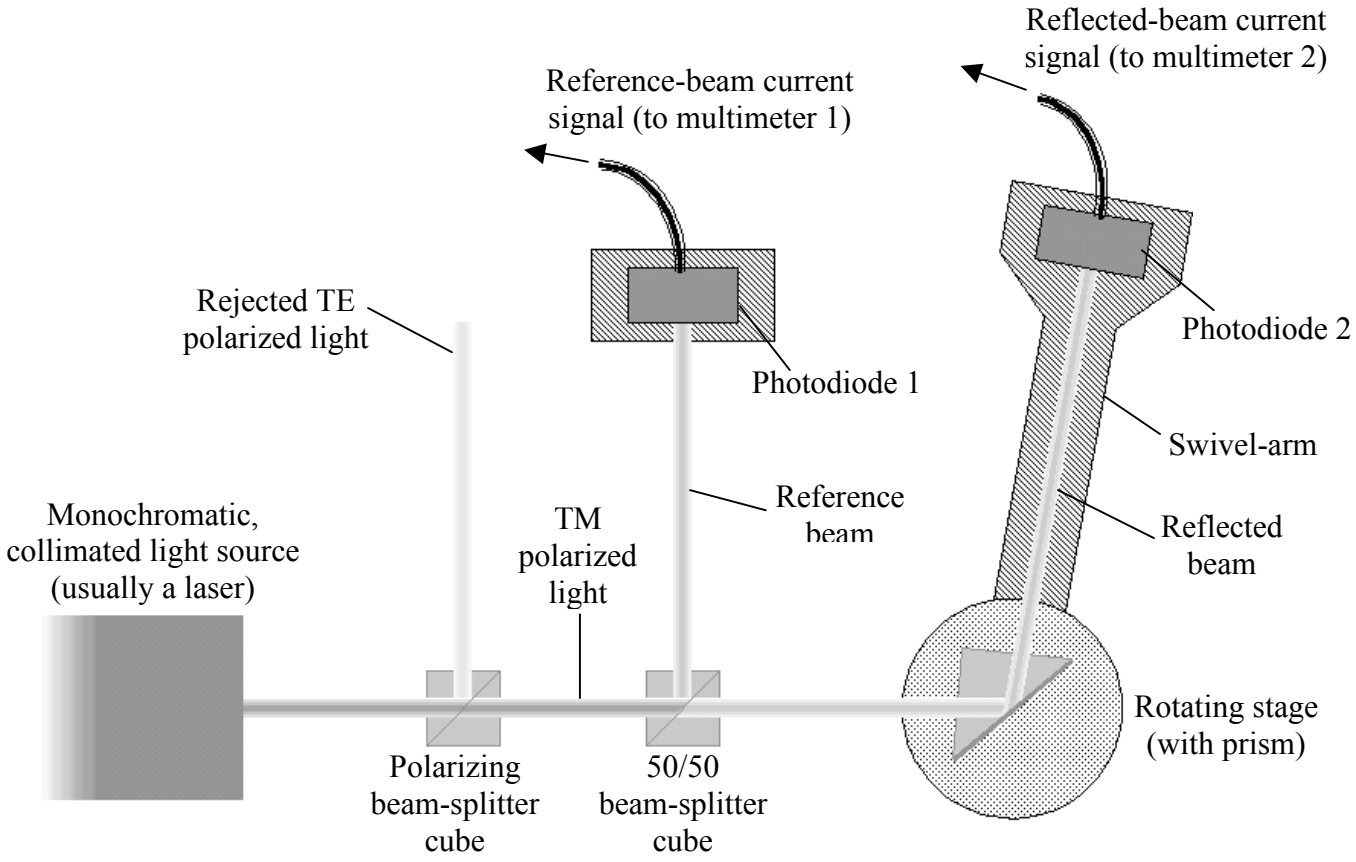
Before reaching the prism, the incident beam is directed to pass through two beam splitter cubes. The first is a polarizing beam-splitter which should be mounted such that only the light with a TM orientation is allowed to pass. If a laser is used as the source, it should be rotated to a TM orientation (approximately) in order to maximize the intensity of light that passes through

the polarizing beam-splitter. Next, the TM polarized light passes through a 50/50 beam-splitter to separate the beam into two equivalent beams. One of these two beams is directed immediately towards a photodiode to be recorded and used as a reference intensity. The second is directed towards the prism so that it can be used to sample the surface plasmon resonance characteristics of the metal-to-air interface.

A second photodiode is mounted on a swivel arm with the axis of rotation centered on the prism. In this way, as we adjust the angle of the prism (i.e. the angle of incidence), we can also adjust the location of the photodiode to pick up the reflected beam intensity.

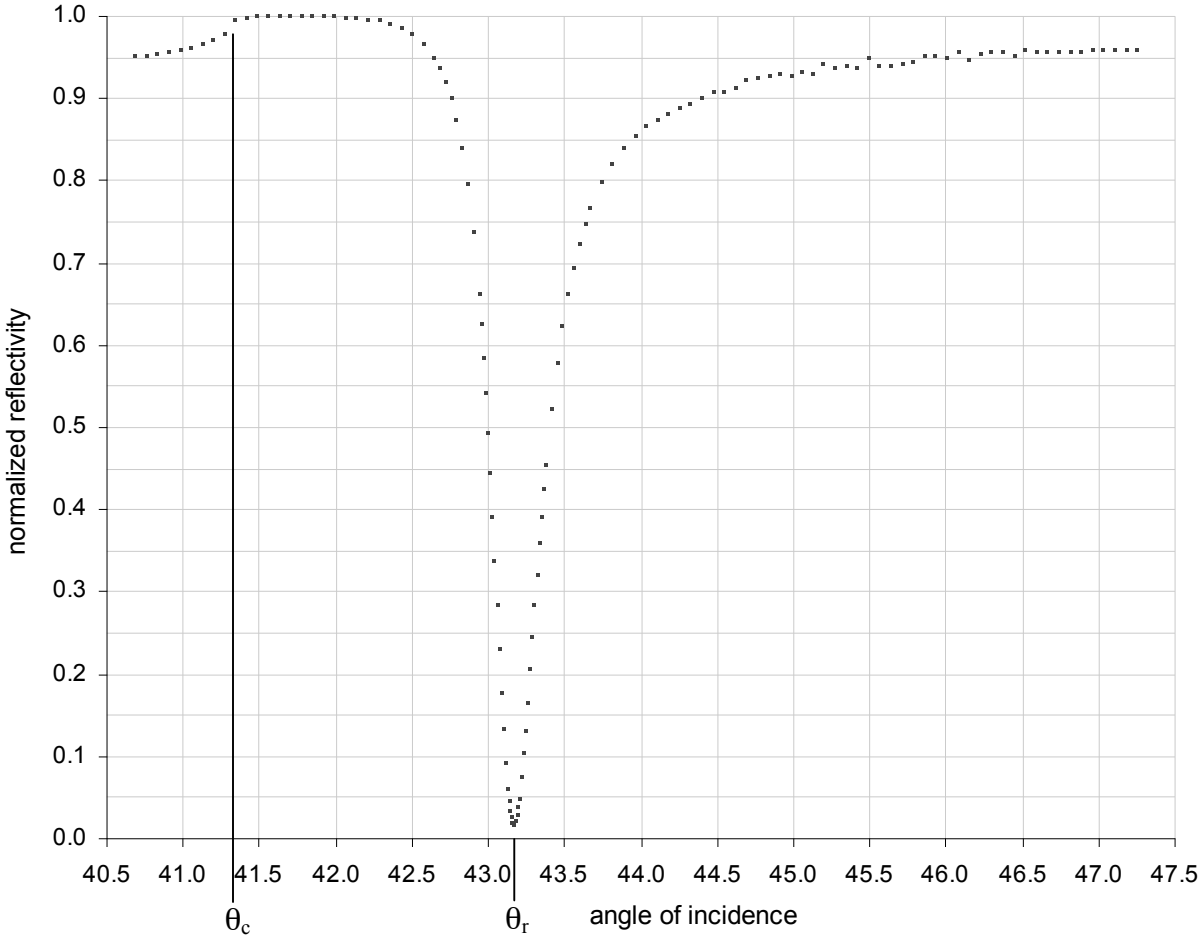
Finally, the reflected beam intensity is divided by the reference beam intensity (for each individual data point) to give a response that is independent of minor fluctuations within the source. This experimental setup is shown pictorially in Figure 10 below.

Figure 10: Basic angle-sweep SPR optical setup



The plot presented in Figure 11 is an example of the results that are typically attainable using the setup shown above. For this particular trial, a 656.5 nm diode laser was used as the light source. The right-angle glass prism had an evaporated layer of silver on the back surface with an approximate thickness of 50 nm.

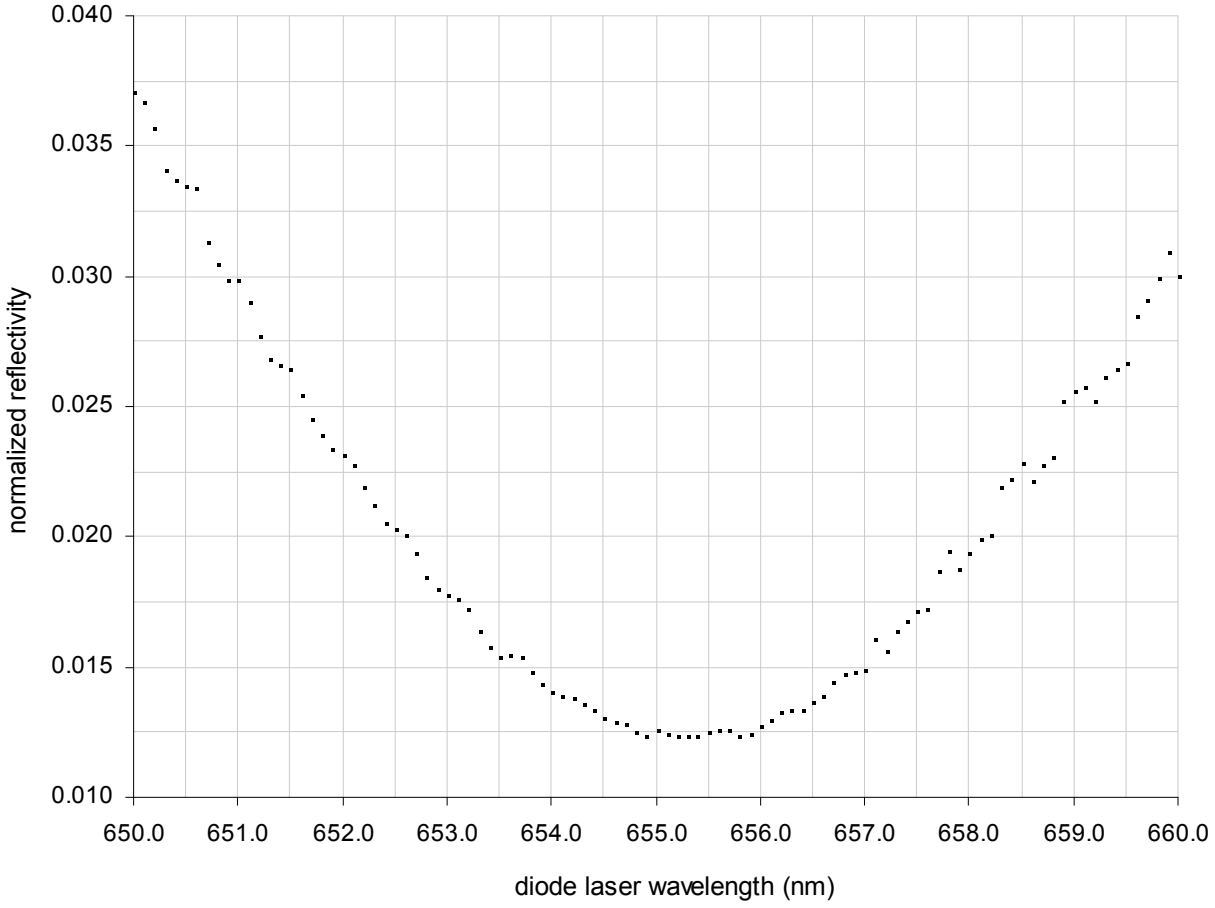
Figure 11: Plot of angle-sweep data (light source: diode laser at 656.5 nm)



These results clearly demonstrate the dramatic effect of surface plasmon resonance. At the resonance angle (which for this setup is approximately 43.2°) the reflectivity drops to about 1% of the maximum reflectivity. Thus, at this angle almost 99% of the incident light intensity is converted into surface plasmons. Also note the small hitch in the upper-left corner of the graph which signals the critical angle of incidence for this prism. In this case, the resonance angle is almost 2 degrees greater than the critical angle, which demonstrates the expected outcome that the conversion of light energy to surface plasmons occurs well within the region of total internal reflection.

The diode laser that was used in the above experiment also had the capability to be tuned over a wavelength range of 650-660 nm. This range is not nearly wide enough to display the full SPR response that is shown in Figure 11 (angle-sweep experiment). However, by fixing the incidence angle at about 43.2° (the resonance angle for the above setup), it was possible to obtain a zoomed-in view of the SPR *wavelength* response in the immediate vicinity of the minimum. A plot detailing this response is shown in Figure 12. For this particular plot, the identical setup was used as in the angle-sweep plot above (the only difference being that the incidence angle was held fixed while the diode laser was swept through a range of wavelengths).

Figure 12: Plot of wavelength-sweep data (light source: diode laser with variable wavelength)



There are a couple of things to note about this plot. First, this plot shows only the very bottom of the SPR minimum reflectivity (the range of reflectivity is from 1% to 4%). To sample the full range of the SPR wavelength response would require a span of more than 100 nm from the input light source. This greatly exceeds the capabilities of the diode laser that was used. Second, the obvious "raggedness" of this plot comes not from any physical effects inherent in the generation of surface plasmons, but rather from inconsistencies in the diode laser source. A close inspection of the graph shows that the bumpiness is actually periodic, which suggests some flaws in the tuning characteristics of the diode laser itself (a fact that will become important in later sections when discussing the wavelength modulation of this same diode laser).

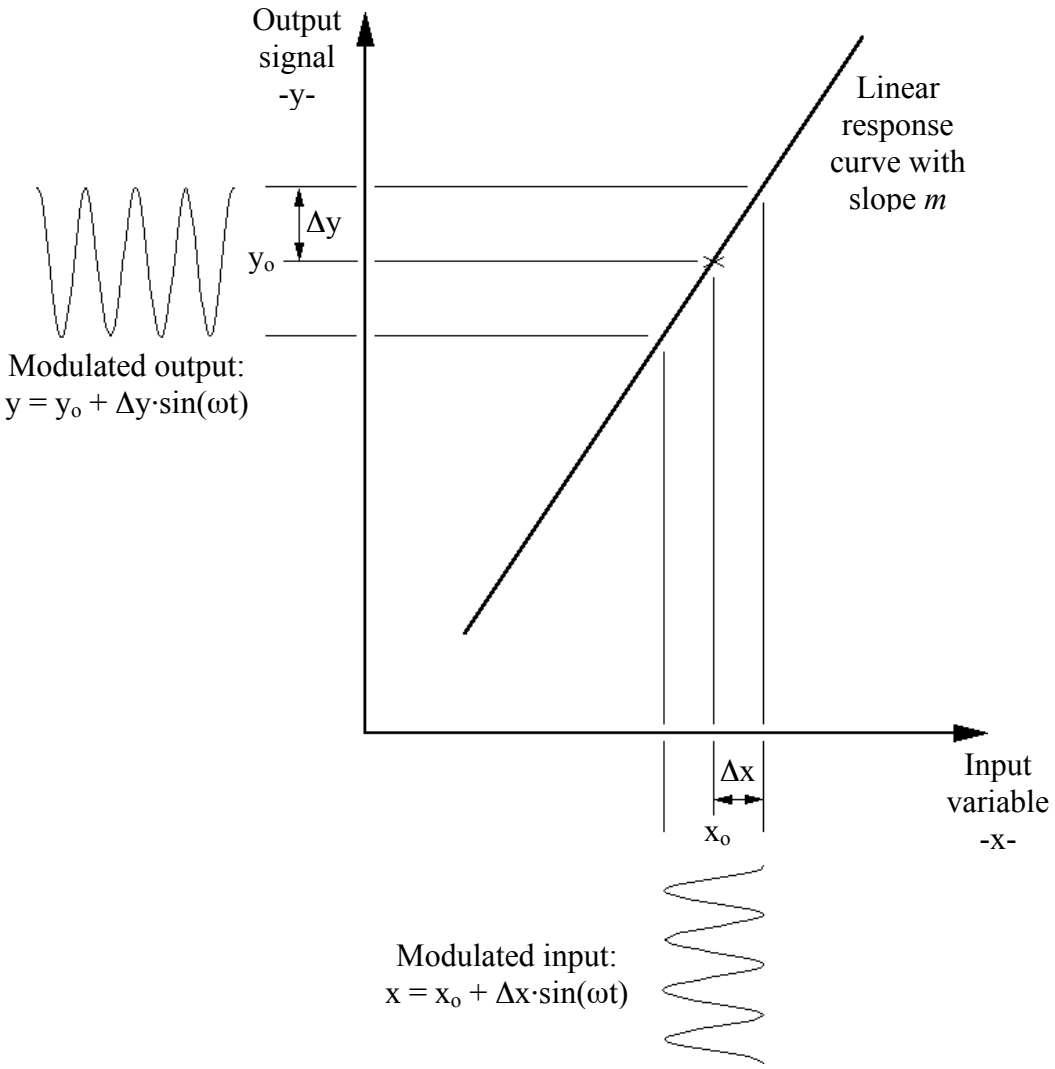
Despite this failure in accuracy, the diode laser wavelength response shows great merit in precision. For the range of 1% to 4% reflectivity in the angle-sweep data set, I was able to record only 8 data points, while pushing the precision of the rotating stage equipment (which included a micrometer fine-adjust) virtually to its limit: about 0.007° between adjacent data points. However, for the same range of 1% to 4% reflectivity in the wavelength-sweep data, I was able to easily record 100 data points while utilizing only one tenth of the attainable precision of the diode laser (data points were recorded at intervals of 0.1 nm. The diode laser fine adjust is 0.01 nm between steps). Thus, an extremely high-end rotating stage (angular precision ~0.00005°) would be required to rival the precision provided by the diode laser.

Section 4: Modulation Techniques Applied to SPR

For any experiment in which it is possible to sinusoidally modulate one of the input variables at a set frequency while holding all other variables fixed, a powerful technique becomes available. The technique is based on the fact that a modulation of the input variable will result in a modulated output signal which can be detected quite easily with a lock-in amplifier. One advantage of this system is that for a modulation frequency of say 100 Hz, the lock-in amplifier is effectively averaging over about 100 data points per second to produce a much more precise final measurement. The 1/f noise is also significantly reduced at higher frequencies. I will discuss how a lock-in amplifier accomplishes this task in the following pages.

First, imagine an output signal (y) that is *linearly* dependant upon some given input variable (x). By sinusoidally modulating the input variable about a point x_o at a frequency ω and with an amplitude Δx, we should see a similar modulation of the *output signal* about a point y_o at the same frequency ω but with an amplitude Δy. See Figure 13 below:

Figure 13: First derivative input modulation



By measuring the amplitude of the modulated output signal (Δy) for a given input modulation amplitude (Δx), we can easily determine the slope of the response curve (which is given by $m = \Delta y/\Delta x$). This technique can even be used for a non-linear response curve provided that we make Δx small enough so that the local response can be reasonably approximated by a straight line. However, this also means that Δy will likely be quite small in comparison to the mean output signal, y_o. Thus we must use a lock-in amplifier to accurately measure the amplitude Δy (otherwise the modulated part of the signal will be lost in the noise).

Digital lock-in amplifiers operate according to the following system[8]: First, a reference signal that oscillates with the same frequency as the input modulation is sent to the lock-in amplifier. The lock-in creates an amplified sinusoidal voltage wave which is timed to the reference signal. Then, when the actual output signal (y) is received, the lock-in multiplies it by this amplified reference wave. The result of this multiplication can be understood by analyzing the case in which two pure sinusoidal functions are multiplied together, given as follows:

$$V_r \sin[\omega_r t + \theta_r] \cdot V_s \sin[\omega_s t + \theta_s] = \tag{45}$$

$$\tfrac{1}{2} V_r V_s \cos[(\omega_r - \omega_s)t + \theta_r - \theta_s] \; - \; \tfrac{1}{2} V_r V_s \cos[(\omega_r + \omega_s)t + \theta_r + \theta_s]$$

The term with angular frequency ($\omega_r - \omega_s$) is the important one here because its frequency drops to zero when $\omega_r = \omega_s$, which results in a non-oscillating term. Thus, the portion of the output signal that is oscillating at the same frequency as the reference signal (which is exactly the part we are hoping to measure) gives a non-oscillating (DC) term when the two are multiplied together. All other portions of the output (noise, DC offset, etc...) give AC signals which can be easily removed with a low-pass filter. After filtering, the net result is a DC signal of the value:

$$V_1 = \tfrac{1}{2} V_r V_s \cos[\theta_r - \theta_s] \tag{46}$$

There remains one slight problem in that this DC voltage is still dependant on the phase difference between the reference and output signals. However, this problem is easily overcome by performing the multiplication described above a second time, except this time with a reference signal that is 90° out of phase from the first. The result is a second DC signal given by:

$$V_2 = \tfrac{1}{2} V_r V_s \sin[\theta_r - \theta_s] \tag{47}$$

The magnitude can then be determined by performing the following manipulation upon V_1 and V_2:

$$\sqrt{V_1^2 + V_2^2} = \tfrac{1}{2} V_r V_s \tag{48}$$

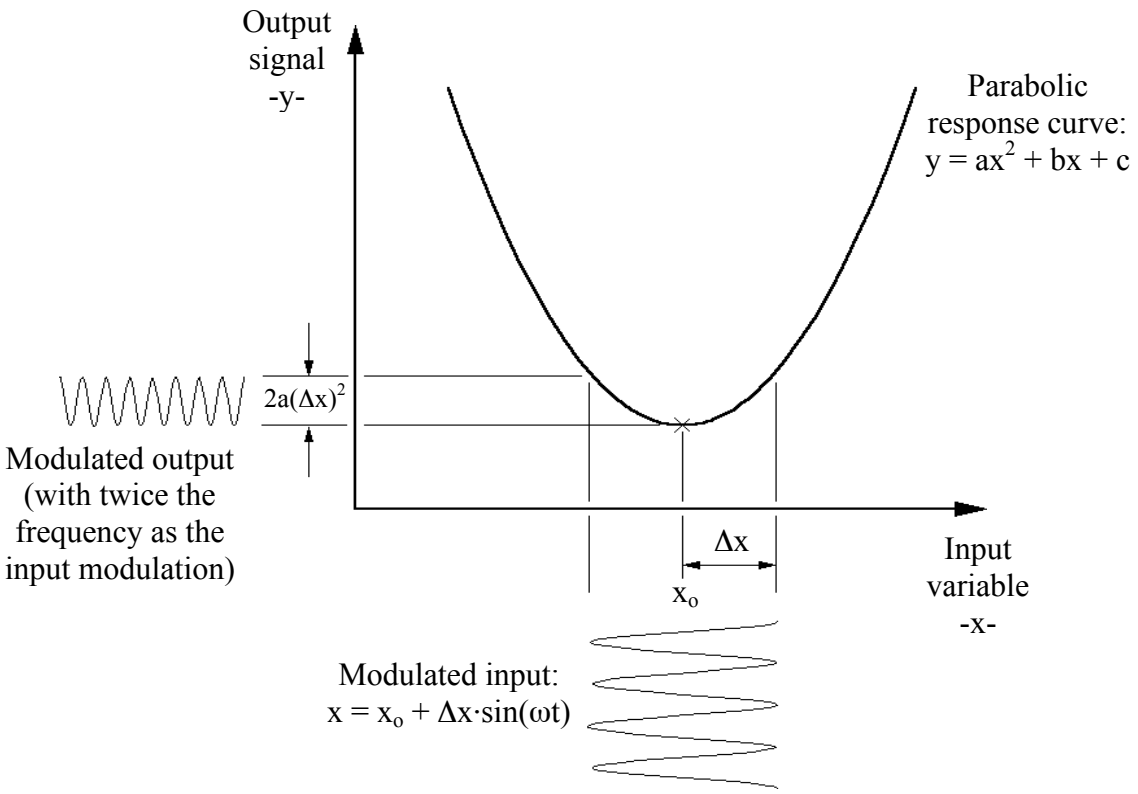
Multiplying by two and dividing by V_r will give the actual amplitude of the modulated signal (V_s).

Now, if we imagine a *parabolic* response curve (which happens to be a decent approximation of the SPR response near the minimum), the analysis is only a little bit more complicated. We start by defining the response as a standard parabola: $y = ax^2 + bx + c$. Then, we plug in the same sinusoidal input modulation as before ($x = x_o + \Delta x \cdot \sin(\omega t)$) which gives the following expanded result:

$$y = [ax_0^2 + bx_0 + c] + [(2ax_0\Delta x + b\Delta x)(\sin(\omega t))] + [a(\Delta x)^2(\sin^2(\omega t))] \quad (49)$$

The first term ($ax_o^2 + bx_o + c$) is simply the mean (un-modulated) response. The second term is the modulated response based upon the slope of the curve (note that this time, the slope changes linearly with x_o as would be expected for a parabola). The third term, however is something new. This term is considered to be second-order because the square of a sinusoid is another sinusoid with *twice* the frequency. The significance of the third term is that it's magnitude is dependant only on Δx and *a*. Thus, for a fixed input modulation amplitude (Δx), it can be easily used to determine the concavity (second derivative) of the response: *2a*. This concept is demonstrated graphically in Figure 14 below:

Figure 14: Second derivative input modulation



By imagining the input variable swinging back and forth near to the bottom of the curve, we can see how the resulting output oscillation should have twice the frequency: When the input is to the far right, the output is a maximum. When the input is at the midpoint (in this diagram

the bottom of the well), the output is a minimum. When the input is to the far left, the output is back to a maximum again. Thus, a half cycle of the input translates into a full cycle for the output. As predicted by equation (49), this second order oscillation will be present throughout the response curve. In Figure 14, I have just depicted the oscillation at the bottom of the well because it is easier to visualize when the slope is zero (the first order term goes away at this point).

We can once again measure the amplitude of this second-order oscillation (in a manner very similar to that used in measuring the first) with the lock-in amplifier. The only difference is that in this case, the lock-in must be programmed to multiply the output signal by the *second-order harmonic* of the reference wave. Thus, oscillations at all other frequencies (including the first-order modulation) can be easily filtered out, leaving only a DC voltage that is proportional to the amplitude of the second-order modulation.

As stated previously, the amplitude of the second-order modulation is a valuable characteristic to measure because it tells us the concavity of the response curve centered at the point x_o. The perfect parabola depicted in Figure 14 should actually yield a pretty boring outcome in regards to second-order modulation, since the concavity of this curve is by definition constant. However, for a response curve that changes its behavior rapidly with respect to some given variable (such as the reflectivity near to the surface plasmon resonance point), the concavity can be an enlightening characteristic to measure. In particular, the inflection points (of which the SPR reflectivity response has two) are defining characteristics since it is at these points that the concavity equals zero. For this same reason, measurement of the *first-order modulation amplitude* becomes most relevant at the very bottom of the SPR minimum reflectivity, since it is at this point that the *slope* equals zero.

As a side note, it turns out that by setting the lock-in amplifier to the third, fourth, fifth, and higher order harmonics of the reference wave, we can measure the third, fourth, fifth, etc... derivatives of the response curve (for a generic mathematical demonstration of this, see Dharamsi[9]). However, for the experiments I performed (see section 5) the higher order harmonics were more susceptible to noise and thus it was impractical to attempt to measure them.

The technique of input variable modulation has already been used by Albrecht, et al.[10] to obtain greater sensitivity with the angle-sweep method of SPR characterization. In order to accurately modulate the angle of incidence, they used a mirror mounted on a piezo-motor to redirect the collimated light from the source. This was subsequently mounted upon a rotating stage controlled by a stepper motor to regulate the *mean* value of the angle of incidence. A series of two lenses were placed along the optic axis to direct the light consistently to the same point on the prism (regardless of incidence angle). A final lens was used to direct the reflected light to the photosensitive detector (regardless of reflection angle). See Figure 15.

In their published article, Albrecht, et al.[10] state that their setup is able to detect shifts in the surface plasmon resonance angle of less than 0.001°. This advertised sensitivity of the angle-modulation method is a full order of magnitude better than the sensitivity attainable by the simple angle-sweep SPR setup that I used to record the data presented in Figure 11. Thus, it was hoped that performing a *wavelength-modulation* experiment might yield similar improvements upon the simple wavelength-sweep method (which, as stated previously, already demonstrates some advantages over the simple angle-sweep method). The next section describes the variety of techniques that I devised and tested to utilize wavelength modulation for the characterization of the SPR response.

Figure 15: Schematic of angle-modulation setup (from Albrecht et. al.[10])

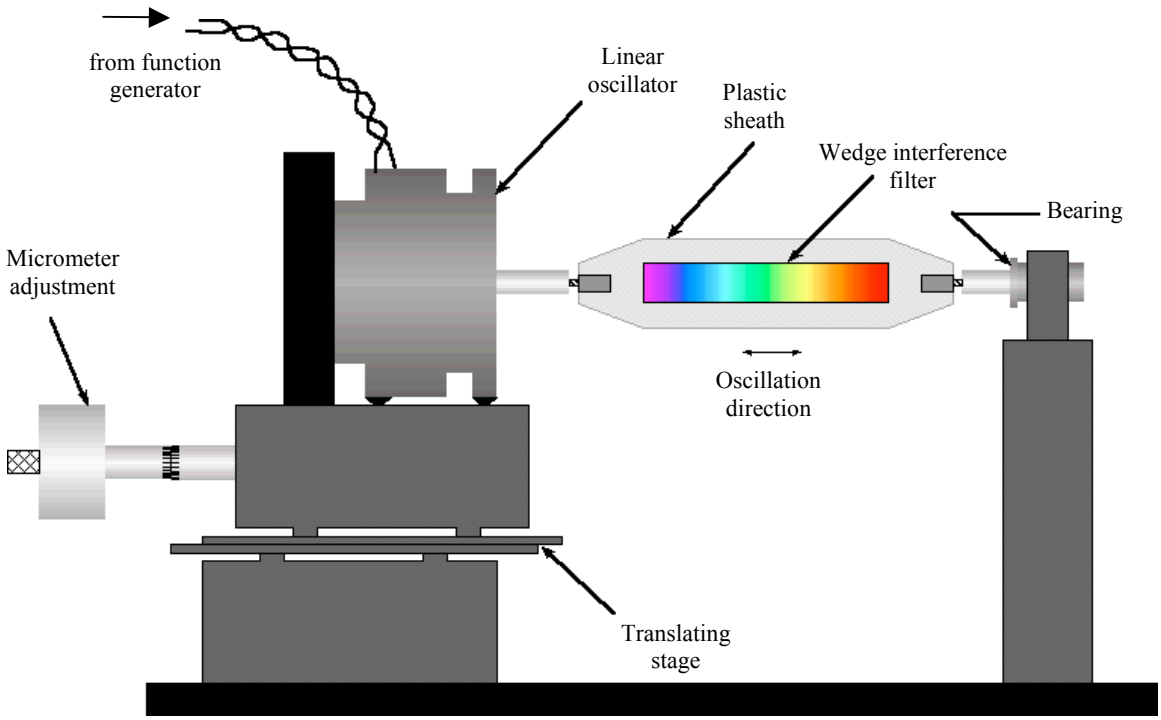
Section 5: Wavelength-Modulation Techniques

Wavelength-modulated SPR can be performed using a variety of different techniques. I decided to build three different setups that would test and compare the merit of three of these techniques. The three setups that I built can be easily described as wide-bandwidth, medium-bandwidth, and narrow-bandwidth (referring to the bandwidth and the modulation amplitude of the incident light).

5-1. Wide-Bandwidth Setup

The wide-bandwidth setup involved a linearly-variable wedge interference filter which could be slid back and forth within the path of light to obtain the desired wavelength. With the filter mounted inside a specially built plastic sheath, I attached one end to a linear oscillator (which provided the modulation) and suspended the other end by a linear bearing. The linear oscillator was placed on a translating stage (controlled by a micrometer) to allow the ability to sweep through a range of wavelengths. See Figure 16.

Figure 16: Wide-bandwidth wavelength modulation apparatus (side view)



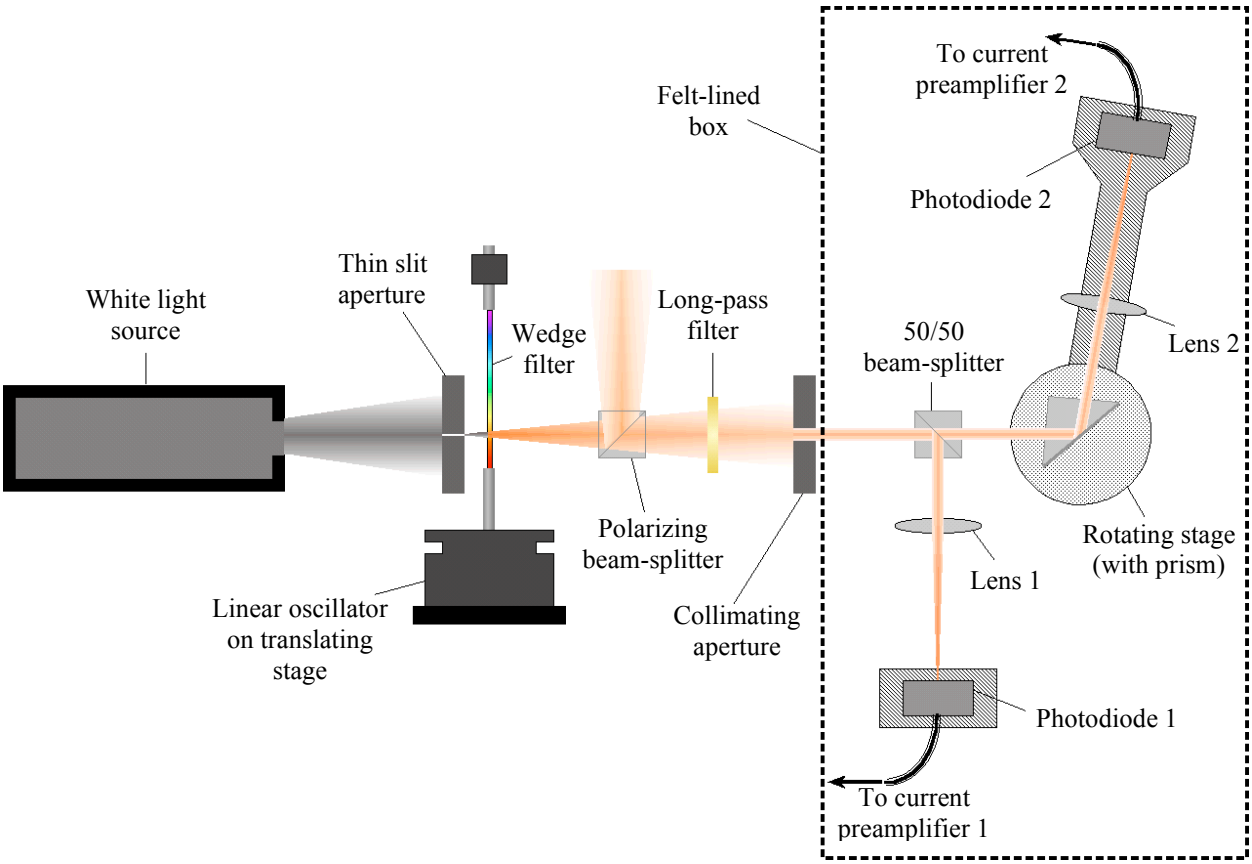
The wedge interference filter I used was about 3 inches long with a linear rate of about 6 nm (passed wavelength) per mm along the length of the filter. By placing a narrow slit aperture immediately in front of the wedge filter, it is possible to select a specific wavelength of light. The minimum possible bandwidth of light passed by the filter is about 10-15 nm (regardless of how narrow a slit aperture is used). It is for this reason that I refer to this as the

wide-bandwidth setup. While oscillating at a frequency of 100 Hz, the linear oscillator was capable of producing a maximum peak-to-peak amplitude of about 1 mm, which translates to a wavelength modulation amplitude of about 6 nm peak to peak. An oscillation frequency of 150 Hz produced about half this amplitude, or 3 nm peak to peak. Since the exact modulation amplitudes were not measured to any better precision than this, it was critical to maintain the linear oscillator at a consistent modulation frequency and amplitude throughout any given experiment.

For the experiments that I performed, the passed wavelength was swept through a range from 530 nm to 640 nm by adjustment of the translating stage (all the while the passed wavelength was being modulated by the linear oscillator ever so slightly about the particular selected value). The position of the translating stage was controlled by a metric micrometer with a linear precision of 0.01 mm. This corresponds to a precision of 0.06 nm for the central wavelength of the passed band of light. The precise conversion of linear position to passed wavelength was calibrated using two He-Ne lasers (one at 543.5 nm and the other at 632.8 nm). The dependence of passed wavelength to filter position was assumed to be exactly linear.

Near the red end of the spectrum, the wedge interference filter that I used happens to let through a fair amount of violet and ultraviolet light as a second order bandpass. Thus, for the duration of these experiments I used a 500 nm long pass filter to ensure that only light of the

Figure 17: Wide-bandwidth complete optical arrangement

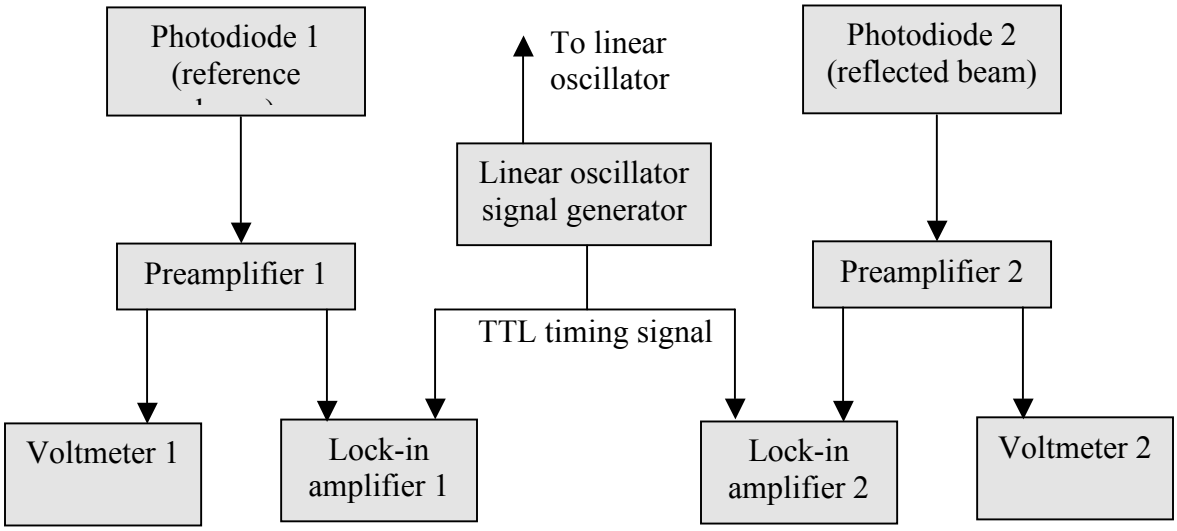


primary bandpass wavelength was allowed through. A second aperture was also placed immediately following the long pass filter to ensure that the remaining beam was very nearly collimated (this is to diminish the range of incidence angles, which would introduce an undesirable extra variable). The photodiode detectors, beam-splitters, and prism used in this experiment were the same as those used in the standard angle-sweep SPR experiment illustrated in Figure 10.

The light source was a 240 watt quartz-halogen bulb. Although the high temperature of this source provides a moderately flat spectral response over the noted range (530-640 nm), the use of a reference detector is critical (even more so than in the angle-sweep SPR experiment). Despite the high power of the 240 watt bulb, the final light intensity that reached the photodiodes was quite small (the two filters, two apertures, and two beam-splitters managed to claim the vast majority of light). Thus, lenses were used to focus the remaining light onto each detector and a felt-lined box was erected to prevent the signal from being washed out by residual light. Even with these precautions, high-gain current preamplifiers were needed to boost the current signal from each photodiode to within a measurable range. The complete optical setup for the wide-bandwidth setup is shown in Figure 17.

The reflected and reference signals were processed separately, each going to its own preamplifier to be converted to a voltage. Each of the two voltages was then sent separately to its own voltmeter and lock-in amplifier as shown in Figure 18.

Figure 18: Electronic components for wavelength modulation setup



Normalization of the SPR reflectivity was a simple case of dividing the reflected signal (from voltmeter 2) by the reference signal (from voltmeter 1). In the wavelength-sweep experiment this turned out to be necessary because the light source, wedge filter, and photodiodes all had a strong systematic dependence on the wavelength of light. For this same reason, normalization of the *slope* of the SPR reflectivity (from the lock-in amplifiers) was equally important. However, accomplishing this effectively required significantly more manipulation, as I will describe next.

We can imagine the reference signal as having some well-defined dependence upon wavelength, f(λ), which incorporates the wavelength dependence of the light source, wedge filter, and photodiode. We can also imagine an ideal SPR reflectivity (for the specific prism, dielectric and metallic materials, metal thickness, and angle of incidence) given as a function of wavelength, g(λ). This SPR reflectivity, g(λ), should constitute the only difference between the reflected and reference beams (excepting a constant intensity correction which is very close to one for an ideal 50/50 beam splitter). Therefore, the reflected signal, h(λ) should simply be the product of the reference signal and the ideal SPR reflectivity,

$$h(\lambda) = f(\lambda)g(\lambda). \tag{50}$$

This, of course, explains why the ideal SPR reflectivity, g(λ), can be obtained from dividing the reflected signal, h(λ), by the reference signal, f(λ). It also gives an idea as to how the *slope* of the ideal SPR reflectivity, g′(λ), should be obtained. Taking the first derivative of equation (50) gives

$$h'(\lambda) = f'(\lambda)g(\lambda) + f(\lambda)g'(\lambda). \tag{51}$$

Then, solving equation (51) for g′(λ) leaves the following expression:

$$g'(\lambda) = \frac{h'(\lambda) - f'(\lambda)g(\lambda)}{f(\lambda)}. \tag{52}$$

In this expression, f ′(λ) represents the slope of the reference signal, while h′(λ) represents the slope of the reflected signal (which are recorded as voltages by lock-in amplifiers 1 and 2, respectively). Thus, the slope of the reflectivity can be obtained by multiplying the normalized reflectivity, g(λ), with the slope of the reference signal (lock-in 1), subtracting this result from the slope of the reflected signal (lock-in 2), and then dividing the whole thing by the reference signal, f(λ).

This method has a couple of minor problems that can be solved with some manipulation. First, the lock-in amplifiers significantly boost the AC portion of the signal. This means that the reference and reflected signal voltages (f(λ) and h(λ)) as recorded by the voltmeters will incorporate some large scaling factor in comparison to their respective slopes (f ′(λ) and h′(λ)) as recorded by the lock-in amplifiers. Thus, the step in equation (52) in which the whole right side is divided by f(λ) introduces a significant (but constant) scaling factor. This scaling factor can be determined by numerically integrating the uncorrected g′(λ) and then adjusting the multiplication factor until a best fit with g(λ) is obtained. However, as long as one is only trying to obtain the form of the slope, it is not really necessary to determine the value of this scaling factor. In my particular case, it did not seem necessary and thus the several plots of the reflectivity slope that you will see on the following pages remain uncorrected. As a side note, it turns out that the step in equation (54) in which f ′(λ) and g(λ) are multiplied together is actually not problematic because g(λ) is the *normalized* SPR reflectivity (h(λ)/f(λ)) and thus no scaling factors are introduced (g(λ) is unitless).

The second problem is a phase difference between the AC voltages recorded by the two lock-in amplifiers. In equation (52), the step in which f ′(λ)g(λ) is subtracted from h′(λ) is not

legitimate unless the reference lock-in voltage (f′(λ)) and the reflected lock-in voltage (h′(λ)) have the same phase. This phase difference can be eliminated by recording both R and θ from each lock-in (or X and Y) and then separately correcting the phase angle for both the reference and reflected data per the following geometric equations:

$$X = R\cos(\theta), \tag{53}$$

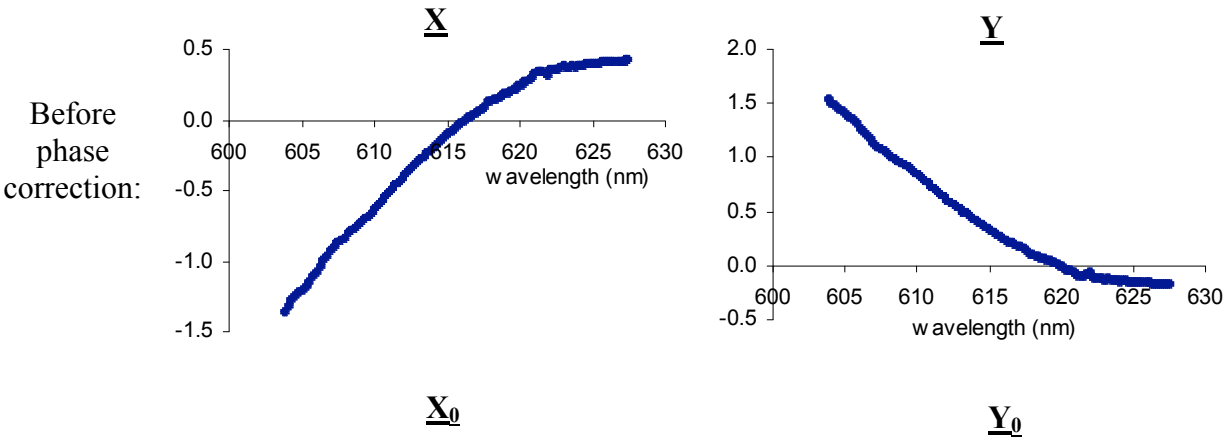
$$Y = R\sin(\theta), \tag{54}$$

$$R = X^2 + Y^2, \tag{55}$$

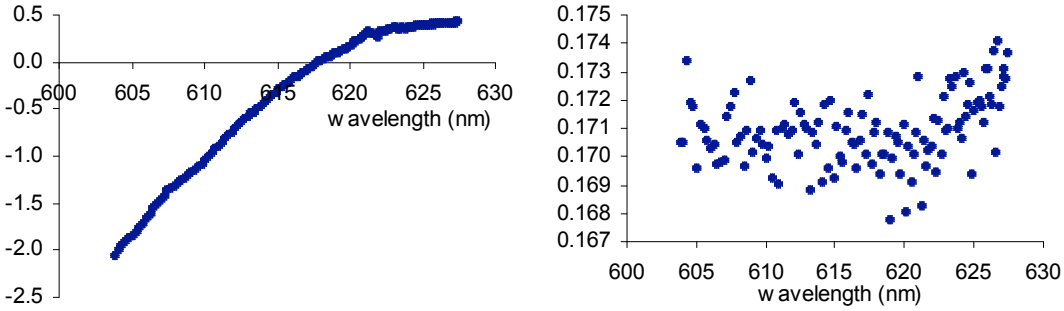
$$\theta = \arctan\left(\frac{Y}{X}\right). \tag{56}$$

Using equations (53) through (56), it is possible to convert to phase corrected components X_0 and Y_0, and then plot both with respect to wavelength. It turns out that the component that is *in-phase* with the wavelength oscillation (corresponding to f′(λ) or h′(λ)) is usually systematic and large, while the *out-of-phase* component is random and relatively small. Thus, the corrected phase angle is adjusted until the plot of Y_0 appears most random. When this occurs, X_0 becomes perfectly in-phase with the wavelength modulation. The resulting X_0's from both the reference and reflected lock-in amplifier data can then be legitimately compared with each other. Plots detailing this process are shown in Figure 19.

Figure 19: Phase correction of slope data



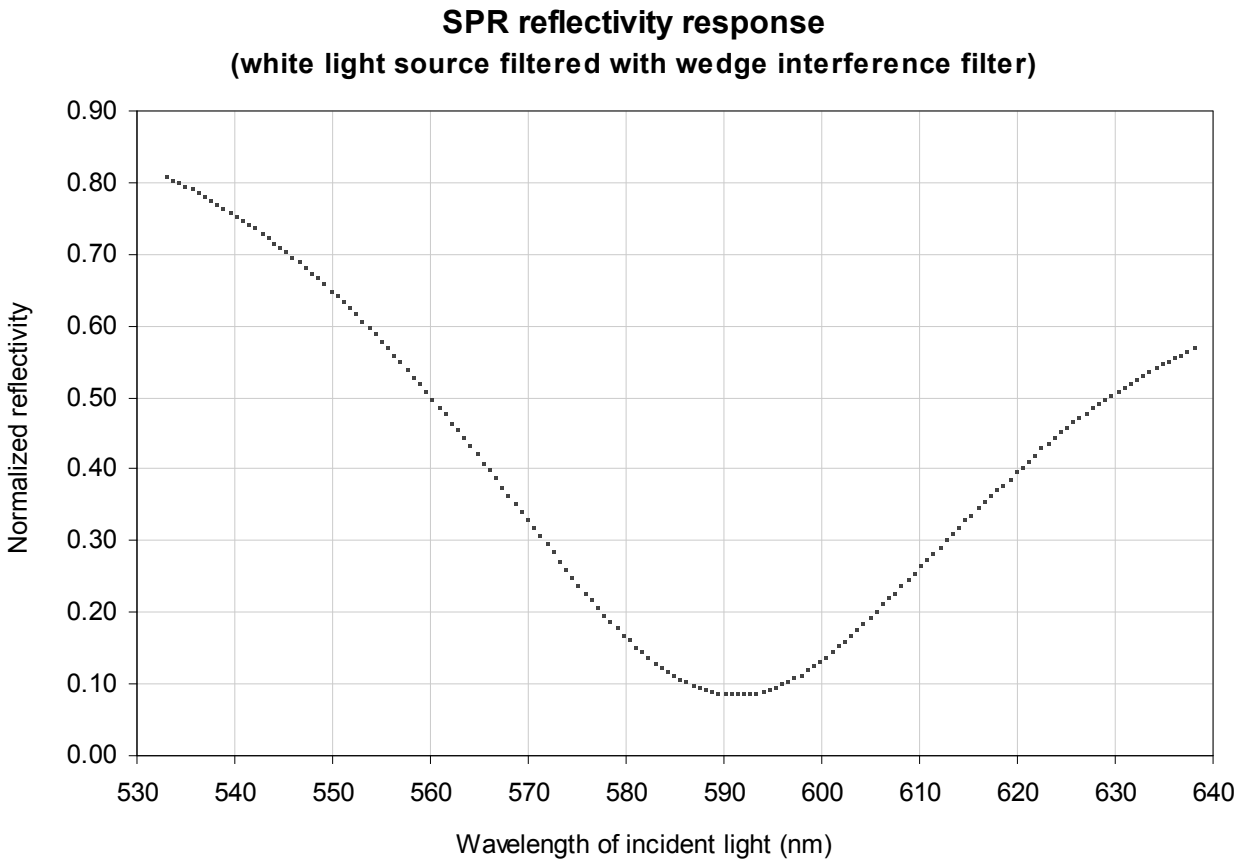
After phase correction:



For the setup I built, the phase difference between the AC components of the reflected and reference signals turned out to be about 3° (perhaps because of a slight difference between the two photodiodes). This phase difference is really not large enough to create any significant errors, however, to be rigorous I performed the phase correction described above on all of the data included in this report.

The first experiment that I performed with the wide-bandwidth setup was a full range sweep (530 nm to 640 nm) to obtain both the SPR reflectivity as a function of input wavelength and the slope of this SPR reflectivity response (with the lock-in amplifier set to the reference frequency). The concavity of the SPR reflectivity response was not attainable because the signal was too noisy when the lock-in was set to the second-order harmonic. For this sweep, the linear oscillator was operating at 100 Hz with maximum amplitude. The prism was fixed such that the angle of incident light was 43.74°. Figure 20 shows the normalized SPR reflectivity measured in this experiment.

Figure 20: Wide Bandwidth – Experiment 1 (graph 1)



As shown by this graph, the sweep technique using the wedge interference filter is capable of recording nearly the entire span of the SPR wavelength dependence (in contrast to the diode laser which has a span of only 10 nm; see Figure 12). On the next page are graphs of the slope of the SPR reflectivity (Figure 21) and a combination of the SPR reflectivity and slope (Figure 22). The slope data points were calculated using the method described on the preceding pages. Figures 20, 21, and 22 all refer to the same set of data.

Figure 21: Wide Bandwidth – Experiment 1 (graph 2)

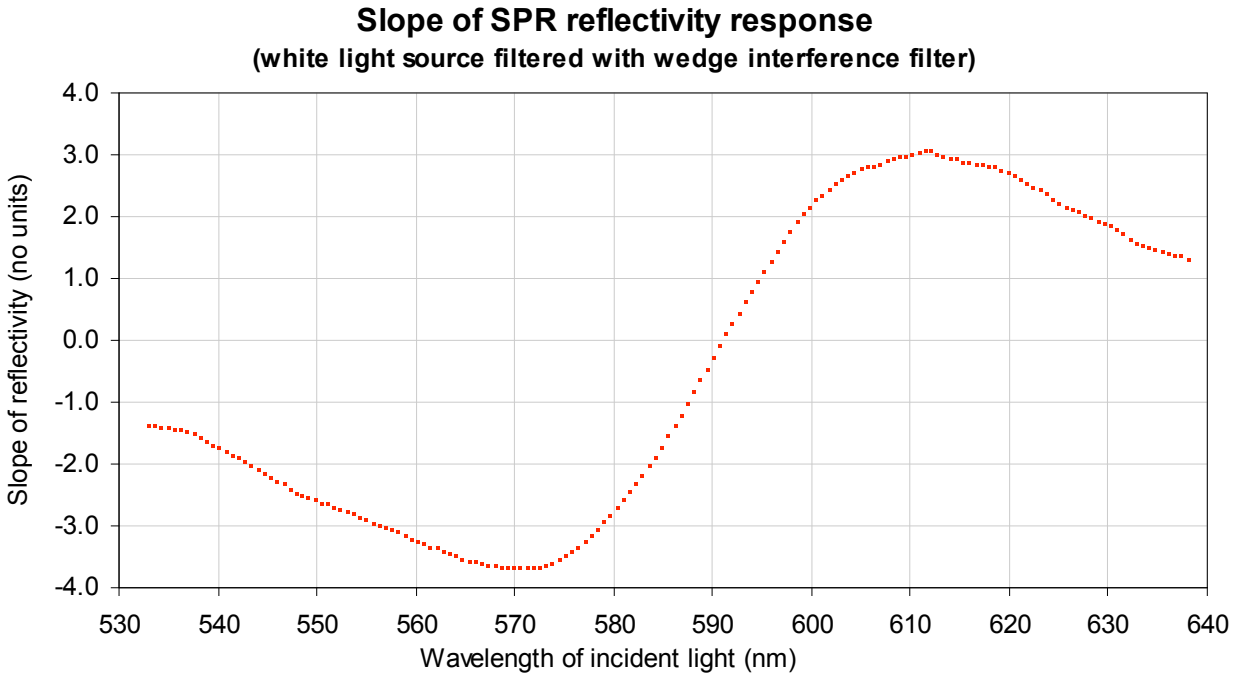
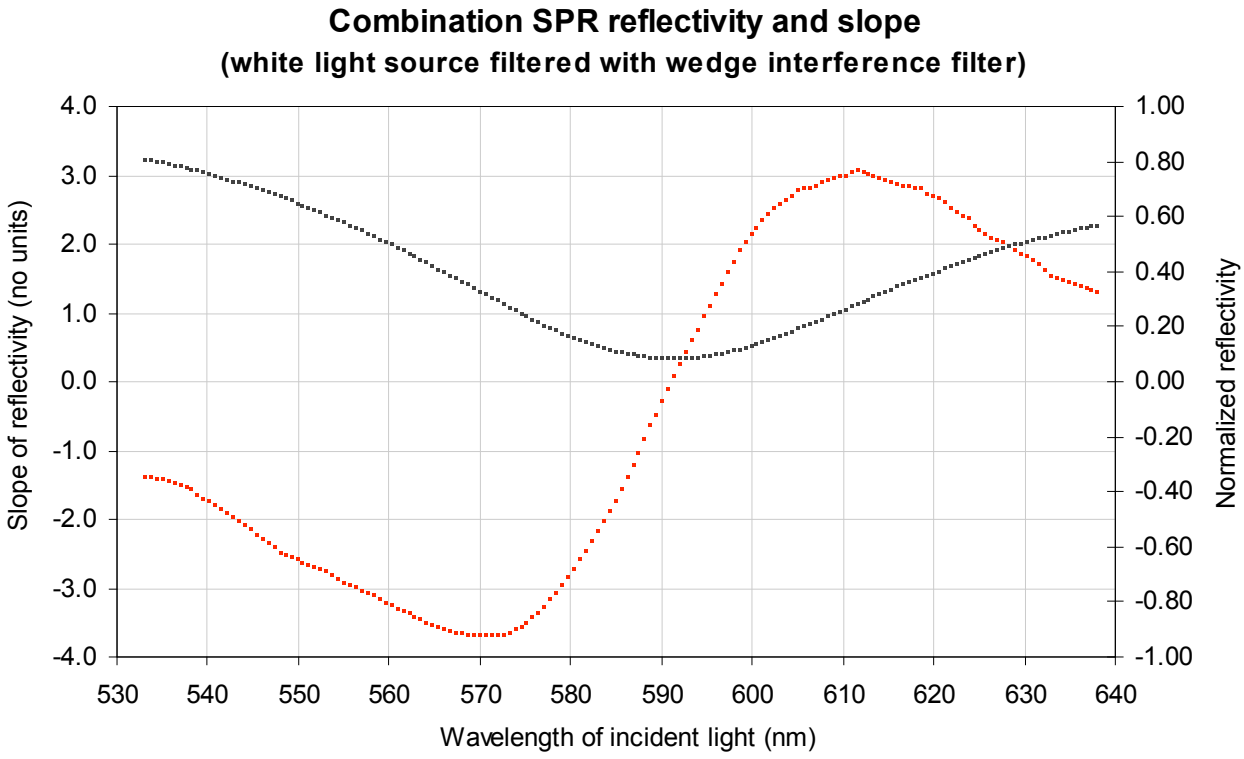


Figure 22: Wide Bandwidth – Experiment 1 (graph 3)



The graphs shown in Figures 20, 21, and 22 have a small amount of bumpiness which is a result of the error prone nature of the low intensity beams and highly amplified signal. Despite this, the data from the high bandwidth setup is at least as clean as that of the diode laser.

In contrast to the diode laser sweep, note that the reflectivity shown in Figure 20 reaches a minimum of about 10% (the diode laser sweep reached a minimum reflectivity of about 1%; see Figure 12). There are two reasons for this significantly increased minimum. First, the bandwidth of incident light is at least 10 nm, which results in a slight flattening of the reflectivity curve. Second, there is a range of incident angles due to the difficulty involved with collimating light from an incandescent source. Since the intensity of the incident light was already quite weak, I was forced to allow a somewhat broad beam of light to reach the prism which further allowed for this imperfect collimation. The beam was 5 mm wide and 1.5 m long, corresponding to an angular range of ~0.2°. Similar to the effect caused by the range of wavelengths, this moderate range of incidence angles has the effect of flattening the reflectivity curve.

As discussed previously, the minimum spacing between data points for the wide-bandwidth setup was 0.06 nm (which is limited by the precision of the translating stage). For the data set shown in Figures 20, 21, and 22, the separate data points were recorded at intervals of 0.6 nm (10 times the minimum possible spacing allowed by this setup). However, in the second experiment (see below), the sensitivity of the wide-bandwidth setup was meticulously tested with data points recorded at the minimum possible interval of 0.06 nm.

Before discussing this experiment though, a quick preface: As shall be discussed in the final section of this report, the practical merit of any technique for examining surface plasmon resonance is its ability to detect very tiny shifts in the SPR reflectivity curve. Normally, such shifts are the result of some minor modification to the surface itself, such as the binding of a thin layer of molecules. However, I did not have any practical means of modifying the prism surface, so I chose instead to impose a very slight shift in the incidence angle. It was hoped that this might provide some reliable assessment of the sensitivity of the wavelength-modulation technique in comparison to the incidence angle sweep technique.

Figures 23 and 24 display the results of the second wide-bandwidth experiment, which involved a much narrower sweep of the SPR reflectivity minimum. Actually a total of three sweeps were performed at three different incidence angles. The first sweep was performed at an approximate incidence angle of 43.24°. For the second sweep, the incidence angle was shifted in the positive direction by 0.007° (which is the minimum quantifiable shift allowed by the rotating stage upon which the prism was mounted). For the third sweep, the incidence angle was shifted by an additional 0.014° (for a 0.021° total shift from the initial setting). In this way, the reflectivity and slope could be compared for incidence angle shifts of 0.007°, 0.014°, and 0.021°. As mentioned previously, data points for this second experiment were recorded at the minimum possible interval of 0.06 nm.

For the tiny incidence angle shift of 0.007°, the SPR reflectivity curve (shown in Figure 23) exhibits a significant wavelength shift (about 0.6 nm, which is equivalent to the spacing between ten data points). The ability to accurately measure this shift is impeded by the conspicuous noisiness of the reflectivity data (the drastically "zoomed-in" view of the reflectivity minimum allows the noise inherent in this system to become much more noticeable). However, the slope of the reflectivity (as measured by the lock-in amplifiers) is much less susceptible to this noise and thus it is much easier to gauge the actual value of the reflectivity shift (see Figure 24). In addition, curve fitting is made a simple matter by the fact that the slope of the reflectivity is almost linear near to the reflectivity minimum. Assuming that it would be possible to observe

Figure 23: Wide Bandwidth – Experiment 2 (graph 1)

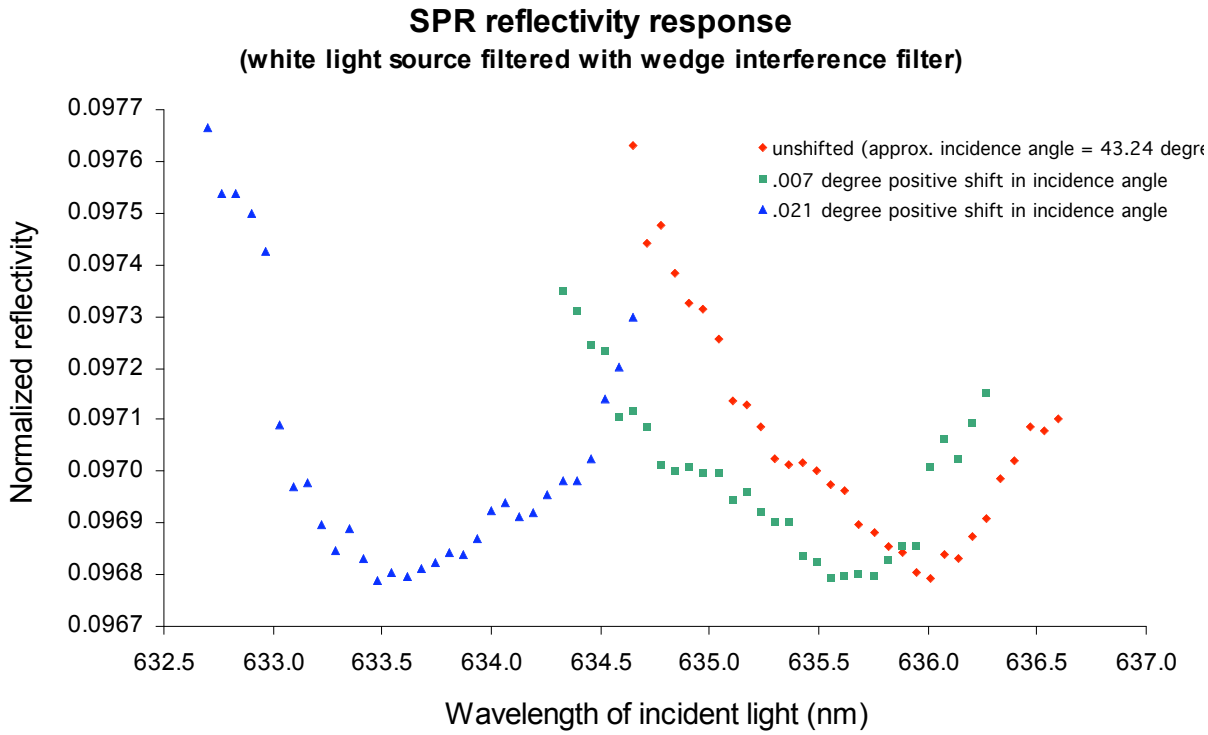
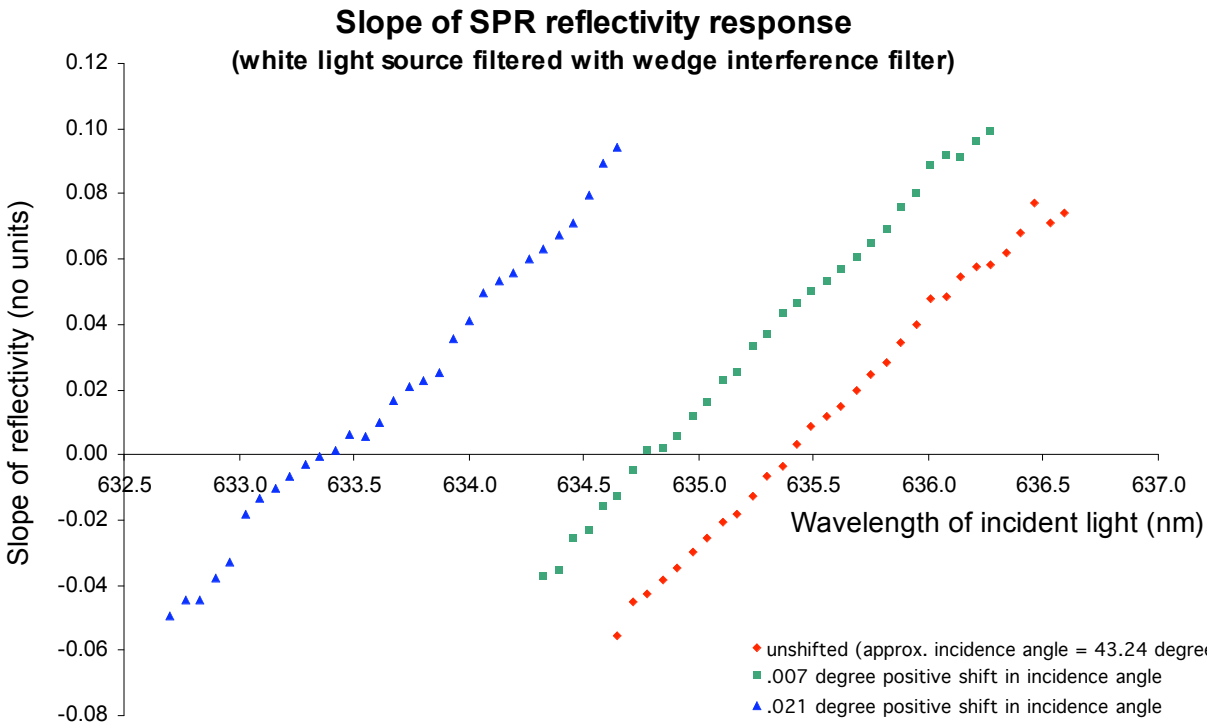


Figure 24: Wide Bandwidth – Experiment 2 (graph 2)



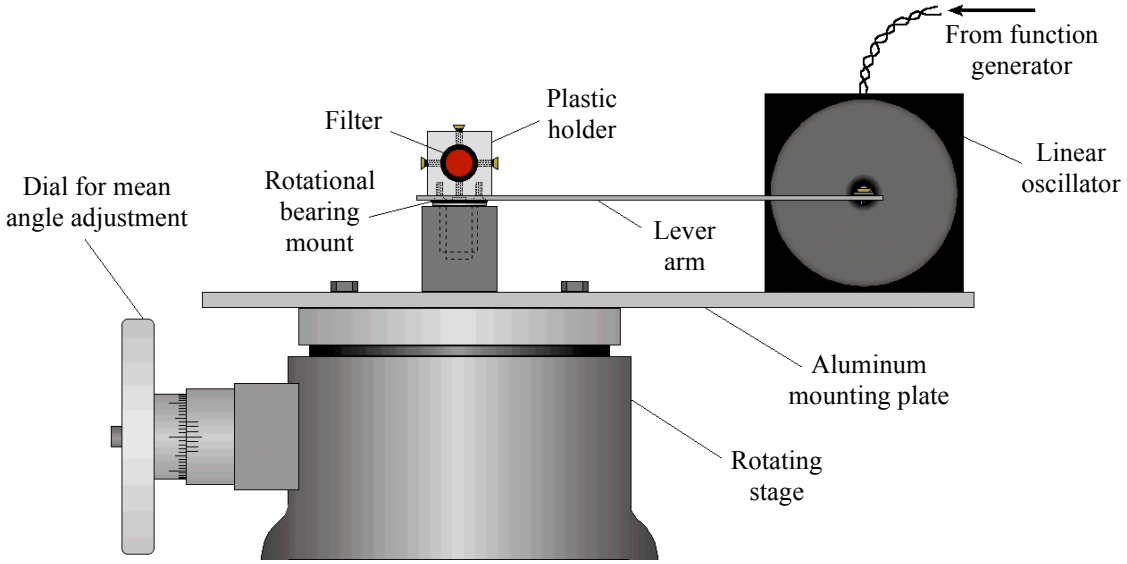
a reflectivity shift on the order of the spacing between two adjacent data points (0.06 nm), it should be possible to detect a shift in angle of incidence of only 0.0007°! This is actually an improvement upon the sensitivity of the angle modulation technique of Albrecht et. al.[10], which is stated to be around 0.001° (see discussion on page 25). It is likely that by using linear curve fits to the reflectivity slope data we could detect a shift of even less than 0.0007°. However, in order to test this, one must first have the equipment to increment the incidence angle by such a small amount (which I did not).

One final thing to note about the wide-bandwidth setup is that the actual zero crossing of the reflectivity slope data (Figure 24) does not coincide exactly with the minimum of the reflectivity (Figure 23). Take for example the "unshifted" sweep, for which the reflectivity slope data crosses the x-axis at about 635.4 nm. For this same sweep, the actual reflectivity (Figure 23) seems to bottom out around 636 nm, which is about half a nanometer above the slope data. The likely cause of this discrepancy is another source of oscillation in the system which results in the lock-in amplifier recording a slightly offset voltage. However, as long as one is only trying to detect a systematic shift in the data, this isn't really a problem because the offset is constant.

5-2. Medium-Bandwidth Setup

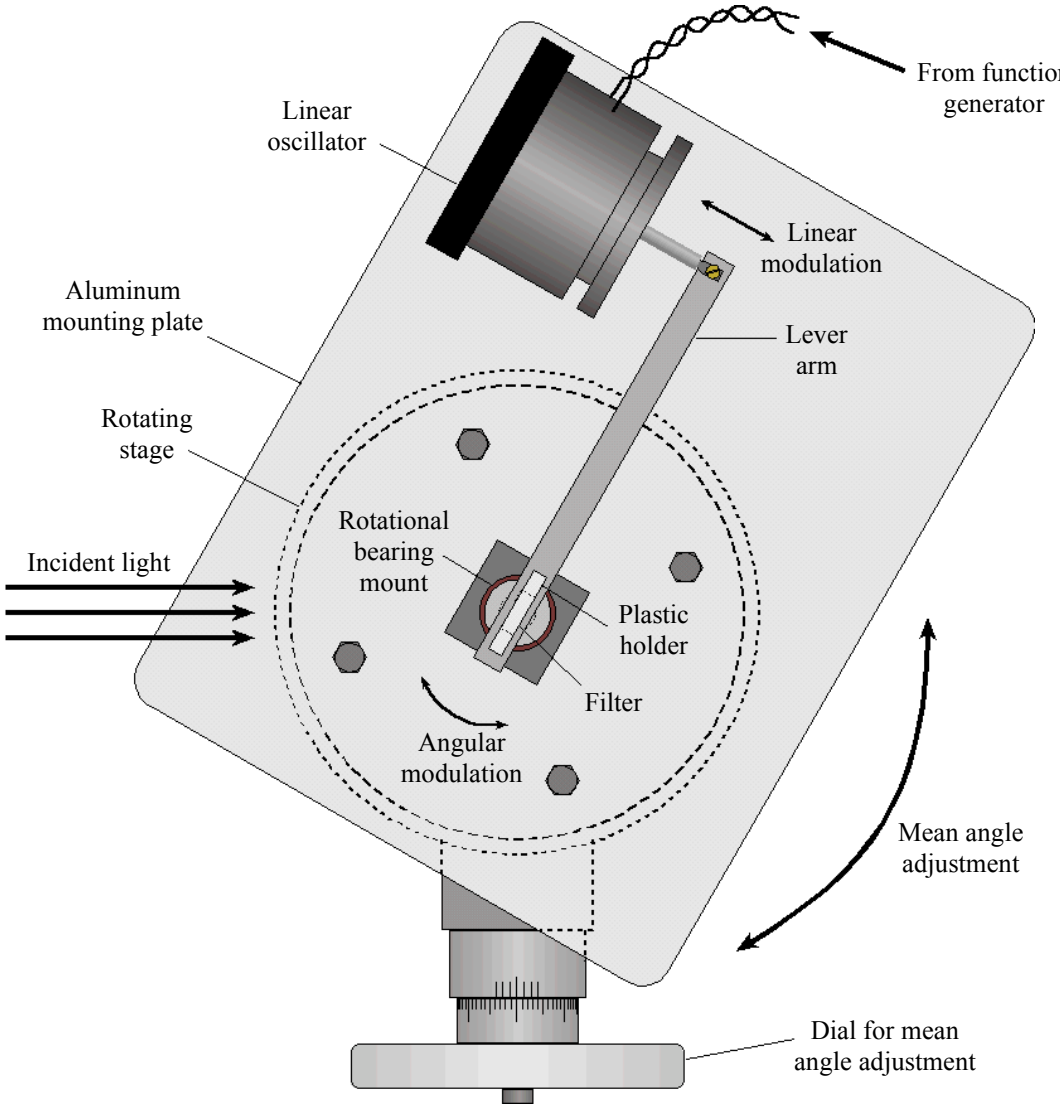
The medium-bandwidth setup was similar to the wide-bandwidth setup, except that this time a small 1 nm bandwidth interference filter was used to control the wavelength of the incident light. It turns out that the pass wavelength of a standard interference filter can be accurately adjusted simply by changing the incidence angle of the filter. Thus, a system was devised that would modulate the angle of the filter (with respect to the optic axis) while at the same time allowing precise control over the mean angle of the filter. The modulation was provided by the same linear oscillator that was used for the wide-bandwidth setup. However, this time the oscillator was mounted at a right angle to the filter and connected by a 6 inch lever arm. This allowed the precise conversion of the linear oscillation into an angular oscillation. The interference filter was secured inside a custom made plastic holder and mounted directly

Figure 25: Medium-bandwidth wavelength modulation apparatus (side view)



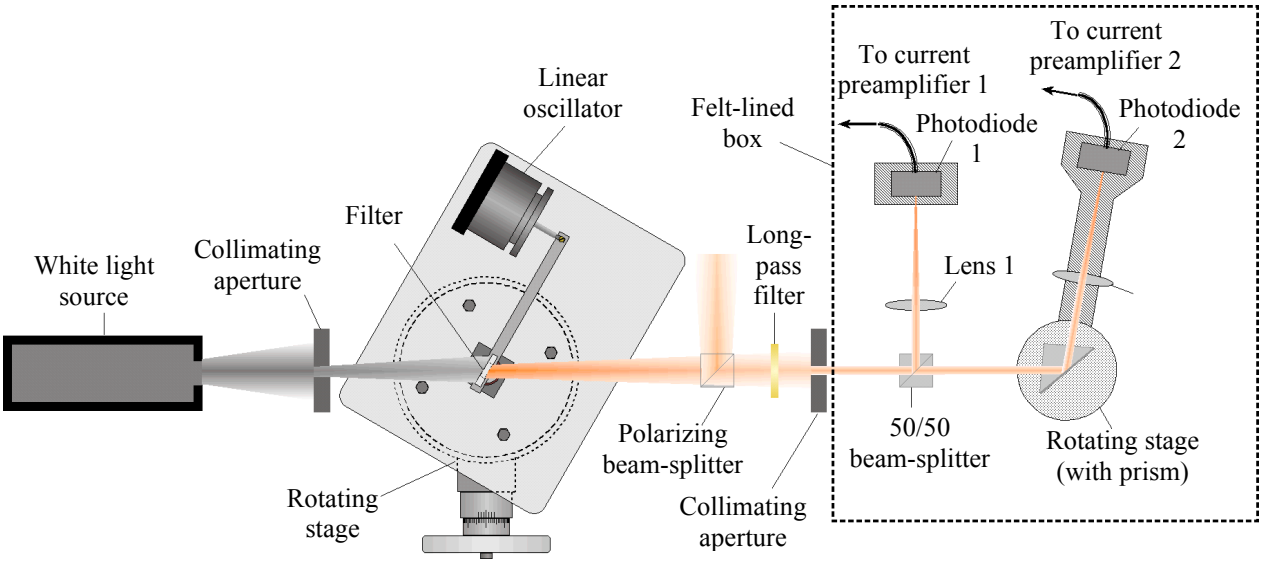
above a rotational bearing which constrained the filter to oscillate about its own centroid. The modulation setup was mounted on a rotating stage with an angular precision of 20 seconds. By stepping the rotating stage through a range of angles, the mean wavelength of incident light could be swept through the SPR minimum. This apparatus is illustrated in Figures 25 and 26. Figure 25 presents a side view of the apparatus (along the optic axis), while Figure 26 shows a top view.

Figure 26: Medium-bandwidth wavelength modulation apparatus (top view)



On the following page, Figure 27 shows the complete optical arrangement for the medium-bandwidth setup. This is nearly identical to the wide-bandwidth setup, except for the replacement of the oscillating wedge filter by the angle-modulated interference filter apparatus shown in Figures 25 and 26. The signal processing was also identical to that used for the wide-bandwidth experiment (see Figure 18).

Figure 27: Medium-bandwidth complete optical arrangement



The interference filter that was used for this experiment was ½" in diameter and allowed a 1 nm bandwidth of light centered at 633 nm to pass at normal incidence. For angles other than normal, the bandwidth remains at about 1 nm, however the central pass-wavelength decreases as a function of the angle. I obtained the following expression for calculating the wavelength as a function of incidence angle (for a standard interference filter) from the Oriel optics catalog[11]:

$$\lambda_\theta = \lambda_0 \left(1 - \left(\frac{n_0}{n} \right)^2 \sin^2 \theta \right). \qquad (57)$$

For equation (57):

λ_θ is the central pass-wavelength at an incidence angle θ
λ_o is the central pass-wavelength at normal incidence
n_o is the index of refraction of the surrounding medium
n is the effective index of refraction of the filter

I used a spectrometer to measure the central pass-wavelength of the filter for incidence angles ranging from -45° to +45° in 5° increments. I then fit a curve to this data based on equation (57) so that the values of λ_o and (n_o/n) could be accurately determined for my specific filter. The resulting best fit curve was then used to calculate the central pass-wavelength for the medium-bandwidth setup experiments described on the following pages.

Using the medium-bandwidth setup, it was not possible to achieve as wide a range of wavelengths as was possible with the wedge interference filter. Thus, for any given wavelength sweep, I could only sample the bottom portion of the SPR minimum. Despite this limitation in range, the medium-bandwidth setup exceeds the wide-bandwidth setup in sensitivity. In my first medium-bandwidth experiment, the wavelength of incident light was swept through the range from 604 nm to 628 nm. The mean filter angle increment between adjacent data points was 10 minutes, which is 30 times the minimum possible increment allowed by the rotating stage upon

which the filter was mounted (20 seconds). Because the relationship between the filter incidence angle and the pass-wavelength is not linear (see equation (57)), the corresponding wavelength increment varies throughout the experiment from 0.13 nm (at the 628 nm upper limit) to 0.25 nm (at the 604 nm lower limit). The prism incidence angle was fixed for this experiment at 43.68°. The linear oscillator was operating at 150 Hz with maximum amplitude (~0.5 mm peak to peak, corresponding to an angular modulation of about 0.2° peak-to-peak). Figure 28 shows the normalized reflectivity measured over this wavelength range:

Figure 28: Medium bandwidth – Experiment 1 (graph 1)

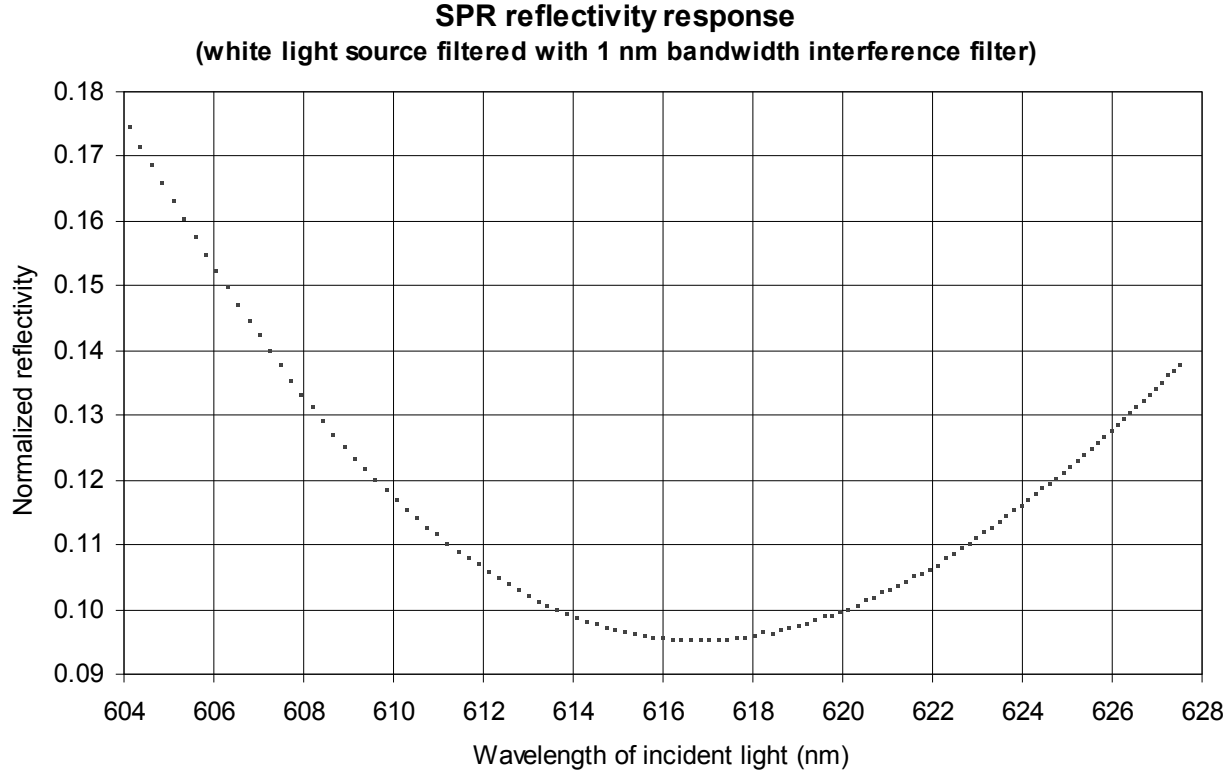


Figure 28 demonstrates that the medium-bandwidth setup yields very well-defined results with very little "bumpiness". As expected, the restricted wavelength range limits the sweep to the bottom 20% of the reflectivity minimum. An unexpected result of this data was that the absolute reflectivity minimum was not any less than that of the wide-bandwidth setup (about 10%). I had expected that with a 90% narrower bandwidth of incident light, the flattening effect discussed in the wide-bandwidth experiment would be significantly reduced. However, this was not the case. Thus, it seems likely that the range of incident angles played a key role in the flattening effect of the medium-bandwidth experiment (especially since the range of incidence angles has a secondary effect of widening the bandwidth of light passed by the interference filter).

Figures 29 and 30 show the measured slope of the reflectivity and a combination of the normalized reflectivity and slope, respectively. The slope of the reflectivity data points were calculated using the same method described in the wide-bandwidth section. However, there was

Figure 29: Medium bandwidth – Experiment 1 (graph 2)

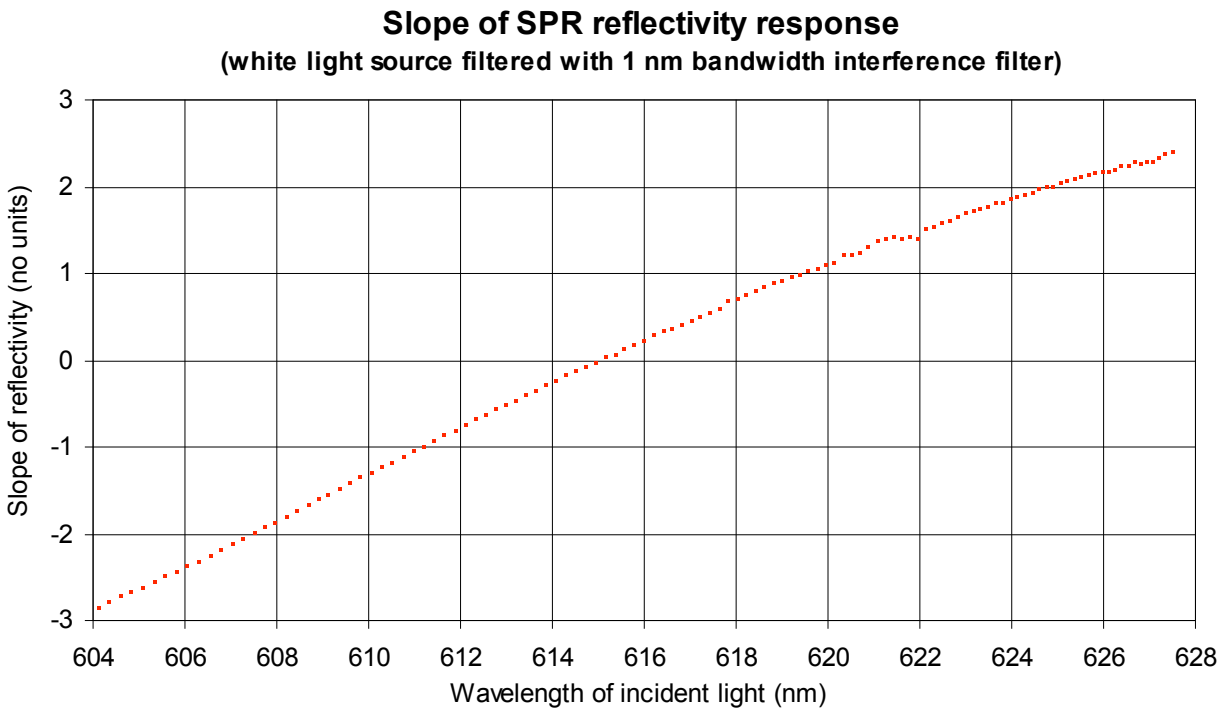
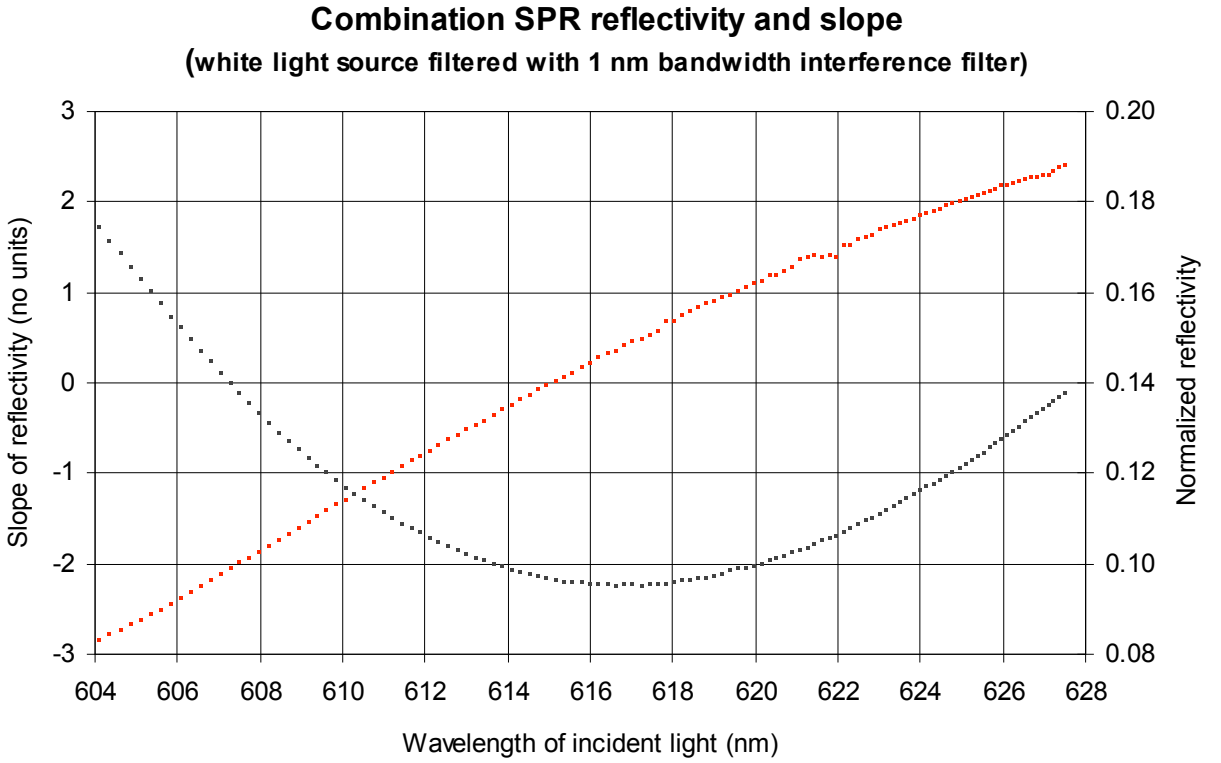


Figure 30: Medium bandwidth – Experiment 1 (graph 3)



one additional manipulation that needed to be performed on the medium-bandwidth data due to the non-linear relationship between the filter incidence angle and the pass-wavelength of light. The amplitude of the filter angle modulation was held constant throughout the experiment at about 0.2°. However, the amplitude of the corresponding *wavelength* modulation depends upon the mean incidence angle of the filter. For example, an angle modulation amplitude of 0.2° corresponds to a wavelength modulation amplitude of 0.29 nm at about 604 nm (~35° incidence), but only 0.16 nm at about 627 nm (~15° incidence). Thus, in the terms of the theoretical analysis on pages 22 through 24, the amplitude of the input variable (Δx) is not constant throughout the sweep. However, it is quite easy to resolve this problem simply by dividing each slope data point (as calculated per the method described in the wide-bandwidth section) by the slope of equation (57) evaluated at the angular setting of the particular data point in question. This counterbalances the effect of a variable Δx and the result is a proportional measurement of the reflectivity slope throughout the wavelength sweep. There is still a constant scaling factor that can be determined by numerical integration, however, as mentioned in the wide-bandwidth section, I did not find it important to perform this step.

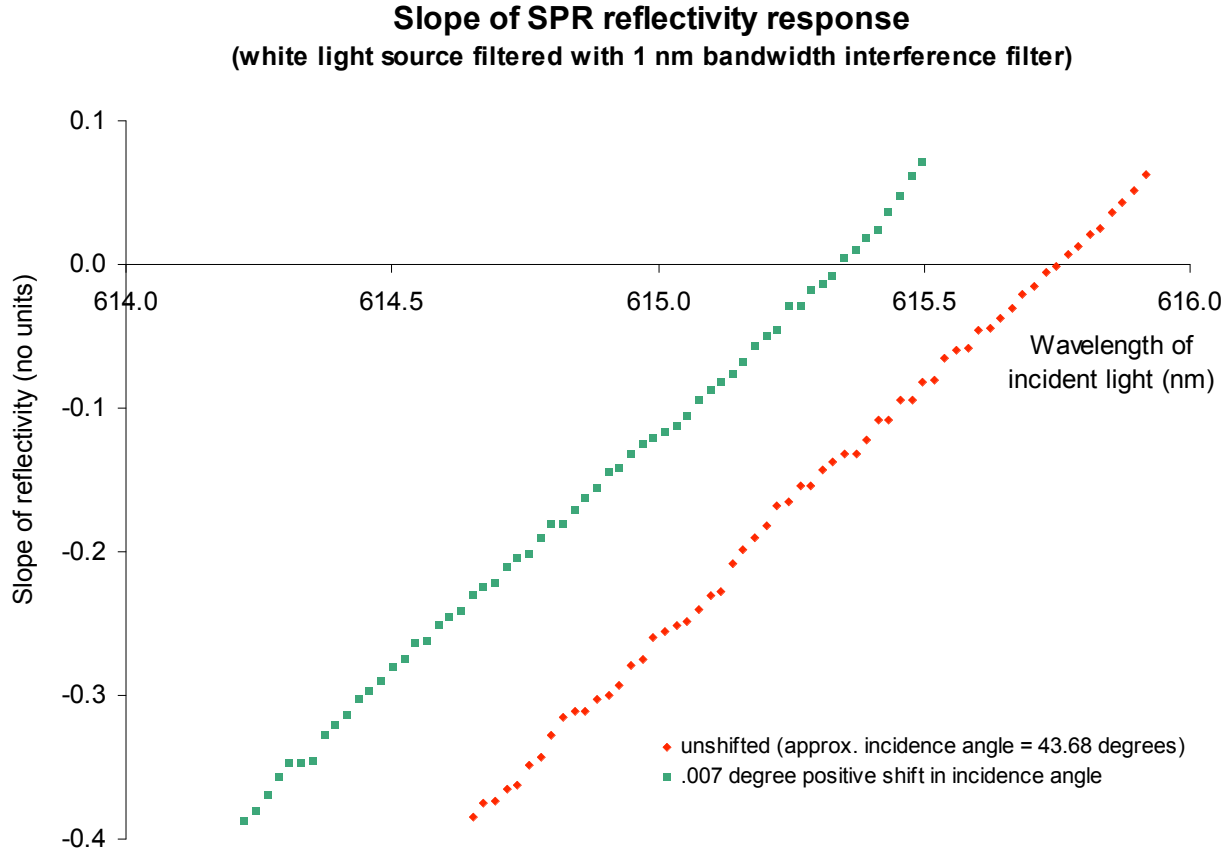
In studying the combination plot (Figure 30) we again observe a slight discrepancy between the apparent minimum of the reflectivity data and the zero crossing of the slope data (this effect was also noticed in the wide-bandwidth data). I am unsure of the exact cause of this offset in the slope data, but I am confident that it remains constant for the duration of any given experiment provided that all the electronics equipment are maintained at a constant setting. Thus, the offset is not an issue as long as one is only attempting to measure a systematic shift in the data (which is precisely my goal as I will discuss next).

In the second experiment with the medium-bandwidth setup, I wanted to compare the sensitivity of this setup to that of the wide-bandwidth setup. So, I performed two consecutive wavelength sweeps with the only difference between the two being a tiny shift in the prism incidence angle. For this experiment, the linear oscillator was operating at 150 Hz with full amplitude. The prism incidence angle setting was fixed at approximately 43.68° for the first sweep with a 0.007° positive shift for the second (similar to the experiment performed with the wide-bandwidth setup). The spacing between adjacent data points was 1 minute for the rotating stage upon which the interference filter was mounted, which corresponds to an average wavelength increment of 0.021 nm between adjacent data points (Note: the wavelength increment still varies due to the non-linear relationship between the filter angle and the pass-wavelength, however, this time the range was only from .0208 nm to .0215 nm due to the much narrower scope of this sweep). In actuality, the filter angle increment for this sweep was still three times the minimum possible allowed by the rotating stage (20 seconds). However, at this resolution, the visible noise in the system did not seem to offer much merit to zooming in any further. Despite this noise, the resulting shift in the recorded data is relatively well-defined and seems even to improve upon the resolution of the wide-bandwidth setup. Figure 31 shows the shift in the slope of the reflectivity for this experiment.

In terms of the wavelength of the incident light, the magnitude of the shift is about 0.5 nm. As should be expected, this is nearly equal to the wavelength shift observed in the wide-bandwidth data for the same incidence angle shift of 0.007°. However, with the medium-bandwidth experiment, this shift is equivalent to the spacing between 20 adjacent data points (due to the smaller increment between adjacent data points). If we make the same assumption as with the wide-bandwidth experiment (namely that it would be possible to detect a wavelength shift equivalent to the interval between two adjacent data points), then the estimated sensitivity

of the medium-bandwidth setup to a shift in the angle of incidence is 0.007° divided by 20, or ~0.0004°. Thus, by this estimation, the medium-bandwidth setup is more sensitive than the wide-bandwidth setup by a factor of ½.

Figure 31: Medium bandwidth – Experiment 2



The key to improving this sensitivity further is to reduce the noise recorded by the lock-in amplifiers. Figure 31 already reveals a significant amount of noise in this system which makes the pursuit of higher resolution seem unpromising. However, the lock-in amplifiers easily have the sensitivity to zoom in much closer. The problem is that the input light intensity tends to fluctuate for a variety of reasons. First, the light source itself had a tendency to go on swings either up or down on the order of 2%-3%. Most of the problems caused by this type of fluctuation could be removed from the data by normalization. However, if the data from different electronic components (for example the reference beam lock-in versus the reflected beam lock-in) are recorded at slightly different times, normalization cannot completely correct for this. Second, since the sensitivity of this system is already quite high, any outside light has the tendency to drastically effect the results. Even with a darkened room and the felt-lined box as a shield, a small amount of ambient light was able to sneak in through the holes allowed for interconnect cables and for the actual input beam. I noticed that simply by shifting my position in my chair (and thus re-directing the reflection of ambient light in the room) I was able to send the output from the lock-in amplifiers on wild swings of 5%-10%. Finally, the fact that the

signal beam (for both the wide-bandwidth and medium-bandwidth setups) is of such a small intensity to begin with makes it that much more susceptible to being affected by ambient light.

Nonetheless, even if further improvements were not feasible, the existing sensitivity of the medium-bandwidth setup (equivalent to a 0.0004° shift in incidence angle) is a full order of magnitude better than traditional angle-sweep SPR techniques (sensitivity ~0.007°) and at least twice as good as the angle modulation technique of Albrecht et. al.[10] (sensitivity ~0.001°).

5-3. Narrow-Bandwidth Setup

The third wavelength-modulated SPR technique that I worked with used the diode laser discussed in Section 3. Not only did this diode laser have the ability to sweep through a wavelength range of 10 nm, but it could also be modulated very slightly (amplitude ~0.01 nm) about the selected wavelength. Because of the small modulation amplitude of the laser, I refer to this as the narrow-bandwidth setup.

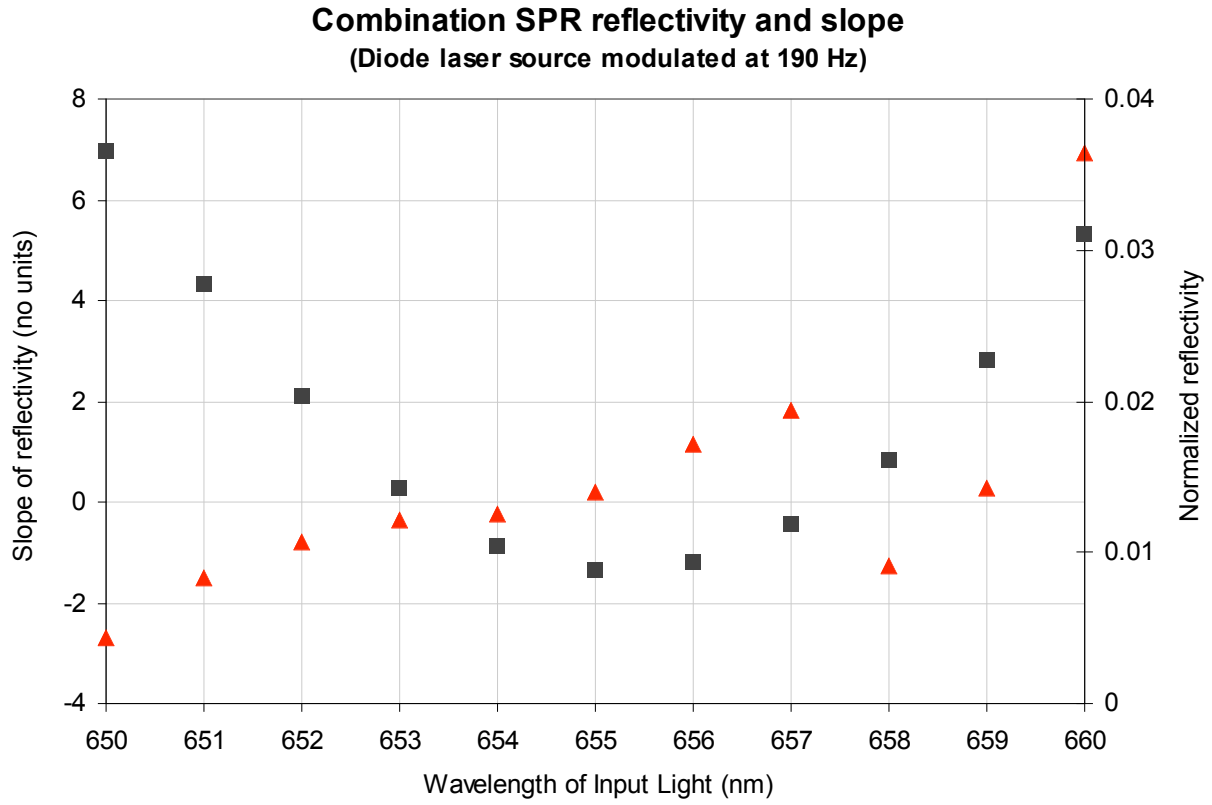
The modulation was accomplished by hooking up a voltage signal generator to the diode laser controller. This resulted in a tiny angular oscillation of a piezo controlled mirror inside the laser cavity which tunes the laser. The nice thing about this setup is that it was "ready-made" in that the laser was already manufactured with the desired tuning and modulation capability (in contrast to the wide and medium-bandwidth setups which both required my home-built mechanics to control the wavelength of input light). The unfortunate thing is that the laser tuning characteristics were not very stable because it had the tendency to "mode-hop" which resulted in some unpredictable data as I will discuss below.

For the wavelength-modulated diode laser experiments (which were performed by Dr. Larry Sorensen, Jin Li, and myself), we used an optical setup very similar to the basic SPR setup shown in Figure 10. The only additional equipment was the diode laser, the diode laser controller, and the signal generator. For this experiment we used a very coarse interval between adjacent data points (1 nm). The diode laser is capable of stepping through it's wavelength range at a much finer interval (0.01 nm), however, the fact that our results were so unpredictable did not seem to offer much merit to zooming in any closer. A combination SPR reflectivity and slope plot of this experiment is shown in Figure 32.

In Figure 32, the black squares are a plot of the SPR reflectivity while the red triangles represent the slope of the reflectivity. The reflectivity itself seems to be pretty consistent, however, the slope of the reflectivity is quite unpredictable. For such a narrow range about the SPR minimum (10 nm), the theoretical slope should be almost perfectly linear. Unfortunately, the measured slope is somewhat non-linear from 650 nm to 657 nm and turns to complete randomness above 657 nm. The one redeeming quality about this is that it seems to correlate with the bumpiness in the simple diode laser wavelength sweep shown in Figure 12. In Figure 12, the reflectivity shows some minor periodic fluctuations at the lower end of the wavelength range which grows into more drastic fluctuations at the upper end. One can easily imagine that the measurement of the slope of such a curve would yield wildly unpredictable results, especially at the upper end.

It's unfortunate that the diode laser seems to mode-hop because it has some distinct advantages over the wide and medium-bandwidth setups. First, the diode laser is electronically controlled which would make automation a simple matter (no dials or micrometers to adjust). Second, the light is much more intense than the techniques involving filtered white light. This

Figure 32: Data from the narrow-bandwidth setup



makes it much less susceptible to the effects of ambient light. Third, the beam is quite narrow in cross section with the near perfect collimation common to lasers which makes it possible to sample surfaces with a small cross-sectional area. Finally, the precision of the diode laser is very fine (minimum interval of 0.01 nm). Unfortunately, as Figure 32 demonstrates, this fine precision could not be utilized because of the laser's tendency to mode hop.

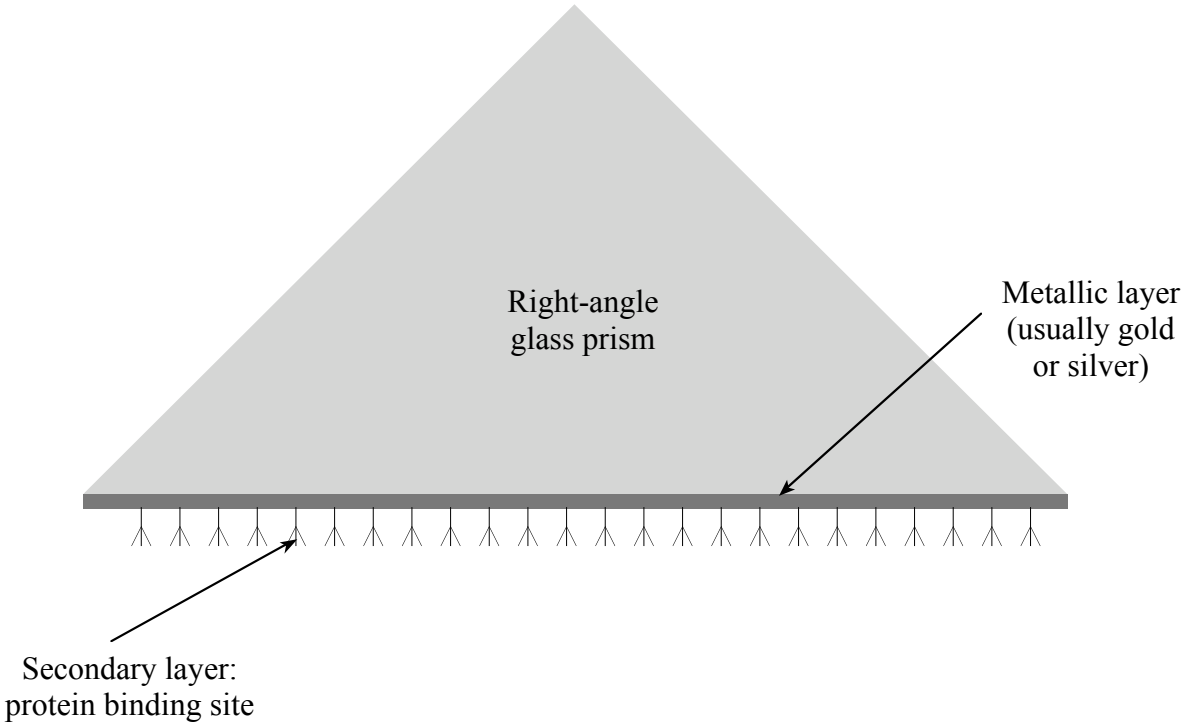
Section 6: Biosensor Applications of SPR

About thirty-five years ago, the discovery of surface plasmon resonance began simply as a novel optical phenomenon. However, in the last ten years, its practical use as a technique to analyze thin film layers has grown dramatically. One publication[12] actually charts this growth by graphing the number of published research papers on SPR sensors over the last few years. According to the data they collected, what started as just a handful of papers in 1992 grew well into the hundreds just five years later. The reason that surface plasmon resonance has become such a fast-growing topic is because of the high level of sensitivity the technique yields towards small geometrical or compositional changes in the thin film layers at the surface itself.

With this high level of sensitivity, SPR can be used to accurately detect and measure a variety of different surface features or changes that occur at the surface including film thickness[10], bubble nucleation[13], solute concentration[14], and molecular bonding to the surface itself[15]. This last application is perhaps the most interesting because of the vast number of proteins and other biomolecules that the scientific community will be attempting to characterize over the coming years.

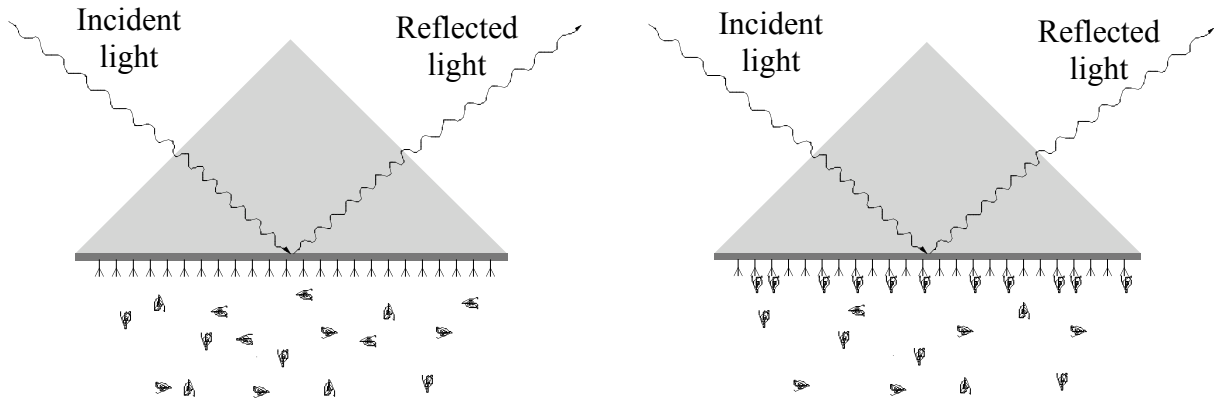
Using SPR, it is possible to detail the specific binding characteristics of a given protein according to the following simplified scheme. First, the metallic layer of the standard right-angle prism shown in Figure 33 is coated with a secondary layer of receptors which acts as a potential binding site for the given protein. The surface is then exposed to a solution which contains the isolated protein. Depending on the characteristics of the receptor and the given protein, the

Figure 33: Right-angle prism with binding sites for proteins



protein will bind to the surface to varying degrees. If the protein binds, the effective refractive index and total thickness of the combined layers will change slightly (See Figure 34). This causes a subsequent shift in the SPR reflectivity curve which can be detected using the techniques discussed in the preceding sections of this report. The magnitude of the shift in the SPR reflectivity corresponds to the extent that binding occurs.

Figure 34: Protein-binding (b) causes a shift in the SPR reflectivity curve



(a) Solution with isolated protein is introduced to the surface.

(b) Protein begins to bind to the surface which can be detected by a shift in the reflectivity.

One of the great advantages of this system is that the detection portion of the setup is completely passive, ie. it has no effect on the proteins or the receptors. Thus, the extent of protein binding can be monitored in real time without affecting the outcome. It is also possible to create an array of binding sites along the prism surface so that the protein's binding characteristics could be tested with a number of different receptors. Unfortunately, there are some limitations to this system. For example, the minimum surface area of the individual array components is of course limited by resolution of the incident light. There are also some difficulties involved with coating the metal surface with receptors in the first place. In fact, it is often the case that a number of different layers are needed to create a film that will resist washing off when exposed to solution (this, however, is a problem that is beyond the scope of this paper).

One difficulty that this paper does hope to address is simply the ability of the system to detect very small shifts in the SPR reflectivity. By its very nature, SPR tends to be quite sensitive to small changes at the surface. However, some of the important surface changes such as protein or DNA binding can still be too subtle for detection by the traditional techniques of SPR. In cases such as these, people have resorted to signal enhancing methods such as tagging. One such research group attached small gold colloids (12 nm diameter) to the ends of the DNA strands they were studying[16]. In this way, when the DNA strands in question bound to the complementary strands on the prism surface, the resulting SPR reflectivity shift would be greatly amplified by the complementary effect of the gold colloids. According to their published results,

this seemed to work quite well, however, it is a very undesirable extra step that one wishes to avoid if possible. Also, by tagging a protein with such large colloids as these, there is the risk that the binding characteristics of the protein molecule will be significantly altered.

Thus it was hoped that the wavelength-modulation techniques discussed in Section 5 might provide some of the extra sensitivity required without the need for intrusive tagging. In fact, both the wide-bandwidth and medium-bandwidth wavelength modulation setups show significant sensitivity improvements over the traditional methods. This improved sensitivity makes either the wide or medium-bandwidth setup an ideal method of monitoring the extent of protein binding that occurs at the surface. The major difficulties with these two setups is the weak intensity of the filtered beam and imperfect collimation, which causes the spacial resolution at the surface to be somewhat deficient. For this reason, it would be difficult to accurately sample a particularly fine micro-array of binding sites. In contrast, the diode laser light source used in the narrow-bandwidth wavelength modulation setup does not have this problem. Unfortunately, due to stability problems with the diode laser that I used, the sensitivity was not sufficient. Nevertheless, it does not seem far-fetched that an accurately tunable diode laser would be available somewhere on the market. Thus, the narrow-bandwidth setup still demonstrates significant potential.

Section 7: Conclusion

At the beginning of this paper, I stated that I had two major goals. My first goal was to present the theory of surface plasmon resonance in a clear and detailed manner. Not being much of a theorist by nature, this was a big challenge for me and by attempting it I have learned quite a lot. In doing so, I've also come to the conclusion that surface plasmon resonance has great value as an educational pathway into understanding advanced physics topics. From a theoretical standpoint, it incorporates elements of electricity and magnetism, quantum mechanics, wave theory, and optics. Furthermore, as a hands-on physics experiment, it's both interesting and reliable (in all my research, it never failed to work!).

My second goal was to discuss my own research on wavelength-modulated SPR. For me, this was the most enjoyable and gratifying part of the project and thus I will discuss it now in a bit more depth. Two of the three wavelength-modulation setups that I experimented with (wide-bandwidth and medium-bandwidth) were about as successful as I could have hoped. I estimate the sensitivity of these two setups (to a small shift in the angle of incidence) to be 0.0007° and 0.0004°, respectively. Both of these improve upon the stated sensitivity of the angle modulation technique of Albrecht et al[10] (0.001°) and are significantly better than the simple angle-sweep technique (~0.007°) discussed in section 3. There is also the potential to improve this sensitivity even more if some of the excess noise could be eliminated (refer to Figure 31).

The key limitation of the wide and medium-bandwidth setups is in the spatial resolution at the surface itself. Due to the low intensity per unit area of the filtered and collimated light, these setups both require a wide beam (1 to 5 mm) for accurate detection. Thus, neither of these two setups are practical as a gauge for protein binding along a finely spaced micro-array of binding sites. However, due to their high level of sensitivity, they are more than capable of detecting minute changes at a surface of at least a few millimeters in diameter.

The one disappointment of this project was with the narrow-bandwidth setup which did not achieve as good a sensitivity as I had hoped. The main reason for this is the diode laser's tendency to mode-hop which is most clearly demonstrated by Figure 12. However, the diode laser's high intensity, near perfect collimation, and sharply defined beam offer distinct advantages over the wide and medium-bandwidth setups. Thus, if a more consistent diode laser were found, the narrow-bandwidth setup could yield positive results.

However, the success of the wide-bandwidth and medium-bandwidth setups discussed above more than compensate for the failure of the diode laser and prove that wavelength-modulated SPR can be a valuable detection technique.

References

[1] J. R. Sambles, G. W. Bradbery, and Fuzi Yang, "Optical excitation of surface plasmons", *Contemporary Physics* **32**, pp. 173-183 (1991).

[2] C. Kittel, Introduction to Solid State Physics, 5th Ed., John Wiley & Sons, New York, p. 289 (1976).

[3] E. Kretschmann, and H. Raether, "Radiative decay of non-radiative surface plasmons excited by light", *Z. Naturforsch* **23a**, pp. 2135-2136 (1968).

[4] A. Otto, "Excitation of non-radiative surface plasma waves in silver by the method of frustrated total reflection", *Z. Phys.* **216**, pp. 398-410 (1968).

[5] G. J. Sprokel and J. D. Swalen, "The attenuated total reflection method", Handbook of Optical Constants of Solids, Vol II, Academic Press, New York, pp. 75-95 (1991).

[6] W. Hansen, "Electric fields produced by the propagation of plane coherent electromagnetic radiation in a stratified medium", *Journal of the Optical Society of America* **58**, pp. 380-390 (1968).

[7] F. de Fornel, Evanescent Waves: From Newtonian Optics to Atomic Optics, Springer-Verlag, Berlin (2001).

[8] "Operating manual: SR830 DSP lock-in amplifier", Stanford Research Systems, Sunnyvale CA (1993).

[9] A. N. Dharamsi, "A theory of modulation spectroscopy with applications of higher harmonic detection", *J. Phys D: Appl. Phys.* **29**, pp. 540-549 (1996).

[10] U. Albrecht, H. Dilger, P. Evers, P. Leiderer, "High resolution surface plasmon measurements – a sensitive probe for thickness and structural information of ultrathin films", *SPIE* **1594**, pp. 344-350 (1991)

[11] "Optics and Filters", vol. III, Oriel Corporation, Stratford CT (1990).

[12] J. Homola, S. Yee, G. Gauglitz, "Surface plasmon resonance sensors: review", *Sensors and Actuators, B,* **54**, pp. 3-15 (1999).

[13] O. Yavas, A. Schilling, J. Bischof, J. Boneberg, P. Leiderer, "Bubble nucleation and pressure generation during laser cleaning of surfaces", *Appl. Phys., A,* **64**, pp. 331-339 (1997).

[14] F. Markey, "Measuring concentration: why Biacore is best", *Biajournal* **2**, pp. 8-11 (1999).

[15] A. Frutos, R. Corn, "SPR of ultrathin organic films", *Analytical Chemistry, A,* **70**, pp. 449-455 (1998).

[16] L. He, M. Musick, S. Nicewarner, F. Salinas, S. Benkovic, M. Natan, C. Keating, "Colloidal Au-enhanced surface plasmon resonance for ultrasensitive detection of DNA hybridization", *J. Am. Chem. Soc.* **122**, pp. 9071-9077 (2000).

Credits

Thank you to Larry Sorensen, Jason Alferness, John Stoltenberg, and Ron Musgrave for all your assistance!

Appendices

Appendix A:

Demonstration that surface plasmons cannot exist with electric field parallel to surface.

The following demonstration traces the path of the derivation worked out in section 2, beginning with equation (9). Each equation in this section (distinguished by an 'A'), is akin to the comparably numbered equation in section 2. The only difference is that the polarization of the electromagnetic waves is chosen to be such that the *electric* field lies parallel to the surface of propagation (along the y-axis). Note that in the derivation provided in section 2, it was the *magnetic* field that was assumed to lie along the y-axis. In this new scenario B_y, E_x, and E_z are equal to zero. Thus, equations (7) and (8) simplify further to:

$$\left(-\frac{\partial E_y}{\partial z}\right)\hat{x}+\left(\frac{\partial E_y}{\partial x}\right)\hat{z}=\left(-\frac{\partial B_x}{\partial t}\right)\hat{x}+\left(-\frac{\partial B_z}{\partial t}\right)\hat{z} \tag{A-9}$$

$$\left(\frac{\partial B_x}{\partial z}-\frac{\partial B_z}{\partial x}\right)\hat{y}=\left(\frac{\mu\varepsilon}{c^2}\cdot\frac{\partial E_y}{\partial t}\right)\hat{y}\,. \tag{A-10}$$

Splitting these equations into components and executing the noted derivatives upon equations (1) and (2), leaves the following three equations:

$$\frac{\partial B_{0x}}{\partial z}\left(e^{i(k_{sp}\cdot x-\omega t)}\right)-ik_{sp}B_{0z}\left(e^{i(k_{sp}\cdot x-\omega t)}\right)=-\frac{i\omega\mu\varepsilon}{c^2}E_{0y}\left(e^{i(k_{sp}\cdot x-\omega t)}\right) \tag{A-11}$$

$$-\frac{\partial E_{0y}}{\partial z}\left(e^{i(k_{sp}\cdot x-\omega t)}\right)=i\omega B_{0x}\left(e^{i(k_{sp}\cdot x-\omega t)}\right) \tag{A-12}$$

$$ik_{sp}E_{0y}\left(e^{i(k_{sp}\cdot x-\omega t)}\right)=i\omega B_{0z}\left(e^{i(k_{sp}\cdot x-\omega t)}\right). \tag{A-13}$$

The oscillating terms of equations (A-11), (A-12), and (A-13) cancel due to the fact that both the wave vector (k_{sp}) and the frequency (ω) remain constant throughout the surface plasmon wave, yielding the following:

$$\frac{\partial B_{0x}}{\partial z}-ik_{sp}B_{0z}=-\frac{i\omega\mu\varepsilon}{c^2}E_{0y} \tag{A-14}$$

$$\frac{\partial E_{0y}}{\partial z}=-i\omega B_{0x} \tag{A-15}$$

$$B_{0z}=\frac{k_{sp}}{\omega}E_{0y}\,. \tag{A-16}$$

Substituting equation (A-16) into equation (A-14) and simplifying, leaves the following expression:

$$\frac{\partial B_{0x}}{\partial z} = \frac{i}{\omega}\left(k_{sp}^2 - \frac{\omega^2 \mu\varepsilon}{c^2}\right)E_{0y}\,. \tag{A-17}$$

Taking the partial derivative of both sides of equation (A-17) with respect to *z* and then plugging in equation (A-15) results in the following differential equation:

$$\frac{\partial^2 B_{0x}}{\partial z^2} = \left(k_{sp}^2 - \frac{\omega^2 \mu\varepsilon}{c^2}\right)B_{0x}\,. \tag{A-18}$$

The solution of this differential equation is given below, with A being the amplitude of the electromagnetic wave.

$$B_{0x} = Ae^{-\left(\sqrt{k_{sp}^2 - \frac{\omega^2\mu\varepsilon}{c^2}}\right)z}\,. \tag{A-19}$$

This result can then be applied to equation (A-17) in order to determine E_{0y}:

$$E_{0y} = (i\omega)\left(k_{sp}^2 - \frac{\omega^2 \mu\varepsilon}{c^2}\right)^{-\frac{1}{2}} Ae^{-\left(\sqrt{k_{sp}^2 - \frac{\omega^2\mu\varepsilon}{c^2}}\right)z}\,. \tag{A-20}$$

Which can subsequently be applied to equation (A-16) to determine B_{0z}:

$$B_{0z} = (ik_{sp})\left(k_{sp}^2 - \frac{\omega^2 \mu\varepsilon}{c^2}\right)^{-\frac{1}{2}} Ae^{-\left(\sqrt{k_{sp}^2 - \frac{\omega^2\mu\varepsilon}{c^2}}\right)z}\,. \tag{A-21}$$

As before, we can apply these solutions to any sharp interface between two different isotropic and homogenous media. The general arrangement for the interface between the two media is shown in Figure 2 (see section 2), with the axes defined such that medium 1 is on the positive z side of the origin and medium 2 is on the negative z side of the origin.

Substituting equations (A-19), (A-20), and (A-21) into equations (1) and (2) gives the electric and magnetic fields at all points near the surface, as follows:

In Medium 1

$$B_{1x} = A_1 \left(e^{-\left(\sqrt{k_{sp}^2 - \frac{\omega^2 \mu_1 \varepsilon_1}{c^2}} \right) z} \right) \left(e^{i(k_{sp} \cdot x - \omega t)} \right) \tag{A-22}$$

$$B_{1z} = (ik_{sp}) \left(k_{sp}^2 - \frac{\omega^2 \mu_1 \varepsilon_1}{c^2} \right)^{-\frac{1}{2}} A_1 \left(e^{-\left(\sqrt{k_{sp}^2 - \frac{\omega^2 \mu_1 \varepsilon_1}{c^2}} \right) z} \right) \left(e^{i(k_{sp} \cdot x - \omega t)} \right) \tag{A-23}$$

$$E_{1y} = (i\omega) \left(k_{sp}^2 - \frac{\omega^2 \mu_1 \varepsilon_1}{c^2} \right)^{-\frac{1}{2}} A_1 \left(e^{-\left(\sqrt{k_{sp}^2 - \frac{\omega^2 \mu_1 \varepsilon_1}{c^2}} \right) z} \right) \left(e^{i(k_{sp} \cdot x - \omega t)} \right) \tag{A-24}$$

$B_{1y} = E_{1x} = E_{1z} = 0$ (by definition).

In Medium 2

$$B_{2x} = A_2 \left(e^{\left(\sqrt{k_{sp}^2 - \frac{\omega^2 \mu_2 \varepsilon_2}{c^2}} \right) z} \right) \left(e^{i(k_{sp} \cdot x - \omega t)} \right) \tag{A-25}$$

$$B_{2z} = (-ik_{sp}) \left(k_{sp}^2 - \frac{\omega^2 \mu_2 \varepsilon_2}{c^2} \right)^{-\frac{1}{2}} A_2 \left(e^{\left(\sqrt{k_{sp}^2 - \frac{\omega^2 \mu_2 \varepsilon_2}{c^2}} \right) z} \right) \left(e^{i(k_{sp} \cdot x - \omega t)} \right) \tag{A-26}$$

$$E_{2y} = (-i\omega) \left(k_{sp}^2 - \frac{\omega^2 \mu_2 \varepsilon_2}{c^2} \right)^{-\frac{1}{2}} A_2 \left(e^{\left(\sqrt{k_{sp}^2 - \frac{\omega^2 \mu_2 \varepsilon_2}{c^2}} \right) z} \right) \left(e^{i(k_{sp} \cdot x - \omega t)} \right) \tag{A-27}$$

$B_{2y} = E_{2x} = E_{2z} = 0$ (by definition).

As before, there are three things to note:

1. This time, the magnetic permeability (μ) has been purposefully retained in equations (A-22) through (A-27). The reason for this is that the magnetic permeability turns out to be the key physical characteristic that prevents the existence of surface plasmons with electric fields along the y-axis.
2. The dielectric constants (ε) and magnetic permeabilities (μ) for the two media are distinguished by subscripts 1 and 2.
3. For the set of equations defining the fields in medium 2, all appearances of z merit an extra negative sign (since medium 2 lies below the origin). This also causes the overall signs of B_{2z} and E_{2y} to be reversed from that of B_{1z} and E_{1y}, respectively, due to the partial derivative with respect to z taken when applying equation (A-17).

Now, we take the boundary condition for magnetic fields *parallel* to a surface, $\frac{1}{\mu_1} B_{1\|} = \frac{1}{\mu_2} B_{2\|}$ (at z = 0), and apply it to equations (A-22) and (A-25) to determine that

$$A_1 = \frac{\mu_1}{\mu_2} A_2 \tag{A-28}$$

The boundary condition for magnetic fields *perpendicular* to a surface, $B_{1\perp} = B_{2\perp}$ (at z = 0), can then be applied to equations (A-23) and (A-26), which gives

$$A_1 \left(k_{sp}^2 - \frac{\omega^2 \mu_1 \varepsilon_1}{c^2} \right)^{-\frac{1}{2}} = -A_2 \left(k_{sp}^2 - \frac{\omega^2 \mu_2 \varepsilon_2}{c^2} \right)^{-\frac{1}{2}} \tag{A-29}$$

Since $A_1 = \frac{\mu_1}{\mu_2} A_2$ per equation (A-28), this simplifies to

$$\mu_1 \left(k_{sp}^2 - \frac{\omega^2 \mu_1 \varepsilon_1}{c^2} \right)^{-\frac{1}{2}} = -\mu_2 \left(k_{sp}^2 - \frac{\omega^2 \mu_2 \varepsilon_2}{c^2} \right)^{-\frac{1}{2}} \tag{A-30}$$

Once again, we know that the $\left(k_{sp}^2 - \frac{\omega^2 \mu_1 \varepsilon_1}{c^2} \right)^{-\frac{1}{2}}$ term and the $\left(k_{sp}^2 - \frac{\omega^2 \mu_2 \varepsilon_2}{c^2} \right)^{-\frac{1}{2}}$ term must both be positive, otherwise equations (A-22) through (A-27) would diverge as z approached either negative or positive infinity. Therefore, in order for equation (A-30) to be true, either μ_1 or μ_2 (but not both) must be negative. However, whereas this was possible for the dielectric constant (ε), it is not possible for the magnetic permeability (μ) to be negative (at least for any known material). Thus, the existence of a surface plasmon with an electric field vector along the y-axis is not physically possible.

Appendix B:

Electronic Equipment List

(1) 6202 External Cavity Tunable Diode Laser (New Focus Inc.)

(1) 6200 External Cavity Tunable Diode Laser Controller (New Focus Inc.)

(2) SR830 DSP Lock-in amplifiers (Stanford Research Systems)

(2) SR570 Low Noise Current Pre-amplifiers (Stanford Research Systems)

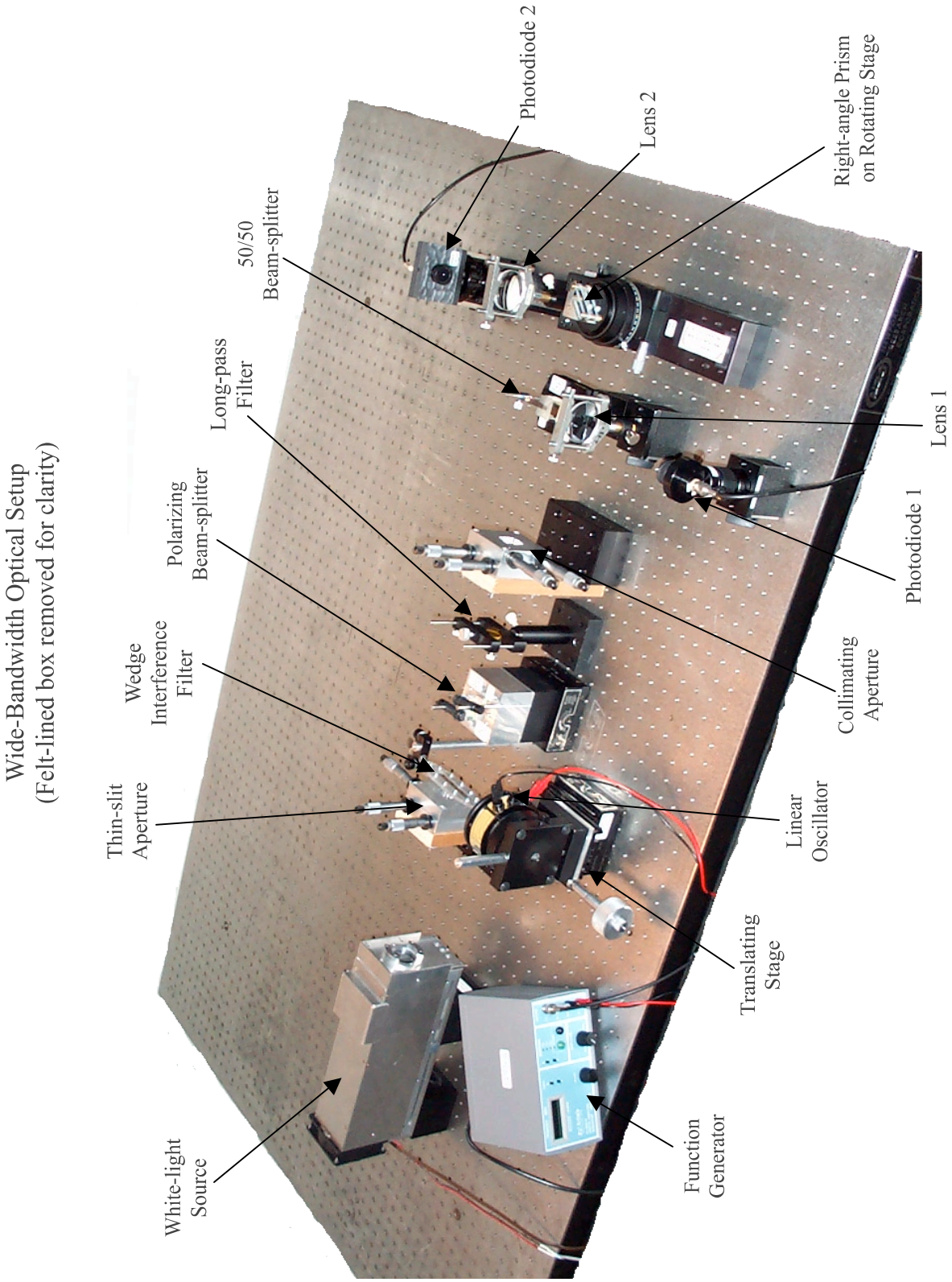
(2) 2000 Multimeters (Keithley)

(1) PI-9587C Digital Function Generator – Amplifier (Pasco Scientific)

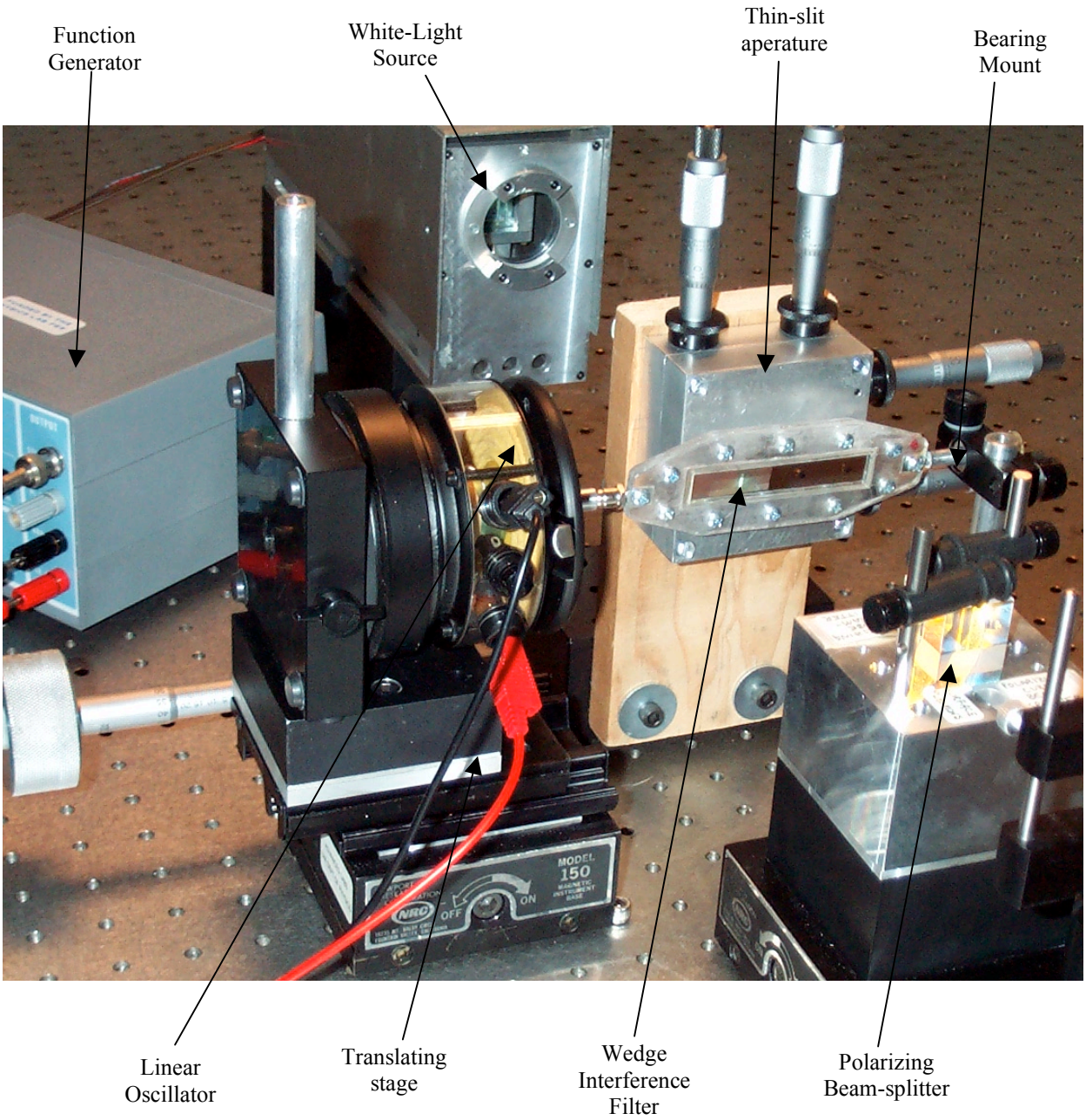
(1) SF-9324 Mechanical Vibrator (Pasco Scientific)

(2) 3CDPI Photodiodes (UDT)

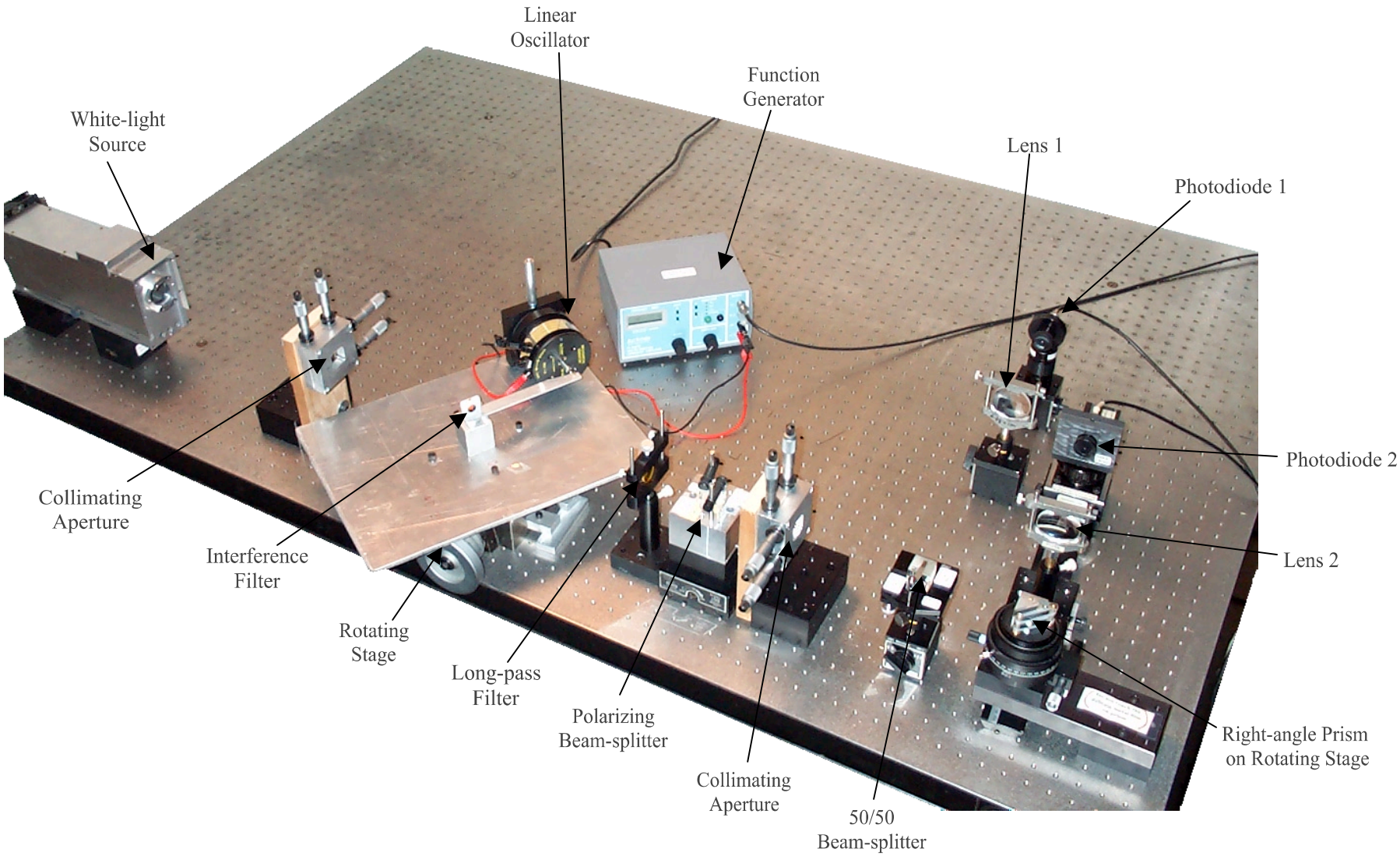
Appendix C: Photos of Wide-Bandwidth and Medium-Bandwidth Setups



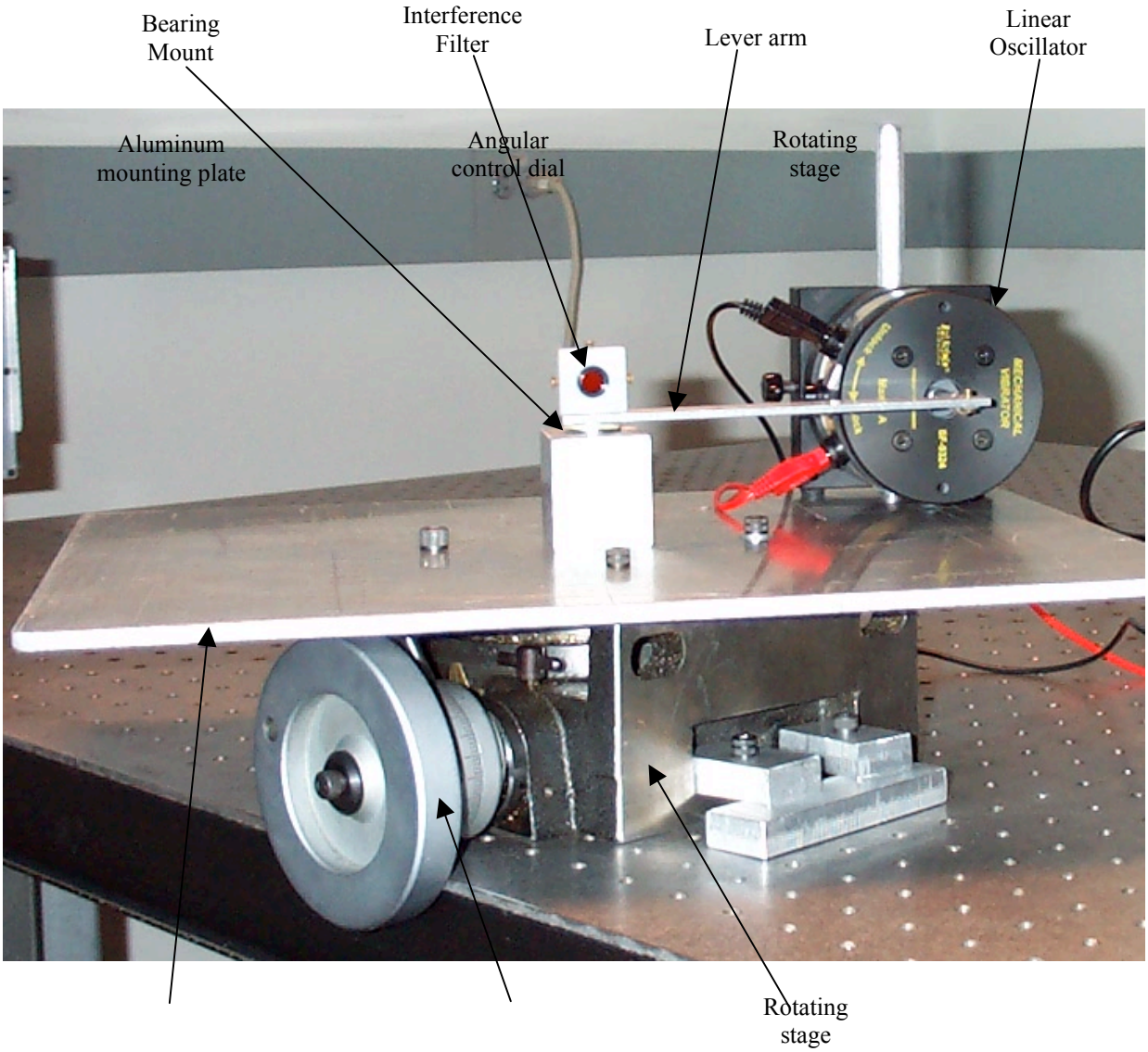
Close-up of Linear Oscillator and Wedge Interference Filter (wide-bandwidth setup)



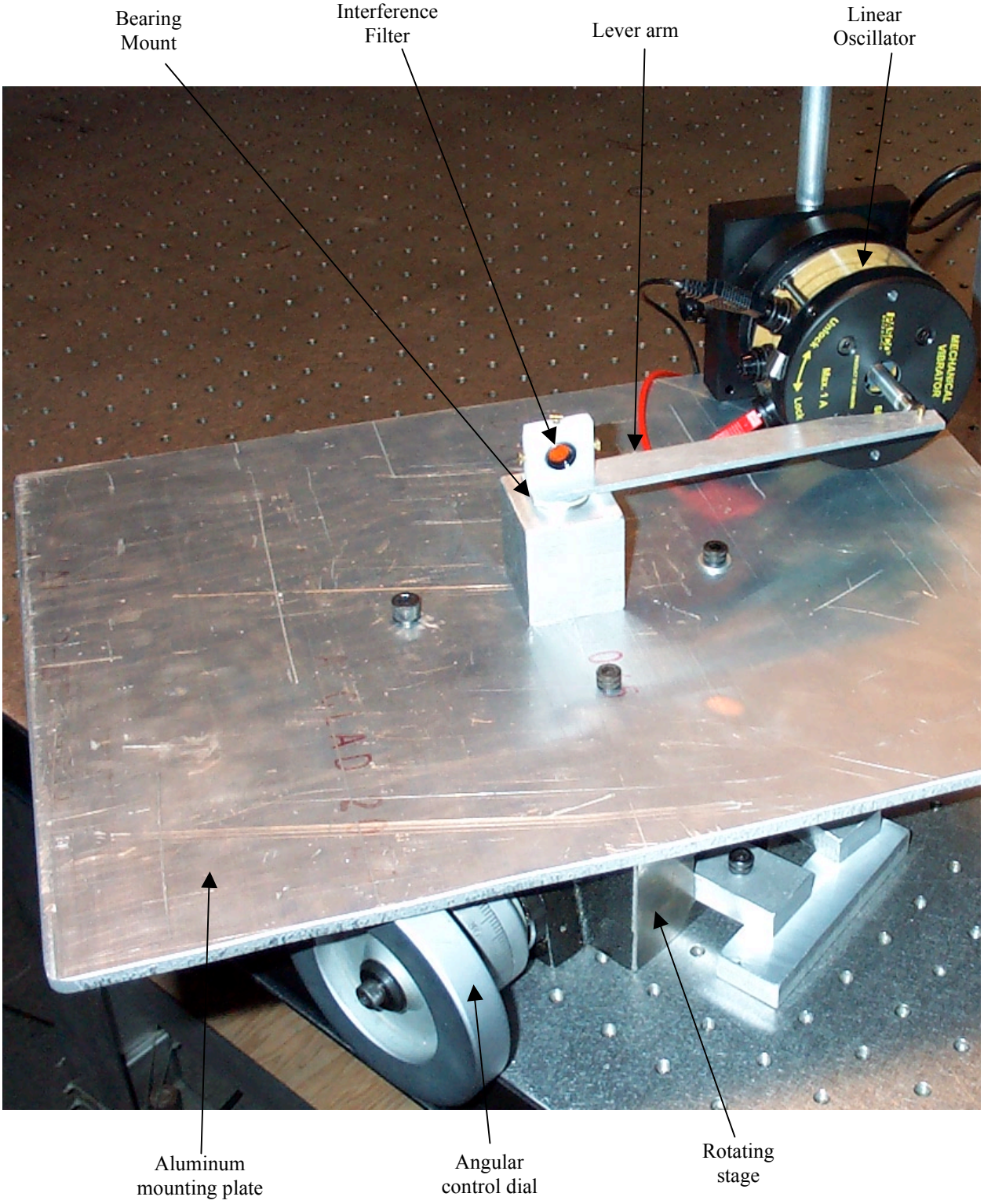
Medium-Bandwidth Optical Setup
(Felt-lined box removed for clarity)



Close-up of Rotating Stage, Interference Filter, and Linear Oscillator
(medium-bandwidth technique)



Top View of Rotating Stage, Interference Filter, and Linear Oscillator
(medium-bandwidth technique)



Appendix D:

Alternative Method for Wavelength Modulation – The Spinning Filter

I also worked on an alternative method for wavelength modulation which was never built. The principle of this method is to use the same filter as in the medium-bandwidth setup, except to spin the filter with an electric motor instead of twisting it with the linear oscillator. This can be accomplished by mounting the filter in the middle of the bore of a standard spur gear. The filter must be mounted at a slight angle (the skew angle, θ_o), so that when the gear rotates, the incidence angle of the filter oscillates back and forth. Two drums are fastened to the spur gear (one on each face). The drums must each have a hole through the center to allow light to pass through the filter. The drums are then mounted on bearings inside a reamed tube. A slot is machined in the tube to allow the gear to be driven from outside the tube. (Note: normally a spur gear is mounted on a shaft through its central bore. However, in this case we need to keep the central bore open to allow light to pass. This necessitates the more elaborate mounting technique just described.) Finally, the whole system is mounted on a rotating stage so that the mean filter incidence angle (θ_f) can be swept through a range of values.

The advantage of this system is that it is less bulky than either the medium-bandwidth or wide-bandwidth setups described in the main body of this paper. It should also be just as sensitive as the medium-bandwidth setup which uses the same filter. The only reason it was not ever built was because the construction was a bit more elaborate and I had time constraints.

On the following page is shown a derivation of the angle modulation (α) as a function of the mean filter incidence angle (θ_f) and filter skew angle (θ_o). It turns out that the final equation of this derivation is approximately sinusoidal as long as θ_f is significantly bigger than θ_o. This creates the desired result which is a sinusoidal wavelength modulation with angular frequency ω (the same frequency as the angular frequency of the spinning filter).

Derivation of filter angle with respect to the optic axis (α) for a mean angle of incidence (θ_f) and a skew angle (θ_o). The filter spins inside a gear with angular frequency (ω).