Physical origin of photonic energy gaps in the propagation of surface plasmons on gratings

W. L. Barnes, T. W. Preist, S. C. Kitson, and J. R. Sambles
Department of Physics, Exeter University, Exeter, Devon, EX4 4QL, United Kingdom
(Received 21 December 1995)

We present an analytic model to describe the existence of photonic energy gaps in the propagation of surface plasmon polaritons on corrugated surfaces. We concentrate on elucidating the physical origin of the band gap, and accordingly we place strong emphasis on the physical reasoning and assumptions that we use. Our model is designed to give direct access to expressions for the electromagnetic field and surface charge distributions associated with modes at the band edges, thus allowing their physical character to be easily appreciated. Having established why a band gap occurs we then find expressions for the central position and width of the gap. We compare the results of our model for the gap width with those already in the literature, and find excellent agreement. Our results for the central position of the gap, notably the prediction that it should fall as the corrugation amplitude rises, contradicts one prediction made in the literature. We also reexamine the comparisons made in the literature between experiment and theory for the gap width, and find them inadequate because the theories have been compared to inappropriate experimental data. Consequently we present our own recent experimental data, enabling us to validate our theoretical results, in particular confirming our prediction that the central position of the gap falls as the corrugation amplitude is increased. The limitations of our model are discussed, as well as possible extensions and areas for future research. [S0163-1829(96)07433-4]

I. INTRODUCTION

Photonic materials are currently the subject of intensive and widespread study (Refs. 1, 2, and references therein). These materials are based on the interaction between an optical field and a material exhibiting periodicity on the scale of the wavelength of light. The periodicity modifies the propagation of the optical wave within the material, and under appropriate circumstances may prohibit propagation over some range of optical frequencies—a photonic band gap. Interest in such systems stems from the potential they offer to control the optical properties of materials, particularly spontaneous emission, since this has important applications in such areas as the reduction of noise in laser diodes and light-emitting diode (LED) emission.[3]

The photonic materials generally considered are bulk in nature, for example the quarter wave dielectric stack used as both mirror and filter, and the face centered cubic lattice.[4] In such systems the photon is dressed by the periodic material—this dressed state is called a polariton mode of the system. One can also consider a system that involves surface rather than bulk modes; if the surface is metallic then the relevant mode is a surface plasmon polariton (SPP) (Ref. 5) and a corrugated surface may be used to provide the periodicity. Just as in the bulk case, under appropriate conditions this periodicity may result in an energy band gap in the propagation of the surface modes. In a recent Brief Report[6] we outlined an analytic description of the photonic gap that exists for SPP's propagating on a metallic grating, and used it to show the physical origin of the band gap. In this paper we provide the detailed formulation of our theory and examine the effect of the surface profile on the gap. We also discuss the merits of our theory in comparison with previous work, discuss the theory's limitations, and present experimental data of previously untested aspects of SPP energy gaps that support our analysis.

The paper is organized as follows. In Sec. II we present a brief summary of the nature of photonic band gaps with reference to a particularly simple system, the quarter wave dielectric stack, and place our subsequent discussion of SPP energy gaps in this context. In this section we also review previous work in the field. In Sec. III we discuss the importance of the detailed nature of the surface profile on the energy gap and in particular look at the implications this has in interpreting experimental data. In Sec. IV we discuss previous theoretical work, thus setting our own in context. In Sec. V we develop in detail our analytic theory, ultimately finding expressions for the central position, the width of the gap and the field and surface charge distributions. By incorporating the results of some numerical modeling we provide a simple physical picture for the nature of the energy gap for SPP's. In Sec. VI we compare the results from our model with experimental data, and discuss the limitations of our model. Section VII provides a summary together with suggestions for future work.

II. BACKGROUND

We start by looking at the simplest periodic photonic material, the quarter wave dielectric stack; see Fig. 1.[7,8] Consider light propagating normal to the interface planes. When the optical wave vector is equal to half of the Bragg vector corresponding to the stack periodicity, Bragg scattering results in both forward and backward traveling waves that interfere constructively to set up a standing wave. We can use simple symmetry arguments to find where the standing wave is positioned with respect to the dielectric stack. If, at some point in the stack the two waves (forward and backward) are in phase the subsequently Bragg scattered waves must arrive back at this point still in phase. Since both waves will suffer an identical phase change on scattering they must travel an equal optical path length, thus requiring that the original

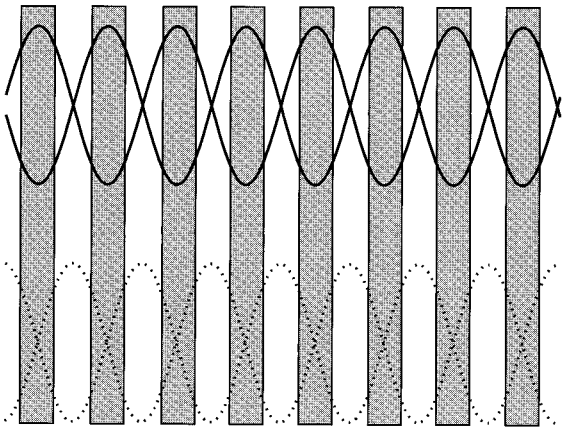

FIG. 1. A sketch of the standing waves in the dielectric stack. The boxed regions are of high-refractive index. The low frequency standing wave, top, has field extrema concentrated in the high index regions; the high frequency solution has the extrema in the low index regions. (Note that the fields drawn are only to give an indication of the distribution; in practice they will not be sinusoidal, e.g., the low frequency field will be more tightly concentrated in the high index region.)

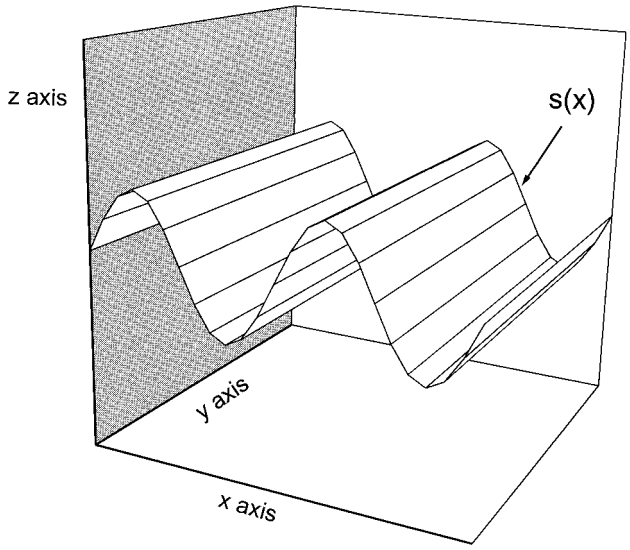

FIG. 2. Sketch of the corrugated surface considered, together with the spatial axes. The surface is described by a shape function $s(x)$.

point under consideration be in the middle of a low index region or the middle of a high index region; see Fig. 1.

The origin of an energy difference between these two standing wave configurations becomes apparent when we consider the nature of the modes involved. Light within a material is no longer just an optical field, it is now intimately linked with the optical response of the material, the mode is a polariton rather than a photon. The interaction between the optical field and the material is represented by the complex dielectric permittivity and thus the index of refraction. As we have just seen the standing wave within the stack can have two configurations, one when the standing wave has the optical field concentrated in the high index layers, the other when it is concentrated in the low index layers. The different refractive indices of the two regions mean that the two modes have different energies (and therefore frequencies) associated with them whilst still having the same periodicity—a band gap has been opened up. Frequencies between these two values, i.e., in the gap, are unable to propagate since they correspond to forward and backward traveling waves that destructively interfere within the stack. Such an energy gap is exploited in the manufacture of dielectric stacks for use as filters and mirrors since, when propagation is prohibited, an incident optical beam is totally reflected rather than transmitted.

To summarize then, if the internal optical wavelength is twice the periodicity of the stack a standing wave may be formed and two possible standing wave configurations or modes exist, having different energies. This difference arises because the optical field of the two modes are concentrated in regions of different refractive index.

We can now go on to consider the situation for surface modes; see Fig. 2. We restrict ourselves to nonradiative surface plasmon polaritons (SPP's) propagating on a corrugated surface, although other surface modes, particularly acoustic, have also been considered.[9] As with the dielectric stack, if the surface mode propagates on a corrugated surface and the mode wave vector is half the value of the grating wave vector this results in the formation of a standing wave and again the mode distribution on the surface may then take two configurations, having different energies. We may qualitatively see the origin of this energy difference once we consider the nature of SPP modes. A nonradiative SPP mode is bound to the interface between a dielectric and a metal, it consists of an electromagnetic field oscillation coupled to an oscillating surface charge density. The energy of an SPP standing wave will thus depend on the energy stored both in the electromagnetic field and the surface charge distribution. Since the two standing wave solutions take different positions with respect to the peaks and troughs of the grating (these are the analogues of the high and low index regions of the dielectric stack) it is not unreasonable to suppose that the electromagnetic field and surface charge distributions will differ in the two modes. As we shall show below, it is by considering the nature of these standing wave solutions in detail that we are able to determine quantitatively the magnitude and central position of this energy gap. The experimental existence of energy gaps, alternatively described as frequency or ω gaps, in the propagation of surface plasmons on corrugated metallic surfaces, is now well established[10–15] (note that these are sometimes referred to as minigaps in the literature since the ratio of gap width to central position has usually been small, typically 0.02).

To understand what follows we must be clear about the effect the corrugated surface has on the dispersion of the SPP mode. For an ideal metal the dispersion curve for SPP propagation on a flat surface takes on a particularly simple form, as shown in Fig. 3. If the modulation depth of the grating is small then the SPP mode wave vector, k_{SPP}, will only be perturbed by the surface modulation when it is close to half the value of the Bragg wave vector, $2K$. The Bragg wave vector is defined as $2K=2\pi/\lambda_g$ where λ_g is the pitch of corrugation; the reason for not defining it as K will become clear in Sec. III. At this value of the mode wave vector a energy gap opens up in the dispersion curve, see Fig. 3, in direct analogy with the energy gaps found in the dispersion

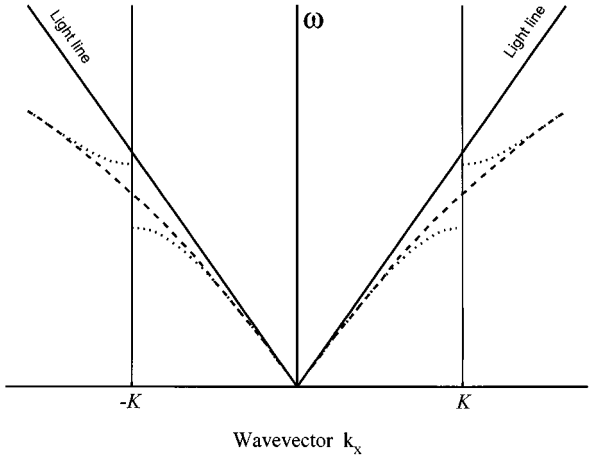

FIG. 3. The surface plasmon polariton dispersion curve. The dashed line shows the dispersion curve for a flat surface, the dotted line that for a corrugated surface. Notice how a frequency gap is opened up in the case of the corrugated surface. The gap occurs at $\pm K$, the zone boundary, the Bragg vector of the corrugation being $2K$. Also shown are the light lines. These are the dispersion curves for photons traveling at grazing incidence to the interface between the metal and the dielectric, i.e., those having the largest possible value of k_x: we see that it is not possible to couple the surface modes directly to photons; the surface modes always have more momentum than the photon of the same frequency.

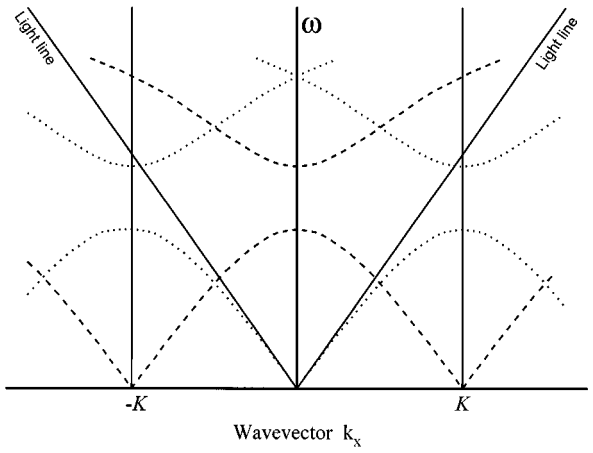

FIG. 4. The surface plasmon dispersion curve for a doubly corrugated surface, i.e., one having two grating components, one with Bragg vector $2K$, the other with Bragg vector K. The dispersion curve scattered by the $2K$ component (dotted curve) has gaps outside the light lines, and is therefore unable to couple to photons. To examine the gap experimentally a second grating component K is used. The dispersion curve scattered by the K component (dashed curve) exhibits a gap within the light lines thus allowing experimental investigation.

curve of electrons propagating in crystalline lattices and that discussed for the dielectric stack above.

It is the principle aim of this paper to explain in detail the physical origin of this gap. There was for some time concern about whether such gaps were gaps in frequency or wave vector. As we shall see below, this confusion arose from a poor interpretation of experimental data rooted in a lack of understanding of the way in which the SPP modes couple to photons. Although this problem has now been resolved in favor of frequency gaps, its legacy has persisted when comparison between experimental and theoretical work has been attempted by some authors; it is therefore important to understand the nature of the problem and see how it is resolved. To do this we need to look at how experimental data on SPP energy gaps is obtained.

III. THE INTERPRETATION OF EXPERIMENTAL DATA ON SPP ENERGY GAPS

Our aim in this section is to look at the way in which experimental data on SPP energy gaps has been obtained, with particular emphasis on how the data may be interpreted. We summarize the results of previous investigations and highlight the importance of having a detailed knowledge of the surface profile of the grating in making such interpretations; we use numerical modeling (discussed later in this paper) to emphasize the relevant points.

The observation of SPP energy gaps can be achieved by studying the resonant interaction between SPP's and radiative modes, e.g., photons. However, an important property of SPP's is that their momentum is greater than that of a free space photon of the same frequency, i.e., $k_{\mathrm{SPP}} > k_0$; photons can only access the region within the light lines of Fig. 2. Scattering from the corrugated surface provides an easy momentum matching route. The SPP's can gain or lose momentum in integer multiples of $2K$ by such scattering, thus providing coupling to photons. However, the portion of the dispersion curve near the first Brillouin zone boundary always lies outside the light lines, even after scattering from the grating, and so cannot couple to photons; see Fig. 3. A common solution to this problem is to introduce another modulation onto the surface. If the modulation has a longer pitch than the original then it may couple the energy gap region to photons; see Fig. 4.

Corrugated surfaces are commonly made by exposing photoresist to a holographic interference pattern. Nonlinearities in the exposure and development process lead to a grating profile that contains higher harmonics in addition to the fundamental. Typically only the lowest harmonic is important (the importance of the existence of higher harmonic components in the surface profile has been recognized by many authors[13,14]) and the surface profile, $s(x)$, may be represented as

$$s(x) = d_1 \sin(Kx) + d_2 \sin(2Kx + \phi_2), \tag{3.1}$$

where x is the spatial coordinate, d_1 and d_2 are the amplitude of the two harmonic components, and ϕ_2 is their relative phase. It is important to be clear on the role that the different components play. The K component of the surface modulation provides the coupling to photons whilst the $2K$ component produces the energy gap. We note that the modulation with Bragg vector K will also produce a band gap, but in a different frequency region, this fact is ignored in Fig. 4. Having thus identified the role of the two components it is clear that there is no physical requirement for them to be harmonic. We should also note that a pure sinusoid can carry out both functions, i.e., gap creation and momentum matching. Second order scattering from the K component can give

rise to a gap, however, this second order process is weak (see Sec. V) and need not be considered further in the present discussion.

We need now to consider the experimental details of the coupling between SPP's and photons, since the way the experiment is carried out has important implications for the interpretation of the data. The purpose of such experiments is to reconstruct the dispersion curve, thus allowing the width and central position of the gap to be determined. The principle method by which this has been achieved is as follows.

Light of a given frequency is incident on the grating in the plane containing the grating vector and the surface normal, at some angle θ with respect to the surface normal and the reflectivity monitored. If this angle is such that

$$k_{\mathrm{SPP}} = \pm k_0 \sin\theta \pm nK, \tag{3.2}$$

then the light may couple to the SPP mode. The reflection coefficient contains components due to specular reflection and reradiation by the SPP mode. Typically the inclusion of the SPP reradiation results in a significant reduction in the reflected intensity due to the phase difference between the specular and reradiated light.[16,17] Data are acquired by measuring the reflection coefficient as a function of incident frequency and angle. It is the interpretation of the reflection data so obtained that we wish now to concentrate on. In the following we make use of numerical modeling to explore the reflectivity under various experimental conditions. The numerical modeling is based on the same techniques that we use later in Sec. V to develop our analytic model. The details of the numerical models have been reported elsewhere.[18]

We start by mapping the reflectivity of a purely sinusoidal silver grating as a function of incident wavelength and angle; see Fig. 5. The crossing between the branches scattered by $\pm K$ occurs for light at normal incidence. For this sinusoidal profile there is no significant interaction between the two branches of the dispersion curve as shown by the absence of a gap in this case. Adding a $2K$ component to the surface profile provides an interaction between the two branches, the forward and backward traveling surface modes are coupled by this component and an energy gap opens up as shown in Fig. 6.

Experimentally there are two distinct ways in which data of this type may be obtained. One is to fix the angle of incidence and scan the reflectivity as the frequency is changed. The other is to fix the frequency and scan the angle of incidence. As has been noted by others, the two techniques can produce rather different results.[13,19] Figure 7 shows sections through Fig. 6 corresponding to the two different types of scan. The wavelength scan, Fig. 7(a), for a constant incident angle, in this case 0°, shows two clear minima indicating the presence of an energy gap. By contrast, the angle scan at a constant wavelength, Fig. 7(b) is quite different, showing only one reflectivity minimum. At first sight this is rather surprising since we chose our fixed wavelength to be in the middle of the gap. We can see why this is from Figs. 5 and 6 where, due to the finite width of the resonances, a clear saddle point exists in the gap. The minimum reflectivity in this case [Fig. 7(b)] which is only about 5% deep, represents coupling to the wings of the resonances. Thus if we scan the wavelength for a range of fixed angles we always see two minima. If we scan the angle for a range

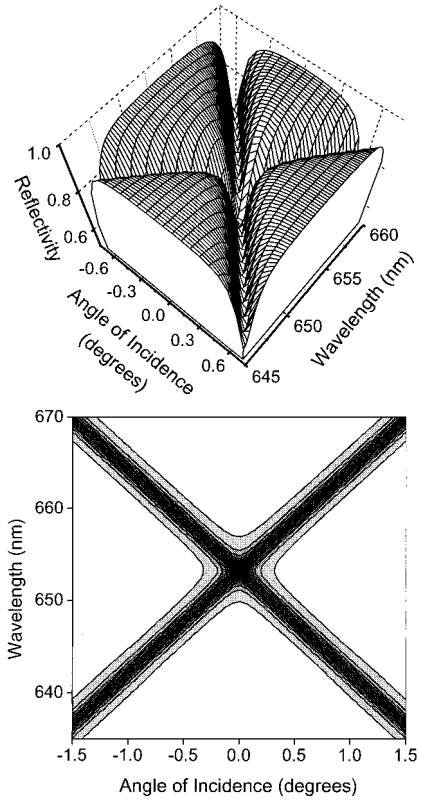

FIG. 5. Numerically modeled reflectivity of a singly corrugated surface. The upper picture shows a reflectivity surface map, reflectivity being plotted as a function of both angle of incidence and wavelength. The surface has Bragg vector K, thus allowing coupling of photons to surface modes, but not providing the scattering required to set up a gap. The lower picture shows this clearly; the reflectivity minimum is a continuous function of wavelength, there is no gap where the branches cross. (Note that in the lower picture the darker the region the lower the reflectivity.) The parameters used in the modeling were grating pitch=634 nm, amplitude d_1=5 nm; the metal parameters were fixed at $\varepsilon_r=-17.5$, $\varepsilon_i=0.7$, characteristic of those for silver in this wavelength range.

of wavelengths there are circumstances in which we will only see one minimum. Consequently, using the minima from angle scanned data will not in general yield the true dispersion curve of the SPP modes. It is better to scan the wavelength for fixed angles of incidence, or, better still, to obtain the reflectivity as a function of both so as to be able to examine the data as we have done here for the numerically modeled data presented in Fig. 4–6.

There was for some time confusion in the literature concerning the existence of k gaps, i.e., gaps that occurred in momentum rather than energy.[13,20,21] Although it has now been established[13,22–24] that there are no momentum gaps for the propagation of SPP's on grating surfaces and that the appearance of such gaps in reflectivity data is an artifact of the coupling between the SPP and a photon, it is still worth revisiting this problem as it has important consequences for the type of model that should be used in looking at SPP band gaps.

That k gaps are due simply to over coupling is easily seen by reproducing the data used to produce Fig. 5 but with an increased amplitude d_1 of the fundamental surface profile component, K; see Fig. 8. To understand the origin of this k

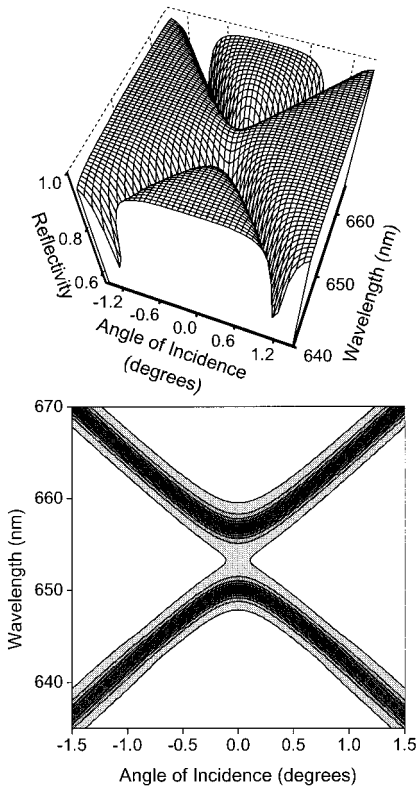

FIG. 6. As Fig. 5, except that now a second grating component has been added, d_2=2 nm, ϕ_2=0°, into the surface profile. Note how a clear energy gap has now opened up at the intersection between the two branches.

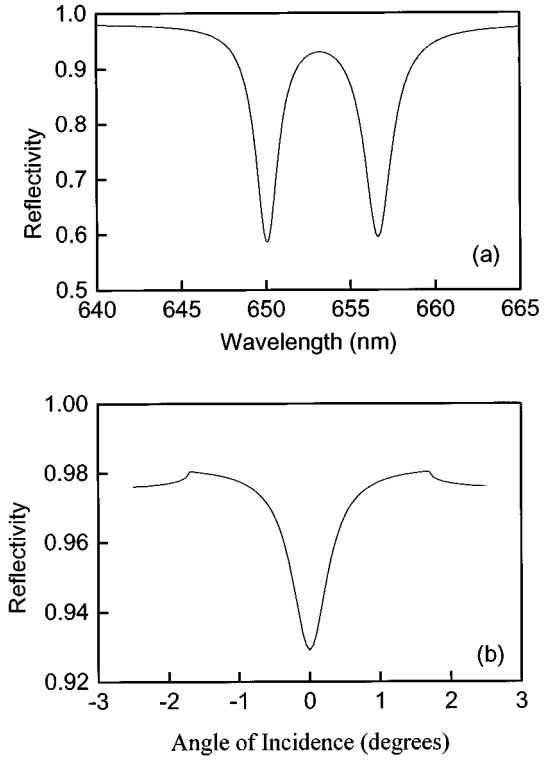

FIG. 7. Slices taken through the data of Fig. 6. The upper scan (a) shows the reflectivity that would be recorded if a wavelength scan was made for fixed incident angle through the intersection between the two branches of the dispersion curve; two minima are clearly seen. The lower scan (b) shows the reflectivity that would be recorded if an angle scan was made for a fixed wavelength, chosen to be at the center of the gap—only one minimum is seen, and there is no evidence in this type of data for an energy gap.

gap we must consider what the reflectivity we measure represents. As discussed above, the reflected light comprises a component due to specular reflection and one due to reradiated SPP emission. By increasing the depth of the grating we increase the coupling between the photon and the SPP thereby decreasing the reflected signal and increasing the reradiated. At the optimum coupling depth the reflected and reradiated components are equal in amplitude and out of phase, resulting in zero net reflectivity—100% coupling. Now suppose that we have two such perfectly coupled modes, one either side of the crossing point of Fig. 5. If we now bring these modes closer in frequency to the crossing point then the two 100% coupled modes will overlap. Clearly they cannot add to produce 200% coupling, instead an increased reflectivity is recorded. This is clearly visible on a constant wavelength scan through the middle of Fig. 8. There now appears to be a reflectivity maximum between two weaker minima and a ''momentum gap'' has appeared. The foregoing highlights the care with which reflectivity minima must be treated when trying to construct a dispersion curve. Ideally experiments should be conducted with very shallow gratings so that this type of distortion of the apparent mode position does not occur.

To complete our examination of the care with which experimental data must be acquired and analyzed we need to consider one further detail of the surface profile, the relative phase, ϕ_2, of the K and $2K$ components. For reasons of clarity this discussion will be delayed until Sec. V F.

We can summarize this section by noting that the details of the surface profile are critical in determining the reflectivity data that will be obtained and how it should be interpreted. The presence of two Fourier components are required if SPP band gaps are to be observed optically. Their magnitude and phase have a significant effect on the reflectivity, through their effect on the coupling between the SPP modes and photons, and must therefore be considered carefully in evaluating data obtained in this way. In fact, as we shall see in Sec. VI C it is possible to use prism coupling the SPP modes to photons, thus avoiding the need for a second corrugation component.

We also note at this stage that information concerning SPP band gaps can be obtained by examining emission rather than reflectivity data.[25] In this case a layer of excited molecules immediately above the metal surface lose their energy by generating an SPP mode of the appropriate frequency. We will not discuss this work further here since it introduces an extra complication into our model, i.e., it requires the inclusion of a third medium, the layer of excited molecules. We shall however return to this subject in Sec. VII when considering directions for future work.

IV. THEORETICAL APPROACHES

The goal of any model for the SPP band gap is to allow us to calculate how the gap width and central position depend upon the grating profile, in particular on the amplitude of the modulation. One approach is to set up a model for the inter-

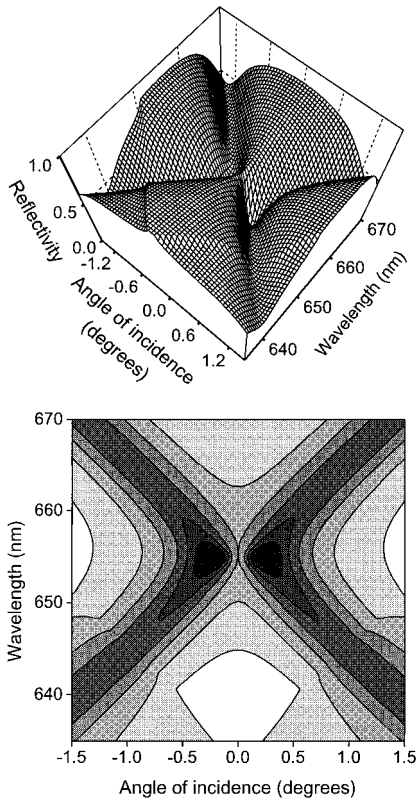

FIG. 8. As Fig. 5, except that now the grating amplitude has been increased from 5 to 30 nm. Now rather than an energy gap, we see a gap in k_x. This is thus seen to be an artifact of over coupling between SPP's and photons.

action between electromagnetic radiation and the grating, as has been undertaken by many authors,[26,27] and to then evaluate the scattering coefficients for the different diffracted orders of the reflected light, the dispersion relation can then be obtained numerically from the poles of the scattering coefficients.[28] (Zeros in the scattering coefficients could not be used since, as we have seen above, it is not always possible to identify reflectivity minima with mode solutions.) Although this method allows the numerical calculation of the gap parameters, it provides little physical insight, the very thing we wish to concentrate on here; we must therefore seek another approach, one that can lead us more directly to the dispersion relation in the vicinity of the band gap, and in particular that can improve our physical understanding.

The crucial factor determining the existence of a SPP mode is that of satisfying the electromagnetic boundary conditions across the interface. In the case of a flat surface this is straightforward,[16,29] but for a corrugated surface the determination is more involved. Many authors have addressed the problem of solving for the SPP gap parameters, and several approaches have been pursued; they can be divided into four categories, detailed below.

A. The Rayleigh method

The essential details of this method are these. Space is divided into three regions, one that is all dielectric (usually air or vacuum), another that is all metal, and a selvedge region in which both metal and dielectric exist, the composition of this latter region being periodically modulated. The electromagnetic fields in the two half spaces are constructed as Fourier sums that possess the Bloch periodicity property. The boundary conditions are satisfied by assuming that the expressions for the fields in the two half spaces are valid all the way in to the interface—this is the Rayleigh hypothesis. This approach has been discussed by many authors,[22,30,31] and can be used to obtain an analytic expression for the width of the band gap.[30] The method is however only approximate—it can only be proved to be valid for small amplitude gratings up to $d_2/\lambda_g \sim 0.1$,[32] although it appears that the range of validity is actually greater than expected from Ref. 32; see the following section.

B. Green's function method (also known as the extinction theorem method)

In this method the boundary conditions are included exactly, and analytic expressions in the form of two simultaneous matrix equations containing the Fourier coefficients of the fields[30,31,33] are obtained. These equations have been solved numerically.[33] Further, the identity of the solutions obtained by this method with those found using the Rayleigh hypothesis allows the range of validity of the Rayleigh hypothesis to be extended beyond the expected range to $d_2/\lambda_g \sim 1$. We should emphasize that agreement here is with another theory, not with experiment. Although the Green's function method has not been applied specifically to band gaps in the dispersion of SPP's it has been used to study band gaps in the propagation of Rayleigh surface waves on a grating.[9]

C. Perturbation approach

This approach takes the solution for the plane surface and expands the solutions using a standard form of perturbation theory. The perturbation approach produces not one equation for the mode solutions, but a whole hierarchy of equations, making solutions to any particular problem tedious unless limited to lowest, i.e., first, order. Such a perturbation treatment was given by Mills,[9] the use of only the first order limiting the range of validity of the technique to small amplitude gratings. The limited range of applicability of this approach was also shown by Da Silva *et al.*[34] Seshadri[35] has developed a perturbation treatment that is applicable to larger amplitude gratings although the validity of this later work must be in question as it makes predictions about the central position of the gap that are not borne out either by other theoretical work,[36] our own theoretical work or experiment—see Secs. V and VI.

Even if we could ignore the problems with the perturbation treatment of Ref. 35, none of the above techniques are ideally suited to our purpose. The reason for this is as follows. To gain physical insight we wish to study the field and charge distributions associated with the mode solutions and we would ideally like simple analytic expressions for these. In all of the above treatments the fields comprise a Fourier expansion involving the Bragg vectors. We could write an expanded field component as

$$H(x,z)=\sum_m a_m \exp(i\beta_m z)\exp(i(k_x+mK)x), \quad (4.1)$$

where β_m are the Bragg vectors and where $K=2\pi/\lambda_g$, λ_g is the pitch of the fundamental component of the surface profile. As the amplitude of the grating increases, producing a greater distortion of the fields, more terms need to be included to represent the resulting fields. We are particularly interested in the SPP modes that decay evanescently away from the corrugated interface. The field components associated with these modes will thus contain Fourier coefficients with imaginary values of β_m, given by Ref. 37,

$$\beta_m=i\left[(k_{\mathrm{SPP}}+mK)^2-\frac{\omega^2}{c^2}\right]^{1/2}. \quad (4.2)$$

Thus as the value of m rises the associated field component decays very rapidly away from the interface. Further, the β_m are independent of the surface shape. As a consequence of these two facts, the number of terms required to construct a solution to a given precision increases extremely rapidly as the depth of the grating rises. Mode solutions thus contain a large number of components and can only really be investigated numerically. It is for this reason that the above methods are not a very direct way of examining the fields associated with the mode solutions, even in the small grating amplitude regime. An alternative technique that does provide a more direct route to the fields is discussed next.

D. Chandezon approach

Here an altogether different approach developed by Chandezon *et al.*[27] is adopted. First the spatial coordinates are transformed to a system in which the surface is flat. Maxwell's equations are then expressed in these new coordinates and solved for the SPP existence and boundary conditions (we shall see what these are in the next section). Using the Chandezon framework we have developed a perturbation approach to provide analytic expressions for the band gap.[6]

The primary advantage of this approach for our purpose is that it yields eigenmode solutions for the fields which can be expanded in the form

$$H(x,z)=\sum_m \sum_q f_m^q \exp(i\lambda^q(z-s(x))\exp(i(k_x+mK)x), \quad (4.3)$$

where $s(x)$ is the shape function of the surface, i.e., *on the surface* $z=s(x)$, λ^q the eigenvalue of the mode, and f_m^q the amplitude of the mth Fourier component of the H field in the qth eigenmode; these details will be dealt with more thoroughly in the next section. This expression for the field, Eq. (4.3), already contains information on the surface profile in the first exponent. Another important point not immediately obvious at this stage (but discussed in Sec. V B) is that the eigenvalues λ^q also depend on the surface profile. This should be contrasted with the Rayleigh expansion where the equivalent of the eigenvalues are the Bragg vectors β_m which, as discussed above, are independent of the surface profile. Thus, when we need to perform the Fourier sum to evaluate the relevant fields the summation based on the Chandezon technique needs fewer terms than the summation based on the Rayleigh method (Ref. 27, Table 1 and associated text). In fact, as we shall see, a single dominant term describes the SPP standing wave on a small amplitude grating and satisfies the boundary conditions directly, i.e., only one term is needed in the summation. We are thus able to obtain simple analytic expressions for the fields associated with the SPP modes.

In summary, whilst several methods exist, we have based our approach on the method of Chandezon *et al.*[27] Although the range of validity of the model for the SPP band gap as we develop it below is no better than some of the other techniques used, notably the Rayleigh and Green's function methods, it does allow the fields and charge distributions associated with the modes to be readily obtained; the technique is thus advantageous in improving our physical insight into the mechanism behind the formation of SPP band gaps.

V. ANALYTIC THEORY

The Chandezon approach involves solving Maxwell's equations in the vicinity of a corrugated surface by making use of a coordinate transformation technique. The procedure we adopt is as follows.

(A) Set up the scheme for the mode solutions following Chandezon. We thus flatten the surface by making use of an appropriate coordinate transformation. We then express Maxwell's equations in the new coordinate system, making use of the periodicity to expand the fields as eigenmode solutions.

(B) Seek solutions to Maxwell's equations for the case when the surface modulation Bragg scatters the SPP mode, assuming that we need only consider first order Bragg scattering.

(C) Apply the boundary conditions appropriate for the existence of SPP's on a grating.

(D) Examine the field distributions and surface charge density, thus identifying the physical origin of the gap.

(E) Derive expressions for the central position and gap width.

(F) Look at the importance of the phase of the grating in the context of optical examination of the SPP band gap.

(G) Examine the effect of refining the model to include higher order terms.

A. The Chandezon technique

The crucial factor determining the existence of a surface plasmon excitation on a flat surface is the necessity of satisfying the electromagnetic boundary conditions across the interface. If the interface is nonplanar then the solution of Maxwell's equations in rectangular coordinates is not a great deal of help in ensuring that the corresponding boundary conditions are exactly satisfied, instead we adopt the method of Chandezon *et al.*[27] The detail of their method is set out in their paper; in our case we are particularly interested in the transverse magnetic (TM polarized) solutions since the SPP mode is TM polarized; an **E** field normal to the metal surface is required to generate the surface charge density variations that constitute a component of the SPP mode. Further, we consider only that surface profile component that gives direct rise to the SPP band gap, i.e., the component with Bragg vector $2K$.

In essence the method used by Chandezon *et al.* proceeds in a number of stages.

(i) The coordinate system is transformed from x,y,z to

$$u=z-s(x),$$
$$v=x, \tag{5.1}$$
$$w=y,$$

where $s(x)$ is a periodic function defining the grating surface. In our case the profile of the grating is given by Eq. (3.1) with d_1 set to zero; thus $s(x)=d_2\sin(2Kx+\phi_2)$. Note that we retain the phase term ϕ_2 for future use even though it is not strictly required here. The position of the interface between the regions of relative permittivity ε_1 and ε_2 is defined as $u=0$; ε_2 is assumed to be real and negative, i.e., metallic.

(ii) For convenience the field variables **E** and **H** are reexpressed as **F** and **G**, where **F** and **G** contain the field components in a form appropriate to the new coordinate system. **F** and **G** are defined so that **F** contains the field components perpendicular to the symmetry plane, i.e., the y or w component, whilst **G** is related to the field component in the symmetry plane and tangential to the local surface. Expressing the fields in this way simplifies the application of the boundary conditions, i.e., that the local tangential field components are continuous across the corrugated interface. For TM polarization **F** and **G** take the form

$$F=Z_0H_y,$$
$$G=-k_0\varepsilon E_{\|}\sqrt{1+s'^2}, \tag{5.2}$$

where

$$Z_0=\left(\frac{\mu_0}{\varepsilon_0}\right)^{1/2}, \quad s'=\frac{\partial s}{\partial v}, \quad k_0=\frac{\omega}{c} \tag{5.3}$$

$E_{\|}$ is the component of the electric field locally parallel to the surface, and H_y is the component of the magnetic field normal to the symmetry plane.

(iii) Maxwell's equations, expressed in terms of **F** and **G** in the new coordinate system u,v,w give, for the TM polarized case under consideration,

$$\frac{\partial F}{\partial u}=\frac{s'}{1+s'^2}\frac{\partial F}{\partial v}+\frac{iG}{1+s'^2},$$
$$\frac{\partial G}{\partial u}=ik_0^2\varepsilon F+\frac{\partial}{\partial v}\left[\frac{s'}{1+s'^2}G\right]+\frac{\partial}{\partial v}\left[\frac{i}{1+s'^2}\frac{\partial F}{\partial v}\right], \tag{5.4}$$

where ε is the relative permittivity of the appropriate medium. Once these two coupled equations are solved we will have found the tangential components of the E and H fields, i.e., F and G, from which the normal E-field component can be found by differentiation.

(iv) In finding the solutions to the above equations we recognize the periodicity of the fields in the v variable and express F and G as Fourier expansions of the form

$$F(u,v)=\sum_m F_m(u)\exp(i\alpha_m v),$$
$$G(u,v)=\sum_m G_m(u)\exp(i\alpha_m v), \tag{5.5}$$

where

$$\alpha_m=k_x+mK, \quad m=0,\pm1,\pm2,\ldots, \tag{5.6}$$

and k_x is the x component of the wave vector of the mode under consideration.

(v) Substitution of these expansions [Eqs. (5.5)] into the new form of Maxwell's Eqs. (5.4) yields an infinite set of equations for $F_m(u)$ and $G_m(u)$ that can be written in the form

$$-i\frac{d\zeta(u)}{du}=\mathbf{T}\zeta(u), \tag{5.7}$$

where **T** is a matrix of infinite size that is independent of u and $\zeta(u)$ is a column vector of the form $\zeta(u)=(F_{+N},F_{+N-1},\ldots,F_{-N},G_{+N}/\varepsilon,G_{+N-1}/\varepsilon,G_{-N}/\varepsilon)$, i.e., in the limit $N\to\infty$ it contains the field components from the Fourier expansion (5.5).

(vi) The normal mode solutions have a u dependence of the form $\exp(i\lambda^q u)$, and we thus write

$$\zeta^q(u)=\psi^q\exp(i\lambda^q u). \tag{5.8}$$

The normal mode solutions can be found by extracting the eigenvalues of **T**, i.e., λ^q and the associated eigenvectors, ψ^q, from

$$(\mathbf{T}-\lambda^q\mathbf{I})\psi^q=0, \tag{5.9}$$

where λ^q are the eigenvalues of **T**.

Writing

$$\psi^q=\begin{pmatrix} f_m^q \\ g_m^q \end{pmatrix}, \tag{5.10}$$

then

$$F_m^q(u)=f_m^q\exp(i\lambda^q u),$$
$$G_m^q(u)=g_m^q\exp(i\lambda^q u). \tag{5.11}$$

We can now express the mode solutions in the form quoted in Sec. IV D. First we find the **F** vector corresponding to the mode solution by substituting (5.11) into (5.5) to give

$$F^q(u,v)=\sum_m f_m^q\exp(i\lambda^q u)\exp(i\alpha_m v). \tag{5.12}$$

For TM polarization $F(u,v)=Z_0H(x,z)$; see Eq. (5.2). If we further convert back to the x,y,z coordinate system and substitute for α_m through Eq. (5.6) then we find

$$H(x,z)=\sum_m\sum_q f_m^q\exp[i\lambda^q(z-s(x))]\exp(i(k_x+mK)x), \tag{5.13}$$

which is identical with Eq. (4.3) (ignoring the constant Z_0).

(vii) To obtain a solution to Eq. (5.9) we must truncate the Fourier expansion at some finite value of m, i.e., $-N \leqslant m \leqslant N$.

B. Solutions that are coupled by Bragg scattering

We are considering the case in which there is a degeneracy between the right and left traveling SPP's, having wave vectors mK where $m=\pm 1$, i.e., $k_{\mathrm{SPP}}=\pm K$. [Note that this situation corresponds to $k_x=0$ in Eq. (5.6), i.e., a photon coupled to this SPP mode would have to propagate normal to the surface.] This degeneracy can be removed by coupling the two modes via the surface Fourier components that have wave vectors $\pm 2K$. The dominant effect occurs through the coupling of the $m=\pm 1$ components via this $\pm 2K$ term in the surface profile and is analogous to the degenerate perturbation theory calculation in quantum mechanics. In the following we will assume that the splitting is dominated by the coupling of the $m=\pm 1$ terms and will omit all other Fourier components associated with the plasmon.

The $\mathbf{T}$ matrix in the eigenvalue equation (5.9), is now a 4×4 matrix given by

$$\mathbf{T}=\begin{bmatrix} KD_0 & -KD_2 & C_0 & C_2 \\ KD_{-2} & -KD_0 & C_{-2} & C_0 \\ -K^2C_0+\varepsilon\left(\frac{\omega}{c}\right)^2 & KC_2 & KD_0 & KC_2 \\ KC_{-2} & -K^2C_0+\varepsilon\left(\frac{\omega}{c}\right)^2 & -KD_{-2} & -KD_0 \end{bmatrix}. \tag{5.14}$$

C_m and D_m are defined by

$$\frac{1}{1+s'^2}=\sum_m C_m \exp(imKv), \qquad \frac{s'}{1+s'^2}=\sum_m D_m \exp(imKv), \tag{5.15}$$

with

$$s'=\frac{\partial s}{\partial v}=2Kd_2\cos(2Kv+\phi_2). \tag{5.16}$$

Notice that the shape of the surface only enters the matrix $\mathbf{T}$ through the coefficients of C and D. We now assume that for our lowest order scattering we need only retain terms to order $(Kd_2)^3$; this assumption is discussed further in Sec. V G. Equation(5.15) then becomes

$$C_0=1-2(Kd_2)^2\equiv 1-\rho, \qquad C_{\pm 2}=0, \qquad D_0=0, \tag{5.17}$$

$$D_{\pm 2}=Kd_2(1-3\rho/2)\exp(\pm i\phi_2)=\xi\,\exp(\pm i\phi_2)/K,$$

with ρ and ξ defined as

$$\rho=2(Kd_2)^2, \qquad \xi=K^2d_2(1-3\rho/2). \tag{5.18}$$

The eigenvalues of $(\mathbf{T}-\lambda\mathbf{I})$ are then found to be

$$\lambda^2=\varepsilon\left(\frac{\omega}{c}\right)^2(1-\rho)-K^2(1-\rho)^2-\xi^2 \quad \text{(twice)}. \tag{5.19}$$

The eigenvectors are found by substituting Eq. (5.17) into (5.9). There is some flexibility in the choice of the two independent eigenvectors due to the degeneracy of the eigenvalues; see Eq. (5.19). We choose to express them in the form shown below since, as we shall see in Sec. V D, this choice gives quickest access to the physics involved,

$$\psi_\sigma=\begin{pmatrix} 1 \\ e^{-i\phi_2} \\ \dfrac{\lambda+\xi}{1-\rho} \\ \dfrac{\lambda-\xi}{1-\rho}\,e^{-i\phi_2} \end{pmatrix}, \qquad \psi_\tau=\begin{pmatrix} 1 \\ -e^{-i\phi_2} \\ \dfrac{\lambda-\xi}{1-\rho} \\ \dfrac{-\lambda-\xi}{1-\rho}\,e^{-i\phi_2} \end{pmatrix}. \tag{5.20}$$

With suitable normalization factors (see later) the first two elements determine the amplitudes of the tangential $H_{m=\pm 1}$ fields and the second two determine the amplitudes of the tangential $E_{m=\pm 1}$ fields. Our task in the next section is to find those combinations of the above eigenvectors that represent solutions to the situation under consideration, i.e., under the appropriate boundary conditions.

C. The boundary conditions

In each region, i, on either side of the interface, the surface mode solution is a mixture of the two eigenvectors, i.e.,

$$\psi^i=(\psi^i_\sigma+\mu_i\psi^i_\tau)b^i, \tag{5.21}$$

where b^i is an overall amplitude and μ_i is the relative strength of ψ_τ to ψ_σ in the mixture. In addition, λ must be chosen appropriately to give solutions that decay away from the interface.

At the boundary $u=0$ and the matching conditions on the tangential H components derived from Eqs. (5.5) and (5.11) give

$$b^1(1+\mu_1)=b^2(1+\mu_2),$$
$$b^1(1-\mu_1)=b^2(1-\mu_2), \tag{5.22}$$

so that $b^1=b^2$ and $\mu_1=\mu_2=\mu$.

Similarly, noting that $E=G/\varepsilon$, the matching conditions on the tangential E components yield

$$\frac{1}{\varepsilon_1}(\lambda_1+\xi)+\frac{\mu}{\varepsilon_1}(\lambda_1-\xi)=\frac{1}{\varepsilon_2}(\lambda_2+\xi)+\frac{\mu}{\varepsilon_2}(\lambda_2-\xi), \tag{5.23}$$
$$\frac{1}{\varepsilon_1}(\lambda_1-\xi)+\frac{\mu}{\varepsilon_1}(-\lambda_1-\xi)=\frac{1}{\varepsilon_2}(\lambda_2-\xi)+\frac{\mu}{\varepsilon_2}(-\lambda_2-\xi).$$

These coupled equations replace the two (identical) boundary conditions,

$$\frac{\lambda_1}{\varepsilon_1}=\frac{\lambda_2}{\varepsilon_2}, \tag{5.24}$$

that apply for a planar surface ($\xi=0$). As mentioned above, the surface plasmon solution must have fields that decay away from the interface, so that

$$\lambda_1=i\eta_1, \quad \lambda_2=-i\eta_2, \tag{5.25}$$

where, using Eq. (5.19),

$$\eta_i(\omega)=\left[K^2(1-\rho)^2+\xi^2-\varepsilon_i\left(\frac{\omega}{c}\right)^2(1-\rho)\right]^{1/2}. \tag{5.26}$$

The solution of the boundary conditions, Eq. (5.23) together with Eq. (5.25) has two solutions,

$$\mu=\mp i. \tag{5.27}$$

Further, we find that

$$\frac{\eta_1^{\pm}}{\varepsilon_1}+\frac{\eta_2^{\pm}}{\varepsilon_2}=\mp\xi\left(\frac{1}{\varepsilon_1}-\frac{1}{\varepsilon_2}\right), \tag{5.28}$$

where

$$\eta_i^{\pm}\equiv\eta_i(\omega_{\pm}). \tag{5.29}$$

Combining Eqs. (5.26–5.29) gives

$$\left(\frac{\omega_{\pm}}{c}\right)^2(1-\rho)=K^2(1-\rho)^2\left(\frac{1}{\varepsilon_1}+\frac{1}{\varepsilon_2}\right)+\frac{2\xi^2}{\varepsilon_1}\pm\frac{2\xi}{\varepsilon_1}\eta_1^{\pm}$$
$$=K^2(1-\rho)^2\left(\frac{1}{\varepsilon_1}+\frac{1}{\varepsilon_2}\right)+\frac{2\xi^2}{\varepsilon_2}\mp\frac{2\xi}{\varepsilon_2}\eta_2^{\pm}. \tag{5.30}$$

In deriving the above we find that $\omega=\omega_{\pm}$ for $\mu=\mp i$. That the two different solutions have different energies is now clear. We could now proceed immediately to find algebraic expressions for the gap width and central position; however, we prefer at this stage to develop expressions for the field distributions since it is these that provide the insight into the origin of the gap.

D. Field distributions, surface charge density, and mode energy

In this section we find explicit expressions for the field distributions and the surface charge density. We then use these expressions to evaluate the electromagnetic energy associated with the modes. Apart from an overall normalization factor the tangential H and E field amplitudes, which we write as $H_{\|}$ and $E_{\|}$, are obtained by dividing F by $Z_0=(\mu_0/\varepsilon_0)^{1/2}$ and G by $(-\omega\varepsilon/c)(1+s'^2)^{1/2}$. Hence, using Eq. (5.5), together with (5.9), (5.20), and (5.21), we find that at the surface

$$H_{\|}^{\pm}=A((f_1^{\sigma}\mp if_1^{\tau})e^{iKv}+(f_{-1}^{\sigma}\mp if_{-1}^{\tau})e^{-iKv})/Z_0$$
$$=H_0\cos(Kv+\phi_2/2\mp\pi/4), \tag{5.31}$$

with $H_0=2\sqrt{2}A/Z_0$. We also find

$$E_{\|}^{\pm}=-A((g_1^{\sigma}\mp ig_1^{\tau})e^{iKv}$$
$$+(g_{-1}^{\sigma}\mp ig_{-1}^{\tau})e^{-iKv})\frac{c}{\omega_{\pm}\varepsilon}\left(\frac{1}{(1+s'^2)^{1/2}}\right)$$
$$=-\frac{(\lambda^{\pm}\pm i\xi)c}{\varepsilon(1-\rho)\omega_{\pm}}\frac{Z_0H_0}{(1+s'^2)^{1/2}}\cos(Kv+\phi_2/2\mp\pi/4). \tag{5.32}$$

In evaluating this expression it is important to recall Eq. (5.25), i.e.,

$$\lambda_1^{\pm}=i\eta_1(\omega_{\pm}), \quad \lambda_2^{\pm}=-i\eta_2(\omega_{\pm}). \tag{5.33}$$

The component of E, normal to the surface, $E_N^{\pm}$ may be deduced from $H_{\|}^{\pm}$ by differentiation, viz.,

$$E_N^{\pm}=\frac{i}{(1+s'^2)^{1/2}}\frac{Z_0c}{\varepsilon\omega_{\pm}}\frac{\partial H_{\|}^{\pm}}{\partial v}$$
$$=-\frac{i}{\varepsilon}\frac{Kc}{\omega_{\pm}}\frac{Z_0H_0}{(1+s'^2)^{1/2}}\sin(Kv+\phi_2/2\mp\pi/4). \tag{5.34}$$

The expressions (5.31), (5.32), and (5.34) determine the fields on the boundary $u=0$, whilst the surface charge density, $\sigma_{\pm}$, can be found from $E_N^{\pm}$ since

$$\sigma^{\pm}=\varepsilon_0(E_{N_1}^{\pm}-E_{N_2}^{\pm})$$
$$=-i\frac{K}{\omega_{\pm}}\left(\frac{1}{\varepsilon_1}-\frac{1}{\varepsilon_2}\right)\frac{H_0}{(1+s'^2)^{1/2}}\sin(Kv+\phi_2/2\mp\pi/4), \tag{5.35}$$

where $E_{N_1}^{\pm}$ and $E_{N_2}^{\pm}$ are the normal field components at the surface in the two media. We have also used the fact that $\varepsilon_0cZ_0=1$. We can summarize the spatial dependence of the fields and surface charge density along the surface as

$$H_{\|}^{\pm},E_{\|}^{\pm}\propto\cos(Kv+\phi_2/2\mp\pi/4),$$
$$E_N^{\pm},\sigma^{\pm}\propto\sin(Kv+\phi_2/2\pm\pi/4). \tag{5.36}$$

It is from these expressions that we can see the origin of the energy gap between the two modes. The extrema of the nor-

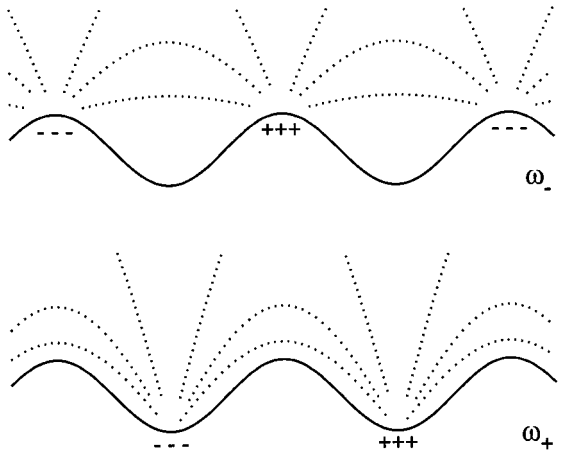

FIG. 9. Sketch of the field (E) and surface charge distributions for the two standing wave solutions at the gap boundaries. The upper sketch is for the low frequency solution, the lower sketch is for the high frequency solution. Notice that the field lines are more distorted in the lower sketch, illustrating the greater energy stored in the fields by this mode.

mal field component and surface charge distribution for the high frequency solution occur at the troughs of the surface component that has periodicity $2K$, whereas for the low frequency solution they occur at the peaks. These distributions are illustrated in Fig. 9, from which it is clear that, owing to the relative distortion of the fields between the two solutions and the associated difference in location of the surface charge, a different energy will be associated with the two distributions. Before finding an expression for this energy difference, we can gain further physical insight by considering the decay lengths of the modes into the surrounding media.

Using Eqs. (5.8) and (5.25) we see that away from the surface the fields are modulated by the factors $\exp(-\eta_1^{\pm}u)$ in region 1 ($u>0$) and by $\exp(\eta_2^{\pm}u)$ in region 2 ($u<0$); $\eta_1^{\pm}$ and $\eta_2^{\pm}$ are thus the inverse decay lengths of the modes away from the interface. Combining Eqs. (5.26), (5.29), and (5.30), we find the following expressions for them:

$$\eta_1^{\pm}=K(1-\rho)\left[-\frac{\varepsilon_1}{\varepsilon_2}\right]^{1/2}\mp\xi,$$
$$\eta_2^{\pm}=K(1-\rho)\left[-\frac{\varepsilon_2}{\varepsilon_1}\right]^{1/2}\pm\xi. \tag{5.37}$$

Thus comparing with $\eta_i^0=\eta_i(d_2=0)$, the decay lengths for the flat surface, (i) for the ω_+ solution,

$$\eta_1^{+}<\eta_1^0, \quad \eta_2^{+}>\eta_2^0, \tag{5.38}$$

so that this plasmon field distribution is ''shifted'' to the dielectric side of the interface whilst (ii) for the ω_- solution

$$\eta_1^{-}>\eta_1^0, \quad \eta_1^{+}<\eta_1^0; \tag{5.39}$$

that is the plasmon field distribution is ''shifted'' towards the metal side of the interface. These shifts in distribution are depicted in Fig. 10, showing again how the two modes differ.

Using the expressions (5.37) for $\eta_1^{\pm}$ the expression for $E_{\parallel}^{\pm}$ simplifies to

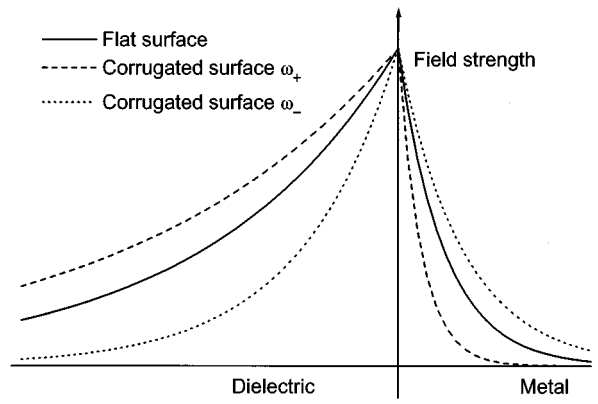

FIG. 10. A plot of the way in which the fields decay away from the interface. Decays are shown for the flat surface, and for the two solutions (high frequency dashed, low frequency dotted) of the corrugated surface. Notice how the field in the dielectric for the high frequency solution becomes less well confined to the interface, as one might expect since this branch is pushed closer to the light line; see Fig. 3.

$$E_{\parallel}^{+}=-i\,\frac{Kc}{\omega_{\pm}}\left(-\frac{1}{\varepsilon_1\varepsilon_2}\right)^{1/2}\frac{Z_0H_0}{(1+s'^2)^{1/2}}\cos(Kv+\phi_2/2 \mp\pi/4), \tag{5.40}$$

which is valid on both sides of the interface (as required by the matching condition).

The total electromagnetic energy associated with the modes is comprised of that stored in the fields on either side of the interface and the energy associated with the surface charge distribution. The latter surface energy density is given by

$$\Sigma_S=\frac{1}{2\lambda_g}\int\sigma V\,dl, \tag{5.41}$$

where the integration is carried out along the surface for a full period $\lambda_g(=2\pi/K)$. Hence

$$\Sigma_S=\frac{K}{2(2\pi)}\int_0^{\lambda_g}\sigma V(v)(1+s'^2)^{1/2}dv, \tag{5.42}$$

where

$$V(v)=-\int E_{\parallel}dl=-\int_0^{v}E_{\parallel}(1+s'^2)^{1/2}dv. \tag{5.43}$$

Substitution of the relevant, previously derived, expressions [Eqs. (5.35) and (5.40)] yields the result that the time averaged surface energy density is

$$\overline{\Sigma_S^{\pm}}=-\frac{1}{8}\left(\frac{Kc}{\omega_{\pm}}\right)^2\frac{\mu_0H_0^2}{K}\left(\frac{1}{\varepsilon_1}-\frac{1}{\varepsilon_2}\right)\left(-\frac{1}{\varepsilon_1\varepsilon_2}\right)^{1/2}. \tag{5.44}$$

The corresponding time averaged energies stored in the fields per unit (flat) surface area is

$$\overline{\Sigma_E^{\pm}}=\frac{1}{2\lambda_g}\int_{-\infty}^{\infty}du\int_0^{\lambda_g}\varepsilon_0\varepsilon[|\overline{E_{\parallel}^{+}}|^2+|\overline{E_N^{\pm}}|^2]dv$$
$$=\frac{\mu_0H_0^2}{16}\left(\frac{Kc}{\omega_{\pm}}\right)^2(1-\rho)\left(\frac{1}{\varepsilon_1}-\frac{1}{\varepsilon_2}\right)\left(\frac{1}{\eta_1^{\pm}}-\frac{1}{\eta_2^{\pm}}\right) \tag{5.45}$$

and, similarly,

$$\overline{\Sigma_H^{\pm}}=\frac{1}{2\lambda_g}\int_{-\infty}^{\infty}du\int_0^{\lambda_g}\mu_0|\overline{H_{\parallel}^{\pm}}|^2dv$$
$$=\frac{\mu_0H_0^2}{16}\left(\frac{1}{\eta_1^{\pm}}+\frac{1}{\eta_2^{\pm}}\right). \tag{5.46}$$

These appear to differ but algebraic manipulation using the relationship between $\omega_{\pm}$ and $\eta_1^{\pm}$ previously derived [Eq. (5.26)] shows that $\overline{\Sigma_H^{\pm}}=\overline{\Sigma_E^{\pm}}$ and so the total electromagnetic field energy per unit area is then $2\overline{\Sigma_E^{\pm}}$.

Hence the total energy per unit area associated with the modes is

$$\overline{\Sigma^{\pm}}=\frac{\mu_0H_0^2}{8K}\left(\frac{Kc}{\omega_{\pm}}\right)^2\left(\frac{1}{\varepsilon_1}-\frac{1}{\varepsilon_2}\right)\left[\left(\frac{K}{\eta_1^{\pm}}-\frac{K}{\eta_2^{\pm}}\right)(1-\rho)-\left(\frac{1}{-\varepsilon_1\varepsilon_2}\right)^{1/2}\right]. \tag{5.47}$$

After further manipulation it is possible to show that for small ξ

$$\overline{\Sigma^{+}}-\overline{\Sigma^{-}}\propto\left(\frac{\omega_+}{Kc}\right)^2-\left(\frac{\omega_-}{Kc}\right)^2, \tag{5.48}$$

as one might expect.

Although this section has allowed us to examine the physical origin of the band gap it yields little that allows us to test our model against experiment. In the next section we derive expressions for the dependence of the gap width and central position on the surface profile that allow such tests to be made.

E. Expressions for the central position and gap width

We first define the parameters for which we shall find analytic expressions since they are not immediately obvious if the above discussion is not familiar. Following the mode solutions derived in Sec. V C, in particular Eq. (5.30), we define the *normalized gap width* to be

$$\left[\left(\frac{\omega_+}{c}\right)^2-\left(\frac{\omega_-}{c}\right)^2\right]$$

and the *normalized central position* to be

$$\frac{1}{2}\left[\left(\frac{\omega_+}{c}\right)^2+\left(\frac{\omega_-}{c}\right)^2\right].$$

Later, when we need the more conventional definition of gap width, i.e., $\delta\omega=\omega_+-\omega_-$, we shall derive it from our expression for the normalized gap width. With these definitions we find from Eq. (5.30) that

$$\left[\left(\frac{\omega_+}{c}\right)^2-\left(\frac{\omega_-}{c}\right)^2\right](1-\rho)=\frac{2\xi}{\varepsilon_1}(\eta_1^{+}+\eta_1^{-})=-\frac{2\xi}{\varepsilon_2}(\eta_2^{+}+\eta_2^{-}) \tag{5.49}$$

and

$$\frac{1}{2}\left[\left(\frac{\omega_+}{c}\right)^2+\left(\frac{\omega_-}{c}\right)^2\right](1-\rho)$$
$$=K^2(1-\rho)^2\left(\frac{1}{\varepsilon_1}+\frac{1}{\varepsilon_2}\right)+\frac{2\xi^2}{\varepsilon_1}+\frac{\xi}{\varepsilon_1}(\eta_1^{+}-\eta_1^{-})$$
$$=K^2(1-\rho)^2\left(\frac{1}{\varepsilon_1}+\frac{1}{\varepsilon_2}\right)+\frac{2\xi^2}{\varepsilon_2}+\frac{\xi}{\varepsilon_2}(\eta_2^{+}-\eta_2^{-}). \tag{5.50}$$

Further manipulation shows that

$$\eta_1^{+}-\eta_1^{-}=-(\eta_2^{+}-\eta_2^{-})=-2\xi, \tag{5.51}$$

so that

$$\frac{1}{2}\left[\left(\frac{\omega_+}{c}\right)^2+\left(\frac{\omega_-}{c}\right)^2\right]=K^2(1-\rho)\left(\frac{1}{\varepsilon_1}+\frac{1}{\varepsilon_2}\right)=\left(\frac{\omega_0}{c}\right)^2(1-\rho)$$
$$=\left(\frac{\omega_0}{c}\right)^2(1-2(Kd_2)^2), \tag{5.52}$$

where

$$\left(\frac{\omega_0}{c}\right)^2=K^2\left(\frac{1}{\varepsilon_1}+\frac{1}{\varepsilon_2}\right). \tag{5.53}$$

Also

$$\frac{(1-\rho)^2}{4\xi^2}\left[\left(\frac{\omega_+}{c}\right)^2-\left(\frac{\omega_-}{c}\right)^2\right]^2+\left(\frac{2\xi}{\varepsilon_1}\right)^2=\left(\frac{\eta_1^{+}}{\varepsilon_1}+\frac{\eta_1^{-}}{\varepsilon_1}\right)^2+\left(\frac{\eta_1^{+}}{\varepsilon_1}-\frac{\eta_1^{-}}{\varepsilon_1}\right)$$
$$=2\left(\frac{\eta_1^{+}}{\varepsilon_1}\right)^2+2\left(\frac{\eta_1^{-}}{\varepsilon_1}\right)^2$$
$$=2\left\{2\frac{K^2(1-\rho)^2}{\varepsilon_1^2}+\frac{2\xi^2}{\varepsilon_1^2}-\left[\left(\frac{\omega_+}{c}\right)^2+\left(\frac{\omega_-}{c}\right)^2\right]\left(\frac{1-\rho}{\varepsilon_1}\right)\right\}$$
$$=4\frac{K^2(1-\rho)^2}{\varepsilon_1^2}+\frac{4\xi^2}{\varepsilon_1^2}-4\left(\frac{\omega_0}{c}\right)^2\frac{(1-\rho)^2}{\varepsilon_1} \tag{5.54}$$

giving, for the normalized gap width,

$$\left(\frac{\omega_+}{c}\right)^2-\left(\frac{\omega_-}{c}\right)^2=4\xi\sqrt{-K^2/\varepsilon_1\varepsilon_2}$$

$$=4(Kd_2)\,\frac{K^2}{\sqrt{-\varepsilon_1\varepsilon_2}}\,(1-3(Kd_2)^2). \tag{5.55}$$

Further, using Eq. (5.52), we find that

$$\left(\frac{\omega_\pm}{c}\right)^2=\left(\frac{\omega_0}{c}\right)^2(1-2(Kd_2)^2)$$

$$\pm 2(Kd_2)\,\frac{K^2}{\sqrt{-\varepsilon_1\varepsilon_2}}\,(1-3(Kd_2)^2). \tag{5.56}$$

In this expression ε_1 and ε_2 should be interpreted as the local values of relative permitivities at $\omega_\pm$ as appropriate. If the variation in their values over the range of the band gap is unimportant then the equations above represent a solution to the problem. If the frequency variation is significant then it remains to solve each equation self-consistently for ω_+ or ω_- given a functional form for $\varepsilon_1(\omega)$ and $\varepsilon_2(\omega)$. We can now express the central position as

$$\frac{\overline{\omega^2}}{c^2}=\frac{1}{2}\left[\left(\frac{\omega_+}{c}\right)^2+\left(\frac{\omega_-}{c}\right)^2\right]=\left(\frac{\omega_0}{c}\right)^2[1-2(Kd_2)^2]. \tag{5.57}$$

We now have expressions for the normalized gap width and the normalized central position, Eqs. (5.55) and (5.57). These expressions will be compared with experimental data in Sec. VI. In examining these two equations we note two important facts in the small modulation limit, i.e., $2Kd_2 \ll 1$.

First, the gap width, $\delta\omega$, is linear in modulation amplitude, d_2. This is not at first sight obvious from Eq. (5.55), but if the normalized gap width is reexpressed in terms of $\delta\omega$ then we find $\delta\omega \propto d_2$. We leave this derivation until Sec. VI A so that we can include the higher order terms of Sec. V G in our model. Having already examined the field and surface charge distributions it is clear why $\delta\omega \propto d_2$. As the modulation depth increases, so the distortion of the fields, and thus the energy associated with them will also increase. To a first order approximation the frequency difference $\delta\omega$ will therefore be linear in d_2.

Secondly, we see from Eq. (5.57) that the central position falls as the corrugation amplitude, d_2, is increased. The physical reason lying behind this fall is seen by considering the energy associated with each mode. As we have seen the energy is a consequence of the field and surface charge distributions. In particular, the high energy solution has fields that extend further into the half space above the metal; see Fig. 10. This mode thus becomes less well bound to the surface as the grating modulation increases. This change is limited since the greatest distortion with respect to the planar situation occurs when the fields extend without decay into the half space above the metal, i.e., the high frequency branch approaches the light line (Fig. 3). There is thus an upper limit on the frequency of this mode. The low frequency branch is not affected in this way and consequently the central position of the two modes falls as the grating modulation is increased.

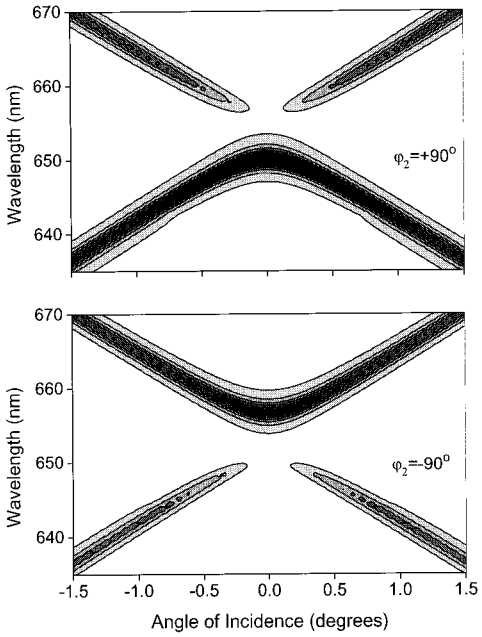

FIG. 11. Numerically modeled reflectivity plots showing the effect of the phase ϕ_2 between the two grating components. Although the size of the gap is unaffected by the phase, the coupling is.

F. The phase of the grating

As we mentioned at the end of Sec. III, it is important to consider the role of the relative phase, ϕ_2, between the fundamental (the K grating) and first harmonic component (the $2K$ grating). In Sec. V E we showed that the frequencies $\omega_\pm$ are independent of the phase of the $2K$ grating. However, phase does become important when we consider the coupling of SPP modes at the edge of the band gap to photons.

We can best illustrate this by numerically examining the reflectivity of a corrugated surface with the two grating components present with various values of ϕ_2. Figure 11 shows such reflectivity contour plots for the same grating as Fig. 6, but with $\phi_2=\pm 90°$. Both cases still exhibit energy gaps of the same magnitude and position as before (Fig. 5), but the coupling strength of one of the branches is now reduced to zero. Whilst the importance of the relative phase of the surface components in determining the coupling strength of the two modes has been known for some time,[14,38,39] it is worth applying our analytic model to uncover the physics involved.

As discussed in Sec. III, for a corrugated surface possessing grating components K and $2K$, photons normally incident on the surface will couple to the SPP modes at the edges of the band gap. For coupling between the photons and the standing wave to occur there must be some component of the incident optical field normal to the surface, at the appropriate points on the surface, to generate the surface charges necessary for the standing wave; see Fig. 12. For normal incidence there will be no component of the optical field normal to the surface where the surface has zero gradient. As Fig. 12 shows, when $\phi_2=90°$, the troughs of the $2K$ component cor-

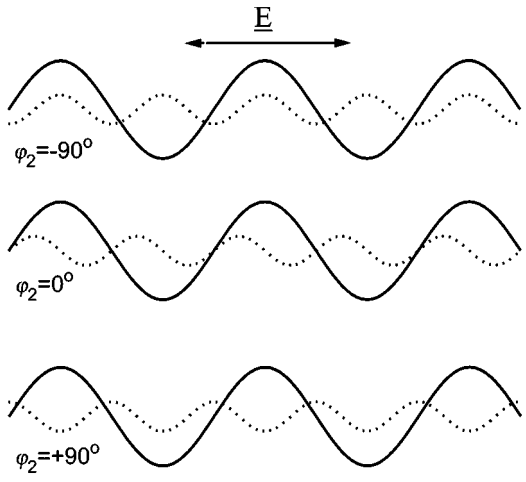

FIG. 12. Sketch of light at normal incidence coupling to the modes of a corrugate surface containing the K (solid line) and $2K$ (dotted line) components. There is a component of the incident field normal to the surface of the K component at all points except at the maxima and minima of the surface profile.

respond to flat regions of the surface, the mode having maxima at the troughs will not therefore couple to photons; see Fig. 11(a). When $\phi_2=-90°$, the peaks of the $2K$ component correspond to flat regions of the surface and the mode having maxima at the peaks is now uncoupled; see Fig. 11(b). For $\phi_2=0°$ the peaks and troughs of the $2K$ component occur at equivalent points with respect to the K component; the coupling is therefore the same for both; see Fig. 6.

The reasoning given in the above paragraph can be reinforced using the results we derived in Sec. V D. We can express the normal E field component of the SPP modes, $E_N^{\pm}$, using Eq. (5.34), as

$$E_N^{\pm} \propto \sin(Kv + \phi_2/2 \mp \pi/4). \tag{5.58}$$

In the particular situation under consideration, i.e., the amplitude of the K grating is much larger than the amplitude of the $2K$ grating, the component of the incident field normal to the corrugated surface is primarily that normal to the K component. The spatial dependence of this field will thus be $\cos(Kv)$. Comparing this with Eq. (5.58) we see that for $\phi_2=90°$, E_N^- (i.e., the low frequency branch) will have the same spatial dependence as the incident field, i.e., $\cos(Kv)$, and will thus be coupled whilst E_N^+ (the high frequency branch) will have an orthogonal spatial dependence, i.e., $\sin(Kv)$, and will thus not be coupled. For $\phi_2=-90°$ the spatial dependence, and thus coupling, will be reversed. This explains the feature found experimentally by Nash *et al.*[39] where simply inverting the grating reversed the strengths of the coupling to the two branches.

G. Extension of the model

Some care has been taken to retain all terms of order $(Kd_2)^3$ correctly within the model *as defined* so as to establish clearly the role of the $2K$ component in causing the band gap. However, at this level of accuracy there are terms of order $(Kd_2)^2$ which have been excluded. These arise because the $2K$ components of the grating will also couple the $\pm K$ to the $\pm 3K$ modes yielding an additional contribution of order $(Kd_2)^2$ to the results previously derived. It is however relatively straightforward to correct for this in a perturbative way which we outline below.

To include these extra terms we extend the eigenvalue matrix of Eq. (5.14) to contain $m=\pm 3$ as well as $m=\pm 1$ terms. The corresponding eigenvalue matrix is now 8×8 rather than 4×4 and has the structure

$$\begin{pmatrix} \mathbf{T} & \mathbf{X} \\ \mathbf{X}' & \mathbf{T}' \end{pmatrix} \begin{pmatrix} \psi \\ \psi' \end{pmatrix} = \lambda' \begin{pmatrix} \psi \\ \psi' \end{pmatrix}, \tag{5.59}$$

where $\mathbf{T}$ is the original 4×4 matrix, λ' is the modified eigenvalue, ψ contains the eigenvector field components associated with the $\pm K$ modes, and ψ' contains those associated with the $\pm 3K$ modes. The latter may be eliminated to yield a modified version of Eq. (5.9),

$$(\mathbf{T} + \mathbf{X}(\lambda' - \mathbf{T}')^{-1}\mathbf{X}')\psi = \lambda' \psi. \tag{5.60}$$

Since $\mathbf{X}$ and $\mathbf{X}'$ contain only terms that couple $\pm K$ modes to $\pm 3K$ modes, their leading order terms are of order Kd_2 and hence the correction to Eq. (5.9) will be of order $(Kd_2)^2$. In this approximation

$$\mathbf{X} = \begin{pmatrix} 3\xi & & & 0 \\ & -3\xi & & \\ & & \xi & \\ 0 & & & -\xi \end{pmatrix}, \quad \mathbf{X}' = \begin{pmatrix} \xi & & & \\ & -\xi & & 0 \\ & & 3\xi & \\ 0 & & & -3\xi \end{pmatrix} \tag{5.61}$$

and

$$\mathbf{T}' = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ W & 0 & 0 & 0 \\ 0 & W & 0 & 0 \end{pmatrix}, \tag{5.62}$$

with $W = 9K^2 - \varepsilon(\omega/c)^2$. The modified eigenvalues take the form

$$\lambda_1' = \frac{i\hat{\eta}_1}{(1-(3/8)(Kd_2)^2)}, \tag{5.63}$$

$$\lambda_2' = \frac{i\hat{\eta}_2}{(1-(3/8)(Kd_2)^2)}, \tag{5.64}$$

with

$$\hat{\eta}_i = K^2(1-\rho')^2 + \xi^2 - \varepsilon_i\left(\frac{\omega}{c}\right)^2 \chi, \tag{5.65}$$

where

$$\rho' = \frac{7}{8}(Kd_2)^2 \tag{5.66}$$

and

$$\chi = (1-\rho')\left(1 + \frac{1}{8}(Kd_2)^2\right). \tag{5.67}$$

The modified boundary conditions take exactly the same form as in Eq. (5.23) with η_i replaced by $\hat{\eta}_i$ and ξ replaced by $\xi' = \xi(1-(3/8)(Kd_2)^2)$. Algebraic manipulation identical

to that used in the preceeding sections, taking care to distinguish between ξ' in the boundary conditions and ξ in the definition of $\hat{\eta}_i$ [equation (5.65)] and noting that $(\xi^2-\xi'^2)$ is of order $(Kd_2)^4$, yields the result that up to and including terms of order $(Kd_2)^3$ the normalized central position and gap width are given by

$$\frac{1}{2}\left[\left(\frac{\omega_+}{c}\right)^2+\left(\frac{\omega_-}{c}\right)^2\right]=\left(\frac{\omega_0}{c}\right)^2(1-(Kd_2)^2) \quad (5.68)$$

and

$$\left(\frac{\omega_+}{c}\right)^2-\left(\frac{\omega_-}{c}\right)^2=\frac{4K^2}{\sqrt{-\varepsilon_1\varepsilon_2}}(Kd_2)\left(1-\frac{7}{2}(Kd_2)^2\right). \quad (5.69)$$

VI. COMPARISON WITH OTHER THEORIES, AND BETWEEN THEORY AND EXPERIMENT

A. Comparison with other theories in the literature

Previous investigations using general optical response theory have investigated the size of the SPP gap. Mills[9] deduced that for a SPP traveling along a corrugated metal/air interface i.e., $|\varepsilon_2|\gg\varepsilon_1=1$, with the SPP propagation direction making an angle $\gamma/2$ with the grating grooves, the gap was given by

$$\frac{\delta\omega}{\omega_0}=\frac{4K}{\sqrt{|\varepsilon_2|}}u\left(\sin\frac{\gamma}{2}\right)^2, \quad (6.1)$$

where u was the amplitude of the wave components $\exp(\pm iKx)$; thus in our notation $u=d_2/2$.

More recently Seshadri[40] has investigated the size of the gap including the frequency dependence of the ''ideal metal'' dielectric constant. This latter effect changes the gap by approximately 5% and if we ignore this contribution his analysis yields

$$\frac{\delta\omega}{\omega_0}=\frac{\eta K_s\cos^2\theta}{\sqrt{|\varepsilon_2|}}, \quad (6.2)$$

where η may be identified with d_2, θ is the angle the SPP makes with respect to the *normal* to the grooves, i.e., $\theta=90°-\gamma/2$, and K_s is the wave vector of the grating component that couples the two modes, i.e., $K_s=2K$.

To compare these results with ours we approximate Eq. (5.55) assuming the value of the band gap, $\delta\omega$, to be small, and find

$$\frac{\delta\omega}{\omega_0}=\frac{2Kd_2}{\sqrt{-\varepsilon_1\varepsilon_2}}\left(\frac{\varepsilon_1\varepsilon_2}{\varepsilon_1+\varepsilon_2}\right)\left(1-2(Kd_2)^2+\frac{(Kd_2)^2}{2(-\varepsilon_1\varepsilon_2)}\left(\frac{\varepsilon_1\varepsilon_2}{\varepsilon_1+\varepsilon_2}\right)^2+O(Kd_2)^4\right), \quad (6.3)$$

which in the limit that $|\varepsilon_2|\gg\varepsilon_1=1$ and $Kd_2\ll1$ yields

$$\frac{\delta\omega}{\omega_0}=\frac{2Kd_2}{\sqrt{|\varepsilon_2|}}. \quad (6.4)$$

This result agrees with that of Seshadri (at $\theta=0°$) and Mills (at $\gamma/2=90°$), in the limit $|\varepsilon_2|\gg1$. Thus all three analyses are consistent in their predictions of the gap width for SPP propagation normal to the grating grooves.

B. Comparisons between experiment and theory in the literature

The comparison with experiment has until recently been somewhat confusing. Raether[20] compared the results of Pockrand[41] with the theory of Mills[9] and reported the theoretically predicted gap to be greater than the experimentally measured one by a large factor. Seshadri[40] compared his theoretical results directly with the experimental data of Pockrand[41] and reported better agreement, obtaining half the experimental value. The gap measured by Pockrand was for an SPP propagation angle of $\theta=65°$ ($\gamma/2=25°$), and Seshadri[40] noted that in spite of the apparent discrepancy between his results and those of Mills, their two expressions agreed at $\theta=0°$. Since the gap at oblique incidence is simply related to that at normal incidence by an angular factor upon which they agree there is clearly an inconsistency somewhere.

This inconsistency arises because in applying Mills' formula to Pockrand's results Raether made two mistakes. First he identified u with d_2 rather than $d_2/2$ producing an overestimate of 2; further he took $\gamma/2=65°$ rather than 25° producing a further factor of $\tan^2(65°)=4.6$. Making these modifications Mills' [9] and Seshadri's[40] expressions produce a consistent estimate of the gap that is roughly half the experimental value. However, Raether makes it clear that Pockrand measured $\delta k/k$ and assumes that $\delta\omega/\omega\approx\delta k/k$. Weber and Mills[19] have however highlighted the danger of making such an interpretation, one that we discussed extensively in Sec. III, and so one must conclude that the often quoted disagreement is not well founded. For this reason we have carried out a series of detailed experiments to precisely determine the dependence of the gap, for SPP propagation at normal incidence to the grooves, upon the amplitude of the grating, d_2, that gives rise to the gap.

C. Comparisons with our own experimental work

As described above, the comparisons between experiment and theory that have been made to date have been unsatisfactory. To remedy this we have recently conducted a series of experiments to allow a proper comparison between theory and experiment to be made. The details of these experiments have been reported elsewhere,[6,42] so that here we examine the results and only describe those aspects of the experiments that are relevant to the present discussion.

We have used two methods to examine the SPP band gap as a function of the modulation depth of the grating that gives rise to the gap.

In the first[6] we used the double corrugation method discussed in Sec. III. The data obtained allowed us to verify the validity of Eq. (5.69), but not (5.68). This is because the amplitude of the $2K$ component, d_2, is always small compared to that of the K component, d_1. To achieve a d_2 big enough to show a noticeable change in the mean frequency of the gap would require a value of d_1 that would allow only

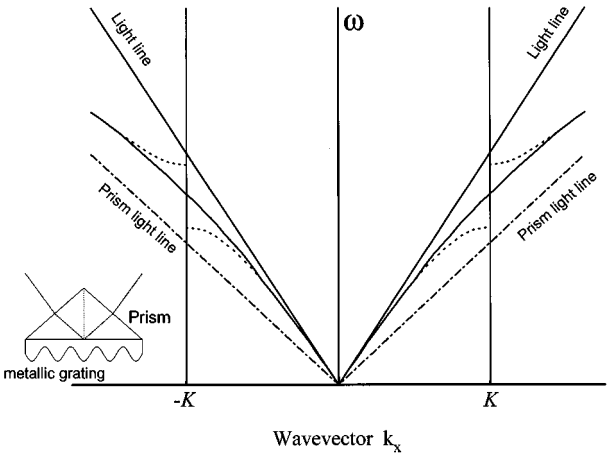

FIG. 13. Dispersion curve for surface plasmon polaritons examined with prism coupling. Surface modes are excited from the prism side, coupling taking place via the evanescent field that occurs on total internal reflection; see inset. Owing to the relatively high index of the prism, the light line for photons in the prism is shifted out in momentum. It is now possible to couple to modes in the region of the gap that arises from a corrugated surface with just one grating component, $2K$.

very poor coupling between the SPP modes and photons, thus making accurate interpretation of the experimental data impossible.

Our second approach[42] has been to use a prism rather than a second grating to couple the SPP modes to photons. The prism coupling technique is outlined in Fig. 13. Since the fundamental of the grating is no longer being used to couple light in and out of the sample, it may now be used as the grating that gives rise to the gap. Consequently, this technique allows the study of much greater corrugation depths. We have been able to go as far as $2Kd_2 \approx 0.23$ (note that for consistency in the prism coupling work we define the Bragg vector of the grating that gives rise to the gap as $2K$; a useful measure of the grating modulation is the product of the Bragg vector and the amplitude, i.e., $2Kd_2$). The double grating technique was limited to $2Kd_2 \approx 0.06$, compared to the $2Kd_2 \approx 0.23$ of the prism technique discussed above. These correspond to gaps (defined as $\delta\omega/\omega_0$) of 7% and 36%, respectively. The values of $2Kd_2$ and $\delta\omega/\omega_0$ are not in direct proportion owing to the different metal used in the two cases; see Eq. (6.3). In the small gap case gold was used and in the large gap case, silver.

The data obtained using the prism technique, together with theoretical expectations based on both our first order [Eqs. (5.55) and (5.57)] and our extended theory [(5.68) and (5.69)], are shown in Fig. 14. The agreement between our extended theory and the data is seen to be good over the range of grating modulation studied. Looking at the data of Fig. 14, the inclusion of higher order terms in our extended theory is particularly important for the central position. Our model, Eqs. (5.68) and (5.69), therefore appears to be valid for values of $2Kd_2$ up to at least 0.23. We note that the agreement at large $2Kd_2$ is good despite the fact that the model neglects the effect of dispersion in the permittivity of the silver over the measured frequency range. To describe still larger gaps it may be necessary to take account of this

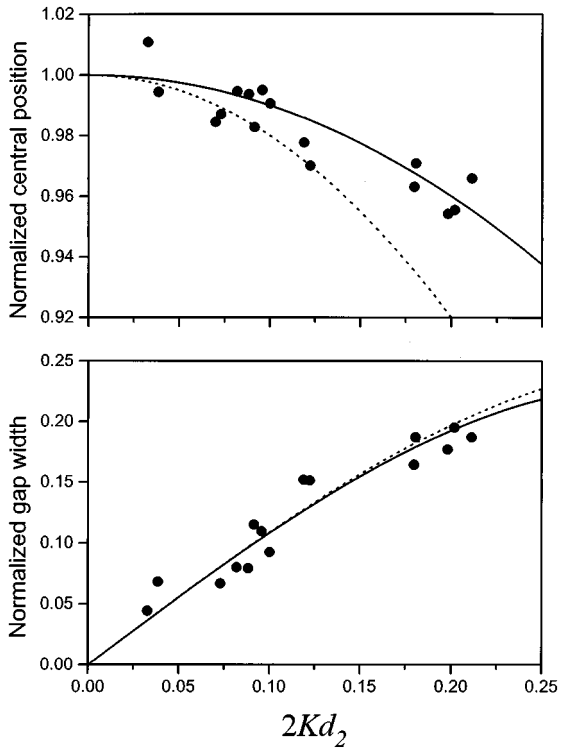

FIG. 14. The normalized central position and width of the gap (as defined in Sec. V E) as a function of surface modulation. The experimental data, taken from a prism coupling experiment (Ref. 42) are shown as dots, whilst the results of the analytic model are shown as lines, dotted for the first order theory, Eqs. (5.55) and (5.57), and solid for the extended theory that includes higher order terms, Eqs. (5.68) and (5.69).

dispersion. Also, the imaginary component of the optical permittivity of silver, $\varepsilon_i/|\varepsilon_r| \approx 0.03$, does not appear to have had a significant effect on the width or central position of the gap.

Figure 14 shows that there is a decrease in the central position of the gap as the grating depth increases, in good agreement with our theory. This result has not been experimentally verified before.

VII. FUTURE WORK

There remain a whole range of questions still to be addressed concerning the physics of SPP band gaps. It is not our purpose to provide a full discussion of these issues here, rather we mention each briefly to indicate areas that future research may take.

A. Bigger band gaps—problems with the light lines

We can see from Fig. 3 that as the band gap increases the high frequency branch approaches the light line, as we discussed at the end of Sec. V E. This provides a limit on how far this upper cutoff can be raised. As the mode approaches the light line it becomes more radiative in character, and is progressively less well confined to the surface; see Fig. 10. The limits on the width of the band gap that can be obtained experimentally have still to be investigated.

B. Band gaps for other surface modes

We have discussed here only the surface plasmon polariton mode; there are other surface modes that we could consider, notably the surface exciton polariton. We expect that the basic physics will remain the same; whether a substantial surface mode gap will be seen is likely to depend on whether low loss surface exciton polariton modes can be examined, i.e., the natural width of each mode can be made significantly less than the gap width. Such investigations may be particularly important in the context of wavelength scale optoelectronic semiconductor devices.

C. Band gap vs propagation angle

We have considered only the case of surface modes propagating in a direction normal to the grating grooves. As the propagation direction of the surface modes moves away from this direction, so the mean frequency of the associated band gap increases. This occurs because it is the component of the surface mode momentum in the direction normal to the grating grooves that must match the appropriate grating vector. Thus if the surface mode propagation direction makes an angle χ with the direction normal to the grating grooves, then we expect that the central position will have risen from ω_0 at $\chi=0°$ to $\omega_0/\cos\chi$. This together with predictions concerning the effect of the propagation direction on the band gap width[9,40] have yet to be checked experimentally.

D. The photonic surface

Following on from the above, it is clear that no matter how big the gap for a propagation angle of $\theta=0°$, there will not be a band gap that covers all possible directions of propagation. An obvious extension of the geometry considered so far is to examine a bigrating, i.e., two gratings, one rotated with respect to the other by 90°, or a trigrating, the gratings being rotated 60° with respect to one another. In this way we hope that it may be possible to construct a textured surface on which, for some range of frequencies, no surface modes may propagate. In analogy with developments in bulk photonic band gap work we call this the *photonic surface*.

E. The effect of SPP band gaps on other optical processes

Spontaneous emission from dye layers above surfaces exhibiting SPP band gaps,[25] together with their influence on the surface enhanced Raman[11] effect have been reported. However, the effect of such phenomena on the decay kinetics of excited molecules in close proximity to such surfaces has not so far been investigated. If the excited molecule/surface separation is appropriate then the dominant decay route for the molecule may be to excite an SPP mode. If, at the emission frequency of the molecule, the surface morphology precludes the existence of SPP modes then decay of the molecular excitation via this route will be blocked. This should alter the decay rate of the molecule, an alteration that should be observable. (Note, to properly describe this situation we would need to extend our analytic model to take account of the overlayer.)

VIII. CONCLUSIONS

We have seen how surface modes may scatter from the grating on which they propagate to form a standing wave. The standing wave has two solutions, one with field extrema at the grating peaks, the other with extrema at the troughs. We have shown that these two solutions have different energies, thus opening up a band gap in the propagation of the surface modes. This has been done by developing an analytic model based on illustrating the underlying physics by concentrating on finding analytic expressions for the spatial field and surface charge distributions associated with the modes. The predictions of our theory have been compared with the theory of others and they have been found to be consistent for the band gap, provided various misunderstandings in the literature are recognized.

By comparing the results of our experimental work with our theory the validity of our model has been verified for gratings with a modulation as high as amplitude/pitch=0.05 ($2Kd_2=0.23$). Further, we have shown that the central position of the gap falls as the grating depth rises, a result previously unverified. The good agreement that we find between experiment and theory provides a good test of the Chandezon technique when applied to the type of problem discussed here.

In addition we have highlighted some areas that are in need of further investigation; we are vigorously pursuing these.

ACKNOWLEDGMENTS

The authors are grateful for the financial assistance received from EPSRC (UK), the Defence Research Agency (Malvern, UK), and The University of Exeter in undertaking this work.

[1] J. Opt. Soc. Am. B **10** (2) (1993), special issue on photonic band gaps.

[2] J. Mod. Opt. **41** (2) (1994), special issue on photonic band structures.

[3] E. Yablonovitch, J. Opt. Soc. Am. **10**, 283 (1993).

[4] E. Yablonovitch and T. J. Gmitter, J. Opt. Soc. Am. A **7**, 1792 (1990).

[5] For a survey, see *Surface Polaritons*, edited by V. M. Agranovitch and D. L. Mills (North-Holland, Amsterdam, 1982).

[6] W. L. Barnes, T. W. Preist, S. C. Kitson, J. R. Sambles, N. P. K. Cotter, and D. J. Nash, Phys. Rev. B **51**, 11 164 (1995).

[7] P. Yeh, A. Yariv, and C. S. Hong, J. Opt. Soc. Am. **67**, 423 (1977).

[8] P. St. J. Russell, Phys. World (Aug), 37 (1992).

[9] D. L. Mills, Phys. Rev. B **15**, 3097 (1977).

[10] R. H. Ritchie, E. T. Arakawa, and J. J. Cowan, Phys. Rev. Lett. **21**, 1530 (1968).

[11] W. Knoll, M. R. Philpott, J. D. Swalen, and A. Girlando, J. Chem. Phys. **77**, 2254 (1982).

[12] Y. J. Chen, E. S. Koteles, R. J. Seymour, G. J. Sonek, and I. M.

Ballyantyne, Solid State Commun. **46**, 95 (1983).
[13] D. Heitmann, N. Kroo, C. Schulz, and Z. Szentirmay, Phys. Rev. B **35**, 2660 (1987).
[14] B. Fischer, T. M. Fischer, and W. Knoll, J. Appl. Phys. **75**, 1577 (1994).
[15] H. Lochbihler, Phys. Rev. B **50**, 4795 (1994).
[16] H. Raether, *Surface Plasmons* (Springer-Verlag, Berlin, 1988).
[17] S. Herminghaus, M. Klopfleisch, and H. J. Schmidt, J. Opt. Lett. **19**, 293 (1994).
[18] See, for example, N. P. K. Cotter, T. W. Preist, and J. R. Sambles, J. Opt. Soc. Am. A **12**, 1097 (1995); T. W. Preist, N. P. K. Cotter, and J. R. Sambles, *ibid*. **12**, 1740 (1995).
[19] M. G. Weber and D. L. Mills, Phys. Rev. B **34**, 2893 (1986).
[20] *Surface Plasmons*, Ref. 16, p. 113.
[21] V. Celli, P. Tran, A. A. Maradudin, and D. L. Mills, Phys. Rev. B **37**, 9089 (1988).
[22] E. Popov, Surf. Sci. **222**, 517 (1989).
[23] P. Halevi and O. Mata-Méndez, Phys. Rev. B **39**, 5694 (1989).
[24] S. H. Zaidi, M. Yousef, and S. R. J. Brueck, J. Opt. Soc. Am. B **8**, 1348 (1991).
[25] S. C. Kitson, W. L. Barnes, and J. R. Sambles, Phys. Rev. B **52**, 1441 (1995).
[26] For a general review of the theory of diffraction from dielectric gratings see, for example, *Electromagnetic Theory of Gratings*, edited by R. Petit (Springer-Verlag, Berlin, 1980); see also M. G. Moharam and T. K. Gaylord, J. Opt. Soc. Am. **11**, 1780 (1986); L. Li, J. Opt. Soc. Am. A **11**, 2816 (1994); J. Opt. Soc. Am. A **12** (5) (1995), special feature issue on diffractive optics modeling.
[27] J. Chandezon, M. T. Dupuis, and G. Cornet, J. Opt. Soc. Am. **72**, 839 (1982).
[28] E. Popov, Surf. Sci. **222**, 517 (1989).
[29] For a general review, see *Electromagnetic Surface Modes*, edited by A. D. Boardman (Wiley, New York, 1978).
[30] F. Tiogo, A. Marvin, V. Celli, and N. R. Hill, Phys. Rev. B **15**, 5618 (1977).
[31] A. A. Maradudin, in *Surface Polaritons*, Ref. 5, Chap. 10.
[32] R. Petit and M. Cadilhac, C. R. Acad. Sci. B **262**, 468 (1966).
[33] B. Laks, D. L. Mills, and A. A. Maradudin, Phys. Rev. B **23**, 4965 (1981).
[34] E. P. Da Silva, G. A. Farias, and A. A. Maradudin, J. Opt. Soc. Am. A **4**, 2022 (1987).
[35] S. R. Seshadri, J. Appl. Phys. **57**, 4874 (1985).
[36] N. E. Glass, R. Loudon, and A. A. Maradudin, Phys. Rev. B **24**, 6843 (1981).
[37] *Surface Polaritons*, Ref. 31, p. 426, Eq. 3.8b. Note that the β_m of our notation is related to the α_p of Maradudin by $\alpha_p = -\beta_m$.
[38] M. G. Weber and D. L. Mills, Phys. Rev. B **31**, 2510 (1985).
[39] D. J. Nash, N. P. K. Cotter, E. L. Wood, G. W. Bradberry, and J. R. Sambles, J. Mod. Opt. **42**, 423 (1995).
[40] S. R. Seshadri, J. Appl. Phys. **58**, 1733 (1985).
[41] I. Pockrand, Ph.D. thesis, Hamburg, 1978.
[42] S. C. Kitson, W. L. Barnes, G. W. Bradberry, and J. R. Sambles, J. Appl. Phys. **79**, 7383 (1996).