Low temperature phonon-drag thermoelectric power calculations in GaAs/GaAlAs heterojunctions and Si MOSFETs

by

Mark John Smith.

A thesis

presented to the University of Warwick

in partial fulfillment of the requirements

for entry to the degree of

Doctor of Philosophy.

Department of Physics

September 1989.

Contents

List of Figures

List of Tables

Acknowledgements

The author wishes to acknowledge his gratitude to the following people:

Professor P. N. Butcher, for his excellent supervision;

Dr. B. L. Gallagher (University of Nottingham) and Dr. S. S. Kubakaddi (Karnatac University, India), for much fruitful discussion;

his parents, for their constant support and encouragement throughout his education;

but particularly to his wife Tania, for her help and understanding.

Declaration

Unless where stated otherwise the work in this thesis is the authors' own original research work, performed in the Department of Physics at the University of Warwick between Oc-

ober 1986 and September 1989 under the supervision of Professor P.N. Butcher.

the work has been published previously:

lculation of the effect of screening on Phonon-drag Thermoelectric Power in a

FET", Smith M.J. and Butcher P.N., 1989 *J.Phys.Condens. Matter* **1** 1261,

"Inelastic scattering and the temperature dependence of Phonon-drag Thermoelectric Power in Quasi-2D systems", Smith M.J. and Butcher P.N., 1989 *J.Phys.Condens.Matter* **1** 4859,

3."Simple models of Phonon-drag in 3D and quasi-2D", Smith M.J. and Butcher P.N. (submitted to *J.Phys.Condens.Matter*),

and presented as posters at conferences:

1."New calculations of Phonon-Drag Thermoelectric Power in GaAs/GaAlAs heterojunctions", Smith M.J. and Butcher P.N.,*I.O.P. Solid State Physics Conference* 1987 (Bristol).

2."Low temperature dependence of thermopower in GaAs/GaAlAs heterojunctions" Smith M.J. and Butcher P.N., *I.O.P Sold State Physics Conference* 1988 (Nottingham).

3."The Thermopower of Si inversion layers", Gallagher B.L., Oxley J.P., Galloway T., Smith M.J., Butcher P.N. and Pepper M., *I.O.P. Solid State Physics* Conference 1988 (Nottingham).

Summary

The effect on the electron transport of the confinement of the electrons to a narrow channel in GaAs/GaAlAs heterojunctions and Si MOSFETs is reflected in quantities like the thermopower (S) which is sensitive to the transport of both heat and charge. The calculations described here confirm that in these systems S is dominated by phonon-drag (S_g) at temperatures (T) around 1-10K and reveals more sensitivity than previously imagined.

Simple models and the Boltzmann transport formalism have been investigated. The formalism enhances the predictions of the simple models and reproduces the simple S_g formulae in appropriate limits. Amplification of S_g in quasi-2D arises from the loss of the momentum conservation constraint across the channel at small widths δ.

Earlier calculations were numerically inaccurate and greatly overestimate $-S_g$ by ignoring screening. An effective multi-subband screening dielectric function is defined which reduces to the single subband approximation at small δ and low electron density (n). Non-degeneracy has also been included. It is an important consideration despite the low temperatures of most of the data. The treatment of electron confinement has been improved and the temperature dependence of the polarizability investigated. It is unimportant in the current experimental systems but significant at lower n and higher T.

The piezoelectric scattering mechanism has been introduced and dominates S_g in the heterojunction for T <1K. A dominant 2D wavevector component has been defined for the phonon population at given T which is very helpful in understanding S_g. A correction for the energy dependence of the electron relaxation-time is necessary and demonstrates the dependence of S_g upon the dominant electron scattering mechanism.

The calculations of S_g in the quantum-limit and boundary scattering regime now explain the measured S in heterojunctions and peaks in $-S_g/T^3$ in the MOSFET up to an accuracy better than 10% without adjustable parameters.

Chapter 1

Introduction to phonon-drag thermoelectric power in quasi-2D.

1.1 Introduction to the chapter.

The objective of the work described in this thesis has been to improve the understanding of phonon-drag in quasi-2D. More specifically, certain aspects of the phonon-drag contribution to the thermoelectric power ("thermopower") of a quasi-2D electron gas in GaAs/GaAlAs heterojunctions and in Si MOSFETs have been studied. Much use will be made of insight gained from simple models. It is helpful, then, to begin by developing the macroscopic transport equations phenomenologically. In this way thermopower and phonon-drag are introduced. In the following sections what is meant by a quasi-2D electron gas is explained and why there should be particular interest in systems of reduced dimensionality. Some motivation towards, and historical background into, studies of thermopower and phonon-drag are then given before the main work of the thesis is presented.

1.2 Introduction to thermopower and phonon-drag.

Thermoelectric power or "thermopower" (S) is one of the four commonly measured coefficients of linear transport in solids along with resistivity (ρ), thermal conductivity (κ) and

the Peltier coefficient (Π). By "transport" is meant transport of heat and charge and in elementary work these are considered separately. Thus for the heat flux ($\mathbf{Q}$) in response to a gradient in temperature (∇T) the relation defining (κ) would be written:

$$\mathbf{Q} = -\kappa \nabla T \tag{1.1}$$

Similarly, the electrical conductivity σ (the inverse of ρ) would be written in terms of the electric current density ($\mathbf{J}$), in response to an emf $\mathbf{E}'$. This quantity is given by the gradient in the electrochemical potential upon the magnitude of the electronic charge ($\nabla \mu_{ec}/e$) and can be written:

$$\mathbf{E}' = \rho \mathbf{J}. \tag{1.2}$$

In a homogeneous isothermal system $\mathbf{E}'$ is equal to the electric field ($\mathbf{E}$). However, whenever charge flows there is an associated transport of heat. Thus there is a second term to add to the right hand side of (1.1) when an electric current flows and the total heat flux is written:

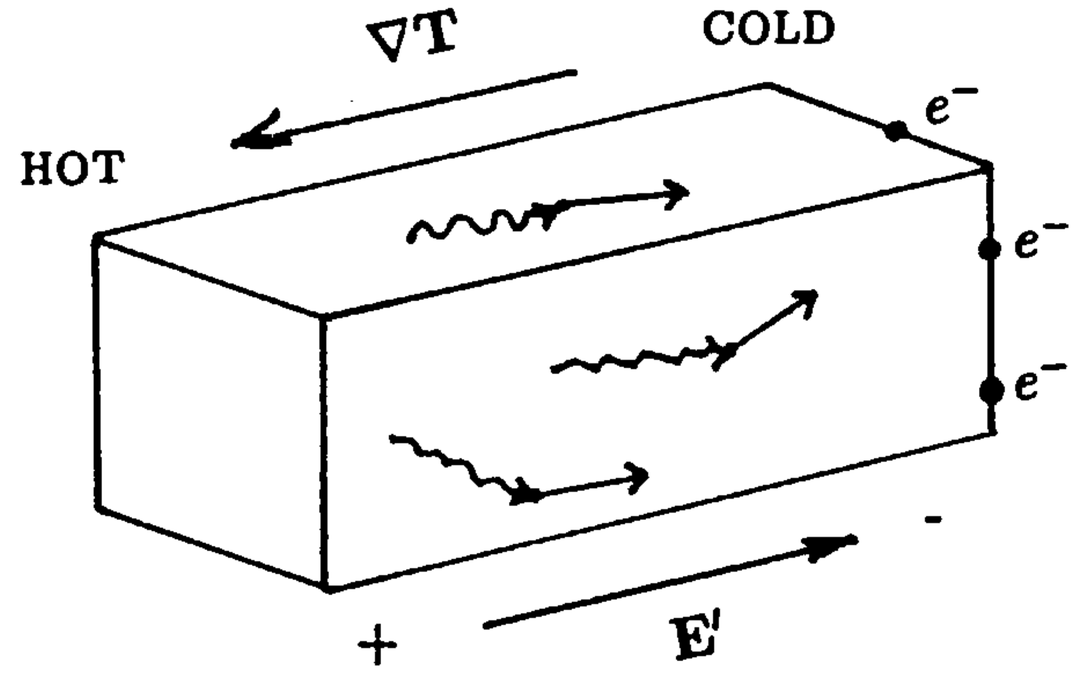
$$\mathbf{Q} = \Pi \mathbf{J} - \kappa \nabla T \tag{1.3}$$

At constant T the heat flux $\Pi\mathbf{J}$ accompanying an electric current gives rise to the Peltier effect. The corresponding effect (Seebeck) is the transport of charge accompanying the flow of heat in a temperature gradient (see, for example, Ziman 1963, Blatt 1968). It is this effect which is responsible for a contribution to $\mathbf{E}'$ when $\mathbf{J} = 0$, and therefore (1.2) is rewritten more generally as:

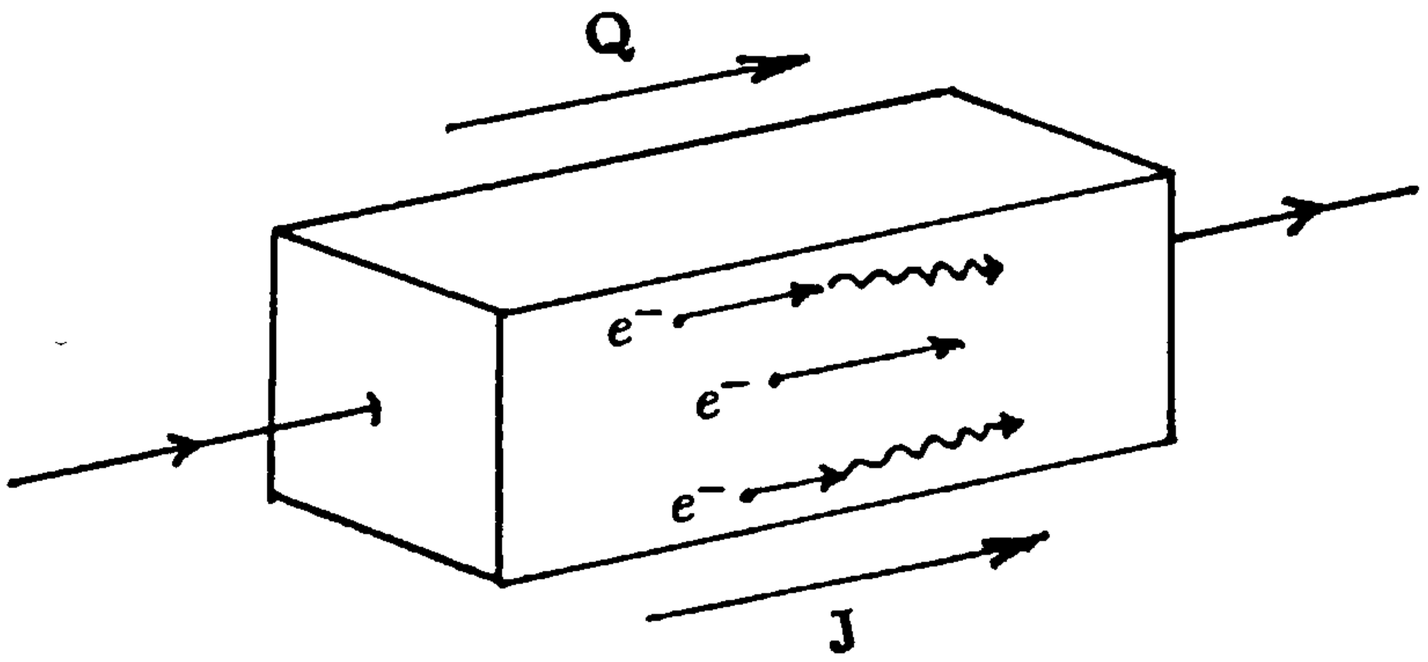
$$\mathbf{E}' = \rho \mathbf{J} + S \nabla T \tag{1.4}$$

Cubic systems are considered for simplicity because the tensor quantities ρ, S, Π and κ become scalars. Then Π from (1.3) is the isothermal heat flux per unit current density or loosely "the heat per charge". From (1.4) the Seebeck coefficient (S) is the emf per unit temperature gradient with no net charge flow. Hence for an isolated homogeneous "isothermal" crystal (as in Figure 1.1a) S is the differential voltage $-\Delta V/\Delta T$ which is measured in response to a small temperature difference ΔT. This difference initially causes a diffusion of charge carriers which results in charge separation and gives rise to

Figure 1.1: Origins of S, Π, and phonon-drag.



(a)Drag of carriers by phonons. In an isolated crystal ΔT causes charge separation as the carriers drift ($\longrightarrow$) down the gradient and give rise, eventually, to the opposing field $\mathbf{E}'$ which prevents further flow. Then $\mathbf{E}' = S\nabla T$ and hence S is usually negative. A net phonon flux ($\sim\!\!\sim\!\!\rightarrow$) increases the momentum of the carriers (e^-) which are, in effect, dragged, along. S is written as S_d(diffusion of carriers) + S_g(drag of carriers).



(b)Drag of phonons by carriers. In an isothermal system there is a heat flux $\mathbf{Q}$ accompanying current density $\mathbf{J}$ given by: $\mathbf{Q} = \Pi\mathbf{J}$. Some momentum is imparted to phonons ($\curvearrowright\!\!\rightarrow$) by the net carrier drift ($\longrightarrow$) and so Π is written as Π_d(diffusion of carriers) + Π_g(drag of phonons).

an electric field. Eventually charge flow ceases when the emf balances the thermal force on the carriers. There is therefore a contribution to S arising from carrier diffusion (S_d). In addition ΔT gives rise to a heat flow which, as a net flux of non-equilibrium phonons in the same direction, results in a mutual energy and momentum exchange between the carriers and the phonons. The carriers are dragged along as in Figure 1.1a. This is the origin of the phonon-drag correction (S_g) to S and the total thermopower is written:

$$S = S_d + S_g. \tag{1.5}$$

Hence, from (1.4), a temperature gradient gives rise to a contribution to $\mathbf{E}'$ from charge carriers dragged by phonons. As shown in the next Chapter the Onsager relation (see, for example, Ziman 1963) between Π and S:

$$\Pi = ST, \tag{1.6}$$

provides further insight. It is to be noted from (1.5) that there is a contribution to the isothermal heat flux in (1.3) arising from phonons dragged by the carriers (Π_g) in addition to that from carrier diffusion (Π_d) as shown in Figure 1.1b. Thus the mutual exchange of energy and momentum in general adds to the interest of phonon-drag and provides two approaches to calculations (see section 2.2).

In treating charge and heat transport together the transport equations (1.3) and (1.4) are more naturally written in terms of the charge and heat current densities arising linearly from small gradients in μ_{ec} and T:

$$\mathbf{J} = \sigma\mathbf{E}' + L\nabla T \tag{1.7}$$

$$\mathbf{Q} = M\mathbf{E}' + N\nabla T. \tag{1.8}$$

In this form $\mathbf{E}'$ and ∇T are considered as stimuli producing the responses $\mathbf{J}$ and Q. However, the form (1.3) and (1.4) is more useful here to make contact with experiment as it is ρ, S, Π, and κ which are measured directly. It is readily seen that $S = -L/\sigma$, $\Pi = M/\sigma$ and $\kappa = LM/\sigma - N$ but there is no generally accepted terminology for L, M and N.

1.3 Some essentials of low dimensional systems and conduction in quasi-2D.

Before considering what is meant by "low dimensional systems" (LDS) and quasi-2D conduction, which is of particular interest here, it is worth briefly considering the more usual case of 3D metallic conduction. In an elementary view the free electrons are three-dimensional in the sense that they are imagined to move equally freely in three dimensions on a macroscopic scale D, say. The potential barriers which confine the electrons to the solid are equally far apart (δ) in all directions and are sufficiently distant to be largely unimportant for the conduction (eg Ashcroft and Mermin 1981). A low dimensional system arises when these barriers are close together ($\delta \ll D$) in one or more dimensions, on a scale comparable to the de Broglie electron wavelength λ (ie $\delta \sim \lambda$). A quasi-2D system arises when the electrons are confined to such a scale in one direction, z say, but otherwise remain free. Conduction then takes place within a quasi -2D plane of width δ. The pure 2D limit is approached as $\delta \rightarrow 0$, where no freedom remains in this dimension. Quasi-1D systems may be created by additional confinement in a further dimension (y say). The electrons are then free in only one dimension (x) as they are confined to a "quantum wire". The electron motion is linear and the 1D conduction is along the wire. A quasi-0D case arises when the electrons are confined to a "quantum box" ie confined to $\delta \sim \lambda$ in three dimensions. Conduction is then only possible by electrons jumping between such boxes. For a review of the physics of 2D see Ando et al 1982. For 1D, 0D and aspects of the physics and technology of LDS see Heinrich et al (1988), Kagoshima et al (1982) and Bernasconi and Schneider (1981).

It is quasi-2D which is of particular interest here and in Chapter 3 the practical realization of quasi-2D confinement will be discussed in Si MOSFETs and GaAs/GaAlAs heterojunctions in order to treat the confinement quantitatively in the systems of interest. It is the MOSFET which inspired the majority of the early work on LDS although quasi-2D conduction has also been studied on the surface of liquid helium, in thin films, and in

layer compounds and graphite (see Ando et al, 1982). The MOSFET is particularly convenient to study experimentally as the surface charge density can be varied over a wide range and because it has an established technology. The review by Ando et al (1982) is quite exhaustive. Structures such as the GaAs/GaAlAs heterojunction manufactured by molecular beam epitaxy (MBE), metal-organic chemical vapour deposition (MOCVD) and similar epitaxial growth techniques, are not as flexible *after* manufacture but properties like potential barrier heights, widths and profiles can be largely determined by choice of material and doping, during growth. For a review see Kelly and Nicholas (1985) and also Heinrich et al (1988).

The essential feature of quasi- 2D is the creation of the narrow ($\delta \ll D$) potential well in the conduction band for motion in one particular direction. This can arise from an applied potential (MOSFET) or from the band gap discontinuity at the junction between different materials (heterojunction). In reality it is some complicated function with a non-trivial shape $V(z)$ but there are simplifications which allow progress. Most importantly, the carrier motion is described using effective mass (m) theory (Ando et al 1982). In the conducting plane 2D plane waves can be taken ($e^{\mathbf{ik.r}}$) having 2D wavevector $\mathbf{k}$, where $\mathbf{r} = (x, y)$. Perpendicular to the plane the potential seen by the electrons is a consequence of the crystal potential and the electron distribution, given by solving Poisson and Schrodinger equations self-consistently.

In an infinite square well (ISW) model potential:

$$V(z) = \begin{cases} 0 & 0 < z < \delta \\ \infty & elsewhere, \end{cases} \tag{1.9}$$

the electron eigenfunctions are a series of envelope functions $\phi_\alpha(z)$ (α = 1,2,....) with energies ε_α given by $\alpha^2\pi^2\hbar^2/(2m\delta^2)$. For a finite well or more realistic models, discrete energy levels ε_α are still expected but of finite number and (inevitably in practice) having some finite width. These levels occurring within the conduction band are therefore referred to as subbands and alpha as the subband index. For a conduction plane of area A the

single particle electron eigenfunction in the independent electron picture is taken as:

$$\psi_{\alpha,\mathbf{k}}(z) = A^{-1/2} e^{i\mathbf{k}.\mathbf{r}} \phi_\alpha(z) \tag{1.10}$$

Each state is labelled by the subband index and 2D wavevector (α, κ). The envelope function for the z direction $\phi_\alpha(z)$ satisfies:

$$-\frac{\hbar^2}{2m}\frac{d^2}{dz^2}\phi_\alpha(z) + V(z)\phi_\alpha(z) = \varepsilon_\alpha \phi_\alpha(z) \tag{1.11}$$

and has the form of sine functions for the simplest (ISW) case. The corresponding band structure is illustrated in Figure 1.2 where all states are filled up to the Fermi level (ε_f) at $T = 0$. Since the density of states $\mathbf{k}$ for each subband α is the constant $m/\pi\hbar^2$ per unit area, the total state density is:

$$g(\varepsilon) = \sum_\alpha \frac{Am}{\pi\hbar^2}\Theta(\varepsilon - \varepsilon_\alpha), \tag{1.12}$$

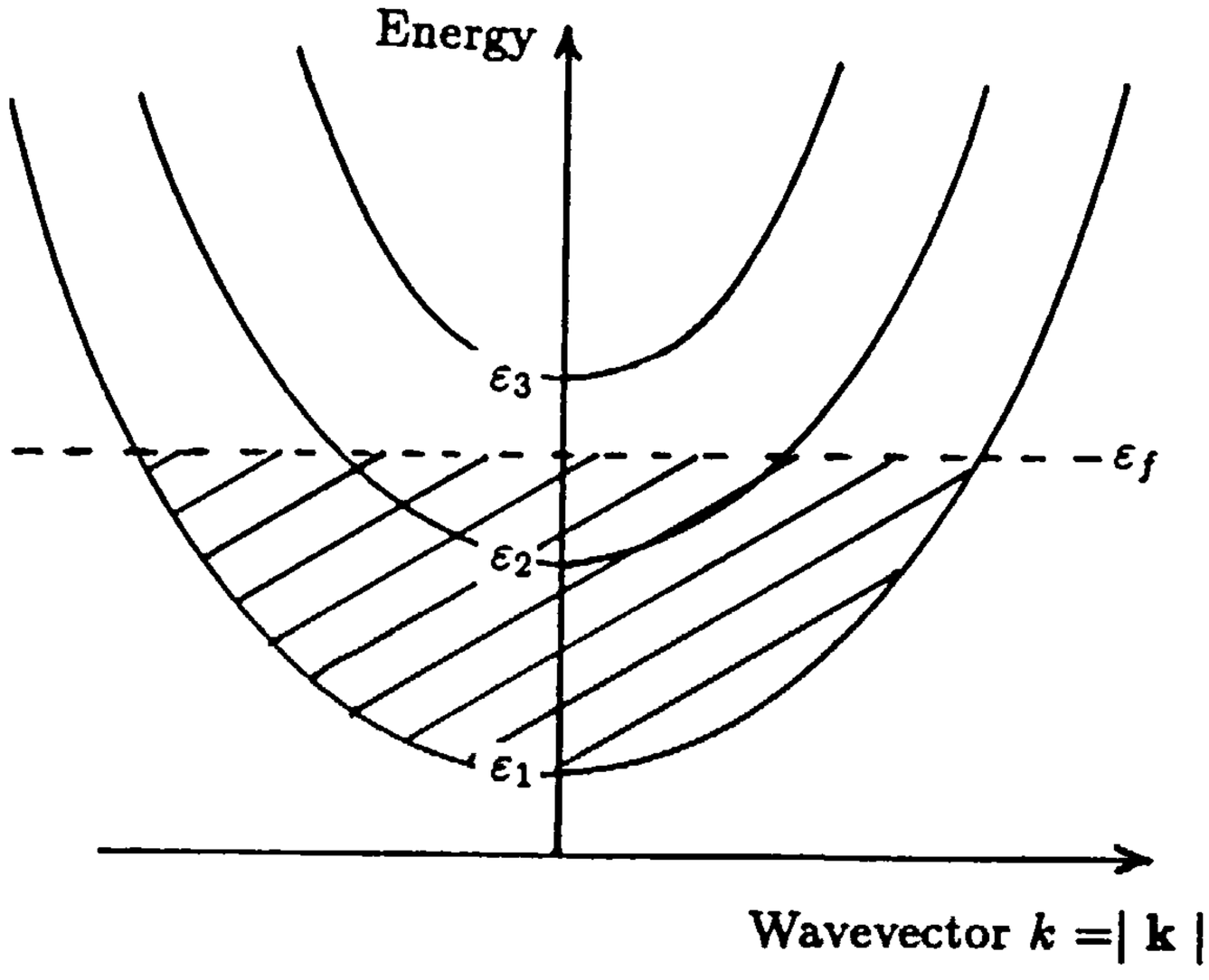
where $\Theta(x)$ is the unit step function, and takes the form as in Figure 1.2. The so-called "quantum limit" is the case when only the first (lowest, $\alpha = 1$ or "ground") subband is occupied. This limit is approached when the subband energy separation becomes large enough for ε_f to lie below ε_2, eg in the ISW model when δ becomes very small. Then $g(\varepsilon)$ takes the 2D limiting value $m/\pi\hbar^2$.

The exact form of $\phi_\alpha(z)$ is only important for quantitative work and hence only the most significant contributions to $V(z)$ are normally of importance to transport calculations. Hence the independent particle picture provides a sound base from which to begin. Accounting for screening improves calculations within the single particle picture as shown in Chapter 4 but it is not normally necessary to go beyond this to obtain agreement with experimental data.

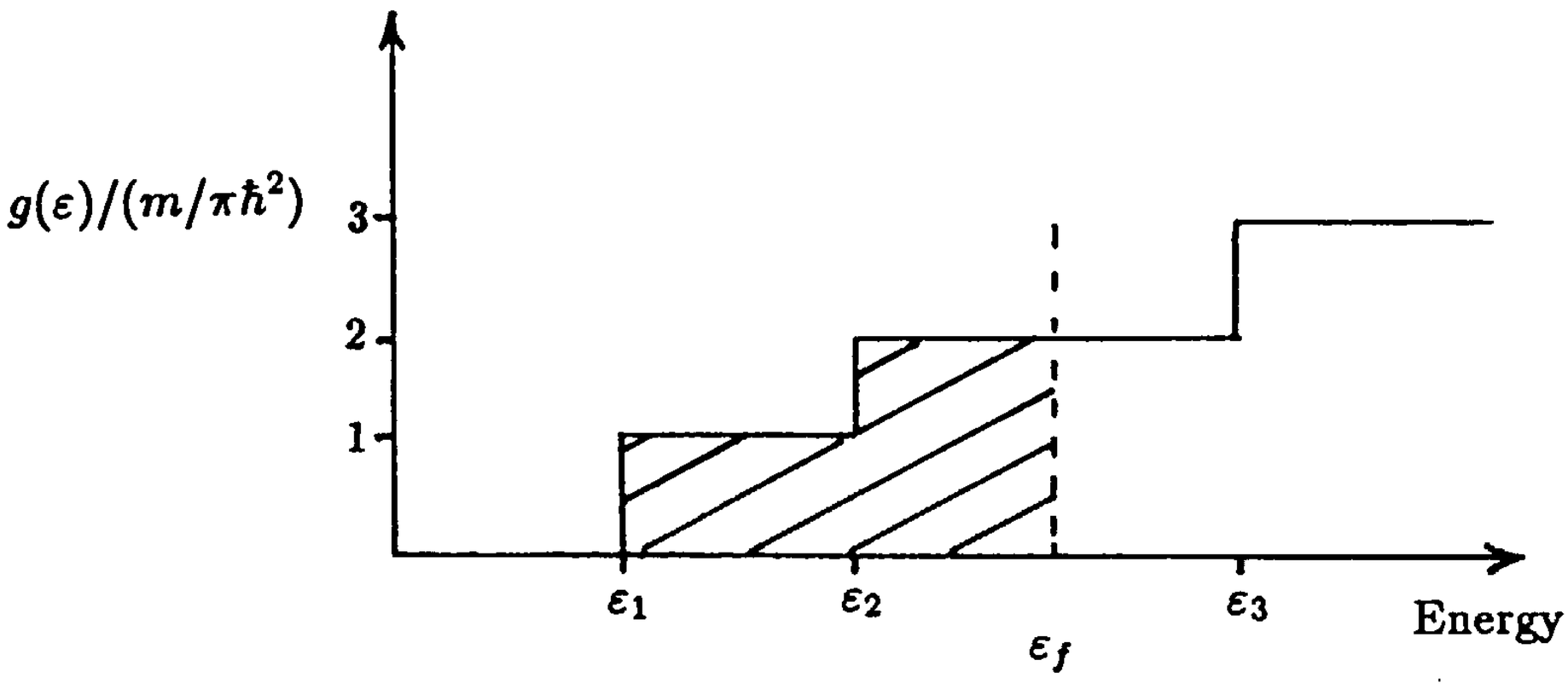
1.4 Interest in LDS and quasi-2D.

Following the elementary ideas mentioned above it is likely that LDS may show some interesting features. Figure 1.2 suggests novel phenomena may arise when ε_f approaches

Figure 1.2: Quasi-2D band structure and density of states.



(a) Quasi-2D band structure and filling of the states $(\alpha, \mathbf{k})$ up to the Fermi level ε_f at $T = 0$.



(b) Quasi-2D density of states $g(\varepsilon_\alpha(\mathbf{k}))$ corresponding to (a).

subband energies ε_α because of the sharp increase in the number of available empty states. In 1D such features are even more pronounced (Kearney and Butcher 1986). Although the conduction has reduced dimensionality in LDS, the conducting channels are generally embedded in or attached to some material host of macroscopic dimensions. The carriers may therefore have restricted motion but still interact with defects, phonons and other features of interest in a 3D solid. The resulting interactions between 3D scatterers and reduced dimensionality carriers therefore adds further interest and a particular example is considered here, the interaction of quasi-2D electrons with 3D phonons. In addition the crystal properties providing low dimensional conduction can affect the properties associated with the bulk. An example is the confinement of phonons in GaAs/GaAlAs superlattices (see, for example, Cardona 1989, Dharssi and Butcher 1989). However, this extra complication will not be of concern here.

The considerable industrial interest in the applications of the physics of LDS to microelectronics is evident from the number of papers published and presented in condensed matter physics journals and at conferences, by authors from industrial research groups. Clearly the ever-decreasing size and improving performance of electronic components eventually requires the models of transport used in their design to be modified to take into account quantum-mechanical effects resulting from extreme miniaturization. The advent of sophisticated high purity fabrication techniques such as MBE and MOCVD have stimulated further interest since it is apparent that device performance can be improved and new devices proposed which use the quantum-mechanical effect of LDS. Already devices are fabricated on the LDS length scale and, in some, such effects are essential to device performance (see for example Kelly and Nicholas 1985, Kelly and Wiesbuch 1986). Furthermore, the physics of LDS can be most readily studied by performing experiments on Si MOSFETs and devices like GaAs/GaAlAs HEMTS (High electron mobility transistors) because then the device properties can be designed to suit the particular experiment. A pertinent example is that of Ruf et al (1989) who required a small subband energy difference $(\varepsilon_2 - \varepsilon_1)$ to see the sign change in S_d predicted by Cantrell and Butcher (1985).

Further examples are in the phonon-drag imaging experiment of Karl et al (1988) and the growing interest in tunnelling and resonance phenomena (see, for example, Chang et al 1974, Mendez et al 1986, Eaves et al 1988) where, for example, the double barriers have heights, widths and separation of particular design. Advances in technology allow experiments to be conceived which would otherwise be impossible and correspondingly, industry must understand the features governing device performance in order to improve it. There is ,therefore, room for a fruitful interplay between pure research and industrial applications.

1.5 Interest in thermopower and phonon-drag in quasi-2D.

The thermopower of a system is an interesting quantity because its sensitivity to the details of the system means that it reveals much information (see, for example, Ziman 1963, Blatt 1968). A simple example of this follows directly from considering (1.3 and 1.4) under isothermal conditions. Regarding Π as the isothermal heat flux per unit current density, or loosely as the "heat carried per charge", it is clear that whilst **J** has the same sign in an applied **E**-field whether the carriers are holes or electrons, the carrier motion, and hence the heat transported (measured from ε_f), occurs in opposite directions. The sign of the coefficient (Π or S) changes ,then, according to the sign of the dominant carriers. Using (1.6) it also follows that S is a measure of the average energy of the carriers measured with respect to ε_f and this gives an indication of the dominant conduction mechanism. Naively, and in the absence of phonon-drag, the carrier (electron, say) energy in the nondegenerate limit, for example, will be about $K_BT + \varepsilon_c - \varepsilon_f$ and hence:

$$S_{(non-deg.)} \approx -\frac{K_B}{e} - \frac{\varepsilon_c - \varepsilon_f}{eT}, \tag{1.13}$$

where ε_c is the conduction band-edge. Thus for the degenerate case:

$$S_{(deg.)} \approx -\frac{K_BT}{\varepsilon_f}.\frac{K_B}{e} \tag{1.14}$$

may be expected. For activated conduction, in which the electrons have to be given energy ΔE to overcome some potential barrier in order to conduct, the result expected is:

$$S_{(act.)} \approx -\frac{\Delta E}{eT} \tag{1.15}$$

In this simple picture ,therefore, S tends to a constant independent of T for non-degenerate conduction at high T, is proportional to T for degenerate conduction and for the activated case to T^{-1}. Furthermore (Mott 1969), for variable range hopping conduction S has a $T^{1/3}$ dependence whereas the signature of phonon-drag in the degenerate case is a T^3 dependence, as shown in Chapter 2. In addition S is also sensitive to the scattering mechanisms involved since exchange of energy and momentum between carrier and scatterer influences the amount and ease with which the "heat per charge" is transported. Finally S also reflects the density of states.

Interest in S increased following the measurements of Geballe and Hull (1953) on single crystals of germanium from room temperature to 77K. They observed an increase in the size of S as T decreased which could not be explained by existing models (Herring 1953). It was explained, however, when Herring (1954) invoked phonon-drag, as in (1.5), in the first detailed calculation of S_g in semiconductors. Here the idea of saturation in the degenerate limit, which is discussed in the next chapter, was introduced for the first time. The S_d term in (1.5) is seen to result when the phonon gas (3D) remains in equilibrium (no net phonon flux) or if there is no electron-phonon interaction (no drag). The term S_g thus arises from the interaction of carriers with non-equilibrium phonons and is sensitive to the electron-phonon coupling.

Calculations of phonon-drag in metals were performed by Bailyn (1958, 1967) following general interest in the effect of non-equilibrium phonons on transport coefficients (Gurevich, 1945). Coupled, linearized, electron and phonon Boltzmann equations are solved using relaxation time approximations in order to calculate S_g which is a sensitive function of the anisotropy of the electron scattering. This result is confirmed by Guenault (1971) who provides insight by showing that Bailyn's formula can be written as a sum over phonon wavevectors $\mathbf{Q}$ of a quantity proportional to $\alpha(\mathbf{Q})$, the fraction of phonon

collisions involving electrons. For the simplest case of free electrons in a pure metal at low temperatures (in which $\alpha \rightarrow 1$), the result is:

$$S_g = -\frac{1}{3}.\frac{C_v}{N_v e} \tag{1.16}$$

where C_v is the lattice specific heat and N_v is the volume density of conduction electrons. This "metallic" formula predicts the low temperature T^3 dependence of S_g in metals. When electrons and holes are both present or when anisotropy is important, S_g is the sum of similar terms of opposite sign and becomes sensitive to their relative magnitude. When $\alpha(Q)$ is small S_g can be quenched by, for example, the increasing dominance of phonon-impurity or phonon-phonon scattering as discussed in Chapter 2.

Interest in the thermopower of quasi-2D electrons followed investigations into quantum size effects (QSE) on electron transport in LDS and the discovery of the Quantum Hall effect (von Klitzing et al 1980, von Klitzing 1986) in particular. QSE are effects on macroscopic measurements resulting from the quantum-mechanical nature of transport and are manifested when aspects of the system become small enough to be comparable to the electron wavelength. Recent interest is typified by the ballistic transport and quantum interference effects observed in novel electronic devices, such as the quantized conductance of a short narrow channel (van Wees et al 1988, Wharam et al 1988, but see also Heinrich et al 1988). Streda (1989) has shown that the thermopower in these systems may also be quantized. Inevitably some interest in QSE was focused on thermopower measurements because, being a sensitive quantity, such effects were expected to be significant. In particular, S_d was predicted to show a sign change (Cantrell and Butcher 1985) when the Fermi level approaches subband energies above ground, to within a few K_BT. This can be understood from the Mott formula (see, for example, Blatt 1968):

$$S_d = -\frac{\pi K_B^2 T}{3e}\left[\frac{1}{\sigma(\epsilon)}\frac{\partial\sigma(\epsilon)}{\partial\epsilon}\right]_{\epsilon_f}. \tag{1.17}$$

The conductivity is normally an increasing function of energy but when ϵ_f is near subband minima the scattering rate increases quickly and hence $\partial\sigma(\epsilon)/\partial\epsilon$ changes sign. Measurements of S on the quasi-2DEG (quasi-2D electron gas) of the conducting surface of

germanium (Zavaritsky and Zavaritsky 1982) in p and n inversion layers on silicon (Zavaritsky and Kvon 1984) and at interfaces in bicrystals and cleavage planes (Zavaritsky 1984), were made typically at $T < 10K$. Such low temperatures are used to avoid the masking of QSE by thermal broadening. The results revealed the T dependence characteristic of drag in 3D metals (1.16) rather than that of degenerate diffusion (1.14) and the magnitudes of S were up to two orders greater than expected. The expectation (Zavaritsky 1984) was that the 2D character of the carriers enabled coupling to a larger phonon population following the loss of the requirement for momentum conservation in the confinement (z) direction (see Chapter 2). The magnitude of S_g appeared to be influenced principally by the phonon mean free path (L) and the dominant value of the parallel phonon wavevector component ($\bar{q}$). Also a much sharper resonance was observed (Zavaritsky, 1984) in the electron scattering rate for $\bar{q}$ around twice the Fermi wavevector ($2k_f$) than seen in 3D electronic conduction. Probing of the electron-phonon interaction by ballistic phonon absorption/emission experiments also began at about this time (eg Hensel et al 1983).

The remarkably accurate quantization of the transverse resistivity in Quantum Hall measurements soon stimulated both experimental and theoretical effort towards similar effects in the thermopower. It was initially thought that S would exhibit similar features and calculations by Girvin and Jonson (1982), Streda (1983), Zawadsky and Lassnig (1984) and Jonson and Girvin (1984) seemed to confirm this. Oscillations in the diagonal component of the S_d tensor are predicted in the disorder-free limit in high magnitude fields (B) at low T, with maxima given by exact multiples of K_B/e. This appeared to be observed in early experiments on heterojunctions (see, for example, Obloh et al 1984) but later measurements revealed conflicting results and in particular the magnitudes recorded were much larger than predicted apparently due to phonon-drag (Fletcher et al 1986, Fletcher et al 1988a(b), D'Iorio et al 1988, Ruf et al 1988). Although the recorded behaviour of S for large B is qualitatively as predicted the oscillation maxima do not generally follow the K_B/e relation. Some of the conflict is due to difficulties in the measurement thermometry. It has been pointed out (Fletcher et al, 1988a) that in at least one experiment (Davidson

et al 1986) the somewhat lower values for S, recorded when $B = 0$, arise because the thermal gradient measured is not totally across the specimen itself. Phonon-drag is also the likely candidate for the disagreement when $B \neq 0$. For $B = 0$ the measured thermopower in heterojunctions (eg Fletcher et al 1986) has roughly the T dependence associated with S_g at low T. This is confirmed by the measurements of S in a heterojunction by Ruf et al (1988), in Si MOSFETs by Gallagher et al (1988), who attribute their results wholly to 'drag, and by Syme and Pepper (1989) in silicon on sapphire inversion layers. The consensus is of a dominance by phonon-drag at around liquid helium temperatures which is lost to the diffusion mechanism at higher and lower temperatures

The first detailed calculations of S_g in quasi-2D were performed by Cantrell and Butcher (1986 and 1987a,b). The expected behaviour, such as the T^3 dependence, seemed to be confirmed approximately but was complicated by more subtle effects such as the enhancement of S_g for $\bar{q}$ around $2k_f$ due to the increased electron-phonon scattering observed by Zavaritsky (1984). Their approach is based on coupled electron and phonon Boltzmann equations in the manner of Bailyn (1967) but accounts for the quasi-2D electron character. Whilst the qualitative behaviour obtained for both heterojunctions and MOSFETS agrees fairly well with the experimental data (Fletcher et al 1986, Gallagher et al 1987) the magnitudes predicted are up to forty times larger than found. This provides an interesting challenge to the understanding of S_g since with such a large discrepancy it is unclear whether the formalism can be extended and improved to provide a good description or whether a new approach is required.

In conclusion then, interest in S in general and S_g in particular has been inspired by: the wealth of information revealed about the conduction processes involved, the search for QSE in LDS, the unexpected results in strong magnetic fields and the tantalizingly good qualitative (but poor quantitative) agreement with the experimental data of the first quasi-2D S_g calculations.

1.6 The work in this thesis.

In the preceding sections the phenomenological macroscopic transport equations and the four commonly measured transport coefficients have been introduced, along with the phenomenon of phonon-drag. Some essential introductory physics of LDS and quasi-2D electron gases (2DEGs) has been described and a brief introduction given to interest in LDS. An historical summary of work leading to measurements of S in quasi-2D and some insight into why it was thought that phonon-drag may be important here have been provided. Some of the behaviour observed in the experiments has been discussed and compared with the first calculations of S_g. The resulting integral formulae are not readily interpreted in terms of simple dependences on say temperature or electron surface density n. Therefore, before considering this work in detail it is worth examining the expected behaviour of S_g which can be deduced from much simpler models. This is described in the next chapter in which the theory and its initial results are reviewed in the light of the experimental data.

In later chapters it is found that when the systems studied are represented more faithfully the formalism can describe the phonon-drag thermopower very well. In particular it is found that screening is very important but the electron confinement, additional scattering mechanisms, non-degeneracy and inelastic scattering all play significant roles.

In Chapter 3 variational approaches to obtaining an envelope function for the ground subband are described for Si MOSFETS and GaAs/GaAlAs heterojunctions. This work allows the formation of the 2DEG and the influences on its properties to be better understood and enables a better description to be used in calculations. Previous estimates of the quantum well width δ are found to be very inaccurate.

In Chapter four multi-subband screening in a quasi-2D system is discussed in the random phase approximation. An effective multi-subband screening dielectric function is derived for screening matrix elements of potentials such as from the electron-phonon interaction. The 2D and 3D limits reproduce the standard results but for general use in quasi-2D it is shown to be more convenient than the full multi-subband screening equation of Siggia and Kwok (1970).

In Chapter five the effect of further improvements to the theory are discussed such as the inclusion of temperature dependent screening, inelasticity, additional (piezoelectric) acoustic phonon scattering mechanisms, non-degeneracy and a correction for the energy dependence of the electron momentum relaxation time. The accuracy of comparisons made with the experimental data is discussed and the effect of making further simplifying approximations is investigated. The dominant phonon wavevector is defined and found to be useful in interpreting the results, which represent a significant improvement in understanding over the preceding calculations.

The final conclusions are presented in Chapter 6 along with some discussion of directions for further work.

Chapter 2

Simple models of phonon-drag in quasi-2D and the first calculations.

2.1 Introduction to the chapter.

In this chapter the calculations and initial results of Cantrell and Butcher (1987a,b), hereafter referred to as I and II, are reviewed in detail. First it is helpful to consider the information which can be obtained about S_g from simple models. In this way the nature of the problem is better understood and the model behaviour may serve as a guide to the calculations. The Boltzmann transport approach is used in which coupled electron and phonon Boltzmann equations are solved for the non-equilibrium electron and phonon distribution functions. Thus it is necessary to briefly consider some aspects of Boltzmann transport. The derivation of the S_g formula in I is then reviewed and the result compared with the predictions from the simple models. An alternative method of evaluating the formula is presented and a significant improvement over the results of II is noted.

2.2 Insight into S_g from simple models.

Simple models are useful in physics for the insight they provide into the behaviour expected from more accurate and complicated models and as a guide to methods of calculation. They may also be misleading by glossing over details. With this in mind it is interesting to inquire what insight can be gained into the behaviour of S_g without performing the full calculations. One approach is to consider the balance between the thermoelectric force (arising from ∇T) and the opposing emf (arising from $\nabla \mu_{ec}$) acting on the carriers which see a zero resultant when $J = 0$ (see equation 1.4). An alternative to this "balance approach" is the "Π approach" (Herring 1954) in which Π is calculated in order to use the relation (1.6) to obtain S. The Π approach is conceptually easier because it involves a ratio of currents (heat to charge) rather than a balance of forces. From (1.6) and (1.5) the quantity $\Pi_g (= S_g T)$ is the contribution to Π arising from the energy flux of phonons dragged by an isothermal charge flux. Either method can be used to obtain simple formulae for S_g although the results, naively, appear to be in conflict as described in what follows.

In his pioneering calculation Herring obtains a simple formula from the Π approach. (In what follows it is supposed, for convenience, that all the vectors considered are parallel and only their components in this direction are considered). The phonon heat flux Q_g (say) in response to an isothermal current density J is written:

$$Q_g = v_s^2 P \tag{2.1}$$

where v_s is an averaged speed of sound and P is a net phonon momentum density which is assumed to follow:

$$P = P_0 e^{-t/\tau_p} \tag{2.2}$$

when disturbed from equilibrium, where τ_p is the phonon momentum relaxation time (see section 2.3). On differentiating (2.2) with respect to t the net phonon force density is obtained. This is the force arising from the dragging of phonons by the charge carriers and thus P is given by:

$$\frac{P}{\tau_p} = f_{ep} N_v E e. \tag{2.3}$$

Here f_{ep} is the fraction of carrier momentum lost which is delivered to the phonons and $-N_v Ee$ is the force per unit volume acting on a gas of particles with charge $-e$ having volume density N_v. Using (2.3) in (2.1) and $-N_v e v_e$ for J, equation (1.3), with $\nabla T = 0$, gives:

$$\Pi_g = -\frac{v_s^2 f_{ep} \tau_p E}{v_e} \tag{2.4}$$

The result for S_g is obtained by using (1.6), writing the carrier drift velocity as $v_e = \mu E$, with μ the mobility, and writing the phonon mean free path L as $v_s \tau_p$. The result is:

$$S_g = -f_{ep} . \frac{L v_s}{\mu T} \tag{2.5}$$

Nicholas (1985) used such an equation to estimate S_g in a GaAs/GaAlAs heterojunction at liquid helium temperatures, although the quasi-two-dimensionality is not explicitly accounted for. At such low T, L is limited only by the size of the specimen through diffuse phonon scattering at the specimen boundaries. This is the "boundary scattering limit" (see, for example, Ziman 1963). Hence L is expected to be approximately constant and since μ is independent of N_v the principal dependences of S_g in this model arise from f_{ep}/T. However, unless f_{ep} is known, this result is difficult to interpret. If f_{ep} is assumed constant then S_g will be independent of N_v although Gallagher et al (1987) find the dependence $S_g \propto N_v^{-1}$ for the Si MOSFET case. Hence it appears that more detail is required.

An apparently different result is obtained from the balance approach which is normally used for metals (see, for example, Blatt 1968, Guenault 1971). The phonon pressure G, say, is written in terms of the phonon (lattice) internal energy density $U(T)$ as:

$$G = \frac{1}{3} U(T). \tag{2.6}$$

Differentiating with respect to displacement in the direction (x, say) of a temperature gradient, under the balance condition $J = 0$, the phonon force density is obtained. It is then supposed that some fraction f_{pe} of the force exerted by the phonons is exerted upon the carriers. Thus f_{pe} is the fraction of momentum lost by the phonons which is delivered

to the carriers and, since $J = 0$, the carrier force balance condition is then:

$$-N_v Ee = \frac{1}{3} f_{pe} C_v \frac{dT}{dx} \tag{2.7}$$

where the lattice specific heat C_v is taken as dU/dT. Comparing with (1.4) when $J = 0$, the result (Blatt 1968) for S_g is:-

$$S_g = -f_{pe} \cdot \frac{1}{3} \cdot \frac{C_v}{N_v e} \tag{2.8}$$

This is very similar to the result obtained by Guenault (1971) with f_{pe} playing the role of an averaged $\alpha(\mathbf{Q})$ (see section 1.5). The low temperature T dependence arising from C_v is modified only by f_{pe} and it seems difficult, therefore, to reconcile this "metallic" result for S_g with the Herring formula (2.5) without knowledge of f_{ep} and f_{pe}. Some attempts have been made in this direction. Zavaritsky (1984), for example, uses a balance approach applied to the metallic conduction of quasi-2D electrons coupled to 3D phonons (ie bulk) to obtain (2.8) with f_{pe} replaced by L/L_{pe} where L_{pe} is the phonon mean free path (longer than L) for phonon scattering of electrons alone. The factor L/L_{pe} can be obtained if f_{pe} is taken as the ratio of the phonon scattering rate due to scattering by electrons (τ_{pe}^{-1} say) to the total (τ_p^{-1}), ie:

$$f_{pe} = \frac{\tau_p}{\tau_{pe}} = \frac{L}{L_{pe}} \tag{2.9}$$

which follows by multiplying both the τ's, by v_s. Similarly f_{ep} can be written as:

$$f_{ep} = \frac{\tau_e}{\tau_{ep}} \tag{2.10}$$

where τ_e is the total electron momentum relaxation time from all mechanisms and $\tau_e p$ is that due to scattering of electrons by phonons. Moreover, a further expression obtained by Herring is given by using (2.9) in (2.5), ie:

$$S_g = -\frac{m v_s^2}{eT} \cdot \frac{\tau_p}{\tau_{ep}}, \tag{2.11}$$

where L is replaced by $v_s \tau_p$ and μ by $e\tau_e/m$. There still remains the problem of calculating the relaxation ratios however, in such formulae, before progress can be made.

When considering S_g in semiconductors Blatt (1968) describes another Π approach which can be applied more generally and the result, obtained more directly, is close to that from balance arguments but without using relaxation ratios. The phonon flux due to drag by an isothermal current density J is written as:

$$Q_g = U(T)v_p, \tag{2.12}$$

which defines v_p as the net phonon drift velocity for energy. Since the low temperature limit is of interest here, it follows from the T dependence of C_v that:

$$U(T) = \frac{1}{4}C_vT. \tag{2.13}$$

This result can also be obtained, for example, from the low temperature limit in the Debye model (see, for example, Kittel, 1976). The 'drag thermopower follows by substituting these last results into (1.3) with ∇T zero and N_vev_e for J and hence:

$$S_g = -\frac{1}{4}.\frac{C_v}{N_ve}.\frac{v_p}{v_e} \tag{2.14}$$

This last expression is helpful since the variation of v_p with N_v is more readily understood than f_{ep} or f_{pe} arising in (2.5) and (2.8). These two expressions should give the same result for S_g when f_{ep} and f_{pe} are written in full since although (2.8) is regarded as a metallic formula, no mention has been made of the carrier statistics in either derivation. However, writing these fractions in full defeats the object of formulae derived from simple physical ideas. It would be more helpful if f_{ep} and f_{pe} could be regarded as constants in particular limits, which must be different since one predicts a T/N_v dependence and the other a $1/T$. The quantity v_p in (2.14) helps shed light on this problem.

First consider the "saturation effect", described by Herring (1954) in both approaches. In the Π approach there can be no net phonon flux (v_p and Q_g are zero) when N_v is zero and therefore Q_g, and hence v_p, must initially increase with N_v. For low N_v it can therefore be supposed that $v_p \propto N_v$ and, from (2.14), it is then predicted that S_g should be initially independent of carrier density. By "low" of course is meant in comparison to the phonon population. This limit therefore corresponds to assuming non-degenerate

carrier statistics or a constant f_{ep} in (2.5). Suppose now that a phonon flux does exist and that N_v is increased further. Since there are more phonons with momenta parallel to the charge flow than against, collisions of carriers with phonons of opposite momenta become less frequent. The resistance offered by the lattice to the carrier flow is therefore reduced. Thus, the rate of transfer of carrier momentum to the phonons is less and, thereby, there is less drag. Hence, the greater the phonon flux (and hence v_p) the more difficult it becomes to increase, ie the drag effect (and v_p) are saturating at large N_v. The same conclusion is reached from the balance approach since when N_v is sufficiently large the phonon flux (which causes the drag) is reduced by electron scattering. Hence there is less flux to cause the drag and so it becomes more difficult to increase. Thus, for some large N_v, v_p becomes independent of N_v and (2.14) predicts that $S_g \propto N_v^{-1}$. This is the metallic limit and corresponds to assuming a constant f_{pe} in (2.8).

Reconsider now (2.9) and (2.10) for f_{pe} and f_{ep} in the light of the insight arising from saturation and v_p. For large N_v, τ_p^{-1} will be dominated by scattering from electrons and f_{pe} will approach unity. This is the conclusion reached by Guenault (1971) and does indeed give $S_g \propto N_v^{-1}$. For low N_v, τ_{ep}^{-1} is independent of N_v as there is always an excess of phonons, ie the scattering environment of the carriers is not much altered by low electron densities. Since τ_e^{-1} is also independent of N_v the conclusion is that f_{ep} is constant at low N_v and hence, from (2.5), S_g is independent of N_v. It seems therefore, that whilst (2.14) in terms of v_p is helpful in obtaining a general understanding (2.5) and (2.8) are more helpful when f_{ep} and f_{pe} are simple constants.

The case of 3D metallic conduction has been assumed in the above discussion but insight is required here into the quasi-2D case at low T. Here, the phonon heat flux is parallel to the conducting layer, which forms a very small fraction (about 10^{-6}) of the specimen cross-section. Hence τ_{pe}^{-1} is a small fraction of τ_p^{-1} which is dominated by boundary scattering and, therefore, electron scattering is unlikely to reduce the phonon heat flux enough to cause v_p to saturate. For the same reason τ_{pe}^{-1} will be independent of N_v since there is always an abundance of phonons from the bulk which are unaffected

by the carriers. Here f_{pe} in (2.10) can be usefully written as $\tau_{pe}^{-1}L/v_s$ and S_g becomes proportional to the phonon mean free path. The quantity τ_{pe}^{-1} is a measure of the phonon momentum transfer rate to electrons. This is returned to in Chapter 5 since it transpires that phonon absorption is favoured when the dominant phonon wavevector exactly crosses the Fermi circle. The electron density in (2.14) or (2.8) is a (3D) volume density but in quasi-2D the (2D) surface density n is a more natural quantity. It is necessary then, to decide whether N_v should be replaced by n/δ, to give the volume density of carriers in the channel, or n/L_z, to give the density with respect to the volume of the specimen. The answer from the Π approach must be n/L_z since in writing $\Pi = Q/J$ it has been assumed that Q is the *resultant* phonon heat flux, not merely that occurring within the channel, and J the charge flux through unit area of the *specimen*. Similarly, in the balance approach, multiplying both sides of (2.7) by the specimen volume (V) gives the balance of the *total* force on the charges. In quasi-2D the total charge $-N_vVe$ on the left side is replaced by $-nAe$ (where $AL_z = V$) but on the right side f_{pe} is still the fraction of the total phonon momentum delivered to the electron gas. The low dimensionality of the conducting channel is thereby already accounted for. Hence dividing by V it is clear that n/L_z replaces N_v. All the dependence of S_g upon the (confinement to the) channel width therefore arises from the dependence of τ_{pe}^{-1} on the width δ. Finally then, the simple expression for S_g in quasi-2D is:

$$S_g = -\frac{1}{3}.\frac{C_v}{(n/L_z)e}.\frac{L}{v_s}\tau_{pe}^{-1} \tag{2.15}$$

which in terms of simple dependencies predicts $S_g \propto LT^3/n$, to be modified by channel width dependence, and some possible enhancement due to favoured phonon absorption, arising from τ_{pe}^{-1}. Furthermore, in quasi-2D the $1/n$ behaviour is not lost at low n since τ_{pe}^{-1} is already independent of n. At an extremely high density f_{pe} and hence the $1/n$ behaviour might be affected, in principle, by saturation but such densities may not be possible in practice.

From these very simple pictures, therefore, some of the behaviour expected of S_g has

been predicted. In particular the T^3 dependence allows S to be written as:

$$S = S_d + S_g = aT + bT^3 \tag{2.16}$$

at low T. Hence at very low T, S_d (which is linear in T) will most likely dominate but at higher T, depending on the values of a and b, it will be S_g which dominates. At higher T still, L will no longer be limited by boundary scattering and will fall. Then S_g will also be reduced and S_d may again dominate. The dominance changeover $S_d \rightarrow S_g \rightarrow S_d$ is apparently that which is observed in the results of Ruf et al (1988). Whether S_g ever dominates in practice clearly depends upon the size of "b" although from the discussion in section 1.5 it appears likely that this is the case.

2.3 Some aspects of Boltzmann transport in 3D systems.

When a crystal is in uniform thermal equilibrium there can be no net flow of either heat or charge. In terms of the macroscopic transport equations, when μ_{ec} and T are constant both terms on the right side of (1.7) and (1.8) vanish. Therefore, to determine the transport coefficients a perturbation must be applied to stimulate non-vanishing current densities in response to variation in μ_{ec} and T. The disturbance is assumed to be small since in writing down (1.7), say, and defining quantities like the conductivities and thermopower in (1.3) the responses J and Q are taken as linear in the stimuli ∇T and $\nabla \mu_{ec}$. Charge flow occurs as the electrons lower energy by moving to vacant states. The distribution of charges amongst the available eigenstates under varying conditions is therefore of central concern. Labelling the eigenstates by wavevectors $\mathbf{K}$, an electron distribution function f can be defined which gives the probability that state $\mathbf{K}$ is occupied around position $\mathbf{R}$ at time t by an electron of particular spin. Spatial and temporal dependence is permitted in order to account for variation with the stimuli and the eigenstate energy $\varepsilon(\mathbf{K})$. Different temperatures for example affect $\varepsilon(\mathbf{K})$ by changing the local value of the lattice constant. In the uniform equilibrium case $f(\mathbf{K}, \mathbf{R}, t)$ reduces to the Fermi-Dirac distribution function $f(\varepsilon(\mathbf{K}))$ but in general will be perturbed.

Since Boltzmann transport is well known (see, for example, Blatt, 1968) and widely used for transport calculations in LDS (Berggren, 1988) only a brief outline is necessary here to introduce the approach used in what follows. The object is the calculation and use of $f(\mathbf{K},\mathbf{R},t)$. Although the physics is treated from an intuitive standpoint much is gained without resorting to detailed analysis of crystal structure, defects or impurities. A continuity equation is written down, and the microscopic nature of the electron system is considered in writing down the forces acting on the charges. Crystal structure is accounted for by adopting the effective mass approximation and calculating scattering rates for all the important mechanisms. The continuity (Boltzmann) equation can then be written:

$$\frac{\partial f}{\partial t} + \mathbf{v}.\nabla f - \frac{e}{\hbar}\mathbf{E}.\nabla_{\mathbf{K}} f = \left[\frac{\partial f}{\partial t}\right]_c, \tag{2.17}$$

and follows from conservation of electrons in $(\mathbf{K},\mathbf{R})$ space (Butcher 1986). Thus $\nabla_{\mathbf{K}}$ is the gradient in $\mathbf{K}$ space, $\mathbf{v}$ the electron velocity, $\mathbf{E}$ the electric field and $[\partial f/\partial t]_c$ the total rate of change due to collisions (scattering) of all types. It is this term on the right which tends to randomize the state occupancy and restore equilibrium when the applied fields are removed.

Once (2.17) is solved the electron density (N_v), current density ($\mathbf{J}$) and electronic heat flux ($\mathbf{Q}_e$) follow directly from sums over all contributing $\mathbf{K}$:

$$N_v = \frac{2}{V}\sum_{\mathbf{K}} f(\mathbf{K},\mathbf{R},t) \tag{2.18}$$

$$\mathbf{J} = \frac{-2e}{V}\sum_{\mathbf{K}} f(\mathbf{K},\mathbf{R},t)\mathbf{v}(\mathbf{K}) \tag{2.19}$$

$$\mathbf{Q}_e = \frac{2}{V}\sum_{\mathbf{K}} f(\mathbf{K},\mathbf{R},t)[\epsilon(\mathbf{K}) - \epsilon_f]\mathbf{v}(\mathbf{K}) \tag{2.20}$$

These equations can be written in integral form when the possible wavevectors form a quasi-continuum by writing the sum over $\mathbf{K}$ as $\int d\mathbf{K}.V/(2\pi)^3$. Here and elsewhere the density of states of particular spin in $\mathbf{K}$ space is taken as $1/(2\pi)^3$ per unit volume (V) and $f(\mathbf{K},\mathbf{R},t)$ is assumed the same for both spin orientations. Hence the sums and integrals are taken merely over the wavevectors and "state $\mathbf{K}$" is understood to mean with given spin, accounted for by taking $2f(\mathbf{K},\mathbf{R},t)$ for the state occupancy.

To solve the Boltzmann equation (2.17) some simplification is necessary. In the limit of weak gradients in μ_{ec} and T, f can be linearized as:

$$f = f_0 + f_1 \tag{2.21}$$

Furthermore, if the scattering rates for state **K** due to the various mechanisms are assumed proportional to the perturbation (f_1) from equilibrium (f_0), the relaxation time (τ) approximation arises whereby:

$$\left[\frac{\partial f}{\partial t}\right]_{c,i} = -\frac{f_1}{\tau_i}. \tag{2.22}$$

Here the index i labels the contribution made to $[\partial f/\partial t]_c$ by the particular mechanism. It is this approximation which is implicit in the discussion of the previous section and which will be used later to solve the Boltzmann equation in the systems of interest. For further discussion on Boltzmann transport and the use and validity of the relaxation time approximation (ie when the scattering is either elastic or randomizing) see, for example, Butcher (1986). Here it is shown that the method can provide insight even when the approximation is not strictly valid.

2.4 The derivation of an expression for S_g in quasi-2D.

Before considering the application of and extensions and improvements to the preceding calculations (I and II) it is necessary to consider the foundation and derivation of the general formula for S_g derived in I. This is applied to particular cases by making certain assumptions and approximations in II. Since the results are found to be qualitatively encouraging but with much larger magnitudes than found it must be first established that there are no trivial numerical errors or difficulties with the initial formalism. Hence, in this section the derivation of the formula for the phonon-drag thermoelectric power of quasi-2D electrons arising from 3D phonons, of I, is considered in detail.

The initial approach is to consider the relation between the steady 2D electric current density ($\mathbf{J}_{2D}$) and the temperature gradient when ∇T is applied along the conducting

plane (x, y) of the specimen. If the current is allowed to flow freely around a closed circuit then, when steady flow is reached, there can be no net emf. Hence from (1.4):

$$\mathbf{J}_{2D} = -\sigma_{2D} S \nabla T, \tag{2.23}$$

can be written, in which σ_{2D} is the 2D electrical conductivity. This current density can also be obtained from a sum such as in (2.19) for $\mathbf{J}$ by replacing $\mathbf{K}$ with (α,$\mathbf{k}$), as in section 1.3, and integrating across the channel (z) weighted by $\mid \phi_\alpha(z) \mid^2$ (see (1.10)) to give:

$$\mathbf{J}_{2D} = -2e \sum_{\alpha,\mathbf{k}} f_\alpha(\mathbf{k}) \mathbf{v}_{\alpha,\mathbf{k}} \tag{2.24}$$

Here and hereafter the spatial dependence of $f_\alpha(\mathbf{k})$ is left as understood. Hence the integral over 3D wavevectors is replaced by $\sum_\alpha \int d\mathbf{k}.A/(2\pi)^2$. The thermopower follows by comparing (2.23) with (2.24) and S_g is that part of S which arises from the departure of the phonon distribution from equilibrium.

The Boltzmann transport approach of the previous section is used to determine $f_\alpha(\mathbf{k})$ in the manner of Bailyn (1967). In the steady state of interest the Boltzmann equation is:

$$\left[\frac{\partial f_\alpha(\mathbf{k})}{\partial t}\right]_c - \mathbf{v}_\alpha(\mathbf{k}).\nabla f_\alpha(\mathbf{k}) = 0 \tag{2.25}$$

since both $\mathbf{E}$ and $\partial f_\alpha(\mathbf{k})/\partial t$ are zero. Hence the rate of change due to collisions (c) is balanced by that due to drift (diffusion) ie ∇T. The scattering of electrons by phonons is known to be weak (Hensel et al 1983) but is responsible for the drag. Hence three contributions to the scattering are accounted for: phonon absorption (a) and emission (e), and collisions with static defects (s); ie:

$$\left[\frac{\partial f_\alpha(\mathbf{k})}{\partial t}\right]_c = \left[\frac{\partial f_\alpha(\mathbf{k})}{\partial t}\right]_a + \left[\frac{\partial f_\alpha(\mathbf{k})}{\partial t}\right]_e + \left[\frac{\partial f_\alpha(\mathbf{k})}{\partial t}\right]_s \tag{2.26}$$

The last term accounts for all stationary scatterers such as dopants, ionized impurities and crystal imperfections. The exact form of such contributions is unimportant except that the term is dominant and can be written in a relaxation-time approximation (2.22) as:

$$\left[\frac{\partial f_\alpha(\mathbf{k})}{\partial t}\right]_s = -\frac{f_\alpha^1(\mathbf{k})}{\tau_\alpha(\mathbf{k})}. \tag{2.27}$$

where:

$$f_\alpha(\mathbf{k}) - f_\alpha^0(\mathbf{k}) = f_\alpha^1(\mathbf{k}) \equiv -g_\alpha(\mathbf{k})\frac{df_\alpha^0(\mathbf{k})}{d\varepsilon_\alpha(\mathbf{k})} \tag{2.28}$$

This is certainly valid when ionized impurity scattering dominates as is normally the case under the low temperature conditions of interest. Interface roughness scattering (Ando et al 1982) can also be important in the MOSFET (Stern 1980) and in very narrow channels in the heterojunction case (Sakaki et al 1987) but the particular mechanism affects only the value of τ, not the validity of the approximation, providing the scattering remains elastic or randomizing (Butcher 1986).

To obtain the two phonon terms in (2.26) the electron transition rates due to acoustic phonon scattering are calculated. Optical phonons are neglected since it is the low temperature limit which is of interest. The electron-phonon coupling is described by a spherically symmetric longitudinal acoustic phonon deformation potential E_1:

$$U(\mathbf{r}, z) = E_1 \nabla . \mathbf{u}(\mathbf{r}, z) \tag{2.29}$$

where U is the interaction energy and $\mathbf{u}(\mathbf{r}, z)$ is the usual expression (Jensen 1984, Kittel 1987) for the lattice displacement from equilibrium due to acoustic phonons, written:

$$\mathbf{u}(\mathbf{r}, z) = \sum_{\mathbf{Q}} \left(\frac{\hbar}{2\rho V \omega_{\mathbf{Q}}}\right)^{1/2} \mathbf{n}_{\mathbf{Q}} \left[a_{\mathbf{Q}}^\dagger e^{i\mathbf{q}.\mathbf{r}} e^{iq_z z} + a_{\mathbf{Q}} e^{-i\mathbf{q}.\mathbf{r}} e^{-iq_z z}\right] \tag{2.30}$$

Here $n_{\mathbf{Q}}$ is the polarization vector of phonons with wavevector $\mathbf{Q} = (\mathbf{q}, q_z)$, frequency $\omega_{\mathbf{Q}}$ and annihilation and creation operators $a_{\mathbf{Q}}$ and $a_{\mathbf{Q}}^\dagger$, and ρ is the density of the bulk. In I the general S_g expression is derived without including any contribution from the two transverse acoustic phonon modes or from other mechanisms of acoustic phonon absorption such as the piezoelectric interaction. In II this is generalized to allow for the coupling of electrons with the transverse modes which is possible in Si because of the anisotropic bands (see, for example, Ridley 1982) as described in section 2.6. Moreover, the author (Smith and Butcher 1989a) and Lyo (1988) both point out the importance of the piezoelectric acoustic phonon scattering mechanism (discussed in Chapter 5).

Consider, first, the longitudinal deformation potential contribution which is the simplest. The results are readily generalized to the other cases. The transition rate for the

system when one phonon $\mathbf{Q}$ is absorbed (and thereby destroyed) by an electron in state $(\alpha, \mathbf{k})$ which is promoted to $(\beta, \mathbf{k}')$, is obtained from the Golden Rule. The result is written as $P^a_{\mathbf{Q}}(\alpha,\beta)$ where:

$$P^a_{\mathbf{Q}}(\alpha,\beta) = \frac{\pi E_1^2 Q^2 N_{\mathbf{Q}}}{\rho V \omega_{\mathbf{Q}}} \mid Z_{\alpha\beta}(q_z) \mid^2 \delta(\varepsilon_\beta - \varepsilon_\alpha - \hbar\omega_{\mathbf{Q}})\delta_{\mathbf{k}',\mathbf{k}+\mathbf{q}} \tag{2.31}$$

and:

$$Z_{\alpha\beta}(q_z) = \int \phi^*_\beta(z) e^{iq_z z} \phi_\alpha(z) dz. \tag{2.32}$$

(For clarity the labels $\mathbf{k}$ and $\mathbf{k}'$ have been dropped from α and β which they always accompany respectively, but they are reinstated in the final formulae). The result for $[\partial f_\alpha/\partial t]_a$ is obtained from the difference of $P^a_{\mathbf{Q}}(\alpha,\beta)$ and the competing process, $P^a_{\mathbf{Q}}(\beta,\alpha)$, by accounting for state occupancy and summing over all $(\beta, \mathbf{k}')$ and $\mathbf{Q}$, ie:

$$\left[\frac{\partial f_\alpha}{\partial t}\right]_a = \sum_{\beta,\mathbf{Q}} f_\beta[1 - f_\alpha]P^a_{\mathbf{Q}}(\beta,\alpha) - f_\alpha[1 - f_\beta]P^a_{\mathbf{Q}}(\alpha,\beta). \tag{2.33}$$

The phonon emission case follows directly and $P^e_{\mathbf{Q}}(\alpha,\beta)$ is obtained from (2.31) by changing the signs of $\hbar\omega_{\mathbf{Q}}$ and $\mathbf{q}$ in the delta symbols and putting $N_{\mathbf{Q}} + 1$ for $N_{\mathbf{Q}}$.

The coupled nature of the electron and phonon distribution functions is evident from the presence of $N_{\mathbf{Q}}$ in (2.31). Since S_g results from the departure of $N_{\mathbf{Q}}$ from the equilibrium value $N^0_{\mathbf{Q}}$ (the Planck distribution) the phonon Boltzmann equation is used to find $N_{\mathbf{Q}}$ and this determines $f_\alpha(\mathbf{k})$. Following similar arguments to those used above, for the electron case, the phonon result is:

$$\left[\frac{\partial N_{\mathbf{Q}}}{\partial t}\right]_c - \mathbf{v}_p(\mathbf{Q}).\nabla N_{\mathbf{Q}} = 0 \tag{2.34}$$

in which:

$$\left[\frac{\partial N_{\mathbf{Q}}}{\partial t}\right]_c = \left[\frac{\partial N_{\mathbf{Q}}}{\partial t}\right]_a + \left[\frac{\partial N_{\mathbf{Q}}}{\partial t}\right]_e + \left[\frac{\partial N_{\mathbf{Q}}}{\partial t}\right]_p \tag{2.35}$$

where:

$$\left[\frac{\partial N_{\mathbf{Q}}}{\partial t}\right]_{a(e)} = -(+)2 \sum_{\alpha\mathbf{k},\beta\mathbf{k}'} f_\alpha[1 - f_\beta]P^{a(e)}_{\mathbf{Q}}(\alpha,\beta) \tag{2.36}$$

and:

$$\left[\frac{\partial N_{\mathbf{Q}}}{\partial t}\right]_p = -\frac{N^1_{\mathbf{Q}}}{\tau_{pp}(\mathbf{Q})} \equiv \frac{G_{\mathbf{Q}}}{\tau_{pp}(\mathbf{Q})} \frac{dN^0_{\mathbf{Q}}}{d\hbar\omega_{\mathbf{Q}}} \tag{2.37}$$

with:

$$N_{\mathbf{Q}} - N^0_{\mathbf{Q}} = N^1_{\mathbf{Q}} \equiv -G_{\mathbf{Q}} \frac{dN^0_{\mathbf{Q}}}{d\hbar\omega_{\mathbf{Q}}} \tag{2.38}$$

Here the subscripts denote phonon adsorption (a) and emission (e) by electrons and the scattering by all other mechanisms (p), such as boundary (or phonon-phonon) scattering which dominates. The first p subscript on τ indicates that it is a phonon relaxation time.

By collecting and combining all the separate contributions, the collision terms can be written to linear order, in the form:

$$\left[\frac{\partial f_\alpha}{\partial t}\right]_c = \frac{df^0_\alpha}{d\varepsilon_\alpha}\frac{g_\alpha}{\tau_\alpha} + \frac{1}{K_B T}\sum_{\beta,\mathbf{Q}}(g_\beta - g_\alpha)(\Gamma_{\beta\alpha} + \Gamma_{\alpha\beta}) - \frac{G_{\mathbf{Q}}}{K_B T}\sum_{\beta,\mathbf{Q}}(\Gamma_{\beta\alpha} - \Gamma_{\alpha\beta}) \tag{2.39}$$

and

$$\left[\frac{\partial N_{\mathbf{Q}}}{\partial t}\right]_c = \frac{dN^0_{\mathbf{Q}}}{d\hbar\omega_{\mathbf{Q}}}\frac{G_{\mathbf{Q}}}{\tau_{pp}(\mathbf{Q})} + \frac{2}{K_B T}\sum_{\alpha\beta}(g_\beta - g_\alpha)\Gamma_{\beta\alpha} - \frac{2G_{\mathbf{Q}}}{K_B T}\sum_{\alpha\beta}\Gamma_{\beta\alpha}. \tag{2.40}$$

The quality $\Gamma_{\beta\alpha}$ is the equilibrium transition rate from α to β by phonon absorption:

$$\Gamma_{\beta\alpha} = f^0_\alpha[1 - f^0_\beta]P^{a0}_{\alpha\beta}(\mathbf{Q}). \tag{2.41}$$

In deriving (2.39) and (2.40) the "detailed balance principle" (see Lax 1974) has been used whereby:

$$f^o_\alpha(1 - f^0_\beta)P^{a0}_{\alpha\beta} = f^0_\beta(1 - f^0_\alpha)P^{e0}_{\beta\alpha}. \tag{2.42}$$

This principle states that in equilibrium (signified by the 0 superscript) the scattering rate for any transition is exactly balanced by its opposite. This enables much cancellation in deriving the collision terms above. The two Boltzmann equations are now equations for g_α and $G_{\mathbf{Q}}$ but since $G_{\mathbf{Q}}$ appears in $[\partial N_{\mathbf{Q}}/\partial t]_c$ only as a numerical factor, with a coefficient $-F$, say, it can be written as:

$$G_{\mathbf{Q}} = \frac{1}{F}\left\{-\left[\frac{dN_{\mathbf{Q}}}{dt}\right]_c + \frac{2}{K_B T}\sum_{\alpha\beta}(g_\beta - g_\alpha)\Gamma_{\beta\alpha}\right\} \tag{2.43}$$

with:

$$F = -\frac{dN^0_{\mathbf{Q}}}{d\hbar\omega_{\mathbf{Q}}}\frac{1}{\tau_{pp}(\mathbf{Q})} + \frac{2}{K_B T}\sum_{\beta\alpha}\Gamma_{\beta\alpha} \tag{2.44}$$

Hence the electron Boltzmann equation becomes:

$$(L_1 + L_2 + L_3)g_\alpha - \left[\frac{\partial f_\alpha}{\partial t}\right]_c + U_\alpha = 0 \tag{2.45}$$

where:

$$U_\alpha = \frac{1}{K_B T} \sum_{\beta,\mathbf{Q}} \frac{1}{F} (\Gamma_{\beta\alpha} - \Gamma_{\alpha\beta}) \left[\frac{\partial N_\mathbf{Q}}{\partial t}\right]_c . \tag{2.46}$$

The terms $L_1 g_\alpha$ and $L_2 g_\alpha$ are the first two terms of (2.39) and $L_3 g_\alpha$ is given by:

$$L_3 g_\alpha = -\frac{2}{(K_B T)^2} \sum_{\beta\alpha} \left[\frac{1}{F} (\Gamma_{\beta\alpha} - \Gamma_{\alpha\beta}) \sum_{\alpha'\beta'} \Gamma_{\alpha'\beta'} (g_{\beta'} - g_{\alpha'})\right] . \tag{2.47}$$

With no electron-phonon coupling there can be no phonon-drag, and $G_\mathbf{Q}$ is zero. Moreover $N_\mathbf{Q}$ is $N^0_\mathbf{Q}$ and $[\partial N_\mathbf{Q}/\partial t]_c$ is zero so that U_α and $L_3 g_\alpha$ (through (2.43)) both vanish and (2.45) reduces to the conventional linearized Boltzmann equation. Thus U_α and $L_3 g_\alpha$ describe the phonon-drag but since the electron scattering rate due to static defects (ie $L_1 g_\alpha$) is dominant (see (2.26) and (2.27)) L_2 and L_3 can both be dropped from (2.45).

From the expressions for $\mathbf{J}_{2D}$, using (2.28):

$$S\frac{dT}{dx} = -\frac{2e}{A\sigma} \sum_\alpha g_\alpha \frac{df^0_\alpha}{d\varepsilon_\alpha} v^x_\alpha , \tag{2.48}$$

can be written for a temperature gradient in the x direction. There is no contribution to $\mathbf{J}_{2D}$ from the equilibrium part (f^0_α) of f_α because there can be no net current in equilibrium. Recalling the definition of $L_1 g_\alpha$ and substituting into equation (2.49) using the final electron Boltzmann equation, neglecting the diffusion term (equal to $[\partial f_\alpha/\partial t]_c$ from (2.25)) to obtain S_g and substituting for U_α from (2.46) the result is:

$$S_g = -\frac{2e}{A\sigma K_B T} \sum_\alpha \tau_\alpha v^x_\alpha \sum_{\beta\mathbf{Q}} \frac{1}{F} (\Gamma_{\beta\alpha} - \Gamma_{\alpha\beta}) v^x_p \frac{dN^0_\mathbf{Q}}{d\hbar\omega_\mathbf{Q}} \frac{\hbar\omega_\mathbf{Q}}{K_B T} . \tag{2.49}$$

The quantity F consists of the two terms in (2.44) although the first dominates, as shown in the next section, and hence the second can be dropped. Further simplifying the result by interchanging α and β in one of the two terms and adding the same result in (2.50) to both sides, but with x replaced by y, and dividing by two, and finally restoring the full notation the final result for S_g as in II, is:

$$S_g = \frac{e}{A\sigma K_B T^2} \sum_{\alpha\mathbf{k}} \sum_{\beta\mathbf{k}'} \sum_\mathbf{Q} \hbar\omega_\mathbf{Q} f^0_\alpha(\mathbf{k})[1 - f^0_\beta(\mathbf{k}')] P^{a0}_\mathbf{Q}(\alpha\mathbf{k}, \beta\mathbf{k}' \tau_{pp}(\mathbf{Q}) \mathbf{v}_p(\mathbf{Q})$$
$$\cdot [\tau_\alpha(\mathbf{k}) \mathbf{v}_\alpha(\mathbf{k}) - \tau_\beta(\mathbf{k}' \mathbf{v}_\beta(\mathbf{k}')] . \tag{2.50}$$

2.5 Some comparisons with the simple models.

Before the final formula of the last section is applied and evaluated in particular cases it is interesting to ask how far it agrees with the simple models of Section 2.2. An answer to this has been developed in conjunction with the authors' supervisor and is described in this section. Most convenient for this purpose is the 3D case which is obtained from (2.48) by replacing (α, $\mathbf{k}$) by $\mathbf{K}$, (β,$\mathbf{k}'$) by $\mathbf{K}'$, A by V and by performing a generalization corresponding to that leading to (2.50). The sum over $\mathbf{K}'$ can be dropped by replacing $\mathbf{K}'$ by $\mathbf{K}+\mathbf{Q}$, following the delta symbols in (2.40). Simplification is possible if $\tau_{\mathbf{K}}$ is assumed constant (τ_e) for all $\mathbf{K}$, which are 3D plane wave states. Hence $\mathbf{v}(\mathbf{K})$ is simply $\hbar\mathbf{K}/m$, the 3D conductivity is written $N_v e^2\tau_e/m$ and $\mathbf{v}_p(\mathbf{Q})$ is taken as $v_s\mathbf{Q}/Q$ to give:

$$S_g = \frac{2}{3N_v eVK_BT^2}\sum_{\mathbf{K},\mathbf{Q}}\frac{(\hbar\omega_{\mathbf{Q}})^2\Gamma_{\mathbf{K}+\mathbf{Q},\mathbf{K}}}{F}\frac{dN^0_{\mathbf{Q}}}{d\hbar\omega_{\mathbf{Q}}} \tag{2.51}$$

Writing C_v in the form:

$$C_v = \frac{1}{V}\sum_{\mathbf{Q}}\hbar\omega_{\mathbf{Q}}\frac{dN^0_{\mathbf{Q}}}{dT} \tag{2.52}$$

S_g can be reduced to:

$$S_g = -\frac{1}{3}\frac{C_v}{N_v e}\overline{\alpha} \tag{2.53}$$

when $\alpha(\mathbf{Q})$ is defined by:

$$\alpha(\mathbf{Q}) = \frac{\tau_{pe}^{-1}(\mathbf{Q})}{\tau_{pe}^{-1}(\mathbf{Q})+\tau_{pp}^{-1}(\mathbf{Q})} \tag{2.54}$$

with the phonon scattering rate due to absorption and emission by electrons given by:

$$\tau_{pe}^{-1}(\mathbf{Q}) = \sum_{\mathbf{K}}2\Gamma_{\mathbf{K}+\mathbf{Q},\mathbf{K}} \Big/ -K_BT\frac{dN^0_{\mathbf{Q}}}{d\hbar\omega_{\mathbf{Q}}} \tag{2.55}$$

and:

$$\overline{\alpha} = \frac{1}{C_vV}\sum_{\mathbf{Q}}\hbar\omega_{\mathbf{Q}}\frac{dN^0_{\mathbf{Q}}}{dT}\alpha(\mathbf{Q}). \tag{2.56}$$

Thus the general quasi-2D S_g formula reproduces the "metallic" formula discussed in Section 2.2 when the relaxation time and average over $\mathbf{Q}$ are defined as above. The argument used here to derive (2.53), however, is much stronger than those used to derive the simple formulae and, furthermore, the various τ make it possible to determine the conditions under which the Herring formula may also be valid.

Writing $\Gamma_{\mathbf{K+Q,K}}$ out in full in equation (2.51) S_g can be written in the form:

$$S_g = -\frac{2}{3}\frac{1}{N_v eTV}\sum_{\mathbf{K}} f(\mathbf{K})[1-f(\mathbf{K}+\mathbf{Q})]\frac{\overline{\tau_p}_{\mathbf{K}}\overline{(\hbar\omega_{\mathbf{Q}})^2}_{\mathbf{K}}}{\tau_{ep}(\mathbf{K})}. \tag{2.57}$$

Here, an averaged total phonon scattering relaxation time $\overline{\tau_p}_{\mathbf{K}}$ has been defined by:

$$\overline{\tau_p}_{\mathbf{K}} = \frac{\sum_{\mathbf{Q}}(\hbar\omega_{\mathbf{Q}})^2 P^{a0}(\mathbf{K},\mathbf{K}+\mathbf{Q})\tau_p(\mathbf{Q})}{\sum_{\mathbf{Q}}(\hbar\omega_{\mathbf{Q}})^2 P^{a0}(\mathbf{K},\mathbf{K}+\mathbf{Q})} \tag{2.58}$$

where:

$$\tau_p^{-1}(\mathbf{Q}) = \tau_{pp}^{-1}(\mathbf{Q}) + \tau_{pe}^{-1}(\mathbf{Q}) \tag{2.59}$$

and is the total phonon scattering rate given by $F/[-dN^0_{\mathbf{Q}}/d\hbar\omega_{\mathbf{Q}}]$ and:

$$\overline{(\hbar\omega_{\mathbf{Q}})^2}_{\mathbf{K}} = \tau^a_{ep}(\mathbf{K})\sum_{\mathbf{Q}}(\hbar\omega_{\mathbf{Q}})^2 P^{a0}(\mathbf{K},\mathbf{K}+\mathbf{Q}) \tag{2.60}$$

in which:

$$\frac{1}{\tau^a_{ep}(\mathbf{K})} = \sum_{\mathbf{Q}} P^{a0}(\mathbf{K},\mathbf{K}+\mathbf{Q}), \tag{2.61}$$

and is the total electron scattering rate in state $\mathbf{K}$ by phonon absorption. For both the non-degenerate and degenerate limits this allows S_g to be written in a form similar to the Herring formula (2.11) providing the averages over $\mathbf{Q}$, at given $\mathbf{K}$, and the scattering times in (2.57), can be replaced by constant average values and moved to in front of the summation ie:-

$$S_g = -\frac{2}{3}\frac{1}{N_v eTV}\frac{\tau_p}{2\tau_{ep}}\overline{(\hbar\omega_{\mathbf{Q}})^2}_{\mathbf{K}}\sum_{\mathbf{K}} f(\mathbf{K})[1-f(\mathbf{K}+\mathbf{Q})]. \tag{2.62}$$

where:

$$\tau^{(e)-1}_{ep} + \tau^{(a)-1}_{ep} = \tau^{-1}_{ep} \approx 2\tau^{(a)-1}_{ep}, \tag{2.63}$$

is assumed. For the non-degenerate limit it is assumed that $\overline{(\hbar\omega_{\mathbf{Q}})^2}_{\mathbf{K}} \approx 2mv_s^2 K_B T$. Since $f(\mathbf{K}) \ll 1$, $f(\mathbf{K}+\mathbf{Q})$ can be dropped and on using (2.18) for N_v the result is:

$$S_{g(non-deg)} \approx -\frac{1}{3}\frac{mv_s^2}{eT}\frac{\overline{\tau_p}}{\overline{\tau_{ep}}}. \tag{2.64}$$

For the degenerate case $\overline{(\hbar\omega_{\mathbf{Q}})^2}_{\mathbf{K}} \approx 2mv_s^2\varepsilon_f$. Assuming elastic scattering, whereby $\varepsilon(\mathbf{K}') = \varepsilon(\mathbf{K})$, and differentiating (2.18) for N_v to obtain:

$$\frac{1}{V}\sum_{\mathbf{K}} f(\mathbf{K})[1-f(\mathbf{K})] = \frac{1}{2}K_B T\frac{dN_v}{d\varepsilon_f}, \tag{2.65}$$

and taking $N_v \propto \varepsilon_f{}^{3/2}$, the result for S_g reduces to:

$$S_{g(deg)} \approx -\frac{1}{2}\frac{mv_s^2}{eT}\frac{\overline{\tau_p}}{\overline{\tau_{ep}}}. \tag{2.66}$$

Hence it seems that the more general formula of the previous section can reproduce the expected results, when the averaged τ's are appropriately defined. Both the metallic formula (2.8) and Herring formula (2.11) are obtained, approximately, although it seems that the latter is valid only for the non-degenerate limit, or the degenerate limit when the scattering is elastic. These results support the validity of the approach described in the previous section. The difference between (2.64) and (2.66) in terms of trivial numerical constants is of no significance in view of the approximate averaging used in their derivation.

2.6 Initial applications and results.

The first application of I appeared in II but an alternative method of evaluating the final formula gives conflicting results from the same data (see Smith and Butcher 1989a). Therefore, some care is taken in this section to point out the approximations and simplifications used, some of which are expanded and generalized in later chapters. The case in which the quasi-2DEG is accommodated in GaAs (GaAs/GaAlAs heterojunction) is treated first. The result is then adapted to account for the anisotropy of the Si MOSFET case). The initial simplifications used are:

1. Quantum limit. From Figure 1.2 it is apparent that when the conducting channel is narrow, providing that n is not too large ($\varepsilon_f < \varepsilon_2$) and that T is low ($K_BT \ll \varepsilon_2 - \varepsilon_f$), the probability of occupying any state in a subband above ground is so small that the higher subbands may be neglected. Hence subband labels are dropped and the calculations performed in the quantum limit.

2. Boundary Scattering. When T is low enough $L(\mathbf{Q}) = v_s\tau_{pp}(\mathbf{Q})$ is limited, for all $\mathbf{Q}$, by the dimensions of the sample and can be taken as constant, L.

3. Debye Phonons. Since only values of $T \ll \Theta_D$ (Debye temperature, eg Kittel 1976) are of concern, Debye phonons are assumed for which $\hbar\omega_{\mathbf{Q}} = v_s\mathbf{Q}$.

4. $\tau(\mathbf{k}) \rightarrow \tau(\varepsilon_f)$. The electron (static impurity) relaxation-time is assumed to be a function of $\varepsilon(\mathbf{k})$ only. Then, since $f(\varepsilon(\mathbf{k}))[1 - f(\varepsilon(\mathbf{k}'))]$ is very peaked near ε_f at low T, $\tau(\varepsilon)$ is replaced by, the constant, $\tau(\varepsilon_f)$.

In addition, the Kronecker delta of (2.31) allows the sum over $\mathbf{k}'$ to be dropped when $\mathbf{k}+\mathbf{q}$ is substituted for $\mathbf{k}'$; the 2D conductivity is written as $ne^2\tau(\varepsilon_f)/m$, $\mathbf{v}(\mathbf{k})$ as $\hbar\mathbf{k}/m$ and $\mathbf{v}_p(\mathbf{Q})$ as $v_s\mathbf{Q}/Q$ and the result for S_g becomes:

$$S_g = \frac{-\pi\hbar^2 E_1^2 L}{ne\rho A K_B T^2 V} \sum_{\mathbf{Q}} Q q^2 N_{\mathbf{Q}}^0 \mid Z_{11}(q_z) \mid^2 \sum_{\mathbf{k}} f^0(\varepsilon(\mathbf{k}))[1 - f^0(\varepsilon(\mathbf{k}+\mathbf{q}))] \\ .\delta(\varepsilon(\mathbf{k}+\mathbf{q}) - \varepsilon(\mathbf{k}) - \hbar\omega_{\mathbf{Q}}). \qquad (2.67)$$

To make progress an approximation is made for:

5. State Occupancy factors : whereby $\varepsilon(\mathbf{k}+\mathbf{q})$ is replaced by $\varepsilon(\mathbf{k}) + \hbar\omega_{\mathbf{Q}}$ owing to the Dirac delta function in the above, and:

$$f^0(\varepsilon(\mathbf{k}))[1 - f^0(\varepsilon(\mathbf{k}) + \hbar\omega_{\mathbf{Q}})] \approx W(\mathbf{Q})\delta(\varepsilon(\mathbf{k}) - \varepsilon_f) \qquad (2.68)$$

is written. The weighting $W(\mathbf{Q})$ is determined by integrating this approximation over $\varepsilon(\mathbf{k})$, which can be performed exactly. The result is:

$$W(\mathbf{Q}) = \frac{K_B T}{e^{-\gamma} - 1} \ln\left[\frac{e^{-\varepsilon_f/K_B T} + e^{-\hbar\omega_{\mathbf{Q}}/K_B T}}{1 + e^{-\varepsilon_f/K_B T}}\right] \qquad (2.69)$$

where γ is $\hbar\omega_{\mathbf{Q}}/K_B T$. Hence, for the low temperatures of interest $W(\mathbf{Q})$ can be approximated by:

$$W(\mathbf{Q}) \doteq \frac{\hbar\omega_{\mathbf{Q}}}{1 - e^{-\gamma}} \qquad (2.70)$$

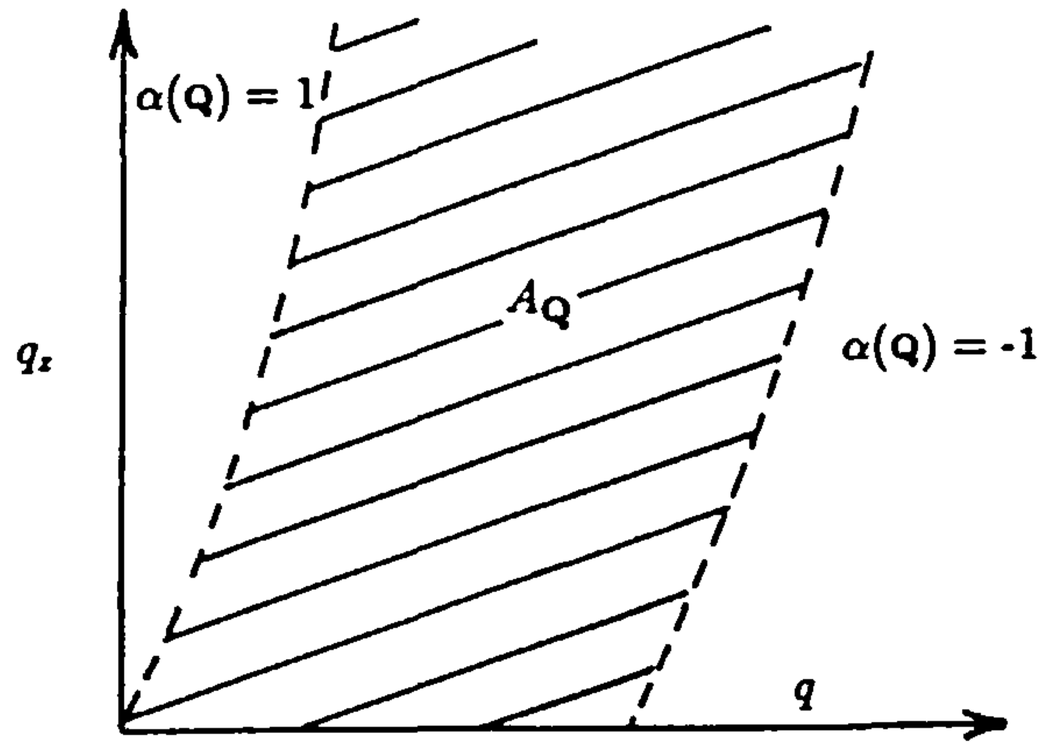
The sums are now transformed to integrals and the integral over $\mathbf{k}$ is performed by converting to an energy integral by writing $\varepsilon(\mathbf{k})$ as $\hbar^2\mathbf{k}.\mathbf{k}/2m$. The result can be written as:-

$$S_g = \frac{-g_v L m^2 E_1^2 v_s}{4(2\pi)^3 ne\rho K_B T^2 \hbar\rho k_f} \int_{A_{\mathbf{Q}}} \frac{\mid Z_{11}(q_z) \mid^2 q^2 Q^2 \qquad dq\, dq_z}{(\sinh^2\gamma/2)\sqrt{1 - \alpha(\mathbf{Q})^2}} \qquad (2.71)$$

where g_v is the valley degeneracy and the integration field is determined by the set (q, q_z) for which the argument of the Dirac delta function in (2.31) is zero. Thus $A_{\mathbf{Q}}$ spans the set (q, q_z) which satisfies $\mid \alpha(\mathbf{Q}) \mid < 1$ where:

$$\alpha(\mathbf{Q}) = \frac{2m\hbar\omega_{\mathbf{Q}} - \hbar^2 q^2}{\hbar^2 q k_f} \qquad (2.72)$$

Figure 2.1: A schematic diagram of the field of integration.



The S_g integral (2.72) is performed over the field defined by $|\alpha(\mathbf{Q})| < 1$ and is bounded by the chain curves along which $|\alpha(\mathbf{Q})| = 1$, where the integrand becomes singular.

which is illustrated schematically in Figure 2.1. The integral (2.71) must now be performed numerically but care is necessary because the integrand becomes singular all along the boundaries $|\alpha(\mathbf{Q})| = 1$ defining $A_\mathbf{Q}$, as shown in the figure. In II a Lorentzian is introduced, to broaden the energy delta function in (2.67), of width Γ taken to be small enough for S_g to be independent of its precise value ie:

$$\delta(x) = \lim_{\Gamma \to 0} \frac{\Gamma}{\pi(x^2 + \Gamma^2)} \tag{2.73}$$

An alternative (Smith and Butcher 1989a) is to evaluate the integral directly as it stands taking particular care towards the boundaries.

In II the result is generalized to apply to Si by accounting for both LA and TA phonon modes which couple differently to the electrons. This is treated following Ridley (1982) by taking E_1 under the integral signs and replacing it with $\Xi_u(q_z^2/Q^2 + D)$ and $\Xi_u q q_z/Q^2$ for LA and TA phonon modes respectively. Here Ξ_u is the deformation potential for pure shear strain and $D = \Xi_d/\Xi_u$, with Ξ_d denoting that for pure dilation. The resultant S_g is then the sum of the two contributions from separate integrals for each mode, of the form (2.71) with the corresponding averaged speeds of sound (v_L and v_T) and replacements for E_1.

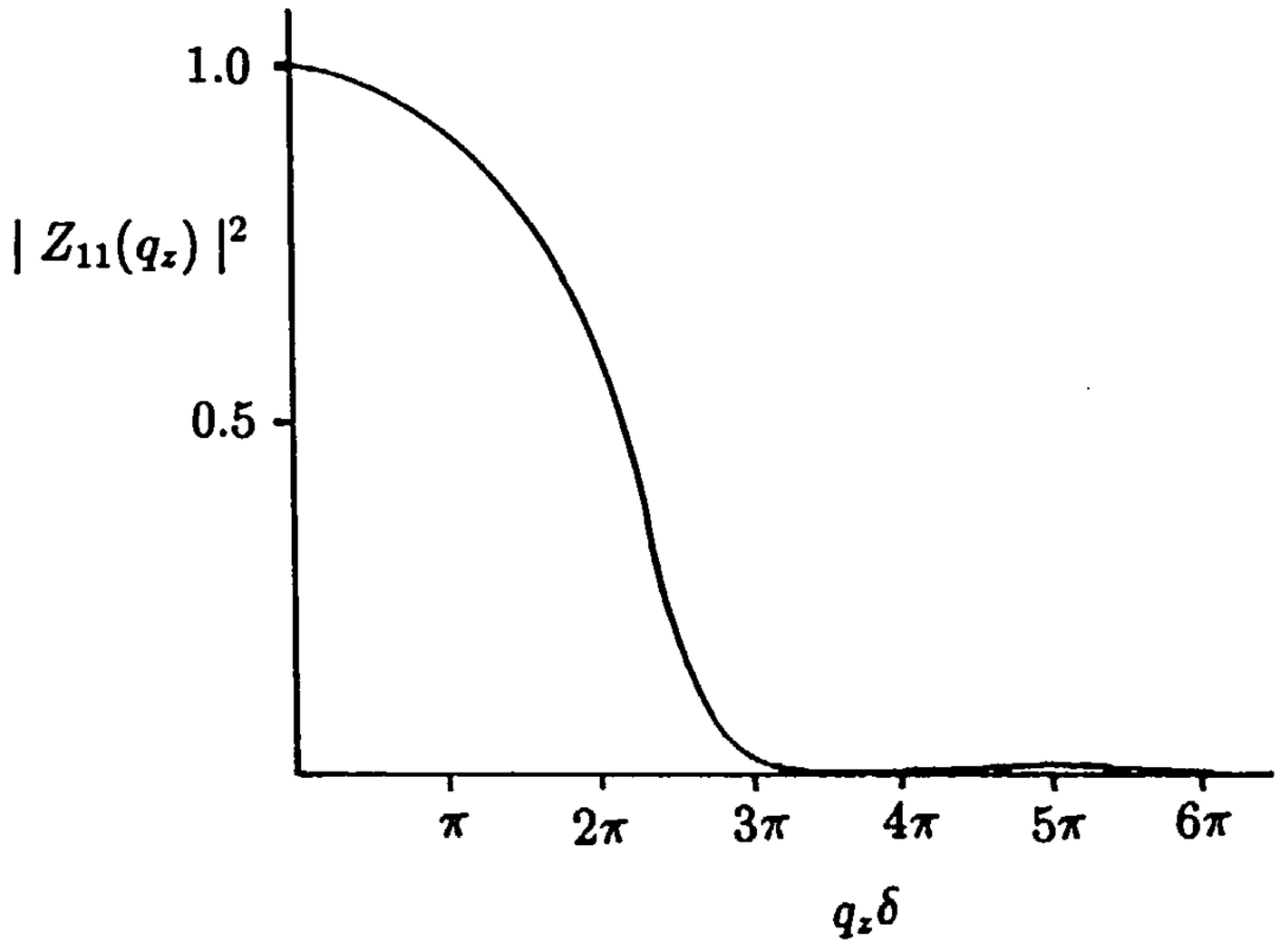
As discussed in II all the dependence of S_g upon the channel width enters via the matrix

element $Z_{11}(q_z)$ defined in (2.32). The ground envelope function of an ISW of width δ is assumed for $\phi_1(z)$ and in Figure 2.2 the corresponding plot of $\mid Z_{11}(q_z) \mid^2$ is illustrated. When $q_z\delta$ is about 3π an effective cut-off is evident. This condition ($q_z\delta < 3\pi$) is the origin of the enhancement of S_g in quasi-2D suggested by Zavaritsky (1984). For large δ the 2D vectors in the Kronecker delta of the transition rate for phonon absorption (2.31) become 3D. Thus (3D) momentum is conserved in the 3D limit and correspondingly there is just one $\mathbf{Q}$ linking the states $\mathbf{K}$ and $\mathbf{K}'$. For small δ the energy conservation condition expressed by the Dirac delta function still holds but $\mid Z_{11}(q_z) \mid^2$ replaces the conservation condition, expressed by the Kronecker delta (which is now 2D), upon the k_z component of the wavevector because k_z is no longer a good quantum number. Hence, in a narrow channel a larger phonon population can couple to, and hence drag, the electrons as now a number of $\mathbf{Q}$ have the required $\mathbf{q}$ component to link $\mathbf{k}'$ and $\mathbf{k}$ and still satisfy energy conservation. A larger S_g results as shown in Figure 2.2. In the figure an extrapolation to very large δ might be expected to yield the 3D limiting value of S_g. However, the quantum limit no longer applies here and contributions from higher subbands should be accounted for.

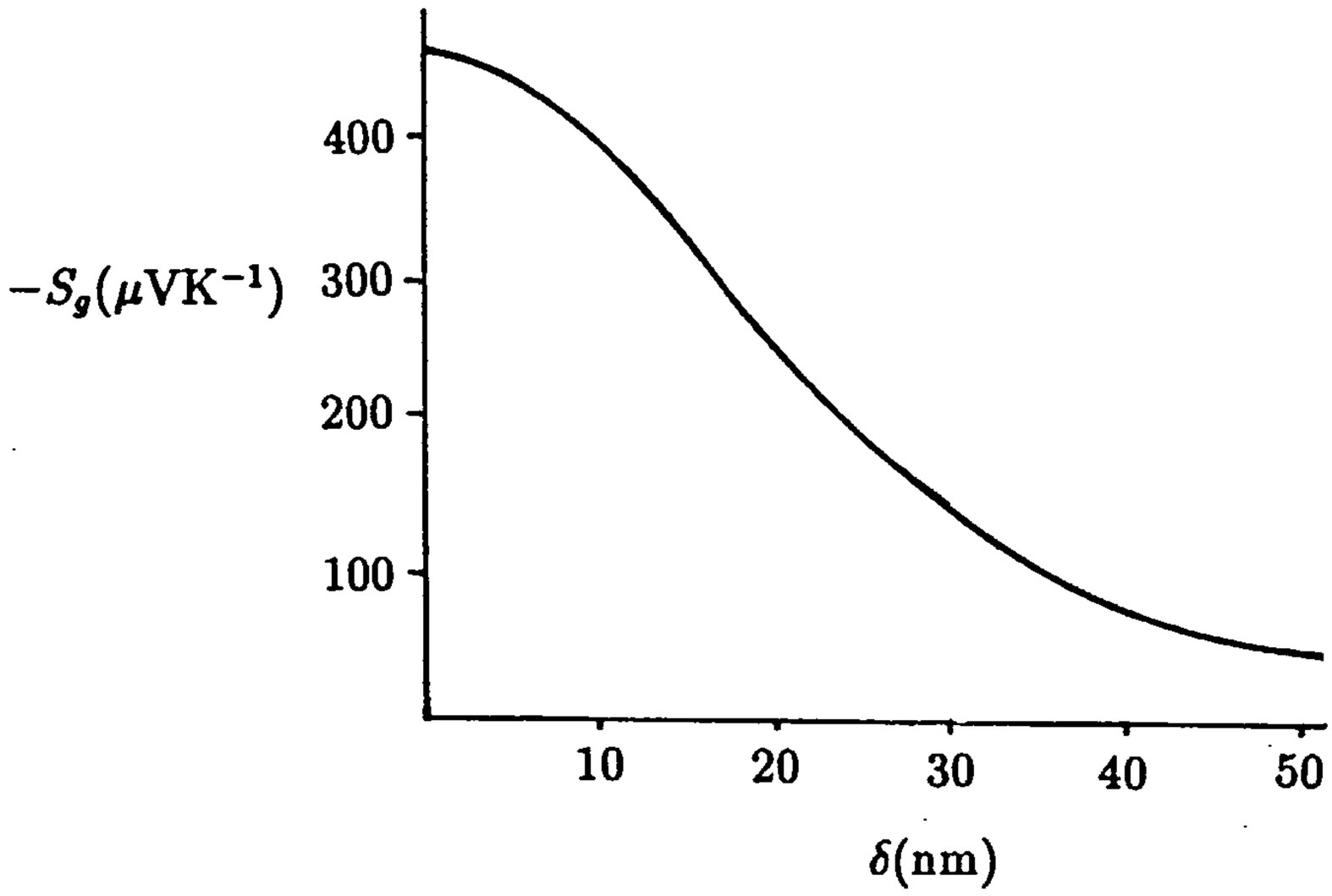
The qualitative agreement found between the calculated results of II for $S_g(T)$ and the experimental data is good in both GaAs/GaAlAs heterojunctions and Si MOSFETs. The accuracy of the position of the peaks found in plots of $-S_g/T^3$ against T (as in Figure 2.3) is particularly interesting. This type of plot is a simple test of the metallic formula (2.15) in which the peak would be attributed solely to a maximum in τ_{pe}^{-1} at some T. It is shown in II, however, that it is the coincidence of such a maximum with that of phonon distribution factors which is responsible. This is discussed in Chapter 5, where a simplified formula is derived by assuming the electron scattering to be elastic, because this formula is much more readily interpreted. The parameter values used in the calculations are given in Table 2.1, including the constant value 2nm taken for δ in the Si MOSFET. This rough estimate was used in II to calculate $S_g(T)$ for a range of n.

The linear dependence of S_g upon the constant L predicted in (2.71) has been tested

Figure 2.2: The effect of channel width on the thermopower in the ISW model.

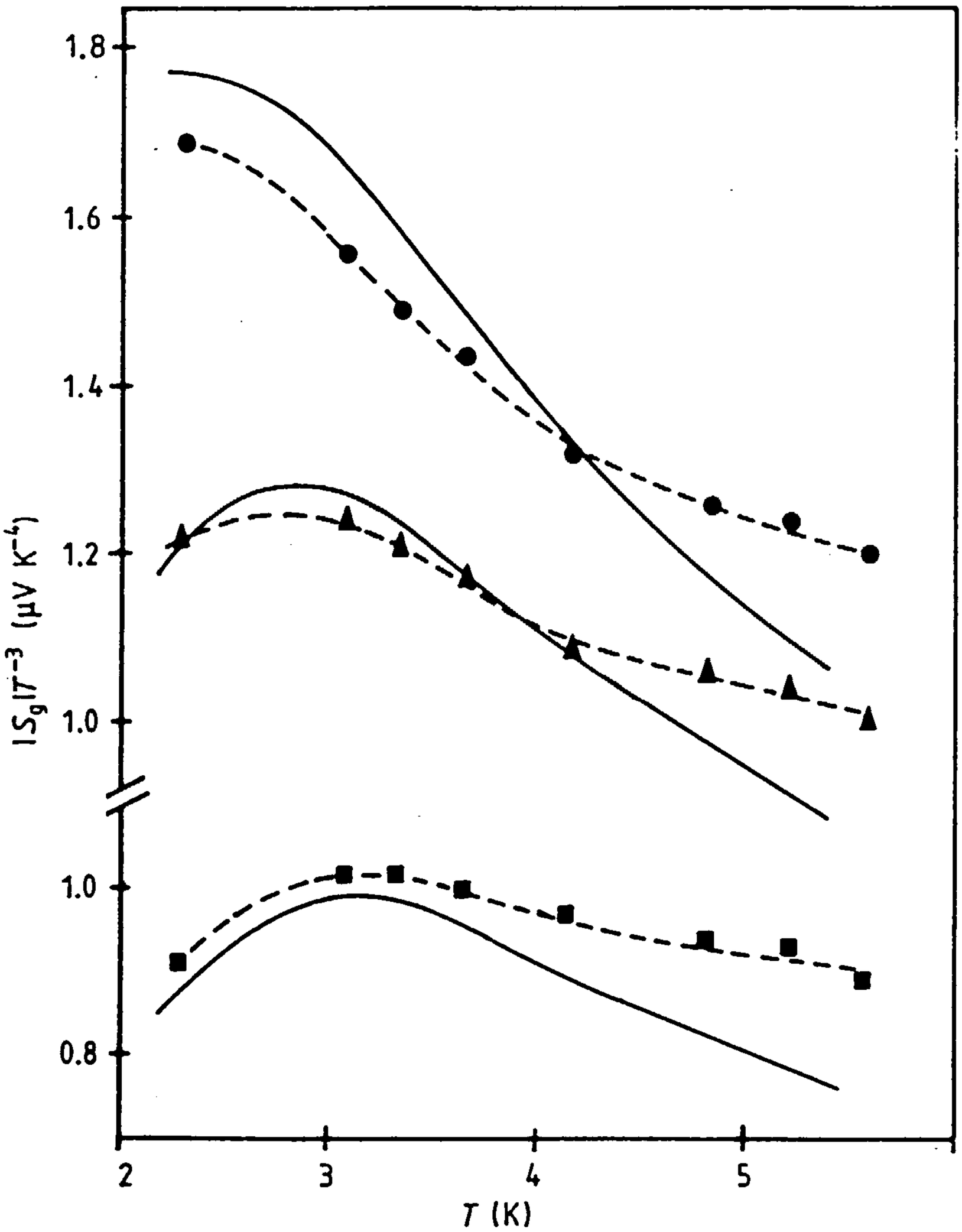


(a)Plot of $|Z_{11}(q_z)|^2$ against $q_z\delta$. The effective cut-off at $q_z\delta = 3\pi$ restricts phonon drag to phonons with $q_z < 3\pi/\delta$. Hence in narrow channels a larger phonon population can drag electrons.



(b)Effect on the thermopower. The q_z cut-off at $3\pi/delta$ is reflected in the variation of S_g with δ. The 3D limit is complicated by contributions from higher subbands.

Figure 2.3: Peaks in plots of $-S_g/T^3$ for a Si MOSFET.



Plots of $-S_g/T^3$ against T for various electron densities. Full curves: theory with deformation potentials scaled by 0.16. Symbols: experimental points of Gallagher et al (1987). Electron densities in units of $10^{15}\mathrm{m}^{-2}$ are 9.8(■), 7.8(▲) and 6.1(●).

Symbol	Units	Value	
		GaAs	Si
m	m_e	0.067 (0.07$_L$)	0.2 (0.19)
m_z	m_e	-	(0.916)
E_1	eV	8.0 (-9.3$_L$,16.0)	-
Ξ_u	eV	-	8.0 (9.0)
Ξ_d	eV	-	1.6 (-6.0)
h_{14}	$10^{-2}Vm^{-1}$	1.2	-
v_L	10^3ms^{-1}	5.1 (5.14$_L$)	8.5 (8.831)
v_T	10^3ms^{-1}	- (3.04$_L$)	5.0 (5.281)
ρ	10^3Kgm^{-3}	5.3	2.39
δ	nm	10.0 (eg. 30.0)	2.0 (eg. 7.0)
L	mm	0.3 (eg. 0.1)	0.5 (0.6)

The values given are those used in II. Changes made are in brackets with an "L" subscript for the values used by Lyo (1988), where different. For discussion see the text.

Table 2.1: Parameter values used in the calculations.

in an elegant experiment by Fletcher et al (1988a). The thermopower of a GaAs/GaAlAs heterojunction specimen measuring $12\times6\times0.42\text{mm}^3$ was measured over 1-6K. The boundary scattering limited value of L in such a specimen with one dimension (L_z) particularly small must be principally determined by the size of L_z. Since the scattering of phonons from all specimen faces is nominally diffuse, simple geometric averaging demands $L \geq L_z$. Polishing one large specimen face so that phonons are reflected, is thence equivalent to doubling L_z and, approximately, L. When this was performed the measured S was found to double but the resistivity change remained "insignificant" over the range investigated. This is to be expected for weak electron-phonon coupling in which case S_d should also be unchanged. The observation that S doubled therefore suggests that S_g is indeed domi-

nant and provides experimental support for the proportionality $S_g \propto L$. The almost exact doubling reported is surprising though, because the L measured from the $\kappa(T)$ data reveals a value of about $L_z/2$ for the polished specimen. The presence of other scattering mechanisms is thereby suggested and is supported by the reported "strong departures" of $\kappa(T)$ from the T^3dependence expected from boundary scattering alone.

Whilst the qualitative behaviour of S_g with L, T and n, and that of $-S_g/T^3$ in particular, agree well with the experimental data, the magnitudes calculated in II are much larger than found. In Figure 2.3, for example, the results (for Si) have been scaled by 1/40. (The necessary scaling in the GaAs case is about 1/3.) The author repeated these calculations using the same data (Table 2.1) by the method of direct evaluation (ie without the introduction of broadening) and obtained similar results but with corresponding scalings of 1/16 for Si and 1/1.5 for GaAs. These new results were later confirmed, to within an error of less than 5% over the entire parameter range, when the approach in II was adopted (see Smith and Butcher 1989a). It must be concluded therefore, that the numerical integration in II lacked accuracy in the final evaluation.

2.7 Status and likely improvements.

Although the final results of the last section show a dramatic improvement over those in II, the outstanding large overestimate of $-S_g$ by a factor of up to 16 (in Si) represents a gap in understanding which warrants further investigation. The good qualitative features of the theory in comparison with both the simple models and the experimental data, however, suggest that this error factor and its difference in size between Si and GaAs may be understood within the existing framework. An obvious point to consider is the effect upon S_g of a more realistic treatment of the electron confinement, ie $\phi_1(z)$. The review by Ando et al (1982), for example, gives typical values for the channel width in a Si MOSFET as in the range 2-20nm. The results of Figure 2.3 show the corresponding variation in S_g to be around 50% although this is only in the crude ISW approximation. Thus the effect of a closer treatment of electron confinement upon the thermopower will be investigated.

Gallagher et al (1987) suggest that screening may explain the overestimate as it is claimed that a reduction by a factor of about 1/20 may result in Silicon which is of the necessary size. To discuss screening however, an estimate of the importance of higher subbands is necessary as will be shown in Chapter 4.

In the next chapter it is considered how the quasi-2D electron confinement arises in GaAs/GaAlAs heterojunctions and Si MOSFETs in order that the effect of a better treatment of $\phi_1(z)$ on the 'drag thermopower can be assessed. The effect of screening is discussed in Chapter 4 and the validity and effect of relaxing the approximations listed in the previous section are examined in Chapter 5.

Chapter 3

Quasi-two dimensionality in Si MOSFETs and GaAs/GaAlAs heterojunctions.

3.1 Introduction to the chapter.

The object here is to look in more detail at the practical realization of the quasi-2D confinement in a Si MOSFET and GaAs/GaAlAs heterojunction in order that this be treated more faithfully in calculations of S_g. Quasi-2D behaviour is shown to arise in the MOSFET from the potential well created at the Si/SiO_2 interface on application of a gate voltage and at a heterojunction from the well arising from the difference in band gap. Calculations based on a many-body variational principle show that a model variational envelope function $\phi(z)$ can be obtained from an effective 1D variational condition in such systems. The Fang and Howard, and Ando variational envelopes are then determined when model potentials are introduced. The effect of $\phi(z)$ on S_g is examined and compared with the infinite square well (ISW) model used in II.

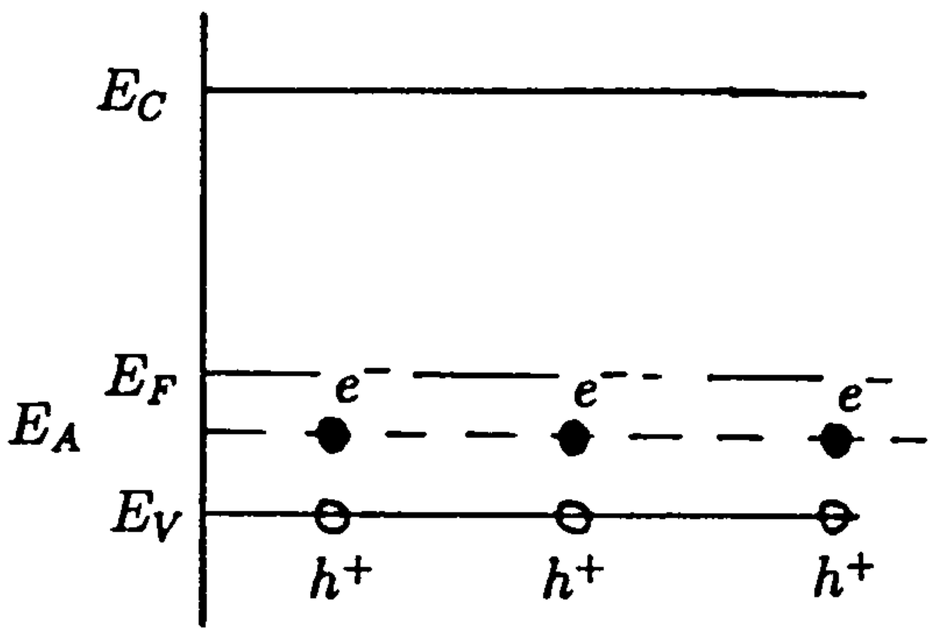
3.2 Occurrence of the confining potential.

For the present purposes the n-channel Si MOSFET may be reasonably considered as a capacitor. The "gate" forms one plate and is a metallic layer deposited on an oxidized Si substrate (usually the high mobility (100) surface). The substrate forms the second plate and is doped p-type for an n-channel device (p-channel follows analogously). The oxide (SiO_2) between the plates remains insulating up to high fields and enables a range of voltages V_G to be applied to the gate before breakdown occurs (for a review of the MOSFET see Ando et al 1982). Consider now the effect on the band profile of the substrate as V_G increases, as illustrated in Figure 3.1. At low temperatures with $V_G = 0$ the acceptor states (with doping level N_A, say) are occupied by electrons (e^-). Some holes (h^+) lie below the valence band edge, but the conduction band is empty. As V_G is increased from zero, some of the holes are repelled from near the Si/SiO_2 interface ($z = 0$) creating a "depletion layer" of thickness z_d assumed to be depleted of all holes. Electron energies would be lower in this region as V_G attracts electrons to the interface and, as in Figure 3.1b, the bands bend down. Electrons far from the interface (large z) see a smaller potential so the bending here is less. If V_G is increased so that E_C drops below E_F near $z = 0$, electrons can leave acceptor sites and reach the conduction band. The vacated acceptor sites are filled by the creation of further holes which are repelled by V_G, leaving the situation in Figure 3.1c. A layer of electrons is formed, parallel to $z = 0$, in which electron motion is restricted in the z direction but otherwise remains free. This quasi-2D system in which the carrier sign is inverted is referred to as the "inversion layer" and can be considered to arise simply from the creation of the self-consistent confining potential $V(z)$ shown in Figure 3.2a.

A similar potential well is created at the heterojunction of oppositely doped GaAs and GaAlAs. If the GaAs is doped with acceptors (at level N_A) and the GaAlAs with donors (at level N_D) the band profile is similar to that of a p-n junction, as shown in Figure 3.3. For the heterojunction the two halves have different band gaps E_g (1.52eV at 4K in GaAs and about 2eV in GaAlAs) which gives rise to a discontinuity ΔE_g shared unequally by

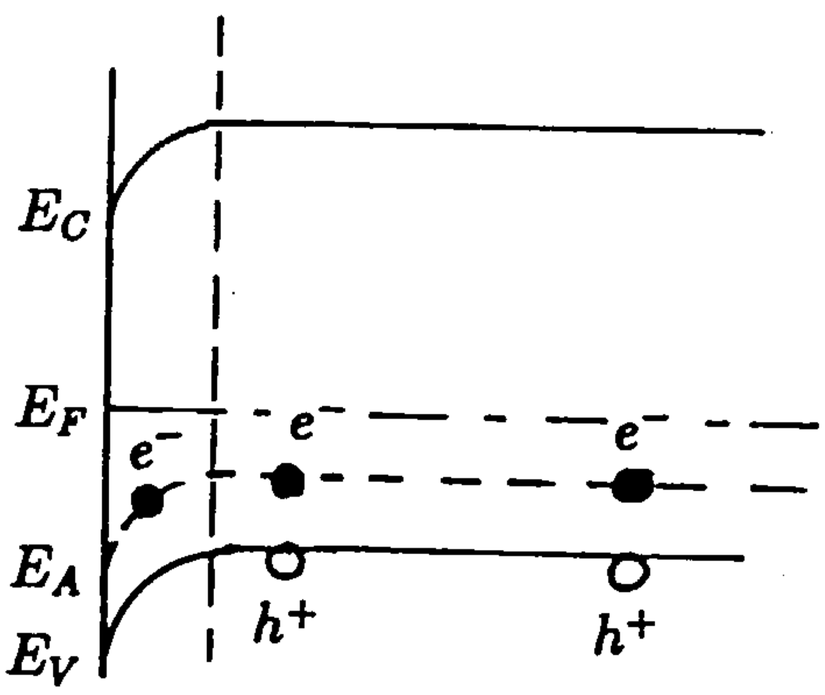
Figure 3.1: Formation of the inversion layer at low T.

(a) $V_G = 0$



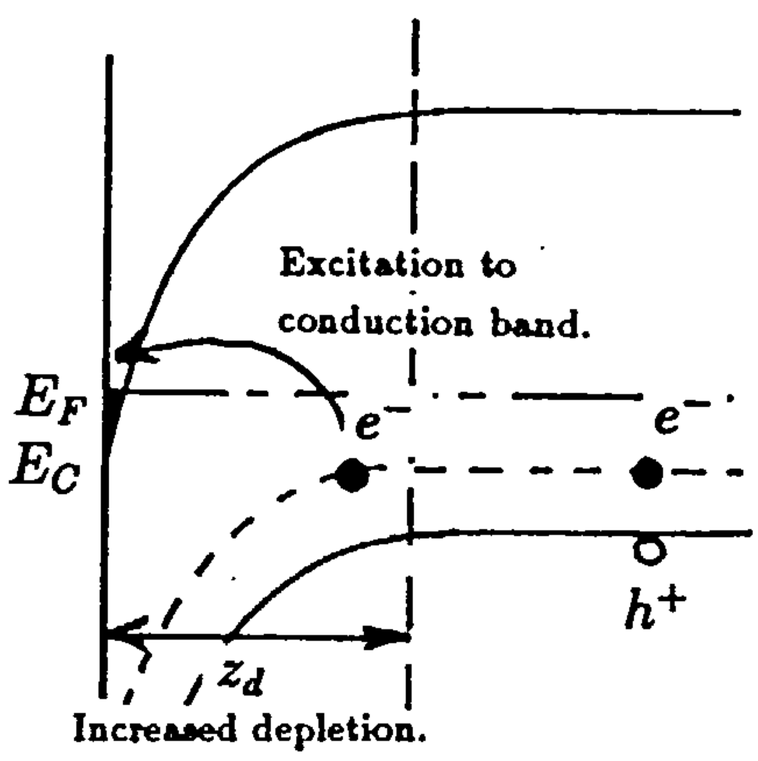
With $V_g = 0$ acceptor sites are occupied and conduction is via holes in the valence band.

(b) Depletion.



As V_G increases some holes are repelled creating a depletion layer. Band bending begins but conduction is still via holes.

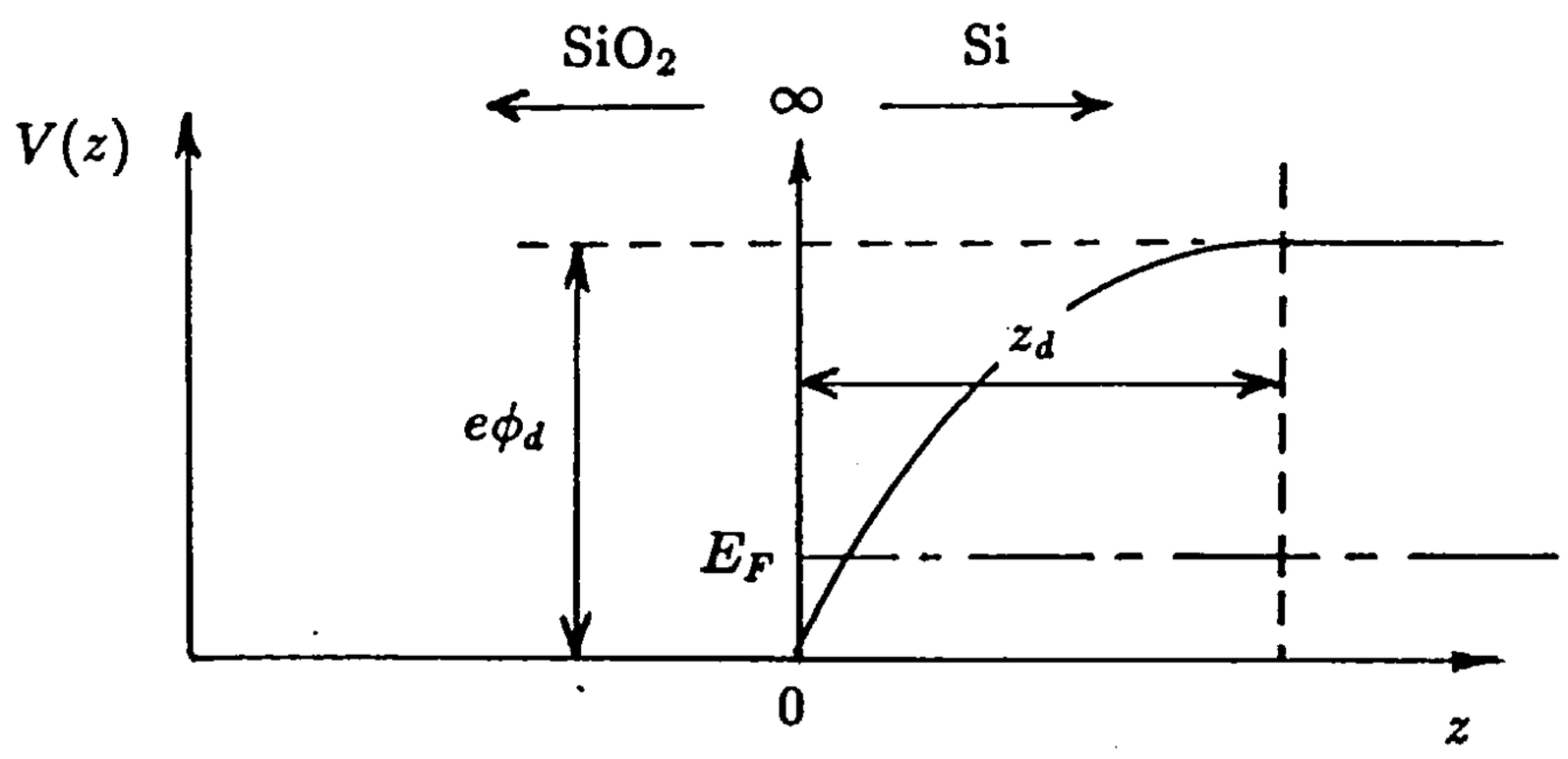
(c) Depletion and Inversion



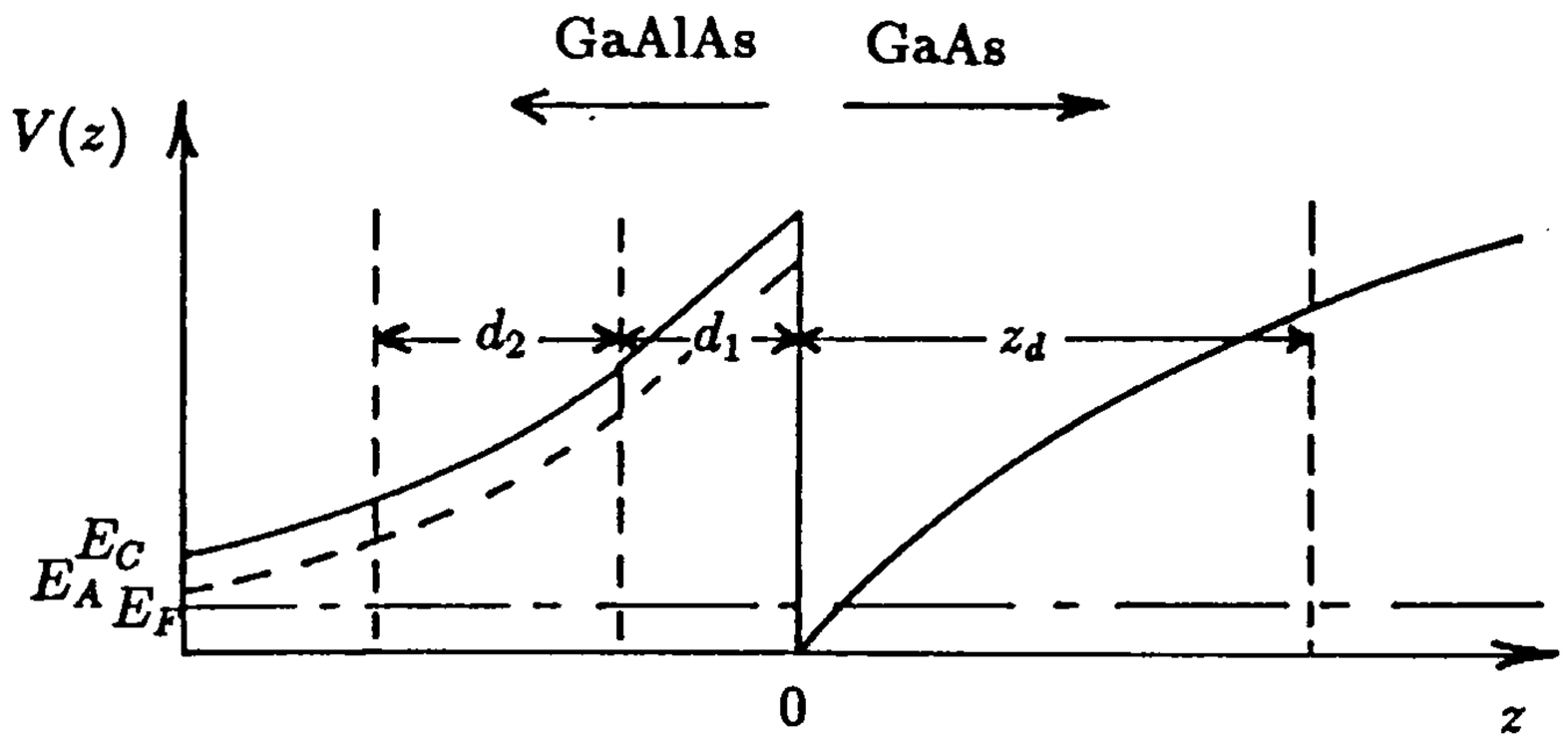
Eventually E_C drops below E_F, near $z = 0$, and electrons able to conduct appear in the conduction band. The depletion layer widens.

NB. The energy differences $E_A - E_V$ and $E_F - E_V$ have been exaggerated for clarity.

Figure 3.2: The confining potential.

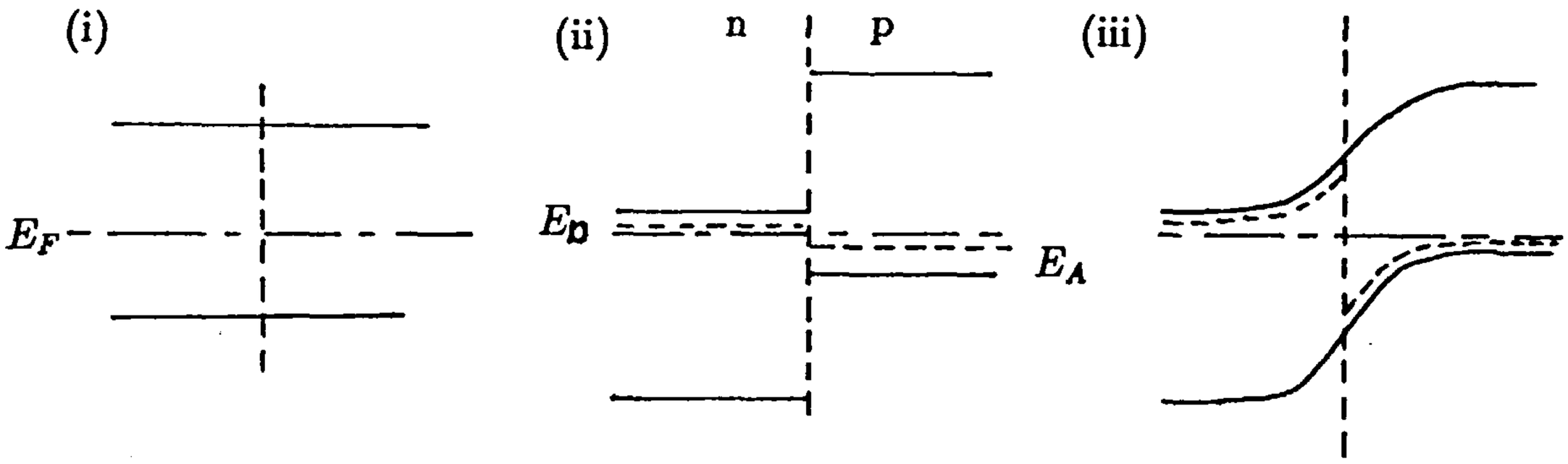


(a)Si MOSFET A potential well roughly triangular in shape is formed below E_F. For a perfect insulating oxide $V(z < 0)$ is infinite.



(b)GaAs/GaAlAs heterojunction. The conduction band edge discontinuity gives rise to a potential well without any fields applied. Since the barrier height is finite some penetration into the spacer layer (d_1) is possible.

Figure 3.3: Conceptual formation of the confining potential in a GaAs/GaAlAs heterojunction.

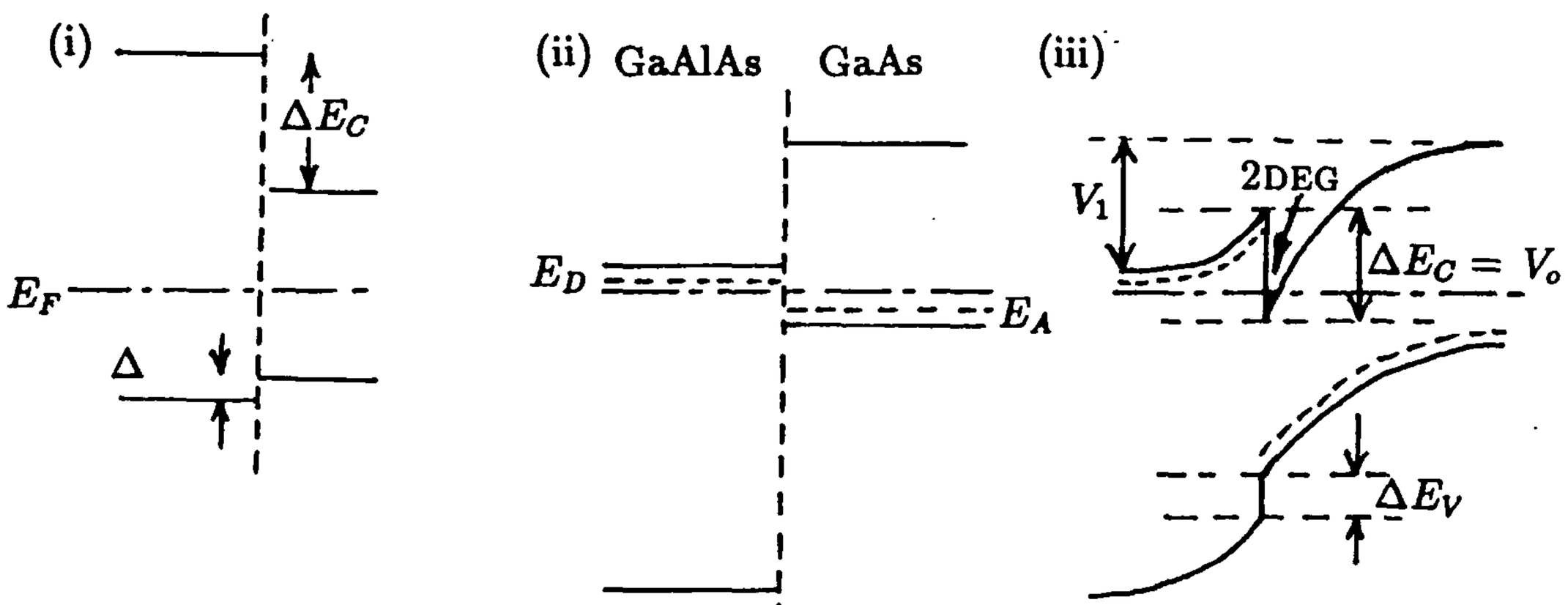


(a)Band Profile in a p-n junction.

(i) Before doping.

(ii)After doping, by matching E_F.

(iii)The observed profile.



(b)Band profile in the heterojunction.

(i) Before doping (including discontinuities).

(ii)After doping, by matching E_F.

(iii)The observed profile (including discontinuities).

the conduction band edge (ΔE_C) and valence band edge (ΔE_V). A quasi-2DEG forms when, as in Figure 3.2b, electrons from ionized donors (in region $z < 0$) see a nearby potential well to fall into.

It is not necessary to go into much further detail here as there are many reviews (see, for example, Sze 1981, Schulz 1986). A few points are worth noting, however. The main advantages of the MOSFET in LDS experiments are a consequence of the oxide properties of stability and low surface state density, and the breakdown field of $10^9 \mathrm{Vm}^{-1}$ allows a variation of the electron surface density over the range $10^{15} - 10^{17} \mathrm{m}^{-2}$. The lower limit is determined by the influence of disorder. The advantages of the heterojunction on the other hand lie in the interface quality and the possibility of modulation doping which allow much higher mobilities to be achieved at liquid helium temperatures. The MOSFET mobility is limited by surface roughness and ionized impurities caught in the oxide (Stern 1980). In Figure 3.3b the donors appear adjacent to the quasi-2DEG but they may be separated from the electrons by providing a nominally undoped spacer of width d_1, as in Figure 3.2b. The electron scattering effect of the ions is then reduced. The spacer should not be made too wide or few electrons will reach the potential well. In the calculations which follow it is assumed that the region of donors of width d_2 is completely ionized. Illumination can be used to increase n by exciting electrons out of deep traps (see for example, Fletcher et al 1988a). The GaAs is usually nominally undoped but unintentional acceptor doping may be present. The very large potential barrier at $z = 0$ in the MOSFET prevents the quasi-2DEG from penetrating the oxide ($z < 0$) but as shown in Figure 3.2b the corresponding barrier in the heterojunction is much smaller and so some penetration occurs. This is why the spacer layer has a profound effect in reducing the ionized impurity scattering and increasing the mobility.

3.3 The variational approach to an envelope function for the ground subband.

Within the independent particle effective-mass picture the desired feature of the envelope function $\phi(z)$, is that it should represent the electron confinement in a way which faithfully reflects the influence of the specimen properties (eg. effective mass, doping levels and barrier heights and widths). Such a picture is already rather approximate and idealized so that, taken together with the uncertainties in the physical properties, great sophistication may not be justified. Ideally, $\phi(z)$ should be expressible in terms of a minimum number of parameters which reflect, and are determined by, the properties of the specimen, in an analytical form convenient for transport calculations. A variational approach in which ϕ is written in terms of variational parameters b, say, $(\phi_b(z))$ is therefore a sensible choice.

For interface problems, the translational invariance perpendicular to the interface ($z = 0$) demands that the single particle Hamiltonian ($\hat{H}_s$) be only a function of z. When $\phi_b(z)$, say, is taken as the ground state envelope function in (1.10), and $\hat{H}_s$ does not depend upon b, the variational condition reduces to:

$$\frac{d}{db} < \phi_b(z) \mid \hat{H}_s \mid \phi_b(z) >= 0 \tag{3.1}$$

which is the form of a variational condition for a purely 1D particle. (The contribution to the kinetic energy arising from xy motion has been dropped because it is independent of b). The total potential energy in the systems of interest arises from potential fields external to the electron gas $V_{EX}(z)$, from ionized impurities and potential barriers for example, and, in the Hartree approximation, from the total field of the electron gas itself $V_s(z)$. The latter depends upon b because it is determined by the electron distribution through Poisson's equation. Therefore the above condition (3.1) no longer holds because the Hamiltonian does depend upon b. However, a similar result does hold.

Consider the quasi-2DEG many-body Hamiltonian $\hat{H}_M$ in the form:

$$\hat{H}_M = \hat{T}_M + \hat{V}_{MS} + \hat{V}_{M,EX} \tag{3.2}$$

where the kinetic energy contribution is:

$$\hat{T}_M = \sum_i \frac{-\hbar^2 \nabla_i^2}{2m}, \tag{3.3}$$

and

$$\hat{V}_{MS} = \frac{1}{2} \sum_{i,j \neq i} \frac{e^2}{4\pi\epsilon \mid \mathbf{R}_i - \mathbf{R}_j \mid} \tag{3.4}$$

replaces $V_s(z)$. The third term is the resultant potential from all sources other than the 2DEG itself. Construct a many-body variational ground state $\mid G_b >$ by taking an $N \times N$ Slater determinant of product type functions, made from the product of N single electron quasi-2D state functions in the ground subband of the form (1.10) using the variational envelope $\phi_b(z)$, say. The variational theorem can then be applied to the many-body ground-state energy $E_M(b)$, ie:

$$\frac{d}{db} E_M(b) = 0 \tag{3.5}$$

where:

$$E_M(b) = < G_b \mid \hat{H}_M \mid G_b > = < \hat{T}_M > + < \hat{V}_{\mu,s} > + < \hat{V}_{M,EX} >, \tag{3.6}$$

because $\hat{H}_M$ does not depend on b. These two equations could be used, in principle, to determine b but it is more convenient to work in second quantized form. Thus:

$$\hat{T}_M = \sum_{s,t} < \psi_s \mid \hat{T} \mid \psi_t > \alpha_s^\dagger \alpha_t \tag{3.7}$$

may be written, where $\hat{T}$ is the single particle kinetic energy operator and s and t label individual particle states. The result is then:

$$< \hat{T}_M > = \sum_i < \psi_i \mid \hat{T} \mid \psi_i >, \tag{3.8}$$

where the sum is over occupied states. Similarly:

$$\hat{V}_{M,S} = \sum_{s,t,u,v} < \psi_s(1)\psi_t(2) \mid \frac{e^2}{4\pi\epsilon \mid \mathbf{R_1} - \mathbf{R_2} \mid} \mid \psi_v(1)\psi_u(2) > \alpha_s^\dagger \alpha_t^\dagger \alpha_u \alpha_v \tag{3.9}$$

and

$$\hat{V}_{M,EX} = \sum_{s,t} < \psi_s \mid \hat{V}_{EX} \mid \psi_t > \alpha_s^\dagger \alpha_t, \tag{3.10}$$

so that, finally, an equation analogous to (3.1) is obtained (Fang and Howard 1966):

$$\frac{d}{db}(E_b/N) = \frac{d}{db}\left\{< T_z >_z + \frac{1}{2} < V_s(z) >_z + < V_{EX}(z) >_z \right\} = 0, \qquad (3.11)$$

Exchange has been neglected, $< \hat{O} >_z$ is the matrix element $< \phi_b(z) \mid \hat{O} \mid \phi_b(z) >$ of the one body operator $\hat{O}$ and:

$$T_z = \frac{-\hbar^2}{2m_z}\frac{d^2}{dz^2}, \qquad (3.12)$$

with m_z the effective mass for motion in the z direction (assumed constant). The form of (3.11) is very similar to the 1D variational expression but here the single particle Hamiltonian depends upon b and the presence of the 1/2 indicates that the quantity E_b/N is the 2DEG energy *per electron.*

The Hartree approximation becomes valid at electron surface densities sufficiently high that the kinetic energy dominates over the interaction between electrons (see, for example, Ando et al 1982). Generally, though, a contribution to (3.11) should arise to account for the effect of the Exclusion Principle (exchange) which reduces the net electron density around a given electron and thereby, also, the net negative charge seen by another electron. A correlation correction to the simple Coulomb energy in the Hartree approach is necessary because the electron motion is not really uncorrelated. The introduction of a single exchange/correlation potential $V_{xc}(z)$ seeks to account for these corrections. This potential is derived through the Density Functional Formalism of Hohenberg and Kohn (1964) and Kohn and Sham (1965) who have shown that the mininum value of the many-body ground state energy (in the general problem) can be obtained from an equation analogous to (3.11) when the "Schrodinger Equation" is augmented by an additional term $V_{xc}(z)$. The formalism does not guarantee that the ε_i and ϕ_i which result from this Hohenberg-Kohn-Sham equation are the one-body energies and envelope functions, although they are normally regarded as such (see, for example, Ando 1976 and 1982a, Das Sarma and Vinter 1982, Stern and Das Sarma 1984). What is guaranteed is that they yield the best value of the many-body ground state electron density which can be obtained from the variational form chosen.

An important feature of the formalism is that it provides a formula for V_{xc} and an approximation scheme to evaluate it. Thus, $V_{xc}(z)$ is given as an unknown functional of the ground state electron density and is approximated by the exchange/correlation part of the chemical potential of a uniform electron gas having the same value as the local electron density (the Local Density Approximation or LDA). There are different ways in which this potential may be parameterized (see the references previously cited) but its exact value is not normally very important (Ando et al 1982). Stern and Das Sarma (1984) make the point that whilst the formalism (in the LDA) has had great successes in comparing calculations of electronic structure in a wide variety of systems such as bulk solids, surfaces and molecules, the "condition for its validity is seldom met in the physical systems of interest". The LDA requires the electron density to vary over distances large compared to the local Fermi wavelength. This condition is often violated in practice but nevertheless results like those of Stern and Das Sarma compare well with the experimental data.

For present purposes, equation (3.11) may be regarded as an effective 1D single particle variational condition determining the ground subband envelope function in a given parameterization. This equation is used to obtain the Fang and Howard envelope for the MOSFET and, generalizing to a two parameter energy minimization, the Ando envelope for the heterojunction (Ando 1982b). The approximations which make these envelopes convenient, however, may have more of an effect upon the final form of $\phi(z)$ than the most accurate representation of the full potential $V(z)$. Therefore, comparisons with full self-consistent calculations including all the necessary contributions would be of interest in estimating the size of likely errors incurred in adopting the envelopes used here. For the present however, $V_{xc}(z)$ will be neglected because the effect of its inclusion is an unnecessary complication here compared to the increased accuracy obtained (see Ando 1982a,b, Stern and Das Sarma 1984).

3.4 The Fang and Howard envelope for MOSFETs.

The simplest approximate variational envelope function, and the most widely used in Si MOSFETs (see, for example, Ando et al 1982 and many of the references therein), is that proposed by Fang and Howard (1966) where:

$$\phi_b(z) = \begin{cases} (b^3/2)^{1/2} z e^{(-bz/2)} & z \geq 0 \\ 0 & z < 0. \end{cases} \tag{3.13}$$

This form is naturally very convenient for calculations and involves just the one parameter b which can be readily determined. Moreover, the position expectation value is $3/b$ for this model envelope and $\delta/2$ in the ISW model. Therefore, an equivalent value of δ can be estimated from:

$$\delta \approx 6/b, \tag{3.14}$$

and hence the value of δ taken, for example, in II can be assessed. The derivation of b is worth considering briefly here to appreciate the features determining $\phi(z)$.

In addition to the kinetic energy and $V_s(z)$ terms, two contributions to the external potential in (3.10) are accounted for, due to the ionized charges of the depletion layer $V_{dep}(z)$ and the image charge $V_I(z)$ arising from the permittivity difference between the semiconductor (ε) and its oxide (ε_{ox}). Trivially,

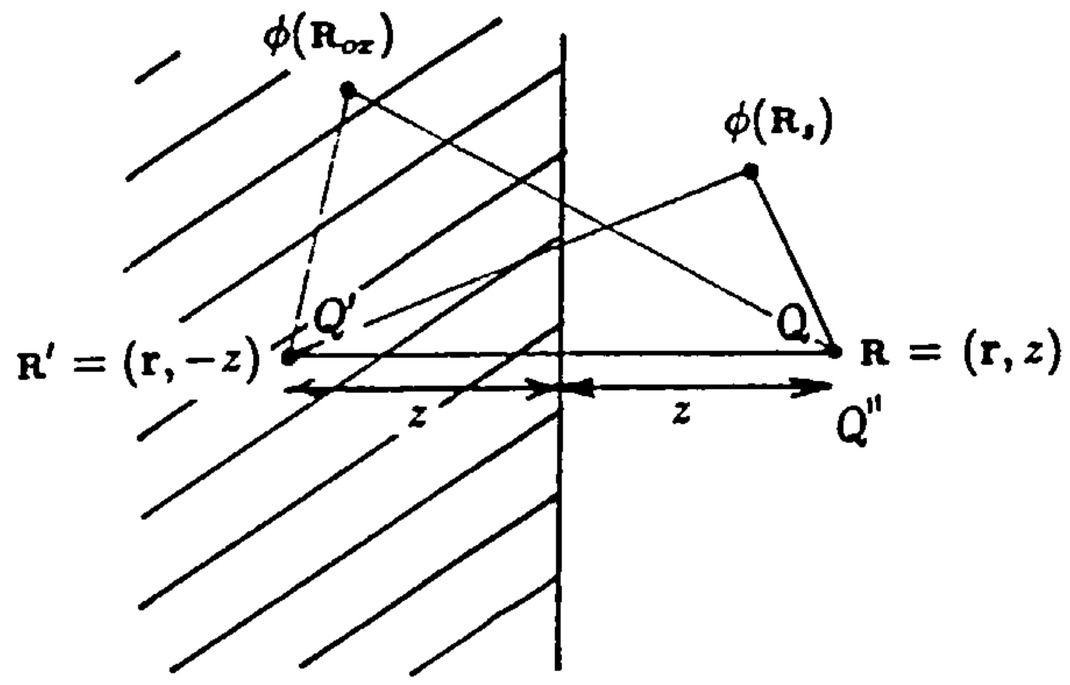
$$< \hat{T} >_z = \frac{\hbar^2 b^2}{8 m_z}. \tag{3.15}$$

The energy $V_s(z)$ is determined by integrating the Poisson equation using $-ne \mid \phi_b(z) \mid^2$ for the charge density. Similarly, $V_{s,dep}(z)$ is obtained using the constant charge density $-N_A e$ over the depletion region of width z_d, shown in Figure 3.2a. The boundary conditions imposed are: $dV_s(\infty)/dz = 0$ and $V_s(0) = n_0 e^2 < z > /\varepsilon$ (which follows trivially) for $V_s(z)$ and $V_{dep}(0) = 0$, $V_{dep}(z_d) = e\phi_d$ and $dV_{dep}(z_d)/dz = 0$, for the depletion energy. The results are then:

$$V_s(z) = \frac{ne^2}{\varepsilon} \left[z + \int_0^z (y - z) \mid \phi_b(y) \mid^2 dy \right], \tag{3.16}$$

$$< V_s(z) > = \frac{33}{16} \frac{ne^2}{\varepsilon b}, \tag{3.17}$$

Figure 3.4: The image system for the image potential at a dielectric boundary.



The potential seen in the silicon $\phi(\mathbf{R}_s)$ is due to Q at $\mathbf{R}$ and its image charge Q' at $\mathbf{R}'$, but that in the oxide $\phi(\mathbf{R}_{ox})$ is due only to Q'' at $\mathbf{R}$. Q' and Q'' are determined by imposing continuity conditions along the boundary.

and:

$$V_{s,d}(z) = \frac{N_A z_d e^2 z}{\varepsilon}\left(1 - \frac{z}{2z_d}\right), \tag{3.18}$$

$$< V_{s,d}(z) >= \frac{3e^2 N_A z_d}{\varepsilon b}\left(1 - \frac{2}{bz_d}\right) \tag{3.19}$$

where:

$$z_d = (2\phi_d \varepsilon / N_A e)^{1/2}. \tag{3.20}$$

Comparing Figures 3.1c and 3.2a and, noting that $V_s(0)$ and $(E_F - E_V)$ are small fractions of the band gap E_g, the band bending $e\phi_d$ can be approximated by E_g, which is about 1.1eV in Si (Landolt-Bornstein 1982). The quantity $-N_A e z_d$ is the depletion layer areal charge density and as the band bending increases the width of the well of Figure 3.2a increases as the square root of its depth. In deriving the above expectation values it is assumed that $z_d >> 1/b$. Typical values are: z_d about $1\mu m$ and b^{-1} about 1nm.

The image potential arises from the dielectric boundary. The image system produced is illustrated in Figure 3.4 where it is supposed that the potential seen in the Silicon at $\mathbf{R}_s$,say , due to a charge Q at $\mathbf{R} = (\mathbf{r}, z)$, has a contribution from an image charge Q' at $\mathbf{R}' = (\mathbf{r}, -z)$. The potential seen in the oxide at $\mathbf{R}_{oz}$ is that from an effective charge Q'' at $(\mathbf{r}, z)$ since there is no charge in the oxide. The potentials and perpendicular components

of the displacement vector are then matched along $z = 0$ to determine Q' and Q''. The result for an electron at $(\mathbf{r}, z)$ is then:

$$V_I(z) = \frac{e^2}{8\pi\varepsilon z}.\frac{\varepsilon - \varepsilon_{ox}}{\varepsilon + \varepsilon_{ox}}. \tag{3.21}$$

and:

$$< V_I(z) >= \frac{e^2 b}{16\pi\varepsilon}.\frac{\varepsilon - \varepsilon_{ox}}{\varepsilon + \varepsilon_{ox}}. \tag{3.22}$$

where ε_{ox} is the permittivity of the oxide.

On substituting into (3.10) for all the expectation values the variational condition for b becomes:

$$\frac{\hbar^2 b}{4m_z} + \frac{e^2}{16\pi\varepsilon}.\frac{\varepsilon - \varepsilon_{ox}}{\varepsilon + \varepsilon_{ox}} - \frac{e^2}{32\varepsilon}\left(\frac{96e^2 N_A z_d + 33n_0}{b^2}\right) + \frac{12e^2 N_A}{\varepsilon}.\frac{1}{b^3} = 0 \tag{3.23}$$

which can be solved numerically without difficulty. Account for an exchange and correlation potential merely adds a further term here. In practice the above is simplified further by dropping the image term and quadratic depletion layer potential term (the so-called "triangular approximation"), so that the conveniently simple analytic result (Stern and Howard 1967),

$$b = \left[12e^2 m_z N^*/\varepsilon\hbar^2\right]^{1/3}, \tag{3.24}$$

can be obtained, where:

$$N^* = N_A z_d + (11/32)n \tag{3.25}$$

and may be considered as an effective areal density of negative charge.

In Table 3.1 values of $6/b$, taken as a measure of the channel width, obtained by using (3.24) and (3.23), in full, are compared with the results obtained by dropping the image and quadratic depletion layer terms separately for the MOSFET data of Gallagher et al (1987). The value 2nm taken in II is clearly a large underestimate, by up to a factor of 5. The Table also illustrates the variation of the channel width with n (gate voltage) which is not accounted for in the ISW where δ is treated as an independent variable. The quadratic depletion layer potential term has little effect and the triangular approximation is thereby justified. This is to be expected as with $\delta \ll z_d$ the electrons see only $V_{dep}(z \ll z_d)$, where

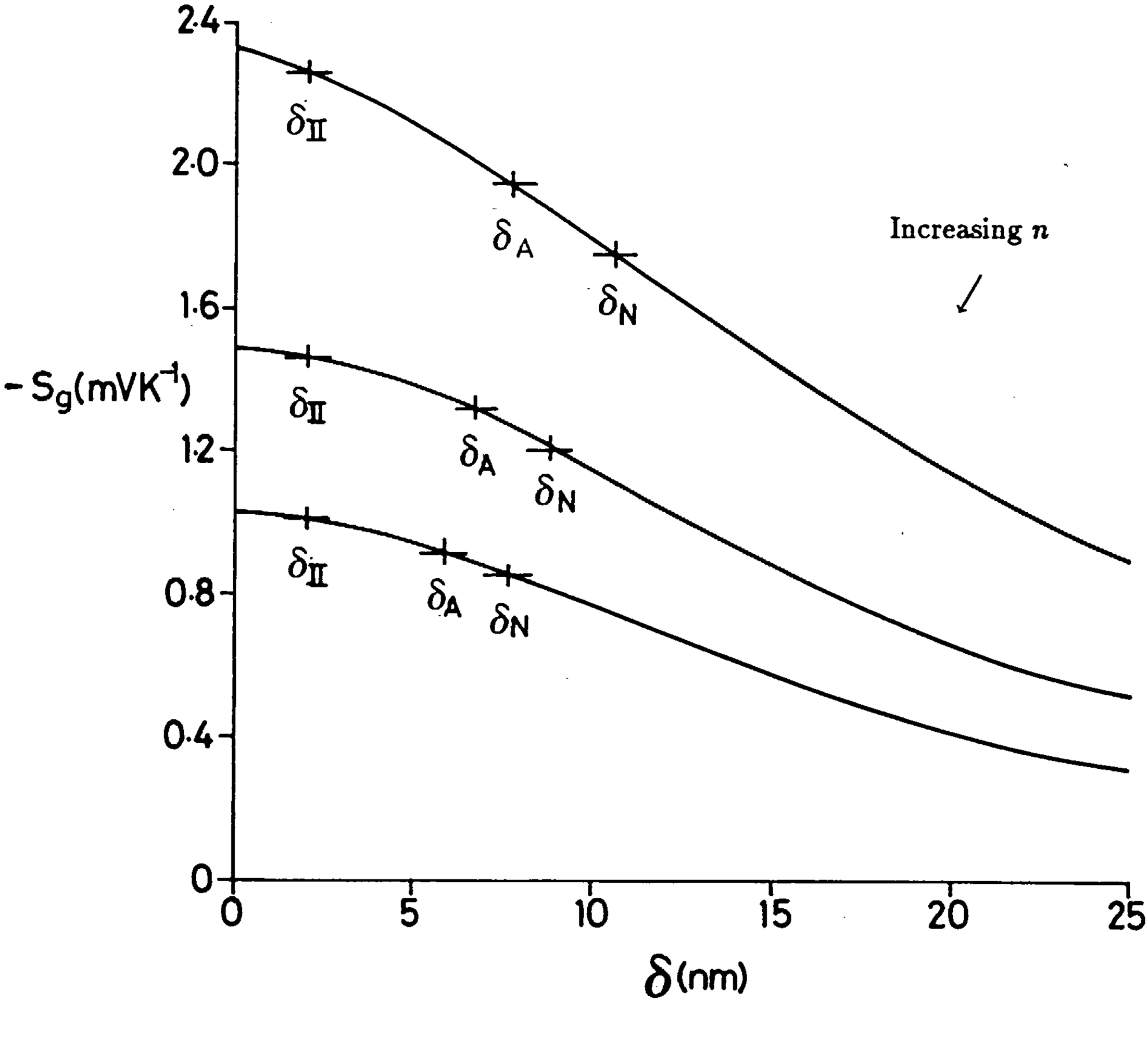
$n/10^{15}\mathrm{m}^{-2}$	Channel width ($\delta = 6/b$) /nm			
	δ_I	δ_Q	δ_A	δ_N
3.2	8.01 (24)	10.59 (0)	7.75 (27)	10.59
6.1	6.94 (22)	8.88 (0)	6.74 (24)	8.89
9.8	6.13 (20)	7.66 (0)	7.66 (22)	7.66

Different estimates of channel width δ (ie $6/b$) for a range of electron densities n for $N_A = 4 \times 10^{20}\mathrm{m}^{-3}$: δ_I is obtained by dropping the image term from (3.23) and δ_Q the quadratic depletion layer potential term, δ_A is the analytic result (3.24) obtained by dropping both and δ_N is without dropping either. The value in brackets is the percentage underestimate in comparison with δ_N.

Table 3.1: Example results for the MOSFET channel width.

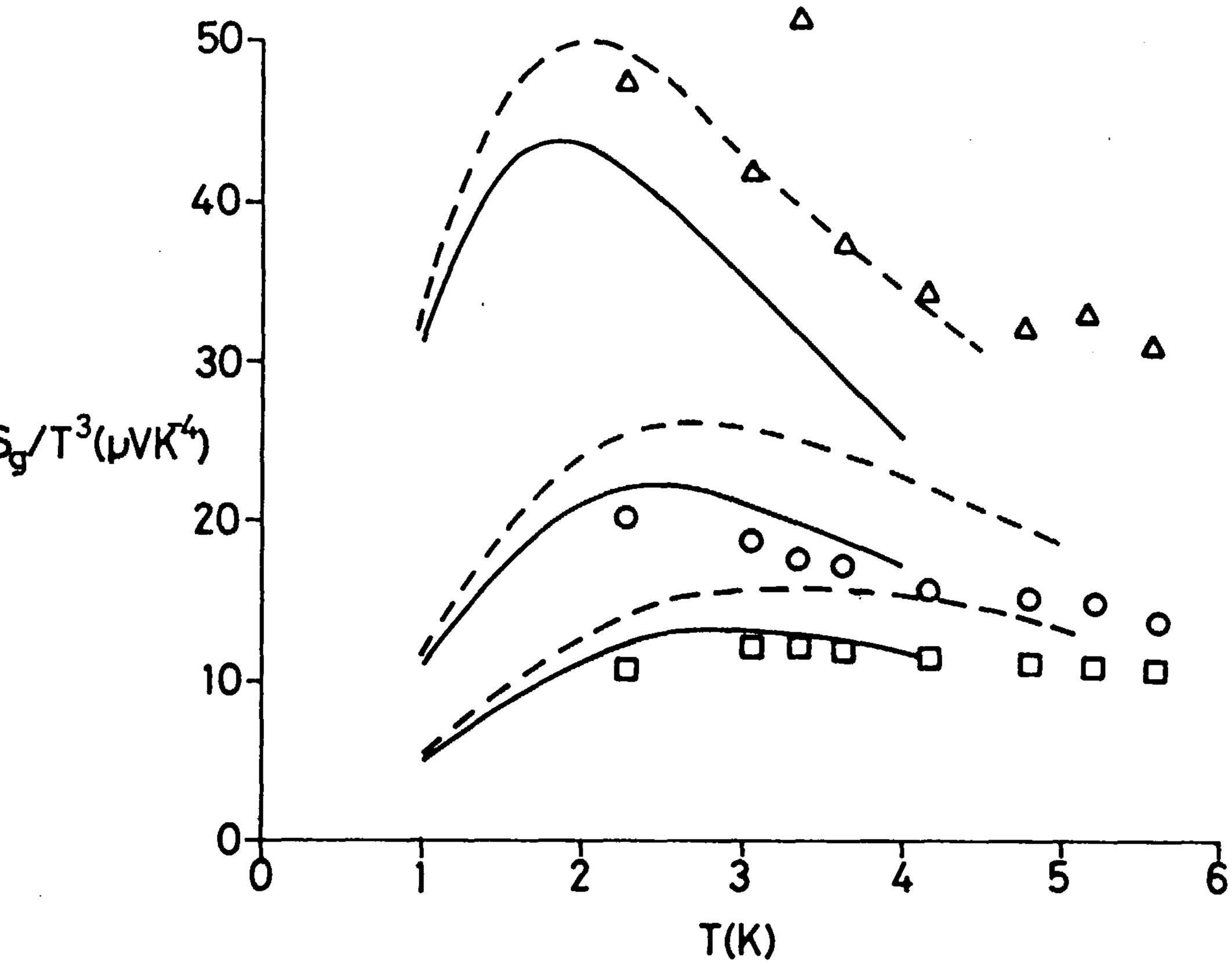
V_{dep} is linear. The image potential term contributes about 20% of δ but δ_A still reflects the variation with n. In Figure 3.5 the effect of varying δ upon S_g in the ISW model is illustrated. In the region of interest (δ_N), the dependence of S_g upon δ is not strong enough to change S_g by more than about 40% even if δ is halved or doubled. The effect of a closer treatment of the dependence of S_g upon channel width is, therefore, likely to be important only when the large discrepancy between the calculated and measured values is explained by other means (this is taken up in Chapter 5). For the present purposes, however, the more convenient analytic expression (3.23) for b may as well be used. The effect of using $\phi_b(z)$ in this way upon plots of $-S_g/T^3$ against T is compared with the ISW model and the data of Gallagher et al (1987) in Figure 3.6. The peak in such plots is a more severe test than the behaviour of $S_g(T)$ alone. In this, and previous plots, the system parameters are taken as in II for comparison (see Table 2.1). The experimental data is more closely followed in the variational model than the ISW but the large scaling and lack of data makes interpretation less clear.

Figure 3.5: The effect of channel width on the thermopower.



Plots of S_g against δ at T =4K for the data of Table 2.1. The values of delta indicated are derived from the full numerical expression (δ_N), the analytic formula for b (δ_A) and the constant value used in II (δ_{II}).

Figure 3.6: The effect of $\phi_b(z)$ on the thermopower.



Plot of $-S_g/T^3$ against T for S_g calculated using $\phi_b(z)$ (solid curves) and the ISW with δ=2nm (chain curves) compared to the corresponding experimental data points of Gallagher et al (1987) for n = 9.8(□),6.1(○) and 3.2(△)$\times 10^{15}$m^{-2}, scaled up by a factor of 12.

3.5 The Ando envelope function for heterojunctions.

A variational envelope function for heterojunctions is complicated by the need to allow for the penetration into the spacer region shown in Figure 3.2b. Such a function has been proposed by Ando (1982b) whereby:

$$\phi_{a,b}(z) = \begin{cases} Aa^{1/2}e^{(az/2)} & z < 0 \\ Bb^{1/2}(bz+\beta)e^{(-bz/2)} & z \geq 0. \end{cases} \tag{3.26}$$

The author has performed calculations for a GaAs/GaAlAs heterojunction (Fletcher et al 1988a) in the spirit of the preceding section and the discussion in Section 3.2. The properties of the envelope may then be examined and the results compared to the ISW model used in II.

The Ando function (3.25) initially appears to create a five parameter variational problem but normalization and continuity conditions, applied to the envelope and its derivative at $z = 0$, reduce this to two, through the relations:

$$\beta = 2b/(a+b), \tag{3.27}$$

$$B^2 = 1/[\beta^2 + 2\beta + 2 + \beta^3/(2-\beta)], \tag{3.28}$$

and:

$$A^2 = \beta^3 B^2/(2-B). \tag{3.29}$$

The resulting energy terms contributing to the total $E_{a,b}$ in (3.10) are expressible, then, in terms of a and b alone and the corresponding two-fold energy minimization condition results. The calculation is very similar to that for the MOSFET but there is no image term, when it is assumed that the relative permittivities of GaAs and GaAlAs (Landolt-Bornstein 1982) are the same (ϵ_r=12.9), and the isotropic effective mass m is used for m_z. The total potential must generate the $V(z)$ shown in Figure 3.2b. Contributions arise from: ionized donors $V_{don}(z)$ in the region of width d_2, ionized acceptors in the depletion layer (if any) $V_{depl}(z)$, $V_s(z)$ and the band gap discontinuity:

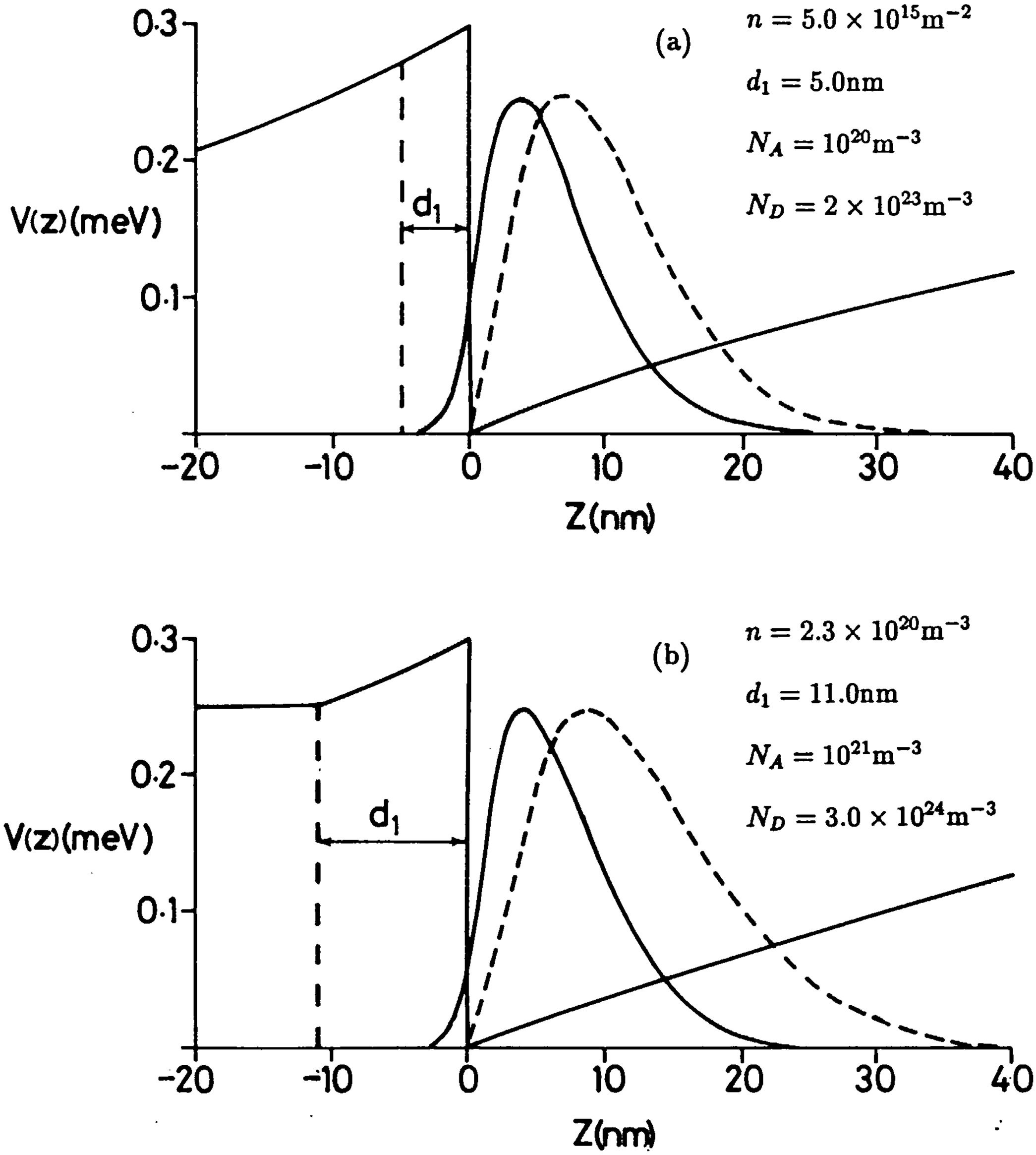
$$V_0(z) = \begin{cases} V_0 & z < 0 \\ 0 & z \leq 0. \end{cases} \tag{3.30}$$

The barrier height V_0 is taken as $x\ eV$ where x is the alloy concentration appearing in $GaAs/Ga_{1-x}Al_xAs$ (see, for example, Ando (1982), Stern and Das Sarma (1984)), which is otherwise taken as understood here.

The kinetic energy contribution to $E_{a,b}$ follows as for the MOSFET and $< V_0(z) >_z$ follows by direct integration in the two regions. The remaining potential energy contributions can be derived by using the known charge density, $\rho(z)$ say, in the corresponding Poisson equation and integrating twice. The first integration generates arbitrary constants for both sides of $z = 0$ which are matched and then determined by assuming that from infinity each separate charge distribution appears as a plane sheet of charge. For surface charge density σ the electric field at infinity is then $\sigma/2\varepsilon = e^{-1}dV/dz$. Overall charge neutrality ensures that the total field at infinity is zero because the resultant σ seen from infinity is zero. The final integration gives a constant which is determined by the zero of potential for the particular $\rho(z)$. Its precise value is unimportant, except for $V_s(z)$, because it is determined by the distribution $\rho(z)$ and the choice of origin, and is consequently independent of a and b. For $V_s(z)$ more care is necessary because $\rho(z)$ is determined by $| \phi_{a,b}(z) |^2$. The MOSFET case is more difficult to interpret this way because of the applied potential used to create the 2DEG.

The minimization of $E_{a,b}$ is performed numerically to determine a, b and thereby $\phi_{a,b}(z)$ and $V(z)$. Example results are illustrated in Figure 3.7 in order to compare with Ando (1982b) and examine the case of Fletcher et al (1988a) in which S_g results are reported. The behaviour of both $V(z)$ and $| \phi_{a,b}(z) |^2$ is qualitatively as expected with a large depletion width and some penetration into $z < 0$. Unfortunately though, there are no simple analogues of (3.24), for the Fang-Howard parameter, for a and b and more specimen information is required including V_0, d_1, N_D and N_A. Ando (1982b) compares the use of $\phi_{a,b}(z)$ and $\phi_b(z)$ in a GaAs/GaAlAs heterojunction in calculations of mobility and subband structure and Stern and Das Sarma (1984), for example, compare their self-consistent numerical results for the electron energy levels and envelopes with results using $\phi_b(z)$. Despite the latter neglecting the penetration into the spacer the net effect on the

Figure 3.7: Results of using the Ando envelope in heterojunctions.



Plots of the conduction band edge and electron distribution $| \phi(z) |^2$ calculated using the Ando envelope function $\phi_{ab}(z)$ for different specimens: (a)Ando (1982b) and (b)Fletcher et al (1988a),with data as shown. The chain curves represent $| \phi_b(z) |^2$ for comparison).

mobility is very small when compared to what is obtained using $\phi_{a,b}(z)$. This is also to be expected in the S_g calculations because the envelope function affects only $\mid Z_{11}(q_z) \mid^2$ and through (2.32) can be seen to depend upon the shape of $\mid \phi(z) \mid^2$ rather than the exact location (As seen in Figure 3.7 the effect on $\mid \phi(z) \mid^2$ of the two specific envelopes is broadly speaking to provide only a shift away or into the spacer). Thus the effect of finite barrier height upon the transport is relatively weak. For present purposes it is also interesting to compare the two variational models by calculating an equivalent ISW channel width. For the Ando envelope this is obtained in a similar way to the MOSFET and is given by:

$$\delta = 2B^2(\beta^2 + 4\beta + 6)/b + 2A^2/a \tag{3.31}$$

Values of 24.3 and 27.3(nm) are obtained for the data of Ando and Fletcher et al respectively (corresponding to the Figure). These should be compared with 20.4 and 26.4(nm) obtained from $\phi_b(z)$. The difference in these values of 16% and 4% is also indicative of the negligible effect on calculated S_g values which results from the rather weak dependence of S_g on the channel width. This suggests that in view of the large difference between the measured and calculated S_g values it may be an unnecessary complication to adopt $\phi_{a,b}(z)$ and that $\phi_b(z)$ is sufficient for present purposes. The corresponding estimate of 28.2 nm for δ in the heterojunction case considered in II should be compared with the value used there, of 10nm. The effect on S_g of this underestimate is similar to that arising in the case of the MOSFET which is illustrated in Figures 3.5 and 3.6.

3.6 Variational envelope functions.

The assumption of an abrupt change in material parameters at an interface is unrealistic because material in this region sees bulk material of a similar type on only one side. Furthermore, when $V(z)$ is discontinuous on an atomic scale the definition of an envelope function through the effective mass theorem, is violated. Interface grading of the potential barrier, effective mass and dielectric constant, for example, is expected then and is accounted for by Stern and Das Sarma (1984) in their heterojunction energy level

calculations. Transport calculations are not so sensitive to such details (Ando 1982b), and uncertainties in material parameter values are normally much more important. Some examples of this are discussed in Chapter 5 and similarly, exchange and correlation are ignored. Whereas the image potential is repulsive and tends to expand the envelope, the effect of $V_{xc}(z)$ is the opposite. Therefore, it is found to some extent that it is better to omit both terms in the variational calculation than to include either alone. This result makes the analytic formula for $\phi_b(z)$ which does neglect both, much more accurate than would appear from the results of Table 3.1. The resulting envelope function therefore proves to be a very close approximation to the self-consistent result with both terms included (see Ando et al (1982) in which numerical calculations by Stern are presented).

Alternative one parameter functions to those described here, such as proposed by Takada and Uemura (1977), and multiparameter functions, have been adopted but these are generally more complicated (Ando et al 1982). Problems can arise when the explicit functional form is important, (Matsumoto and Uemura 1974). This is expected because the functional form is only modelled approximately. Forms for higher subbands have been adopted (see for example, Mori and Ando 1979, Takada and Uemura 1977) but these are of no interest in the quantum limit except in accurate treatments of screening (see the following Chapter). Subband energy differences determined by self-consistent calculations are obviously important in spectroscopy but in transport are principally used to determine the onset of intersubband scattering.

Chapter 4

Screening and its effect upon phonon-drag in quasi-2D.

4.1 Introduction to the Chapter.

In this chapter the idea of screening is introduced to account for the effect of the Coulomb interaction within an electron gas upon an applied potential. Screening in 3D is used to introduce the formalism and show how a dielectric function can be defined and used to obtain values for free-electron matrix elements arising in the scattering rates of transport problems. The pure 2D limit is obtained from the same formalism and shown to be very similar to 3D. The quasi-2D case is complicated by the loss of translational invariance. The standard formulae are derived for the quasi-2D case and the single subband approximation (SSA) is introduced and shown to be valid in the quantum limit for sufficiently narrow channels. An effective quasi-2D multi-subband dielectric function is introduced and compared with the SSA in the systems of interest. It is then shown how screening may be introduced into the calculations of S_g of Chapter 2. It has a profound effect on the magnitude of the 'drag thermopower.

4.2 Introduction to screening.

In the calculation of S_g in Chapter 2 the potential field of the electron-phonon interaction is critical because it is the potential energy perturbation $U(\mathbf{r}, z)$ of equation (2.29) which causes the transition between electron states. This is a common feature of electron transport problems because it is an energy perturbation which is used to calculate scattering rates through the Golden Rule (see, for example, Butcher 1973). The assumption of independent electrons here is a common and convenient one although it is clear that correlation must be present in reality (see previous Chapter). Correlation changes the behaviour of otherwise free electrons and thereby the electron-phonon interaction. Ideally, then, the electron states would be treated more realistically but a simpler approach, adopted here, is to leave the electron states as independent and make a correction for correlation in a manner analogous to simple electrostatics.

Consider the application of an electrostatic potential to a simple (ie linear, isotropic, homogeneous) dielectric material (see, for example, Duffin 1980). The potential $\phi(\mathbf{R})$ which is observed can be related to that which would be seen in a vacuum $\phi_0(\mathbf{R})$ by using the (constant) relative permittivity κ_r of the medium.

$$\phi(\mathbf{R}) = \phi_0(\mathbf{R})/\kappa_r. \tag{4.1}$$

The effect of the medium is to modify the "bare" potential $\phi_0(\mathbf{R})$, which would otherwise be recorded, through the polarization $\mathbf{P}$ which results from charge separation in the applied field and can be related to the total field $\mathbf{E}$ by:

$$\mathbf{P} = \chi\varepsilon_0\mathbf{E} \tag{4.2}$$

The quantities κ_r and (the electric susceptibility) χ are related through:

$$\kappa_r = 1 + \chi. \tag{4.3}$$

An electron gas is similarly polarized by an applied field and the resultant potential which appears is therefore also modified from its "bare" value (ie that which would be taken if

the gas was unaffected). The potential from a positively charged impurity in a metal, for example, attracts a screen of electrons around it so that the net field seen at a distance is reduced. Conventionally it is said that the potential field is "screened" by the electron gas. An applied potential $\phi_0(\mathbf{R})$ also produces polarization. That which is observed $\phi(\mathbf{R})$ is the screened value of $\phi_0(\mathbf{R})$ and accounts for the redistribution of charge. Screening is naturally more important at high electron densities and to describe it a relation similar to (4.1) can be sought. The simplest linear relationship is:

$$\phi(\mathbf{R}) = \int \varepsilon(\mathbf{R}, \mathbf{R}')\phi(\mathbf{R}')d\mathbf{R}', \tag{4.4}$$

because the potentials at all points $\mathbf{R}'$ affect the potential at $\mathbf{R}$. The electronic potential energy is then $U(\mathbf{R}) = -e\phi(\mathbf{R})$. If the electron gas is homogeneous $\varepsilon(\mathbf{R}, \mathbf{R}')$ must be translationally invariant and can depend only upon relative $(\mathbf{R}-\mathbf{R}')$, rather than absolute, position. Equation (4.4) then has the form of a convolution integral and can be Fourier Transformed to leave the simpler relation:

$$\tilde{\phi}(\mathbf{Q}) = \tilde{\phi}_0(\mathbf{Q})/\varepsilon(\mathbf{Q}). \tag{4.5}$$

(The Transform of the potentials here is denoted by the tilde superscript but is left as understood in $\varepsilon(\mathbf{Q})$). A susceptibility $\chi(\mathbf{Q})$ can be similarly defined and is related to the dielectric function $\varepsilon(\mathbf{Q})$ through an equation analogous to (4.3):

$$\varepsilon(\mathbf{Q}) = 1 + \chi(\mathbf{Q}). \tag{4.6}$$

A density-density correlation function (or "response function") $\rho(\mathbf{Q})$ can also be defined in the manner of (4.4) to relate linearly the change $\delta n(\mathbf{R})$ in the electron density $n(\mathbf{R})$ at the point $\mathbf{R}$ to the total potential at all other points $V(\mathbf{R}')$. It is related to $\varepsilon(\mathbf{Q})$ and $\chi(\mathbf{Q})$ through:

$$\chi(\mathbf{Q}) = -\frac{e^2}{\varepsilon_0 Q^2}\rho(\mathbf{Q}) \tag{4.7}$$

and is given here for completeness because it is considered by some to be a more natural quantity to work with (see for example, Ashcroft and Mermin 1981, or Devrese and Brosens 1983).

Time t dependence can be accommodated in (4.4) by associating the labels t and t' with $\mathbf{R}$ and $\mathbf{R}'$ and performing a further integration (over t'). A frequency ω dependence is thereby introduced into the quantities such as $\varepsilon(\mathbf{Q}), \chi(\mathbf{Q})$ and $\rho(\mathbf{Q})$. When $\varepsilon(\mathbf{Q},\omega)$ is known and $\tilde{\phi}_0(\mathbf{Q},\omega)$ is calculated, $\phi(\mathbf{R},t)$ can be determined by inverting (4.5), with frequency dependence included.

The function $\varepsilon(\mathbf{Q},\omega)$ has considerable importance in solid-state physics (see, for example, Kittel 1976, Ashcroft and Mermin 1981 and Madelung 1981). For example, in the long wavelength limit ($Q \rightarrow 0$) the frequency dependence of $\varepsilon(\mathbf{0},\omega)$ describes the collective oscillations which can be supported by the electron gas (plasmon frequencies) and the dispersion relation for electromagnetic wave propagation. In the low frequency limit ($\omega \rightarrow 0$), the wavevector dependence of $\varepsilon(\mathbf{Q},0)$ describes the screening properties which are of interest here. Further discussion is given in the books by Ziman (1972) and Inkson (1984) and more advanced texts such as Devrese and Brosens (1983) and Mahan (1981). In the next section it is shown that the foregoing discussion in terms of Fourier Transforms is particularly suited to transport calculations. Providing that translational invariance is preserved, the results hold equally well for 2D as for 3D. Quasi-2D is more interesting because this symmetry is destroyed in the z direction.

There are many ways of considering the screening problem as the references already cited show. The Thomas-Fermi approximation (see, for example, Ashcroft and Mermin 1981) is the simplest and illustrates the basic result. The electrochemical potential $\mu_{ec}(\mathbf{R}) = \varepsilon_f(\mathbf{R}) - e\phi(\mathbf{R})$ of an electron in the potential $\phi(\mathbf{R})$ is approximated by the constant value of ε_f at $T = 0$. The result is:

$$\varepsilon(\mathbf{Q}) = 1 + Q_s^2/Q^2. \tag{4.8}$$

Where the Thomas-Fermi screening wavevector Q_s is obtained from:

$$Q_s^2 = e^2 g(\varepsilon_f)/\varepsilon_0 \tag{4.9}$$

with $g(\varepsilon_f)$ the density of states at the Fermi level. The screened potential at $\mathbf{R}(\neq 0)$ due

to an impurity with charge e at the origin, say, is then obtained by inverting (4.5). Hence:

$$\phi(\mathbf{R}) = \frac{e}{4\pi\varepsilon_0 R} e^{(-Q,R)}. \tag{4.10}$$

which shows how the Coulomb potential is screened out at large R. This is a physically appealing result in 3D because $g(\varepsilon_f)$ increases with the electron density n and thereby the screening gets stronger. In 2D, however, $g(\varepsilon_f)$ is constant (see Section 1.3) and the model gives the physically unacceptable result that the screening is independent of n. In the RPA (see below) this puzzle is resolved when it becomes clear that (4.9) breaks down when $Q > 2k_f$.

The simplest approximation which goes beyond Thomas-Fermi is the RPA (random phase approximation) or "self-consistent approach" which is used to obtain the well-known Lindhardt dielectric function (see Ehrenreich and Cohen (1959) for an early derivation). Each electron is assumed to move in a screened potential given by $\phi_0(\mathbf{R})$ plus a potential which is induced by the redistribution of all the electrons and is obtained from the solution of Poissons' equation. The RPA remains valid at large Q and reduces to the Thomas-Fermi result at small Q. It also has the advantage that the formalism can be readily adapted to treat 2D and quasi-2D. This was first treated in full by Siggia and Kwok (1970) but see also the paper by Mori and Ando (1979). The RPA is the approach used in the following to screen the calculations of S_g. Before the quasi-2D case is considered the basic formalism is developed for 3D and then applied to 2D. It should be noted that the electrons are assumed to respond as free particles to the mean field. In reality the electrons are not free and the field seen locally is not the mean value. Many improvements have been suggested and the shorter range exchange and correlation interactions can be included in addition to the Coulomb. For the general case see Mahan (1981) and for some discussion in quasi-2D see Ando et al (1982). These effects are of minor importance in transport calculations for the same reasons discussed in Chapter 3 but see also Mori and Ando (1979) and Ando et al (1982).

4.3 3D screening in the RPA.

The RPA dielectric function can be obtained by calculating the first order perturbation $\delta n(\mathbf{R})$ in the electron density arising from the application of a bare potential energy $V^b(\mathbf{R})e^{iwt}e^{\eta' t}$. When the unperturbed electron states are of the form $u_\lambda(\mathbf{R})e^{iw_\lambda t}$, with corresponding energies $\epsilon_\lambda = \hbar\omega_\lambda$, $\delta n(\mathbf{R})$ is given by:

$$\delta n(\mathbf{R}) = \sum_{\mu,\lambda} P_{\mu\lambda} < \mu \mid V(\mathbf{R}) \mid \lambda > u_\mu u^*_\lambda, \tag{4.11}$$

with:

$$P_{\mu\lambda} = \frac{2(f_\lambda - f_\mu)}{\epsilon_\lambda - \epsilon_\mu + i\eta - \hbar\omega}. \tag{4.12}$$

The * superscript denotes the complex conjugate, f_λ the occupation probability of the state with label λ by an electron with given spin and $\eta(= \hbar\eta')$ is some arbitrary small parameter indicating the slow growth of the perturbation. The potential $V(\mathbf{R})$ is the screened value of $V^b(\mathbf{R})$ and is written as:

$$V(\mathbf{R}) = V^b(\mathbf{R}) + V^i(\mathbf{R}), \tag{4.13}$$

where $V^i(\mathbf{R})$ is the potential induced by $\delta n(\mathbf{R})$ and satisfies-

$$\nabla.[\kappa(\mathbf{R})\nabla V^i(\mathbf{R})] = -e^2\delta n(\mathbf{R}). \tag{4.14}$$

The permittivity $\kappa(\mathbf{R})$ of the background medium enters the equation in this way to allow for inhomogeneity. Writing:

$$\nabla.[\kappa(\mathbf{R})\nabla G(\mathbf{R},\mathbf{R}')] = -e^2\delta(\mathbf{R}-\mathbf{R}') \tag{4.15}$$

the induced potential is obtained from:

$$V^i(\mathbf{R}) = \int G(\mathbf{R},\mathbf{R}')\delta n(\mathbf{R}')d\mathbf{R}', \tag{4.16}$$

and can be used in (4.13). Moreover, substituting for $\delta n(\mathbf{R}')$ from (4.11) and forming the matrix element between the states with labels α and β:

$$< \alpha \mid V^b(\mathbf{R}) \mid \beta > = < \alpha \mid V(\mathbf{R}) \mid \beta > - \sum_{\mu\lambda} P_{\mu\lambda} < \mu \mid V(\mathbf{R}) \mid \lambda > I(\alpha\beta,\mu\lambda) \tag{4.17}$$

where:

$$I(\alpha\beta,\mu\lambda) = \int\int G(\mathbf{R},\mathbf{R}')u^*_\alpha(\mathbf{R})u_\beta(\mathbf{R})u_\mu(\mathbf{R}')u^*_\lambda(\mathbf{R}')d\mathbf{R}d\mathbf{R}'. \tag{4.18}$$

These last two equations give the general result which can be applied to derive the dielectric function for particular cases.

Consider first the case of a homogeneous 3D free-electron gas with a homogeneous background of constant permittivity κ. Free-electron states are assumed so that the state labels such as λ become associated with wavevectors ($\mathbf{K}$). In a convenient form then:

$$u_\lambda(\mathbf{R}) = V^{-1/2}e^{i\mathbf{K}.\mathbf{R}} \tag{4.19}$$

$$u_\mu(\mathbf{R}) = V^{-1/2}e^{i(\mathbf{K}+\mathbf{Q}').\mathbf{R}} \tag{4.20}$$

$$u_\beta(\mathbf{R}) = V^{-1/2}e^{i\mathbf{K}'.\mathbf{R}} \tag{4.21}$$

and

$$u_\alpha(\mathbf{R}) = V^{-1/2}e^{i(\mathbf{K}'+\mathbf{Q}).\mathbf{R}}, \tag{4.22}$$

with V the system volume. The matrix elements in (4.17) are then equal to the Fourier Transforms of the potentials, eg:

$$< \alpha \mid V^b(\mathbf{R}) \mid \beta >= \tilde{V}^b(\mathbf{Q}) \tag{4.23}$$

The solution to (4.15) is the Coulomb energy at $\mathbf{R}$ due to a charge $-e$ at $\mathbf{R}'$, and is a function of $\mid \mathbf{R}-\mathbf{R}' \mid$ only, ie:

$$G(\mathbf{R},\mathbf{R}') \equiv G(\mathbf{R}-\mathbf{R}') = \frac{e^2}{4\pi\kappa \mid \mathbf{R}-\mathbf{R}' \mid}. \tag{4.24}$$

It therefore has Fourier Transform:

$$\tilde{G}(\mathbf{Q}) = \frac{e^2}{V\kappa Q^2} \equiv V_c(\mathbf{Q}), \tag{4.25}$$

where V is the system volume. The quantity $I(\alpha\beta,\mu\lambda)$ can now be evaluated and is equal to $V_c(\mathbf{Q})\delta_{\mathbf{Q},\mathbf{Q}'}$. The screening equation (4.17) can therefore be written:

$$\tilde{V}^b(\mathbf{Q}) = \tilde{V}(\mathbf{Q})[1 - V_c(\mathbf{Q})\Pi(\mathbf{Q})], \tag{4.26}$$

where the "polarizability" $\Pi(\mathbf{Q})$ is given by:

$$\Pi(\mathbf{Q}) = \frac{2}{V} \sum_{\mathbf{K}} \frac{f_{\mathbf{K}} - f_{\mathbf{K+Q}}}{\varepsilon_{\mathbf{K}} - \varepsilon_{\mathbf{K+Q}} + i\eta - \hbar\omega}. \tag{4.27}$$

Comparing (4.26) with (4.5) the dielectric function is:

$$\varepsilon(\mathbf{Q}) = 1 - V_c(\mathbf{Q})\Pi(\mathbf{Q}). \tag{4.28}$$

This is the standard Lindhardt result. (The frequency dependence arising from (4.27) is left as understood here because in this chapter the static case (where η and ω are zero) is of interest and is assumed hereafter). Derivations and some discussion of the properties of this dielectric function are given by Ehrenreich and Cohen (1959), Madelung (1981) and Mahan (1981).

The result (4.26) can be written as:

$$\tilde{V}(\mathbf{Q}) = \tilde{V}^b(\mathbf{Q})/\varepsilon(\mathbf{Q}). \tag{4.29}$$

Its importance in transport calculations can be seen by considering the scattering rate $R(\beta, \alpha)$, say, for the transition from state u_β to u_α caused by some potential energy perturbation which takes the value $V^b(\mathbf{R})$ when screening is ignored. It is the screened value $V(\mathbf{R})$ which the electrons see. Hence:

$$R(\beta, \alpha) = \frac{2\pi}{\hbar} |< \alpha \mid V(\mathbf{R}) \mid \beta >|^2 \, \delta(\varepsilon_\alpha - \varepsilon_\beta - \hbar\omega), \tag{4.30}$$

when the energy difference between the states is $\varepsilon_\alpha - \varepsilon_\beta = \hbar\omega$. Following (4.23) and (4.29) the result is:

$$R(\beta, \alpha) = \frac{2\pi}{\hbar} \mid \tilde{V}^b(\mathbf{Q})/\varepsilon(\mathbf{Q}) \mid^2 \delta(\varepsilon_\alpha - \varepsilon_\beta - \hbar\omega). \tag{4.31}$$

Screening is accounted for very conveniently then by dividing the known Transform of the bare potential by the dielectric function, with both evaluated at the wavevector $\mathbf{Q}$ which links the states. A corresponding result is now sought for 2D and quasi-2D.

4.4 Screening in 2D and quasi-2D : theory.

To apply the formalism to the case of a quasi-2D electron gas the 3D electron states of the previous section must be replaced by quasi-2D states as in (1.10). Hence the state label λ

is characterized by a 2D wavevector $\mathbf{k}$ and subband index λ so that, for example:

$$u_\lambda(\mathbf{R}) = A^{-1/2} e^{i\mathbf{k}.\mathbf{r}} \phi_\lambda(z), \tag{4.32}$$

replaces (4.19). The states with labels μ, β and α follow analogously. The 3D matrix elements in (4.17) are then equal to 1D matrix elements between subbands of the 2D Fourier transforms of the potentials. For example:

$$< \alpha, \mathbf{k}' + \mathbf{q} \mid V(\mathbf{r}, z) \mid \beta, \mathbf{k}' > = < \alpha \mid \tilde{V}(\mathbf{q}, z) \mid \beta > \equiv \tilde{V}_{\alpha\beta}(\mathbf{q}), \tag{4.33}$$

where the 2D Transform $\tilde{V}(\mathbf{q}, z)$ is defined by:

$$V(\mathbf{r}, z) = \sum_{\mathbf{q}} e^{i\mathbf{q}.\mathbf{r}} \tilde{V}(\mathbf{q}, z). \tag{4.34}$$

and $\mid \beta >$, say, represents the $\phi_\beta(z)$ envelope. For the MOSFET or heterojunction the permittivity is a function of z and, writing $\tilde{g}(\mathbf{q}, z, z')$ for the 2D transform of $G(\mathbf{R}, \mathbf{R}')$, (4.15) can be written:

$$\frac{d}{dz}\left[\kappa(z)\frac{d}{dz}\tilde{g}(\mathbf{q}, z, z')\right] - q^2\kappa(z)\tilde{g}(\mathbf{q}, z, z') = -e^2\delta(z - z'). \tag{4.35}$$

When $\kappa(z)$ is a constant κ, say, the solution is:

$$\tilde{g}(\mathbf{q}, z, z') = V_c'(q) e^{-q|z-z'|}, \tag{4.36}$$

where $V_c'(q) = e^2/2\kappa q A$ and is the 2D transform of the Coulomb potential between electrons restricted to the 2D plane. For the general case:

$$I(\alpha\beta\mu\lambda) = V_c'(q) F(\alpha\beta\mu\lambda, q) \delta_{\mathbf{q},\mathbf{q}'} \tag{4.37}$$

where:

$$V_c'(q) F(\alpha\beta, \mu\lambda, q) = \int\int \tilde{g}(\mathbf{q}, z, z') \phi_\alpha^*(z) \phi_\beta(z) \phi_\mu(z') \phi_\lambda^*(z') dz dz' \tag{4.38}$$

and the screening equation becomes:

$$\tilde{V}_{\alpha\beta}^b(\mathbf{q}) = \tilde{V}_{\alpha\beta}(\mathbf{q}) - V_c'(q) \sum_{\mu\lambda} F(\alpha\beta\mu\lambda, q) \Pi_{\mu\lambda}(\mathbf{q}) \tilde{V}_{\mu\lambda}(\mathbf{q}). \tag{4.39}$$

The quasi-2D polarizability is defined analogously to that for 3D by:

$$\Pi_{\mu\lambda}(\mathbf{q}) = \frac{2}{A}\sum_{\mathbf{k}} \frac{f_{\lambda,\mathbf{k}} - f_{\mu,\mathbf{k+q}}}{\varepsilon_{\lambda,\mathbf{k}} - \varepsilon_{\mu,\mathbf{k+q}} + i\eta - \hbar\omega}. \tag{4.40}$$

The screening equation can also be written as:

$$\tilde{V}^{b}_{\alpha\beta}(\mathbf{q}) = \sum_{\mu\lambda} \varepsilon_{\alpha\beta\mu\lambda}(\mathbf{q})\tilde{V}_{\mu\lambda}(\mathbf{q}) \tag{4.41}$$

where:

$$\varepsilon_{\alpha\beta\mu\lambda}(\mathbf{q}) = \delta_{\mu\lambda}\delta_{\lambda\beta} - V'_{c}(\mathbf{q})F(\alpha\beta\mu\lambda, q)\Pi_{\mu\lambda}(\mathbf{q}). \tag{4.42}$$

This is the multi-subband screening (MSS) result obtained by Siggia and Kwok (1970) who evaluate $\Pi_{\mu\lambda}(\mathbf{q})$. It can be written more compactly in the form:

$$\underline{\tilde{V}}^{b} = \hat{\varepsilon}\,\underline{\tilde{V}} \tag{4.43}$$

where $\underline{\tilde{V}}^{b}$ is a matrix whose elements are given through (4.41). Each depends upon *all* of the elements of the matrix $\underline{\tilde{V}}$ through the action of $\hat{\varepsilon}$. To see how this equation may be inverted to find elements of $\underline{\tilde{V}}$ consider $\underline{\tilde{V}}^{b}$ and $\underline{\tilde{V}}$ as column vectors by writing their elements in order and link them with $\hat{\varepsilon}$ in matrix form. For only two subbands for example, the following equation is inverted:

$$\begin{pmatrix} \tilde{V}^{b}_{11} \\ \tilde{V}^{b}_{12} \\ \tilde{V}^{b}_{21} \\ \tilde{V}^{b}_{22} \end{pmatrix} = \begin{pmatrix} \varepsilon_{1111} & \varepsilon_{1112} & \varepsilon_{1121} & \varepsilon_{1122} \\ \varepsilon_{1211} & \varepsilon_{1212} & \varepsilon_{1221} & \varepsilon_{1222} \\ \varepsilon_{2111} & \varepsilon_{2112} & \varepsilon_{2121} & \varepsilon_{2122} \\ \varepsilon_{2211} & \varepsilon_{2212} & \varepsilon_{2221} & \varepsilon_{2222} \end{pmatrix} \begin{pmatrix} \tilde{V}_{11} \\ \tilde{V}_{12} \\ \tilde{V}_{21} \\ \tilde{V}_{22} \end{pmatrix} \tag{4.44}$$

(The $\mathbf{q}$ dependence of all terms has been omitted for clarity.) For n subbands $\hat{\varepsilon}$ has $n^2 \times n^2$ elements. The different subband labels' contributions to each element of $\underline{\tilde{V}}$ arise because the perturbed electron density is expanded in the complete set of subbands. This is true, therefore, independent of whether the states are occupied or not. That a separate expression arises for each $\tilde{V}_{\alpha\beta}(\mathbf{q})$ should be expected. Whereas in 3D only the wavevector ($\mathbf{Q}$) is required to link states, in quasi-2D both $\mathbf{q}$ and the two subband indices are required. In pure 2D however, $\mathbf{q}$ is sufficient and an analogous dielectric function to 3D, $\varepsilon(\mathbf{q})$, can be defined.

Consider the quasi-2D case in the limit of small channel width δ. First it is necessary to solve (4.35). To maintain contact with the heterojunction and MOSFET let the electrons occupy material of permittivity $\kappa_2(z \geq 0)$ at the interface $(z = 0)$ with material of permittivity $\kappa_1(z < 0)$, corresponding to Figure 3.2. The solution is then:

$$\tilde{g}(\mathbf{q}, z, z') = \begin{cases} V_c'(q) e^{-q|z-z'|} & z < 0 \\ \frac{1}{2} V_c'(q) \left\{ \frac{(\kappa_2+\kappa_1)}{\kappa_2} e^{-q|z-z'|} + \left(\frac{\kappa_2-\kappa_1}{\kappa_2} \right) e^{-q|z+z'|} \right\} & z \geq 0, \end{cases} \tag{4.45}$$

with $V_c'(\mathbf{q})$ evaluated with the mean permittivity $\bar{\kappa} = (\kappa_1 + \kappa_2)/2$. This result (4.45) reduces to (4.36) when the permittivities are the same. The general result gives the $\mathbf{q}$ component of the net electrostatic potential, including the image interaction, between electrons at $\mathbf{r}$ and $\mathbf{r}'$ in the electron gas at z and z' across the channel. For a narrow channel the energies of the subband minima become widely separated so that the denominators of polarizability terms (4.40) other than with $\mu = \lambda$ become very large. Moreover, $f_{\lambda,\mathbf{k}}$ and $f_{\mu,\mathbf{k+q}}$ become very small other than for the ground subband. Furthermore, for $T = 0$, $f_{\lambda,\mathbf{k}}$ is zero for $\varepsilon_f < \varepsilon_\lambda$ and therefore $\Pi_{\mu,\lambda}(\mathbf{q})$ is vanishingly small for all (μ, λ) except (1,1), when δ tends to zero. Only $\tilde{V}_{11}(\mathbf{q})$ is of interest because the probability of scattering to other subbands is very small. The screening equation for narrow channels is therefore:

$$\tilde{V}_{11}(\mathbf{q}) = \tilde{V}_{11}^b(\mathbf{q})/\varepsilon(\mathbf{q}) \tag{4.46}$$

where the dielectric function is $\varepsilon_{1111}(q)$ and is commonly written in the form:

$$\varepsilon(\mathbf{q}) = 1 - V_c'(q) F(1111, q) \Pi_{11}(\mathbf{q}). \tag{4.47}$$

(see, for example, Ando et al 1982). The quantity $\Pi_{11}(\mathbf{q})$ is the 2D polarizability calculated by Stern (1967). It has the value $-(g_v m^*/\pi\hbar)\xi(q)$ where:

$$\xi(q) = \begin{cases} 1 & q \leq 2k_f \\ 1 - [1 - (2k_f/q)^2]^{1/2} & q > 2k_f, \end{cases} \tag{4.48}$$

and follows from direct integration of (4.40). The dielectric function can then be written as:

$$\varepsilon(q) = 1 + (q_s/q)\xi(q) F(1111, q) \tag{4.49}$$

where the 2D screening constant analogous to (4.9) is defined by:

$$q_s = g_v me^2/2\pi\hbar^2\kappa \tag{4.50}$$

The quantity $V_c'(q)F(1111,q)$ is the average of $\tilde{g}(\mathbf{q},z,z')$ weighted by the ground subband electron distribution $\mid \phi_1 \mid^2$ over z and z'. In the 2D limit $\mid \phi_1(z) \mid^2$ becomes the delta function $\delta(z)$, and $F(1111,q)$ equals unity. This is the pure 2D result determined by Stern. It is a natural extension from 2D to obtain (4.47) for the case of a narrow but finite channel but no information is given about when this single subband approximation (SSA) to MSS breaks down. It should be anticipated that further polarizability terms will contribute when the channel width or Fermi energy become sufficiently large. The SSA discards all information about the perturbed electron density carried by higher subbands.

4.5 Screening in quasi-2D: calculations.

In the calculations of S_g in Sections 2.4 and 2.6 the electron-phonon interaction for single phonon absorption replaces $V^b(\mathbf{R})$ and the corresponding value of $\tilde{V}_{11}(\mathbf{q})$ enters the expression (2.31) for the transition rate in the quantum limit. An indication of the effect of screening upon S_g is given, then, by evaluating $\varepsilon^{-2}(q)$ from (4.47) for the 2D limit when $q = 2k_f$. In GaAs for $n = 9.8\times10^{15}\mathrm{m}^{-2}$ (the largest value in the MOSFET data of Gallagher et al 1987) S_g can be expected to be reduced by a factor of about 1/2. For the MOSFET the corresponding factor is 1/20 due to the larger effective mass and the valley degeneracy, which increase q_s. At a given q the Form factor $F(1111,q)$ reduces the screening effect in a finite width channel, because, for the same surface density, the electrons are further apart. A range of wavevectors contribute to S_g including $q < 2k_f$ where screening is stronger. The effect upon the calculated magnitude of S_g for the MOSFET data of Section 2.6, then, is likely to be very important as suggested by Gallagher et al (1987), but less important, although still significant, for the heterojunction.

To determine whether the SSA is valid for the calculations of S_g a measure of the importance of MSS is necessary over the required regime. There is apparently no simple

MSS analogue of $\varepsilon(q)$, however, because in principle all the subbands are required in order to invert (4.43) and evaluate $\tilde{V}_{11}(q)$. The only simple model for which this is practical is the ISW. An *effective* dielectric function *can* be defined however.

Consider first the matrix equation, (4.43), and (4.41) in terms of the inverse of the dielectric matrix:

$$\tilde{V}_{\alpha\beta}(\mathbf{q}) = \sum_{\mu\lambda} \varepsilon^{-1}_{\alpha\beta\mu\lambda}(\mathbf{q}) \tilde{V}^{b}_{\mu\lambda}(\mathbf{q}). \tag{4.51}$$

Expanding the potentials, $\tilde{V}(q,z)$ for example, within the matrix elements, as Fourier series:

$$\tilde{V}(\mathbf{q},z) = \sum_{q_z} \tilde{V}^{b}(\mathbf{Q}) e^{iq_z z}, \tag{4.52}$$

(4.51) can be rewritten in the form:

$$\tilde{V}_{\alpha\beta}(\mathbf{q}) = \sum_{q_z} \frac{\tilde{V}^{b}(\mathbf{Q}) < \alpha \mid e^{iq_z z} \mid \beta >}{\varepsilon_{\alpha\beta}(\mathbf{Q})}, \tag{4.53}$$

where the quantity $\varepsilon_{\alpha\beta}(\mathbf{Q})$ is defined by:

$$\varepsilon_{\alpha\beta}(\mathbf{Q}) = \frac{< \alpha \mid e^{iq_z z} \mid \beta >}{\sum_{\mu\lambda} \varepsilon^{-1}_{\alpha\beta\mu\lambda}(\mathbf{q}) < \mu \mid e^{iq_z z} \mid \lambda >}. \tag{4.54}$$

For the SSA the corresponding equation to (4.53) is:

$$\tilde{V}_{\alpha\beta}(\mathbf{q}) = \sum_{q_z} \frac{\tilde{V}^{b}(\mathbf{Q}) < \alpha \mid e^{iq_z z} \mid \beta >}{\varepsilon(q)}. \tag{4.55}$$

The relation (4.53) is a convenient reformulation of the MSS equation and highlights the similarity of the problem with the simple 3D case of (4.29). Moreover, $\varepsilon_{\alpha\beta}(\mathbf{Q})$ can be considered as an effective MSS dielectric function by comparing (4.53) with (4.55). The influence of higher subbands is reflected by the labels α and β for particular matrix elements and the loss of translational invariance by the q_z dependence. The latter is treated only approximately in quasi-2D by $\varepsilon(q)$ through the Form factor in (4.47) because $\varepsilon(q)$ is the SSA to $\varepsilon_{\alpha\beta}(\mathbf{Q})$. The importance of MSS over the SSA is measured by the quantity:

$$f_{\alpha\beta}(\mathbf{Q}) = \frac{\varepsilon_{\alpha\beta}(\mathbf{Q})}{\varepsilon(q)}. \tag{4.56}$$

This is the amount by which the SSA multiplies each term in the summand of (4.53) and can be used to determine the importance of MSS effects for any $V^b(\mathbf{R})$ and set of subbands.

For the electron-phonon interaction potential (2.29) of interest only one phonon occupation number changes in the expression for the transition rate and from (2.30) only the corresponding coefficient of $a_{\mathbf{Q}}$ contributes to S_g. Hence only the one component ($\mathbf{Q}$) in the Fourier expansion of $V^b(\mathbf{R})$ and only one q_z term in $\tilde{V}_{\alpha\beta}(\mathbf{q})$ need consideration. The effect on the thermopower of taking the SSA to MSS is thus measured by $\mid f_{11}(\mathbf{Q}) \mid^2$. Deviations of this quantity from unity indicate the importance of MSS effects (ie higher subbands).

As described in Smith and Butcher (1989a) calculations have been performed by the author in the ISW model to determine this importance in the Si MOSFET case. The calculations in GaAs (heterojunction) follow analogously but in Si the effect upon S_g is most pronounced because the screening is stronger. A finite number n_s of subbands was assumed for $n_s = 1, 2, 3...$, and calculations of $\epsilon(q)$, $\varepsilon_{\alpha\beta}(\mathbf{Q})$ and $\mid f_{\alpha\beta}(\mathbf{Q}) \mid^2$ were performed. About five subbands or less were required to achieve convergence. A wide parameter range was explored with δ from 0.1 to 30 nm, n from 10^{15} to $10^{17}\mathrm{m}^{-2}$ and q and q_z up to $10k_f$. For the lowest electron densities the SSA is good over the whole range for all δ. For large δ (20-30 nm) deviations from MSS up to 1% begin to appear for $q_z > 4k_f$. These deviations decrease for given q_z as q increases and increase at given q as q_z increases. For $n = 10^{16}\mathrm{m}^{-2}$ they are small for all q,q_z for δ up to 5nm but are up to 10% even at small q when δ is about 20nm or greater. Thus the SSA becomes inaccurate for wide channels, high electron densities and large q_z (which is only accounted for in MSS). However, the regime of interest to S_g in the Si MOSFET calculations for the data of Gallagher et al (1987) is characterized by $n < 10^{16}\mathrm{m}^{-2}$ and δ to 10nm. The SSA is likely to be good then because the dominant contributions to S_g arise around $\mid \mathbf{Q} \mid \leq 2k_f$.

Calculations were performed using both approaches for the MOSFET and the heterojunction. The results are discussed in the next section. The calculation of the quantities

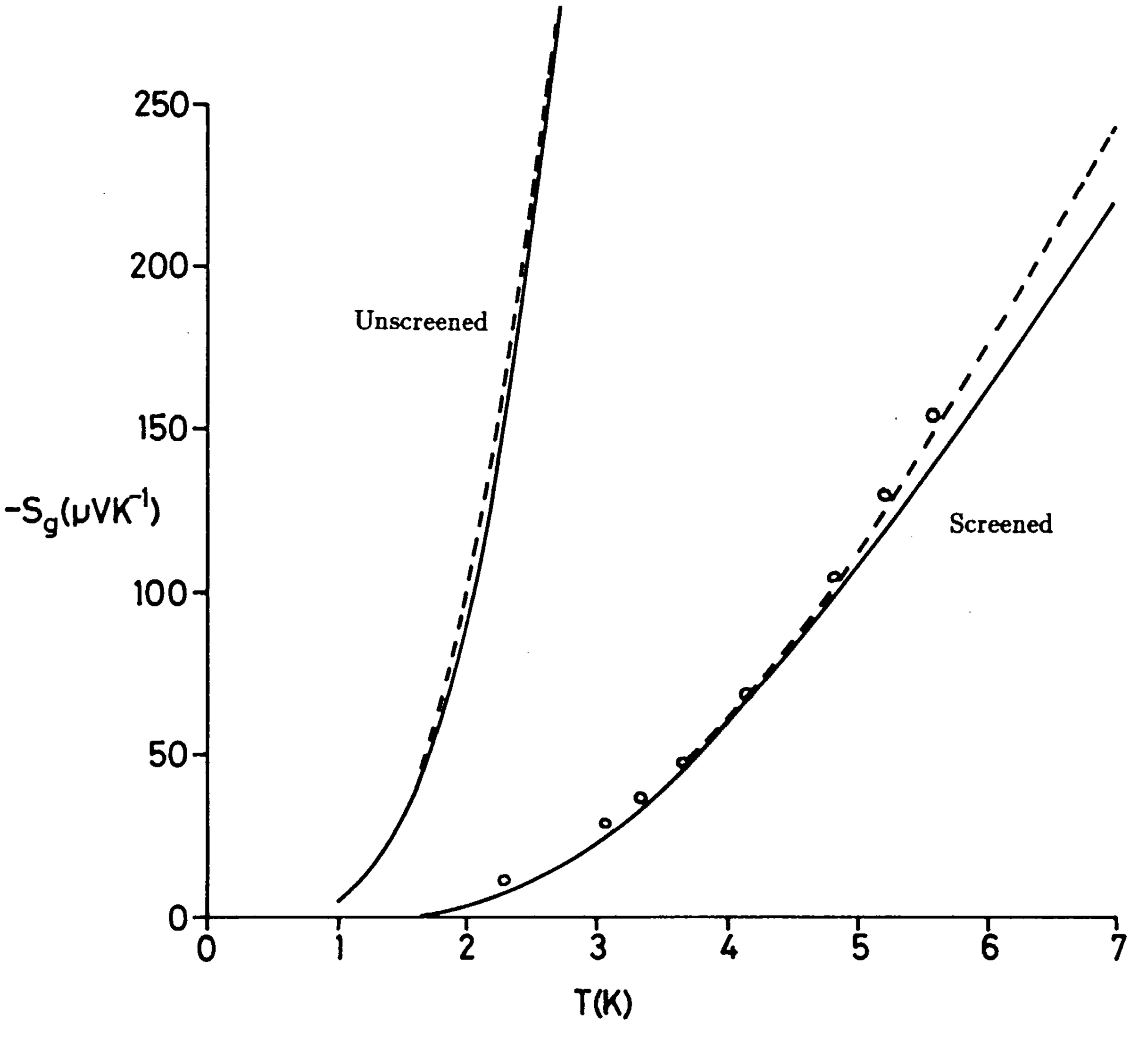
$\Pi_{\mu,\lambda}(\mathbf{q}), F(\alpha\beta\mu\lambda, q)$ and $< \alpha \mid e^{iq_z z} \mid \beta >$, required in the ISW model, is described in the Appendix.

4.6 The effect of screening on the thermopower.

The effect of screening on the calculated values of S_g is illustrated in Figures 4.1 to 4.3. For comparison the same parameter values and experimental data as in II are considered here (see Table 2.1). For Si (the MOSFET) at the highest electron density ($9.8 \times 10^{15}\mathrm{m}^{-2}$), where MSS effects are greatest, the difference between the screened values in MSS and the SSA is negligible for the ISW width of 2nm used in II and is at most 0.2% when $\delta = 8nm$. A similar conclusion is reached for the heterojunction data and is supported by the calculations of Tang (1988). Thus, for the range of the experimental data, the channels are narrow enough and the electron densities low enough for the SSA to be adopted. Moreover, the important contributions to S_g (shown in the next chapter to be from phonons with wavevector components q not much greater than $2k_f$) are not large enough to magnify the deviations of MSS from the SSA, which occur at large Q.

Using the estimates of δ of the previous chapter, S_g can be calculated both with and without taking account of screening. The effect on S_g is a significant improvement in the agreement with the experimental data . The effect on S_g in Si is particularly dramatic with the large overestimate by a factor of about 12 or more, reduced to a difference of about 10% for the largest n. This reduction is about the same size as that expected following the rough estimate of Section 4.5. A closer description of the ground subband envelope function is therefore warranted now because the information discarded by taking an ISW of constant width is now comparable to the difference between theory and experiment. In the figures therefore, results are presented for the ISW and variational envelope functions (which will be assumed hereafter). Figure 4.1 illustrates the large difference between the screened and unscreened S_g results for the largest n and the excellent agreement between the screened results and the data. For clarity other results are not plotted on the same graph as the results are rather similar (see Smith and Butcher 1989a) but see also Figure

Figure 4.1: The effect of screening upon S_g in Si.



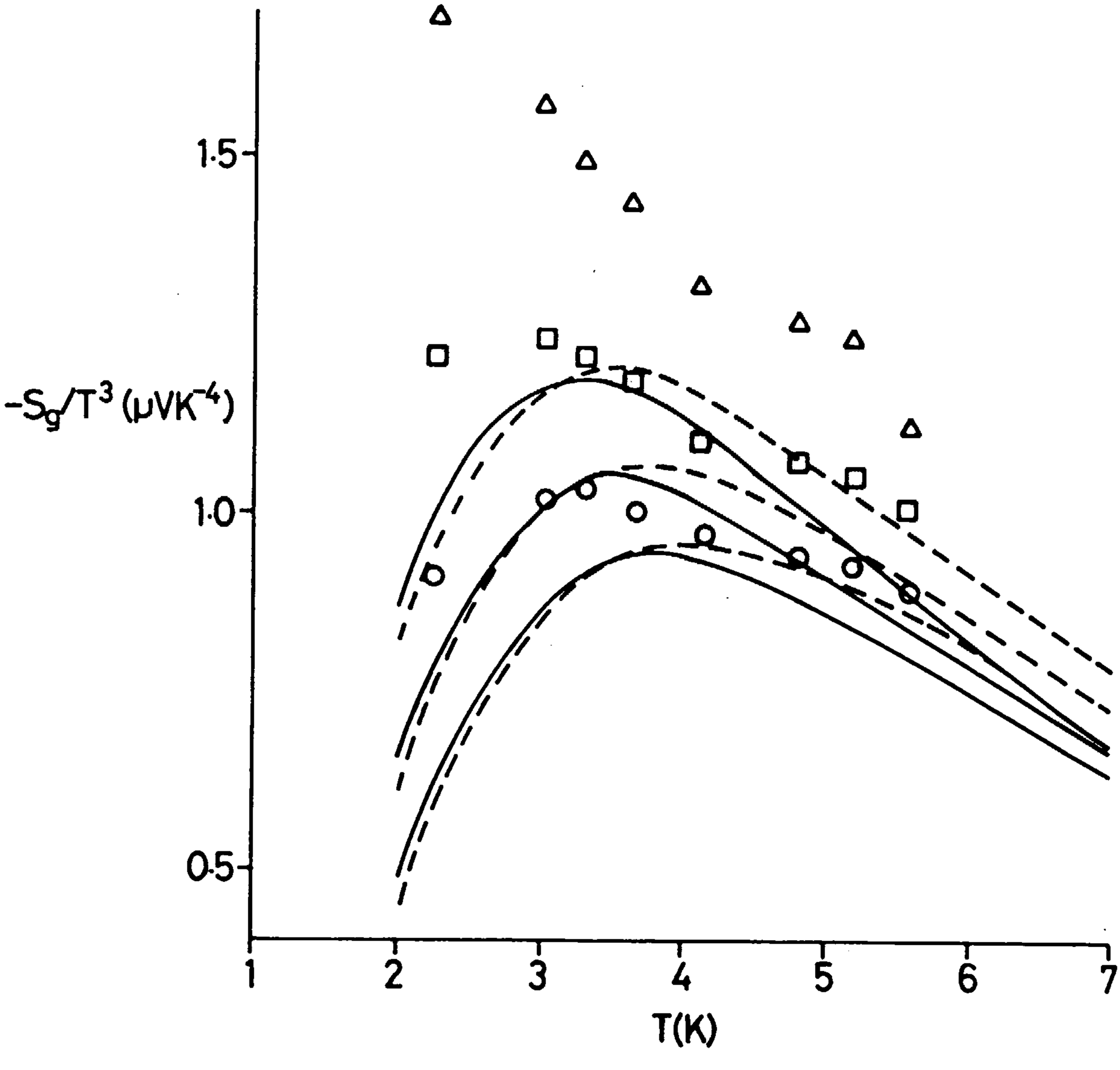
Plots of $-S_g$ against T for $n = 9.8\times10^{15}\mathrm{m}^{-2}$ compared with the experimental data of Gallagher et al (1987). A solid line denotes a calculation using the variational envelope and a broken line the ISW of equivalent width.

4.2. The size and trend in the (screened) calculated values follow the experimental data very closely. A more severe test of the theory are the screened values given for $-S_g/T^3$, illustrated for the three highest n of II in Figure 4.2. The variational results are very similar to those obtained using an ISW of equivalent width. Peaks arise just as when screening is ignored (see Figure 2.3) but the positions of the maxima for given n are shifted by about 0.75K. This is a direct consequence of introducing screening and is returned to in Chapter 5. Quantitatively the results are clearly very close to the data for large n but increasingly underestimate S_g at lower n. Moreover, the experimental peaks appear to be rather flatter than those arising in the calculations and may even be absent at low n judging from Figure 4.2. (This is discussed further in the next chapter following comparison with more complete data).

As shown in Figure 4.3 the results for the heterojunction are qualitatively similar to those for the MOSFET but see also Figure 4.2. The effect of screening is indeed weaker in GaAs and reduces the unscreened values by only a factor of about 1/2. For the case shown, however, screening takes the calculated values *further below* the experimental data. Note that the ISW of equivalent width appears to be a worse approximation to the variational envelope in GaAs than in Si (Figure 4.1) but, in addition to the different materials, n and δ are very different in the two cases (compare the figure captions).

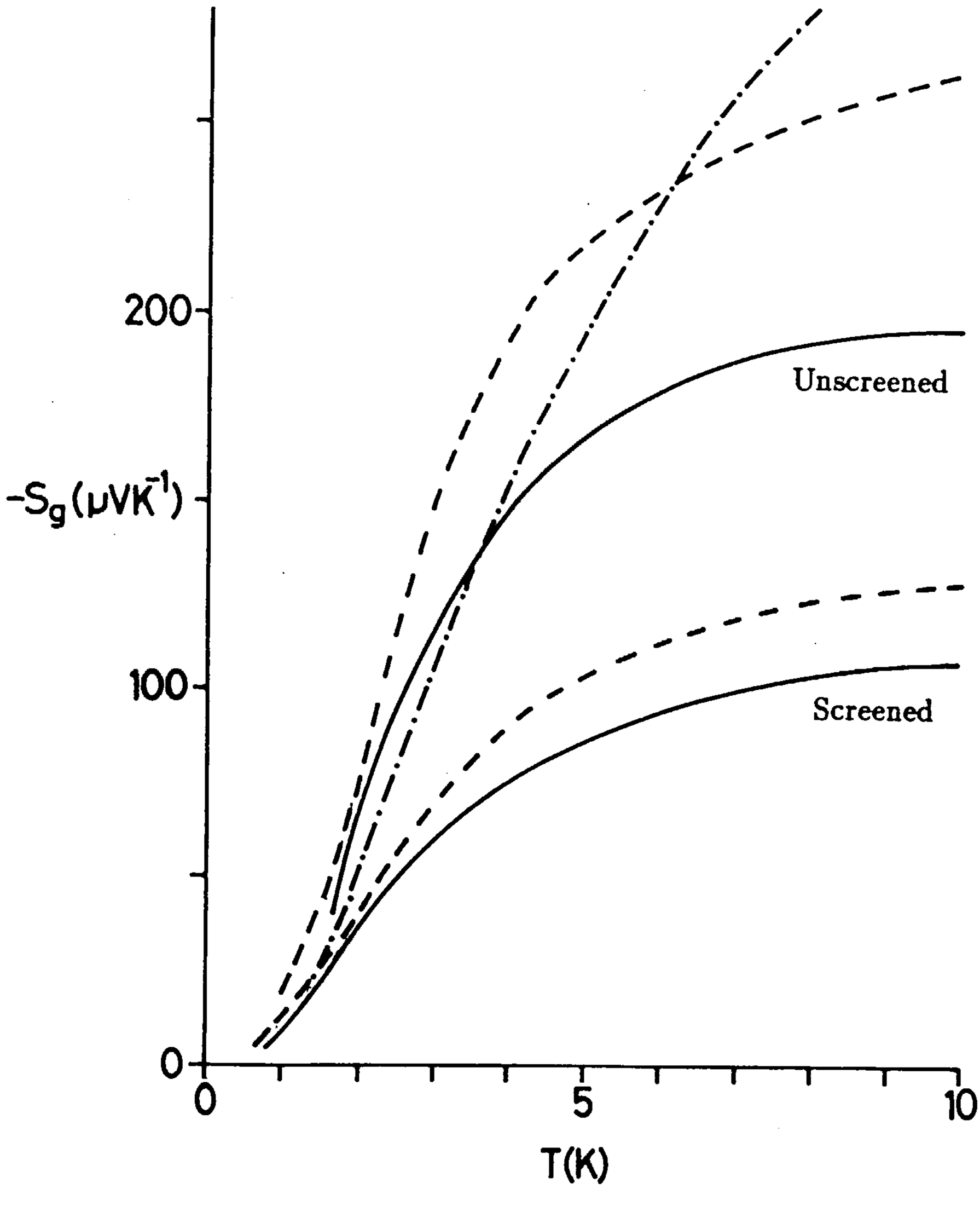
Whilst screening of the electron-phonon interaction has been shown to have a very significant effect upon phonon-drag, then, and has gone some way to explain the experimental data there is still much to explain. The GaAs case in particular appears much less well understood now than before. In Si the peaks in $-S_g/T^3$ are shifted and are sharper than is found experimentally, particularly at lower n. In the following chapter these points are reconsidered in the light of more data, better parameter values and further improvements to the theory.

Figure 4.2: Screened results for $-S_g/T^3$ in Si.



Plots of $-S_g/T^3$ against T for $n = 9.8, 7.8$ and $6.1, \times 10^{15}\mathrm{m}^{-2}$ compared with the experimental data points of Gallagher et al (1987),○,□ and △ respectively. A solid line denotes a calculation using the variational envelope and a broken line the ISW of equivalent width.

Figure 4.3: The effect of screening upon S_g in GaAs.



Plots of $-S_g$ against T for $n = 1.78 \times 10^{15}\mathrm{m}^{-2}$ compared with the experimental data (chain curve) of Fletcher et al (1986). A solid curve denotes a calculation using the variational envelope and a broken line the ISW of equivalent width.

Chapter 5

Further investigations.

5.1 Introduction to the Chapter.

In any pioneering calculation it is helpful to make all reasonable approximations which simplify the task in order to obtain the first results. When the calculations appear possible, and the initial results show some of the expected behaviour, it is then worth considering the formalism more carefully. Thus the inclusion of screening and a better description of the electron confinement has greatly improved the description of phonon-drag in quasi-2D. It is now interesting and worthwhile to consider the calculations and approximations in more detail.

In the GaAs case, for example, it is apparent that the calculated values are a large underestimate. An additional mechanism of electron-phonon coupling is therefore considered in the next section. The effect of non-degeneracy and of the temperature dependence in the screening are also considered in this chapter. The effect of making some further approximations such as elastic scattering is then discussed. The dominant wavevector $\bar{q}$ is introduced to explain the occurrence of peaks in $-S_g/T^3$ and helps, in addition, to interpret other influences upon the phonon-drag. The effect of the energy dependence of the electron relaxation time is considered and the comparison with experiment is investigated in detail.

5.2 Piezoelectric scattering.

In the preceding calculations it has been assumed that the deformation potential is the only mechanism whereby acoustic phonons scatter electrons. It arises through the effect of the local distortion of the lattice upon the crystal potential and is expressed in equation (2.29) in terms of the deformation potential coupling constant E_1 (for some discussion, see for example, Ziman 1963). Lyo (1988) has pointed out that piezoelectric scattering also contributes significantly to S_g in GaAs and this is confirmed by the author (Smith and Butcher 1989b) and by Karl et al (1988). The piezoelectric effect is exhibited by materials which lack inversion symmetry. The application of a strain generates an additional polarization **P** (see, for example, Zawadzki 1982) with components;

$$P_i = \sum_{ijk} p_{ijk}\, \varepsilon_{jk}. \tag{5.1}$$

Here ε_{jk} is the strain $(\partial u_j/\partial x_k + \partial u_k/\partial x_j)/2$, with u_j the j^{th} element of the lattice displacement vector in (2.30) and x_j the corresponding spatial coordinate, and p_{ijk} is an element of the piezoelectric tensor. The scattering rate corresponding to (2.31) can then be obtained by replacing E_1^2 by $e^2 \mid G(\mathbf{Q}) \mid^2 /Q^2$, where:

$$G(\mathbf{Q}) = \frac{1}{\kappa} \sum_{ijk} \frac{Q_j Q_k}{Q^2}\, \xi_j\, p_{ijk}. \tag{5.2}$$

Here ξ_j is the j^{th} component of the polarization vector, with wavevector $\mathbf{Q}$ and component Q_j, and κ is the permittivity. The mode index is left as understood and hence $\mid G(\mathbf{Q}) \mid^2$ is required for each mode. Whereas for the deformation potential alone, only the longitudinal mode couples to electrons in GaAs, there are now four contributions to the total scattering rate. However, translational and rotational symmetry (Zawadzki 1982) requires that $p_{123} = p_{132} = p_{213} = p_{231} = p_{312} = p_{321} \equiv h_{14}$ (the piezoelectric potential coupling constant, analogous to E_1) and all other $p_{ijk} = 0$. Inversion symmetry in materials like Si requires that h_{14} is also zero. Hence there is a piezoelectric scattering contribution to S_g in GaAs, but not in Si. Using the above simplifications the scattering rate is obtained by

replacing E_1^2 with $(h_{14}e/\kappa)^2 A(\mathbf{Q})$, where:

$$A(\mathbf{Q}) = 4\left\{\xi_1 Q_2 Q_3 + Q_1 \xi_2 Q_3 + Q_1 Q_2 \xi_3\right\}^2 / Q^4, \tag{5.3}$$

and is calculated separately for the longitudinal (L) and transverse (T) phonon modes. For the L mode ξ is simply $\mathbf{Q}/Q$ and on averaging over q_x and q_y (where $Q_1, Q_2, Q_3 = q_x, q_y, q_z$) to reflect the symmetry of the 2DEG plane:

$$A_L(\mathbf{Q}) = 9q_z^2 q^4 / 2Q^6 \tag{5.4}$$

For the T modes a sensible choice is to take one mode to have ξ in the 2DEG plane. The result for the sum of the two T modes is then:

$$A_T(\mathbf{Q}) = (8q_z^2 q^4 + q^6)/2Q^6, \tag{5.5}$$

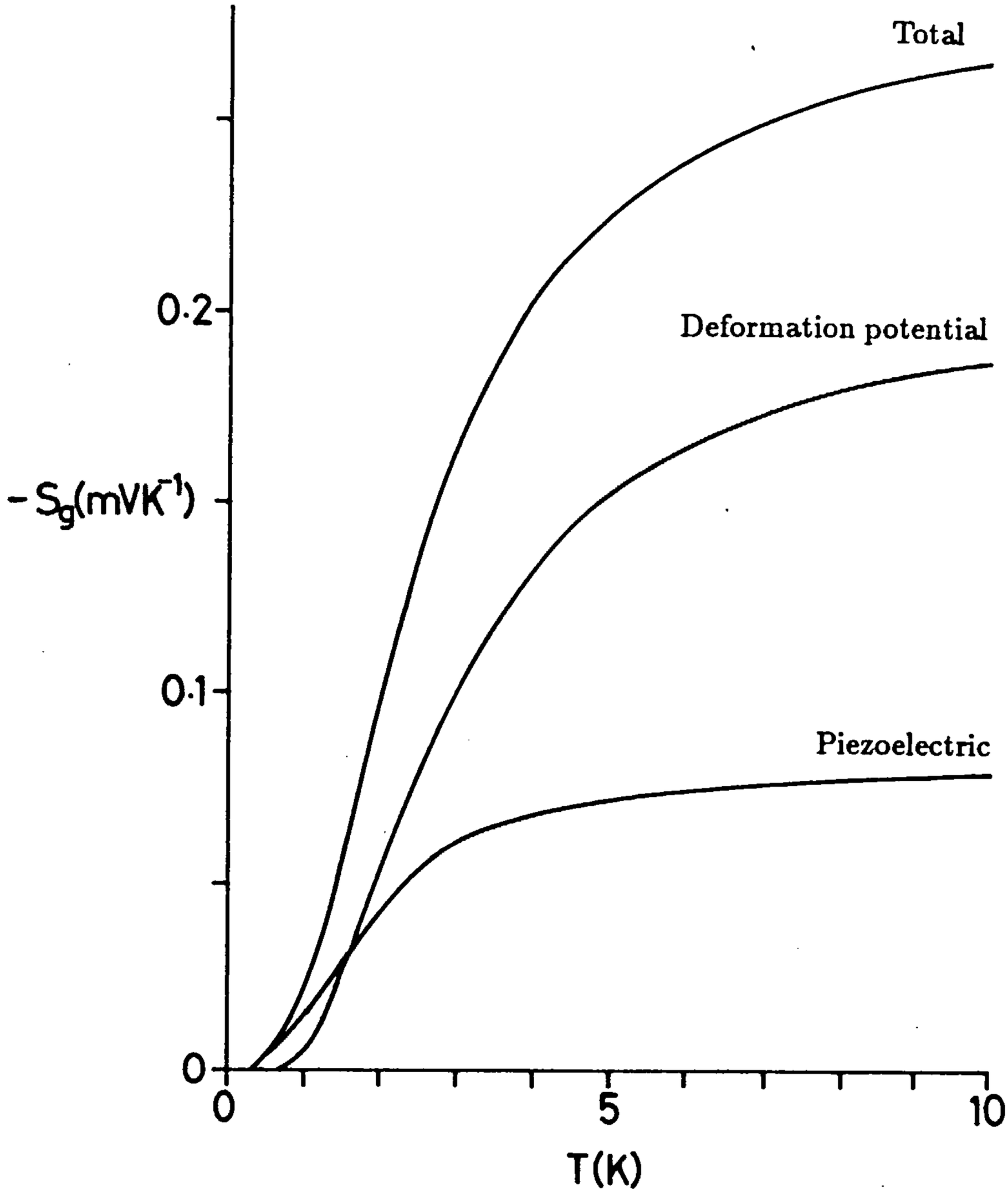
These last two expressions agree with those of Price (1981). The mode types are treated separately because they have different sound speeds (see Table 2.1).

Results of screened calculations of S_g both with and without accounting for this scattering mechanism are illustrated in Figure 5.1 for the parameter values of Lyo (1988) (with $n = 1.78 \times 10^{15}\mathrm{m}^{-2}$). For T <1K the piezoelectric contribution is clearly dominant and for T=5-10K it constitutes about 30% of the total. This goes some way towards explaining the underestimate of the measured values which is apparent in Figure 4.3. The relative importance of the different mechanisms is, however, very sensitive to the magnitude of the specimen parameters such as E_1, as discussed in Section 5.5.

5.3 Further approximations.

In his calculations Lyo (1988) also makes some approximations which make the final evaluation more convenient and easier to interpret. However, on close examination the author finds that important discrepancies are thereby introduced. A "Comment" was submitted to "The Physical Review", where the results are reported, to this effect but was withdrawn when Lyo proposed to publish an "Errata" with an acknowledgement. Some discussion

Figure 5.1: The effect of piezoelectric scattering on the thermopower in GaAs.



Plot of $-S_g$ against T for the data of Lyo (1988) to show the contribution to the total 'drag thermopower from the piezoelectric scattering mechanism ($n = 1.78 \times 10^{15}\mathrm{m}^{-2}$).

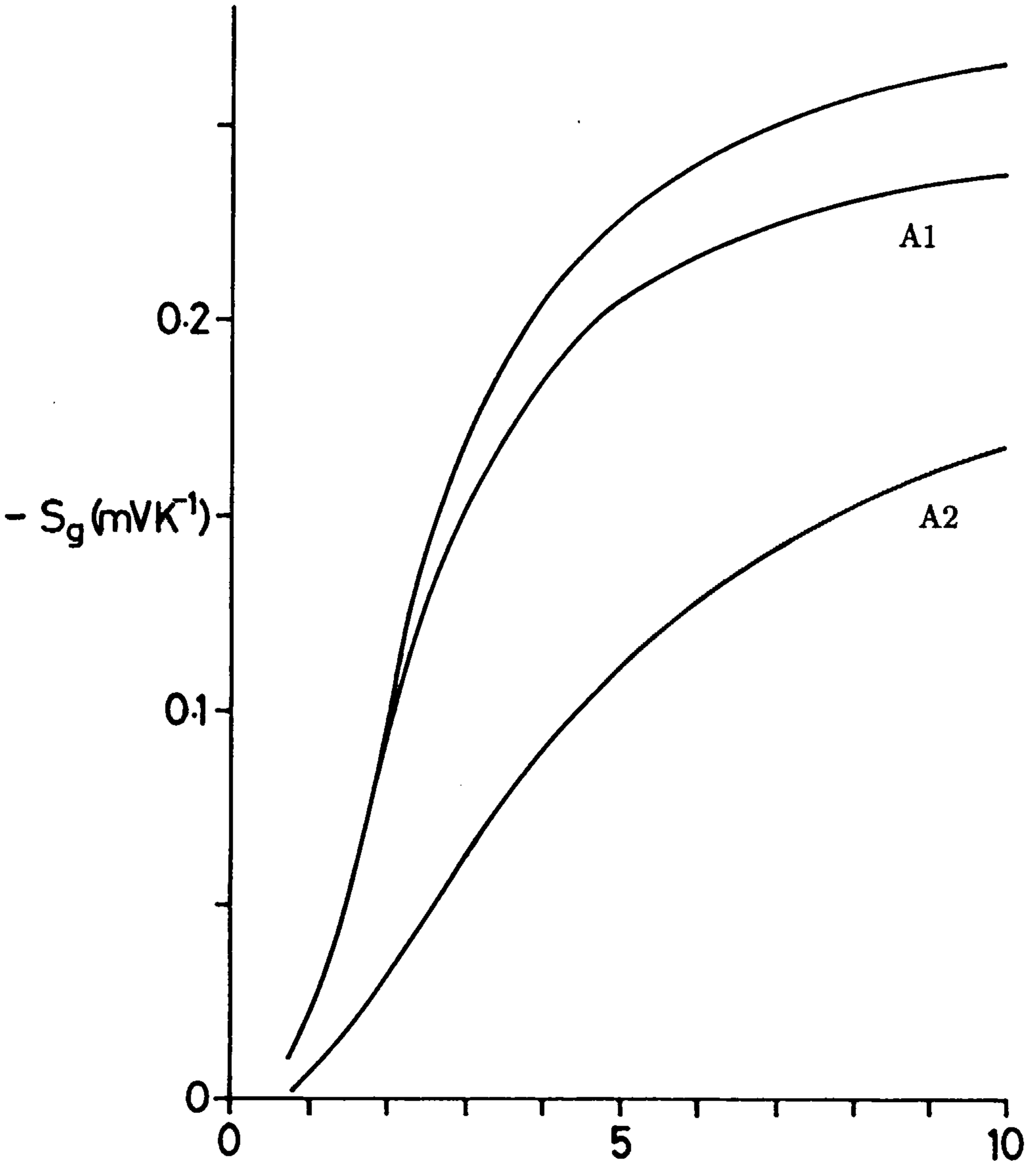
of these discrepancies (and of the importance of the other extensions to the calculations which follow in the remaining sections) has been published by the author (Smith and Butcher 1989b). Most significant is an additional approximation for the product of state occupancy factors (see equations (2.69) and (2.71)). At very low temperatures the energy and momentum conservation conditions discussed in Section 2.6 prevent all but the smallest wavevector phonons from contributing to S_g. The approximation of elastic scattering then becomes valid because the electron energy at ε_f changes little when it absorbs a phonon of energy $\hbar\omega_{\mathbf{Q}}$, as $\hbar\omega_{\mathbf{Q}} \ll \varepsilon_f$. Thence $\hbar\omega_{\mathbf{Q}}$ can be dropped from the Dirac delta function in the expression for the transition rate (see equation 2.31). Lyo goes further and takes $\hbar\omega_{\mathbf{Q}} \ll K_B T$ to simplify the prefactor $W(\mathbf{Q})$ of the delta function in (2.71), to $K_B T$. However, even at 10K, $\hbar\omega_{\mathbf{Q}}/K_B T$ at $Q = 2k_f$ is about 0.8 and the effect of the elastic approximation is to underestimate S_g by about 50% as illustrated in Figure 5.2. This goes unnoticed in Lyo's final results, however, because an unnecessary factor of two "for spin" is introduced in error. Moreover, the correction varies with the temperature through the changing phonon (wavevector) population and thus the temperature dependence of S_g is also a little misrepresented.

The elastic approximation itself, is not so seriously in error but $\hbar\omega_{\mathbf{Q}}/\varepsilon_f$ is still about 0.1 for the same data and leads to an underestimate of about 10% as shown in the Figure. The advantage of the approximation is that the necessary integrals are greatly simplified. For example, the expression (2.72) for the unscreened longitudinal acoustic phonon deformation potential contribution to S_g becomes:

$$S_g = \frac{-Lm^2 v_s E_1^2}{2(2\pi)^3 K_B T^2 n\, e\, \rho\, \hbar\, k_f} \int_0^{\infty} \int_0^{2k_f} \frac{q^2 Q^2 \mid Z_{11}(q_z) \mid^2}{\sinh^2(\gamma/2)\sqrt{1-(q/2k_f)^2}}\, dq\, dq_z \qquad (5.6)$$

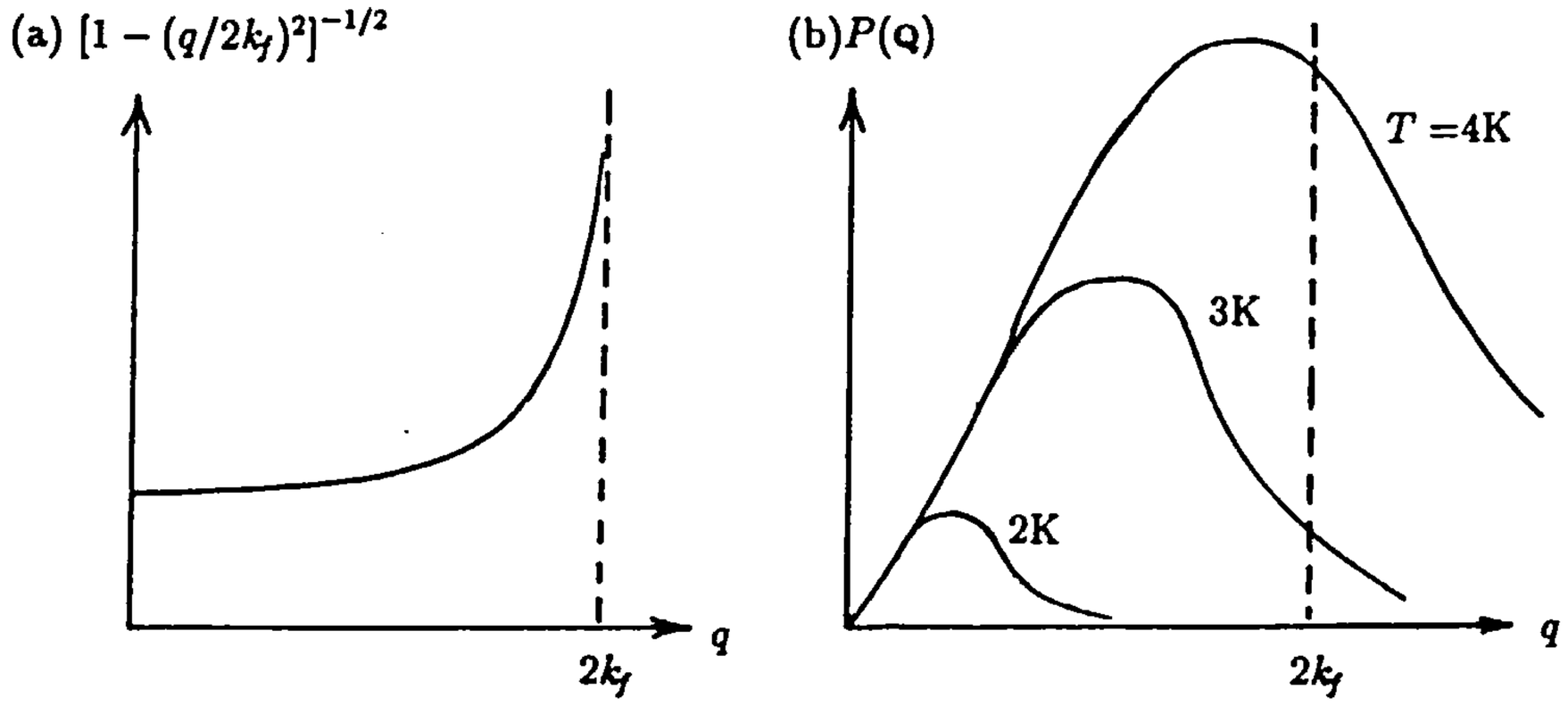
The field of integration of Figure 2.1 is now defined more simply and $\alpha(\mathbf{Q})$ is replaced by $(q/2k_f)$ in the denominator. By absorbing a phonon the electrons now only change their momentum. The momentum transferred, and the scattering rate, rise as q approaches $2k_f$ where the integrand diverges. Contributions to the integrand from around $q = 2k_f$ are most significant then. This is discussed further in the next section. The divergence in the inelastic case is shifted from $2k_f$ and is that which is broadened out by Cantrell and

Figure 5.2: The effect of further approximations.



Plot of $-S_g$ against T corresponding to Figure 5.1 to show the effect of dropping $\hbar\omega_{\mathbf{Q}}$ from the delta function (A1) and assuming $\hbar\omega_{\mathbf{Q}} \ll K_B T$ in the prefactor $W(\mathbf{Q})$, (A2).

Figure 5.3: Factors in the simplified S_g integrand.



The resultant integrand can be considered as composed of two factors which depend differently upon q.

Butcher in II.

For the Si case the aforementioned approximations can be more severe. At $Q = 2k_f$ the quantities $\hbar\omega_{\mathbf{Q}}/K_BT$ and $\hbar\omega_{\mathbf{Q}}/\epsilon_f$ are given by $2\hbar v_s(2\pi n/g_v)^{1/2}/K_BT$ and $4m(g_v/2\pi n)^{1/2}$. For the same n and v_s, then, $\hbar\omega_{\mathbf{Q}}/K_BT$ is smaller by 0.7, and $\hbar\omega_{\mathbf{Q}}/\epsilon_f$ is bigger by about 0.7 times the ratio of the effective masses 0.2/0.07; ie a factor of 2. Thus inelasticity is more important for given n in Si, and, although $\hbar\omega_{\mathbf{Q}}/K_BT$ is smaller, n can be larger and hence both approximations are better left unmade.

5.4 The "dominant" phonon wavevector.

Interpretation of S_g results is made easier by considering the "dominant" value $\bar{q}$ of the 2D wavevector component q at a given T. Consider the simplest unscreened degenerate case of the phonon-drag of electrons by longitudinal phonons via deformation potential coupling which is described by (5.6). As described for the inelastic case in II, the integrand can be separated into two factors but the simpler elastic case is better suited to interpretation. The two factors have the approximate form illustrated in Figure 5.3. The first, $[1-(q/2k_f)^2]^{-1/2}$, arises from the scattering rate, depends upon n and provides the wavevector cut-off at $2k_f$. The second, $P(\mathbf{Q})$ say, depends only upon the phonon wavevector and

distribution at given T. Consider their product at given n. At very low T the peak in $P(\mathbf{Q})$ lies well below $2k_f$ and the cut-off factor goes unnoticed because its value is not much greater than unity. At higher T the peak in $P(\mathbf{Q})$ moves to larger q and the product is enhanced much more by the other factor. The initial behaviour of S_g can therefore be expected as an increasingly strong rise in temperature. When the peak in $P(\mathbf{Q})$ moves to above $2k_f$, however, only that part below $2k_f$ contributes to S_g, which consequently rises less quickly. The position at which the dominant value of q in $P(\mathbf{Q})$ is around $2k_f$ can therefore be expected to provide the greatest enhancement of S_g and may be assumed responsible for the peak in $-S_g/T^3$ noted in Si.

To verify this assumption $P(\mathbf{Q})$ can be considered as a weighting function and used to calculate a "dominant", or representative value, of q at a given T by taking:

$$\bar{q} = \frac{\int_0^\infty qP(q)\,dq}{\int_0^\infty P(q)\,dq} = \frac{2K_BT}{\hbar v_s}.\frac{\int_0^\infty x^5 \sinh^2 x\,dx}{\int_0^\infty x^4 \sinh^2 x\,dx}. \tag{5.7}$$

For convenience q_z is taken as zero here, so that $Q = q$ and $\mid Z_{11}(q_z) \mid^2 = 1$. Referring to standard integral tables (Gradshteyn and Ryzhik 1980) the value of $\bar{q}$ is $5K_BT/\hbar v_s$ (to within about 4%) and is in agreement with the value used by Gusev et al (1984). Screening increases $\bar{q}$ above this value by reducing the contribution to $P(\mathbf{Q})$ from small q. Thus:

$$\bar{q} \geq 5K_BT/\hbar v_s \tag{5.8}$$

The equality is approached at low n, where screening is less important, and at higher T, where the contribution to $P(q)$ from larger q is more dominant. The location, $T = T_p$, of peaks in $-S_g/T^3$ is then determined by the condition $\bar{q} = 2k_f$ and, assuming the equality, values of T_p can be compared with the measured location of the peak T_M. Example results are illustrated in Table 5.1 for more complete MOSFET data (Gallagher et al 1988) than given previously (Gallagher et al 1987), for a range of n. The close agreement between T_p and T_M appears to confirm the given explanation of the peaks in $-S_g/T^3$ values in Si. Peaks would also be predicted by the above for the GaAs results but the interpretation is complicated here by the presence of piezoelectric scattering. This could be explored further.

$n/10^{15}\mathrm{m}^{-2}$	$T_p(K)$	$T_M(K)$
12.7	4.5	4.2
6.9	3.3	3.3
4.0	2.5	2.6

Table 5.1: Positions of the peak in $-S_g/T^3$ in a Si MOSFET.

The quantity $\bar{q}$ is affected by changes to S_g which alter the **Q** dependence of the S_g integrand, such as the approximations already discussed. Inelasticity allows phonons with $q > 2k_f$ to drag electrons by shifting the cut-off to a higher q. At sufficient T, then, inelasticity becomes important as $\bar{q}$ exceeds $2k_f$ and the contribution from larger wavevector phonons is weighted by an increasing phonon population. The background value of T thence determines the influence of inelasticity through the value of $\bar{q}$. This observation can be compared with the situation arising in ballistic phonon absorption and emission experiments such as the phonon-drag imaging experiment of Karl et al (1988), (see also Hensel et al 1983, Rothenfusser 1986 and Kent et al 1988) in which there is no temperature gradient, and hence no net phonon momentum flux, prior to phonon injection. If the phonon momentum pulse is characterized by a representative q, then inelastic effects are small if $q \ll 2k_f$ and the situation is analogous to an S_g measurement at very low T ($\ll$ 1K, say). Furthermore, the observation of the dominance of the piezoelectric mechanism by Karl et al, using phonon frequencies of about 120GHz and $n = 6 \times 10^{15}\mathrm{m}^{-2}$ is to be expected as $\bar{q}/2k_f \approx 4 \times 10^{-4}$.

The latter experiment is of particular interest and relevence here. Phonon-drag is produced by a phonon pulse generated at the back of a GaAs/GaAlAs heterojunction specimen using a laser targeted upon an Al coating. With the specimen held at 1.2K the phonon energies are characteristic of the superconducting gap of Al (T_c=2K). The phonon pulse arrives ballistically at a small region of 2DEG in the form of a bridge between larger contact areas produced by selective etching. The thermoelectric voltage generated be-

tween the contacts is then a sensitive function of the incident phonon flux, the absorption mechanism and the phonon focussing (see Bron 1985). There is no need to account for electron diffusion as the specimen is isothermal. An image of the "drag" is created by shading points representing the crystal trajectory, according to the sign and magnitude of the thermoelectric voltage. Focussing of the phonon group velocities $\mathbf{v}_g$ into certain preferred directions is a consequence of the non-spherical "slowness-surface" in real systems. This is the constant frequency surface in $\mathbf{k}$-space, to which $\mathbf{v}_g$ is perpendicular. Concave regions on the surface naturally channel phonons towards certain directions more than others. Thus focussing is important in interpreting the phonon-drag images which are different for the different absorption mechanisms.

In the calculations of S_g the slowness surface has been assumed spherical. It may be interesting then to investigate whether focussing has an affect upon the phonon-drag thermopower.

5.5 Comparing with experiment.

The comparison between calculated values of S_g and the experimental data is now worth examining more closely because the agreement is much closer than before. A simple example is the need to subtract S_d from the measured S. This contribution has been estimated (see, for example, Nicholas 1985, Fletcher et al 1986, Lyo 1988) by using the Mott formula (see, for example, Blatt 1968) in the form:

$$S_d = \frac{1}{3}(p+1)\frac{\pi^2 K_B}{e}.\frac{K_B T}{\varepsilon_f}. \tag{5.9}$$

This formula gives the $K_B T/\varepsilon_f$ dependence predicted by the simple arguments of Section 1.5 and the magnitude is determined by the energy dependence of the electron relaxation time in the neighbourhood of ε_f, which is assumed to follow:

$$\tau(\varepsilon) = \tau_o\, \varepsilon^p. \tag{5.10}$$

For the heterojunction Lyo (1988) takes the approximate value $p = 1$ suggested by Fletcher et al (1986). The explicit calculation of $\tau(\varepsilon)$ by Kundu et al (1987) supports this estimate

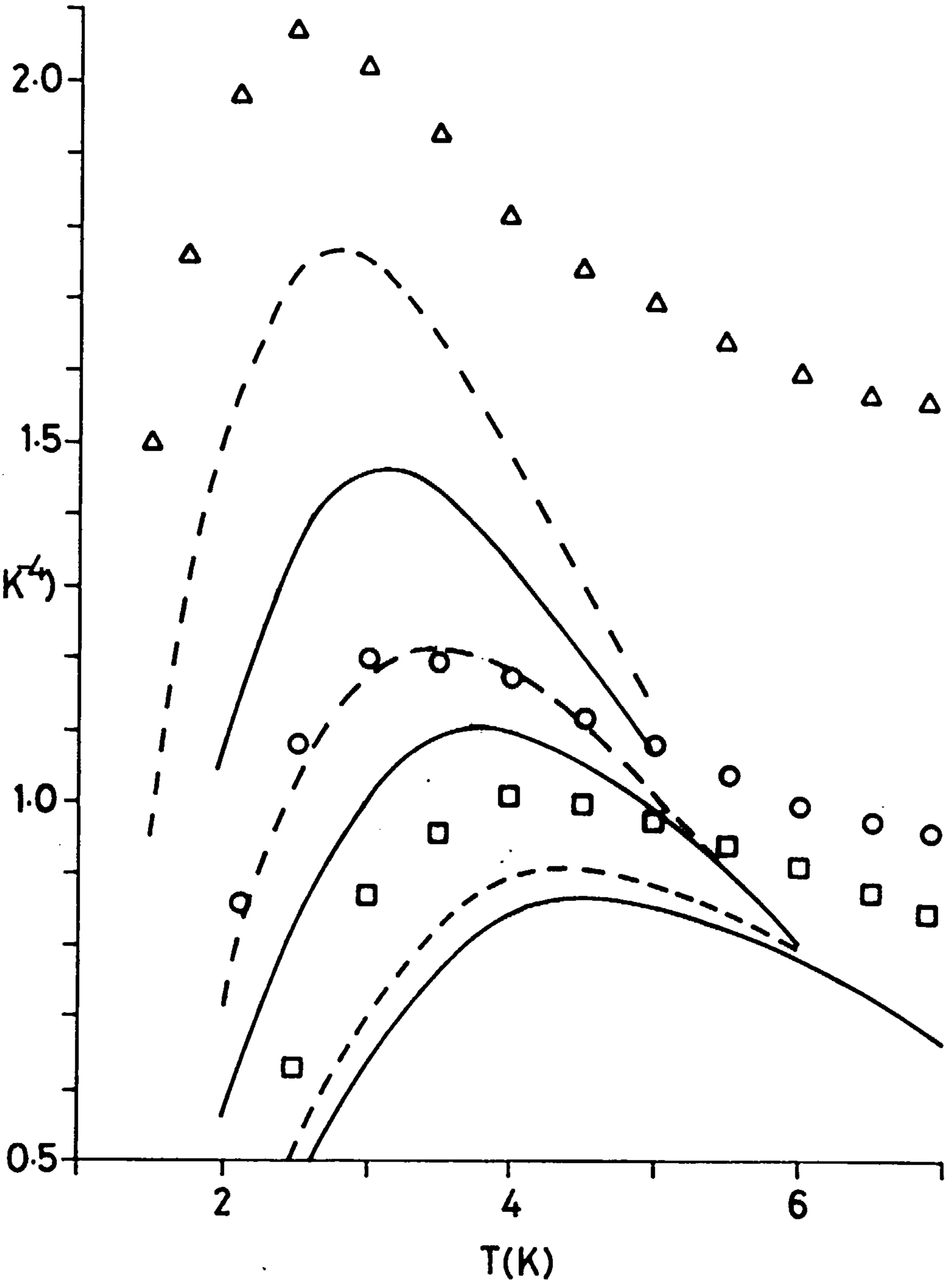
but these authors also remark that a more refined calculation is needed to properly describe $\tau(\varepsilon)$. Gallagher et al (1987) suggests a value of about -1 for their MOSFET data to account for the absence of a linear T dependence in S at the lowest temperature investigated. They suggest that screened ionized impurity scattering may be responsible for such a negative value but the calculations of Stern (1980) show surface roughness scattering to be a likely candidate.

The value of p is clearly important because, with $p = 1$, S_d constitutes about 10-20% or more of the measured S. For the data of Fletcher et al, for example, the value of S_d for $n = 1.78 \times 10^{15}\mathrm{m}^{-2}$ is 24% of the total at 10K. For Si the corresponding contribution is greater due to the smaller Fermi energy resulting from the increased effective mass and the valley degeneracy. However, such estimates are only approximate because the accuracy of (5.9) is limited both by (5.10) and the assumption of pure two-dimensionality. On this basis the accuracy with which S_g can be compared with experiment should not be exaggerated. In the results which follow later in the chapter plots are given for a variety of p values.

For the Si case Gallagher et al also provide plots of S^{-1} against n for a range of T. These plots are linear in n and intercept the n-axis at the same value, $n_{MIT} = 1.2 \times 10^{15}\mathrm{m}^{-2}$, which is interpretted as the transition value between strongly and weakly localized electron states (see, for example, Kramer et al 1985, Nagaoka and Fukyama 1981). Conduction by "free electrons" is considered to occur only when n exceeds this minimum. The free electron density to be used in the MOSFET calculations of S_g, then, is $n_{\mathrm{free}} = n - n_{MIT}$ and is assumed hereafter. The results for S_g are consequently raised and the positions of peaks in $-S_g/T^3$ move to lower T as shown in Figure 5.4. This behaviour follows from the 1/n dependence of S_g in (2.15) and because as the electron density falls the condition $\bar{q} = 2k_f$ is satisfied earlier. The problem of the shift in the peak value observed in Section 4.6 is thereby partially solved.

Lyo takes a different set of parameter values in his calculations than those used in II. Most significant is the value for E_1 taken as 9.3eV (instead of 8.0eV) because the

Figure 5.4: The effect of using n_{free} in the MOSFET calculations.



Plot of $-S_g/T^3$ against T for the data of Gallagher et al (1988) for $n = 14.1$(□),8.35(○), and 4.90(△)$\times 10^{15}\text{m}^{-2}$ compared to the calculated values using n(solid) and n_{free}(chain).

deformation potential contribution to S_g is proportional to E_1^2. The difference in these values represents a 26% increase in E_1^2 and accounts for a large fraction of the difference between theory and experiment. The precise value to be taken is the subject of some debate (see, for example, Manion et al 1987 and references therein, and likewise for Nolte et al 1987). The values quoted range from 7eV (for bulk GaAs) to 16eV. (Manion et al). The accuracy of these measurements are complicated by the degree to which the deformation potential mechanism of scattering is isolated and that to which the comparison can be made with the corresponding theory. Care is necessary, for example, in the account taken for screening, the electron confinement and the presence of other scattering mechanisms. The corresponding values in Si for Ξ_u and Ξ_d appear to be more certain (Ando et al 1982) and in Smith and Butcher (1989a) and the calculations which follow the commonly used values of 9.0 and -6.0eV are adopted. The other parameter values taken are given in Table 2.1. The phonon velocities are obtained from the standard reference (Landolt Bornstein, New Series 1982) and are averaged over the phonon modes of appropriate propagation and displacement directions.

The value taken for the phonon mean free path L deserves consideration because it appears in the S_g expression (2.72) as a prefactor and is assumed constant. The kinetic formula for the thermal conductivity κ is commonly used with the asymptotically exact expression for the low temperature (harmonic) specific heat (see, for example, Ashcroft and Mermin 1981) to yield:

$$L = \frac{15\hbar^3}{2\pi^2 K_B^4} \cdot \frac{c^2 \kappa}{T^3}. \tag{5.11}$$

Here $1/c^3$ is the average inverse cube speed of sound for the three acoustic modes. The value of L is thence obtained from the measurement of $\kappa(T)$, which gives a constant L when $\kappa \propto T^3$. Lyo, for example, uses the value 0.30mm for the data of Fletcher et al, obtained at 3K. However, the values 0.2, 0.2, 0.1 and 0.07mm can also be obtained, for example, at T= 2, 5, 7, and 10K respectively. The T^3 dependence is clearly, not followed and the value of L to be used in the calculations is correspondingly uncertain. A similar problem arises with the data of Fletcher et al (1988a) but Ruf et al (1988) find a much

closer T^3 dependence and a more consistent value of L, results (0.12mm). For other cases it remains uncertain whether L can be assumed to be constant. For the MOSFET data a value constant over the experimental range is recorded to within about 10% (Gallagher, private communication).

5.6 Energy dependence of $\tau(\varepsilon(\mathbf{k}))$.

In addition to determining the value of S_d, $\tau(\varepsilon)$ can also be seen to influence S_g. Consider the factor arising in the general S_g expression of section 2.4 for the difference of $\tau(\mathbf{k})\mathbf{v}(\mathbf{k})$ for the electron states $\mathbf{k}$ and $\mathbf{k}+\mathbf{q}$, in the quantum limit. This factor has been evaluated previously (Smith and Butcher 1989a, Lyo 1988) according to the approximation discussed in Section 2.6, whereby τ is assumed to be a function of $\varepsilon(\mathbf{k})$ alone and is evaluated at ε_f due to the peak in $f(\varepsilon(\mathbf{k}))[1-f(\varepsilon(\mathbf{k}+\mathbf{q}))]$. Using the conservation conditions, then,:

$$\tau(\mathbf{k}+\mathbf{q})\mathbf{v}(\mathbf{k}+\mathbf{q})-\tau(\mathbf{k})\mathbf{v}(\mathbf{k})=\tau(\varepsilon_f+\hbar\omega_{\mathbf{Q}})\mathbf{v}(\mathbf{k}+\mathbf{q})-\tau(\varepsilon_f)\mathbf{v}(\mathbf{k}) \tag{5.12}$$

and is simplified further (as in II) to $-\hbar\mathbf{q}\tau(\varepsilon_f)/m$. The latter is exact when $\tau(\varepsilon(\mathbf{k}))$ is constant and for the elastic case in which $\varepsilon(\mathbf{k}+\mathbf{q})=\varepsilon(\mathbf{k})$. It is a good approximation when $\tau(\varepsilon)$ varies slowly near ε_f and when $\hbar\omega_{\mathbf{Q}} \ll \varepsilon_f$. However, for the heterojunction case already discussed $\hbar\omega_{\mathbf{Q}}/\varepsilon_f$ is about 0.1 at $Q=2k_f$ and with p of (5.10) non-zero, it is unclear whether the approximation is valid.

Using (5.10) to expand $\tau(\varepsilon_f+\hbar\omega_{\mathbf{Q}})$ linearly about $\tau(\varepsilon_f)$, however, the correction factor:

$$\lambda(\mathbf{Q})=1+\frac{1}{2}\,p\,\frac{\hbar\omega_{\mathbf{Q}}}{\varepsilon_f}\left[1+\left(\frac{k_f}{q}\right)^2\frac{\hbar\omega_{\mathbf{Q}}}{\varepsilon_f}\right] \tag{5.13}$$

is obtained, which should be introduced under the integral sign in the expression for S_g. The factor is unity for the elastic case, when $p=0$ or when $\hbar\omega_{\mathbf{Q}} \ll \varepsilon_f$; but, for $n=2\times10^{15}\mathrm{m}^{-2}$ in GaAs the correction at $q=2k_f$ and $q_z=0$ amounts to 5.5% for $p=1$ and 11% for $p=2$. The corresponding corrections in Si are 23% and 46%. They are larger due to the smaller value of ε_f for the same n (already discussed). The temperature dependence of S_g is affected as $\mathbf{q}$ and the mean value of $\hbar\omega_{\mathbf{Q}}$ increase with T. A negative

value of p, however, due to surface roughness scattering in Si, for example, gives a reduction rather than an increase in S_g. The effect of taking different values for p is illustrated in Figures 5.6 to 5.8.

5.7 Temperature dependent screening.

Temperature dependence enters screening in the RPA through the polarizability $\Pi(\mathbf{Q})$. This has been evaluated at $T = 0$ (Smith and Butcher 1989a, Lyo 1988) by taking the zero temperature value of the Fermi function $f_0(\varepsilon(\mathbf{k}), \varepsilon_f)$ and obtaining $\Pi_0(\mathbf{q})$, say, which is the result of Stern (1967). At higher temperatures the procedure of Maldague (1978) can be adopted in which it is noted that the finite T value of the Fermi function can be written as:

$$f(\varepsilon, \varepsilon_f) = \frac{1}{4K_BT} \int_0^\infty \frac{f_0(\varepsilon, \mu)}{\cosh^2(\varepsilon_f - \mu)/2K_BT} d\mu. \tag{5.14}$$

Using this expression explicitly in (4.40) for $\Pi(\mathbf{Q})$ and transforming the sum over wavevectors $\mathbf{k}$ to an integral, but performing the integrals over $\mathbf{k}$ and μ (in the above) in reverse order,

$$\Pi(\mathbf{q}) \equiv \Pi(\mathbf{q}, \varepsilon_f, T) = \frac{1}{4K_BT} \int_0^\infty \frac{\Pi(\mathbf{q}, \mu, 0)}{\cosh^2(\varepsilon_f - \mu)/2K_BT} d\mu. \tag{5.15}$$

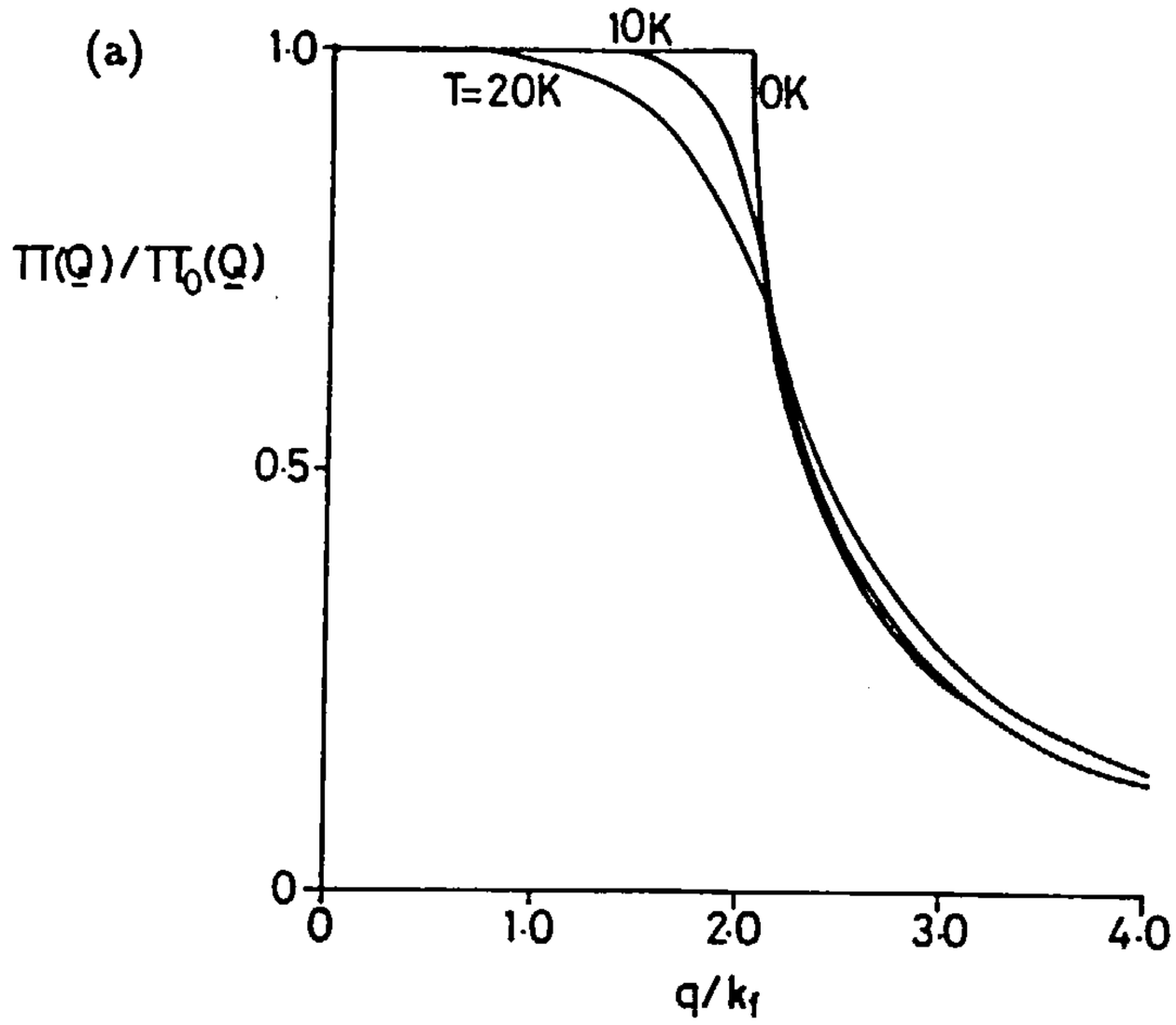
Hence $\xi_0(q)$ in (4.46) and (4.47), for the value of $\xi(q)$ at $T = 0$, can be replaced by:

$$\xi(q) = \frac{1}{e^{-\varepsilon_f/K_BT} + 1} - a \int_0^1 \frac{x^2}{\cosh^2 a(x^2 - b)} dx, \tag{5.16}$$

in which $a = \hbar^2 q^2/16mK_BT$ and $b = 1 - (2k_f/q)^2$. An example result is illustrated in Figure 5.5a (for $n = 2 \times 10^{15}\mathrm{m}^{-2}$ in GaAs).

The effect on the polarizability at 10K is a reduction of about 20% at $q = 2k_f$. This might be expected to have a correspondingly large effect upon S_g but $\varepsilon^{-2}(q)$ at $2k_f$ increases by just 6% and S_g by 3%. This follows because $\varepsilon^{-2}(q)$ depends more weakly on the polarizability and because a range of wavevectors q contribute to S_g. Furthermore the value $q = 2k_f$ is the value for which the effect of finite temperature is greatest. Away from $2k_f$ the change in the polarizability is much less. For $q > 2k_f$ the polarizability is slightly increased whereas below it is decreased, thus the resultant effect upon S_g depends upon

Figure 5.5: Comparison of the 2D polarizability at finite T in GaAs and Si.



Plots of $\Pi(\mathbf{q})/\Pi_o(\mathbf{q})$ against q for $n = 2 \times 10^{15}\mathrm{m}^{-2}$ in (a)GaAs and (b)Si for a range of temperatures T.

$\bar{q}$. Hence, for low T, $\bar{q} < 2k_f$ and K_BT/ε_f is reduced so that $\Pi(\mathbf{q})$ approaches $\Pi_0(\mathbf{q})$. Hence the 3% change seen at 10K is doubly reduced. In the reported results of Fletcher et al (1988) and Ruf et al (1988a) the value of ε_f is larger due to higher n. Consequently K_BT/ε_f is reduced and the effect of finite T on the screening even smaller.

For Si the large value of K_BT/ε_f for given n and T makes the correction more important as illustrated in Figure 5.5b for the same electron density as in Figure 5.5a. The large difference between the materials is apparent and the corresponding effect on S_g at 10K is a decrease of 39%. This large change results because $\bar{q} > 2k_f$ and here $\Pi(\mathbf{Q})$ is increased considerably whereas in GaAs the effect goes unnoticed. At the higher densities of Gallagher et al (1988) in the range 3.5 to $12.7\times10^{15}\mathrm{m}^{-2}$ however, this non-degeneracy is much less important. The resulting change at $n = 6.9$ and 12.7 ($\times10^{15}\mathrm{m}^{-2}$) is $< 1\%$ but the effect increases quickly with T/n. Hence for $n = 3.5 \times 10^{15}\mathrm{m}^{-2}$ and $T = 7\mathrm{K}$ the change in S_g amounts to 6%. For such low densities temperature dependence in the screening is important but is negligible at higher n and is relatively unimportant in the experimental data.

5.8 Non-degeneracy.

Non-degeneracy also affects the approximation (2.69) for the product of state occupancy factors in the transition rate, in addition to the screening. As K_BT/ε_f increases, the product becomes less like a delta function at ε_f. Hence, to properly account for S_g at higher temperatures or low n the effect of non-degeneracy in the calculations should be examined by relaxing the approximation. The sum over $\mathbf{k}$ in equation (2.67) can no longer be transformed to an integral which can be performed analytically, with the loss of the Dirac delta $\delta(\varepsilon - \varepsilon_f)$. This represents a significant complication. Retaining the explicit form of the Fermi function:

$$\sum_{\mathbf{k}} f\left(\varepsilon(\mathbf{k})\right)\left[1 - f\left(\varepsilon(\mathbf{k}+\mathbf{q})\right)\right]\delta\left(\varepsilon(\mathbf{k}+\mathbf{q}) - \varepsilon(\mathbf{k}) - \hbar\omega_{\mathbf{Q}}\right) \;=\; \frac{A(2m)^{3/2}}{(2\pi)^2\hbar^3 q}\int_0^\infty f(u^2+\gamma)\left[1 - f(u^2+\gamma+\hbar\omega_{\mathbf{Q}})\right]du \qquad (5.17)$$

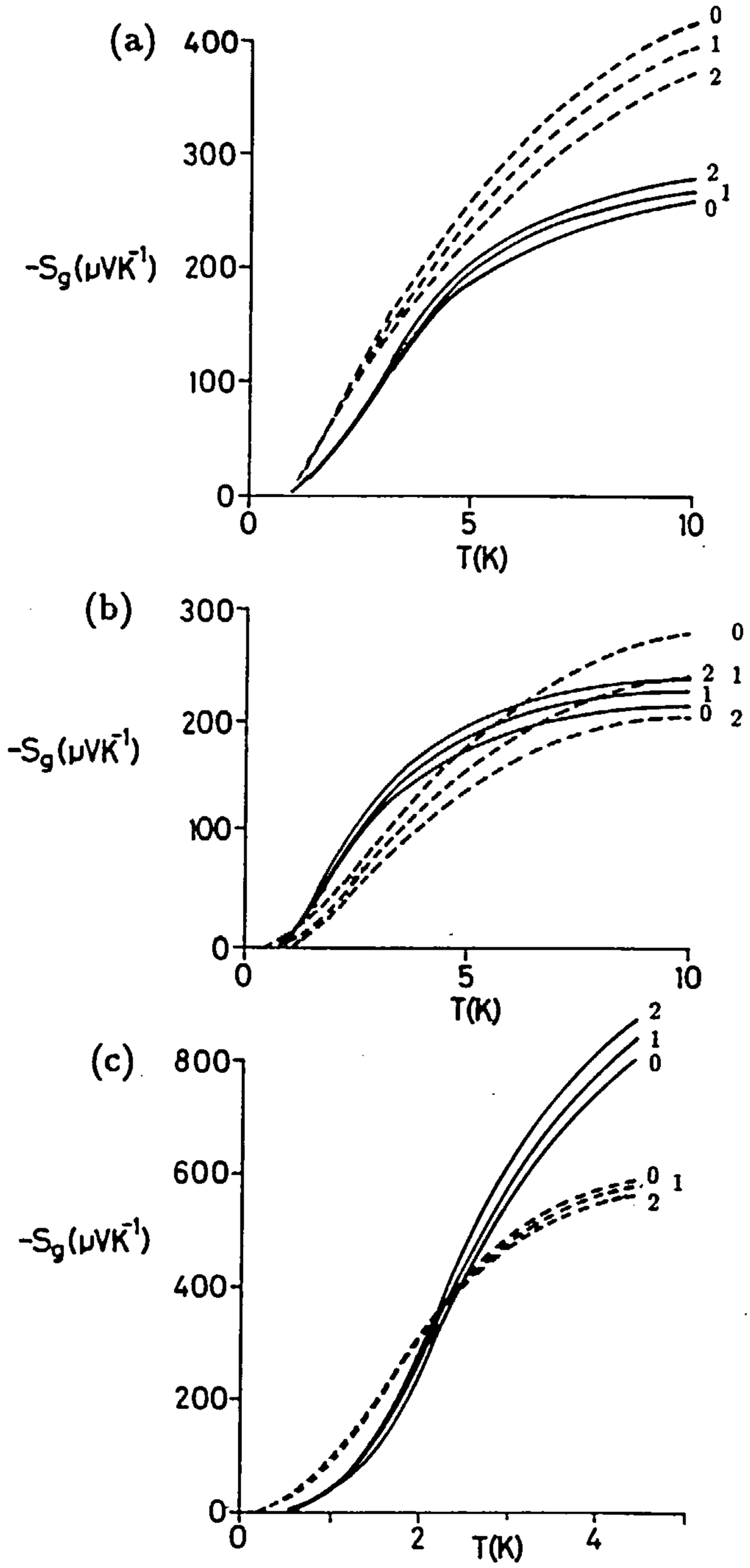
is obtained where $\gamma = (\hbar\omega_{\mathbf{Q}} - \alpha)^2/4\alpha$, with $\alpha = (q/k_f)^2\varepsilon_f$, and is evaluated numerically. The evaluation of S_g in the final formula (2.72) is therefore, in effect, a 3D integral or higher (see (3.16)).

The effect upon S_g values for the data of Fletcher et al (1986, 1988a) and Ruf et al (1988a) for which $n = 1.78 - 3.3 \times 10^{15}\mathrm{m}^{-2}$ is a decrease of up to 8% for the largest value of K_BT/ε_f, at 10K. For the largest n the decrease at 10K is 5.7%. At these temperatures then, non-degeneracy is relatively unimportant in GaAs compared to the uncertainties in E_1 and L. For the MOSFET however, the effect is much larger due to the smaller Fermi energy for given n in Si. At $T = 7$K, for example, $-S_g$ is reduced by between 30 and 40% over $n = 3.5$ to $12.7 \times 10^{15}\mathrm{m}^{-2}$. At $T = 3$K the corresponding reductions are between 14 and 50%. Hence this non-degeneracy effect is a most important consideration in the interpretation of the MOSFET data and its neglect in the preceding calculations represents a large overestimate.

5.9 Discussion of final results.

Calculations of S_g for the three most recent measurements of S in a GaAS/GaAlAs heterojunction already cited, have been performed and compared with the experimental data. The results are illustrated in Figure 5.6. The uncertainties in L, E_1 and $\tau(\varepsilon(\mathbf{k}))$ are avoided by presenting results fitted to the data in the middle of the temperature range by varying L for different values of E_1 (when it is uncertain that L is constant). The corrections previously discussed are included and the parameter values taken as in Table 2.1. In this way the value E_1 =16.0eV is found to be the most favoured because the fit-values of L for the data of Ruf et al (1988a) are then 0.162, 0.149 and 0.138 (mm) for $p = 0$,1 and 2 respectively. These should be compared with the value 0.124mm determined from $\kappa(T)$ and the values obtained when E_1 =9.3eV, which are three times larger. (The constant value $L = 0.124$ mm is used in the Figure). For the data of Fletcher et al (1986) E_1 =9.3eV gives L =0.28 mm. This is very close to the value 0.30 mm used by Lyo (1988). Taking E_1 =16.0eV gives L =0.10 mm which is much smaller but remains within the range

Figure 5.6: Full calculations of S_g for three GaAs/GaAlAs heterojunctions.



Plots of full calculations of $-S_g$ against T (solid) compared to the data (chain) of: (a)Ruf et al (1988a), (b)Fletcher et al (1986) and (c)Fletcher et al (1988a). The data is corrected for S_d and the theory for $\tau(\varepsilon)$ using the values of p as indicated.

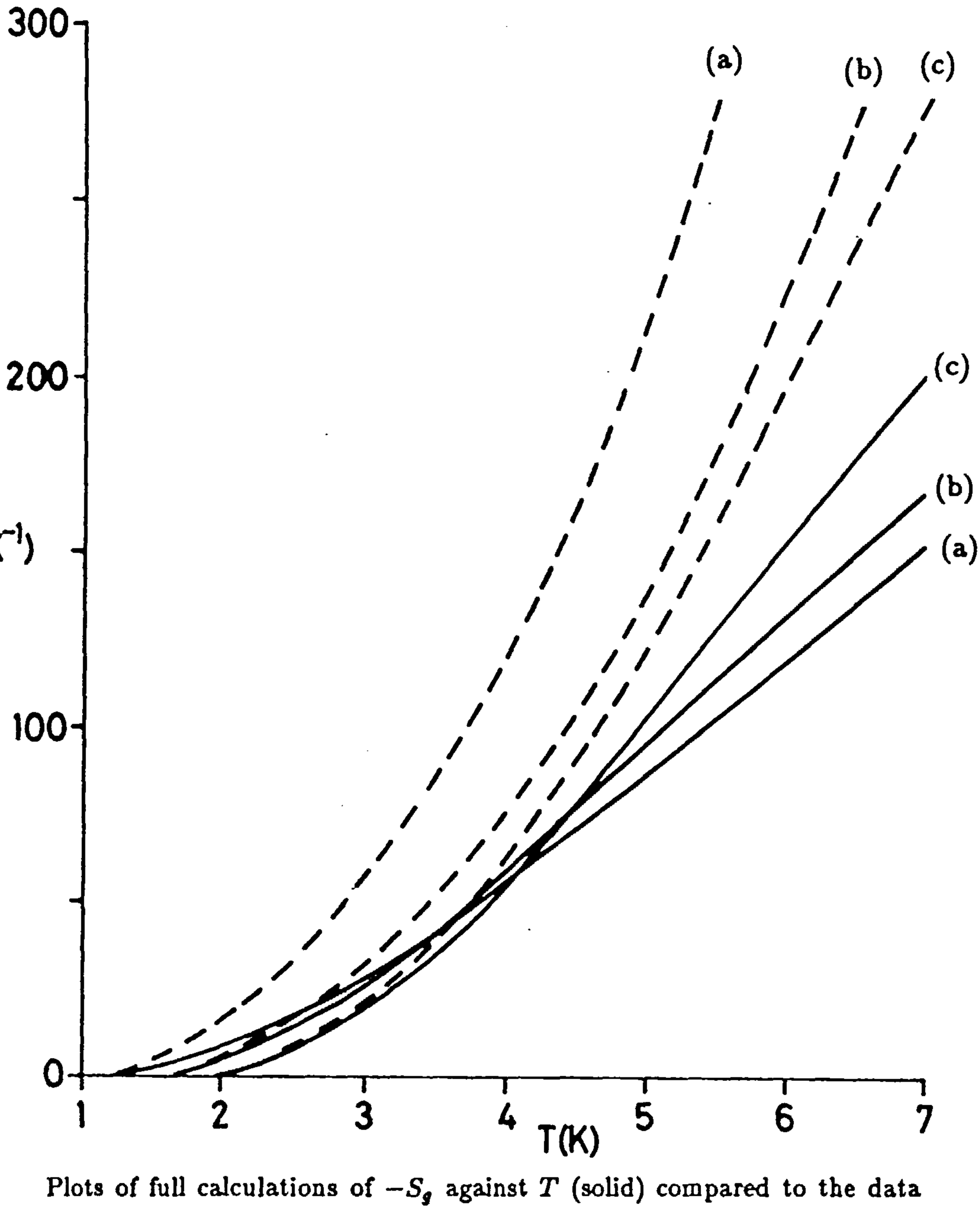
discussed in Section 5.5. Similarly, a fit value of about 0.60 mm is obtained for the data of Fletcher et al (1988a) and should be compared with the value 0.42 mm for L_z of the sample.

The qualitative agreement between the calculated and measured curves is very good. The measured values corrected for S_d, using the corresponding value of p in (5.9), are used in the Figures. For comparison, the same fit value of L (at $p = 0$) is used in the calculations at different p. This shows up clearly the importance of the correction for $\tau(\varepsilon(\mathbf{k}))$ discussed in Section 5.6. Although an exact quantitative comparison is not justified, for the aforementioned reasons, the level of comparison is very close, with the fit values of L taking the expected size. In the case of Ruf et al (1988a) the difference at $p = 2$ between the calculated and "measured" curves is about 20% or less. This difference could easily be accomodated by small changes in p, L and/or E_1. Not much significance should be attached to the exact fit-values or apparent agreement in the other cases, though, because $\kappa\alpha T^3$ is not followed.

The indication, then, is clearly that E_1 should be taken at the high end of the range 7.0-16.0eV for the heterojunction and that the value of p requires further investigation. That a simple scaling of the calculated results for the latter two cases would not fit the calculated values to the data, is to be expected for a value of L which is not constant. A varying value of p would have a similar effect but, as shown in the diagram, is unlikely to make a significant difference as the effect is smaller. Finally, some sensitivity to the value of the effective mass m should be noted. Taking the more approximate value of $0.07m_e$ (used by Lyo) rather than the 0.067 m_e (used in the Figures) leads to an underestimate of S_g by about 7%. This sensitivity is perhaps not surprising because m affects ε_f and $2k_f$ and the value of most of the corrections discussed in addition.

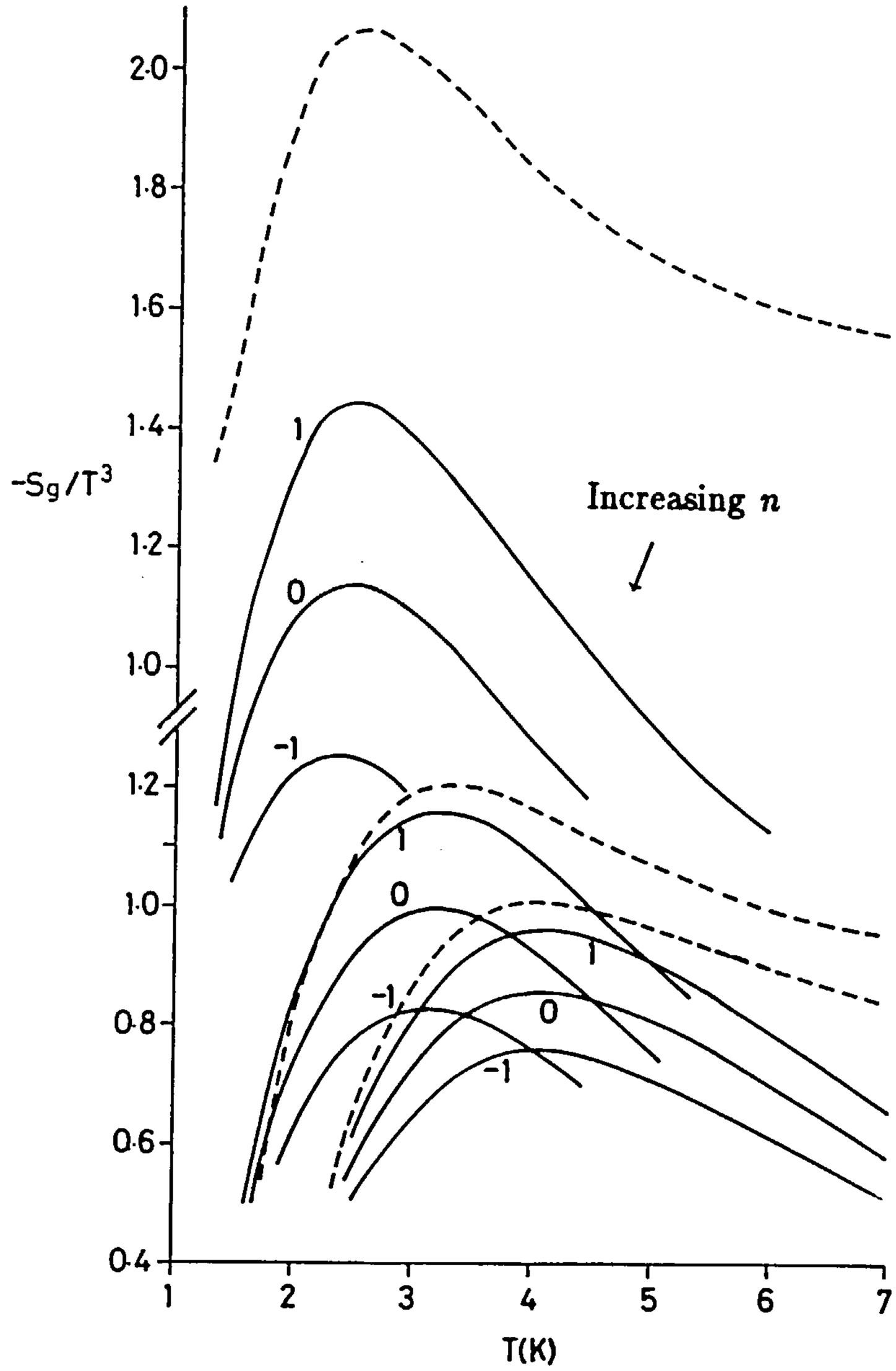
Results for the MOSFET are illustrated in Figures 5.7 and 5.8 for a range of n from the data of Gallagher et al (1988). The Figures show that the agreement with the calculated values is particularly good in the peaks in $-S_g/T^3$ at high n. For simplicity the experimental curves are shown only for the uncorrected S data. It is evident that the value

Figure 5.7: Full calculations of S_g for the MOSFET.



Plots of full calculations of $-S_g$ against T (solid) compared to the data (chain) of Gallagher et al (1988) for n= (a)3.5, (b)6.9 and (c)12.7$\times 10^{15}$m^{-2} without any corrections for S_d or $\tau(\epsilon)$.

Figure 5.8: Full calculations of S_g/T^3 for the MOSFET.



Plots of full calculations of $-S_g/T^3$ against T corresponding to Figure 5.7 with the theory curves corrected for $\tau(\varepsilon)$ using the values of p as indicated.

of p is more important here than for the heterojunction (the non-degeneracy in the results is greater). Although a direct comparison is made difficult without knowledge of p, which may vary (Stern 1980), the Figures show that the calculations are underestimating the thermopower, increasingly, at higher T. Most interesting, perhaps, is the crossover in the $-S_g(T)$ curves at about 4K apparent in Figure 5.7 but which is not seen in the measured values of S. Thus, at T >5K S_g increases with n whereas for T <3K the behaviour is the opposite. This may be an inevitable consequence of the present theory. Consider the simple model using $\bar{q}$ discussed in Section 5.4. At very low T $\bar{q}$ lies below $2k_f$ for all n. For the smallest n however, $\bar{q}$ lies closest to $2k_f$ and thus the rate of increase of $-S_g(T)$ is the greatest. The initial n^{-1} dependence (Section 2.2) may be modified, though, at higher T. Here $q > 2k_f$. For low n, S_g now increases much more slowly. For large n, $\bar{q}$ remains below $2k_f$ for longer (higher T) and hence S_g can continue to rise quickly. This mechanism could, then, allow the curve for high n to cross that for low n as in the Figure.

It should be noted that all similar formulae for S_g will also be separable into two factors like those described in Section 5.4, because $P(q)$ arises from features such as the phonon distribution function whilst the other arises from the conservation conditions in the transition rate. Consequently the crossover behaviour observed may be a true reflection of the behaviour of S_g unless the theory has overlooked some further features. The inclusion of both temperature dependence in the screening and non-degeneracy has resulted in decreases in S_g of about 40%, for low n, compared to the more approximate results. Accounting for both leads to a significant departure from the excellent agreement obtained previously (Smith and Butcher 1989a) but perhaps this should be expected because here it was assumed (Gallagher et al 1987) that the contribution to S (with which the calculations were compared) from S_d was absent or negligible. If there are no additional features to be accounted for, then, a more significant contribution from S_d may provide the explanation. With $p = 1$ in (5.9), for example, S_d at 6.5 K is 432 (240) $\mu V k^{-1}$ for n =3.5(12.7) $\times 10^{15} m^{-2}$ which represents 33(16.5)% of the measured S. Furthermore, such a value of p gives rise to a large correction to the calculated S_g and changes the slope of $-S_g(T)$,

because the correction factor $\lambda(\mathbf{Q})$ is the most different from unity when $K_B T/\varepsilon_f$ (ie $\propto T/n$) is large. Thus for $p = 1$ the crossover is exaggerated but for $p \geq 2$ the crossover begins to unravel because, although the $-S_g(T)$ curves rise for all n, the increase is greater for the lowest n. For $p = 1$, for example, at T =7K, $-S_g$ is increased by 35(13)% for $n = 3.5(12.7) \times 10^{15} m^{-2}$. Clearly then, a close description of the thermopower may only be achieved when S_d and $\tau(\varepsilon(\mathbf{k}))$ are more closely described.

Chapter 6

Conclusions and suggestions.

6.1 Introduction to the chapter.

The objective of the work described here has been to improve the understanding of phonon-drag in quasi-2D systems, and to examine the validity of the existing formalism used to describe it. In achieving these aims some interesting questions have arisen. In this Chapter the final conclusions are presented, some outstanding problems are discussed and some suggestions are made for their solution. Finally some prospects for further developments in the field are indicated.

6.2 Conclusions.

It is clear from Chapters 1 and 2 that the physics of thermopower in LDS shows many interesting features. The sign change as ε_f passes subband minima and the enhanced coupling of the electron and phonon systems in quasi-2D are just two pertinent examples. Important implications for device technology and stimulating new areas of physics research are provided by new structures designed to enhance particular phenomenon in experiments. Interactions between systems of different dimensionality (between quasi-2D electrons and 3D phonons, for example) adds further interest.

Phonon-drag thermopower has been shown here to be sensitive to : the electron-phonon

coupling, screening, m, $\tau(\varepsilon(\mathbf{k}))$ and the degeneracy level. It is a mutual phenomenon which can be considered from either the viewpoint of electrons dragging phonons or the opposite process. In general both processes occur and the equations describing the perturbed distribution functions of the two populations are coupled. Simple models help to shed much light on the physics. The existing "metallic" and "Herring" formulae are not really in conflict but are most helpful in different regimes. They may be understood by introducing the phonon drift velocity and the idea of "saturation". The $1/n$ dependence of S_g observed in a MOSFET, for example, is due to the absence of the saturation effect which occurs in 3D metallic conduction. Here the behaviour $S_g \propto T^3/N_V$ is predicted but for non-degenerate statistics S_g is independent of N_V in 3D. In quasi-2D the underlying behaviour remains as LT^3/n (where L is the phonon mean free path) but is modified according to the rate of momentum transfer between the electron and phonon systems. This is confirmed by the more complete calculations (see, for example, Section 5.4). The fraction of momentum which is exchanged $\alpha(\mathbf{Q})$ (introduced by Guenault (1971)) is consequently of central importance in addition to the quantities L and $\bar{q}$ (Zavaritsky 1984).

The quantity $\hbar\bar{q}$ is a measure of the dominant value of the phonon momentum component $\hbar\mathbf{q}$ which can be transferred to electrons. It ($\bar{q}$) explains features in ballistic phonon absorption experiments in addition to much of the behaviour of S_g. The condition $\bar{q} = 2k_f$, for example, allows more phonons to scatter electrons at a given n and gives rise to the peak in $-S_g/T^3$. Hence the simple temperature dependence $S(T) = aT + bT^3$ is approximately correct but is too naive and neglects some more interesting structure.

The desire to understand S in the extreme quantum limit (large magnetic field (B) and low temperature) is one source of the interest in similar measurements of S at $B = 0$. Such temperatures (around 4K, say) reduce the masking of sharp features by thermal broadening and simplify calculations considerably. In the calculations of S_g in II (for T up to 10K) full advantage is taken of the quantum limit and boundary scattering assumptions. Hence L is taken as constant, the electron distribution as degenerate, higher subbands are ignored and complications like screening are neglected. The assumptions of quasi-2D free electron

states and Debye phonons also prove to be helpful. Many corrections and improvements have been made here and it is clear that the formalism now provides a close description of the physics. Screening, for example, is very important and can be accounted for in the RPA by calculating a dielectric function in the SSA using the same quasi-2D states. MSS effects have been explored and an effective MSS dielectric function has been defined but in the systems of interest MSS effects are not important. For larger n, wider channels or higher T this is no longer true.

Although other influences upon S_g have been explored, screening is the single most important feature which is missing from II. Some of the large overestimate of S_g reported there is now understood to arise from inaccurate numerical integration, but taken with more careful evaluation, the inclusion of screening provides a much closer description of S_g. The overestimate by a factor of 40 is thereby reduced to a difference of tens of percent (Smith and Butcher 1989a). Consequently an investigation in finer detail has been justified. Further simplification of the formulae, such as gained by assuming elastic scattering, does aid understanding and allows a convenient definition of $\bar{q}$ to be made (see section 5.4). However, inaccuracies of 10-20% in S_g are thereby introduced in the experimental regime, which indicates the scale of the importance of inelastic effects.

The dependence of S_g upon the channel width δ has explained the apparent enhancement of S_g in quasi-2D. The increase in $-S_g$ as the 2D limit ($\delta \rightarrow 0$) is approached is a consequence of losing the transverse momentum conservation condition but $-S_g$ still remains limited by the requirement to conserve energy and the 2D momentum component. For wide channels the more general multi-subband formula of I should be applied because the quantum limit is lost. Hence the 3D limit of S_g ($\delta \rightarrow \infty$) cannot be explored by the formulae used here. The electron confinement has been described more closely than in II by using variational envelope functions. These follow the variation with n and allow an equivalent ISW width (δ) to be estimated. The constant values used in II were a large underestimate for both the GaAs (heterojunction) and Si (MOSFET) cases. Improvements made by using more accurate treatments are unlikely to be important compared to other

uncertainties, in E_1 or $\tau(\varepsilon)$, for example, because S_g depends rather weakly on the exact form of $\phi_1(z)$.

It has been interesting to compare the calculations in the two materials GaAs and Si because, until now, the influence of the structure used in the measurement was insignificant compared to the difference in material parameters. The piezoelectric mechanism of acoustic phonon scattering, for example, is an important consideration in GaAs but is absent in Si. The prediction of peaks in $-S_g/T^3$ is complicated in GaAs because the two coupling mechanisms, and hence the S_g integrand, depend differently upon the phonon wavevector. Furthermore, the different values of m and the valley degeneracy in Si result in a much smaller value for ε_f for given n, than in GaAs. This is most significant for many of the corrections which have been made because conduction at low T is obviously dominated by the behaviour of electrons with $\varepsilon(\mathbf{k})$ around ε_f. The importance of non-degeneracy in the scattering rate and the T dependence in the screening, for example, increase with K_BT/ε_f. Both have been neglected in previous publications but it has been shown here that this can lead to overestimates of $-S_g$ by 40% or more. (Smith and Butcher 1989b). The large effect of non-degeneracy in the polarizability which is apparent in Figure 5.5 is not necessarily reflected in S_g. This is a consequence of the variation of $\bar{q}$ with T and the resulting contribution to S_g which is made from phonons with wavevectors away from $2k_f$, where the effect on the polarizability is greatest.

6.3 Outstanding problems.

In principle the inclusion of non-degeneracy in the scattering rate and the temperature dependence in the screening described here allows the calculation of S_g to higher temperatures than were possible before. This would be of interest because the measurements of S by Fletcher et al (1986) for T up to 30K, for example, show broad maxima in $S(T)$ at T =10 to 15K. The expectation is that S_g falls and the dominance of S by phonon-drag is increasingly lost to diffusion. However, this is most likely to be the result of increasing phonon scattering which reduces L below the boundary scattering limit. This limit is a

reasonable first approximation for T <10K but may already need correction (see Section 5.5). The fit value (see Chapter 5) of L at T =5K of 0.1mm for the data of Fletcher et al (1986), for example, is already below the minimum specimen dimension of L =0.36nm. This is just one of the problems which arise in comparing the theory with the experimental data. Another is the uncertainty in material parameters such as E_1. This will be resolved by further measurement and investigation but to account for $L(T) = v_s\tau_{pp}(\mathbf{Q})$ the detail of the phonon scattering must be considered.

The most important problem in comparing theory and experiment is now the uncertainty in $\tau(\varepsilon)$. The value of p used to estimate S_d (see (5.10) and (5.9)) is increasingly important as T increases because of the growing contribution which S_d makes to S. Its importance has been shown to be significant even over the range 1-10K. Calculation of $\tau(\varepsilon)$ for T up to 30K to establish the value of p for the systems of interest as a function of n and accounting for surface roughness and ionized impurity scattering in a MOSFET, for example, is likely to prove very interesting. In addition, the correction to S_g for $\tau(\varepsilon)$ (Section 5.6) introduces a dependence upon the dominant electron scattering mechanism into the phonon-drag whereas previously only the electron-phonon coupling mechanism was believed to be important. The corrections can be large and, for sufficiently large $\hbar\omega_{\mathbf{Q}}/\varepsilon_f$, a negative p can, in principle, change the sign of both S_d and $\lambda(\mathbf{Q})$, (see (5.9) and (5.13)). This is obviously difficult to reconcile with a net transfer of phonon momentum to electrons and may be prevented in practice because, when $\hbar\omega_{\mathbf{Q}}/\varepsilon_f$ is large, the linear expansion of $\tau(\varepsilon_f + \hbar\omega_{\mathbf{Q}})$ about $\tau(\varepsilon_f)$ is not good enough. Nevertheless further investigation of $\tau(\varepsilon)$ and its effect upon S_d and S_g will be worthwhile now that its extra significance has been demonstrated here. The effect of $\tau(\varepsilon)$ upon S_d near subband minima has been explored by Cantrell and Butcher (1985) and interesting results involving the sign change mentioned in Section 1.5 have been reported by Ruf et al (1988b), but the study of the effect of $\tau(\varepsilon)$ upon S_g is a new development. Another important problem in treating S_g at higher T is the growing influence of higher subbands. Clearly MSS effects increasingly influence the screening but the possibility of intersubband scattering also requires more

subbands to be included in the S_g formula. The absorption of higher energy phonons at higher T can take electrons from one subband to another. These effects become more important as $K_BT/(\varepsilon_2 - \varepsilon_f)$ increases. Thus, as with $\tau(\varepsilon)$, S_g will increasingly reflect the physics of the structure, in addition to that of the quasi-2D confinement. The influence of the spacer layer, for example, and the form of the confining potential, which determines the subband structure, will be of great significance.

For the present calculations the outstanding difficulty is the crossover in the $S_g(T)$ curves for different n in a MOSFET (see Figure 5.7) but this is likely to be resolved by the calculation of $\tau(\varepsilon)$ and S_d. The agreement between the curves calculated for the heterojunction, and the peaks in $-S_g/T^3$ for the MOSFET, and the experimental data is very good indeed. Some of the agreement obtained using the corrections described in Chapter 5 is better than the published results (Lyo 1988, Smith and Butcher 1989a,b). Furthermore, the new data for the MOSFET (Gallagher et al 1988) shows much sharper peaks and at lower n than previously published (Gallagher et al 1987) which have moved closer to the theoretical values calculated here (compare Figures 2.3 and 5.8). The shift in the position of the peaks in $-S_g/T^3$ which was noted when screening was introduced into the MOSFET calculations, has been removed by using n_{free} and the other corrections.

In conclusion it appears that, whilst there are still some outstanding difficulties (particularly if the calculations are to be taken to higher T) when $L(T)$ is accounted for and E_1 and $\tau(\varepsilon)$ are known with more certainty, the theoretical description of phonon-drag in quasi-2D will be extremely good.

6.4 Prospects for further developments.

To take the calculations of S_g in quasi-2D beyond the quantum limit and boundary scattering regime it has been shown that $\tau(\varepsilon)$, and $L(T)$ must be calculated and higher subband occupation accounted for. With the corrections and improvements introduced here the prospects for this development are most favourable because the necessary multi-subband formalism is provided in I. More detailed knowledge of the subband structure will be nec-

essary to compare these calculations with experimental data though, because the energy differences, $\varepsilon_2 - \varepsilon_1$ etc, determine the onset of intersubband scattering. The 2D $\rightarrow$ 3D dimensionality crossover may then be explored and the results for S_g compared with the 2D and 3D simple models. In addition, the approach of the 3D limit as $\delta \rightarrow \infty$ should demonstrate the gradual recovery of the conservation requirement for the transverse momentum component in an electron-phonon collision. The correction to S_g for $\tau(\varepsilon)$ is of particular interest in a widening channel due to the increasing role of the intersubband scattering. Features as strong as those predicted by Cantrell and Butcher (1985) in S_d are not likely because S_d is proportional to the derivative of $\tau(\varepsilon)$ but the possibility of some such structure is a new and interesting development.

The same formalism (I and II) has been applied by Kubakaddi and Butcher (1989a) to the quasi-1D case by considering thin wires of various geometries. Their results for a thin cylindrical wire in GaAs show very similar features to that of quasi-2D, with comparable importance of screening and piezoelectric scattering. This is attributed to the dominance of S_g by the 3D character of the phonons. Kundu et al (1988) have performed corresponding calculations of S_d.

When the S_g calculations are taken to higher temperatures some more interesting questions may be answered. It should be resolved, for example, whether the peaks in $-S(T)$ observed by Fletcher et al (1986) are due to the increase in phonon collisions decreasing L or whether other phenomena, such as electron scattering, are reducing the drag. Whatever the explanation, the dominance changeover in $S : S_d \rightarrow S_g \rightarrow S_d$ as T is raised, should be observed but S_d must be known more precisely. The calculation of $\tau(\varepsilon)$ will help in this respect but the validity of the 2D Mott formula (5.9) in widening channels will have to be examined. The onset of drag by optic phonons will become increasingly important as T is raised still further but should be accomodated by the same formalism, given suitable corrections for the new coupling mechanisms and dispersion relation.

It is now clear that the thermopower measurements in the extreme quantum limit are complicated by the presence of a large contribution from phonon-drag. Observation of

accurate quantization of the peak maxima (see Chapter 1) is not possible then, unless S_g is subtracted but the current formalism cannot be applied to the $B \neq 0$ case. Some effort has been made in this direction (Karyagin et al 1988, Kubakaddi and Butcher 1989b, Lyo 1989) and the results look promising but further work is necessary. Uncertainties in the parameters (such as E_1, for example) however, will not allow an accurate subtraction of S_g from the data and such accurate quantization to be revealed. However, it would be a very exciting development if such studies of S_g were able to show how the phonon-drag complication could be suppressed in experiments.

References

Ando T. 1976 *Phys.Rev.B* **13** 3468

Ando T. 1982a *J.Phys.Soc.Jpn.* **51** 3893

Ando T. 1982b *J.Phys.Soc.Jpn.* **51** 3900

Ando T., Fowler A.B. and Stern F. 1982 *Rev.Mod.Phys* **54** 437

Ashcroft N.W. and Mermin N.D. 1981 *Solid-State Physics* (Philadelphia: Holt-Saunders)

Bailyn M. 1958 *Phys.Rev.* **112** 1587

Bailyn M. 1967 *Phys.Rev.* **157** 480

Berggren K.F. 1988 *Int.J.Qu.Chem.* **33** 217

Blatt F.J. 1968 *Physics of Electronic Conduction in Solids* (New York: McGraw-Hill)

Bauer G., Kuchar F. and Heinrich H. eds. 1984 *Two dimensional systems: heterostructures and superlattices.* (Berlin: Springer- Verlag)

Bauer G., Kuchar F. and Heinrich H. eds. 1986 *Two dimensional systems: Physics and new Devices* (Berlin: Springer- Verlag)

Bernasconi J. and Schneider J. 1981 *Physics in one dimension* (Springer-Verlag)

Butcher P.N. 1986 In:*Crystalline semiconducting materials and devices* Eds. Butcher P.N., March N.H. and Tosi M.P. (New York: Plenum)

Cantrell D.G. and Butcher P.N. 1985 *J.Phys.C:Solid State Physics* **18** L587

Cantrell D.G. and Butcher P.N. 1986 *J.Phys.C:Solid State Physics* **19** L429

Cantrell D.G. and Butcher P.N. 1987a *J.Phys.C:Solid State Physics* **20** 1985

Cantrell D.G. and Butcher P.N. 1987b *J.Phys.C:Solid State Physics* **20** 1993

Cardona M. 1989 *Superlat. and Microstructs.* **5** 27

Chang L.L., Esaki L. Tsui R. 1974 *Appl.Phys.Lett.* **24** 593

Das Sarma S. and Vinter B. 1982 *Phys.Rev.B* **26** 960

Davidson J.S., Dahlberg E.D., Valois A.J. and Robinson G.Y. 1986 *Phys.Rev.B* **33** 8238

Devrese J.T. and Brosens F. 1983 (Eds.) *Electron correlation in solids, molecules and atoms. (NATO advanced study series B, Physics;* **81***)* (New York: Plenum)

Dharrsi I. and Butcher P.N. 1989 (Submitted to *J.Phys.:Condens. Matter)*

D' Iorio M., Stoner R. and Fletcher R. 1988 *Solid State Commun.* **65** 697

Duffin W.J. 1980 *Electricity and Magnetism* (Maidenhead: McGraw-Hill)

Eaves L., Alves E.S., Foster T.J., Henini M., Hughes O.H., Leadbeatter M.L., Sheard F.W., Toombs G.A., Chan K., Celeste A., Portal J.C., Hill G. and Pate M.A. 1988 In: *Physics and technology of submicron structures* Eds. Heinrich H., Bauer G. and Kuchar F. (Berlin: Springer-Verlag)

Ehrenreich H. and Cohen M.H. 1959 *Phys.Rev.* **115** 786

Fang F.F. and Howard W.E. 1966 *Phys.Rev.Lett.* **16** 797

Fletcher R., Maan J.C., Ploog K. and Weimann G. 1986 *Phys.Rev.B* **33** 7122

Fletcher R., D'Iorio M., Sachrajda A.S., Stoner R., Foxon C.T. and Harris J.J. 1988a *Phys.Rev.B* **37** 3137

Fletcher R., D'Iorio M., Moore W.T. and Stoner R. 1988b *J.Phys.C: Solid State Physics* **21** 2681

Gallagher B.L., Gibbings C.J., Pepper M. and Cantrell D.G. 1987 *Semicond.Sci.Technol.* **2** 456

Gallagher B.L., Oxley J.P., Galloway T., Smith M.J., Butcher P.N. and Pepper M.: Poster presentation I.O.P.Solid-State Physics Conference, Nottingham 1988 (unpublished to date).

Geballe T.H. and Hull G.W. 1954 *Phys.Rev.* **94** 1134

Girvin S.M. and Jonson M. 1982 *J.Phys.C: Solid State Phys.* **15** L1147

Gradshteyn I.S. and Ryzhik I.M. 1980 *Tables of Integrals, series and products* (New York: Academic Press)

Guenault A.M. 1971 *J.Phys.F: Metal.Phys.* **1** 373

Gurevich L. 1945 *J.Phys. (USSR)* **9** 477

Gusev G.M., Zavaritsky N.V., Kvon Z.D. and Yurgens A.A 1984 *JETP Lett.* **40** 1057

Heinrich H., Bauer G. and Kuchar F. eds. 1988 *Physics and technology of submicron structures* (Berlin: Springer-Verlag)

Hensel J.C., Dynes R.C. and Tsui D.C. 1983 *Phys.Rev.B* **28** 1124

Herring C. 1953 *Phys.Rev.* **92** 857

Herring C. 1954 *Phys.Rev.* **96** 1163

Hohenberg P. and Kohn W. 1964 *Phys.Rev.* **136** B864

Inkson J. C. 1984 in *Many-body theory of solids. An Introduction* (New York: Benjamin)

Jensen H.H. 1964 In: *Phonons and phonon interactions* Ed. Bak T.A. (New York: Plenum)

Jonson M. and Girvin S.M. 1984 *Phys.Rev.B* **29** 1989

Kagoshima S., Nagasawa H. and Sambongi T. 1982 *One dimensional conductors* (Springer-Verlag)

Karl H., Dietsche W., Fischer A. and Ploog K. 1988 *Phys.Rev.Lett.* **61** 2360

Karyagin V.V., Lyapalin I.I. and Dyakin V.V. 1988 *Soviet Phys.Semicond.* **22** 954

Kearney M.J. and Butcher P.N. 1986 *J.Phys.C:Solid State Physics* **19** 5429

Kelly M.J. and Nicholas R.J. 1985 *Rep.Prog.Phys.* **48** 1699

Kelly M.J. and Wiesbuch C. 1986 *The Physics and Fabrication of Microstructures and devices* (Berlin: Springer-Verlag)

Kent A.J., Hardy G.A., Hawker P., Rampton V.W., Newton M.I., Russel P.A. and Challis L.J. 1988 *Phys.Rev.Lett.* **61** 180

Kittel C. 1976 *Introduction to solid state physics.* (New York: Wiley)

Kittel C. 1987 *Quantum Theory of solids.* (New York: Wiley)

Kohn W. and Sham L.J. 1965 *Phys.Rev.* **140** A1133

Kramer B., Bergmann G. and Bruynserade Y. 1985 (Eds.) *Localization, interaction and transport phenomena.* (Berlin: Springer-Verlag)

Kubakaddi S.S. and Butcher P.N. 1989a *J.Phys.Condens.Matter* **1** 3939

Kubakaddi S.S. and Butcher P.N. 1989b *Phys.Rev.B* 40 1377

Landolt-Bornstein 1982 *Numerical data and functional relationships in science and technology, New Series, Group III: Crystal and solid-state physics Volume 17: semiconductors.* (Berlin: Springer-Verlag)

Lax M. 1974 In *Symmetry principles in solid state and molecular physics* (New York: Wi-

ley)

Lyo K. 1988 *Phys.Rev.B.* **38** 6345

Lyo K. 1989 (Private communication)

Madalung O. 1981 *Introduction to Solid State Theory* (Berlin: Springer-Verlag)

Mahan G.D. 1981 *Many-particle Physics* (New York: Plenum)

Maldague P.F. 1978 *Surf.Sci.* **73** 296

Manion S.J., Artaki M., Emanuel M.A., Coleman J.J., Hess K., Nolte D.D., Walukiewicz W. and Haller E.E. 1987 *Phys.Rev.Lett.* **59** 501

Matsumoto Y. and Uemura Y. 1974 *Jpn.J.Appl.Phys.Suppl.***2** 367

Mendez E.E., Esaki L. and Wang W.I. 1986 *Phys.Rev.B* **33** 2893

Mori S. and Ando T. 1979 *Phys.Rev.B* **19** 6433

Mott N.F. 1969 *Philos.Mag.* **19** 835

Nagaoka Y. and Fukyama H. 1982 (Eds.) *Anderson localization* (Berlin: Springer-Verlag)

Nicholas R.J. 1985 *J.Phys.C:Solid State Phys.* **18** L695

Northrop G.A. and Wolfe J.P. 1985 In: *Nonequilibrium phonon dynamics* Ed. Bron W.E. (New York: Plenum)

Obloh H., Von Klitzing K. and Ploog K. 1984 *Surf.Sci.* **142** 236

Price P.J. 1981 *Ann.Phys.* **133** 217

Ridley B.K. 1982 *Quantum processes in semiconductors* (Oxford: Clarendon)

Rothenfusser M., Koster L. and Dietsche W. 1986 *Phys.Rev.B* **34** 5518

Ruf C., Obloh H., Junge B., Gmelin E. and Ploog K. 1988 *Phys.Rev.B* **37** 6377

Ruf C., Brummel M.A., Gmelin E. and Ploog K. 1989 *Superlat. and Microstructs.* **6** 175

Sakaki H., Noda T., Hirakawa K., Tanaka M. and Matsuesue T. 1987 *Appl.Phys.Lett.* **51** 1934

Schulz M. 1986 In: *Crystalline semiconducting materials and devices* Eds. Butcher P.N., March N.H. and Tosi M.P. (New York: Plenum)

Siggia E.D. and Kwok P.C. 1970 *Phys.Rev.B* **2** 1024

Smith M.J. and Butcher P.N. 1989a *J.Pys.: Condens.Matter* **1** 1261

Smith M.J. and Butcher P.N. 1989b *J.Pys.: Condens.Matter* **1** 4859

Stern F. and Das Sarma S. *Phys.Rev.B* **30** 840

Stern F. and Howard W.E. 1967 *Phys.Rev.* **163** 816

Stern F. 1967 *Phys.Rev.Lett.* **18** 546

Stern F. 1980 *Phys.Rev.Lett.* **44** 1469

Streda P. 1983 *J.Phys.C: Solid State Phys.* **16** L369

Syme R.T. and Pepper M. 1989 *Superlat. and Microstructs.* **5** 103

Sze S.M. 1981 *Physics of semiconductor devices (Second Edition)* (New York: Wiley)

Takada Y. and Uemura Y. 1977 *J.Phys.Soc.Jpn.* **43** 139

Tang D.S. 1988 *Phys.Rev.B* **37** 8319

van Wees B.J., van Houten H., Beenaker C.W.J., Williamson J.G., Kouwenhoven L.P.,

van der Marel D. and Foxon C.T. 1988 *Phys.Rev.Lett.* **60** 848

von-Klitzing K., Dorda G. and Pepper M. 1980 *Phys.Rev.Lett.* **45** 494

von-Klitzing K. 1986 *Rev.Mod.Phys* **58** 519

Wharam D.A., Thornton T.J., Newbury R., Pepper M., Ahmed H., Frost J.E.F., Hasko

D.G., Peacock D.C., Ritchie D.A. and Jones G.A.G. 1988 *J.Phys.C: Solid State Phys. 21*

L209 Zavaritsky N.V. and Kvon Z.D. 1983 *JETP Lett.* **38** 97

Zavaritsky and Zavaritsky 1983 *Sov.Phys. JETP* **56** 674

Zavaritsky N.V. 1984 *Physica* **126** B369

Zawadzki W. 1982 in *Handbook on Semiconductors Vol.1. (Band theory and Transport properties* Ed. Paul W. (Amsterdam: North Holland)

Zawadzki W. and Lassnig R. 1984 *Surf.Sci.* **142** 225

Ziman J.M. 1963 *Electrons and Phonons - the theory of transport phenomena in solids* (Oxford: Clarendon)

Ziman J.M. 1972 *Principles of the theory of solids.* (Cambridge: University Press)

Appendix A

Formulae for the Ando envelope function.

The potential functions contributing to $V(z)$ are:

$$V_0(z) = \begin{cases} V_0 & z < 0 \\ 0 & z \geq 0 \end{cases} \tag{A.1}$$

$$V_{s,dep}(z) = \begin{cases} \frac{N_{dep}e^2 z}{2\varepsilon} & z < 0 \\ \frac{-N_{dep}e^2}{2\varepsilon} z\left(\frac{z}{z_d} - 1\right) & 0 \leq z < z_d \\ \frac{-N_{depl}e^2}{2\varepsilon}(z - z_d) & z \geq z_d \end{cases} \tag{A.2}$$

$$V_{s,don}(z) = \begin{cases} \frac{-N_D e^2 d_2 z}{2\varepsilon} & z < -(d_1 + d_2) \\ \frac{N_D e^2}{2\varepsilon}[z^2 + (2d_1 + d_2)z + (d_1 + d_2)^2] & -(d_1 + d_2) \leq z < -d_1 \\ \frac{N_D e^2 d_2}{2\varepsilon}[z + 2d_1 + d_2] & z > -d \end{cases} \tag{A.3}$$

and:

$$V_s(z) = \begin{cases} \frac{ne^2}{\epsilon} \left\{ \begin{array}{r} \left[\frac{B^2}{2b}(6 + 4\beta + \beta^2) - \frac{A^2}{2a} - \frac{1}{2}z - \frac{B^2}{b}e^{-bz}\{b^2z^2 + \right. \\ \left. 2bz(2 + \beta) + 6 + 4\beta + \beta^2\}\right] \end{array} \right\} & z \geq 0 \\ \frac{ne^2}{\epsilon}\left[\frac{1}{2}z + \frac{A^2}{a}\left(\frac{1}{2} - e^{az}\right) - \frac{B^2}{2b}(6 + 4\beta + \beta^2)\right] & z < 0, \end{cases} \tag{A.4}$$

where:

$$N_{dep} = N_A z_d \tag{A.5}$$

and:

$$d_2 = \frac{n + N_{depl}}{N_D}. \tag{A.6}$$

The quantities V_0 (barrier height), d_1 (spacer thickness), N_D (net donor density in the doped region of GaAlAs $z < -d_1$) and N_A (net acceptor density in the GaAs) are specimen parameters. The width d_2 follows from overall charge neutrality. The depletion width z_d is determined by refering to Figure 7. Hence the quantity V_1 is given from the difference between $V(z)$ at plus and minus infinity and can be equated to the band gap E_G in GaAs (about 1.52 eV). Strictly, corrections arise from the departures of E_F from E_C in the GaAlAs and E_v in the GaAs but these departures are typically $<< 0.1E_G$ at 4K (see, for example, Landolt-Bornstein 1982). The equation to be solved for z_d is then:

$$E_G = \frac{N_A e^2 z_d^2}{2\varepsilon} + \frac{N_D e^2 d_2}{2\varepsilon}(2d_1 + d_2) - V_0 - \frac{ne^2}{\varepsilon}\left[\frac{A^2}{a} - \frac{B^2}{b}(6 + 4\beta + \beta^2)\right] \tag{A.7}$$

Strictly, this equation requires a and b to be known but this is unnecessary in practice because the term involving a and b can be neglected.

The quantity to be minimized is:

$$E_{a,b} = < T(z) >_z + < V_0(z) >_z + < V_{don}(z) >_z + < V_s(z) >_z + < V_{dep}(z) >_z, \tag{A.8}$$

where A, B and β are related to a and b through (3.27) to (3.29), with:

$$< T(z) >_z = \frac{\hbar^2}{8m}(B^2 b(2\beta + 2 - \beta^2) - A^2 a), \tag{A.9}$$

$$< V_0(z) >_z = V_0 A^2, \tag{A.10}$$

$$\begin{aligned} < V_{don}(z) >_z \;=\; & \frac{-N_A e^2 A^2}{2\varepsilon} e^{-ad_1}\left[d_2(d_1 + d_2) - \frac{d_2}{a} - e^{-ad_2}\left\{d_2^2 - d_1 d_2 - 2d_1^2 - \frac{2d_1 + d_2}{a}\right\}\right] \\ & + \frac{N_A e^2 d_2}{2\varepsilon}\left[A^2(2d_1 + d_2 - 1/a) + B^2\left\{(2d_1 + d_2)(2 + 2\beta + \beta^2)\right.\right. \\ & \left.\left. + (6 + 4\beta + \beta^2)/b\right\}\right], \end{aligned} \tag{A.11}$$

$$\begin{aligned} < V_s(z) >_z \;=\; & \frac{ne^2}{\varepsilon}\left[A^2(A^2 - 2)/2a - A^2 B^2(6 + 4\beta + \beta^2)/b - B^2(6 + 4\beta + \beta^2)/2b\right. \\ & \left. - B^4(30 + 60\beta + 60\beta^2 + 24\beta^3 + 4\beta^4)/8b\right] \end{aligned} \tag{A.12}$$

and:

$$< V_{dep}(z) >_z = -\frac{N_{depl}e^2}{2\varepsilon}\left[\frac{A^2}{a} + B^2\left\{\frac{(24+12\beta+2\beta^2)}{z_d b^2} - \frac{(\beta^2+4\beta+6)}{b}\right\}\right]. \quad \text{(A.13)}$$

Appendix B

Calculations of MSS quantities in an ISW.

The multisubband polarizability $\Pi_{\mu\lambda}(\mathbf{q})$ is calculated in the quantum-limit directly from (4.40), which can be rewritten as:

$$\Pi_{\mu\lambda}(\mathbf{q}) = \pi_1(\omega) - \pi_2(\omega) \tag{B.1}$$

where:

$$\pi_1(\omega) \equiv \frac{2g_v}{A} \lim_{\eta\to 0} \sum_{\mathbf{k}} \frac{f_{\lambda,\mathbf{k}}}{\varepsilon(\mathbf{k}) - \varepsilon(\mathbf{k}+\mathbf{q}) - \omega} \tag{B.2}$$

and

$$\pi_2(\omega) \equiv \frac{2g_v}{A} \lim_{\eta\to 0} \sum_{\mathbf{k}} \frac{f_{\mu,\mathbf{k}+\mathbf{q}}}{\varepsilon(\mathbf{k}) - \varepsilon(\mathbf{k}+\mathbf{q}) - \omega}. \tag{B.3}$$

Here $\varepsilon(\mathbf{k})$ is $\hbar^2(\mathbf{k}.\mathbf{k})/(2m)$ and ω is $\varepsilon_\mu - \varepsilon_\lambda - i\eta$ with ε_i the i^{th} subband energy. It can be observed that $\pi_2(\omega) = -\pi_1(-\omega)$ and $\pi_1(\omega)$ can be evaluated by transforming to an integral over $\mathbf{k}$ up to $\mid \mathbf{k} \mid = k_f$ in the quantum limit. Here $f_{\lambda,\mathbf{k}}$ is zero for $\lambda > 1$ or $\lambda = 1$ with $\mid \mathbf{k} \mid > k_f$, and:

$$\pi_1(\omega) = \pm\frac{g_v m}{2\pi}\delta_{\lambda,1}\left\{\left[\left(1+\frac{2m\omega}{q^2}\right)^2 - \left(\frac{2k_f}{q}\right)^2\right]^{1/2} - \left[\left(1+\frac{2m\omega}{q}\right)^2\right]^{1/2}\right\} \tag{B.4}$$

which gives Sterns' result (4.48) for $\Pi_{11}(\mathbf{q})$. The sign is determined by taking expansions of $f(\varepsilon(\mathbf{k}+\mathbf{q}))$ and $\varepsilon(\mathbf{k}+\mathbf{q})$ about $f(\varepsilon(\mathbf{k}))$ and $\varepsilon(\mathbf{k})$ in the limit $q \to 0$. The result for

$\lambda = 1$ and $\mu > 1$ is then:

$$\Pi_{\mu,1}(\mathbf{q}) = \frac{-g_v m}{2\pi\hbar^2}\left\{\left(1+\frac{2m\varepsilon'_\mu}{\hbar^2 q^2}\right)^2 - \left[\left(1+\frac{2m\varepsilon'_\mu}{\hbar^2 q^2}\right)^2 - \left(\frac{2k_f}{q}\right)^2\right]^{1/2}\right\} \tag{B.5}$$

where ε'_μ is $\varepsilon_\mu - \varepsilon_1$ (ie measured from the ground subband energy). The quantity $\Pi_{\mu\lambda}(\mathbf{q})$ is symmetrical in μ and λ and when both are greater than unity is zero when only the ground subband is occupied. The general result is given by Mori and Ando (1979).

The form factor $F_{\alpha\beta\mu\lambda}(q)$ is calculated from (4.38) using (4.45) for $\tilde{g}(q,z,z')$ with:

$$\phi_\alpha = (2/\delta)^{1/2}\sin(\alpha\pi z/\delta) \quad : \alpha = 1,2,3... \tag{B.6}$$

The general result can be written:

$$F_{\alpha\beta\mu\lambda}(q) = F_1 + F_2 \tag{B.7}$$

where:

$$F_1 = 2(1-\kappa_1/\kappa_2)I(\alpha,\beta,q)I(\mu,1,q)/\delta^2 \tag{B.8}$$

and

$$F_2 = q(1+\kappa_1/\kappa_2)\left[G(p_{\alpha\beta}) - G(p'_{\alpha\beta})\right]/\delta^2 \tag{B.9}$$

in which:

$$G(p_{\alpha\beta}) = \left[\frac{1}{2}\Delta_{\alpha\beta\mu\lambda} - I(\mu,1,q) - \cos(p_{\alpha\beta}\delta)I(\mu,1,-q)e^{-q\delta}\right]/(p^2_{\alpha\beta}+q^2) \tag{B.10}$$

with:

$$p_{\alpha\beta} = (\alpha-\beta)\pi/\delta \quad p'_{\alpha\beta} = (\alpha+\beta)\pi/\delta \tag{B.11}$$

$$\Delta_{\alpha\beta\mu\lambda} = \delta_{\alpha-\beta,\mu-\lambda} + \delta_{\alpha-\beta,\mu+\lambda} + \delta_{\alpha-\beta,\mu+\lambda} + \delta_{\alpha-\beta,-\mu-\lambda} \tag{B.12}$$

$$I(\alpha,\beta,q) = [I'(p_{\alpha\beta}) - I'(p'_{\alpha\beta})]/2 \tag{B.13}$$

and

$$I'(p_{\alpha\beta}) = q\left[1-\cos(p_{\alpha\beta}\delta)e^{-q\delta}\right]/(p^2_{\alpha\beta}+q^2) \tag{B.14}$$

For the matrix element $< \alpha \mid e^{iq_z z} \mid \beta >$ the result is

$$M_{\alpha\beta}(q_z) = \int_0^\delta \phi^*_\alpha(z)e^{iq_z z}\phi_\beta(z)dz = M^R_{\alpha\beta} + iM^I_{\alpha\beta}(q_z) \tag{B.15}$$

where both $M^R_{\alpha\beta}(q_z)$ and $M^I_{\alpha\beta}(q_z)$ are real and:

$$M^R_{\alpha\beta}(q_z) = m_2 \sin\theta_2 - m_1 \sin\theta_1 \tag{B.16}$$

and

$$M^I_{\alpha\beta}(q_z) = m_2(1-\cos\theta_1) - m_1(1-\cos\theta_1) \tag{B.17}$$

in which:

$$m_1 = \frac{q_z^2}{\delta(p^2_{\alpha\beta}-q_z^2)} \quad m_2 = \frac{q^2}{\delta(p'^2_{\alpha\beta}-q_z^2)} \tag{B.18}$$

and

$$\theta_1 = (\alpha-\beta)\pi/\delta + q_z\delta \quad \theta_2 = (\alpha+\beta)\pi/\delta + q_z\delta. \tag{B.19}$$