TRIBOLOGY GROUP NOMINATED LECTURE

ONE HUNDRED YEARS OF HERTZ CONTACT

K L JOHNSON, FRS, MA, MScTech, PhD, CEng, MIMechE
Cambridge University, Engineering Laboratory, Trumpington Street, Cambridge

1 INTRODUCTION

Heinrich Hertz graduated *magna cum laude* from the University of Berlin in 1880 at the age of twenty-three. For the next two and a half years he was appointed a research assistant to Helmholtz in Berlin and it was during this time that he did his work on contact mechanics. His interest was stimulated during experiments on optical interference between glass lenses in contact (Newton's rings), by a concern for the possible influence of elastic deformation of the lenses under the action of the contact force. He worked fast and appears to have developed the theory as we know it during the Christmas vacation of 1880. He presented a paper to the Berlin Physical Society in January 1881, which is the commonly quoted date for publication of the theory although it did not appear in print until the following year (1). The paper* was submitted to the *Journal for Pure and Applied Mathematics* edited by Kronecker who sent it to Kirchoff for review. When he had had no reply by May, Hertz made enquiries of Kronecker who sent him to see Kirchoff. It was clear that Kirchoff was impressed, though he claimed to have found an error and had also set about rewriting large sections of the paper. At first Hertz was pleased at Kirchoff's interest but became increasingly annoyed with Kirchoff's 'improvements' and made the recommended changes with the 'greatest reluctance'. (It turned out that the supposed error was Kirchoff's and not Hertz'.) The revised manuscript was submitted at the end of June and Hertz was disappointed to hear that it would not be out before Christmas (1881)!†

Meanwhile the presentation at the Physical Society had aroused some technological interest. A member of the audience worked for the Standards Commission. Through him the theory found its first application in the calculation of the compression of polished glass spheres which were placed between standard rods used for the accurate measurement of distance. A teacher at the Industrial Academy approached Hertz after his presentation suggesting that he communicate his results to a technical journal. Consequently a second paper ‡(2) was prepared and published in *Verhandlungen des Vereins zur Beförderung des Gewerbefleisses* in November 1882. In addition to presenting the elastic theory as before, it included the extension to the two-dimensional case of line contact, a discussion of how the results might be applied to the definition and measurement of hardness, and also the results of experiments in support of the theory. In these experiments the shape and size of the contact area between loaded glass lenses were measured under a microscope by coating one of the surfaces with a thin layer of lamp black.

Turning now to the theory itself, the key step lay in bringing together two different mathematical concepts: (*i*) the geometry of two curved surfaces which touch without deformation and (*ii*) the theory of potential applied to an elastic half-space bounded by a plane surface. When two smooth surfaces touch, without deformation, at the origin O of cartesian axes, in which Oz is their common normal and the x–y plane is their common tangent plane, it can be shown that their separation parallel to the z-axis (with a suitable choice of direction for the x-axis) can be expressed by:

$$h \cong Ax^2 + By^2 \tag{1}$$

where $A(\equiv 1/2R')$ and $B(\equiv 1/2R'')$ are functions of the principal radii of curvature ρ_1', ρ_1'' and ρ_2', ρ_2'' of the two surfaces respectively, and R' and R'' are their principal *relative* radii of curvature. In this expression higher order terms in x and y are neglected, so that it is a good approximation only at distances from O which are small compared with the relative radii of curvature of the surfaces. It is clear from equation (1) that contours of constant separation are ellipses whose ratio of semi-axes $a'/b' = \sqrt{(B/A)}$. It seems probable that Hertz was familiar with this piece of geometry from his work on interference fringes, indeed he treats it in a perfunctory way in his papers as though it were common knowledge. The further proposition: that under load the solids deform such that the contact area is also elliptical, may have been suggested by observation. Under compression surface points on each body undergo normal elastic displacements $w_1(x, y)$ and $w_2(x, y)$ (see Fig. 1) such that, within the contact area:

$$w_1(x, y) + w_2(x, y) = \delta - Ax^2 - By^2 \tag{2a}$$

and outside the contact area:

$$w_1(x, y) + w_2(x, y) > \delta - Ax^2 - By^2 \tag{2b}$$

so that the surfaces do not overlap. In these equations δ is the approach of those points in each body which are

* 'On the contact of elastic solids'.
† These details are taken from Hertz' letters to his family (3) which provide a fascinating personal commentary on the progress of his work and, indeed, on his life as a whole.
‡ 'On the contact of rigid elastic solids and on hardness'.

0020-3483/82/0196-0363 $02.00

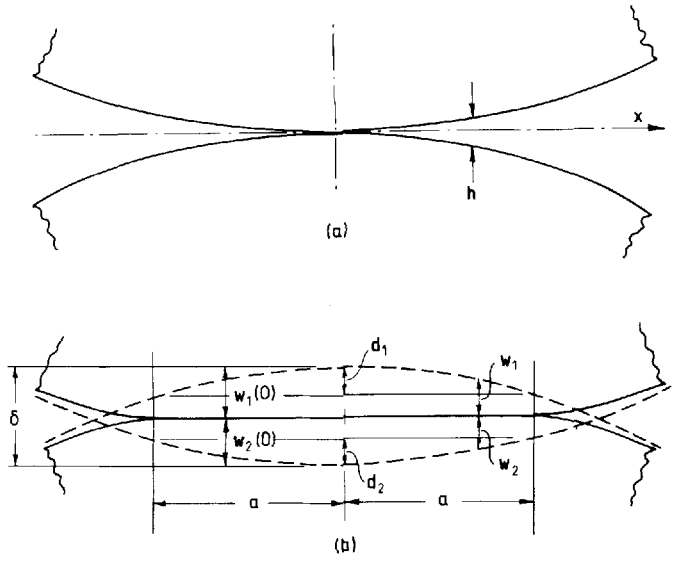

Fig. 1 The geometry of contact of non-conforming bodies (a) Unloaded (*b*) Loaded

chosen as datum for elastic displacements, so that:

$$\delta = w_1(0) + w_2(0) \tag{2c}$$

Hertz recognized that the contact of non-conforming bodies was a case of stress concentration which could be analysed independently of the geometry and state of stress in the bodies as a whole. He wrote: 'We can confine our attention to that part of each body which is very close to the point of contact, since here the stresses are extremely great compared with those occurring elsewhere, and consequently depend only to the smallest extent on the forces applied to other parts of the bodies.' By confining attention to that part of the surface close to the contact region, the stresses and deformations could be found by neglecting the slight curvature of the surfaces of the two bodies and regarding each as an elastic half-space bounded by the plane surface $z = 0$. Studies of the deformation of such a semi-infinite solid due to pressure exerted on a small area of its plane surface had already been carried out, notably by Boussinesq, who made use of the method of potential. Boussinesq had already found the deflexion of the surface of a half-space produced by a *uniform* pressure acting on an elliptical area and had also shown that the pressure necessary to produce a uniform normal displacement of an elliptical area is of the form:

$$p'(x, y) = p'_0[1 - (x/a)^2 - (y/b)^2]^{-1/2} \tag{3}$$

Hertz must have been familiar with this approach since he proceeds to show that a pressure distribution of the form:

$$p(x, y) = p_0[1 - (x^2/a^2) - (y^2/b^2)]^{1/2} \tag{4}$$

produces normal displacements within the ellipse whose semi-axes are a and b, such that:

$$w_1(x, y) + w_2(x, y) = (L - Mx^2 - Ny^2)/\pi E^* \tag{5}$$

where L, M and N are functions of p_0, a and b, and $1/E^* = (1 - \nu_1^2)/E_1 + (1 - \nu_2^2)/E_2$. In this calculation he draws the analogy between the elastic displacements produced by a distribution of pressure on the surface of a half-space and the electric potential due to a distribution of surface charge, which suggests that he was familiar with potential theory from his studies of electrostatics. Now, by a suitable choice of values for a, b and δ, the displacements given by equation (5) will satisfy the condition of contact given by equation (2a) (see Appendix 1). Hertz then went on to show that the displacement *outside* the ellipse of contact satisfied the condition (2b) that the surfaces should not overlap. The solution was then complete.

In this respect it is interesting to note that the addition of a Boussinesq pressure distribution of the form (3) adds a constant displacement within the loaded ellipse, which does not invalidate the condition (2a) but merely changes the value of the approach δ. It is unacceptable, however, in other ways (see Fig. 2). If this additional pressure, however small, were positive (compressive) it would give rise to a discontinuity in gradient of the deformed surface which would lead to interference outside the contact. If it were negative (tensile) then the infinite tension at the boundary, in the absence of an adhesive, would cause the surfaces to peel apart (we shall return to this point in Section 9). The Hertz pressure distribution is, therefore, the only one which satisfies all the conditions of the problem.

The Hertz theory is an inspired approximation on two obvious counts. First, the geometry of general curved surfaces is described by quadratic terms only and second, the two bodies, at least one of which must have a curved surface, are taken to deform as though they were elastic half-spaces. From time to time attempts are made to improve on Hertz by taking higher order terms in the geometric description of cylindrical or spherical profiles and by calculating the deformations of truly cylindrical or spherical solids. The corrections so found increase with the ratio of contact size a to radius of curvature R. It must not be forgotten, however, that the linear theory of elasticity is itself an approximation confined to small strains. We shall see that the magnitude of the strains in a contact situation is also proportional to the ratio a/R and it has been shown that, for non-conforming surfaces, the errors incurred by the Hertz approximations are of comparable magnitude with the errors of linear elasticity. There is no

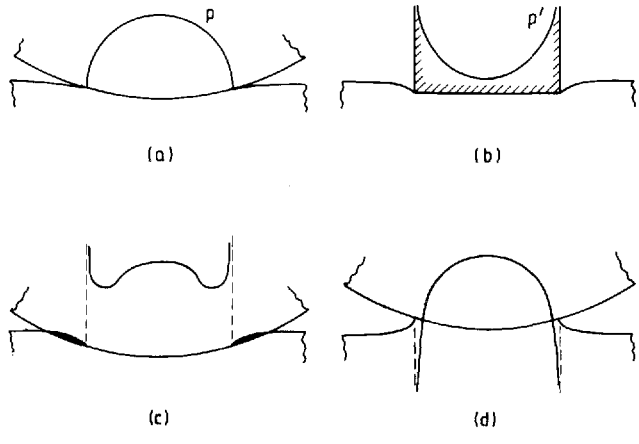

Fig. 2 Contact pressure distributions which satisfy equation (2a)
(a) Hertz: $p = p_0[1 - (x/a)^2 - (y/b)^2]^{1/2}$
(b) Boussinesq: $p' = p'_0[1 - (x/a)^2 - (y/b)^2]^{-1/2}$
(c) $p + p'$
(d) $p - p'$

point, therefore, in making such corrections whilst retaining linear theory. It does follow, however, that doubt must be cast on the Hertz results on all three counts if the ratio a/R becomes too large. With metallic bodies this restriction is ensured by the small strains at which the elastic limit is reached, but a different situation arises with compliant elastic solids like rubber. Rather surprisingly Fessler and Ollerton (**4**), using compliant photoelastic models, report that measurements of the contact dimensions did not depart significantly from Hertz' predictions at values of a/R up to 0.3 (the limit of the strength of the material). This comforting observation must be due to the errors discussed above being compensatory.

A different situation arises with *conforming* surfaces in contact, for example, a pin in a closely fitting hole (**5**) or a ball and socket joint (**6**). Here the arc of contact may be large compared with the radius of the hole or socket without incurring large strains. The contact problem may then be analysed using the linear theory of elasticity provided that the true geometry is used in the elastic calculation.

2 A SIMPLE TREATMENT OF THE HERTZ THEORY

The final equations of the Hertz theory which relate the size of the contact area, the contact pressure and the elastic compression to the load, geometry and elastic moduli, often appear to the practising engineer as magic formulae. A simple qualitative treatment may help to dispel the magic.

Two curved bodies in contact, of radius R_1, R_2 and Young's moduli E_1 and E_2, are shown in Fig. 1. The contact zone has width $2a$. The elastic compression within the contact zone is denoted by d, i.e.

$$d_1 = w_1(0) - w_1(a) = a^2/2R_1$$

and

$$d_2 = w_2(0) - w_2(a) = a^2/2R_2$$

Thus we may write:

$$\frac{d_1}{a} + \frac{d_2}{a} = \frac{a}{2}\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \equiv \frac{a}{2}\left(\frac{1}{R}\right) \tag{7}$$

Now the state of *strain* in each solid varies as the ratio d/a and the state of *stress* in each solid, for a given value of a, will vary in proportion to the mean contact pressure $\bar{p}$. Thus we may now write:

$$\frac{\bar{p}}{E_1} + \frac{\bar{p}}{E_2} \propto \frac{d_1}{a} + \frac{d_2}{a} \propto \frac{a}{R}$$

or

$$\bar{p} \propto \frac{a}{R}\Big/\left(\frac{1}{E_1} + \frac{1}{E_2}\right) \propto \frac{aE}{R} \tag{8}$$

where $1/E \equiv 1/E_1 + 1/E_2$. We must now distinguish between different contact geometries:

(a) In the line contact of cylinders, the load per unit axial length P' is given by:

$$P' = 2a\bar{p} \tag{9}$$

whereupon, from equation (8):

$$a \propto \left(\frac{P'R}{E}\right)^{1/2} \qquad \left[\left(\frac{4P'R}{\pi E^*}\right)^{1/2}\right] \tag{10}$$

and

$$\bar{p} \propto \left(\frac{P'E}{R}\right)^{1/2} \qquad \left[\frac{\pi}{4}\left(\frac{P'E^*}{\pi R}\right)^{1/2}\right] \tag{11}$$

(b) In the point contact of solids of revolution:

$$P = \pi a^2 \bar{p} \tag{12}$$

Thus, from equation (8)

$$a \propto \left(\frac{PR}{E}\right)^{1/3} \qquad \left[\left(\frac{3}{4}\frac{PR}{E^*}\right)^{1/3}\right] \tag{13}$$

and

$$\bar{p} \propto \left(\frac{PE^2}{R^2}\right)^{1/3} \qquad \left[\frac{2}{3}\left(\frac{6}{\pi^3}\frac{PE^{*2}}{R^2}\right)^{1/3}\right] \tag{14}$$

In this case the elastic approach δ is proportional to $d_1 + d_2$, hence:

$$\delta \propto \left(\frac{P^2}{E^2R}\right)^{1/3} \qquad \left[\left(\frac{9}{16}\frac{P^2}{E^{*2}R}\right)^{1/3}\right] \tag{15}$$

The constants of proportionality, of course, can only be found from the complete theory. It transpires that contact deformation is governed by the plane strain modulus $E/(1-\nu^2)$, where ν = Poisson's ratio. We must, therefore, replace E by E^* where $1/E^* = (1-\nu_1^2)/E_1 + (1-\nu_2^2)/E_2$. The Hertz results are shown in square brackets for comparison.

The equivalent expressions for an elliptical contact area are clumsy and involve elliptic integrals whose values must be found from tables. Most textbooks, e.g. (**7**), follow Hertz' method. An alternative procedure suggested by Greenwood, which enables approximate values of the contact shape and size, etc., to be found without recourse to tabulated data, is given in Appendix 1. The relative radii of curvature R' and R'' of the surfaces in contact, as expressed in equation (1), are calculated from the geometry of the individual surfaces in the usual way. We now introduce an 'equivalent relative radius of curvature' $R_e \equiv \sqrt{(R'R'')}$, and an equivalent contact radius $c \equiv \sqrt{(ab)}$, whereupon we can write:

$$c \equiv (ab)^{1/2} = \left(\frac{3PR_e}{4E^*}\right)^{1/3} \cdot F_1(R'/R'') \tag{16}$$

$$p_0 = \frac{3P}{2\pi ab} = \left(\frac{6PE^{*2}}{\pi^3 R_e^2}\right)^{1/3} \cdot [F_1(R'/R'')]^{-2/3} \tag{17}$$

and

$$\delta = \left(\frac{9P^2}{16E^{*2}R_e}\right)^{1/3} \cdot F_2(R'/R'') \tag{18}$$

The leading term in these expressions will be recognized as applying to a sphere of radius R_e in contact with a plane; the second term may be regarded as a correction factor for the ellipticity of the contact. The correction factors are plotted in Fig. 3. Provided that the ratio R'/R'' is not too great the correction factors may be taken to be unity. The shape of the contact ellipse may be expressed:

$$b/a = (R''/R')^{1/2} \cdot F_0(R'/R'') \tag{19}$$

As a first approximation the shape may be taken to be the same as that of a contour of equal separation, i.e. $F_0(R'/R'')$ taken to be unity. As observed by Hertz, the ellipse of contact is thinner than the contours of

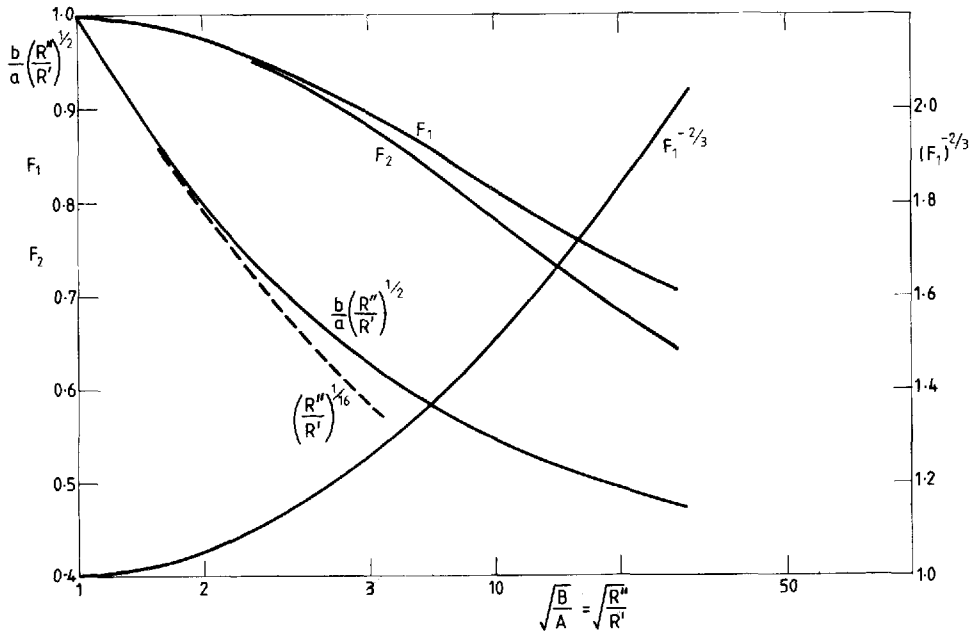

Fig. 3 Elliptical contact functions. Note that all the functions approach unity when $R''/R' \to 1$

separation before deformation, and a better approximation is obtained (see Fig. 3) from:

$$b/a \cong (R''/R')^{2/3} \tag{20}$$

The individual values of a and b are then found from equations (16) and (19).

3 EARLY DEVELOPMENTS: THE INTERNAL STRESS FIELD

Three areas of mechanical engineering in which the Hertz theory found an early application were railways, rolling bearings and toothed gearing. They are still important today. In each case the loads which can be safely carried are governed by the resistance of the material to contact stress particularly under conditions of repeated loading. For a proper appraisal of this problem a knowledge of the stresses beneath the surface is required. In principle the stress components at any point in one of the bodies can be found directly from the potential functions found by Hertz. In practice this requires the evaluation of integrals whose lower limit is the positive root of a cubic equation. In the pre-computer age the recognized procedure was to transform the integrals into a combination of standard tabulated elliptic integrals; not a trivial exercise! Hertz himself only evaluated the stresses at the contact surface and presented a speculative sketch of the subsurface stresses based on his knowledge of the stress field beneath a concentrated load.

Not surprisingly the axi-symmetric case of the contact of solids of revolution was attacked first. Huber (**8**) in 1904 obtained simple expressions for the stress components at the surface, both inside and outside the contact circle, and also for the stresses along the axis of symmetry (z-axis), viz:

$$\sigma_r = \sigma_\theta = -p_0[(1+\nu)\{1 - (z/a)\tan^{-1}(a/z)\} + \tfrac{1}{2}(1 + z^2/a^2)^{-1}] \tag{20a}$$

$$\sigma_z = -p_0(1 + z^2/a^2)^{-1} \tag{20b}$$

These are principal stresses so that the principal shear stress $\tau_1 = \frac{1}{2}|\sigma_r - \sigma_z|$. Its maximum value lies below the surface and has the value $0.31p_0$ at a depth $z = 0.57a$ (taking $\nu = 0.3$). The line contact of cylinders was investigated in 1914 (**9**) with similar conclusions.

The stresses beneath a general elliptical contact were first analysed by Professor N. M. Belajef, presumably stimulated by the railways, since the results were published in the *Bulletin of the Institute of Engineers of Ways of Communication*, in St Petersburg in 1917 and 1929, and introduced to English speaking readers by Timoshenko (**7**). In the US an investigation of transverse fissures in rails led to the work of Thomas and Hoersch (**10**) who obtained the stress components on the axis of symmetry for varying eccentricities of ellipse. The magnitude of the maximum shear stress hardly changes with the shape of the ellipse, but the depth at which it occurs falls from $0.57a$ with a circular point contact to $0.78a$ in line contact.

This work coincided with a growing realization that the yield point of a ductile metal is governed by shear stress; either the absolute maximum (Tresca criterion) or the octahedral shear stress (von Mises criterion). Hertz was under the impression that plastic flow would initiate in the centre of the contact area at the point of maximum compressive stress and would lead to a detectable increase in the contact area. Dissatisfied with the qualitative comparative scales of hardness, based on the ability of one material to scratch another, he proposed that hardness should be defined as that pressure, exerted by an axi-symmetric indenter, which would just cause a specimen of the material to reach its yield point. This is the only feature of Hertz' papers on 'contact' which has failed the '100 year survival' test. The fact that the material elements which yield first are confined to a small volume beneath the surface, fully contained by elastic material, ensures that any measurable quantities, such as the contact area, depart very gradually from their purely elastic values. A fascinating account of an attempt to detect yield by careful optical observations of the first occurrence of a permanent indentation in the surface is given by Davies

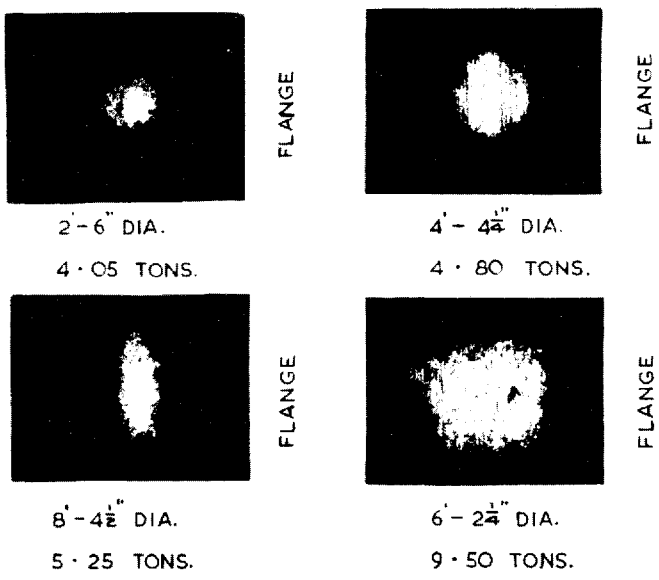

Fig. 4 Measurements of the contact areas between a locomotive driving wheel and a rail [from Andrews (**13**)]
Left Slightly worn tyre
Right Freshly turned tyre

(**11**). The experimental difficulties in detecting reliably the point of first yield made Hertz' suggestion impractical. Hardness continued to be measured by making fully plastic indentations in the surface of the specimen. The derivation by Tabor (**12**) of a rational relationship between hardness H and yield stress Y ($H \cong 3Y$) had to wait for the development of the theory of plasticity.

Before leaving the question of plastic yield it is instructive to consider the expressions for the load P_Y at which yield is first reached beneath the surface. For a circular point contact:

$$P_Y = 21\, R^2 Y^3 / E^{*2} \tag{21a}$$

and for a line contact:

$$P'_Y = 8.8\, RY^2 / E^* \tag{21b}$$

These expressions show the importance of yield stress, i.e. hardness, in carrying contact loads without yield. They also show the benefit of a low elastic modulus; in fact the ratio of yield stress to modulus (Y/E) is the important material property.

Railway engineers have an old working rule:

$$P/D = \text{constant}$$

which relates the safe working axle load P to the diameter of the wheels D. At first sight this rule appears to contradict equation (21a) where P_Y varies as R^2. However, using the concept of an equivalent relative radius of curvature $R_e = \sqrt{(R'/R'')}$, introduced above, we note that $R' = D/2$, whereas R'' depends on the transverse profiles of both wheel and rail and is independent of wheel diameter, so that $R_e \propto \sqrt{D}$ and $P_Y \propto D$.

The contact area between full scale wheels and rails with varying loads and wheel diameters has been measured by Andrews (**13**). Elastic contact areas are notoriously difficult to measure†; in this case a sheet of paper plus a sheet of carbon paper were placed between the wheel and the rail (see Fig. 4). The trend of the results followed the Hertz theory, but the measured area exceeded the theoretical, particularly with freshly turned wheels. The discrepancy was attributed mainly to the effect of surface roughness. This aspect will be considered in Section 8.

The load capacity of rolling bearings is limited by fatigue life rather than by plastic deformation, and, until the recent introduction of vacuum melted steel, the fatigue cracks were observed to initiate at subsurface inclusions. This prompted calculations of sub-surface stresses by Professor Lundberg at Gothenburg (**14, 15**) in collaboration with the SKF Company, in particular the orthogonal shear stress which oscillates between equal positive and negative maximum values on either side of the axis of symmetry. The literature on rolling contact fatigue contains a debate as to whether the orthogonal shear stress which reverses sign, or the maximum shear stress on the centre line which does not, is the stress governing the fatigue life. The relative magnitude of these two stresses changes with the eccentricity of the contact ellipse. Palmgren of SKF favoured the orthogonal shear stress, a choice which is generally supported by modern fatigue research, and on this assumption he developed a theory of bearing life which still forms the basis of rolling bearing design round the world (**16**).

4 PECULIARITIES OF LINE CONTACTS

Application of the Hertz theory to practical line contacts leads to special difficulties not encountered in general three-dimensional contacts. First there is the difficulty of calculating the elastic approach of the two bodies and second there is the question of the end effects in a practical line contact of finite length.

In a point contact the magnitude of the strains in the two bodies decays with distance r from the contact point as r^{-2}. By integration, therefore, the elastic displacements decay as r^{-1}, so that the displacement of the surface relative to a distant point in the body ($r \to \infty$) is finite and leads to equation (15) for the approach δ of distant points. In a two-dimensional contact the strains decay as r^{-1} and the displacements as $\ln(r)$ so that the displacement of the surface increases without limit as the datum for displacements is taken deeper into the solid. The displacement of the surface of an elastic half-space, at the centre (origin) of a two-dimensional Hertz pressure, relative to a point on the z-axis at a depth d is given by:

$$w(0) = P'[(1 - \nu^2)/\pi E][2 \ln(d/a) - \nu/(1 - \nu)] \tag{22}$$

As d increases without limit, so does $w(0)$. Some engineers have greeted this fact with incredulity but Hertz was clear what it meant: 'This means that it (the displacement) depends not merely on what happens at the place of contact, but also on the shape of the body as a whole; and that its determination no longer forms part of the problem we are dealing with'. Calculation of the deflexion taking into account both local contact and bulk deformation would seem an impossible task in Hertz' day. In our own it would appear to call for 'finite elements', but useful engineering answers can be obtained simply by the use of engineering judgement. Since the datum distance d lies within the logarithm in equation

† With steel surfaces, a useful technique is to plate one surface with a matt film of copper using Stead's reagent. On separation the contact area shows up as a bright reflecting surface.

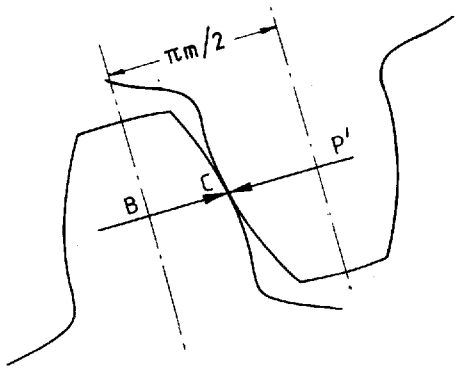

Fig. 5 Contact of gear teeth of module m. Total deflexion is made up of a bending deflexion δ_B and a contact deflexion δ_C

(22), the choice of its value is not critical.† As an example we will consider the deflexion of a pair of involute spur gear teeth (Fig. 5).

The deflexion of each tooth can be regarded as being made up of a deflexion of the centreline of the tooth, bending as a cantilever, and the contact deflexion due to the contact stresses. The combined bending deflexion of a pair of teeth is given by (see Appendix 2):

$$\delta_B = K(P'/E^*) \tag{23}$$

where P' is the load per unit face width and K is a constant ~ 6, independent of the size (i.e. module) of the teeth. To find the contact deflexion it seems reasonable to use equation (22) in which the depth h is taken to be half the width of the tooth, i.e. $h = (\frac{1}{4})\pi m$, where m is the module of the teeth. Thus the combined contact deflexion of both teeth is given by:

$$\delta_c = \frac{P'}{E^*}\frac{\cos\phi}{\pi}\left[\ln\left(\frac{mE^*}{P'}\right) - 1.8\right] \tag{24}$$

The contact deflexion is roughly one-third of the bending deflexion.

A different situation arises when the length of a line contact is limited and the two bodies extend beyond its ends by distances which are significantly greater than the contact width $2a$. This is the case in a roller bearing on account of the rounded ends of the rollers. Neglecting end effects for the moment, the bodies make contact over a narrow rectangle of length $2l$ and width $2a$. This is a three-dimensional rather than a two-dimensional situation so that the contact deflexion is finite. The combined deflexion at the centre of the rectangle has been found by Lundberg (**18**) to be:

$$\delta(0) = (P/\pi lE^*)[\ln(l/a) + 1.886] \tag{25}$$

A much more elaborate procedure for combining local with distant displacements in line contact has been proposed recently by Kalker (**19**).

Line contacts of finite length experience 'end effects' whose nature depends on the elastic moduli of the bodies and their geometry close to the ends of contact. Three different situations are shown diagrammatically in Fig. 6.

(a) Contact is maintained to the edge of the cylinders whose ends are sharp, square and coincident. In this case the pressure falls at the end due to lack of material support at the ends. An estimate of the reduction in pressure may be obtained from the ratio of the plane stress modulus to the plane strain modulus, i.e. $(1 - \nu^2) \cong 0.9$.

(b) If one of the bodies has a sharp square corner and the other continues beyond the end of contact, the contact pressure is theoretically infinite at the end. The form of the singularity varies with the elastic moduli of the contacting bodies (**20**). If the short body is rigid, very close to the end p_0 rises with distance y from the end as $y^{-1/2}$. For bodies having equal elastic contants it varies as $y^{-0.23}$

(c) If one or both bodies are rounded so that there is no discontinuity in surface slope, as in a roller bearing, there will be a finite concentration of pressure close to the end, but the pressure will fall continuously to zero at the end itself. A short distance away from the end ($y > 4a$, say) the transverse pressure profile is approximately Hertzian in which the local maximum $p_0(y)$ is related to the semi-contact width at that point $a(y)$ to good approximation by the Hertz relationship:

$$a/R = 2p_0/E^* \tag{26}$$

General analyses of the axial variation in pressure in line contacts have been made on this basis (**21**, **22**).

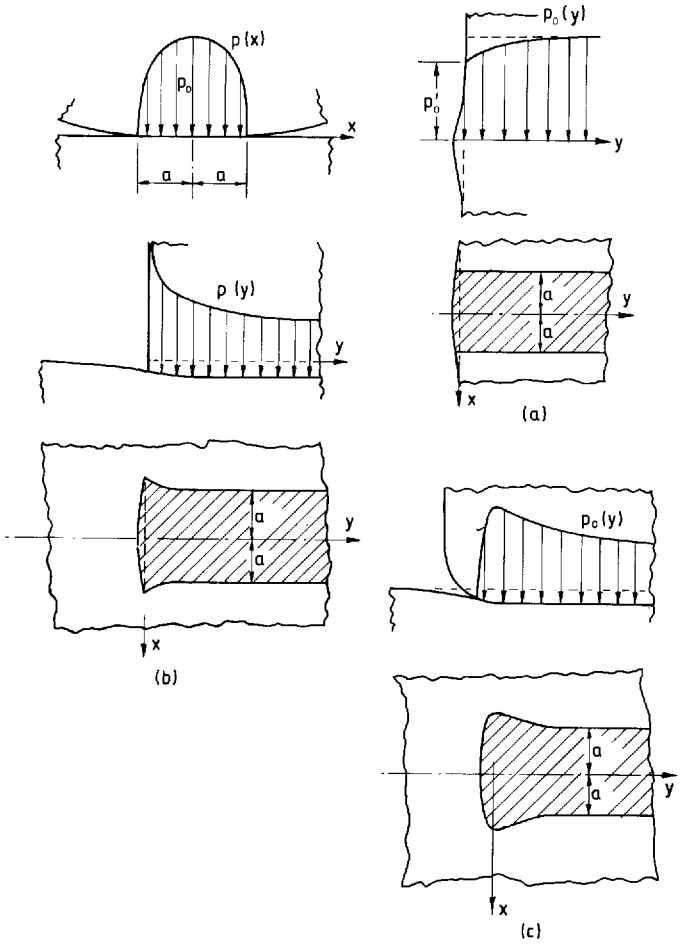

Fig. 6 End effects in line contact
(a) Two square ends
(b) One square end
(c) Two rounded ends

† A review of line contact deflexion formula has been given by Nikpur and Gohar (**17**).

At the ends, however, the stress distribution is truely three-dimensional and a numerical treatment which recognizes the fact is essential to obtain reliable results (**23**).

5 NON-HERTZIAN GEOMETRY: NUMERICAL METHODS

Modern developments in computing have stimulated research into numerical methods of solution of problems in which the contact geometry cannot be described adequately by the quadratic expression of equation (1). The contact of worn wheels and rails, particularly in flange contact, is a case in point (**24a**); so also is the contact of conforming gear teeth (Circarc or Novikov).

In the numerical method the contact area is subdivided into a grid and the pressure distribution represented by discrete 'boundary elements' acting on the elemental areas of the grid. Usually, elements of uniform pressure are employed (Fig. 7a), but overlapping triangular elements (Fig. 7b) offer some advantages. They sum to a piecewise-linear pressure distribution and the fact that the pressure falls to zero at the edge of the contact ensures that the surfaces do not interfere outside the contact area. The three-dimensional equivalent of overlapping triangular elements are overlapping hexagonal pyramids on an equilateral triangular grid.

The influence coefficients $[C_{ij}]$ which express the deflexion w_i at mesh point I due to a unit element of pressure centred at mesh point J are calculated by conventional analytical means, whereupon:

$$w_i = \{(1 - \nu^2)/E\} \sum \{C_{ij} \cdot P_j\} \tag{27}$$

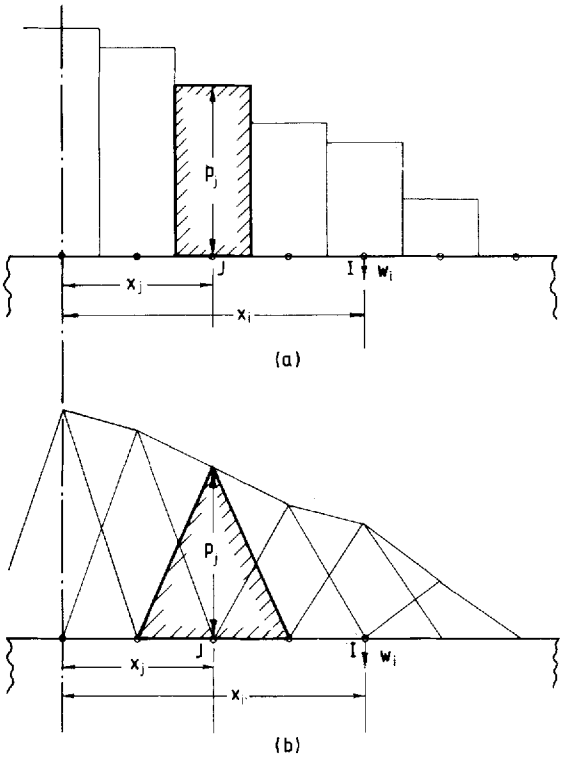

Fig. 7 Boundary elements in the numerical analysis of non-Hertzian contact
(a) Constant pressure elements
(b) Overlapping triangular elements

If $h(x_i, y_i)$ expresses the known separation between the surfaces before deformation, the displacements must satisfy the relationships:

$$(w_i)_1 + (w_i)_2 = (1/E^*) \sum \{C_{ij} P_j\} \begin{cases} = \delta - h(x_i, y_i) & \text{(28a)} \\ > \delta - h(x_i, y_i) & \text{(28b)} \end{cases}$$

where the equality sign applies inside the contact area and the inequality outside. In general the shape and size of the contact area are not known in advance.

Two alternative approaches have been used to determine the contact area and to find the values of p_j which satisfy equation (28). In the first method a contact area is assumed, usually bounded by the 'interpenetration area' (found by putting the left hand side of equation (28a) equal to zero). Equation (28a) is then solved by matrix inversion to find the values of p_j which satisfy it throughout the assumed contact area. Near the edge negative (tensile) values of p_j are found. The pressure is put equal to zero at these mesh points and the equations re-solved. Iteration leads to pressures which are either positive or zero. It is then necessary to check, using equation (28b), that there is no interference in the region of zero pressure. This method has been used by Paul and Hashemi (**24b**) to find wheel-rail contact areas which are far from elliptical in shape.

An alternative method, favoured by Kalker (**25**), is based on the principle that the pressure distribution and area of contact will be such as to minimize the total elastic strain energy, subject to no interference outside the contact and positive pressure inside. Kalker reformulates the principle in terms of complementary energy. The advantage of this approach is that the process of minimizing the energy function subject to the necessary constraints can be carried out by established numerical procedures of quadratic programming. Both methods have been shown to work; their relative merits must turn in the end on the computing time required to achieve comparable precision.

6 FRICTION AT THE INTERFACE

So far we have been considering the problem posed by Hertz: the normal contact of frictionless bodies. Developments of the theory in the second half of the century have been dominated by investigations into the influence of friction at the interface. Under a purely normal load friction affects the contact of dissimilar solids; under tangential loads it influences the contact stresses in ways which are different in static, sliding and rolling contacts. We shall look briefly at each of these situations.

(a) *Sliding contacts* In sliding contacts it is usual to assume that Amonton's law of friction holds, so that the tangential traction:

$$q = \mu p$$

in a direction opposite to the relative motion, at all points in the contact area, where μ is a constant coefficient of friction (Fig. 7a). The plane (two-dimensional) problem with application to the stresses in gear teeth, was the first to be solved completely (**26**, **27***, **28***, **29***). The effect of friction is to bring the point of maximum shear stress closer to the surface. For values of $\mu > 0.30$ the maximum

* Strictly for contacting solids whose elastic properties are similar.

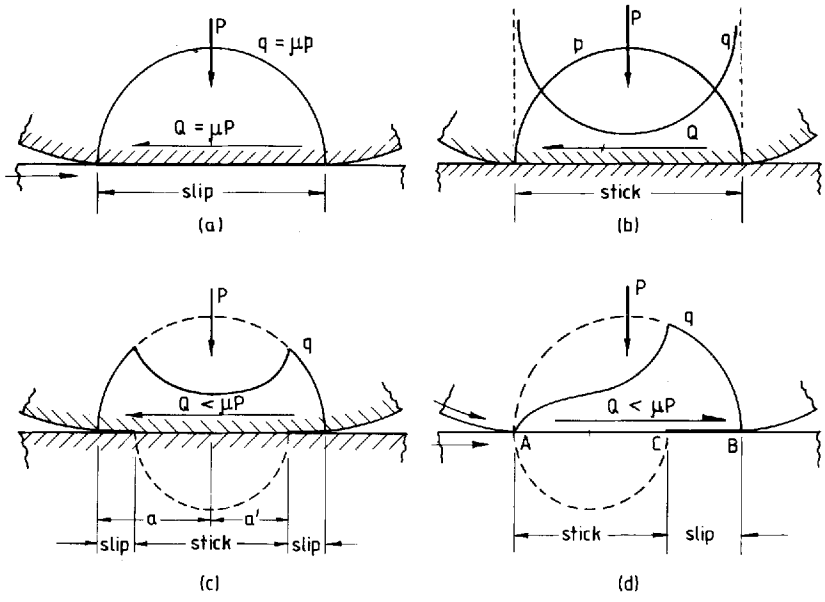

Fig. 8 Distribution of tangential traction
(a) Sliding contact
(b) Static contact with all slip prevented
(c) Static contact with slip at the edges
(d) Rolling contact with slip at the rear

shear stress occurs at the surface and all points within the contact area yield simultaneously. The three-dimensional case is more difficult; the internal stresses due to the combined effect of normal pressure and frictional traction in a sliding circular contact were analysed by Hamilton and Goodman (**30**) in 1966; the corresponding solution for elliptical contacts (**31**, **32**) celebrates the Hertz centenary. For circular contacts first yield was found to occur in tension at the rear edge of the contact area again, when μ exceeds 0.30.

(b) *Static contacts* Tangential forces less than the limiting friction force (i.e. $Q < \mu P$) can be applied to two bodies in contact without equilibrium being disturbed by the initiation of sliding. The absence of sliding, however, does not imply that there should be no slip over at least part of the contact area. This problem was first studied by Cattaneo (**33**) in 1938 and in greater detail by Mindlin and his colleagues (**34**). Cattaneo showed that, under a constant normal load, P, and in the absence of slip, a tangential load would give rise to an infinite tangential traction at the boundary of the contact ellipse† (Fig. 8b). Thus some slip is inevitable at the edge of the contact area under the action of the smallest tangential force. This annular zone of 'microslip' penetrates further into the contact with increasing Q until, when $Q = \mu P$ the whole contact slides. The zone of no-slip is a centrally placed ellipse, having semi-axes a' and b' in proportion to the semi-axes of the contact ellipse a and b. The tangential traction in the slip annulus is given by:

$$q = \mu p = \mu p_0[1 - (x/a)^2 - (y/b)^2]^{1/2} \tag{30}$$

In the no-slip region it is given by the superposition of equation (30) and a similar negative traction reduced in magnitude by a'/a (see Fig. 8c) i.e.

$$q = \mu p_0[\{1 - (x/a)^2 - (y/b)^2\}^{1/2} - (a'/a)\{1 - (x/a')^2 - (y/b')^2\}^{1/2}] \tag{31}$$

The size of the no-slip region is related to the tangential force by:

$$a'/a = (1 - Q/\mu P)^{1/3}. \tag{32}$$

Being a dissipative process, micro-slip has the effect of making the state of contact stress dependent upon the history of loading. Mindlin *et al* (**35**, **36**) examined several sequences of loading which involved simultaneous variations of normal and tangential forces (within limiting friction), in particular, an oscillating oblique force of amplitude $\pm P^*$. They showed that no micro-slip occurs if the angle α between the oscillating force and the common normal to the contact surfaces is less than the angle of friction ($\tan^{-1}\mu$). They calculated the energy dissipated per cycle in slip, ΔW, and showed that ΔW increased with the inclination α of the force. The incentive for this work arose from a study of the transmission of elastic waves though granular materials (**37**) as applied to the carbon microphone. It has also revealed the mechanism of vibration damping in clamped joints. If the size of the slip region increases with the force amplitude, as in Mindlin's problem, then it follows that the energy dissipated per cycle is given by (**38**):

$$\Delta W \propto P^{*3} \tag{33}$$

This result is in contrast with *linear* hysteretic damping in which ΔW is proportional to P^{*2}. Mindlin's work has also thrown light on the mechanics of fretting (**39**). The increase in severity of fretting with the obliquity α of an oscillating force transmitted between a sphere and flat surface in contact is illustrated in Fig. 9 (**40**).

† This stress distribution is the tangential analogue of the normal pressure under a rigid punch given by equation (3).

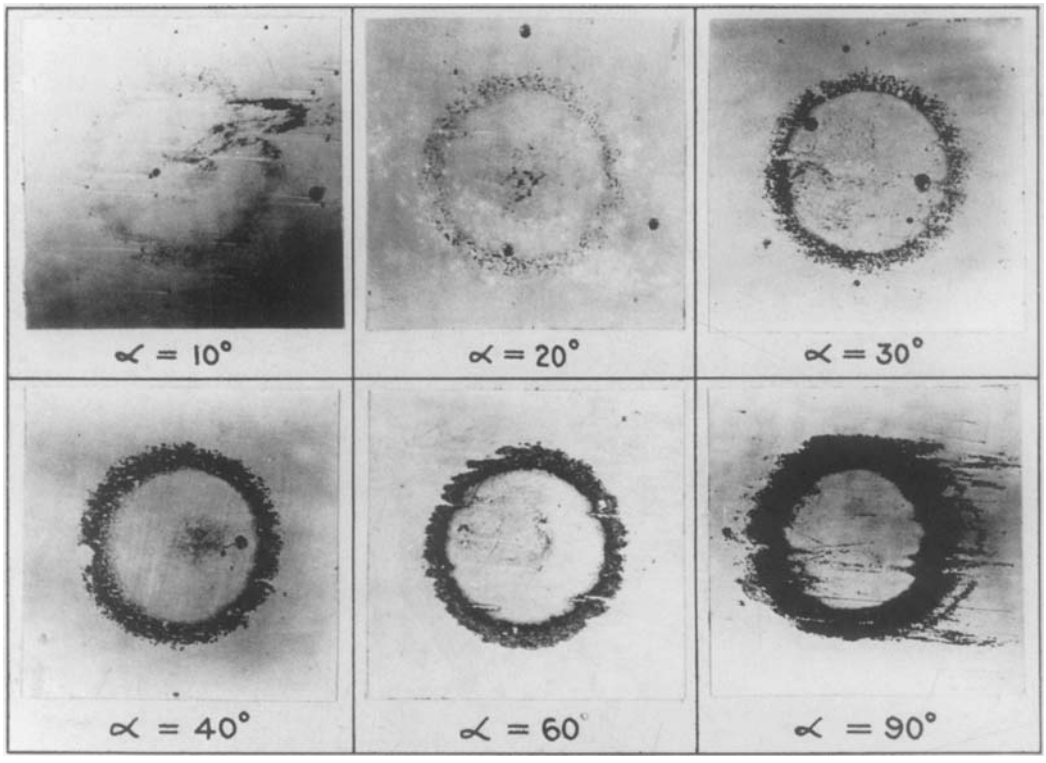

Fig. 9 Annuli of slip and fretting at the contact of a sphere with a flat produced by an oscillating oblique force [from Johnson (**40**)]

(c) *Rolling contacts* Frictional effects also arise in rolling contacts. A driving wheel, or a wheel to which the brakes are applied, exerts a tangential force in the direction (longitudinal) in which it rolls. When rounding a corner transverse tangential forces will normally be brought into play and also a moment about the normal axis due to the angular velocity of the wheel about that axis (spin). As with static contacts, tangential forces give rise to micro-slip and cause the contact area to be divided into regions of slip and no-slip. In rolling, however, the no-slip region is no longer centrally located but is adjacent to the leading edge of the contact. The first analysis of this problem, for line contact, was carried out in 1926 by a British railway engineer, F. W. Carter (**41**). The case of a driving wheel is shown in Fig. 8d. The region of no-slip AC is at the leading edge of the contact; the region of micro-slip CB is towards the rear. The tangential force Q, acting in opposite directions on each surface, puts the material of the wheel in the no-slip zone into circumferential compression and that of the rail into tension. Thus the wheel behaves as though its circumference were reduced and the rail as though its length were increased and, in consequence, the wheel rolls forward in one revolution a distance slightly less than its undeformed perimeter. This phenomenon is known as rolling 'creep' and the fractional difference between the peripheral velocity of the wheel and its forward speed is known as the 'creep ratio' or 'creepage'. The relationship between the tangential force Q and the creepage ξ—the creep curve—is shown in Fig. 10. At small creepages ($Q \ll \mu P$) the relationship is linear and the gradient of the curve is referred to as the 'creep coefficient' f. Creep behaviour of general point contacts has been studied exhaustively by Kalker (**42**, **43**). Longitudinal and lateral forces and creepages are denoted by the suffices x and y respectively, whereupon:

$$Q_x/\mu P = f_{11}\xi_x \tag{34a}$$

and

$$Q_y/\mu P = f_{22}\xi_y + f_{23}\omega\sqrt{(ab)}/U \tag{34b}$$

where ω is the angular velocity of spin, U is the forward rolling velocity and $f_{11}(a/b)$, $f_{22}(a/b)$, $f_{23}(a/b)$ are creep coefficients, calculated by Kalker for different eccentricities of the ellipse of contact. The creep relations given by equations (34) enter the equations of motion of a railway bogie and thereby influence the conditions of

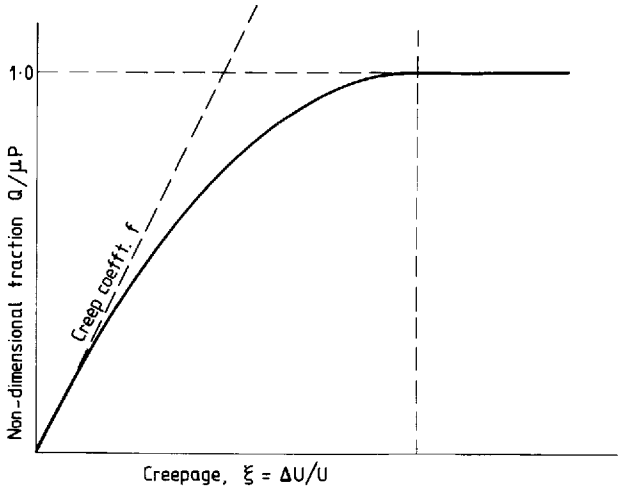

Fig. 10 Rolling contact creep curve due to Carter (**41**)

lateral stability of high-speed railway vehicles. This aspect of contact mechanics has played an important role in the development of high speed trains (**44**).

(d) *Dissimilar materials* Elastic bodies in normal contact undergo tangential as well as normal displacements at the contact surface. If the materials of the two bodies are different, their tangential displacements differ in proportion to the parameter:

$$\beta \equiv \frac{1}{2}\frac{[(1-2\nu_1)/G_1] - [(1-2\nu_2)/G_2]}{[(1-\nu_1)/G_1] + [(1-\nu_2)/G_2]} \tag{35}$$

where G_1 and G_2 are the shear moduli of the two bodies. Some typical values of β for common pairs of materials are given in Table 1. Friction at the interface acts to inhibit such relative tangential displacements.

Table 1

Body 1	Body 2	G_1, GPa	ν_1	G_2, GPa	ν_2	β
Rubber	Metal	small c.f. G_2	0.50	∞ c.f. G_1	—	0
Perspex	Steel	0.97	0.38	80	0.30	0.19
Glass	Steel	22	0.25	80	0.30	0.21
Dural	Steel	28	0.32	80	0.30	0.12
Cast iron	Steel	45	0.25	80	0.30	0.12
Tungsten carbide	Steel	220	0.22	80	0.30	−0.14

The normal contact of dissimilar spherical bodies with friction has been studied by Spence (**45**). Relative motion is prevented by friction in a central circular region, but, at the edge of the contact, slip is inevitable. The width of the annulus of slip is governed approximately by the ratio (β/μ). For realistic values of β and μ the frictional traction is an order of magnitude smaller than the normal pressure; its influence on the intensity of contact stress is therefore small and in general may be neglected. In one respect, however, its effect is significant; it has a major influence on the magnitude of the radial tensile stress just outside the edge of a circular contact—the stress which leads to fracture of brittle materials. This component of Hertz stress is reduced on the more compliant surface and increased on the more rigid one. In this way the fracture load for specimens of plate glass, when compressed by steel balls, was found to be 40 per cent higher than when glass balls were used (**46**). In the case of glass balls, of course, the contacting solids have the same elastic moduli ($\beta = 0$) and no frictional tractions are introduced. The increasing use of ceramics in engineering has led to the strength of brittle and semi-brittle materials under contact and impact loading becoming an active area of current research.

Slip at the interface between dissimilar rolling bodies was recognized by Osborne Reynolds (**47**) in 1875 (seven years before Hertz' paper) as a possible source of rolling friction. He demonstrated 'slip' (strictly 'creep') between an 'india-rubber' cylinder rolling on an 'iron plane' by showing that the roller traversed more than its undeformed circumference in one revolution. This he explained by saying that radial compression of the roller in the contact zone is accompanied by circumferential extension. With the hindsight provided by Hertz, we object that the circumferential strain in a Hertz contact is *compressive* rather than tensile, except with an incompressible solid such as rubber, in which case it is zero. For a rubber wheel in contact with a rigid plane equation (35) gives $\beta = 0$, so that no creep would be expected! The paradox is resolved when it is realized that Reynolds did not use a solid rubber wheel, but a relatively rigid cylinder covered with a thin rubber tyre, in which the circumferential strain would be expected to be tensile. Nearly a century had to elapse before a complete theory of rolling contact of dissimilar solids appeared in 1967 (**48**). The behaviour is quite complex: three regions of micro-slip occur separated by two regions of no-slip. However the rolling resistance due to micro-slip was found to be small compared with that due to the hysteresis of most materials.

7 ELASTIC IMPACT

In his original paper (**1**) Hertz applied his theory to the impact of elastic bodies and showed that the time of impact (i.e. the time when the bodies are in contact) is given by:

$$T_c = 2.94\,\delta^*/V \tag{36}$$

where δ^* is the maximum compression of the two bodies and V is the velocity of impact. For a uniform sphere of radius R striking a flat surface of the same material the time of impact may be expressed by:

$$T_c = 4.5\,(R/c_0)(V/c_0)^{1/5} \tag{37}$$

where c_0 is the longitudinal elastic wave speed in the material $(E/\rho)^{1/2}$. The theory is 'quasi-static' in the sense that elastic wave motion in the bodies is neglected and the static force-compression law, expressed by equation (15), is assumed to apply throughout the impact. This assumption amounts to neglecting the mass of material involved in local deformation close to the area of contact and assuming that the whole of each body is moving as though rigid with the velocity of its centre of mass. Hertz states that this assumption is justified if the time taken for a stress wave to travel the diameter of the body and back, i.e. $\sim 4R/c_0$, is small compared with the total time of impact, thereby maintaining quasi-static equilibrium. Equation (36) shows that this condition requires the velocity of impact V to be small compared with the wave speed c_0,† In most circumstances, however, the Hertz theory of impact is restricted to low velocities by the necessity of remaining within the elastic limit. When a uniform sphere of density ρ strikes a flat massive body, the velocity to initiate yield V_Y is given by

$$\rho V_Y^2/Y = 85(Y/E^*)^4 \tag{38}$$

where Y is the yield stress of the softer body. For a steel in which $Y = 1000$ N/mm^2, $V_Y \cong 0.8$ m/s compared with $c_0 = 5200$ m/s.

The Hertz theory is for frictionless surfaces; it has nothing to say, therefore, about the influence of friction upon the bounce of a ball, which plays such a significant role in the world of sport. The oblique impact with a flat surface of a ball, spinning about an axis perpendicular to

† There is a paradox in this argument since, if one of the bodies is very large, no waves are reflected back to the point of impact. Hunter (**49**) has given a more satisfactory demonstration that the energy absorbed in wave motion in a Hertz impact is small.

its plane of motion (backspin positive), is shown in Fig. 14. The tangential peripheral velocity of the ball at the point of contact is given by:

$$w = u + \omega R \tag{39}$$

where u is the tangential velocity of its centre and ω is the angular velocity of spin. The velocity w will be opposed by friction. If the coefficient of friction is sufficiently high, the point of contact will be brought to rest and then rebound motion will be given by:

$$w_2 = 0 = u_2 + \omega_2 R \tag{40a}$$

where suffices 1 and 2 refer to velocities before and after impact respectively. However, if slip occurs throughout the impact so that $w_2 > 0$, then it is shown in Appendix 3 that:

$$w_2 = w_1 - 7\mu v_1 \tag{40b}$$

provided $w_1 > 7\mu v_1$ in which v_1 is the normal incident velocity.

This simple treatment can be demonstrated to be false by the bounce of a 'superball', a rubber ball which combines a high coefficient of restitution with a high coefficient of friction. When projected obliquely onto a flat surface with back-spin, it is observed to bounce back in the direction from which it came with its angular velocity reversed, i.e. both u_2 and ω_2 are negative. It, therefore, approaches the second bounce like the first, with back-spin and the sequence is repeated, the ball bouncing backwards and forwards as shown in Fig. 11a. Now the change of sign of both u and ω on impact implies that w_2 must be negative, whereas the simple 'rigid body' analysis above results in w_2 being zero or positive (equations 40a and b).

The discrepancy is due to neglect of the *tangential* elastic deformation in the contact zone. During the impact, as the contact area grows to a maximum and subsequently shrinks to zero, the tangential friction force causes shear deformations in the contact zone of the form analysed by Mindlin and Deresiewicz (**35**) and discussed above. If gross sliding does not occur, the contact area will consist of a central region of no-slip surrounded by an annulus of micro-slip. Recently Maw *et al* (**50, 51**) have analysed oblique impact, using the methods of Mindlin and Deresiewicz, to provide a rational extension of the Hertz theory to oblique impact with friction. They obtain a non-dimensional relationship between the peripheral velocity on rebound w_2 and that on impact w_1 which is shown in Fig. 11b. We see that negative values of w_2 are predicted which are consistent with the observed behaviour of elastic bodies such as a 'superball'.

8 ROUGH SURFACES

Up to now we have discussed the contact of elastic bodies as though they made intimate contact at all points throughout the nominal contact area. We know that this is not so. The inevitable topographical roughness of real surfaces ensures that they make true contact over an area which is generally small compared with the apparent area of contact. The real contact pressure at the true contact spots will be much higher than the pressure given by the

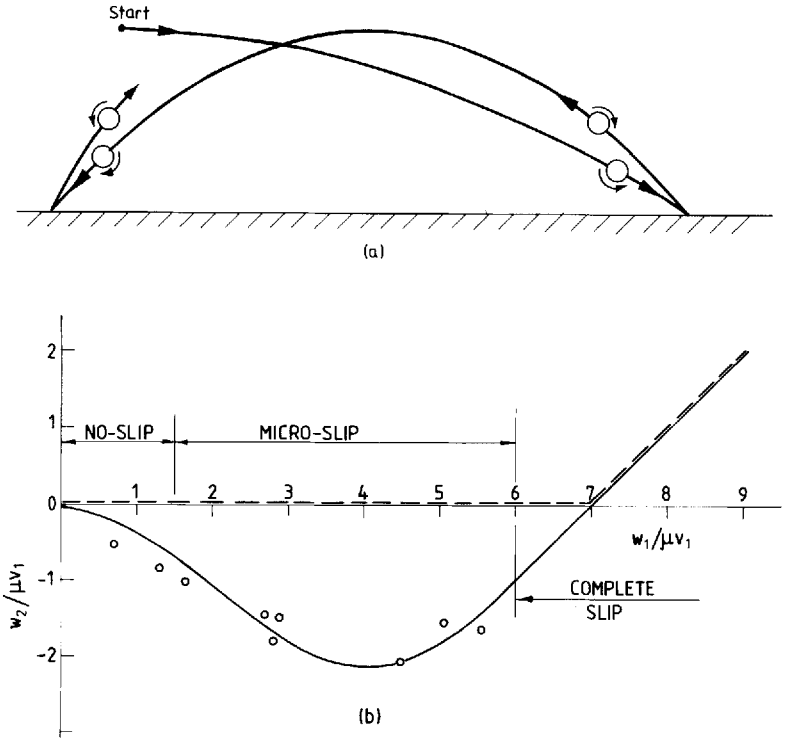

Fig. 11 Oblique impact with friction
(a) The bounce of a superball with back-spin; note that the tangential component of velocity and the angular velocity of spin are reversed on rebound
(b) Relationship between rebound and incident velocities in the oblique impact of spheres [from Maw, Barber and Fawcett (**50**)], experimental measurements of rebound of a superball

Hertz theory applied to the contact as a whole; in between the true contact spots the pressure must fall to zero. We are obliged to ask whether, and in what circumstances, it is justifiable to use a smooth surface theory such as that of Hertz.

There has been considerable effort in recent years to apply statistical methods to quantify the contact of randomly rough surfaces. Various approaches have been adopted but it seems appropriate in the context of this paper to refer to Greenwood and Williamson (**52**) who model a rough surface by spherically tipped asperities of random height which, when brought into contact with a mating surface, themselves deform according to the Hertz theory (equation 14). It is found that, when a rigid flat surface is pressed into contact with a nominally flat rough surface which is modelled in this way, the number of real contact spots and the total real area of contact both grow in direct proportion to the normal load. These predictions are consistent (a) with Amonton's law of friction, and (b) with the experimental observations that the electrical and thermal conductance of nominally flat rough surfaces increase in direct proportion to the load.

We are concerned here, however, with the contact of curved rough surfaces. The contact of a smooth sphere with a nominally flat random rough surface has been considered by Greenwood and Tripp (**53**) and in a simpler way by Johnson (**54**). It is assumed in the statistical treatment that there are a large number of asperity contacts within the nominal contact area. The asperities then act like a compliant layer between the bulk of the two solids. The deformation is found to be governed mainly by the non-dimensional parameter:

$$\alpha \equiv \frac{\sigma}{\delta_0} = \frac{\sigma R}{a_0^2} = \sigma\left(\frac{16E^*R}{9P}\right)^{1/3} \tag{41}$$

where σ is the standard deviation of the surface roughness;† δ_0 and a_0 are the Hertzian (i.e. smooth surface) values of the compression and nominal contact radius under the actual applied load P. The 'smoothed out' pressure distributions for two different values of α are shown in Fig. 12. When α is large the deformation is almost entirely confined to the asperities; the maximum pressure is much less than the Hertz value and contact spreads over an area much greater than the Hertz circle. As α is reduced, more of the deformation occurs in the bulk of the solids and less in the asperities. The Hertz results for a smooth surface are approached as α approaches zero. It would appear that errors in the contact radius a and the maximum pressure p_0 by using the Hertz theory are likely to be less than 10 per cent if the parameter α is less than ~ 0.05. If surface roughness effects are to be correctly modelled in a laboratory test rig, a disc machine for example, the necessity of maintaining this non-dimensional parameter comparable with its full-scale value should be kept in mind.

In view of the statistical nature of surface roughness the smoothed out pressure falls asymptotically to zero and the contact radius cannot be precisely defined. Real contact will be confined to an archipelago of small islands surrounding the region of more intense contact. This state of affairs is illustrated by the observations of wheel/rail contacts shown in Fig. 4.

† In a practical application σ can be taken to be the combined (r.m.s.) value of the roughness (e.g. c.l.a.) of both surfaces.

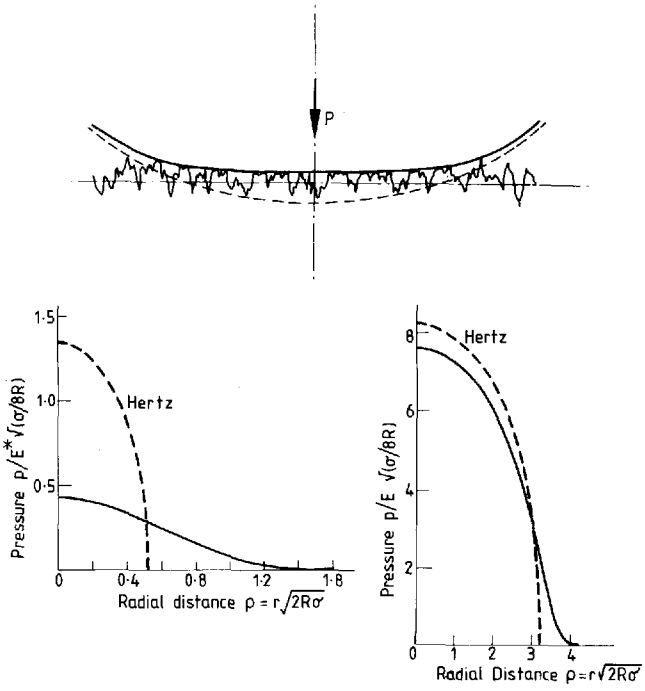

Fig. 12 Contact of a smooth sphere with a nominally flat rough surface
(a) Very rough surface, $\alpha = 1.85$
(b) Slightly rough surface, $\alpha = 0.047$

9 ADHESION BETWEEN ELASTIC BODIES

Two ideally flat clean surfaces, when pressed into intimate contact, should adhere through the action of intermolecular forces of attraction. Such forces are large in magnitude but have a range of action of only a few nanometers. The work which must be done to separate unit area of two adhering surfaces is 2γ where γ is known as the 'surface energy' of each surface. Its magnitude for metals is ~ 0.2 J/m^2 and for van de Waals solids such as polymers or rubber is ~ 0.04 J/m^2. Real surfaces do not exhibit such adhesion and fall apart when the force pressing them into contact is removed on account of their surface roughness. True contact only occurs at the tips of the asperities and, during unloading, adhesion at the lower points of real contact is broken by the compression which is exerted by the higher asperities. However, a different situation arises when one or both surfaces is very compliant such as gelatine or low-modulus rubber. It was discovered, during an investigation into the friction of automobile windscreen wipers, that a smooth hemispherical rubber slider would adhere strongly to a flat glass or perspex plate. When pressed into contact the radius of the contact circle exceeded the value predicted by Hertz; when the load was removed a repeatedly measurable contact area remained, and it was necessary to apply a tensile force to separate the two surfaces. This phenomenon became the subject of a study by Johnson, Kendall and Roberts (**55**).

We saw in the Introduction that the condition of contact within the contact area (equation 2a) is not uniquely satisfied by the Hertz pressure distribution (equation 4); a distribution of the form of equation (3) can be added or subtracted without violating equation (2a). Subtraction of equation (3) was originally ruled out on the grounds that the interface could not withstand

tension (see Fig. 2a). In the presence of adhesive forces this is no longer the case. For the contact of spheres, therefore, we can write:

$$p(r) = p_0[1 - (r/a)^2]^{1/2} - p_0'[1 - (r/a)^2]^{-1/2} \quad (42)$$

where p_0 is given by the Hertz theory (equation 14) and p_0' remains to be determined. This is achieved by considering an energy balance as in the Griffith theory of brittle fracture. It may be shown that the elastic strain energy associated with the pressure distribution of equation (42) is:

$$U_E = \frac{\pi^2 a^3}{E^*}\left(\tfrac{2}{15}p_0^2 - \tfrac{2}{3}p_0 p_0' + p_0'^2\right) \quad (43)$$

The total compression is:

$$\delta = \pi a\,(p_0 - 2p_0')/2E^* \quad (44)$$

In the presence of adhesion the total free energy U_T is $U_E + U_S$ where U_S is the surface energy given by:

$$U_S = -2\gamma\pi a^2 \quad (45)$$

The surface energy U_S is reduced when the surfaces come into intimate contact and is increased when they separate. The equilibrium value of a is given by:

$$\left(\frac{dU_T}{da}\right)_{\delta=\text{constant}} = 0 \quad (46)$$

from which we obtain:

$$p_0' = (4\gamma E^*/\pi a)^{1/2} \quad (47)$$

The net contact force is given by:

$$P = \int_0^a 2\pi r\, p(r)\, dr = (\tfrac{2}{3}p_0 - 2p_0')\,\pi a^2$$

Substituting for p_0 and p_0' and rearranging gives a relationship between P and a, thus:

$$(P - 4E^*a^3/3R)^2 = 16\pi\gamma E^* a^3 \quad (48)$$

This relationship is plotted in Fig. 13, where it is compared with measurements using gelatine. When loaded by a compressive force ($P > 0$) the contact radius a exceeds the Hertz value. When P is reduced to zero the surfaces remain in contact with a value of a given by point C in Fig. 13. The application of a tensile force ($P < 0$) causes the contact to shrink further until, at point B, the system becomes unstable and the contact breaks. The critical load P_c is given by

$$P_c = -3\pi\gamma R \quad (49)$$

This development of Hertz contact theory for adhering elastic solids has thrown light on the mechanism of rubber friction. Similar behaviour would be expected if the adhesion were provided by a specific layer of adhesive, provided that the thickness of the layer were small compared with the elastic compression of the solids.

The infinite tensile stress at the edge of the contact given by equation (42) calls for comment. It is comparable with the infinite tension at the tip of a crack in linear elastic fracture mechanics. In reality the surfaces separate slightly very close to the edge of contact allowing the stress there to fall to the maximum value of intermolecular attraction. The effect of this relaxation upon the strain energy, expressed by equation (43), is negligibly small, so that the resulting equation (48) remains valid to very close approximation.

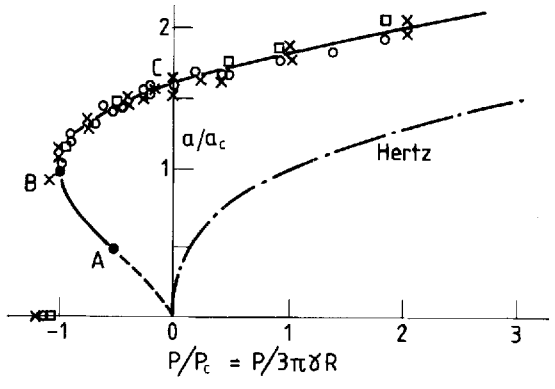

Fig. 13 Contact of spheres with adhesion: $a_c = (9\pi\gamma R^2/4E^*)^{1/3}$, $P_c = 3\pi\gamma R$. Experiments with gelatine spheres of different radii

10 CONCLUSION

In this brief survey emphasis has been placed on those developments which stem directly from Hertz' pioneering work. Inevitably many important aspects of contact mechanics have had to be omitted:

Elastic contact stress analysis of layered solids, such as rubber covered rollers, is now well developed (see Gladwell **56**). The contact of anisotropic solids arises with drawn polymer sheets and fibres (**57**) and also with single crystals. Willis (**58**) has shown that general anisotropic solids also make contact over an elliptical area, although the orientation of the ellipse depends upon the axes of anisotropy of the materials as well as the curvature of the surfaces. The contact of inhomogeneous solids, whose elastic moduli vary with depth (**59**), has been applied to building foundations in soil. The development of 'finite element' methods has opened up the numerical solution of problems of conforming surfaces with extended contact areas (**60**). This development has revealed a class of contact problems—'receding contacts'—in which the contact area decreases with increasing load (**61**). The contact of two bodies whose bulk temperatures are different is influenced by differential thermal expansion in the contact area (**62**).

A discussion of the lubrication of non-conforming contacts—elastohydrodynamic lubrication—has been omitted with regret, but a few remarks may not be out of place. In heavily loaded contacts (where the stress level is of practical importance) the thickness of the film ($\sim 1\ \mu m$) is generally small compared with the elastic compression of the surfaces. The contact stresses are then not very different from those given by the Hertz theory for dry contact. The most significant difference is due to a sharp pressure peak, followed by a rapid drop in pressure at the rear of the contact, which gives rise to a secondary maximum in shear stress close to the surface (**63**). In sliding, the non-Newtonian properties of typical lubricants lead to frictional tractions which are approximately proportional to the normal pressure, with an effective coefficient of friction (i.e. traction coefficient) having a value between 0.05 and 0.10. Finally, stresses at the contact of inelastic solids (**64**) are beyond the scope of the paper.

It is clear that throughout the century since Hertz the stimulus for developments in the theory has come from practical problems. Railway engineering has played an obvious major role, but in some cases the connections have been surprising and unexpected. Mindlin's work on tangential contact sprang from an attempt to understand the mechanics of the carbon microphone; the work of Roberts on rubber adhesion arose from an investigation of windscreen wipers! Study of oblique impact arose out of an investigation of piston slap in internal combustion engines. Hertz himself become involved in contact mechanics through a problem in experimental optics and used a mathematical analogy between elasticity and electrostatics to solve it. In this paper I have concentrated on the intertwined strands of theory and application. For an account of the development of the theory from a mathematical standpoint reference should be made to the book by Gladwell (**56**).

To conclude, the Hertz theory has stood the test of time. The developments outlined in this paper add to the original and extend its scope, particularly in respect of the influence of interfacial friction, but the original theory is in no way supplanted. It remains, as published in 1882, a cornerstone of mechanical engineering.

REFERENCES

(1) HERTZ, H. Uber die Berührung fester elastischer Körper. (On the contact of elastic solids). *J. reine und angewandte Mathematik*, 1882, **92**, 156–171. (For English trans. see *Misc. papers by H. Hertz*, Jones and Schott, Macmillan, London, 1896).

(2) HERTZ, H. Uber die Berührung fester elastische Körper und uber die Härte. (On the contact of rigid elastic solids and on hardness). *Verhandlungen des Vereins zur Beförderung des Gewerbefleisses*, Leipzig, Nov. 1882. (For English trans. see *Misc. papers by H. Hertz*, Jones and Schott, Macmillan, London, 1896).

(3) HERTZ, Johanna. *Heinrich Hertz: memoirs, letters, diaries*, 2nd edition (Mathilde Hertz and Charles Susskind) in English and German, San Francisco Press, Ca, 1977.

(4) FESSLER, H. and OLLERTON, E. Contact stresses in toroids under radial loads. *Brit. J. App. Phys.*, 1957, **8**, 387.

(5) PERSSON, A. On the stress distribution of cylindrical elastic bodies in contact. Doctoral dissertation, Chalmers Univ. of Tech., Gothenburg, 1964.

(6) GOODMAN, L. E. and KEER, L. M. The contact stress problem for an elastic sphere indenting an elastic cavity. *Int. J. Solids & Structures*, 1965, **1**, 407.

(7) TIMOSHENKO, S. P. and GOODIER, J. N. *Theory of elasticity*, McGraw-Hill, 1951.

(8) HUBER, M. T. Zur Theorie der Berührung fester elastische Körper, *Ann. der Phys.*, 1904, **14**, 153.

(9) HUBER, M. T. and FUCHS, S. Spannungsverleitung bei der Berührung zweier elastische Zylinder. *Phys. Zeitschr.*, 1914, **15**, 298.

(10) THOMAS, M. R. and HOERSCH, V. A. Stresses due to the pressure of one elastic solid upon another. Univ. of Illinois Engineering Expt. Sta. Bull. No. 212, 1930.

(11) DAVIES, R. M. The determination of static and dynamic yield stress. *Proc. R. Soc.*, 1949, **A197**, 416.

(12) TABOR, D. *The hardness of metals*, Clarendon, 1951.

(13) ANDREWS, H. I. The contact between a locomotive driving wheel and the rail. *Wear*, 1958–9, **2**, 468.

(14) LUNDBERG, G. and ODQUIST, F. K. G. *Proc. R. Swed. Inst. Engng. Res.*, 1932, No. 116.

(15) LUNDBERG, G. and SJÖVALL, H. Stress and deformation in elastic contacts. Chalmers Univ. of Tech., Gothenburg, 1958.

(16) LUNDBERG, G. and PALMGREN, A. Dynamic capacity of rolling bearings. *Acta Polytechnica*, 1947, No. 7.

(17) NIKPUR, K. and GOHAR, R. Deflection of a roller compressed between platens. *Tribology International*, 1975, **8**, 2.

(18) LUNDBERG, G. Elastiche Berührung zweir Helbraüme. *Forsh. a. d. Geb. das Ing. wesen*, 1939, **10**, 201.

(19) KALKER, J. J. Surface displacement of an elastic half-space in a slender, bounded, curved surface region. *J. Inst. Maths. Applcs.*, 1977, **19**, 127.

(20) DUNDURS, J. and LEE, M-S. Stress concentration at a sharp edge in contact problems. *J. of Elasticity*, 1972, **2**, 109.

(21) NAYAK, L. and JOHNSON, K. L. Pressure between elastic bodies having a slender area of contact and arbitrary profiles. *Int. J. Mech. Sci.*, 1979, **21**, 237.

(22) JOHNS, P. M. and GOHAR, R. Roller bearings under radial and eccentric loads. *Tribology International*, 1981, **14**, 131.

(23) HARTNETT, M. J. and KANNEL, J. Contact stresses between elastic cylinders. *Trans. ASME, J. Lub. Tech.*, 1981, **103**, 40.

(24) PAUL, B. and HASHEMI, J. (a) Contact geometry associated with arbitrary rail and wheel profiles. Symp. on Rolling Contact, AMD-**40**, 93, ASME, New York, 1980.
(b) Contact pressures on closely conforming elastic bodies. Symp. on Solid Contact and Lubrication, AMD-**39**, ASME, New York, 1980.

(25) KALKER, J. J. and van RANDEN, Y. A minimum principle for frictionless elastic contact with application to non-Hertzian half-space contact problems. *J. Engng. Math.*, 1972, **6**, 193.

(26) MUSKHELISHVILI, N. I. *Some basic problems in the mathematical theory of elasticity*, English trans., edited by Radok, Noordhoff, 1953.

(27) McEWEN, E. Stresses in elastic cylinders in contact along a generatrix. *Phil. Mag.*, 1949, **40**, 454.

(28) PORITSKY, H. Stresses and deflexions of cylindrical bodies in contact. *Trans. ASME, J. Appl. Mech.*, 1950, **18**, 191.

(29) SMITH, J. O. and LIU, G. K. Stresses due to tangential and normal loads on an elastic solid. *Trans. ASME, J. Appl. Mech.*, 1953, **21**, 157.

(30) HAMILTON, G. M. and GOODMAN, L. E. The stress field created by a circular sliding contact. *Trans. ASME, J. Appl. Mech.*, 1966, **33**, 371.

(31) BRYANT, M. D. and KEER, L. M. Rough contact between elastically and geometrically similar curved bodies. *Trans. ASME, J. Appl. Mech.*, 1982, **49**, 345.

(32) SACKFIELD, A. and HILLS, D. A. Some useful results in the classical Hertz contact problem. To be published in *J. Strain Analysis*.

(33) CATTANEO, C. Sul contatto di corpi elastici. *Acad. dei Lincei, Rendiconti, Ser. 6*, 1938, **27**, 342–348, 434–436, 474–478.

(34) MINDLIN, R. D. Compliance of elastic bodies in contact. *Trans. ASME, J. Appl. Mech.*, 1949, **17**, 259.

(35) MINDLIN, R. D. and DERESIEWICZ, H. Elastic spheres under varying oblique forces. *Trans. ASME, J. Appl. Mech.*, 1953, **21**, 237.

(36) DERESIEWICZ, H. Oblique contact of non-spherical bodies. *Trans. ASME, J. Appl. Mech.*, 1951, **25**, 623.

(37) DERESIEWICZ, H. Mechanics of granular materials. *Progress in Appl. Mech.*, 1958, **5**, 233.

(38) GOODMAN, L. E. A review of progress in analysis of interfacial slip damping. Proc. ASME Coll. on Structural Damping, Edited by J. E. Ruzicka, Pergamon, 1960.

(39) JOHNSON, K. L. and O'CONNOR, J. J. The mechanics of fretting. *Proc. Instn. Mech. Engrs.*, App. Mech. Convention, Newcastle, 1964.

(40) JOHNSON, K. L. Energy dissipation at spherical surfaces in contact transmitting oscillating forces. *J. Mech. Engng. Sci.*, 1961, **3**, 362.

(41) CARTER, F. W. On the action of a locomotive driving wheel. *Proc. R. Soc. A*, 1926, **112**, 151.

(42) KALKER, J. J. On the rolling contact of two elastic bodies in contact. Doctoral Dissertation, Tech. Univ. Delft, 1967.

(43) KALKER, J. J. A survey of wheel-rail rolling contact theory, *Vehicle System Dynamics*, 1979, **5**, 317.

(44) WICKENS, A. H. and GILCHRIST, A. O. Railway vehicle dynamics—the emergence of a practical theory, CEI MacRobert Award Lecture, 1977.

(45) SPENCE, D. A. The Hertz contact problem with finite friction. *J. of Elasticity*, 1975, **5**, 297.

(46) JOHNSON, K. L., O'CONNOR, J. J. and WOODWARD, A. C. The effect of indenter elasticity on the Hertzian fracture of brittle materials. *Proc. R. Soc. A*, 1973, **334**, 95.

(47) REYNOLDS, O. On rolling friction. *Phil. Trans. R. Soc.*, 1875, **166**, 155.

(48) BENTALL, R. H. and JOHNSON, K. L. Slip in the rolling contact of dissimilar rollers. *Int. J. Mech. Sci.*, 1967, **9**, 389.

(49) HUNTER, S. C. Energy absorbed by elastic waves during impact. *J. Mech. Phys. Solids*, 1956–7, **5**, 162.

(50) MAW, N., BARBER, J. R. and FAWCETT, J. N. The oblique impact of elastic spheres. *Wear*, 1976, **38**, 101.
(51) MAW, N., BARBER, J. R. and FAWCETT, J. N. The role of elastic tangential compliance in oblique impact. *Trans. ASME, J. Lub. Tech.*, 1981, **103**, 74.
(52) GREENWOOD, J. A. and WILLIAMSON, J. B. P. The contact of nominally flat surfaces. *Proc. R. Soc. A*, 1966, **295**, 300.
(53) GREENWOOD, J. A. and TRIPP, J. H. The elastic contact of rough spheres. *Trans. ASME, J. Appl. Mech.*, 1967, **35**, 153.
(54) JOHNSON, K. L. Non-Hertzian contact of elastic spheres. In *The mechanics of the contact between deformable bodies*, Edited by de Pater and Kalker, Delft Univ. Press, 1975.
(55) JOHNSON, K. L., KENDALL, K. and ROBERTS, A. D. Surface energy and the contact of elastic solids. *Proc. R. Soc. A*, 1971, **324**, 301.
(56) GLADWELL, G. M. L. *Contact problems in the classical theory of elasticity*, Sijthoff and Noordhoff, 1980.
(57) PINNOCK, P. K., WARD, I. M. and WOLFE, J. M. The compression of anisotropic fibre filaments. *Proc. R. Soc. A*, 1966, **291**, 267.
(58) WILLIS, J. R. Hertzian contact of anisotropic bodies. *J. Mech. Phys. Solids*, 1966, **14**, 163.
(59) GIBSON, R. E. Some results concerning displacements and stresses in a non-homogeneous elastic half-space. *Z.A.M.P.*, 1967, **17**, 58 and 1969, **19**, 160.
(60) FREDRIKSSON, B. On elastostatic contact problems with friction. Doctoral dissertation No. 6, Univ. of Linköping, 1976.
(61) DUNDURS, J. Properties of elastic bodies in contact. In *Mechanics of contact between deformable bodies*, edited by de Pater and Kalker, Delft Univ. Press, 1975.
(62) BARBER, J. R. Thermoelastic contact problems. In *Mechanics of contact between deformable bodies*, edited by de Pater and Kalker, Delft Univ. Press, 1975.
(63) DOWSON, D., HIGGINSON, G. R. and WHITAKER, A. V. Stress distribution in lubricated rolling contacts. *Proc. Instn. Mech. Engrs.*, Symp. on Rolling Contact Fatigue, London, 1963.
(64) JOHNSON, K. L. Inelastic contact, plastic flow and shakedown. Proc. Int. Symp. on Contact Mech. and Wear of Wheel/rail Systems, Waterloo Univ. Press, 1982.

APPENDIX 1

The gap between two undeformed curved surfaces in contact may be written:

$$h = (1/2R')x^2 + (1/2R'')y^2 \equiv Ax^2 + By^2 \tag{50}$$

where R' and R'' are the principal *relative* radii of curvature. They may be found from the principal radii of curvature of each surface ρ_1', ρ_1'' and ρ_2', ρ_2'' by:

$$A + B = \frac{1}{2}\left(\frac{1}{R'} + \frac{1}{R''}\right) = \frac{1}{2}\left(\frac{1}{\rho_1'} + \frac{1}{\rho_1''} + \frac{1}{\rho_2'} + \frac{1}{\rho_2''}\right) \tag{51a}$$

and

$$B - A = \frac{1}{2}\left[\left(\frac{1}{\rho_1'} + \frac{1}{\rho_1''}\right)^2 + \left(\frac{1}{\rho_2'} + \frac{1}{\rho_2''}\right)^2 + 2\left(\frac{1}{\rho_1'} - \frac{1}{\rho_1''}\right)\left(\frac{1}{\rho_2'} - \frac{1}{\rho_2''}\right)\cos 2\theta\right]^{1/2} \tag{51b}$$

where θ is the angle between the principal axes of curvature of each surface. If the two bodies are now pressed into contact such that distant points in the bodies approach each other by a displacement δ, then, within the contact area, the normal displacements of the two surfaces relative to the distant points are given by:

$$w_1(x, y) + w_2(x, y) = \delta - Ax^2 - By^2 \tag{52}$$

We introduce an equivalent radius R_e defined by:

$$R_e = (R'R'')^{1/2} = \tfrac{1}{2}(AB)^{-1/2} \tag{53}$$

Hertz showed that the pressure distribution:

$$p(x, y) = p_0[1 - (x/a)^2 - (y/b)^2]^{1/2} \tag{54}$$

acting on an elliptical area of semi-axes a and b on the surface of an elastic half-space produces normal surface displacements within that area given by:

$$w(x, y) = \frac{1 - \nu^2}{\pi E}(L - Mx^2 - Ny^2) \tag{55}$$

where L, M and N are functions of the shape of the ellipse as given in equations (56). Writing $1/E^* = [(1 - \nu_1^2)/E_1] + [(1 - \nu_2^2)/E_2]$, where E_1, ν_1 and E_2, ν_2 are Young's moduli and Poisson's ratio of the two bodies, the condition of contact expressed by equation (52) is satisfied provided:

$$A = M/\pi E^* = (p_0/E^*)(b/a^2e^2)[\boldsymbol{K}(e) - \boldsymbol{E}(e)] \tag{56a}$$

$$B = N/\pi E^* = (p_0/E^*)(b/a^2e^2)[(a/b)^2\,\boldsymbol{E}(e) - \boldsymbol{K}(e)] \tag{56b}$$

$$\delta = L/\pi E^* = (p_0/E^*)\,b\,\boldsymbol{K}(e) \tag{56c}$$

where

$$e^2 = 1 - (b/a)^2,\ b < a$$

and $\boldsymbol{K}(e)$ and $\boldsymbol{E}(e)$ are complete elliptic integrals.

Dividing equation (56b) by (56a) gives:

$$\frac{R'}{R''} = \frac{B}{A} = \frac{(a/b)^2\,\boldsymbol{E}(e) - \boldsymbol{K}(e)}{\boldsymbol{K}(e) - \boldsymbol{E}(e)} \tag{57}$$

which may be inverted to find the shape of the contact ellipse, i.e.

$$b/a = (A/B)^{1/2}\,F_0(B/A) \tag{58}$$

Multiplying equations (56a) and (56b) gives:

$$(AB)^{1/2} = 1/2R_e = (p_0/E^*)(b/a^2e^2) \times \{[(a/b)^2\boldsymbol{E}(e) - \boldsymbol{K}(e)][\boldsymbol{K}(e) - \boldsymbol{E}(e)]\}^{1/2}$$

We now write $c = (ab)^{1/2}$ and put $p_0 = 3P/2\pi ab$ to obtain:

$$c^3 = (ab)^{3/2} = \frac{3PR_e}{4E^*}\frac{4}{\pi e^2}(b/a)^{3/2} \times \{[(a/b)^2\boldsymbol{E}(e) - \boldsymbol{K}(e)][\boldsymbol{K}(e) - \boldsymbol{E}(e)]\}^{1/2}$$

i.e.

$$c = (ab)^{1/2} = \left(\frac{3PR_e}{4E^*}\right)^{1/3} F_1(B/A) \tag{59}$$

and

$$p_0 = \frac{3P}{2\pi ab} = \left(\frac{6PE^{*2}}{\pi^3 R_e^2}\right)^{1/3}[F_1(B/A)]^{-2/3} \tag{60}$$

The elastic approach is found from equation (56c) to give:

$$\delta = \frac{3P}{2\pi abE^*}b\,\boldsymbol{K}(e) = \left(\frac{9P^2}{16E^{*2}R_e}\right)^{1/3}\frac{2}{\pi}\left(\frac{b}{a}\right)^{1/2}[F_1(e)]^{-1/3}\,\boldsymbol{K}(e) = \left(\frac{9P^2}{16E^{*2}R_e}\right)^{1/3}F_2(B/A) \tag{61}$$

The functions $F_1(B/A)$, $[F_1(B/A)]^{-2/3}$ and $F_2(B/A)$ are plotted in Fig. 3. When B/A approaches unity these functions also approach unity and the expressions for c, p_0 and δ become those for the contact of solids of revolution, which have a circular contact area of radius c.

APPENDIX 2

The total deflexion of a tooth pair (Fig. 5) is the sum of the bending deflexion δ_B and the contact deflexion δ_c. Consider contact at the pitch point as representative of one pair of teeth in contact. Bending of a cantilever, length l and width w under a load F gives a deflexion:

$$\delta = Fl^3(1 - \nu^2)/3EI$$

For standard involute teeth $l \propto m$, $I \propto m^3 w$, $F = P'w \cos \phi_0$ which, for a tooth pair, gives:

$$\delta_B = KP'/E^* \tag{62}$$

where K is a constant ~ 6.

The contact deflexion of one surface relative to a point at depth h below the surface is:

$$(F/w)\frac{1 - \nu_2}{\pi E}[2 \ln (h/a) - \nu/(1 - \nu)]$$

Now

$$a = 2(P'R/\pi E^*)$$

We take h to be the depth of the centreline of the tooth ($\frac{1}{4}\pi m$) and the relative radius of curvature R to be $2m$ (1 : 1 gear with 24 teeth). Then

$$\frac{h}{a} = \tfrac{1}{4}\pi m \tfrac{1}{2}(\pi E^*/P'2m)^{1/2}$$

$$\cong 0.5\,(mE^*/P')^{1/2}$$

The contact deflexion of a tooth pair is thus:

$$\delta_c = \frac{P'}{E^*}\frac{\cos \phi}{\pi}\left[\ln\left(\frac{mE^*}{P'}\right) - 1.8\right] \tag{63}$$

APPENDIX 3

The bounce of a ball of mass m and radius R in coplanar motion is shown in Fig. 14. The tangential and normal velocities of its centre G are denoted by u and v; its angular velocity (back-spin positive) is ω. The peripheral

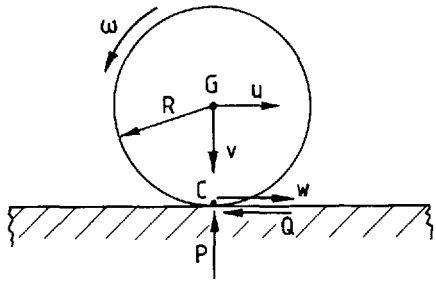

Fig. 14 Oblique impact of a sphere of radius R with a flat surface

velocity w of the point of contact with the ground C is given by

$$w = u + \omega R \tag{64}$$

Normal and tangential forces P and Q are exerted at C. If the ball is perfectly elastic the normal impulse during impact is given by:

$$\hat{P} = 2mv_1 \tag{65}$$

where suffixes 1 and 2 denote velocities before and after impact. The tangential impulse is given by:

$$\hat{Q} = -m(u_2 - u_1) \tag{66}$$

The moment of inertia of a uniform sphere about a diametral axis is $2mR^2/5$.

Moment of momentum about the point of impact C is conserved so that:

$$-mu_2R + \tfrac{2}{5}m\omega_2R^2 = -mu_1R + \tfrac{2}{5}m\omega_1R^2 \tag{67}$$

from which we can write:

$$\hat{Q} = \tfrac{2}{7}m(w_1 - w_2) \tag{68}$$

If slip continues throughout the time of contact:

$$\hat{Q} = \mu\hat{P}$$

i.e.

$$\tfrac{2}{7}m(w_1 - w_2) = \mu 2mv_1$$

so that

$$w_2 = w_1 - 7\mu v_1 \tag{69}$$

If $\mu > w_1/7v_1$, slip ceases during the impact and

$$w_2 = 0 = u_2 + \omega_2 R \tag{70}$$

This lecture is published for presentation at an Ordinary Meeting of the Tribology Group on 1 December 1982. The MS was received on 15 September 1982.