J. Fluid Mech. (1972), *vol.* 56, *part* 4, *pp.* 775–802
Printed in Great Britain

Wave mechanics of breakdown

By M. T. LANDAHL

Department of Aeronautics and Astronautics, Massachusetts Institute of Technology

(Received 21 December 1971 and in revised form 14 July 1972)

Kinematic wave theory is used to determine under what conditions breakdown of a steady or unsteady laminar flow into high frequency oscillations should occur. The analysis of a small-scale secondary wave riding on a large-scale inhomogeneity, such as that produced by a finite amplitude primary instability wave, reveals that the breakdown mechanism has three basic ingredients: (i) a self-excited secondary wave with a group velocity near the propagation velocity (phase velocity) of the primary wave, (ii) space–time focusing of the secondary wave train on the primary wave crest and (iii) a nonlinear filtering mechanism leading to rectification of the secondary wave.

The theory is applied to a laminar shear flow. Good quantitative agreement with the experiments on boundary-layer transition by Klebanoff, Tidstrom & Sargent (1962) is found for the critical condition leading to breakdown. Also, the theory is able to explain all the main qualitative breakdown features observed by Klebanoff *et al.* and others, such as the rapid localized onset, and the formation of a hairpin vortex lifting up from the surface downstream of the primary wave crest.

1. Introduction

The transition of laminar flow to turbulence was one of the very first fluid flow phenomena to receive serious attention in the history of fluid mechanics. In his famous investigation of flow in a pipe, Reynolds (1883), employing dye to visualize the flow, observed that, as the velocity in the tube was gradually increased, the dye streak would, at some distance downstream from the tube entrance, suddenly mix up with the surrounding water and thus indicate the appearance of a rapid unsteady 'sinuous motion' (later termed 'turbulent motion' by Lord Kelvin). Reynolds also noticed that the 'flashes' of turbulence would be intermittent at first, and that the critical velocity at which they first appeared depended strongly on the level of disturbances in the water present when it was drawn into the tube. The suddenness of transition has later been demonstrated to be a feature common to all high Reynolds number shear flows in the presence of walls. In their boundary-layer instability experiments, Schauber & Skramstad (1948) found that the final transition to a fully turbulent flow occurred in the form of intermittent high frequency oscillations of large amplitude. A visual illustration of this process was provided by Emmons (1951), who, in observing the flow of a thin sheet of water down an inclined plane, saw the turbulence first appearing in localized spots which would move downstream with the flow, spread

laterally and eventually merge to produce a fully turbulent flow. Later, more quantitative investigations such as those by Klebanoff *et al.* (1962), Kovasznay, Komoda & Vasudeva (1962), Hama & Nutant (1963) and Obremski & Fejer (1967) attest to the suddenness and localized initial appearance of the high frequency oscillations. We shall here use the term breakdown for the onset of this mode of secondary oscillation in accordance with the terminology of Klebanoff *et al.* (1962) (hereafter referred to as I). Recent experiments in turbulent shear flows (Kline *et al.* 1967; Kim, Kline & Reynolds 1971) give strong indications that 'bursts' of turbulence resembling breakdown appear intermittently in fully developed turbulent shear flows, as well, and may indeed be the main mechanism responsible for maintaining the turbulence.

Although a great deal of experimental information concerning breakdown has accumulated over the last decade or so, no completely satisfactory theoretical explanation of the phenomenon has as yet appeared. The most commonly expressed hypothesis, first proposed by Betchov (1960), is that breakdown is a manifestation of secondary instability in the local, highly inflexional, instantaneous velocity profiles induced by the primary instability wave. The secondary instability hypothesis in its simplest version based on the assumption of a quasi-steady parallel primary flow does not, however, seem to be able to explain the suddenness in the onset of breakdown. Greenspan & Benney (1963) included the effect of unsteadiness in the secondary instability model by considering a parallel shear flow whose thickness varies periodically with time. The analysis, which was carried out for a simplified straight-line free-shear-layer velocity profile, showed that a substantial increase in growth rate over that of a quasi-steady analysis was produced during the contraction phase of the oscillation cycle. However, the theory would not give information regarding what conditions would lead to the appearance of first secondary waves. Obremski & Morkovin (1969) in analysing the experiments of Obremski & Fejer (1967) used quasi-steady stability theory to trace the development of wave packets in a shear flow varying in time and space, but they were not able to deduce a precise breakdown criterion.

Probably the most puzzling feature of the breakdown phenomenon is the exceedingly rapid growth of the secondary disturbance at the onset. This is exemplified in figure 1, reproduced from I, which shows the fluctuation amplitude downstream of an oscillating ribbon measured at a distance from the wall of approximately 0·6 boundary-layer thicknesses. The accompanying oscillograms of the u fluctuations (figure 21 of I, not reproduced here) indicate clearly that the rapid, almost instantaneous rise in amplitude seen at a distance of 9 in. from the ribbon is primarily caused by a 'spike-like' disturbance which in the distance of 0·25 in. from $x = 9$ in. to $x = 9.25$ in. has grown by a factor of about 6. As the frequency measurements indicate that this distance is only about 0·7 of the wavelength of the secondary disturbances, the equivalent exponential growth rate would be much higher than the secondary inflexional instability mechanism could provide at this wavenumber. (The calculations reported in §4, below, based on quasi-parallel theory, give a growth rate close to zero for the wavenumber of interest.) One is therefore led to look for some other explanation for the enhanced growth. One that could be dismissed at the outset is that the rapid growth could be mainly

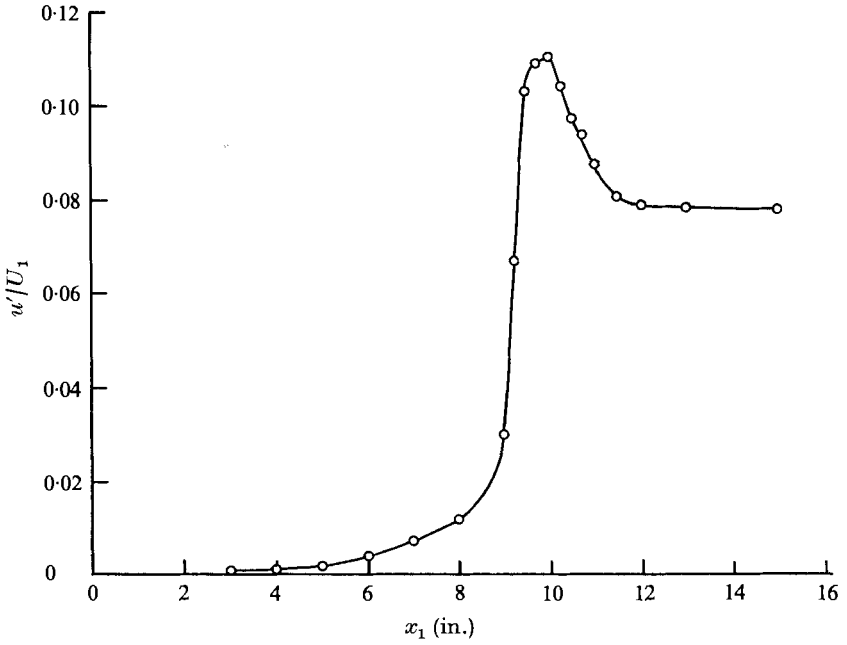

FIGURE 1. Variation of secondary disturbance amplitude with distance from oscillating ribbon (from I).

due to the lifting up of a small-scale but intensive disturbance piercing the $y/\delta = 0{\cdot}6$ level just behind $x = 9$ in. The measurements show that the secondary disturbance has a y extent somewhat greater than δ (I, p. 25), so that once the disturbance appears it will be felt all across the boundary layer with a maximum amplitude to be expected from inviscid stability theory to be found near the inflexion point of the instantaneous profile. In the case of interest this is located at about $y/\delta = 0{\cdot}5$, i.e. close to the point of measurement. (There is also another inflexion point at $y/\delta \simeq 0{\cdot}25$, but this corresponds to a minimum in the shear and would not produce instability.) The lift-up velocity required to produce the whole observed growth rate would therefore be truly enormous, although of course some small contribution could come from this effect. It should be noted that measurements of the intensity closer to the wall such as those of figure 4† in I do not show the rapid rise because (*a*) the amplitude of the secondary instability is small at this position ($y/\delta \simeq 0.2$) owing to the large distance from the inflexion point and (*b*) the primary wave has a higher amplitude there than near its phase reversal point at $y/\delta \simeq 0{\cdot}6$ and therefore tends to mask the high frequency oscillations.

In view of the above we shall therefore instead look for a different mechanism in which the nonlinearity is by necessity strong such as in a shock wave, a hydraulic jump or in vortex breakdown. Suggestions along this line have been made by Lighthill (1970, §IV). Whether a particular flow can remain smooth may be investigated by analysing the development of a small wavelike perturbation of

† One should notice, however, that the experimental conditions underlying figures 4 and 21 are slightly different, so that the location of the breakdown is somewhat different in the two cases.

the flow as in hydrodynamic stability theory. Consider as an illustrative example the supersonic one-dimensional gas flow in a Laval nozzle. An oscillatory disturbance will produce two sets of acoustic waves travelling downstream in the nozzle: an advancing one propagating with a velocity of $U+a$ (where U and a are the local flow and sound velocities, respectively) and a receding one with a propagation velocity of $U-a$. It is the latter that is of interest here. Conservation of wave action in a linear system (see below) will mean that the amplitude of the receding wave train should vary as $(U-a)^{-\frac{1}{2}}$. Hence, according to linear theory, the amplitude of the receding waves would become infinite if the throat area were to be decreased so as to produce sonic flow there.

This is an example of space–time focusing of the waves brought about by consecutive wave fronts catching up with each other and thus producing a concentration of wave energy. Of course, the infinite amplitudes will not occur in reality owing to diffraction and because nonlinear effects will set in once the amplitude becomes large. An essential effect of nonlinearity is to give the positive and negative portions of the wave slightly different propagation velocities. Thus, those of positive pressure perturbation will travel somewhat slower than $U-a$ and therefore tend to get trapped somewhat upstream of the sonic throat, whereas those of negative pressure will have a slightly higher propagation velocity and therefore will be swept downstream and out of the throat. The throat will hence act as a rectifier and collector of acoustic energy radiated from upstream thus making the disturbance pressure build up rapidly and initiate a shock, which will subsequently cause choking of the nozzle. This simple example suggests the existence of a flow which cannot support arbitrary infinitesimal perturbations but which must break down locally into finite amplitude initially small-scale oscillations.

It is the purpose of the subsequent analysis to demonstrate that breakdown should be a distinct possibility for any inhomogeneous continuum system which can support propagating waves of scale small relative to the scale of the inhomogeneities.

2. Critical condition for breakdown

As the experiments on transition indicate that the secondary disturbances appearing at breakdown have a length scale much smaller than that of the primary instability wave, one can in the analysis of them apply the ideas of kinematical wave theory developed in recent years by Whitham (1965), Hayes (1970) and others. We shall here follow Hayes's (1970) exposition of the basic theory.

The kinematic wave theory deals with the propagation of wave trains in a slightly inhomogeneous medium. Provided that the inhomogeneity is slight in the sense that the variation in the medium properties is small within a distance of one wavelength and time of one cycle of the wave considered, the wave propagation characteristics can be described approximately by a local dispersion relationship

$$\omega = \Omega(\mathbf{x}, \mathbf{k}, t), \tag{1}$$

where ω is the frequency, $\mathbf{x}$ the position vector and $\mathbf{k}$ the wavenumber vector. For the moment we shall assume Ω to be real, as would be the case for wave propagation in a conservative system. Later, the changes needed for treatment of a slightly non-conservative system will be incorporated. One may consider either local waves or so-called modal ones (Hayes 1970). For the latter, the behaviour with the cross-space co-ordinate(s) is given by the eigensolution of the local eigenvalue problem. In the problem at hand, the waves are modal ones of the (straight or oblique) Tollmien–Schlichting or Rayleigh (i.e. inviscid) variety propagating in the x, z, t space, and their behaviour with y is given by the solution to the Orr–Sommerfeld problem for the local instantaneous velocity profile†.

Central to the theory is the concept of a wave group trajectory given by

$$d\mathbf{x}/dt = \mathbf{c}, \tag{2}$$

where

$$\mathbf{c} = \Omega_{\mathbf{k}} \tag{3}$$

is the group velocity. The wavenumber varies along the trajectory according to the relation

$$d\mathbf{k}/dt \equiv \partial\mathbf{k}/\partial t + (\mathbf{c}.\nabla)\,\mathbf{k} = -\Omega_{\mathbf{x}}. \tag{4}$$

Solution of (1)–(4) determines a set of rays

$$\mathbf{x} = \mathbf{x}(\mathbf{a}, t), \quad \mathbf{k} = \mathbf{k}(\mathbf{a}, t), \tag{5}$$

where $\mathbf{a}$ is a parameter having the same dimensions as $\mathbf{x}$. It may be regarded as the Lagrangian label of the wave group. Note that there will generally be a different ray for each initial wavenumber considered.

The above relations are purely kinematical and simply express the conservation of the number of wave crests. The variation of wave amplitude along the ray can be determined through Whitham's (1965) principle of conservation of wave action. In a linear problem the wave action density A, which is a quantity having the dimensions energy times time, is proportional to the square of the amplitude, and its flux is $\mathbf{c}A$. For a linear conservative system the following conservation law then holds:

$$\partial A/\partial t + \nabla.(\mathbf{c}A) = 0, \tag{6}$$

or

$$\mathrm{d}A/dt = -A(\nabla.\mathbf{c}). \tag{6a}$$

In evaluating the right-hand side of (6a) with the aid of (3) one must keep in mind the variation of $\mathbf{k}$ along the ray imposed by (4). Therefore, we shall, following Hayes (1970), distinguish between propagation space $(\mathbf{x}, t)$ (in the present cases (x, z, t) with y constituting the cross-space variable) and the augmented space $(\mathbf{x}, \mathbf{k}, t)$. The symbols $\partial/\partial t$, $\partial/\partial\mathbf{x}$ and ∇ are used for derivatives in the propagation space and subscripts for derivatives in the augmented space. Thus the operator d/dt also includes the contribution from the variation of $\mathbf{k}$ along the trajectory.

It can be shown that the conservation law (6) is equivalent to invariance of the quantity AJ, where

$$J = |\nabla_{\mathbf{a}}\mathbf{x}| \tag{7}$$

† Lighthill (1969) has proposed the use of local waves for the analysis of shear flow turbulence. This would be appropriate for very short secondary waves of scales much smaller than the boundary-layer thickness. In the case of breakdown studied by Klebanoff *et al.* (1962) the secondary wave first appearing was found to have a wavelength of about one and a half times the boundary-layer thickness.

is the Jacobian determinant of $\mathbf{x}$ with respect to parameter space (Hayes 1970). J is a measure of the volume as convected by the rays. Conservation of wave action takes into account the changes in wave intensity brought about by convergence or divergence of neighbouring rays. A point where $J = 0$ is known as a focus and is a singular point in kinematic wave theory.

To determine J directly from the solution (5) one needs to compute neighbouring rays and then the determinant (7). Alternatively, J can be calculated by direct quadrature of the equation

$$d(\ln J)/dt = \nabla . \mathbf{c}, \tag{8}$$

which is obtained from (6a) and imposing constancy of AJ.

In order to make use of these ideas for the analysis of shear waves we need to extend the theory to non-conservative systems. For such systems one will have a complex local dispersion relation $\Omega = \Omega^{(r)} + i\Omega^{(i)}$ in which $\mathbf{k}$ may be considered real. For small growth rates $\Omega^{(i)}$, equation (6a) and the linear theory for a homogeneous system suggest that a good approximation to the rate of change of the action variable A should be (see Lighthill 1969, equation 7)

$$\frac{1}{A}\frac{dA}{dt} = -\nabla . \mathbf{c} + 2\Omega^{(i)}, \tag{9}$$

with $\mathbf{c}$ determined from $\Omega^{(r)}$. (For simplicity, the superscript will be omitted from $\Omega^{(r)}$ in the following.) This equation reduces to the correct one in both the conservative and homogeneous limiting cases, in the latter when interpreted in terms of a complex $\mathbf{k}$ producing the familiar Gaster (1968) relationship between temporal and spatial growth rates. As the results that follow indicate that the situation of primary interest is one for which $\Omega^{(i)}$ is equal or close to zero, the approximation (9) should be adequate for the present purpose. Combination of (8) and (9) gives

$$A = \frac{\text{constant}}{J} \exp\left(2\int \Omega^{(i)} dt\right), \tag{10}$$

which shows that a focus will produce a singularity in A for a slightly non-conservative system as well, provided that $\int \Omega^{(i)} dt \neq -\infty$. In reality, the wave amplitude will of course not become infinite. The ray theory ceases to be valid near a focus, and a more complete theory would be required there in which local diffraction as well as nonlinear effects would have to be considered. To make a complete assessment of whether the total growth of the wave is indeed large at the focus it would be necessary to study how the quantity $\int \Omega^{(i)} dt$ behaves (as well as to determine the way in which J approaches zero, i.e. how strong the singularity is). However, as the analysis below indicates that the total amplification also usually becomes the largest near the focus, the search for possible zeros of J would be the most essential step in the task of finding out where nonlinearities are first likely to become strong.

We shall study the propagation of a wave packet through inhomogeneities caused by a locally plane, travelling primary wave whose amplitude may vary along the wave (for example, because of spanwise irregularities) so as to produce a minimum in the component of $\mathbf{c}$ normal to the wave front somewhere along the

wave. Such a minimum will generally occur near that point along the crest of the primary wave where its amplitude has a maximum. By crest we mean that phase of the primary wave cycle at which the deflexion of the outer streamlines away from the wall is the largest. We shall thus be interested in the neighbourhood of the 'top' of the crest along the wave. There the normal components of the local primary instantaneous velocities have their smallest values and, for a given wave-number, the normal component of $\mathbf{c}$ is then usually the lowest, being in a sense a weighted average of the primary velocities. The peak velocity profiles labelled as instantaneous in the experiments by Klebanoff *et al.* (1962) correspond essentially to top of the crest. Near this point, the tangential component of $\mathbf{c}$ can be expected to have an extremum for reasons of local symmetry about this point in the tangential direction, so that along a section normal to the crest through the top of the crest

$$\nabla . \mathbf{c} \simeq \partial c^{(n)}/\partial \xi, \tag{11}$$

where ξ is the co-ordinate normal to the primary wave front (considered positive downstream) and

$$c^{(n)} = \Omega_\alpha \tag{12}$$

is the component of $\mathbf{c}$ normal to the wave front, α being the corresponding wave-number component. Note that the tangential component will not contribute substantially to $\nabla . \mathbf{c}$ as in the section considered its derivative in the tangential direction would be close to zero. Now, if the inhomogeneity caused by the primary wave is a frozen pattern moving with the (phase) velocity $\mathbf{c}_0$,

$$\frac{\partial \mathbf{c}}{\partial t} = -(\mathbf{c}_0 . \nabla)\,\mathbf{c} = -c_0^{(n)} \frac{\partial \mathbf{c}}{\partial \xi}, \tag{13}$$

where $c_0^{(n)}$ is the primary phase velocity component normal to the wave. Thus,

$$\frac{dc^{(n)}}{dt} = \frac{\partial c^{(n)}}{\partial t} + c^{(n)} \frac{\partial c^{(n)}}{\partial \xi} = (c^{(n)} - c_0^{(n)}) \frac{\partial c^{(n)}}{\partial \xi}. \tag{14}$$

By combining (8), (11) and (14) we obtain

$$\frac{d(\ln J)}{dt} = \frac{1}{(c^{(n)} - c_0^{(n)})} \frac{dc^{(n)}}{dt}, \tag{15}$$

or, provided that $c_0^{(n)}$ is independent of time,

$$J = K(c^{(n)} - c_0^{(n)}), \tag{16}$$

where K is a constant for each ray determined by the initial condition. It thus follows that J will have a zero whenever the component of the secondary group velocity normal to the primary wave front becomes equal to the normal component of the phase velocity of the primary wave. Since only the normal components of $\mathbf{c}$ and $\mathbf{c}_0$ need be considered once the appropriate section along the primary wave front has been located, the problem of finding the focus can, through an appropriate rotation of the co-ordinate system, be reduced to the study of wave propagation in one space dimension. Thus, for the modal waves of present interest we need only concern ourselves with two-dimensional waves. As no confusion can arise thereby, we shall for simplicity in writing omit the superscripts on $c_0^{(n)}$ and $c^{(n)}$ in the following.

For the case of a neutrally stable primary wave we obtain from (2)–(4) with $\xi = x - c_0 t$

$$d\xi/dt = c - c_0, \tag{17}$$

$$d\alpha/dt = -\Omega_x = -\Omega_\xi \tag{18}$$

and

$$dc/dt = \Omega_{\alpha t} + c\Omega_{\alpha x} - \Omega_{\alpha\alpha}\Omega_x = (c - c_0)\,\Omega_{\alpha\xi} - \Omega_{\alpha\alpha}\Omega_\xi. \tag{19}$$

Elimination of t gives

$$\frac{d\alpha}{d\xi} = -\frac{\Omega_\xi}{c - c_0}, \tag{20}$$

$$\frac{dc}{d\xi} = \frac{(c - c_0)\,\Omega\alpha_\xi - \Omega_{\alpha\alpha}\Omega_\xi}{c - c_0}, \tag{21}$$

which are convenient forms for studying the ray trajectories leading to simple analytical results in the ξ, c plane. We shall consider the local dispersion relation to have the character usually encountered in inflexional boundary-layer instability. Typical such curves are those of figures 7 and 9. Thus, $\Omega^{(i)}$ is positive for values of the wavenumber α less than that for neutral stability α_s and $c_\alpha = \Omega_{\alpha\alpha}$ is negative in this neighbourhood. There is one such dispersion diagram associated with each value of ξ, with the lowest values of c, for a given α, appearing at the crest, so that Ω_ξ and $\Omega_{\alpha\xi} - c_\xi$ are both zero for $\xi = 0$. (If the primary wave has large phase shifts, these zeros may be located at slightly different points. As the analysis below shows, however, it is the zero of Ω_ξ that should be associated with $\xi = 0$.) An initial point in the ξ, c plane will, through the dispersion relation, define a value† of α, which then can be used as a starting value in the step-by-step integration of (20) and (21). For regions close to the point $\xi = 0$, $c = c_0$ (labelled O in the following) it is possible to find simple approximate solutions by neglecting the first term in the numerator of (21) and setting

$$\Omega_{\alpha\alpha} \simeq (\Omega_{\alpha\alpha})_O = -B, \quad \Omega_\xi \simeq \xi(\Omega_{\xi\xi})_O = C\xi, \tag{22}$$

where B and C are constants, which, upon insertion into (21) and separation of variables, yields the following solution:

$$(c - c_0)^2 - BC\xi^2 \simeq (c_a - c_0)^2 - BC\xi_a^2, \tag{23}$$

$$\Omega - \alpha c_0 = (\Omega - \alpha c_0)_a, \tag{24}$$

where the suffix a refers to the Lagrangian label of the wave group as before. The ray trajectories are thus hyperbolas in the c plane.‡

The relation (24) is exact and states that the frequency of the wave is constant relative to a co-ordinate system moving with the primary wave. The time required for the wave group to reach a point ξ, c is found through integration of (19) using the same approximations as before:

$$t - t_a = \frac{1}{(BC)^{\frac{1}{2}}}\left|(\operatorname{sgn}\xi)\ln\left(\frac{c - c_0 + (BC\xi^2)^{\frac{1}{2}}}{c_m - c_0}\right) - (\operatorname{sgn}\xi_a)\ln\left(\frac{c_a - c_0 + (BC\xi^2)^{\frac{1}{2}}}{c_m - c_0}\right)\right|, \tag{25}$$

† Actually, when $c = c(\alpha)$ has a maximum, there may be two values of α for a given point ξ, c. However, as we are primarily interested in the higher wavenumber region, this complication need not concern us here.

‡ This is always the case when $\Omega_{\alpha\alpha}$ and $\Omega_{\xi\xi}$ have the opposite sign. If not, the trajectories will become ellipses.

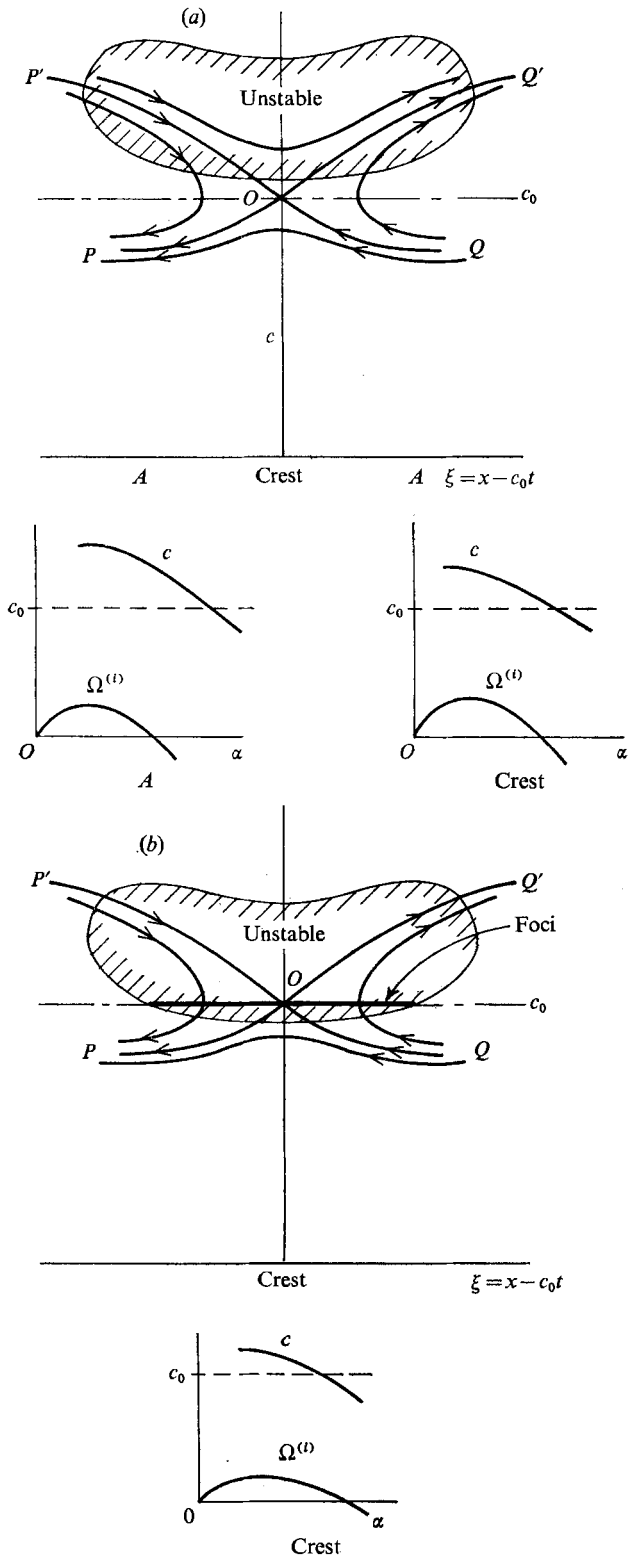

FIGURE 2. Secondary ray trajectories (hypothetical) for a neutrally stable primary wave. (*a*) Before breakdown. (*b*) After breakdown.

where $c_m - c_0$ is the extremum of $c - c_0$ reached in the trajectory. The total amplification of the wave can also be calculated. We may assume that near $c = c_0$ the amplification rate varies approximately linearly with c, so that

$$\Omega^{(i)} \simeq D(c - c_s), \tag{26}$$

where c_s is the group velocity at neutral stability and D is a constant. Using the relations between ξ, c and t we find

$$\int \Omega^{(i)} dt = D[\xi - \xi_a + (c - c_0)(t - t_a)]. \tag{27}$$

With the aid of the above results for the behaviour near O one can easily sketch the ray trajectories. Typical (hypothetical) examples are shown in figure 2. At larger distances from the crest the hyperbolas will become distorted. For example, there would generally be a maximum as well as a minimum value of c possible at each ξ, so that the trajectories would span only a finite range of c values. The unstable region will shrink away from the crest and may disappear altogether as the inflexional regions eventually may vanish closer to the neighbouring troughs. There are only two limiting trajectories that lead to the point O, namely those following the asymptotes of the distorted hyperbolas. The trajectory starting at P' will reach O after an infinite time (see (25)) and then move out either along OP or OQ' taking an infinite time doing so. The trajectory starting at O will proceed either along OQ' or OP. All other trajectories are uniquely defined. The ones to the left of POP' with $c > c_0$ will first begin to catch up with the crest but will slow down while doing so and not quite reach it. They will cross the line $c = c_0$ with infinite slope and will experience space–time focusing there. Similarly, the waves to the right of QOQ' will first recede towards the crest, focus at $c = c_0$, but will then speed up and eventually move away downstream. The waves above $P'OQ'$ will move past the crest since c is always greater than c_0, whereas those below POQ will be overtaken by the crest, as this then always moves faster than the wave.

Consider now the amplification of the secondary waves. Take first the case, depicted in figure 2(*a*), for which the unstable regime is everywhere above the line $c - c_0$. This may be typical of the earlier stages of the primary wave development in which its fluctuation amplitudes are small and hence whatever inflexion points are produced in the instantaneous velocity profile are located in the outer part of the boundary layer. The rays coming through the unstable region will receive a finite amount of amplification but will eventually move out into regions with damping. The wave on the limiting trajectory from P' will become completely quenched as it approaches O. Now, consider a later stage of the primary wave development illustrated in figure 2(*b*), in which the unstable region has moved past O. (This could also come about because of speed-up of the primary wave.) There will then be waves receiving a large amount of amplification. In particular, those coming in along the limiting trajectories will continue to amplify for ever, so that their total growth will become infinite by the time they reach O, however small $\Omega^{(i)}$ is. In addition, focusing will reinforce the disturbance even further, thus making almost certain that strong nonlinear effects will set in. A critical condition is therefore reached when the stability boundary first passes over O, i.e. when

$$c_0 = c_s. \tag{28}$$

The above reasoning would indicate that, before the critical condition is reached ($c_0 < c_s$), all disturbances travelling with the crest would be completely quenched, whereas immediately after c_s has become equal to c_0, disturbances reaching O would have gone through an infinite amplification. A plot of disturbance amplitude at the crest versus time (following the crest) would then show zero until the instant the critical condition is reached and infinity thereafter. This presupposes, however, that the primary wave has existed for an infinite time before the critical condition is attained. Also, we have here altogether neglected the effect of the rate of change of the primary wave properties, which will be examined next.

We begin by considering the condition for focusing. As the primary wave is now unsteady, we shall define a local phase velocity by

$$c_0 = -(\partial c/\partial t)/(\partial c/\partial x). \tag{29}$$

This definition is suggested by the steady case, for which c is a function of $\xi = x - c_0 t$ only. The phase velocity so defined is a measure of the local velocity with which an inhomogeneity travels, as mirrored in the associated temporal and spatial variations of the energy propagation velocity. The value of c_0 will generally depend on the amplitude of the non-steady primary wave (see below). For the three-dimensional case a z component could be defined in an analogous manner. Combining (29) with the definition of dc/dt one finds that

$$\frac{\partial c}{\partial x} = \frac{dc/dt}{c - c_0} \tag{30}$$

and equation (8) for J can thus be written as

$$\ln J = \int \frac{dc/dt}{c - c_0}\, dt = \int \frac{dc}{c - c_0}, \tag{31}$$

which is identical to (15) for the steady case, except that now c_0 is a function of t and hence c. It therefore follows that J will vary near c_0 as

$$J = K(c - c_0)^{1/(1-\kappa)}, \tag{32}$$

where

$$\kappa = \frac{dc_0}{dc} = \frac{dc_0/dt}{dc/dt} \tag{33}$$

evaluated at $c = c_0$. Thus, with the definition (29) the focus is again identified with $c = c_0$, provided that $\kappa < 1$ and $-\kappa \neq \infty$. To obtain an expression for c_0 directly in terms of Ω we rewrite (30) as follows:

$$c - c_0 = \frac{dc/dt}{\partial c/\partial x} = \frac{\Omega_{\alpha t} + c\Omega_{\alpha x} - \Omega_{\alpha\alpha}\Omega_x}{\Omega_{\alpha x} + \Omega_{\alpha\alpha}(\partial\alpha/\partial x)}. \tag{34}$$

When $c = c_0$ we have either that the numerator is zero or the denominator is infinite. In the former case $dc/dt = 0$, so that $|\kappa| = |(dc_0/dt)/(dc/dt)| = \infty$, and J is not zero. In the latter, $\partial\alpha/\partial x = \infty$. By taking $\partial/\partial x$ of (4) we obtain with the aid of (3)

$$\frac{d}{dt}\frac{\partial\alpha}{\partial x} = -\left(\frac{\partial\alpha}{\partial x}\right)^2 \Omega_{\alpha\alpha} - 2\frac{\partial\alpha}{\partial x}\Omega_{\alpha x} - \Omega_{xx}. \tag{35}$$

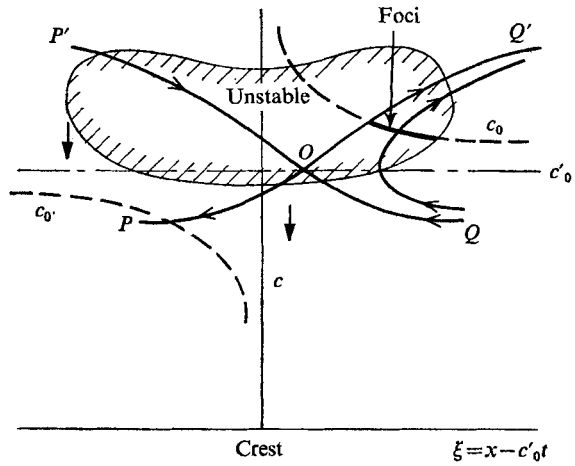

FIGURE 3. Secondary ray trajectories (hypothetical) for a growing primary wave of phase velocity c'_0.

(This is Hayes's (1970) 'derived ray equation' (19) in one space dimension.) Near the singularity the first term dominates, and we have

$$\frac{d}{dt}\left[\left(\frac{\partial\alpha}{\partial x}\right)^{-1}\right] \simeq \Omega_{\alpha\alpha} \tag{36}$$

or, using (4),

$$\left(\frac{\partial\alpha}{\partial x}\right)^{-1} = \int \Omega_{\alpha\alpha}\, dt = -\int \frac{\Omega_{\alpha\alpha}}{\Omega_x}\, d\alpha. \tag{37}$$

Taking Ω_x to be approximately constant near the zero of $(\partial\alpha/\partial x)^{-1}$, and making use of the fact that $c \equiv \Omega_\alpha = c_0$ at this point, we find that

$$\partial\alpha/\partial x = -\Omega_x/(c - c_0) \tag{38}$$

near the focus. Substitution of this into (34) yields

$$c_0 = -\Omega_{\alpha t}/\Omega\alpha_x = -c_t/c_x. \tag{39}$$

A sufficiently general approximate expression describing the variation of c near the crest of a slowly growing primary wave should be the following:

$$c = c_1(\alpha) - f(\xi, \alpha) \exp\left(\int \Omega_0^{(i)} dt\right), \tag{40}$$

where c_1 is of order unity, f is positive and proportional to the primary wave amplitude, $\Omega_0^{(i)}(t)$ is the growth rate of the primary wave, and

$$\xi = x - \int c'_0\, dt, \tag{41}$$

$c'_0(t)$ being the propagation velocity of the primary wave crest. Insertion into (39) gives

$$c_0 = c'_0 + \Omega_0^{(i)}(f/f_\xi) = c'_0 - \Omega_0^{(i)} f/c_\xi. \tag{42}$$

In a ξ, c diagram the $c = c_0$-curves will be independent of time and approximately hyperbolas with $\xi = 0$ and $c = c_0'$ as asymptotes (see figure 3). With the above, the equations determining the trajectories become

$$d\xi/dt = c - c_0', \tag{43}$$

$$dc/dt = (c - c_0')\, c_\xi + \Omega_0^{(i)} f - \Omega_{\alpha\alpha}\Omega_\xi. \tag{44}$$

By introducing

$$\xi' = \xi + [(\Omega_0^{(i)} f)/(\Omega_{\xi\xi}\,\Omega_{\alpha\alpha})]_O = \xi - \Delta \tag{45}$$

one sees that the region near O may be approximated in the same way as in the steady case. Thus, the trajectories are again hyperbolas but now displaced a distance Δ to the right. Typical trajectories are sketched in figure 3. For $\xi > 0$ the focusing takes place at values of c slightly greater than c_0', where for $\xi < 0$ it occurs below $c = c_0'$. An approximate solution of the equation for $\partial\alpha/\partial x$ (see appendix) shows that only trajectories starting to the left of $PO'P'$ and to the right of QOQ' can focus. The line segment along which focusing of unstable waves can take place is shown by the heavy line in figure 3. Evidently, the disturbances that now first become strong originate a small distance downstream of the crest. Also, the first unstable waves that can focus may appear even slightly before $c_0' = c_s$ is reached. For small primary wave growth rates, however, the condition (25) with c_0 replaced by c_0' should again rather accurately identify the critical state at which nonlinear effects become strong.

3. The qualitative effects of nonlinearity

Whereas the analysis carried out above is quite general and should apply to any continuum system, we need in the discussion of nonlinearities to specify what system is considered because the nonlinear characteristics differ considerably from system to system. Therefore, the discussion will be here confined to shear waves. Once the critical condition has been reached and nonlinearity becomes strong, positive and negative portions of the secondary wave cycle will behave differently. We shall consider as positive that part of the secondary wave cycle in which the disturbance vorticity is predominantly positive. As the linear theory identifies the condition of breakdown as one in which the secondary wave propagates with a group velocity very close to the phase velocity of the primary wave, the main direct nonlinear effect on the secondary wave that needs be considered is the added convection velocity in the streamwise direction of the secondary vorticity perturbation induced by the wave itself. To understand this effect qualitatively, we may use a simplified model in which the secondary wave train is approximated by a row of concentrated vortices of alternating sign being convected downstream by the primary flow. The boundary condition of zero normal velocity at the wall can be satisfied by the introduction of mirror image vortices as shown in figure 4. It follows from the figure that the self-induced effect will be such as to slow the positive vortices down and speed the negative ones up.

Now consider the different qualitative behaviour of the positive and negative portions of the secondary waves during passage through the critical region as

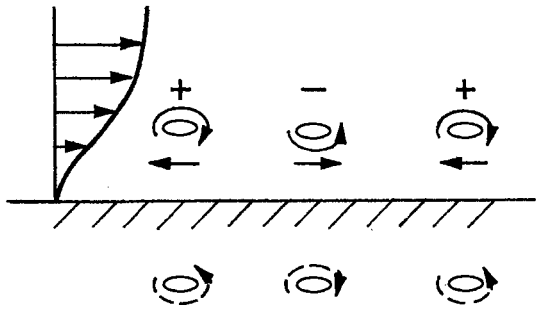

FIGURE 4. Simplified model illustrating the induced convection velocity of secondary vorticity.

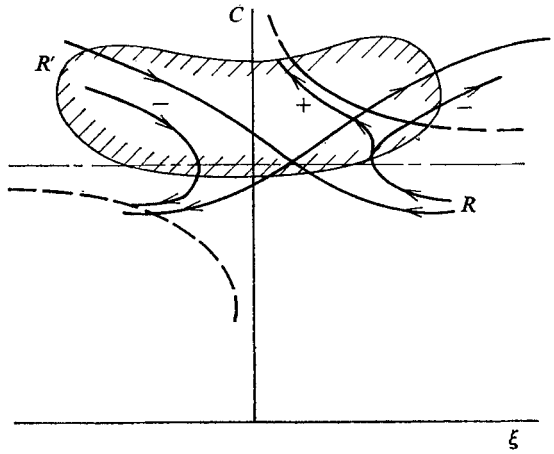

FIGURE 5. Ray trajectories (hypothetical) for positive and negative secondary wave portions illustrating rectification effect.

shown in figure 5. The self-induced effect is such as to add or subtract amounts to the propagation velocity of the group proportional to the amplitude of the wave but is likely not to alter the dispersion relationships substantially. A negative portion of the secondary wave will hence tend to be speeded up relative to the wave crest, whereas the positive one will be slowed down. For a wave starting along the ray R in the figure, the negative part of the wave, once it nears focusing and becomes strong, will veer off rapidly towards the right in the diagram and thus pass through the focus downstream of the crest and eventually into a region of stability. A positive part of the wave, on the other hand, will turn sharply to the left once it gets close to the focus and will therefore never be able to pass through the focus. Thus, it will be trapped on the wave crest in a region in which it can continue to amplify. For the waves originating at R' it will be the negative part that tends to move closer to the crest. However, as it slides along the line of foci, it will eventually get out in the stable range and be quenched. Hence, as the critical condition is reached, there will be a very rapid build-up of positive disturbance vorticity just downstream of the crest.

Because of the very strong concentrating effects both from focusing and wave trapping, this build-up will require only a very short distance, and it is

likely that the secondary vorticity will have built up its full strength already in the first half wave beyond the critical point. For a spatially growing primary wave, the streamwise position of the onset of criticality will fluctuate back and forth during the primary cycle as new crests become critical in succession. A measurement station at a fixed x location just beyond that at which the crest of the primary wave first becomes critical will therefore be swept over by secondary disturbances during only a small portion of the primary cycle, so that at the early stages of breakdown there would be time and room for only one secondary half wave between the point of first criticality and the measurement station. The strong secondary vorticity concentration, which according to the quasi-steady stability theory takes place predominantly around y stations close to the instantaneous inflexion point, would consequently give rise to a single half wave in the measured velocity which would be negative or positive depending on whether the velocity is measured below or above the vortex centre. As the point of measurement is moved downstream there will be room for two or more half waves. Hence, because of the rectification due to nonlinearity the signal will look like a piece of a sinusoid with either the positive or negative portions of the waves missing, i.e. it will have the appearance of a series of spikes as is indeed observed in transition measurements (see further next section).

For the three-dimensional development of the secondary disturbance one can deduce the following qualitative picture guided by general fluid-dynamical principles and experimental evidence. Continuity of vorticity requires that the disturbance vortex lines created at the spanwise position of breakdown terminate in the quasi-steady primary field at spanwise positions outside the critical region. As the secondary disturbance vorticity is thus produced essentially by a local lift-up of the primary vorticity distribution, a 'hairpin disturbance vortex is formed with its 'head' lifted up at the breakdown position and 'legs' extending down into the neighbouring non-critical spanwise stations and attached to the quasi-steady undisturbed vorticity distribution there. When the head is convected downstream, the legs will tilt over more and more while stretching and thereby create an upward velocity component tending to lift the head further away from the surface. Since it is thus blown into a region of higher velocity, the head will move away downstream at a velocity somewhat higher than that predicted by the two-dimensional theory. The spanwise extent of the critical region will increase continuously as the amplitude of the travelling primary wave grows, so that the region of breakdown will attain a planform looking like a growing blunt-nosed wedge with its apex pointing downstream.

4. Comparisons with experiments

In order to compare the theory with experiments on breakdown in the general case of unsteady shear flow, one would need instantaneous velocity profiles as functions of space and time so that the dispersion relation could be obtained for each combination $(\mathbf{k}, \mathbf{x}, t)$ and thence c_0 (equation 39) to determine the location of criticality. Such profiles could be found from perturbation theory or experiments. The calculation is much simplified, however, when the space–time

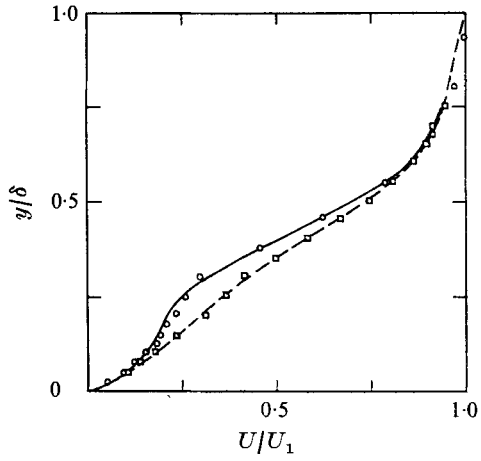

FIGURE 6. Fitted velocity profiles to the measured instantaneous and mean velocities at station C from the experiments in I. ○——○, instantaneous profile; □——□, mean profile.

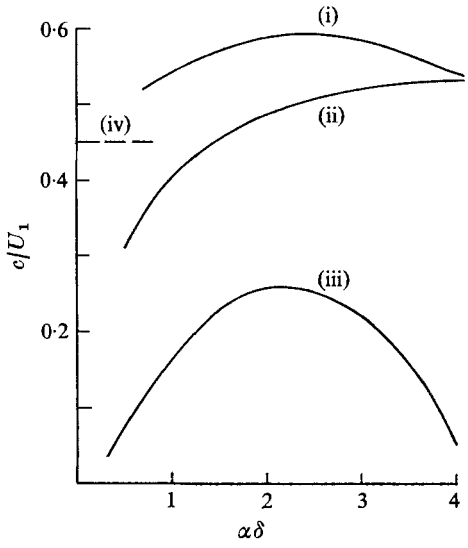

FIGURE 7. Dispersion diagram for instantaneous profile at station C. (i) Group velocity c/U_1. (ii) Phase velocity c_p/U_1. (iii) Growth rate $\Omega^{(i)}\delta/U_1$. (iv) Primary phase velocity c_0'/U_1.

inhomogeneity is a primary wave with a phase velocity and instantaneous velocity distribution that can be obtained directly from the measurements. This is the case with the experiments of I.

In the investigation reported in I controlled spanwise flow irregularities were introduced by means of small strips of tape glued under the oscillating ribbon at regular spanwise intervals. For the higher ribbon amplitudes these produced spanwise variations in the primary wave fluctuation intensities with fluctuation amplitude peaks and valleys (upstream of first breakdown location) in between and at the spanwise positions of the strips, respectively. Breakdown was observed to occur in the peak region and at that primary phase at which the instantaneous

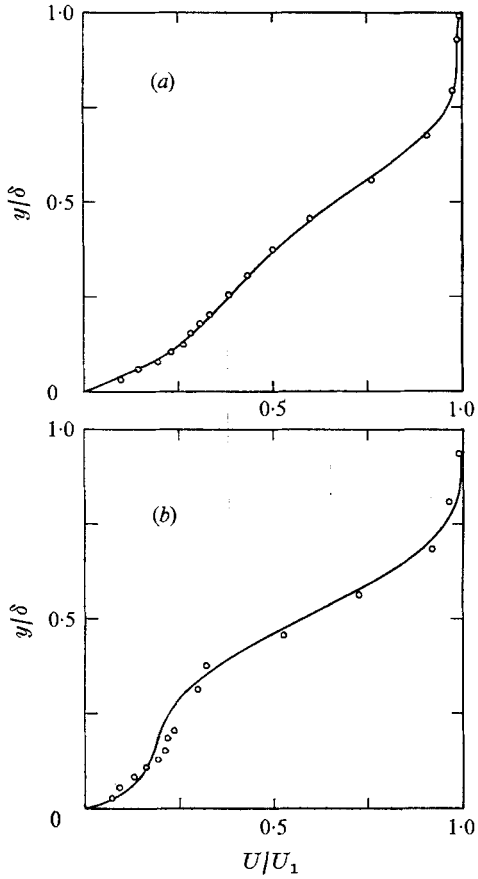

FIGURE 8. Velocity profiles fitted to the measured mean velocities at station D (at breakdown). (*a*) Mean. (*b*) Instantaneous. ——, fitted curve; ○, measurements, from I.

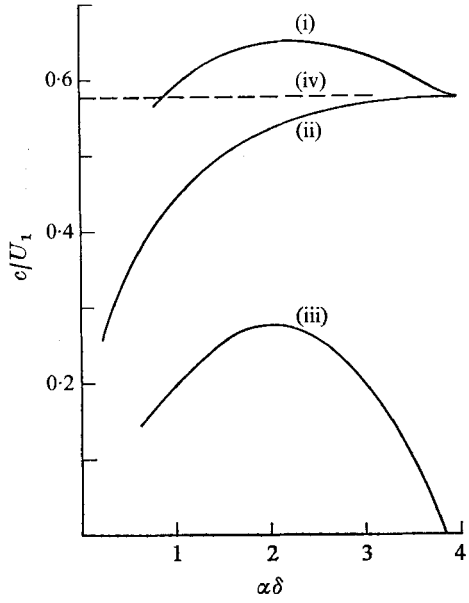

FIGURE 9. Dispersion diagram for instantaneous profile at station D (at breakdown). (i) Group velocity c/U_1. (ii) Phase velocity c_p/U_1. (iii) Growth rate $\Omega^{(i)}\delta/U_1$. (iv) Primary phase velocity c_0'/U_1.

velocities generally were the lowest, i.e. at the primary wave crest. This is in accordance with the present theory since the group velocity of the secondary wave would be the lowest relative to the neighbouring troughs and the neighbouring valley regions where the primary wave velocity fluctuations would have their highest negative values. Because of the symmetry of the primary velocity field about the spanwise position of the peak the critical secondary waves would travel in the free-stream direction, i.e. would be unswept. For the calculations, polynomial expressions were fitted to the instantaneous profiles for each station as shown in figures 6 and 8 and the corresponding dispersion characteristics (figures 7 and 9) obtained by a numerical procedure (Landahl 1969). The computer program gave phase velocities directly (also shown in the figures) and from these group velocities were determined by graphical differentiation. Trial calculations with slightly different polynomial fits than those shown gave variations of the computed phase velocities of approximately $\pm 0{\cdot}02U_1$ (U_1 = free-stream

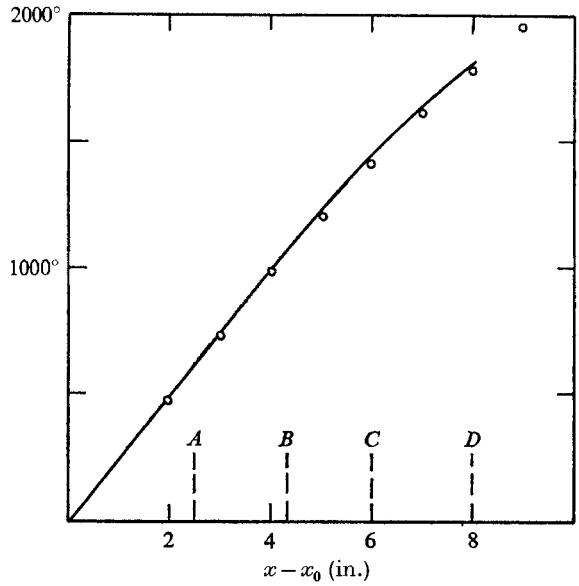

FIGURE 10. Phase of primary wave. ———, calculated from mean profiles; O, measurements, from I.

velocity). Moderate variations of the Reynolds number had very little effect on the results, thus confirming that the secondary waves are essentially inviscid.

The phase velocity of the primary wave could be inferred from the results of figure 7 in I (reproduced in figure 10). As mean profiles were also measured, however, a separate calculation was carried out to determine c_0 on the basis of two-dimensional linear stability theory. The comparison between the theoretically determined phase as function of downstream distance shown in figure 10 indicates good agreement. This might to some extent be fortuitous, however, since measurements of the phase variation across the boundary layer (figure 8 of I) indicate that, because of nonlinear effects, the troughs and crests move at somewhat different speeds. For the application of the present theory it is the phase velocity of the crest region which is of interest, but in the absence of this information or, alternatively, of the variation of the instantaneous profile with time and space around the breakdown region, the linear theory results were used.

It is interesting that the experiments show the breakdown to be preceded by a fairly strong acceleration of the primary wave, so that criticality is reached mainly through the catching up of the primary wave with the secondary wave group rather than through the slowing down of the latter. Because the primary phase speed in the neighbouring valleys is not much changed from its Blasius value, this acceleration will cause a considerable phase lead in the peak region. From figure 10 one can estimate that at breakdown the primary wave front in the peak region has moved about 1 in. ahead of the wave front in the valley. Since the distance from the peak to valley is only about $\frac{1}{2}$ in. there thus results a considerable warping of the primary wave just before breakdown. This phenomenon may be clearly observed in the experiments of Hama & Nutant (1963).

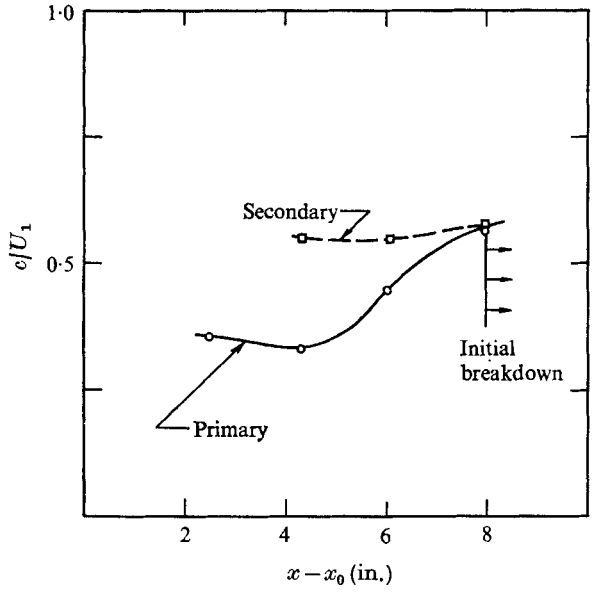

FIGURE 11. Comparison of calculated group velocity of secondary wave at neutral stability with phase velocity of primary wave. – –, c/U_1 at $\alpha = \alpha_s$; ——, c_0'/U_1.

The present calculations show that the warping can be explained on basis of linear stability theory and is a direct consequence of the deformation of the mean velocity profile, presumably due to the Benney–Lin mechanism (Benney & Lin 1960; Benney 1961), leading to a differential in the primary phase velocity between peaks and valleys. That the two-dimensional theory should give such good quantitative agreement with the measured phase velocities is perhaps somewhat surprising in view of the strong three-dimensionality induced by the warping. However, what apparently matters most is that the relative change of the mean profile should be small over a distance of one primary wavelength, in which case the kinematical wave theory described in § 2 would apply. According to this, the propagation velocities at the peak and the valley would be that of a wave with wave front normal to the free stream. To determine the changes in the primary wave amplitude one needs to include also the effects of the convergence and divergence of the rays as taken into account by the denominator J in (10). No calculations of the primary rays to determine this effect have been made, but the qualitative effects can be inferred from the diagrams of figures 10 and 11 since the group velocity follows the same trend as the phase velocity. Upstream of $x-x_0 \simeq 5$ in. the waves at the peak and valley will move at approximately the same speeds, and the effects of ray convergence will be slight, so that the growth rate will be approximately that given by the local dispersion relation for the mean profiles. Thus, the peak profiles, beginning to receive a noticeable inflexional region from station B on, will produce a much higher growth rate than those of the valley which become 'fuller' and more stable than the Blasius profile. From about $x-x_0 = 5$ in. on, however, the wave at the peak is rapidly speeding ahead of that of the valley causing a divergence of the rays from the

peak region and a convergence towards the valley. This may possibly explain why the primary wave in the valley region subsequently begins to grow rapidly (figure 4 of I) despite the linear growth rate of the valley profile being close to zero, and that the growth at the peak from station C on becomes substantially less than what the local dispersion relation would give.

In figure 11 the computed group velocity of the secondary wave at the crest for the wavenumber for neutral stability is plotted as function of downstream distance from the ribbon together with the phase velocity c_0' of the primary wave. Ignoring the effect of the primary wave growth rate, the critical condition should be reached when the two curves cross, which is seen to occur at $x-x_0 = 8$ in. in good agreement with the experiments. If the effect of primary growth rate is included (this rate being quite small at breakdown judging from figure 9 of I) the station for breakdown would move slightly upstream. (Acceleration of the primary wave has no effect on the breakdown condition, equation (42).) The experimental uncertainty in determining the location of breakdown may be about 0·5 in.,† as would be commensurate with the present theoretical model, in which the precision of the breakdown location would be limited to about one secondary wavelength ($\simeq = 0{\cdot}3$ in.). The inaccuracies in the application of the theory are therefore roughly within the experimental uncertainties, so that the agreement between theory and experiments as to the location of breakdown must be judged quite good even though the very precise agreement indicated in figure 11 should be regarded as fortuitous.

A clear indication that ordinary secondary instability cannot explain the observed breakdown location is provided by a comparison of the dispersion diagrams of stations C and D as shown in figures 7 and 9. The maximum amplification rate at station C (which if the effect of spatial inhomogeneities are disregarded is a rough measure of the total amplification during one primary cycle) is about the same as at station D; yet no breakdown is observed at the former, and the calculations do indeed clearly show that the secondary waves are far from the critical state at this station. This is also brought out in figure 12, which gives a comparison between the calculated frequency (which is obtained from the wavenumber and the phase velocity at the point of neutral stability, figure 9) and measured frequency of the secondary disturbance. The agreement is striking, whereas the frequency of the most amplified secondary disturbance is only about half of the measured value. The experimental values of the secondary frequency were obtained in I by counting the number of negative spikes per unit time in the oscillograms such as those of figure 27 in I. A closer inspection of the oscillograms reveals that the wave form indeed resembles that of a sinusoid with the positive half cycles missing as predicted from the discussions of the qualitative effects of nonlinearity presented in § 3.

The behaviour of the secondary disturbance immediately after breakdown can also be determined approximately from wave mechanical considerations. According to the analysis of § 2 the critical secondary wave group should propagate along a trajectory close to the limiting one downstream from the crest (OQ in figure 3). In calculating this trajectory, velocity profiles were employed

† P. S. Klebanoff (private communication).

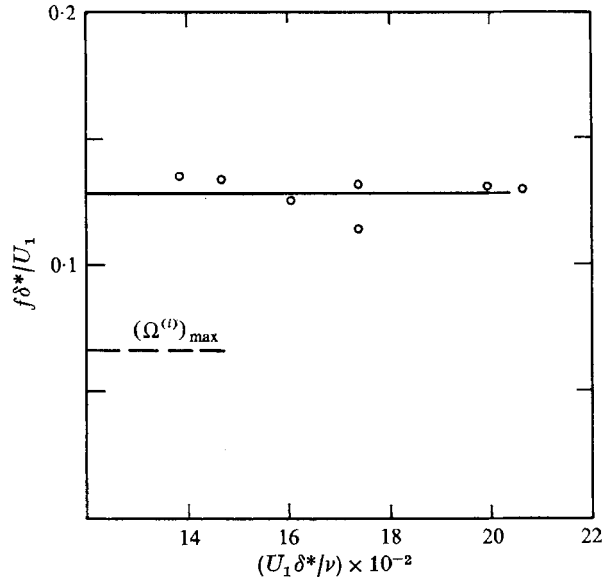

FIGURE 12. Comparison of calculated frequency of critical disturbance with experimental 'hairpin' frequency. f = frequency, δ^* = displacement thickness. ———, theory ($c = c_0'$); O, experiments, from I.

which were obtained by assuming the velocity at each value of y to vary sinusoidally in $\alpha_0\xi$ (α_0 being the wavenumber of the primary wave) between the value at the crest ($\alpha_0\xi = 0$) and that for the mean profile ($\alpha_0\xi = \frac{1}{2}\pi$) downstream of the crest. From the calculated dispersion relation for each such profile the group velocity was then determined for the wavenumber corresponding to the same frequency (as measured in a co-ordinate system following the primary wave) as that of the critical secondary wave according to (24) (hence taking the primary wave to be neutrally stable). The propagation velocity along the trajectory thus obtained is shown in figure 13 together with the estimates in I of the propagation velocities of the spike-like secondary disturbances. The agreement is surprisingly good in view of the crudeness of the calculation procedure with the theoretical values being somewhat on the low side. If one were to consider a Fourier decomposition of a single spike (consisting, say, of a half wave of the critical secondary wave disturbance) instead of an infinite wave train of the critical frequency, the propagation velocity would probably come out somewhat higher as the lower wavenumber regime having higher group velocities (and higher growth rates) would then also come into play. In particular, this may diminish somewhat the apparent deceleration towards the phase of the mean velocity profile of the primary wave ($\alpha_0\xi = \frac{1}{2}\pi$) seen in the diagram. As was pointed out above in § 3, however, the induced velocity due to the 'legs' of the hairpin vortex would have an increasingly stronger vertical component as the hairpin travels downstream, making its 'head' move further away from the wall and hence be accelerated. Thus, in the later stages of hairpin travel, one should expect larger departures from the two-dimensional theory. Nevertheless, the results shown in figure 13 suggest that simple wave kinematics could explain much of the observed

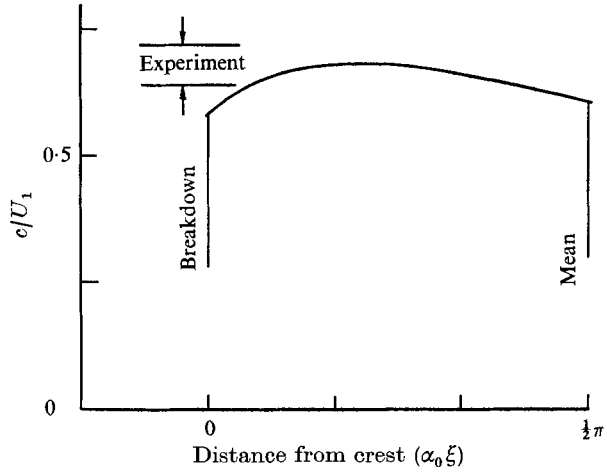

FIGURE 13. Propagation velocity of critical disturbance group compared with experimental 'hairpin eddy' velocity. ——, theory; $\frac{\downarrow\downarrow}{\uparrow\uparrow}$, experiments, from I.

behaviour of the spike, in particular its acceleration away from the crest immediately after it is formed.

Although direct quantitative comparisons of the present theory with experiments have been confined to those of I, some further qualitative comparisons can be made with other experiments on breakdown. The measurements of the instantaneous vorticity distribution carried out in the transition experiments by Kovasznay *et al.* (1962) are particularly revealing in the light of the present theory. In their experiments, controlled three-dimensional disturbances were introduced as in the measurements of I, and an array of hot wires was used to determine the instantaneous velocity profiles of the boundary layer at several instants during the primary oscillation cycle for different amplitudes of the oscillating ribbon. The results were displayed in such a way as to suggest the streamwise variation of vorticity (implicitly assuming that the primary wave growth rate was small so that it could be considered 'frozen' as it is convected past the position of the hot wires). Their figure 10 is reproduced in the present figure 14. In the earlier stages just preceding breakdown a strong concentration of vorticity in the middle part of the boundary layer is evident with values exceeding the maximum in the steady boundary layer. This excess is most likely due to stretching of vortex lines produced by the divergence of the primary wave ray lines from the peak resulting from the warping of the primary wave front as described above. This is borne out by figure 12 in their paper, which shows that the vortex stretching is large and positive only upstream of the crest. In the later stages of breakdown (= higher ribbon amplitude) the region of high vorticity downstream of the crest is seen to compress, to intensify and to develop a 'kink'. The compression and intensification of the downstream shear layer is most readily explained in the light of the present theory as being due to the convection of disturbance vorticity of positive sign by the secondary waves travelling upstream towards the crest during the initial nonlinear phase of the breakdown as described in §3 (see figure 5). The kink is just a manifestation of the development of the first secondary

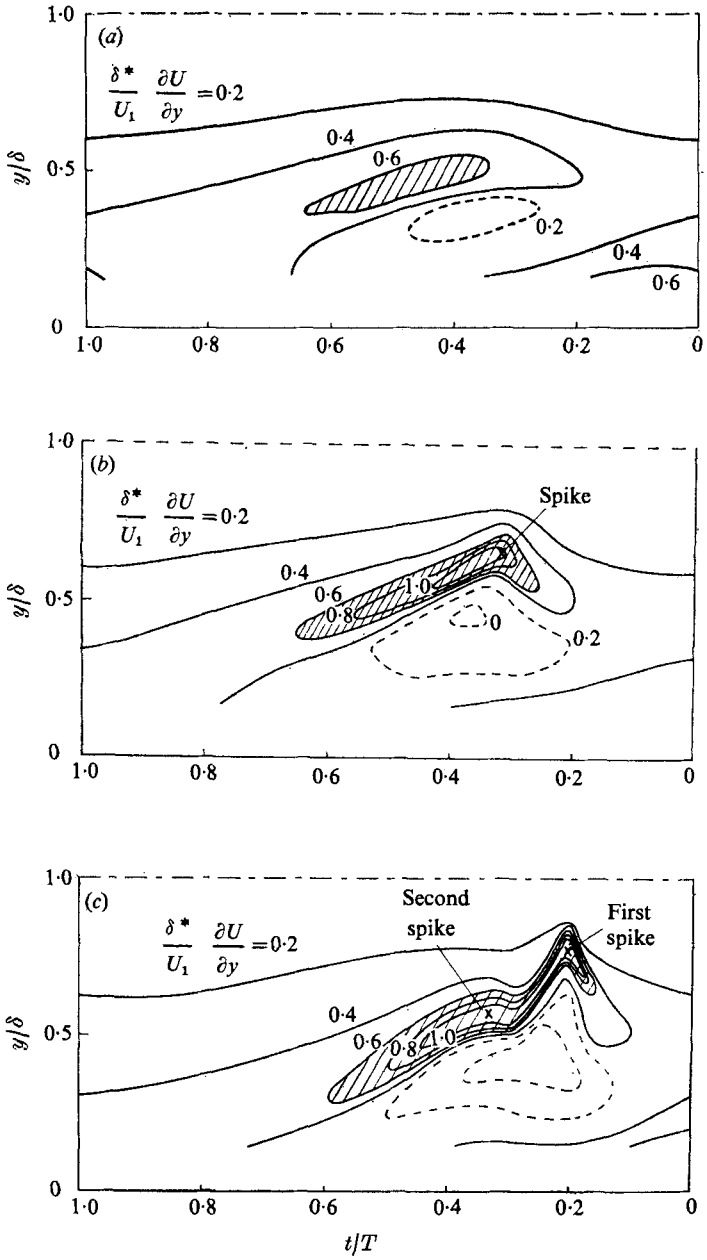

FIGURE 14. Contours of approximate vorticity at peak position (from Kovasznay *et al.* 1962). (*a*) Early development (no spike). (*b*) One-spike stage. (*c*) Two-spike stage. δ^* = displacement thickness, U = instantaneous velocity, T = period for one full cycle of ribbon oscillation.

half wave of large strength. The phase velocity of the critical secondary wave is very close to c_0' (see figure 9), so that to an observer riding with the primary wave it would appear like an almost steady wavelike warping of the shear layer. The hot wire located just below this kink would register a negative spike in the

oscillogram. With higher ribbon amplitude there is room for more than one spike, and the two-spike stage develops, etc. Kovasznay *et al.* (1962) also observed that the breakdown region has the shape of a blunt-nosed delta with its apex pointing downstream. As was pointed out in § 3, this is consistent with the wave picture, since the spanwise extent of the critical region grows when the primary wave moves downstream, and the initial secondary wave at the apex of the delta accelerates away from the crest as described above.

The experiments of Obremski & Fejer (1967) are especially interesting from the point of view of the present theory since the basic large-scale inhomogeneity was there produced by subjecting a Blasius boundary layer to a sinusoidally oscillating pressure gradient rather than by the Tollmien–Schlichting waves themselves. The high frequency disturbance appearing at transition seem to be Tollmien–Schlichting type instability waves modified by the periodic variations of the instantaneous velocity profile. These come in more gradually than in the oscillating ribbon case and do not show any spikes. Calculations of the total amplification of a wave packet through the free-stream oscillation cycle carried out by Obremski & Morkovin (1969) show that the measured frequency at breakdown is fairly close to that for the wave packet receiving the highest total amplification. (In these calculations, the effect of space–time focusing was omitted as was the change of frequency of the packet during the cycle resulting from the application of (4).) This result is not inconsistent with the present theory for the breakdown of an instability wave, since for a neutrally stable primary wave the secondary wave spending the largest time in the unstable region is that of marginal instability (see § 2). To see whether the critical condition of the present theory is satisfied, one would need to determine c_0 from (39), which requires the calculation of time and space derivatives of the group velocity. Some preliminary estimates based on the available calculated dispersion relationships for the family of profiles of interest (Obremski, Morkovin & Landahl 1969) indicate the possibility that c_0 could in some cases be sufficiently small during part of the primary cycle to lead to criticality. However, despite the fairly extensive set of calculated dispersion data, the tables available at present are not sufficiently detailed to allow the required computation of derivatives with any accuracy, so that no definite conclusion can as yet be drawn as to whether the present model applies to these experiments.

5. Discussion

The general theory for breakdown presented here is based on two fundamental assumptions: (i) that there is a disparity of length scales between primary and secondary wavelike disturbances, so that it becomes meaningful to consider the secondary waves as riding on the primary ones, and (ii) that the inhomogeneity due to the primary disturbance is slowly varying, so that kinematical wave theory applies. In the boundary-layer transition case the assumption (i) is satisfied because the primary Tollmien–Schlichting wave has an internal shear layer of small thickness $(\alpha_0 R)^{-\frac{1}{3}}\delta$ (where α_0 is the primary wavenumber) which is also subjected to the thinning effect of vortex stretching through the three-

dimensional development as described above. Inviscid stability theory predicts that the secondary wavelength at marginal stability is of the order of the shear-layer thickness. For a shear layer that is inviscidly unstable such as a jet or wake, there is no internal friction layer in the primary instability wave, and the disparate length scales do not arise. Thus, one would not expect breakdown in such shear flows, and experiments do indeed bear this out. The assumption (ii) is satisfied if $\Delta_1 = \epsilon_0 \lambda_1/\lambda_0$ and $\Delta_2 = \epsilon_0 \omega_1/\omega_0$ are both small, where ϵ_0 is the primary (non-dimensional) amplitude, λ_0 its typical wavelength, λ_1 the wavelength of the secondary disturbance, and ω_0 and ω_1 the corresponding primary and secondary frequencies, respectively. In the experiments of I, the ratio of wavelengths is about 1/8 (using either the streamwise or spanwise primary wavelength as a reference) and the primary fluctuation amplitude is about 0·1, giving for the above non-dimensional quantities $\Delta_1 \simeq 10^{-2} \simeq \Delta_2$, which should provide ample assurance for the validity of the kinematical wave theory. This theory may be considered the first term in an asymptotic series expansion in Δ, except that it does not deal with the first-order effects on the mode function. As we have only used it to determine under what conditions the secondary wave amplitude becomes large and hence the nonlinear effects dominant, however, this shortcoming of the kinematical theory need not concern us here.

The wave theory was shown to account for most of the observed features of breakdown in a boundary layer in surprising amount of detail. In particular, the theory may shed some new light on the question of the somewhat puzzling role of three-dimensionality. In the new theory three-dimensionality is mainly important in setting up the local flow conditions leading to criticality, whereas the breakdown process itself is predominantly two-dimensional and governed by the behaviour of the local instantaneous velocity profiles near the region of breakdown. Besides being consistent with the estimates of the scales of primary inhomogeneities given above, this point of view is also to some extent supported by the observations in I that the measured velocity profile at transition behind a roughness element bears a striking resemblance to the instantaneous one at breakdown behind an oscillating ribbon. However, the roughness-element case requires a great deal of further study before it can be fully understood, particularly in view of the very strong three-dimensionality of the flow in this case. To remove the restriction to weak spatial inhomogeneities in the present theory one would require a considerably more difficult analysis, which would possibly be worthwhile, however, in view of the important role of spatial inhomogeneities revealed by the present analysis. (Strong temporal inhomogeneity is more easily handled, as was demonstrated in the Greenspan & Benney (1963) treatment.)

In fact, the primary instability process itself is only incidental to breakdown and transition, since one can easily produce a breakdown condition by a sufficiently intense local disturbance of the shear flow, for example through free-stream disturbances. If the breakdown is to be self-maintained as it travels downstream, however, hydrodynamic instability is required. In a sense, the breakdown mechanism is the one most essential in transition of a boundary layer to turbulence rather than the classical hydrodynamic instability one since the

former represents a strongly irreversible process. An unstable small amplitude wave packet of the Tollmien–Schlichting type may amplify for a while as it moves downstream, but once it has passed through a particular streamwise position in the boundary layer it will leave the shear flow practically undisturbed as before the passage of the packet, except for a small deformation of the mean flow of second order in disturbance amplitude. At breakdown, on the contrary, there is an irreversible redistribution of the basic shear flow vorticity due to the nonlinear rectification mechanism. This suggests that the various methods proposed for delaying transition (boundary-layer suction, flexible walls, etc.) should be re-examined as to their effectiveness in influencing breakdown.

From the dynamical point of view, the almost instantaneous formation of the more-or-less concentrated hairpin vortex must be associated with a nearly impulse-like concentrated force acting on the flow at the centre of the hairpin in the upstream direction (cf. the concept of Kelvin impulse and its relation to circulation in ideal-fluid mechanics). This must correspond to a strong local shear stress impulse on the wall at the breakdown location. That the breakdown thus will produce a strong local contribution to the Reynolds stress is of interest in connexion with the recent findings by Kline and co-workers (Kline *et al.* 1967; Kim *et al.* 1971) that the major contribution to the Reynolds stress in a fully developed turbulent boundary layer comes from intermittent 'bursts' of turbulence, which seem to have many characteristics in common with the breakdown process in a laminar boundary layer. It has been demonstrated (Landahl 1967) that Tollmien–Schlichting type waves of random amplitude and phase are excited in the turbulent boundary layer, and in view of the present theory it is likely that they can interact in a linear or weakly nonlinear fashion so as to produce a critical condition leading to breakdown. Since the primary waves are found to be damped for the turbulent mean velocity profile, a nonlinear mechanism is required to produce them, and the breakdown mechanism itself is the most likely candidate for acting as a stirrer as it is strongly nonlinear. Considerable efforts would be required to analyse whether waves breaking down and thus triggering new waves can lead to a self-sustaining mechanism of turbulence and the details of this process should provide a fertile field for further study.

Although the main emphasis in this paper has been on shear flows, the theory suggests that breakdown should be a common phenomenon in all continuum systems, be they solid, liquid or gaseous, whenever wavelike disturbances of disparate length scales can arise. Such situations should be particularly abundant in geophysical fluid flows. One possible application from the field of wind-induced water waves can be mentioned. Capillary waves of very short wavelength may focus on a gravitational wave and thereby produce a strong nonlinear mechanism for transferring wind energy to long waves. Preliminary calculations along this line indicate that strong coupling between wind and water then may take place in regions outside the wave velocity regimes in which Miles's (1959) mechanism is operative. By drastically reducing the tension, for example by spreading oil on the surface (see Barger *et al.* 1970; Mollo-Christenson 1971), one can suppress this transfer mechanism and thus becalm the waters.

This work was supported in part by the Air Force Office of Scientific Research under Contract No. F 44620-71-C-0007. The investigation was initiated under the Woods Hole Oceanographic Institution 1970, Geophysical Fluid Dynamics Program supported by the National Science Foundation under contract GZ 1494. I am indebted to the many colleagues and friends who, through lengthy discussions, have helped shape my thoughts on the subject; in particular to Sheila Widnall and to Philip Klebanoff, to the latter also for making available the original data from I.

Appendix. Approximate solution of equation (32)

The approximate solutions of the derived ray equation (32) (Hayes 1970) are sought in the neighbourhood of O (figure 2). In this region $\Omega_{\alpha x}$ is small, so that the second term may be neglected. Furthermore, $\Omega_{\alpha\alpha}$ and Ω_{xx} may be approximated by constants, so that we may instead of (32) consider the simpler equation

$$\frac{d}{dt}\frac{\partial\alpha}{\partial x} = B\left(\frac{\partial\alpha}{\partial x}\right)^2 - C. \tag{A 1}$$

By introducing $p = (\partial\alpha/\partial x)^{-1}$ this can be transformed into the following separable equation:

$$dp/dt = Cp^2 - B. \tag{A 2}$$

Upon integration, the final result may be written as

$$\frac{\partial\alpha}{\partial x} = \left(\frac{C}{B}\right)^{\frac{1}{2}} \frac{1 - S\exp\left[2(BC)^{\frac{1}{2}}(t-t_a)\right]}{1 + S\exp\left[2(BC)^{\frac{1}{2}}(t-t_a)\right]}, \tag{A 3}$$

where

$$S = \frac{(C/B)^{\frac{1}{2}} - (\partial\alpha/\partial x)_a}{(C/B)^{\frac{1}{2}} + (\partial\alpha/\partial x)_a}. \tag{A 4}$$

It follows from this result that $\partial\alpha/\partial x$ can become infinite somewhere along the trajectory only if $-1 < S < 0$, i.e. if

$$(\partial\alpha/\partial x)_a > (C/B)^{\frac{1}{2}}. \tag{A 5}$$

For a ray that starts out along the axis $\xi = 0$, $(\partial\alpha/\partial x)_a = 0$ and $(\partial\alpha/\partial x)$ remains finite. Thus any ray that crosses $\xi = 0$ cannot focus.

REFERENCES

Barger, W. R., Garrett, W. D., Mollo-Christensen, E. L. & Ruggles, K. W. 1970 *J. Appl. Meteor.* **9**, 396.

Benney, D. J. 1961 *J. Fluid Mech.* **10**, 209.

Benney, D. J. & Lin, C. C. 1960 *Phys. Fluids*, **4**, 656.

Betchov, R. 1960 *Phys. Fluids*, **3**, 1026.

Emmons, H. W. 1951 *J. Aeron. Sci.* **18**, 490.

Gaster, M. 1968 *J. Fluid Mech.* **32**, 173.

Greenspan, H. & Benney, D. 1963 *J. Fluid Mech.* **15**, 133.

Hama, F. R. & Nutant, J. 1963 *Proc.* 1963 *Heat Transfer and Fluid Mech. Inst.* p. 77. Stanford University Press.

HAYES, W. D. 1970 *Proc. Roy. Soc.* A **320**, 209.

KIM, H. T., KLINE, S. J. & REYNOLDS, W. C. 1971 *J. Fluid Mech.* **50**, 133.

KLEBANOFF, P. S., TIDSTROM, K. D. & SARGENT, L. H. 1962 *J. Fluid Mech.* **12**, 1.

KLINE, S. J., REYNOLDS, W. C., SCHRAUB, F. A. & RUNSTADLER, P. W. 1967 *J. Fluid Mech.* **30**, 741.

KOVASZNAY, L. S. G., KOMODA, H. S. & VASUDEVA, B. R. 1962 *Proc.* 1962 *Heat Transfer and Fluid Mech. Inst.*, p. 1. Stanford University Press.

LANDAHL, M. T. 1967 *J. Fluid Mech.* **29**, 441.

LANDAHL, M. T. 1969 *Phys. Fluids*, **12** (suppl. II), 146.

LIGHTHILL, M. J. 1969 In *Computation of Turbulent Boundary Layers* (ed. S. J. Kline, M. V. Morkovin, G. Sovram & D. J. Cockrell), vol. 1, p. 511. Stanford University.

LIGHTHILL, M. J. 1970 *Osborne Reynolds and Engineering Science*, p. 83. Manchester University Press.

MILES, J. W. 1959 *J. Fluid Mech.* **6**, 568.

MOLLO-CHRISTENSEN, E. L. 1971 *A.I.A.A. J.* **7**, 1217.

OBREMSKI, H. J. & FEJER, A. A. 1967 *J. Fluid Mech.* **29**, 93.

OBREMSKI, H. J. & MORKOVIN, M. V. 1969 *A.I.A.A. J.* **7**, 1298.

OBREMSKI, H. J., MORKOVIN, M. V. & LANDAHL, M. T. 1969 *AGARDOgraph*, no. 134.

REYNOLDS, O. 1883 *Phil. Trans. Roy. Soc.* **174**, 935.

SCHAUBER, G. B. & SKRAMSTAD, H. K. 1948 *Nat. Adv. Comm. Aeron. Rep.* no. 909.

WHITHAM, G. B. 1965 *Proc. Roy. Soc.* A **283**, 238.