J. Fluid Mech. (1994), *vol.* 276, *pp.* 327–342

On the mechanism of shear flow instabilities

By PETER G. BAINES AND HUMIO MITSUDERA†
CSIRO Division of Atmospheric Research, Aspendale, Australia

(Received 7 April 1994)

In homogeneous and density-stratified inviscid shear flows, the mechanism for instability that is most commonly invoked and discussed is Kelvin–Helmholtz instability, as it occurs for a simple velocity discontinuity. There is a second mechanism, the wave-interaction mechanism, which is much more general, and is the subject of this paper. This mechanism depends on two free waves that propagate in opposite directions in a stratified shear flow, and which may become stationary relative to each other because of the shear. If this occurs, and their relative phase is suitably chosen, the velocity field of each wave increases the displacement of the other, and so the disturbance grows.

We show that this mechanism is responsible for instability in a general class of symmetric but otherwise arbitrary velocity and density profiles, provided that the Richardson number $R_i < \frac{1}{4}$ in a central region of arbitrarily small thickness. A critical layer exists in this central region for the growing disturbance, but its role in the instability process is incidental. When $R_i > \frac{1}{4}$ everywhere, the flow is stable because the free waves described above are absorbed by the critical layer, and hence are heavily damped. The necessary criteria of Rayleigh and Fjortoft for instability in homogeneous fluid are seen to provide a suitable geometry for two interacting waves. Some specific examples are given, including a succinct explanation of Holmboe waves.

1. Introduction

There is now an extensive literature on the nature of instabilities in stratified shear flows (see for example Drazin & Reid 1981), most of which consists of mathematical and numerical studies. Although the mathematics are well-developed, the physical mechanism that causes shear instability is not so well understood, in contrast to convective instability (for example) where the physical cause is obvious.

A number of mechanisms have been identified for various special cases. The first is Kelvin–Helmholtz instability of a velocity discontinuity, a kinematic self-advective process described in detail in Batchelor (1967). The second is the wave interaction mechanism, identified for particular cases of waves on two interfaces by Taylor (1931), Goldstein (1931) and Yih (1974). This process has been developed and generalized by Cairns (1979) with the introduction of the concept of 'negative energy waves' which are stable modes whose introduction into the flow causes a decrease in the total energy. Instability can result if such a mode resonates with another mode that has positive energy, which occurs when the waves have the same speed and wavelength. This can be identified by the crossing of dispersion curves for these modes in a frequency–

† Present address: JAMSTEC, Yokosuka, Japan.

wavenumber diagram. Whether a given wave mode has positive or negative energy depends on the frame of reference used, and the important property is the relation between two such modes. A survey of this topic is provided by Craik (1985), and generalizations using a Hamiltonian formulation have been given by Ostrovskii, Rybak & Tsimring (1986).

A different picture of the mechanism of shear instability has been presented by Lindzen in a series of papers with various colleagues, and this work has been summarized in Lindzen (1988). In this picture, shear flow instability is seen to be a rather complex process, involving over-reflection and the 'Orr mechanism', the latter being an advective process that causes transient growth in uniform shear; instability results because this transient disturbance is continually fed by an over-reflected wave. Unstable flows are therefore seen to be those that have a geometry that permit an over-reflected wave to continually energize the transient advective Orr process. The viability of this mechanism is the subject of current debate (e.g. Smyth & Peltier 1989; Takehiro & Hayashi 1992).

For homogeneous and stratified inviscid flows, general results about whether an arbitrarily chosen profile is or is not stable are limited to a small number of criteria that are *necessary* for instability. Specifically, these state that for the flow to be unstable we must have (e.g. Drazin & Reid 1981):

(i) the Richardson number $R_i = N^2/(\mathrm{d}U/\mathrm{d}z)^2$ less than $\frac{1}{4}$ at some level in the flow (the Miles–Howard criterion);

(ii) if $N = 0$ everywhere, $\mathrm{d}^2U/\mathrm{d}z^2$ changes sign at some level in the flow (Rayleigh's criterion);

(iii) again if $N = 0$, $\mathrm{d}^2U/\mathrm{d}z^2(U - U(z_i)) < 0$ at some level, where z_i is the level of the inflexion point in (ii) (the Fjortoft criterion); $U(z_i)$ may in fact be replaced by any other number.

In themselves, these criteria do not provide much insight into the mechanics of the instability process. Rather, one would expect that an understanding of the mechanism would help to explain their significance, and why they are necessary for instability.

The purposes of this paper are, first, to point out that the destabilizing wave interaction mechanism may be regarded as a purely kinematic advective process, paralleling that for Kelvin–Helmholtz instability. In our opinion, this gives more insight into the instability process than a picture based on waves with positive and negative energies. Secondly, it is shown that the wave interaction mechanism may be generalized to a wide class of general stratified shear flows that contain a layer in which $R_i < \frac{1}{4}$, and that a relatively simple *sufficient* condition for instability may be derived for such flows. In particular, we show how this process explains the above Miles–Howard, Rayleigh and Fjortoft requirements for instability, and why the flows are stable when $R_i > \frac{1}{4}$ everywhere. This leads to the hypothesis that almost all instabilities of statically stable horizontally uniform steady stratified shear flows may be attributed to this advective wave interaction mechanism. Other examples indicate that the same applies to rotating systems (Hoskins, McIntyre & Robertson 1985; Hayashi & Young 1987; Sakai 1989). Thirdly, some applications of this approach are made to the case of Holmboe instability, and to some profiles discussed by Huppert (1973).

2. Preliminary equations

We consider a stably stratified inviscid fluid which has velocity and density profiles $U(z)$ and $\rho_0(z)$ respectively in the undisturbed state, where x and z are horizontal and

vertical coordinates. The equations which govern small disturbances to this basic state are

$$\rho_0\left(\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial x}\right)u'+\frac{\mathrm{d}U}{\mathrm{d}z}w'\right)=-\frac{\partial p'}{\partial x}, \tag{2.1}$$

$$\rho_0\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial x}\right)w'=-\frac{\partial p'}{\partial z}-\rho' g, \tag{2.2}$$

$$\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial x}\right)\rho'+\frac{\mathrm{d}\rho_0}{\mathrm{d}z}w'=0, \tag{2.3}$$

$$\frac{\partial u'}{\partial x}+\frac{\partial w'}{\partial z}=0, \tag{2.4}$$

where u', w', p' and ρ' denote the perturbation quantities of velocity, pressure and density from the mean values U, 0, p_0 and ρ_0. Equation (2.4) implies that we may define a perturbation stream function ψ by

$$u'=-\frac{\partial\psi}{\partial z}, \quad w'=\frac{\partial\psi}{\partial x}. \tag{2.5}$$

If we make the Boussinesq approximation for convenience and eliminate all other variables in favour of ψ, we obtain

$$\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial x}\right)^2\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial z^2}\right)\psi+N^2\frac{\partial^2\psi}{\partial x^2}-\frac{\mathrm{d}^2U}{\mathrm{d}z^2}\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial x}\right)\frac{\partial\psi}{\partial x}=0, \tag{2.6}$$

where N is the buoyancy frequency, defined by $N^2=-(g/\rho_0)(\mathrm{d}\rho_0/\mathrm{d}z)$. As is standard practice for instability studies, we look for disturbances in the form of normal modes, namely

$$\psi=\hat{\psi}(z)\,\mathrm{e}^{\mathrm{i}k(x-ct)}, \tag{2.7}$$

and obtain the Taylor–Goldstein equation for $\hat{\psi}$

$$\mathrm{L}(\hat{\psi})\equiv(U(z)-c)^2\frac{\mathrm{d}^2\hat{\psi}}{\mathrm{d}z^2}+\left(N(z)^2-(U-c)\frac{\mathrm{d}^2U}{\mathrm{d}z^2}-(U-c)^2k^2\right)\hat{\psi}=0. \tag{2.8}$$

For given velocity and density profiles, solutions of this equation give eigenvalues for c and eigenfunctions for $\hat{\psi}$. If the former has a complex value, the flow has a growing mode and is deemed to be unstable. For any given solution, the vertical displacement η is given by

$$w'=\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial x}\right)\eta, \quad \text{so that} \quad \eta=\psi/(U-c). \tag{2.9}$$

2.1. *Free waves on interfaces*

An abrupt change in density $\Delta\rho_0$ of the mean flow at a particular level constitutes a density interface, and similarly an abrupt change in velocity gradient $\mathrm{d}U/\mathrm{d}z$ may be termed a *vorticity interface*. Both types of interface support waves. We consider a situation where two such interfaces coincide at $z=0$, but N and $\mathrm{d}^2U/\mathrm{d}z^2$ are otherwise zero in an infinite fluid. N^2 and $\mathrm{d}^2U/\mathrm{d}z^2$ then both have delta function behaviour at $z=0$, and solutions to (2.8) have the form

$$\hat{\psi}=\mathrm{e}^{-k|z|}, \tag{2.10}$$

with eigenvalues

$$c=U(0)-\frac{\Delta\zeta}{4k}\pm\left(\left(\frac{\Delta\zeta}{4k}\right)^2+\frac{g'}{2k}\right)^{1/2}, \tag{2.11}$$

where $g' = g\Delta\rho_0/\rho_0$, and $\Delta\zeta$, the discontinuity in vorticity, is given by $\Delta\zeta = (\mathrm{d}U/\mathrm{d}z)(-)-(\mathrm{d}U/\mathrm{d}z)(+)$, which will be taken to be positive or zero here. The system is stable, with neutral, propagating waves. If $\Delta\zeta = 0$, (2.10) and (2.11) show that there are two gravity waves of speed $(g'/2k)^{1/2}$, advected by the velocity at the interface. If instead $\Delta\rho_0 = 0$, one solution disappears and there is a single vorticity wave which propagates at the speed $\Delta\zeta/2k$. Associated with the wave is a sinusoidal vortex sheet, in phase with the displacement of the interface, which provides the mechanism of propagation. Positive η implies that upper-layer fluid has been replaced by lower-layer fluid (and hence by lower-layer vorticity) there, etc. This causes a velocity field which is $\frac{1}{4}$-wavelength out of phase with the displacement, resulting in a neutral wave propagating (in this case) to the left.

3. A simple prototype: the case of two vorticity interfaces

We next consider a system with two vorticity interfaces at $z = \pm d$ in a homogeneous fluid, where the fluid velocity varies linearly between them and is constant ($= \pm U_0$) for $|z| > d$, as shown in figure 1 (*a*). The vorticity increments at the interfaces, $\Delta\zeta = \pm U_0/d$, are equal and opposite. The solution (Rayleigh 1896) is

$$\hat{\psi} = \mathrm{e}^{-k|z-d|-\mathrm{i}\vartheta} - \mathrm{e}^{-k|z+d|+\mathrm{i}\vartheta}, \tag{3.1}$$

$$\left(\frac{c}{U_0}\right)^2 = \frac{(1-2kd)^2 - \mathrm{e}^{-4kd}}{(2kd)^2}, \tag{3.2}$$

where ϑ depends on c. If $0 < kd < k_s d \cong 0.64$, c is purely imaginary with $c = \mathrm{i}c_i$ (figure 1 *b*), and the flow is unstable; ϑ is then given by

$$\tan 2\vartheta = -\frac{2kdc_i/U_0}{1-2kd}, \tag{3.3}$$

and lies in the range $0 < \vartheta < \frac{1}{2}\pi$. The disturbance is stationary (in this coordinate system) but grows with time. As figure 1 (*b*) shows, the maximum growth rate lies near the middle of the range of kd, not far from the 'resonant' condition for stationary free waves, $kd = 0.5$.

The mechanism of the instability is essentially kinematic and is illustrated in figure 1 (*a*). The motion may be regarded as a pair of waves propagating in opposite directions on the vorticity interfaces, each being affected by the velocity field of the other. The displacements of each interface imply a sinusoidal vortex sheet perturbation as for (2.10) and (2.11), and (3.1) shows that the total velocity field is the sum of the velocity fields of each interfacial wave as if it acted in isolation. With the stationary phase configuration as shown in figure 1 (*a*), the net velocity field at each interface may be expressed as the sum of a component that is in phase with its displacement, plus a second component that is out of phase with it by $\frac{1}{2}\pi$. The first component acts to increase the amplitude of the displacement of the interface by advection, so that the disturbance grows with time. The instability has a finite bandwidth $0 < kd < k_s d$, because the second component affects the speed of the interfacial wave, acting to keep it stationary relative to the other wave. Because the vertical velocity field of each free wave leads its displacement by $\frac{1}{4}$-wavelength in its direction of propagation, the two waves are able to force each other symmetrically and in sympathy. As figure 1 (*a*) shows, this means that in the resulting growing disturbances the phase for the vertical displacement leans forward with increasing z, whereas that for the vertical velocity leans backward.

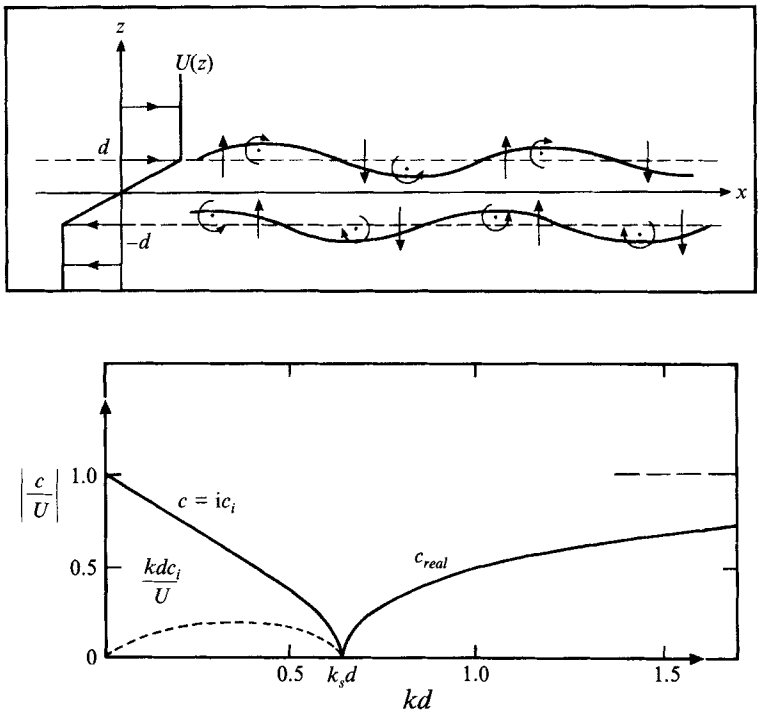

FIGURE 1. (*a*) The instability mechanism for two vorticity interfaces in a homogeneous fluid, with the velocity profile as shown on the left. The diagram shows interface displacements, the associated vorticity perturbation (circular arrows), and the peaks in vertical velocity (vertical arrows) at the interfaces. Note that these are displaced from the positions they would have if they were free waves. Waves on each interface are affected by the other in two ways. First, as in this unstable mode, they may alter the phase speed, so that the two waves are locked together. This occurs for a finite range of k, $0 \leqslant k \leqslant k_s$. Secondly, if this locking is achieved, they may amplify the wave by simple advection. The position of the vertical arrows between the nodes and the antinodes of the displacements indicates that both processes are happening here. (*b*) Wave speeds (real and imaginary, solid line) and growth rates (dashed line), from (3.2).

If the phase between the two waves shown in figure 1 is altered by an increment of π, the resulting disturbance would be damped rather than growing, corresponding to the other complex root for c. For $kd > k_s d$, the interaction between the two waves is too weak to be able to lock them together: ϑ is then pure imaginary, and the solutions are essentially two free waves, one on each interface, each weakly affected by the presence of the other wave. This interaction becomes weaker as kd increases. As $kd \to 0$, the mean flow tends to a vortex sheet, and the situation becomes that of the simplest case of Kelvin–Helmholtz instability, where the mechanism based on self-advection of a vortex sheet (Batchelor 1967) applies.

4. Generalization to arbitrary symmetric profiles

The above mechanism is also valid for a wide class of velocity and density profiles. We can demonstrate this by describing the properties of a specific type of general stratified shear flow. The development follows a framework outlined by Cairns (1979) and Drazin (1989). We consider a system as shown in figure 2, where the flow is confined between horizontal rigid boundaries at z_1, z_2 (which may be at infinity), and

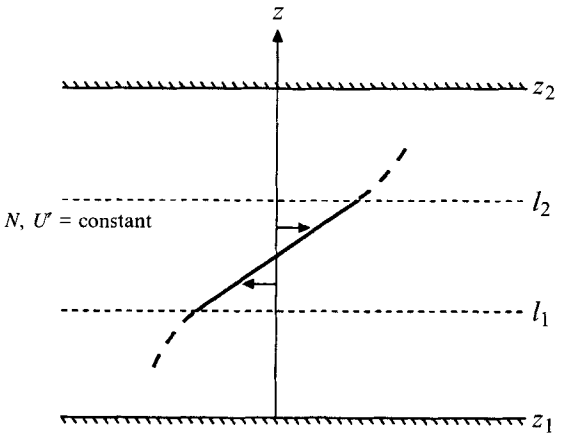

FIGURE 2. Configuration for the general case discussed in §4.

we may identify three flow regions, as follows. There is an uppermost region $l_2 < z < z_2$ within which the flow is stable, in the sense that it fails to satisfy any of the necessary criteria for instability in §1 (for example, $R_i > \frac{1}{4}$, or $\mathrm{d}^2U/\mathrm{d}z^2 \neq 0$ if $N = 0$, at all levels). The flow in the lowest region $z_1 < z < l_1$ is similarly stable, and these two regions are separated by a central region $l_1 < z < l_2$ in which N and $\mathrm{d}U/\mathrm{d}z$ are constant. To investigate the stability of such a flow, we look for eigensolutions $\hat{\psi}$ of (2.8) that satisfy the boundary conditions

$$\hat{\psi}(z_1) = 0 = \hat{\psi}(z_2), \tag{4.1}$$

and for which the flow is unstable if the corresponding eigenvalue c is complex.

For given values of k and c, we may define two solutions $\hat{\psi}_1(z, k, c)$ and $\hat{\psi}_2(z, k, c)$ of (2.8) such that $\hat{\psi}_1$ satisfies the lower boundary condition, and $\hat{\psi}_2$ the upper. Specifically, this means

$$\left.\begin{aligned} L(\hat{\psi}_1) = 0 \quad &\text{with} \quad \hat{\psi}_1(z_1) = 0, \\ L(\hat{\psi}_2) = 0 \quad &\text{with} \quad \hat{\psi}_2(z_2) = 0. \end{aligned}\right\} \tag{4.2}$$

This determines each of $\hat{\psi}_1$ and $\hat{\psi}_2$, apart from a multiplicative factor. For example, for given k and c, one may obtain a representative form for $\hat{\psi}_1$ by integrating (2.8) from $z = z_1$ with some arbitrarily chosen value of $\mathrm{d}\hat{\psi}_1(z_1)/\mathrm{d}z$. For either of $\hat{\psi}_1$ or $\hat{\psi}_2$ to be an eigensolution, it must satisfy both boundary conditions, which means that c and k must have values such that $\hat{\psi}_1 = \text{constant} \times \hat{\psi}_2$, or equivalently

$$W(\hat{\psi}_1, \hat{\psi}_2) = \hat{\psi}_1 \hat{\psi}_2' - \hat{\psi}_1' \hat{\psi}_2 = 0, \tag{4.3}$$

where $\hat{\psi}'$ denotes $\mathrm{d}\hat{\psi}/\mathrm{d}z$.

In the central region where the Richardson number is constant, we consider three specific situations with R_i having different values or ranges, the first of which is zero.

Case 1. $R_i = 0$ *in the central region*

Since $N = 0$ in this region, $\hat{\psi}_1$ and $\hat{\psi}_2$ here have the form

$$\left.\begin{aligned} \hat{\psi}_1 &= D_1(k, c)\exp(k(z-l_1)) + B_1(k, c)\exp(-k(z-l_1)) \\ \hat{\psi}_2 &= B_2(k, c)\exp(k(z-l_2)) + D_2(k, c)\exp(-k(z-l_2)) \end{aligned}\right\} \quad (l_1 < z < l_2). \tag{4.4}$$

The functions D_1 and B_1 depend on unspecified details of the flow in the lower region alone, and similarly for D_2 and B_2 for the upper region. However, if z_1 and l_1 are fixed

and the central region is broadened so that l_2 becomes large, the effects of the upper region on the lower are removed, and we have

$l_2 \to \infty$,

$$\left.\begin{aligned} &\hat{\psi}_1 \to B_1(k,c)\exp(-k(z-l_1)) \quad (z > l_1), \\ &D_1(k,c) = 0. \end{aligned}\right\} \tag{4.5}$$

D_1 is independent of the upper region, and $D_1 = 0$ gives the dispersion relation for the lower region in isolation. Similarly, if z_2 and l_2 are fixed and $l_1 \to -\infty$, we have

$l_1 \to -\infty$,

$$\left.\begin{aligned} &\hat{\psi}_2 \to B_2(k,c)\exp(k(z-l_2)) \quad (z < l_2), \\ &D_2(k,c) = 0. \end{aligned}\right\} \tag{4.6}$$

Hence, when the upper and lower regions become isolated independent waveguides, the dispersion relations for each are that D_1 and D_2 are zero. For the complete system, however, if $\hat{\psi}_1$ or $\hat{\psi}_2$ is to be a solution they must satisfy (4.3), which gives the dispersion relation

$$D_1 D_2 = \exp(-2k(l_2-l_1))\, B_1 B_2. \tag{4.7}$$

Solutions for c are eigenvalues for the complete system, and complex roots of (4.7) imply instability. In the central region $l_1 < z < l_2$, the solution takes the form of the sum of two free waves but with c determined by (4.7), as for the prototype case in §3.

We next consider flows where the velocity profile is anti-symmetric about a particular level (taken as $z = 0$), and the buoyancy is symmetric, so that

$$U(-z) = -U(z), \quad N^2(-z) = N^2(z). \tag{4.8}$$

With these relations one may readily show that solutions of (2.8) satisfy

$$\hat{\psi}(z,k,c) = \hat{\psi}(-z,k,-c), \tag{4.9}$$

and hence

$$D_2(k,c) = D_1(k,-c), \quad B_2(k,c) = B_1(k,-c). \tag{4.10}$$

If the lower region has a finite number of modes (n say), we may write

$$D_1(k,c) = \prod_{j=1}^{n} (c_j(k) - c), \tag{4.11}$$

where c_j is the speed of the jth free wave mode. More generally, for one particular mode, we may write

$$D_1(k,c) = (c_j(k) - c)\, d_1(k,c). \tag{4.12}$$

(4.7) may then be written in the form

$$D_1 D_2 = (c_j(k)^2 - c^2)\, d_1(k,c)\, d_1(k,-c) = \epsilon^2 B_1(k,c)\, B_1(k,-c), \tag{4.13}$$

so that

$$c^2 = c_j(k)^2 - \epsilon^2 \frac{B_1(k,c)}{d_1(k,c)} \frac{B_1(k,-c)}{d_1(k,-c)}, \tag{4.14}$$

where $\epsilon = \exp(-k(l_2-l_1))$. Suppose next that $c_j(k) = 0$, so that two free wave modes are stationary in this reference frame, for this k value. Then if ϵ is sufficiently small, (4.14) must have roots $c = \pm i c_i$, where

$$c_i = \epsilon \left| \frac{B_1(k,c)}{d_1(k,c)} \right|, \tag{4.15}$$

and hence the flow must be unstable. It is clear from the nature of this equation that instability does not depend on ϵ being small, since the growth rate increases with ϵ, and it is shown in the Appendix that if $c_j(k) = 0$, the flow is unstable for ϵ in the range $0 < \epsilon < 1$. If $\epsilon = 1$, this proof does not apply, and the question of instability depends on further details. It is also clear that instability does not depend on $c_j(k)$ being exactly zero. As for the prototype example of §3, in general there will be a finite bandwidth of unstable wavenumbers centred (approximately) on this criterion.

Hence, the essence of this instability process depends on the mutual interaction between two otherwise free waves that propagate in opposite directions, and this process works in the same way as for the simple example in §3. For these general shear flows, the essential characteristic of whether waves are propagating rightward or leftward is determined by whether or not the vertical velocity leads the vertical displacement in the relevant direction. This property is independent of the frame of reference, whereas the energy of the wave is not.

Case 2. $0 < R_i < \frac{1}{4}$ *in the central region*

For this case we may follow the same procedure as in Case 1 to obtain essentially the same result, although the details are more complex. If we define

$$\xi = k(z - c/U_z), \quad \xi_i = k(l_i - c/U_z) \quad (i = 1, 2), \tag{4.16}$$

where $U_z = \mathrm{d}U/\mathrm{d}z$ in the central region, then for $l_1 < z < l_2$ we may write $\hat{\psi} = \xi^{1/2}\phi(\xi)$, where ϕ satisfies

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}\xi^2} + \frac{1}{\xi}\frac{\mathrm{d}\phi}{\mathrm{d}\xi} - \left(1 + \frac{\nu^2}{\xi^2}\right)\phi = 0, \quad \nu = (\tfrac{1}{4} - R_i)^{1/2}. \tag{4.17}$$

The Hankel functions $H_\nu^{(1)}(\mathrm{i}\xi)$, $H_\nu^{(2)}(\mathrm{i}\xi)$ are solutions of (4.17), and these may be used to construct solutions corresponding to (4.4), which on the real axis of the ξ-plane have the form

$$\left.\begin{aligned} \hat{\psi}_1 &= \left(\frac{\xi}{\xi_1}\right)^{1/2}\left(D_1\frac{H_\nu^{(2)}(\mathrm{i}\xi)}{H_\nu^{(2)}(\mathrm{i}\xi_1)} + B_1\frac{H_\nu^{(1)}(\mathrm{i}\xi)}{H_\nu^{(1)}(\mathrm{i}\xi_1)}\right) \\ \hat{\psi}_2 &= \left(\frac{\xi}{\xi_2}\right)^{1/2}\left(D_2\frac{H_\nu^{(1)}(\mathrm{i}\xi)}{H_\nu^{(1)}(\mathrm{i}\xi_2)} + B_2\frac{H_\nu^{(2)}(\mathrm{i}\xi)}{H_\nu^{(2)}(\mathrm{i}\xi_2)}\right) \end{aligned}\right\} \quad (\xi_1 < \xi < \xi_2), \tag{4.18}$$

where $c = c_r + \mathrm{i}c_i$. As for Case 1, $D_1 = 0$ gives the dispersion relation for waves in the lower region, and $D_2 = 0$ for the upper region. Substituting the expressions (4.18) into (4.3) then gives the dispersion relation

$$D_1 D_2 - B_1 B_2 \frac{H_\nu^{(2)}(\mathrm{i}\xi_1) H_\nu^{(1)}(\mathrm{i}\xi_2)}{H_\nu^{(1)}(\mathrm{i}\xi_1) H_\nu^{(2)}(\mathrm{i}\xi_2)} = 0. \tag{4.19}$$

If we again restrict consideration to the symmetric profiles (4.8), the solutions satisfy (4.9), and D_i, B_i in (4.18) satisfy (4.10). In the case where a free mode of the upper or lower region is stationary, so that $c_j(k) = 0$ in (4.12), then writing $c = \mathrm{i}c_i$, we have

$$c_i = \left|\frac{B_1(k, c)}{d_1(k, c)}\right| \left|\frac{H_\nu^{(1)}(\mathrm{i}\xi_2)}{H_\nu^{(2)}(\mathrm{i}\xi_2)}\right|, \tag{4.20}$$

and the arguments for Case 1 again apply. In particular, the flow must be unstable if $2kl_2$ is sufficiently large, and this is due to the mutual interaction of two free modes.

Case 3. $R_i > \frac{1}{4}$ *in the central region*

Here we know from the Miles–Howard theorem that the flow is stable, and the interest centres on the way in which the mechanism of instability of Cases 1 and 2 breaks down. In this case, from (4.16) and (4.17), the general solution to (2.8) has the form

$$\hat{\psi} = \xi^{1/2}(A_1 I_{i\mu}(\xi) + A_2 I_{-i\mu}(\xi)), \tag{4.21}$$

where $\mu = (R_i - \frac{1}{4})^{1/2}$, and A_1 and A_2 are constants. Whereas $H_\nu^{(1)}(\mathrm{i}\xi)$ and $H_\nu^{(2)}(\mathrm{i}\xi)$ have monotonic near-exponential behaviour for ξ real, $I_{\pm i\mu}(\xi)$ are oscillatory, and the shear is weak enough to permit wave propagation in the region. If waves propagating on the upper and lower regions are to be stationary relative to each other, so that they can interact as for Cases 1 and 2, they must have a critical level in the central region, $l_1 < z < l_2$. A single wave on the upper or lower region that has a critical layer in the central region will have energy propagation toward the critical layer, with a decrease in amplitude of $\mathrm{e}^{-\mu\pi}$ and a phase change of $\frac{1}{2}\pi$ across it (Booker & Bretherton 1967). Such a wave embodies a constant flux of momentum $\rho_0 \overline{u'w'}$ toward (or away from, depending on sign) the critical layer, and this flux is discontinuous across it, decreasing in magnitude by the factor $\mathrm{e}^{-2\mu\pi}$. There is no reflection from the critical layer, and for this system to be steady, the incident wave must be maintained by some forcing due to, for example, flow over sinusoidal topography. In the absence of such forcing, there are no neutral wave modes on the upper and lower 'waveguides' because the waves approaching the critical layer are not reflected. If such a system with two waves of equal amplitude on each side of the central region is to be unstable, therefore, the waves must be able to force each other across the critical layer at sufficient strength to overcome the loss due to critical-layer absorption. The $\mathrm{e}^{-\mu\pi}$ factor weakens these effects, so that they are not sufficient to overcome the progressive disappearance of the waves due to critical-layer absorption, and consequently the flow is stable. It should be noted that it is still possible for waves on the upper and lower regions to assume a configuration where they can force each other in a mutually positive manner as in Cases 1 and 2, in spite of the changes in amplitude and phase of each wave across the critical level. Hence these changes alone do not preclude instability. Instead, we attribute the stability of the flows where $R_i > \frac{1}{4}$ everywhere to the fact that the discrete neutral modes do not exist.

5. A general sufficient criterion for instability

In the preceding section it is established that, for stratified shear flows where $N(-z) = N(z)$, $U(-z) = -U(z)$ (equation (4.8)), and there is a central region in which $R_i < \frac{1}{4}$ and is uniform that splits the flow into stable upper and lower waveguides, the flow is unstable if $c_j(k) = 0$ for some jth mode of either wave guide. Each of these wave guides is defined by extending the central region to infinity, in place of the other wave guide. This condition is sufficient for instability and not necessary, and a finite bandwidth of unstable wavenumbers is expected around it. This enables a number of deductions to be made about whether or not a given profile is likely to be unstable, as follows.

We assume that $R_i > \frac{1}{4}$ for $|z| > l_2 = |l_1|$, and consider the upper wave guide alone. From the properties of such waveguides (Bell 1974; Ince 1926), for each of them alone there is an infinite number of modes j that have phase velocities that lie outside the range of fluid speeds for all k, and have limit points at the maximum and minimum

fluid velocities, U_{max} and U_{min}, respectively, in the region where $R_i > \frac{1}{4}$. For the waves propagating leftwards with velocities less than U_{min}, $c_j(k)$ is a monotonic increasing function of k, with $c_j(k) \to U_{min}$ as $k \to \infty$. If $U_{min} = U(l_2)$, then the total flow (with. both wave guides) will be unstable if the fastest leftward-propagating mode of the upper wave guide (mode 1, say) has speed $c_1(0) < 0$. This guarantees that the above criterion for matching mode speeds can be achieved for some k, for this mode. Whether or not this criterion can be achieved for a given profile may often be assessed on a relatively simple basis, or by using hydrostatic layered models.

It follows that simple, monotonic profiles satisfying (4.8) will generally be unstable, unless the wave speeds are severely constrained by such things as horizontal boundaries. The cases studied by Taylor (1931), Goldstein (1931) and Yih (1974) are three simple examples that meet this criterion. These situations may be stabilized by introducing horizontal boundaries above and below the interfaces and close enough to them to restrict the wave speeds sufficiently. Also, it is clear that more complex profiles that fold back so that $U_{min} < 0$ for the upper wave guide will be stable by this mechanism, because an unstable mode can only have critical levels where $R_i < \frac{1}{4}$.

6. Examples

6.1. *One density and one vorticity interface – Holmboe instability*

The prototype configuration for Holmboe instability, as discussed by Holmboe (1962), is the velocity profile of figure 1 with a density interface at $z = 0$. This is the simplest model of a thin region of density variation embedded in a broader region of velocity variation. As shown by Holmboe, this system is subject to two types of instabilities: one stationary (i.e. non-oscillatory), and the other oscillatory, with travelling waves on the vorticity interfaces and a standing wave on the density interface. The essence of this second type may be seen more readily by considering the simpler system of a single vorticity interface and a density interface, as shown in figure 3(*a*). The interfaces are separated by a distance d, and the vorticity interface is advected at velocity U_0 relative to the density interface. The solution for disturbances in this system is

$$\hat{\psi}(z) = A\,\mathrm{e}^{-k|z-d|} + B\,\mathrm{e}^{-k|z|}; \tag{6.1}$$

c satisfies the cubic equation

$$(c^2 - c_1^2)(c - (U_0 - c_2)) + c_1^2 c_2\,\mathrm{e}^{-kd} = 0, \tag{6.2}$$

where $c_1^2 = g'/2k$ and $c_2 = U_0/2kd$, so that c_1 and c_2 are speeds of the free waves on the two interfaces, and

$$B/A = \mathrm{e}^{kd}(2kd(1 - c/U_0) - 1). \tag{6.3}$$

The last term in (6.2) couples the two interfaces together, and the waves are independent and freely propagating if this term is negligible. From the previous sections, we would expect instability to occur when an upper and a lower wave can be stationary relative to each other, and they propagate in opposite directions. Here this can only occur between the rightward-propagating wave on the lower interface, and the leftward propagating wave on the upper. This suggests that the region of instability should be centred on or near the line

$$c_1 = U_0 - c_2, \quad \text{or} \quad J \equiv \frac{g'd}{2U_0^2} = kd\left(1 - \frac{1}{2kd}\right)^2, \tag{6.4}$$

in $J - kd$ space. Figure 3(*b*) shows the computed growth rates, and it is clear that this

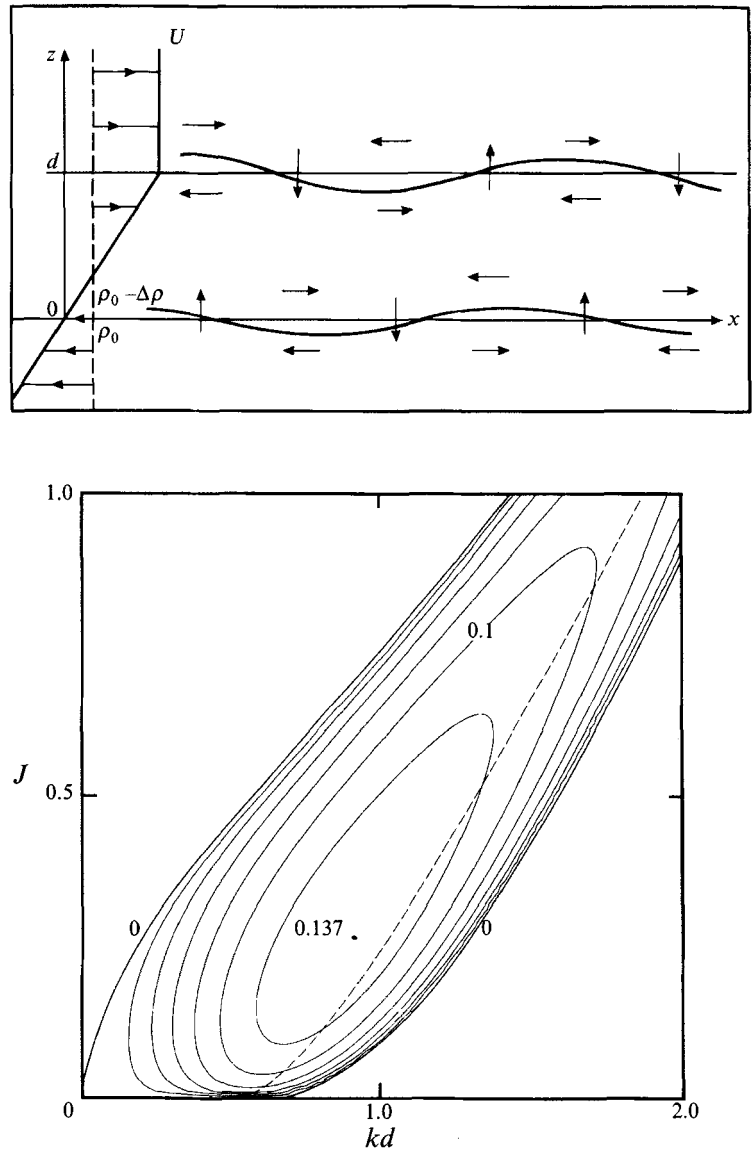

FIGURE 3. (*a*) As for figure 1(*a*) but for a vorticity and a density interface (at $z = 0$), as shown on the left. (*b*) As for figure 1(*b*). The clashed line denotes (6.4).

is as expected. Figure 3(*a*) shows the interfaces for a growing mode in a frame of reference moving with the mode, in which the instability mechanism is essentially the same as for the previous sections.

For the 3-interface case treated by Holmboe, the instability diagram in $J-kd$ space is shown in figure 4 for comparison. Here the oscillatory instability just described applies independently to the upper and lower interface pairs. As for the system of figure 3, for each J, kd, there is at most one unstable mode. The growth rate curves are only changed very slightly from those of figure 3(*b*), except where J is small, and the resulting oscillations on the centre interface are standing waves, for these three-interface modes. As J decreases, the frequency (kc_i) decreases to zero more rapidly than in figure 3(*b*), and there is a region where the unstable mode is stationary ($c_r = 0$). Here

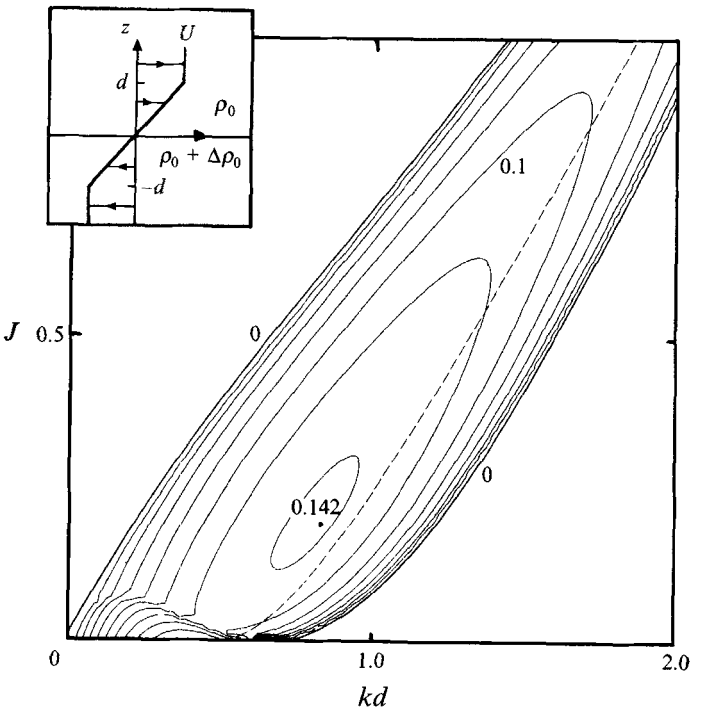

FIGURE 4. Growth rates in (J, kd)-space, as for figure 3(*b*), but for the symmetric Holmboe velocity profile with the density interface at $z = 0$. The clashed line denotes (6.4).

the mechanism is essentially the same as that of §3, and is dependent on the waves on the two vorticity interfaces. The density interface (at $z = 0$) acts to suppress this instability, so that the growth rate is less than that of §3 (figure 1*b*), except where $J = 0$ and they become equal.

6.2. *Huppert's profiles*

These examples are from Huppert (1973), and are included here in response to a challenge from a referee. We discuss the following three cases.

(i)
$$\left.\begin{aligned} U(z) &= U_0 \sin \pi z/d, \quad N^2 = N_0^2, \quad -d < z < d, \\ \psi &= 0 \quad \text{on} \quad z = \pm d, \end{aligned}\right\} \tag{6.5}$$

so that

$$R_i = \frac{J}{\cos^2 \pi z/d} \quad \text{where} \quad J = (N_0 d/\pi U_0)^2. \tag{6.6}$$

(i)
$$\left.\begin{aligned} U(z) &= U_0 \sin \pi z/d, \quad N^2 = N_0^2 \cos^2 \pi z/d, \quad -d < z < d, \\ \psi &= 0 \quad \text{on} \quad z = \pm d, \end{aligned}\right\} \tag{6.7}$$

so that $R_i = J$, given above, which is constant everywhere.

(iii)
$$\left.\begin{aligned} U &= U_0 z/d, \quad N^2 = N_0^2 z^2/d^2, \quad |z| < z, \\ \psi &= 0 \quad \text{at} \quad z = \pm d, \end{aligned}\right\} \tag{6.8}$$

so that $R_i = N_0^2 z^2/U_0^2 = J(\pi z/d)^2$, where $J = (N_0 d/\pi U_0)^2$. Here there is always a region near $z = 0$ where $R_i < \frac{1}{4}$.

Each of these cases has an anti-symmetric U and a symmetric N with respect to

$z = 0$, and hence the results of §4 are directly applicable, with the minor proviso that when $R_i < \frac{1}{4}$ in the central region near $z = 0$, it is non-uniform there. There is no reason to expect that this makes any qualitative difference to the processes acting if the central region is small and U' is approximately uniform within it, and we assume so here.

For case (i), R_i becomes less than $\frac{1}{4}$ near $z = 0$ and adjacent to $z = \pm d$ when $J < \frac{1}{4}$. Hence, $Ri(J)$, (4.20) and the argument in the Appendix yield that there must be instability centred on k values that give $c_r = 0$ in a one-sided problem involving only the upper or lower wave-guide. The region of instability in (α, J)-space (where $\alpha = kd/\pi$) for this profile has been computed by Hazel (1972, figure 14). The line on the (α, J)-plane where $c_r = 0$ for the one-sided problem (where the profile is (6.5) for $z > 0$, but with $U = U_0 \pi z/d$, $N = 0$ for $-\infty < z < 0$) lies within the unstable region as would be expected, not far from the line of maximum growth rate for the full profile (6.5). Case (ii) is essentially the same.

For Case (iii), the regions of instability on the (α, J)-plane have been computed by Huppert (1973, figure 2), and show complicated structure. These characteristics can be understood in terms of resonance between modes obtained from a one-sided problem, in which $N = 0$ is assumed for $R_i < 0.25$. For the full profile (6.8), instability with stationary growing modes is found in the range

$$(n-\tfrac{1}{2})^2+\alpha^2 < J < n^2+\alpha^2, \tag{6.9}$$

where $n = 1, 2, \ldots$, and $\alpha = kd/\pi$. For example, when $\alpha = 2.5$, stationary instability is found in the range $6.5 < J < 7.25$ for the first mode, and within $8.5 < J < 10.25$ for the second mode. For the above one-sided problem, stationary solutions ($c = 0$) are found at $J = 7.14$ and 9.82 for the first and second modes, respectively. These values lie within the bands of instability (6.9), and hence we conclude that the stationary unstable modes are formed by resonance between modes of the same order.

In the darkly shaded bands of Huppert's diagram, instability occurs with growing disturbances with $c_r \neq 0$. Numerical examination of the above one-sided problem shows that, for $\alpha = 2.5$, resonance between the first mode of the upper wave guide and the second mode of the lower wave guide occurs at $J = 8.0$, and between the upper second and lower third modes at $J = 11.4$. These values lie within the first and second dark instability bands, respectively, implying that this instability is again due to resonance but between modes of different order. The situation is similar to the Holmboe mode described in §6.1, where resonance with $c_r \neq 0$ occurs between unequal modes. Another example has been given by Sakai (1989, figure 11).

7. Summary and discussion

We have described a mechanism for the instability of a simple prototype shear flow in §3, in what we believe to be the simplest possible terms: if two free wave modes that propagate in opposite directions can be stationary relative to each other, the total flow can be represented as the sum of each wave, with altered phase and complex wave speed. The disturbance grows exponentially because the velocity field of each wave can advect the other to increase its displacement. The principal attribute of the shear is that it permits oppositely propagating waves to be relatively stationary. Since vorticity and gravity waves have the configuration that maximum upward velocity leads maximum displacement by $\frac{1}{4}$-wavelength, two such stationary waves are able to mutually amplify each other by simple advection. This physical picture is shown in §4 to apply to a broad class of arbitrary stratified shear flows where the velocity profiles are anti-symmetric. On the other hand, two waves that propagate in the same direction cannot force each

other in this fashion, although one may grow at the expense of the other. It is also easy to see why the familiar 'necessary' criteria listed in §1 are in fact necessary.

This kinematic advection mechanism is an alternative viewpoint to that based on energetics described by Cairns (1979) and Craik (1985). These are different ways of interpreting and describing the same process.

The workings of this mechanism help to explain the disjoint regions of instability in the $J-kd$ diagrams for some systems with continuous $U(z)$ and $\rho(z)$ profiles (Howard & Maslowe 1973), which may be understood as applying to wave modes with different vertical structure.

In the unstable flows considered here, a critical layer where $c_r = 0$ exists in the central region between the two wave guides. However, this critical layer plays no part in the instability process provided $R_i < \frac{1}{4}$ in the region, and its presence is an incidental consequence of the geometry.

At the fundamental level, there is no need to invoke the 'Orr mechanism' to account for the growth of the disturbance. The concept of 'over-reflection' is not required either, and over-reflection may be interpreted in terms of the same mechanism that causes instability here (Baines 1994).

On the other hand, the critical layer is important in rendering the flow stable when $R_i > \frac{1}{4}$. In this case, waves may propagate up to the critical layer, where they are absorbed into the mean flow (Booker & Bretherton 1967) without reflection. Neutral wave modes no longer exist in the upper and lower wave guides. Interaction across the critical layer is still possible, but is too weak to compensate for the decay due to critical-layer absorption, and this is the reason why the flow is stable.

The mechanism is different from the familiar mechanism of Kelvin–Helmholtz (KH) instability of a vortex sheet (as described in Batchelor 1967), but the latter is a limiting form of it. The end result of KH instability is the familiar KH billows, but in many cases the products of instability are instead likely to be internal (and vorticity) waves in stratified shear flows, quite possibly growing to nonlinear amplitudes. For example, numerical studies of the case of Holmboe waves (Smythe, Klaassen & Peltier 1988) have shown that some degree of overturning and mixing may occur, but the waves are maintained at finite amplitude, with a nonlinear profile, indefinitely.

The ocean is riddled with internal gravity waves, to the extent that a well-defined universal spectrum (the Garrett–Munk spectrum) has been identified (see for example Gill 1982). These waves may have a number of sources, but their existence in the deep ocean has never been satisfactorily explained. The results obtained here suggest that the primary source of these mid-ocean waves may be shear-flow instability. This suggestion is not new, but it has been substantially discounted in the past because the principal product of this instability has been seen to be KH billows, which are not internal waves as such, and collapsing billows do not generate them efficiently. In the light of the interaction mechanism, shear-flow instability will readily generate internal waves in a wide variety of situations, and the principal limitation would appear to be that we require the gradient Richardson number, R_i to be less than $\frac{1}{4}$ at some level in a sheared ocean current profile.

This framework also applies to barotropic instability in rotating flows, where a β-term is included in (2.8). In rotating baroclinic systems the same process (two stationary waves advecting each other) has been identified in general terms as the mechanism for baroclinic instability (see for example Hoskins *et al.* (1985), Sakai (1989)), with Eady waves being a prime example. Shear instability in other systems such as those discussed by Hayashi & Young (1987) and Takehiro & Hayashi (1992) may also be interpreted in the same fashion.

We emphasize that all of the instabilities described above may be interpreted in terms of one mechanism alone, namely the mutual forcing of two stationary, otherwise free, waves. This suggests that shear instability in general is due to this single mechanism, rather than a family of different processes.

Humio Mitsudera gratefully acknowledges the support of the Australian Research Council and Shell Australia.

Appendix

Proof that (4.14) *has complex roots* $c = \mathrm{i}c_i$ *with* $c_i > 0$ *when* $c_j(k) = 0$, *for* ϵ *in the range* $0 < \epsilon < 1$.

At $z = l_1$, in general we have

$$\hat{\psi}_1 = D_1(k,c) + B_1(k,c). \tag{A 1}$$

For given k and c, $\hat{\psi}_1(z)$ is determined apart from an arbitrary factor G, which may be identified as the initial gradient at $z = z_1$. Hence we may write

$$D_1 + B_1 = GF_1(k,c), \quad D_1 = (c_j - c)\, d_1 = (c_j(k) - c)\, G\bar{d}_1(k,c), \tag{A 2}$$

where F_1 and $\bar{d}_1$ are determined by k and c. With $c_j = 0$, substituting B_1 and d_1 from (A 2) into (4.14) gives

$$c^2 = \epsilon^2 c^2 - \epsilon^2(P + cQ), \tag{A 3}$$

where

$$P = \frac{F_1(k,c)\,F_1(k,-c)}{\bar{d}_1(k,c)\,\bar{d}_1(k,-c)}, \quad Q = \frac{F_1(k,-c)}{\bar{d}_1(k,-c)} - \frac{F_1(k,c)}{\bar{d}_1(k,c)}. \tag{A 4}$$

We look for solutions where $c = \mathrm{i}c_i$, in which case P is real and positive, and Q is imaginary for all real k. Writing $Q = \mathrm{i}\bar{Q}$, we then have

$$c_i = \frac{\epsilon^2\bar{Q} \pm [\epsilon^2\bar{Q}^2 + 4(1-\epsilon^2)\,\epsilon^2 P]^{1/2}}{2(1-\epsilon^2)}, \tag{A 5}$$

so that there is a positive value of c_i for all ϵ in the range $0 < \epsilon < 1$. For $c = \mathrm{i}c_i$, P and $\bar{Q}$ are well-behaved functions. $F_1(k,c)$ is the value of $\hat{\psi}_1$ at $z = l_1$, so that it is always finite, and $\bar{d}_1(k,c)$ is the dispersion relation for the lower wave guide (with one root factored out), so that its zeros for c are all on the real axis. Hence $\bar{d}_1$ is bounded away from zero, and P and $\bar{Q}$ are finite for all c_i.

REFERENCES

Baines, P. G. 1994 *Topographic Effects in Stratified Flows*. Cambridge University Press (to appear).

Batchelor, G. K. 1967 *An Introduction to Fluid Dynamics*. Cambridge University Press, 615 pp.

Bell, T. H. 1974 Effects of shear on the properties of internal gravity wave modes. *Deutsche Hydro. Zeits.* **27**, 57–62.

Booker, J. R. & Bretherton, F. P. 1967 The critical layer for internal gravity waves in a shear flow. *J. Fluid Mech.* **27**, 513–539.

Cairns, R. A. 1979 The role of negative energy waves in some instabilities of parallel flows. *J. Fluid Mech.* **92**, 1–14.

Craik, A. D. D. 1985 *Wave Interactions and Fluid Flows*. Cambridge University Press, 322 pp.

Drazin, P. G. 1989 Internal gravity waves and shear instability. In *Waves and Stability in Continuous Media* (ed. A. Donato & S. Giambò), commenda di Rende, Italy: EditEl, pp. 61–73.

Drazin, P. G. & Reid, W. H. 1981 *Hydrodynamic Stability*. Cambridge University Press, 525 pp.

GILL, A. E. 1982 *Atmosphere–Ocean Dynamics*. Academic Press, 662 pp.

GOLDSTEIN, S. 1931 On the stability of superposed streams of fluids of different densities. *Proc. R. Soc. Lond.* A **132**, 524–548.

HAYASHI, Y.-Y. & YOUNG, W. R. 1987 Stable and unstable shear modes of rotating parallel flows in shallow water. *J. Fluid Mech.* **184**, 477–504.

HAZEL, P. 1972 Numerical studies of the stability of inviscid stratified shear flows. *J. Fluid Mech.* **51**, 39–61.

HOLMBOE, J. 1962 On the behaviour of symmetric waves in stratified shear layers. *Geophys. Publik.* **24**, 67–113.

HOSKINS, B. J., MCINTYRE, M. E. & ROBERTSON, A. W. 1985 On the use and significance of isentropic potential vorticity maps. *Q. J. R. Met. Soc.* **111**, 877–946.

HOWARD, L. N. & MASLOWE, S. A. 1973 Stability of stratified shear flows. *Boundary-Layer Met.* **4**, 511–523.

HUPPERT, H. E. 1973 On Howard's technique for perturbing neutral solutions of the Taylor–Goldstein equation. *J. Fluid Mech.* **57**, 361–368.

INCE, E. L. 1926 *Ordinary Differential Equations*. Reproduced by Dover (1956), 558 pp.

LINDZEN, R. S. 1988 Instability of plane parallel shear flow (towards a mechanistic picture of how it works). *Pageoph.* **126**, 103–121.

OSTROVSKII, L. A., RYBAK, S. A. & TSIMRING, L. SH. 1986 Negative energy waves in hydrodynamics. *Sov. Phys. Usp.* **29**, 1040–1052.

RAYLEIGH, LORD, 1896 *The Theory of Sound*, vol. 2. Reproduced by Dover (1945), 504 pp.

SAKAI, S. 1989 Rossby–Kelvin instability: a new type of ageostrophic instability caused by a resonance between Rossby waves and gravity waves. *J. Fluid Mech.* **202**, 149–176.

SMYTH, W. D., KLAASSEN, G. P. & PELTIER, W. R. 1988 Finite amplitude Holmboe waves. *Geophys. Astrophys. Fluid Dyn.* **43**, 181–222.

SMYTH, W. D. & PELTIER, W. R. 1989 The transition between Kelvin–Helmholtz and Holmboe instability: an investigation of the overreflection hypothesis. *J. Atmos. Sci.* **46**, 3698–3720.

TAKEHIRO, S.-I. & HAYASHI, Y. Y. 1992 Over-reflection and shear instability in a shallow-water model. *J. Fluid Mech.* **236**, 250–279.

TAYLOR, G. I. 1931 Effect of variation in density on the stability of superposed streams of fluid. *Proc. R. Soc. Lond.* A **132**, 499–523.

YIH, C.-S. 1974 Instability of stratified flows as a result of resonance. *Phys. Fluids* **17**, 1483–1488.