Journal of Computational Physics **151,** 597–615 (1999)
Article ID jcph.1999.6211, available online at http://www.idealibrary.com on IDEAL®

On the Cancellation Problem in Calculating Compressible Low Mach Number Flows

Jörn Sesterhenn, Bernhard Müller, and Hans Thomann

Institut für Fluiddynamik, ETH Zürich, Sonneggstr.3, CH-8092 Zürich, Switzerland
E-mail: joern.sesterhenn@lrz.tu-muenchen.de

Received December 24, 1996; revised October 8, 1998

In calculating low Mach number flows one faces the stiffness problem in two different facets:

- the applicable time steps become very small
- the constants of the cancellation errors become very large $(1/\gamma M^2)$.

Usually the first point receives attention. Here we want to *concentrate on the cancellation problem only*. To our knowledge there is no detailed investigation of this problem in the related literature. In primitive variable formulations the problem can be solved by using the pressure coefficient instead of the pressure and a similar variable for the temperature or the internal energy. In conservative variable formulations this is thought not to be possible and therefore is sacrificed. We are able to show that a local reference state can also be introduced into a conservative scheme, if carefully applied to all quantities and applied to all constituent parts of the program. A detailed error analysis is given for all these parts. Finally, we show that we can perform a very low Mach number calculation at $M = 10^{-11}$ with a seven digits arithmetic only and still maintain the set of conservative variables. The governing equations are unaltered and the method depends neither on the time integration scheme nor the specific discretization. The method should be used in connection with the standard strategies like preconditioning, multigrid, or an (semi-)implicit method if acceleration is desired.

1. INTRODUCTION

In recent years the interest of many researchers has focused on the calculation of compressible low Mach number flows. Initially, the reason for doing so was the interest to use commercial codes in the incompressible limit [16], thereby closing a gap in the Mach number range, accessible by incompressible and compressible formulations. Since the time step for integrating the Euler equations is restricted by the fast acoustic waves, but the solution

0021-9991/99 $30.00

sought is usually also depending on the spreading of entropy waves, the number of time steps needed grows with $\mathcal{O}(1/M)$. Thus the point which primarily attracted attention was the stiffness of the equations, resulting in a large number of required time steps. We observe that the interest in compressible low Mach number flows has shifted to the case where density variations (e.g., created by combustion) are not vanishing for $M \to 0$.

Many proposals have been made to treat low Mach number flows. A good and extensive review can be found in an article by R. Klein [7]. Not considered in this review is the branch of methods trying to find a steady state solution by preconditioning the equations, partly in connection with multigrid methods as used by van Leer *et al.* [14] or implicit methods as in [3] by Chen and Pletcher.

Klein [7] requires that large amplitude density variations should be allowed, all waves should be promoted by their correct signal speeds, appropriate pressure variables for the thermodynamic and acoustic fluctuation range should be considered, and long wave acoustics with pressure amplitudes $\delta p/p_\infty = \mathcal{O}(\mathcal{M})$ should be accurately described.

In previous work we focused on the problem of the small time steps. Thereby we observed that even if the convergence can be accelerated substantially by preconditioning or multigrid methods, the accuracy severely deteriorates at very low Mach numbers. A multigrid scheme converges with a sufficient rate but we were only able to reduce the residual a few decades, depending on the Mach number.

Related observations or remarks are given in literature, see, for example, [7, Sect. 3.1; 16, Concluding Remarks; 5; 15]. Some of these authors are aware of cancellation as the source of the problem but to our knowlege no analysis of the numerical situation is reported in the literature.

Therefore this article entirely focuses on the *cancellation of the numerical solution* using modern finite volume schemes for the computation of low Mach number flows.

Numerical analysis is well aware of this process since in the pioneering days of computational sciences not even 6 digits were available to the scientist. The simple remedy is to introduce reference quantities into the equations and perform the calculation only on the fluctuations. A familiar variable of this type is the pressure coefficient $C_p = (p_o - p)/\varrho_o u_o^2$. As shown in Section 2 it is not sufficient to introduce a reference quantity in the momentum equation, but also in the energy equation and, if density changes of $\mathcal{O}(1)$ are present, in the continuity equation as well.

Although the concept sounds simple in principle, application to the equations of compressible fluid flow is not straightforward.

Several attempts in that direction are reported in the literature. Briley *et al.* [2] introduced in 1983 a pressure coefficient as an independent variable such that the Euler equations for constant enthalpy reduce to the incompressible Euler equations in the incompressible limit:

$$\tilde{p} = \frac{p}{\varrho_o u_o^2} - \frac{1}{\gamma M_o^2}.$$

This eliminated parts of the problem since cancellation errors in formulating the pressure flux can be avoided, as well as errors in the time integration scheme. However, they gave no error analysis and failed to see that the contribution of the kinetic energy to the total energy also gives rise to cancellation errors. This was observed by Hafez and Soliman [6] in 1991

and consequently they also introduced a reference temperature,

$$\tilde{T} = \frac{c_p T}{u_o^2} - \frac{1}{(\gamma - 1) M_o^2}.$$

Turkel *et al.* [13] reported a similar approach to that of Hafez *et al.* in 1993 but in connection with preconditioning methods. In the same year, Shuen *et al.* [11] used a pressure coefficient as Briley did, thus also omitting referencing the kinetic energy.

However, all methods need new independent variables and cannot keep the set of conservative variables (ϱ, $\varrho\mathbf{u}$, ϱe) as independent variables which is important when one has all Mach number codes in mind and wants to ensure the correct shock speeds.

These reference formulations have not yet entered formulations using the conservative variables although this would be highly desirable. The situation is further complicated by the fact that the speed of sound is also subjected to fluctuations. Thus it is even questionable whether a Riemann problem can be solved at all, without introducing cancellation errors at low Mach numbers. Guillard and Viozat [5] show theoretically the failure of the popular Roe scheme by means of low Mach number asymptotics.

The major aim of this paper is to carry the referencing idea further and to employ it rigorously in a conservative variable formulation. The proposed idea can be useful in codes designed for all Mach number flow. Since only the accuracy of the computation is addressed, one might want to use it in connection with acceleration techniques as described in the literature. The method can also be used to accurately compute acoustic problems as for example in noise reduction problems in the automobile industry. In these applications very small amplitudes have to be resolved, frequently on an almost incompressible background flow. In this case the proposed idea may be used to apply a conventional compressible flow solver to this kind of problem with moderate effort. (Apart from low Mach number flows a similar situation is encountered in relativistic flows where the rest energy mc_o^2 is tremendously high and plays a similar role as the high internal energy in the acoustic problems considered here. Also, in modelling turbulence with a k-ω model, ω takes a finite, but very high value at walls. In this case the cancellation process buries flux differences from the Euler part of the equations under the large values of ω.)

The present paper is organised as follows. First we demonstrate the mechanism of cancellation in a simple example: A pressure difference between two adjacent locations on a computational grid. Then we identify all occurrences of cancellation within an explicit finite volume code and show how the problem can be avoided by calculating the flow as a perturbation to a reference state. In Section 4 we explain how the reference state should be chosen. In the last section we validate the approach by comparing it with a two-dimensional physical test case.

2. CANCELLATION OCCURRING IN CALCULATING LOW MACH NUMBER FLOWS

In this section we want to recall the concept of numerical cancellation, shown in elementary books from numerical analysis [12]. We want to illustrate the occurrence of cancellation within the CFD-framework using a pressure difference between two grid locations in a smooth flow field. It is worth mentioning that this situation not only occurs in the momentum equation but also affects the energy equation.

The pressure in the left grid location is $\widehat{p_L}$, in the right it is $\widehat{p_R}$. Suppose, both are evaluated using an equation of state from conservative quantities as

$$\hat{p} = (\gamma - 1)\left(\widehat{\varrho e} - \frac{(\widehat{\varrho u})^2}{2\hat{\varrho}}\right). \tag{1}$$

Herein $\hat{\varrho}$ is the density, $\widehat{\varrho e}$ is the total energy per unit volume γ the ratio of the specific heats, and $\widehat{\varrho u}$ the momentum per unit volume. To characterize the order of magnitude of each term we introduce nondimensional quantities $p = \hat{p}/\gamma\hat{\varrho}_o\hat{c}_o^2$, $\varrho e = \widehat{\varrho e}\gamma/\hat{\varrho}_o\hat{c}_o^2$, $\varrho = \hat{\varrho}/\hat{\varrho}_o$, and $\varrho u = \widehat{\varrho u}/\hat{\varrho}_o\hat{u}_o$. The values $\hat{\varrho}_o$, $\hat{u}_o$, and $\hat{c}_o$ are chosen to be characteristic flow quantities. Also introducing a reference Mach number $M_o := \hat{u}_o/\hat{c}_o$, Eq. (1) becomes

$$p = (\gamma - 1)\left(\varrho e - \gamma M_o^2 \frac{(\varrho u)^2}{2\varrho}\right). \tag{2}$$

Now all the variables have the magnitude $\mathcal{O}(1)$, so we can readily see the influence of the Mach number, controlling the relative magnitude of the kinetic energy term. It may become arbitrarily small, provided the Mach number is small enough. With this in mind, we write

$$p_L = p_o \qquad \text{and} \qquad p_R = p_o + \delta p. \tag{3}$$

with $p_o = \mathcal{O}(1)$ and $\delta p = \mathcal{O}(M_o^2)$. Thus, the pressure difference becomes

$$p_R - p_L = (p_o + \delta p) - p_o. \tag{4}$$

Due to the limited number of digits available on a computer, a numerical calculation can only be performed with a limited relative accuracy. For example, evaluation of the pressure $p_R = p_o + \delta p$ yields the numerical result

$$(p_o + \delta p)(1 + \epsilon_1).$$

Every arithmetic operation introduces a relative error, here characterized by ϵ_1. ϵ_1 is not a constant and depends on the floating point representation used on the specific computer as well as on the arguments of the specific operation. It may be estimated as $|\epsilon_1| \leq 5 \cdot 10^{-t}$ when t decimals are available for representation of the mantissa. It is assumed that no exponent overflow occurs. This is a reasonable assumption. This error is called the *roundoff error* in the literature. It is unavoidable and has to be distinguished from the *cancellation error* which occurs as an accumulation effect of roundoff errors. Cancellation *can be avoided* to a certain extent by algebraically manipulating the formulas. This is demonstrated below. Introducing a relative error in every operation, Eq. (4) yields

$$\Delta p = ((p_o + \delta p)(1 + \epsilon_1) - p_o)(1 + \epsilon_2) \tag{5}$$

for the computed pressure difference. Extracting the known exact result δp in Eq. (5) we find

$$\Delta p = \delta p\left(1 + \frac{p_o + \delta p}{\delta p}\epsilon_1 + \epsilon_2 + \mathcal{O}(\epsilon_1\epsilon_2)\right), \tag{6}$$

ϵ_2 being the error introduced in the second operation. The leading error term is

$$\frac{p_o + \delta p}{\delta p}\epsilon_1 \approx \mathcal{O}\left(\frac{1}{M_o^2}\right)\epsilon_1. \tag{7}$$

If this term is comparable to one, the relative error dominates the evaluation. We observe that the relative error is scaled by $1/M_o^2$ which explains the cancellation mechanism at small Mach numbers. Since δp is the pressure difference associated with the distance between two grid points, we also expect an error dependency like $1/\Delta x$: Refining the mesh will lead to even worse results.

We give a numerical example. Calculating a flow at $M_o = 10^{-3}$ with single precision arithmetic, i.e., $|\epsilon_1| \leq 5 \cdot 10^{-7}$, will not preserve a single digit of the pressure difference. The calculation is already harmed at considerably higher Mach numbers as we will show in Figs. 6 and 7.

We want to stress that this does not affect the momentum equation only. The nondimensional expression for the total energy density is given by

$$\varrho e = \varrho\left(\epsilon + \gamma M_o^2 u^2\right), \tag{8}$$

where $\epsilon = \hat{\epsilon}\gamma/c_o^2$ is the nondimensional internal energy and is $\mathcal{O}(1)$ by choice of the reference quantity. We can observe the same situation as in Eq. (2), so that the energy equation will be affected by cancellation as well. This is also true if one uses primitive formulations with the temperature, the speed of sound, or similar quantities as the principal variable.

The situation for the continuity equation is slightly better: Gibbs fundamental equation $dh = T\,ds + dp/\varrho$ can, by use of the equation of state, be written as

$$dp - c^2/\gamma\; d\varrho = p\,ds. \tag{9}$$

Here s in nondimensionalized by the specific heat at constant volume. We can see that density changes have the same order as the pressure changes, as long as no $\mathcal{O}(1)$ entropy changes are present. If there are, referencing the continuity equation is necessary too.

3. HOW TO AVOID CANCELLATION

From Eq. (4) it is apparent how to avoid cancellation. We can rearrange the evaluation of Eq. (4) as

$$p_R - p_L = (p_o + \delta_p) - p_o = (p_o - p_o) + \delta p. \tag{10}$$

The numerical result now yields

$$\begin{aligned}\Delta p &= ((p_o - p_o)(1 - \epsilon_1) + \delta p)(1 + \epsilon_2) \\ &= \delta p\left(1 + \frac{p_o - p_o}{\delta p}\epsilon_1 + \frac{p_o - p_o}{\delta p}\epsilon_2 + \epsilon_2 + \frac{p_o - p_o}{\delta p}\epsilon_1\epsilon_2\right) \\ &= \delta p(1 + \epsilon_2).\end{aligned} \tag{11}$$

Thus the numerical error is as small as the roundoff error. This is the best we can hope for. Please note that this result does not depend on $(p_o - p_o)$ dropping out. If p_o would be different on both sides, p_{oR} and p_{oL} say, the resulting coefficients $(p_{oR} - p_{oL})/\delta p$ on the second line of Eq. (11) would still be of magnitude $\mathcal{O}(1)$ instead of $\mathcal{O}(1/M^2)$

In a common finite-volume code for compressible flows are three operations leading to cancellation:

- the time integration
- the determination of the flux-difference
- the determination of the fluxes.

In the following sections, we first explain the concept and then show how to avoid cancellation for each item. Please observe that the solved equations remain unaltered.

3.1. *The Finite-Volume Setting*

Let the conservative law be denoted as

$$\frac{d}{dt}\int_{\Omega} u\, d\omega + \int_{\partial\Omega} f(u)\, ds = 0. \tag{12}$$

$u = (\varrho, \varrho u_i, \varrho e)^T$ is the vector of conserved quantities and the flux function $f(u)$ has the components ϱu_k, $p\delta_{ik} + \varrho u_i u_k - \tau_{ik}$ and $\varrho u_k(e + p/\varrho) - u_i\tau_{ik} - \lambda(\partial T/\partial x_k)$ in tensor notation, using the common symbols. An explicit finite volume approximation reads

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Omega_i}\sum F^n(U)\cdot S. \tag{13}$$

Readers not familiar with these concepts will find them, e.g., in the book by LeVeque [8]. U_i^n is the approximation to the conserved quantity u within the cell i at the time level n, Ω is the volume of the cell, $F(U)$ the numerical flux function across each interface, and S a vector with the direction of the outward normal. Its length is the absolute value of the surface area. The sum extends over all sides of the control volume.

The basic idea of this paper is to introduce a suitable reference state U_o and a perturbation δU to U_o into the governing equations.

3.2. *The Error Introduced via the Time Integration*

Inserting the perturbation δU to the reference state U_o into Eq. (13) gives

$$(U_o + \delta U)_i^{n+1} = (U_o + \delta U)_i^n - \frac{\Delta t}{\Omega_i}\sum F^n(U_o + \delta U)\cdot S. \tag{14}$$

If U_o is well chosen (cf. Section 4) δU is small. In the following a numerical evaluation is introduced which takes advantage of this. We restrict our considerations to a reference state $U_o = U_o(x, y, z)$ which is constant within time scales imposed by the acoustic waves. It might be readjusted according to the slow entropy waves. We subtract the reference state from both sides of Eq. (14):

$$\delta U_i^{n+1} = \delta U_i^n - \frac{\Delta t}{\Omega_i}\sum F^n(U_o + \delta U)\cdot S. \tag{15}$$

This already results in a numerically better scheme. To estimate the improvement, we look at the magnitudes of the different terms in Eqs. (14) and (15). Suitably scaled $U_o = \mathcal{O}(1)$, $\delta U = \mathcal{O}(\delta)$, and $\frac{\Delta t}{\Omega}\sum F\cdot S = \mathcal{O}(\delta)$. Herein, δ is small and decreases at least with M, for low Mach number flows. Thus all summands in Eq. (15) are of the same order, whereas in the original equation, cancellation occurs. The last formulation can be implemented in an existing code with little effort. We found the numerics of the resulting scheme to be substantially improved at Mach numbers as high as $M = 0.01$.

3.3. *The Error Introduced via the Flux-Difference*

Next, the error introduced via the flux-difference has to be treated. Although we expect $\frac{\Delta t}{\Omega}\sum F\cdot S$ to be small as $\mathcal{O}(\delta)$ it usually contains differences of large numbers. We present two approaches which avoid this situation, one calculating the flux-difference Δf directly, and the other using the wave-propagation idea.

Calculating the flux-balance directly. First, we want to explain the concept by an example in one space dimension on a uniform grid. The continuity equation in this case becomes

$$\varrho^{n+1} = \varrho^n - \frac{\Delta t}{\Delta x}(\varrho_R u_R - \varrho_L u_L). \tag{16}$$

$(.)_L$ and $(.)_R$ denote states at the left- and right-hand cell-interface. They have to be provided by a Riemann solver. Introducing a perturbation to a reference state ϱ_o, u_o we find

$$(\varrho_o+\delta\varrho)^{n+1} = (\varrho_o+\delta\varrho)^n - \frac{\Delta t}{\Delta x}((\varrho_o+\delta\varrho)_R(u_o+\delta u)_R - (\varrho_o+\delta\varrho)_L(u_o+\delta u)_L). \tag{17}$$

After rearrangement and subtraction of ϱ_o from both sides of Eq. (17)

$$\delta\varrho^{n+1} = \delta\varrho^n - \frac{\Delta t}{\Delta x}\left(\underbrace{((\varrho_o u_o)_R - (\varrho_o u_o)_L)}_{\Delta F_o} + \underbrace{((\delta\varrho u_o)_R - (\delta\varrho u_o)_L) + ((\varrho_o\delta u)_R - (\varrho_o\delta u)_L) + ((\delta\varrho\delta u)_R - (\delta\varrho\delta u)_L)}_{\Delta(\delta F)}\right). \tag{18}$$

We can identify mixed products of the reference state and the perturbations. Here we have grouped terms together that we expect to be of equal magnitudes. Thus we subtract quantities of magnitude $\mathcal{O}(1)$, $\mathcal{O}(\delta)$, and $\mathcal{O}(\delta^2)$ separately, thus avoiding cancellation. The example above is one-dimensional but the same procedure can be applied independent of this constraint. Generally, the flux-balance of Eq. (15) is split into

$$\sum F(U_o+\delta U)\cdot S = \sum F_o(U_o+\delta U)\cdot S + \sum \delta F(U_o+\delta U)\cdot S \tag{19}$$

as indicated in Eq. (18). Equation (15) becomes

$$\delta U_i^{n+1} = \delta U_i^n + \frac{\Delta t}{\Omega_i}\sum F_o\cdot S + \frac{\Delta t}{\Omega_i}\sum \delta F_o\cdot S. \tag{20}$$

The sum $\sum F_o\cdot S$ is the finite volume expression for the integral $\int_{\partial\Omega} f_o\,dS$. If the background flux yields $\frac{\partial U}{\partial t}=0$, this part will drop out.

Please note that Eq. (20) is not determined uniquely. It gives only the final form of the integration scheme. Also note that the states at the left- and right-hand cell-interfaces in Eq. (18) are undetermined up to now.

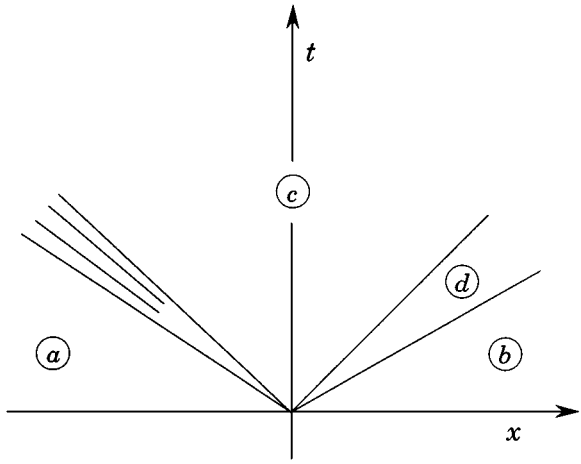

FIG. 1. Evolution of a Riemann problem.

Wave propagation. Reference [8] is an alternative approach for calculating the flux differences. The bulk contribution to the flux flowing homogeneously *through* the control volume drops out in a even more natural fashion. Consider a Riemann problem as depicted in Fig. 1 as it is assumed at each cell interface.

In order to find a suitable formulation for flux-difference splittings, we rewrite the flux difference $f_b - f_a$ across an evolving Riemann problem as

$$f_b - f_a = (f_b - f_d) + (f_d - f_c) + (f_c - f_a) = \sum \Delta f^+ + \sum \Delta f^-. \qquad (21)$$

Δf^+ resp. Δf^- denotes the flux-differences which are associated with right- or left-running waves. a and b are the initial left and the right states. Solving the Riemann problem one finds the intermediate states c and d. We always look for the state on the time axis, which may lie in any of the flow regions a, c, d, or b, depending on the flow direction and Mach number.

Now consider Fig. 2. From the viewpoint of a particular cell i, the fluxes at the left and right cell interfaces may be expressed as

$$f_{i-1/2} = f_i - \sum \Delta f_L^+ \qquad (22)$$

$$f_{i+1/2} = f_i + \sum \Delta f_R^-. \qquad (23)$$

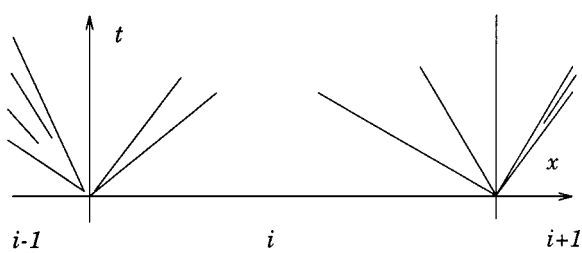

FIG. 2. Riemann problems at different sides of a control volume.

The indices L and R indicate a left and right interface of the cell. The integration may thus be expressed as

$$\delta U_i^{n+1} = \delta U_i^n + \frac{\Delta t}{\Omega_i} \sum_k \left(F_i + \sum \Delta F^{\pm} \right)_k \cdot S_k. \tag{24}$$

k indicates all cell interfaces, and $\Delta F^{\pm}$ is constructed in accordance to Eqs. (22) and (23) with an exact or approximate Riemann solver. Since all expressions are based on cell i, we again can extract $\sum_k F_i \cdot S_k = 0$. The following equation remains,

$$\delta U_i^{n+1} = \delta U_i^n + \frac{\Delta t}{\Omega_i} \sum_k \sum \Delta F_k^{\pm} \cdot S_k. \tag{25}$$

It is still algebraically equivalent to Eq. (13).

Both approaches avoid cancellation and the preference may be given according to the Riemann solver in use and the particular implementation.

3.4. *The Error Introduced via the Flux-Evaluation*

In the above discussion we have not yet answered the question of how to calculate the cell interface values in Eq. (18) or equivalently the jumps ΔF in Eq. (25). It is not obvious that these quantities can be calculated at all for $M \to 0$ without introducing cancellation errors since the correct sound speed and fluid velocity must be kept. Indeed, Viozat *et al.* [15, 5] showed that the Roe-solver is inconsistent with the zero Mach number limit in its original formulation. For that purpose, we propose a characteristic-based Riemann solver. The method is very well suited for the calculation of low Mach number flows. It is not recommended in the presence of strong shocks. The method is very efficient since it employs only a small number of essential operations. (We want to stress that the basic idea of this paper is not restricted to this method in peculiar and can be applied to other Riemann solvers too.)

The following description of the fluxes is given in terms of cell interface values at time $t + \Delta t/2$, rather than in the flux-difference form.

The Euler equations in one space dimension can be decoupled into three ordinary differential equations along distinct lines in time and space, called characteristics. These characteristics have the slopes

$$dx/dt = u - c =: \lambda_1,$$
$$dx/dt = u =: \lambda_2,$$

and

$$dx/dt = u + c =: \lambda_3.$$

The corresponding three ODEs along these characteristics read

$$dp - \varrho c\, du = 0 \tag{26}$$
$$dp - c^2\, d\varrho = 0 \tag{27}$$
$$dp + \varrho c\, du = 0. \tag{28}$$

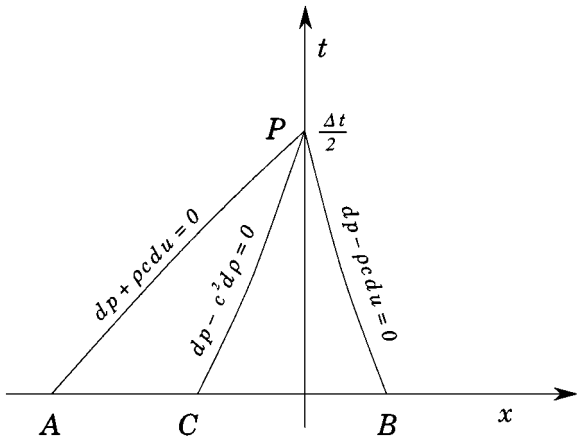

FIG. 3. Method of characteristics.

The situation is illustrated in Fig. 3. Assuming isentropic changes along each characteristic, these ODEs could be integrated and the solution found by solving the resulting nonlinear system in entropy, velocity, and speed of sound. More easily the state at point P can be evaluated by approximating the characteristics with straight lines and freezing the values ϱc and c^2 at their root values, denoted A, B, and C. This approach is more general than the former because no additional assumptions are made. For higher accuracy or if sources are present on the right-hand sides an iteration scheme suggested by Courant and Friedrichs [4, p. 73] can be applied. Details of this linearization in the phase space (u, p, ϱ) are given in [10, Abb. 2.5]. The result is a linear 3×3-system

$$\begin{aligned}(p_P - p_A) + \varrho_A c_A (u_P - u_A) &= 0\\ (p_P - p_B) - \varrho_B c_B (u_P - u_B) &= 0\\ (p_P - p_C) - c_C^2 (\varrho_P - \varrho_C) &= 0.\end{aligned} \tag{29}$$

In matrix notation

$$Av + b = 0 \tag{30}$$

with

$$A = \begin{pmatrix} 1 & \varrho_A c_A & 0 \\ 1 & -\varrho_B c_B & 0 \\ 1 & 0 & -c_C^2 \end{pmatrix}; \qquad v = \begin{pmatrix} p_P \\ u_P \\ \varrho_P \end{pmatrix}; \qquad b = \begin{pmatrix} -p_A - \varrho_A c_A u_A \\ -p_B + \varrho_B c_B u_B \\ -p_C + c_C^2 \varrho_C \end{pmatrix}.$$

v is the unknown state vector at the cell interface.

The very essence of this Riemann solver is to evaluate differences of the flow quantities only along the characteristics. The sign of low Mach number flows is exactly the fact that the ratio of acoustic wave amplitudes to entropy wave amplitudes becomes vastly different. Since the equations are decoupled form each other (according to the concept of a quasi linear system) each equation represents a single wave with its own amplitude and wave speed. In this sense, none of the equations is stiff. Most other Riemann solvers mix the computation of the wave amplitudes along the different characteristics. Guillard and Viozat [5] show how this leads to inconsistencies of the Roe solver in the low Mach number limit.

Now we want to show that this system can indeed be solved accurately as $M \to 0$. We introduce reference values into the equations. Let us consider a case with acoustic waves interacting with a non-homogeneous density field $\varrho = \varrho(x, y, z)$ with variations of $\mathcal{O}(1)$. As reference quantities we chose $\varrho^{ref} = \varrho(x, y, z)$, $p^{ref} = p_o$, and $u^{ref} = u_o$. Equation (29) then becomes

$$\begin{aligned}(\delta p_P - \delta p_A) + \varrho_A c_A (\delta u_P - \delta u_A) &= 0\\ (\delta p_P - \delta p_B) - \varrho_B c_B (\delta u_P - \delta u_B) &= 0\\ (\delta p_P - \delta p_C) - c_C^2 \left(\left(\varrho_C^{ref} + \delta \varrho_P \right) - \left(\varrho_C^{ref} + \delta \varrho_C \right) \right) &= 0.\end{aligned} \tag{31}$$

Please observe how ϱ^{ref} was chosen according to the upwind direction of the entropy wave. Again in matrix notation,

$$A \delta v + \tilde{b} = 0 \tag{32}$$

with

$$\delta v = \begin{pmatrix} \delta p_P \\ \delta u_P \\ \delta \varrho_P \end{pmatrix}; \qquad \tilde{b} = \begin{pmatrix} -\delta p_A - \varrho_A c_A \delta u_A \\ -\delta p_B + \varrho_B c_B \delta u_B \\ -\delta p_C + c_C^2 \delta \varrho_C \end{pmatrix}.$$

The matrix A is of course unaltered since the wave speeds have to be preserved.

Now we choose a nondimensionalisation with

$$\varrho_o,\ p_o,\ l_o,\ t_o, \text{ and } c_o.$$

Additionally we impose

$$\frac{p_o}{\varrho_o c_o^2} = 1 \qquad \text{and} \qquad \frac{c_o t_o}{l_o} = 1.$$

The condition number of the matrix A in the norm of Frobenius is

$$\lim_{M \to 0} \kappa(A) \doteq 3.5,$$

independent of M. The Frobenius norm $\| \cdot \|_F$ is defined as $\|A\|_F := \sqrt{\sum_{i,j} a_{ij}^2}$.

Thus Eq. (32) can be solved even for vanishing M. Note that this result depends on the reference values. If one would introduce u_o as the dimensional reference for the velocity, the condition would be $\mathcal{O}(M^2)$. This is what is sometimes called "the pressure singularity" in the low Mach number limit. When scaling with u_o one seeks to blow up small perturbations to $\mathcal{O}(1)$ and thereby scale the ambient pressure towards infinity. In the present approach, we allow small quantities to become small and exploit the fact that floating point numbers are actually scaled by the computer. This means that the present approach is limited by the exponent of the implemented model for real data. Commonly single precision arithmetic provides an exponent of ± 99 or bigger.

Still A and $\tilde{b}$ have entries of the form

$$\varrho c = \sqrt{\gamma \varrho p} \tag{33}$$

and

$$c^2 = \gamma p/\varrho, \tag{34}$$

which may lead to cancellation at $M \to 0$. In other words, although Eq. (32) can be solved accurately, we still need to ensure that the obtained solution is not spoiled by the poor accuracy of the matrix entries themselves. We cannot extract a reference state here, since the proper speed of sound has to be maintained. Therefore we have to ask what happens if the speed of sound is not calculated properly. If A and $\tilde{b}$ are perturbed by some quantity ΔA and $\Delta\tilde{b}$, how reliable will the solution of Eq. (32) be? How big is $\Delta(\delta v)$ in

$$(A + \Delta A)(v + \Delta(\delta v)) + (\tilde{b} + \Delta\tilde{b}) = 0? \tag{35}$$

The error $\Delta(\delta v)$ may be estimated using standard textbooks on numerical analysis. We assume the worst case with full loss of all significant decimals of the perturbation to Eqs. (33) and (34). Following [9, p. 33], we find

$$\frac{\|\Delta(\delta v)\|}{\|(\delta v)\|} \le \frac{\kappa(A)}{1 - \kappa(A)(\|\Delta A\|/\|A\|)} \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta\tilde{b}\|}{\|\tilde{b}\|} \right).$$

As a compatible vector norm we choose the L_2 norm. For $M \to 0$ we find

$$A = \begin{pmatrix} 1 & \sqrt{\gamma} & 0 \\ 1 & -\sqrt{\gamma} & 0 \\ 1 & 0 & -\gamma \end{pmatrix}; \qquad \Delta A = \begin{pmatrix} 0 & \mathcal{O}(M) & 0 \\ 0 & \mathcal{O}(M) & 0 \\ 0 & 0 & \mathcal{O}(M) \end{pmatrix}$$

$$\tilde{b} = \begin{pmatrix} \mathcal{O}(M) \\ \mathcal{O}(M) \\ \mathcal{O}(M) \end{pmatrix}; \qquad \Delta\tilde{b} = \begin{pmatrix} \mathcal{O}(M^2) \\ \mathcal{O}(M^2) \\ \mathcal{O}(M^2) \end{pmatrix}$$

$$\|\Delta A\|_F \sim \sqrt{\sum^{3} M^2} = \sqrt{3}M$$

$$\|A\|_F \sim \sqrt{\sum_{i,j=1}^{3} a_{ij}^2} \overset{M\to 0}{\doteq} 2.8$$

$$\|\Delta\tilde{b}\|_2 \sim \sqrt{3}M^2$$

$$\|\tilde{b}\|_2 \sim \sqrt{3}M.$$

Thus

$$\frac{\|\Delta(\delta v)\|}{\|(\delta v)\|} \le \frac{\kappa(A)}{1 - \kappa(A)\phi_1 M}(\phi_1 M + \phi_2 M) = \phi M.$$

(ϕ indicates a constant of order one.) The relative error will be $\frac{\|\Delta(\delta v)\|}{\|(\delta v)\|} \le \mathcal{O}(M)$. This means we are able to evaluate the fluxes, using Eq. (29), since the errors $\Delta(\delta v)$, introduced by the erroneous calculation of the speed of sound, are a factor $\mathcal{O}(M)$ smaller than the result (δv) we want to obtain. We omitted in this paper the demonstration that the primitive variables, needed in the above characteristic Riemann solver, can be determined from the conservative variables. This can be shown analogously [10, Sect. 17].

If, for example, $u > 0$ the first component of the numerical flux-difference $\Delta F^+ = F_b - F_c = F_b - F_P$ would now be calculated as

$$\begin{aligned}\varrho_b u_b - \varrho_P u_P &= \left(\varrho_b^{ref} + \delta\varrho_b\right)(u_o + \delta u_b) - \left(\varrho_C^{ref} + \delta\varrho_P\right)(u_o + \delta u_P) \\ &= \underline{u_o\left(\varrho_b^{ref} - \varrho_C^{ref}\right)} + \left(\varrho_b^{ref}\delta u_b - \varrho_C^{ref}\delta u_P\right) \\ &\quad + u_o(\delta\varrho_b - \delta\varrho_P) + (\delta\varrho_b\delta u_b - \delta\varrho_P\delta u_P). \qquad (36)\end{aligned}$$

We have again grouped terms of the same magnitude to avoid cancellation. All terms are at most $\mathcal{O}(M)$. This is not obvious for the first term (underlined) since we allow variations of ϱ^{ref} to be $\mathcal{O}(1)$. But $u_o(\varrho_b^{ref} - \varrho_C^{ref})$ is still at most $\mathcal{O}(M)$ since $\mathcal{O}(u_o) = M$. Normally one would choose $u_o = 0$. In this case the term vanishes.

Klein [7] showed that pressure variations of order $O(M)$ *and* $O(M^2)$ are equally important. The expression for the flux contains terms of both magnitudes. This leads obviously to a numerical limit. We can retain both, if present *at the same time*, unless the $\mathcal{O}(M^2)$ terms become comparable to the order of the smallest floating point number representable on the computer, or both are more than 10^{-t} apart from each other, with t being the mantissa length. This means for single precision, a limit of $M = 10^{-7}$ and double precision of $M = 10^{-14}$ if *both* effects are present.

Please remember that conventional schemes lose the $\mathcal{O}(M^2)$-contribution as $(1/\gamma M^2)\epsilon$ becomes comparable to unity. This means that a conventional scheme is always limited by a *fraction of the mantissa* of the floating point model in use. Our scheme is limited by the exponent for the case when only one effect is present. If both effects are important at the same time, it still can make *full use of the mantissa*.

4. CHOICE OF THE REFERENCE STATE U_o

The choice of the reference state U_o depends strongly on the type of problem in question. The simplest possibility is the computation of a perturbation to a constant state in space and time. Then, U_o consists just of a few constants.

If one is interested in small perturbations to a steady state problem, this steady state solution would be the ideal candidate. The simplest method is to calculate a steady state solution with the available program itself using any suitable state as reference, maybe even on a coarser grid. Then the solution of that run is taken as the new reference state.

If $\mathcal{O}(1)$ entropy changes are present in the flow field the reference field will be time dependent if we consider time intervals which are imposed by the fluid velocity. For example, consider a density jump of $\mathcal{O}(1)$ which is convected with a low Mach number flow. A suitable choice of the reference conditions would be the initial profiles for the density and energy density, and $(\varrho u)_o = 0$ for the momentum density. In an explicit scheme, the jump will move over one grid cell in approximately $1/M$ time steps. In the jump region the variable $\delta\varrho$ will become comparable to $\mathcal{O}(1)$ during this time interval, whereas in the rest of the flow field it will remain $\mathcal{O}(M)$. Thus a reasonable strategy would be to replace

$$\begin{aligned}U_o &\leftarrow (U_o + \delta U) \\ \delta U &\leftarrow 0\end{aligned}$$

wherever δU exceeds a certain threshold, a multiple of M, say, once every $1/M$ time steps.

5. NUMERICAL RESULTS

Physical test case. We now want to describe a physical test case which we will use in order to validate the improvements shown in Section 3, as well as to demonstrate the cancellation as shown in Section 2. The flow to be described was investigated by [1] and others with respect to the stability of the growing boundary layer. Thereby a solution to the simplified Navier–Stokes equations was found and a stability analysis performed on this basis. This solution may be found in [1; 10, Appendix B].

Consider a very long and thin tube filled with a little overpressure compared to the ambient. It is initially closed by a diaphragm. When the diaphragm is removed, an expansion wave runs into the tube expelling some gas. This is depicted in Fig. 4. In the inviscid case the pressure would drop to ambient pressure across the expansion wave. In the presence of friction, viscous effects retard the motion, leading to a quasi stationary outflow.

In the plots to be shown, the longitudinal coordinate is given in terms of $x = \hat{x}/(\hat{c}_o \hat{R}^2/\hat{\nu}_o)$ and $p = [\hat{p}/\hat{p}_o - 1]/(\widehat{\Delta p_o}/\hat{p}_o)$, $\hat{R}$ being the radius of the tube, $\hat{c}_o$ the speed of sound at rest, and $\hat{\nu}_o$ the kinematic viscosity. $\hat{p}_o$ is the ambient pressure and $\hat{p}_o + \widehat{\Delta p_o}$ the initial pressure in the tube.

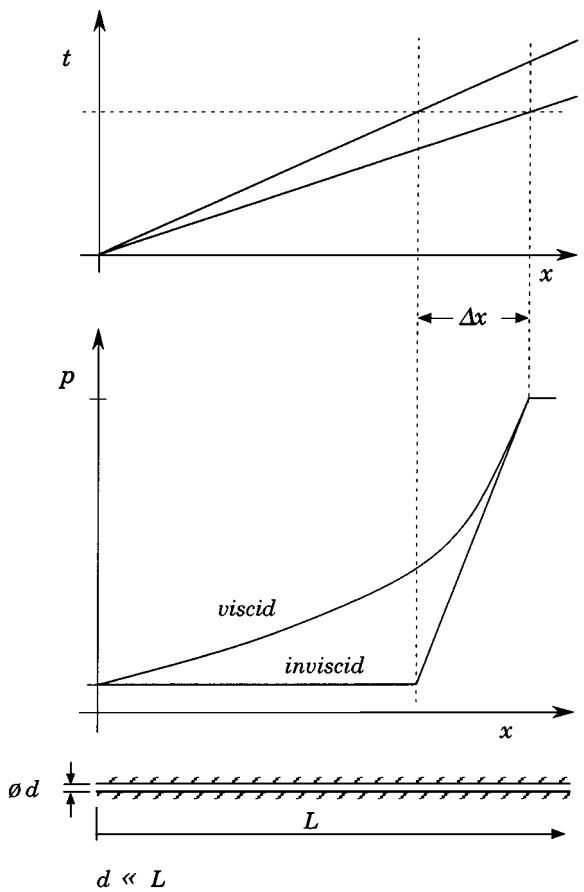

FIG. 4. Sketch of the flow situation. In the upper diagram an expansion fan in the (x-t) plane is shown. The lower diagram shows the corresponding pressure distribution at the time level indicated by a the dashed line in the (x-t) plane. At the bottom, the tube is depicted. It has a diameter d which is small compared to it's length L. The membrane would have been at the left end.

The flow is dominated by the time scale of the running expansion wave. Thus the calculation time is not increased when decreasing the Mach number. Please note that the depth Δx of the expansion wave is proportional to the Mach number. At very small Mach numbers the expansion fan cannot be resolved. It is rather a discontinuity which demands the conservative formulation of the governing PDEs in order to retain the physically correct signal speed.

A numerical experiment proving the concept. The concept presented in Section 2 can be investigated experimentally. The problem explained above was used as a test case. We assumed axial symmetry and thus the problem was reduced to two dimensions. For the waves propagating in radial directions the characteristic flux evaluation was slightly changed to include the effect amplitude changes for radial waves:

$$\text{along } \frac{dr}{dt} = v \pm c, \qquad dp \pm \varrho c\left(dv \pm \frac{cv}{v \pm c}\frac{dr}{r}\right) = 0 \tag{37}$$

$$\text{along } \frac{dr}{dt} = v, \qquad dp - c^2\, d\varrho = 0. \tag{38}$$

We start off with a flow field in primitive variables (ϱ, u, p), extracted from the results of these calculations for a certain radius outside the boundary layer.

We want to investigate the result of a single momentum balance as occurring in every time step. Therefore we look at the data as given at a certain time level and evaluate the predicted error growth, when performing the one single momentum balance. The momentum balance in the axial direction of the tube is

$$\delta f = f_R - f_L = (\varrho u^2 + p)_R - (\varrho u^2 + p)_L. \tag{39}$$

The cell interfaces in the longitudinal direction on a certain radius are of equal area and can be omitted in the present considerations. ϱu^2 is the momentum flux due to convection of mass and p contributes to the pressure force. The main error, as we know from Section 2, originates from first adding the small contribution ϱu^2 to the large value of p on either side and subtracting both sums in a further step. Therefore we neglect the error introduced in calculating ϱu^2 and concentrate on the summation only. To find the theoretical expression of the error corresponding to Eq. (7) we write the flux balance with a relative error term $(1 + \epsilon_i)$ for every summation:

$$\Delta f = \big[\big[(\varrho u^2)_R + p_R\big](1 + \epsilon_1) - \big[(\varrho u^2)_L + p_L\big](1 + \epsilon_2)\big](1 + \epsilon_3). \tag{40}$$

Δf is the numerical value, obtained in our attempt to calculate the exact value δf. Collecting the exact value δf from Eq. (39) yields to leading order

$$\Delta f = \delta f\left[1 + \frac{(\varrho u^2)_R + p_R}{\delta f}\epsilon_1 - \frac{(\varrho u^2)_L + p_L}{\delta f}\epsilon_2 + \epsilon_3\right]. \tag{41}$$

Since the ϵ_i are only known by their magnitudes, the three leading error terms (the terms containing an ϵ in (41)) are lumped together into $\alpha\epsilon_\alpha$ and we define

$$\Delta f_\alpha = \delta f(1 + \alpha\epsilon_\alpha). \tag{42}$$

The subscript in Δf_α indicates that Δf is evaluated with a certain leading error $\alpha\epsilon_\alpha$. The

leading term of the *theoretical expression* for the error is approximately given as

$$\alpha\epsilon_\alpha \approx \frac{\left[(\varrho u^2)_R + (\varrho u^2)_L\right] + [p_R + p_L]}{\left[(\varrho u^2)_R - (\varrho u^2)_L\right] + [p_R - p_L]}\epsilon. \tag{43}$$

In order to figure out the *actual error* in our numerical experiment, we additionally calculate the same balance with a different numerical precision and with a different arithmetic:

$$\Delta f_\beta = \delta f(1 + \beta\epsilon_\beta). \tag{44}$$

If we then calculate

$$\left(\frac{\Delta f_\alpha}{\Delta f_\beta} - 1\right) = \alpha\epsilon_\alpha\left(1 - \frac{\alpha\epsilon_\alpha}{\beta\epsilon_\beta} - \beta\epsilon_\beta + \cdots\right) \tag{45}$$

we are able to compute $\alpha\epsilon_1$ as

$$\alpha\epsilon_\alpha \approx \left(\frac{\Delta f_\alpha}{\Delta f_\beta} - 1\right), \tag{46}$$

provided $\beta\epsilon_2 \ll \alpha\epsilon_1$. To ensure the inequality, we evaluate Δf_α with single, Δf_β with double precision arithmetic in Fortran and with the terms rearranged by the method described in this paper.

Equations (43) and (46) are compared to each other in Fig. 5. The expansion wave front has reached $x = 0.08$. The fluid pushed out behind the wave front has a Mach number of about $M = 0.01$. The roundoff error ϵ in Eq. (43) was prescribed as $\epsilon = 5 \cdot 10^{-8}$. Note, this is the error introduced in the evaluation of the momentum flux balance every single time step. As shown in Section 3, there are several sources, augmenting the numerical error. On top of that, they are accumulated in numerous time steps. The failure of a finite-volume code, calculating with single precision the flow case mentioned above, is shown in Fig. 6. The error shows up dramatically in entropy, density, and temperature. One can observe it already at $M = 0.1$. The error in pressure is shown in Fig. 7. It does not look as dramatic at the present Mach number as the entropy, and might be overlooked in more complicated flow

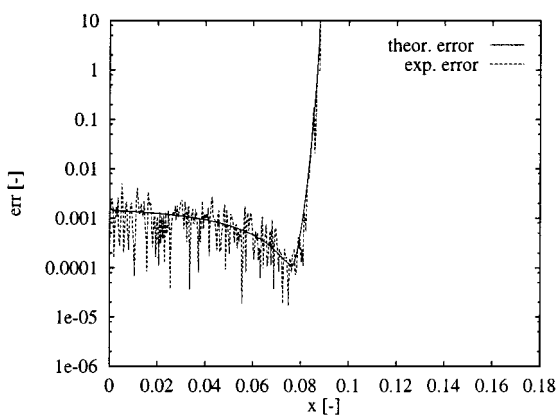

FIG. 5. Comparison of the error Eqs. (43) and (46), $M = 0.01$, 400×8 cells.

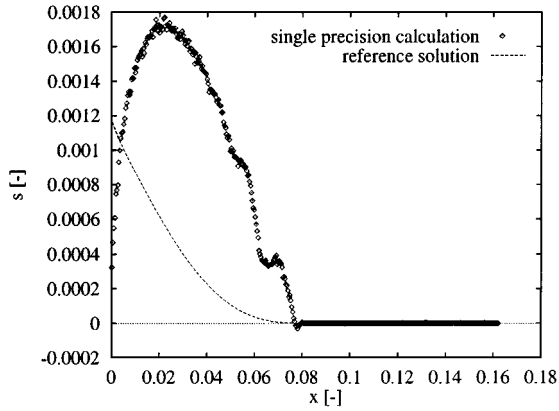

FIG. 6. Entropy along the axis of the tube. Comparison of two calculations, performed with single and double precision arithmetic, 400×8 cells, $M = 0.01$.

situations, where the aid of an analytical solution is not available. At lower Mach numbers the solution becomes useless.

Computation with very low Mach numbers. To demonstrate the desired capability of calculating flows at very low Mach numbers without harm from cancellation errors, we show in Fig. 8 a flow at $M = 10^{-11}$, using single precision arithmetic (seven digits). In this case we chose the reference state to be the ambient conditions ϱ_o, $u_o = 0$, and p_o. They were prescribed constant in space and time. To contrast this result, we show in Fig. 9 a flow-field, calculated with the same numerical precision, but with an unaltered code at $M = 10^{-6}$.

6. CONCLUSION AND REMARKS

We have shown that cancellation errors play a significant role in calculating low Mach number flows. In the past, several authors treated this problem, but no attempts were made to precisely show the role of the computer accuracy and numerical cancellation for this problem. This mechanism was demonstrated and identified in several important steps within a common finite volume code: time integration, flux balance, and flux evaluation.

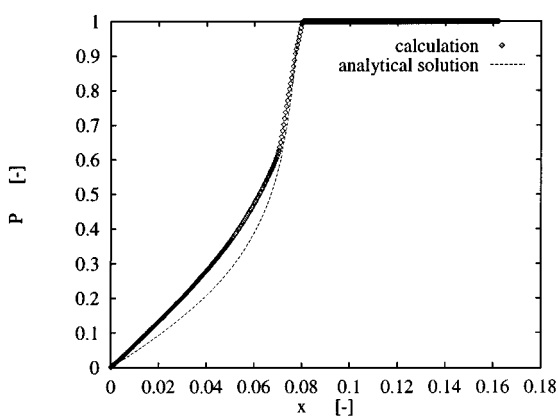

FIG. 7. Failure of a conventional scheme, 400×8 cells, $M = 0.01$.

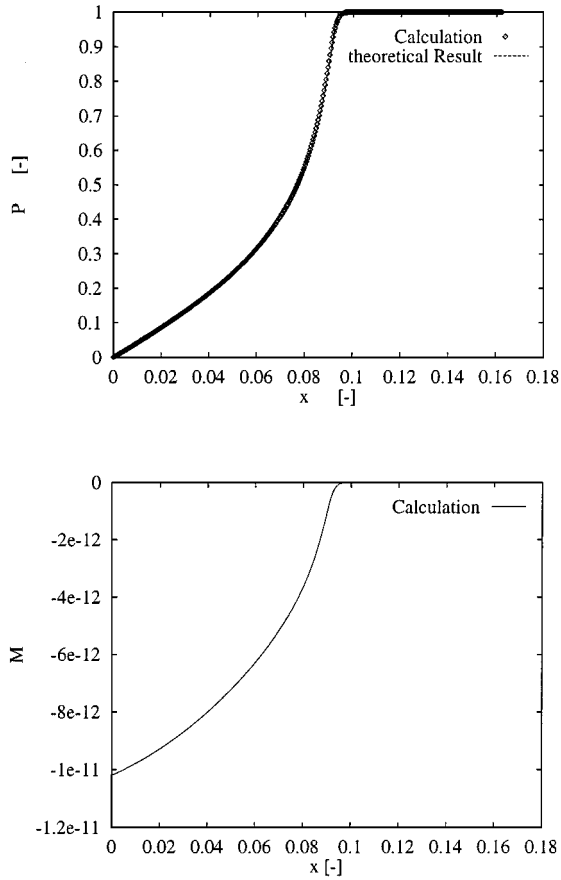

FIG. 8. Calculation at $M = 10^{-11}$, 400×8 cells.

Applying proper numerics, the cancellation problem in calculating low Mach number flows can be cured. This was shown in an example, calculating a flow at $M = 10^{-11}$, using a single precision Fortran on a common workstation. The method used is still conservative and capable of calculating high Mach number flows. *The basic idea is to introduce a*

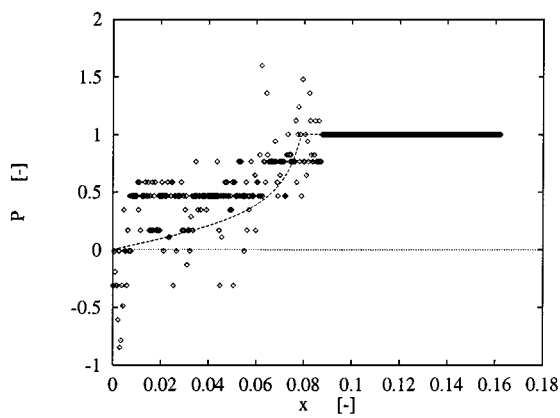

FIG. 9. Failure of an unaltered code at $M = 10^{-6}$, 400×8 cells.

reference state into the governing equation and effectively calculate perturbations to this state, without neglecting any terms of the full original equations. The effort for doing so in an existing program is moderate.

REFERENCES

1. F. Aebli and H. Thomann, Stability of the flow behind a pressure pulse in a tube, *Z. Angew. Math. Mech.* **39**, 387 (1988).
2. W. R. Briley, H. McDonald, and S. J. Shamroth, A low Mach number Euler formulation and applications to time-iterative LBI schemes, *AIAA J.* **21**, 1467 (1983).
3. K.-H. Chen and R. H. Pletcher, Primitive variable, strongly implicit calculation procedure for viscous flow at all speeds, in *AIAA Proceedings* (AIAA, Washington, DC, 1991), No. 91-1652.
4. R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves* (Springer-Verlag, New York/Berlin, 1948).
5. H. Guillard and C. Viozat, *On the Behaviour of Upwind Schemes in the Low Mach Number Limit*, Tech. Rep. No. 3160, INRIA, 1997.
6. M. Hafez and M. Soliman, *Numerical Solution of the Incomressible Navier–Stokes Equations in Primitive Variables on Unstagered Grids*, AIAA Paper 91-1561-CP, 1991.
7. R. Klein, Semi-implicit extension of a Godunow-type scheme based on low Mach number asymptotics. I. One-dimensional flow, *J. Comput. Fluids* **121**(2), 213 (1995).
8. R. J. LeVeque, *Numerical Methods for Conservation Laws* (Birkhäuser Verlag, Basel, 1990), Vol. I.
9. H. R. Schwarz, *Numerische Mathematik*, 2nd ed. (Teubner, Stuttgart, 1988).
10. J. Sesterhenn, *Zur numerischen Berechnung kompressibler Strömungen bei kleinen Mach-Zahlen*, Ph.D. thesis, Diss. ETH No. 11334, 1995.
11. J.-S. Shuen, K.-H. Chen, and Y. Choi, A coupled implicit method for non-equilibrium flows at all speeds, *J. Comput. Phys.* **106**, 306 (1993).
12. J. Stoer and R. Burlisch, *An Introduction into Numerical Analysis*, 2nd ed., Texts in Applied Mathematics (Springer-Verlag, New York, 1993), Vol. 12.
13. E. Turkel, Fiterman, and B. van Leer, *Preconditioning and the Limit to the Incompressible Flow Equations*, Tech. Rep. No. 93-42, ICASE, 1993.
14. B. van Leer, W. T. Lee, and P. L. Roe, Characteristic time-stepping or local preconditioning of the Euler equations, in *10TH Computational Fluid Dynamics Conference* (AIAA, Washington, DC, 1991).
15. C. Viozat, *Implicit Upwind Schemes for Low Mach Number Compressible Flows*, Tech. Rep. No. 3084, INRIA, 1997.
16. G. Volpe, Performance of compressible flow codes at low mach numbers, *AIAA J.* **31**(1), 49 (1993).