JOURNAL OF COMPUTATIONAL PHYSICS 111, 1–14 (1994)

Numerical Solution of the Coupled Mode Equations in Duct Acoustics

LEIF ABRAHAMSSON AND HEINZ-OTTO KREISS

National Defence Research Establishment, S-172 90 Sundbyberg, Sweden

Received October 26, 1992; revised August 17, 1993

We present a fast solver for the coupled mode equations in duct acoustics. It is based on a partitioning of the resulting system of ordinary differential equations into separate subsystems, the number of which increases by the separability of the problem. This structure is obtained by transformations of the dependent variables such that weakly interacting modes are separated. The decoupling process requires the numerical solution of algebraic Riccati equations. However, these computations can be done on a spatial scale much larger than the characteristic wavelength of the problem. © 1994 Academic Press, Inc.

1. INTRODUCTION

We consider the numerical solution of the reduced wave equation for the acoustic pressure in a two-dimensional duct with variable geometry and a heterogeneous fluid. The side walls are assumed to be perfectly reflecting so that the wave propagation is solely along the duct axis. In the coupled mode approach the solution is expressed as a sum of modes

$$p(\xi, \zeta) = \sum_{m=1}^{M} \phi_m(\xi)\, \psi_m(\xi, \zeta), \tag{1.1}$$

where (ξ, ζ) are some suitable axial and transversal coordinates of the duct. Here ψ_m is the mth normalized eigenfunction of a regular Sturm–Liouville problem over the cross section $\xi = \text{constant}$. Substituting (1.1) into the Helmholtz equation leads to a system of M ordinary differential equations of second order for ϕ_m and is called the coupled mode equations. It is a boundary value problem because flow conditions are usually imposed at both ends of the duct. If the problem is separable this is the classical method of separation of variables. Then the eigenvalue problem for ψ_m needs to be solved only once, and each ϕ_m is determined independently of the others. For nonseparable problems we obtain a large system for ϕ_m which could be prohibitively expensive to solve by a conventional finite difference method. The reason is that the solution must be well resolved for each spatial period, the number of which may be in the thousands. The purpose of this paper is to derive a fast solution of the coupled mode equations. It is obtained by an iteration scheme in which the iteration operator is partitioned into separate subsystems, the number of which increases by the separability of the problem. If the mode couplings are weak enough a diagonal iteration operator will do. Otherwise it must be enlarged as to include those mode couplings that are too strong to be iterated on. To begin with we shall try to make the coupled system more diagonally dominant by introducing a transformed system in which some of the mode couplings have been reduced. This is done by transformations of Riccati type [17, 18]. They will be efficient if the eigenvalue separation is large compared to the size of the mode coupling. This approach is related to the method of invariant embedding [6] and has received a lot of attention recently for nonoscillatory systems [19]. The same tool is used to split, whenever possible, each subsystem of the iteration operator into two halves of systems of one-way equations. This transformation of the coupled mode equations is worthwhile only if it can be found by computations on a spatial scale much larger than the characteristic wavelength of the problem. The actual scale, represented by a grid $\{\xi_i\}$, $i = 0, 1, \ldots, N$, must be chosen such that the variation of ψ_m is smooth on this grid. The size of N is a measure of the separability of the problem. If N is larger than the number of wavelengths over the length of the duct, then the use of coupled modes becomes questionable.

Let k_0 and h denote a characteristic wavenumber and some average height of the duct. The number of propagating modes is of the order hk_0 and $M \approx hk_0$ would be a natural choice in the expansion (1.1). The cost of computing the coefficients of the coupled mode equations is proportional to M^3N operations. The eigenfunction sensitivity, which determines N, tends to increase by hk_0. Altogether it means that the ansatz (1.1) might not be computationally tractable if hk_0 is too large. However, sometimes it is possible to choose or remodel the source such that it contains only M_0 modes, $M_0 \ll hk_0$. Then one could try to take M as M_0 plus a few extra modes which are set to zero initially. If not all of

0021-9991/94 $6.00

them turn out to be excited one could be confident that this M would suffice.

For small hk_0 the case with only one propagating mode is of particular importance in aeroacoustics for silencer and ventilation systems [20]. Although this case is included here it is of less interest because it is usually not very demanding computationally.

We shall assume that the coefficients of the coupled mode equations are twice continuously differentiable. This assumption enables the decoupling process, which could be improved further by using still more smoothness. However, in practice the geometry and the medium are often specified at discrete points only and also with some numerical uncertainty. Then it would be dubious to utilize higher derivatives obtained by interpolation.

An alternative approach is to replace the duct with a number of uniform subsections. A constant coefficient problem is solved for each subsection and joined at the boundaries by continuity conditions [12]. The stepwise change of data will excite a large number of modes, even evanescent ones. This is contrary to our endeavor to separate modes and it necessitates the smoothness assumption.

A finite difference (or finite element) method applied directly to the Helmholtz equation leads to a huge linear system of equations in which the number of unknowns is of the order $lhk_0^2P^2$, where l is the length of the duct. P is the number of discretization points per half wavelength and as a thumb rule at least $P \approx 10$ is necessary to achieve two digits of accuracy by a fourth-order method. Therefore reported computations are for rather small domains in terms of lhk_0^2 [3, 14]. The practical applicability of coupled modes is only limited by the size of hk_0 and the degree of separability.

This study was initiated by applications in underwater acoustics. To some extent this is reflected in our choice of notations and computational examples.

2. THE COUPLED MODE EQUATIONS

Two-dimensional harmonic sound propagation is governed by the reduced wave equation

$$\rho \nabla \cdot (\rho^{-1} \nabla p) + k^2 p = 0, \tag{2.1}$$

where

$\mathrm{Re}(p(\mathbf{r})\, e^{-i\omega t})$,	acoustic pressure
$\mathbf{r} = (x, z)$,	cartesian coordinates
$\rho(\mathbf{r})$,	fluid density
$k = \omega(1 + i\beta(\mathbf{r}))/c(\mathbf{r})$,	wave number
ω,	circular frequency
$c(\mathbf{r})$,	sound speed
$\beta(\mathbf{r})$,	sound absorption

Physical quantities are assumed to be expressed in any units suitable for the application under study. We consider (2.1) in a waveguide bounded by two nonintersecting curves,

$$\gamma_i = (x_i(s), z_i(s)), \qquad i = 1, 2, \qquad s_0 \leqslant s \leqslant s_1,$$

see Fig. 1.

The boundary conditions on γ_i are assumed to be of the form

$$a_i p_n + b_i p = 0, \tag{2.2}$$

$i = 1, 2$, $a_i, b_i \in \mathcal{R}$, where p_n denotes the outward normal derivative. We assume that the duct is terminated anechoically at both ends. It implies that the duct extends to infinity in both directions such that γ_1 is parallel to γ_2 and ρ, c, β, a_i, b_i are constants with $\beta = 0$ in the infinite sections. The ends $s = s_0$ and $s = s_1$ are assumed to be part of the uniform sections. Then p can be written as a sum of normal modes around and beyond the ends of the duct. This modal sum is divided into two groups which represent incoming and outgoing waves. As the source condition for (2.1) we shall specify an incoming wave from the left, $s < s_0$, while only outgoing waves are allowed for $s > s_1$. These end conditions are natural when the solution itself is sought in terms of modes. A nonuniform termination of the duct with boundary conditions like (2.2) would cause intermode couplings. It could be handled by coupled modes as well but requires more work.

The above problem is well posed only in the absence of resonance. Often one cannot state in advance whether resonance will occur. In any case, if ω is a resonant frequency it will show up in a controlled way during the solution process.

The numerical solution is facilitated by the use of a boundary fitted orthogonal coordinate system. Then we obtain a rectangular computational domain and a normal derivative at the boundary goes over into differentiation along a coordinate axis. We shall assume that such a transformation can be realized numerically without too heavy calculations. This imposes a loosely defined constraint on the geometry of the duct. For example, if the radius of curvature of γ_i is not too small compared to the height of the duct, one could use the program [1, 2]. Ducts with side cavities, narrow throats, branches, etc. must be treated differently, for example, by domain decomposition methods. Let

$$\begin{aligned} &x = x(\xi, \zeta), \qquad z = z(\xi, \zeta), \qquad 0 \leqslant \xi \leqslant l, \qquad 0 \leqslant \zeta \leqslant h, \\ &a^2 = x_\xi^2 + z_\xi^2, \qquad b^2 = x_\zeta^2 + z_\zeta^2, \end{aligned} \tag{2.3}$$

denote an orthogonal coordinate transformation. For convenience we shall retain the same unit for (ξ, ζ) as for

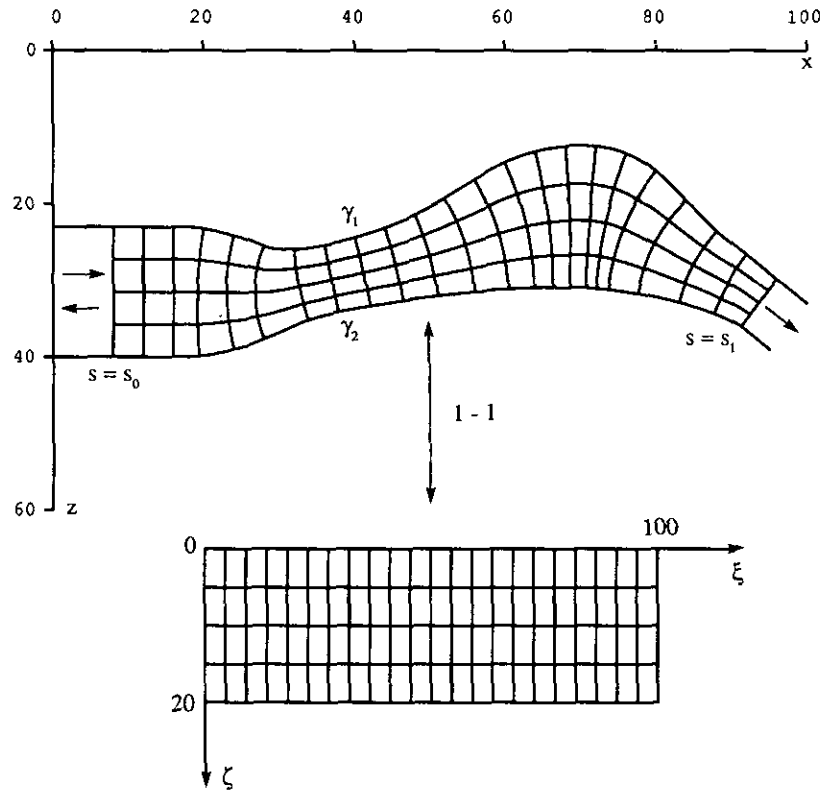

FIG. 1. Geometry of duct with an orthogonal mapping onto the computational domain in the (ξ, ζ)-plane.

(x, z) so that $l \approx$ length of duct, $h \approx$ mean height. In the new coordinates Eq. (2.1) after multiplication by a^2 takes the form

$$p_{\xi\xi} - d^{-1} d_\xi p_\xi + d(a(\rho b)^{-1} p_\zeta)_\zeta + a^2 k^2 p = 0, \quad (2.4)$$

where $d = \rho a b^{-1}$. For every ξ, $0 \leqslant \xi \leqslant l$, we introduce a transversal eigenvalue problem with ξ as a parameter:

$$\begin{aligned} -d(a(\rho b)^{-1}\psi_\zeta)_\zeta - a^2 \operatorname{Re}(k^2)\psi &= \lambda(\xi)\psi, & 0 \leqslant \zeta \leqslant h, \\ a_1 b^{-1}\psi_\zeta + b_1\psi &= 0, & \zeta = 0, \\ a_2 b^{-1}\psi_\zeta + b_2\psi &= 0, & \zeta = h. \end{aligned} \quad (2.5)$$

This is a regular Sturm–Liouville problem with simple eigenvalues $\lambda_1 < \lambda_2 < \cdots$, and corresponding eigenfunctions $\psi_m(\xi, \zeta)$ which are orthonormal with respect to the scalar product

$$(u, v) = \int_0^h \bar{u} v \, d^{-1} \, d\zeta.$$

The eigenvalue separation implies that ψ_m is differentiable with respect to ξ to the same degree as the coefficients of (2.5). Substituting the ansatz (1.1) into (2.4) and taking the scalar product with ψ_n, $n = 1, 2, \ldots, M$, leads to the system

$$\phi_m'' = \tilde{\lambda}_m(\xi)\phi_m + \sum_{\substack{n=1 \\ n \neq m}}^{M} \{\alpha_{mn}\phi_n' + \beta_{mn}\phi_n\}, \quad 0 \leqslant \xi \leqslant l, \quad (2.6)$$

$$\begin{aligned} &m = 1, 2, \ldots, M, \\ &\alpha_{mn} = -\alpha_{nm} = (\psi_{m\xi}, \psi_n) - (\psi_m, \psi_{n\xi}), \\ &\beta_{mn} = -(\psi_m, d(d^{-1}\psi_{n\xi})_\xi + ia^2 \operatorname{Im}(k^2)\psi_n), \\ &\tilde{\lambda}_m = \lambda_m + \beta_{mm}. \end{aligned} \quad (2.7)$$

For $\xi \leqslant 0$, α_{mn} and β_{mn} vanish and $\lambda_m(\xi) = \lambda_{0m}$, a constant. Then for $\xi \leqslant 0$ the normal mode solution of (2.4) can be written as

$$p = \sum_{m=1}^{M} \{u_{0m}^{+} \exp(k_{0m}\xi) + u_{0m}^{-} \exp(-k_{0m}\xi)\}\psi_m(0, \zeta) \quad (2.8)$$

and correspondingly for $\xi \geqslant l$ with the notational change $0 \to 1$. Here the longitudinal wavenumbers k_m are defined by

$$k_m = \lambda_m^{1/2}, \qquad \tfrac{1}{4}\pi \leqslant \arg k_m < \tfrac{1}{4}\pi + \pi.$$

In what follows we shall adopt this convention for the principal branch of the square root. It implies that $(+)$-terms of the sum (2.8) are exponentially decaying by ξ for $\lambda_m > 0$ (evanescent waves) and also for $\lambda_m < 0$ (propagating waves) if positive wave attenuation is added. The exceptional case $\lambda = 0$ is avoided by introducing a tiny perturbation of ω when necessary. The source condition of an incident wave from the left implies that the amplitudes u_{0m}^+ are given and the nonreflective condition at $\xi = l$ is equivalent to $u_{1m}^- = 0$, $m = 1, 2, ..., M$. In terms of ϕ_m these conditions translate into

$$\begin{aligned} k_{0m}^{-1}\phi_m' + \phi_m &= 2u_{0m}^+, \qquad \xi = 0, \\ k_{1m}^{-1}\phi_m' - \phi_m &= 0, \qquad \xi = l, \end{aligned} \tag{2.9}$$

$m = 1, 2, ..., M$. The determination of the reflected and transmitted wave amplitudes u_{0m}^- and u_{1m}^+ is part of the solution.

The boundary conditions (2.2) imply that no energy escapes through the walls and the energy flux along the duct is, apart from absorption, conserved. The same is true for the coupled mode solution (1.1). By integration of the imaginary part of the scalar product of p and (2.4) over $[\xi_1, \xi_2]$ we obtain

$$\operatorname{Im}(p, p_\xi)|_{\xi=\xi_2} = \operatorname{Im}(p, p_\xi)|_{\xi=\xi_1} - \operatorname{Im}\int_{\xi_1}^{\xi_2} (p, a^2k^2p)\, d\xi. \tag{2.10}$$

The left-hand side, multiplied by $\frac{1}{2}\omega^{-1}$, is the energy flux through the cross section $\xi = \xi_2$. In particular, by inserting (2.8) into (2.10) for $\xi_1 \leqslant 0$ and $\xi_2 \geqslant l$ we obtain

$$\operatorname{Im}\sum_{m=1}^{M} |u_{0m}^+|^2 k_{0m} = \operatorname{Im}\sum_{m=1}^{M} \{|u_{0m}^-|^2 k_{0m} + |u_{1m}^+|^2 k_{1m}\} \tag{2.11}$$

in the case of no dissipation. Physically it means that the incident energy flux is equal to the sum of the backscattered and transmitted ones. For example, if the end at $\xi = l$ is so narrow that all $\lambda_{1m} > 0$, then all energy is backreflected. Verification of (2.11) provides a first crosscheck upon computed results. However, it does not necessarily mean that the solution is accurate. The identity (2.11) is a direct consequence of the way Eqs. (2.6) were derived and it is valid for any M, even $M = 1$. The energy relation (2.11) is just one instance of a number of useful reciprocity principles in dust accoustics [11, 13, 21].

The coefficients (2.7) must be computed numerically on a grid

$$0 = \xi_0 < \xi_1 < \cdots < \xi_N = l, \tag{2.12}$$

such that they are well resolved. The relative variation of λ_m is expected to follow that of the coefficients of (2.4). The eigenfunction sensitivity is usually more critical. It might depend on the relative variation of the difference of the eigenvalues which increases both by frequency and the width of the duct. In practice we have determined $\{\xi_i\}$ by requiring

$$\begin{aligned} \max_{1 \leqslant m \leqslant M} (\Delta\psi_m, \Delta\psi_m) &\leqslant 0.05, \\ \Delta\psi &= \psi(\xi_{i+1}) - \psi(\xi_i), \end{aligned} \tag{2.13}$$

with the additional constraint $\Delta\xi_i \leqslant \delta$. Here δ should be a suitable scale for the variation of the coefficients of (2.4).

The eigenvalue problem (2.5) is solved numerically by the standard three-point difference formula and Richardson extrapolation is applied once to obtain accuracy of the fourth order [16]. By the Sturm oscillation theorem ψ_M has $M - 1$ simple zeros in $(0, h)$ which are almost uniformly distributed for large M. Let P denote the number of gridpoints between two consecutive zeros of ψ_M. Some $P \approx 10$ is necessary to ensure three digits of accuracy. Then the dimension of the resulting tridiagonal eigensystem is of the order PM. The eigenpairs are computed one by one using Rayleigh quotient iterations according to [15, 22]. By experience the eigensolver spends less than $100PM^2$ operations for M eigenpairs. The cost to form the scalar products (2.7) is $1.5PM^3$ operations. Let us assume that (2.6), (2.9) is solved numerically by a compact Numerov method. This would lead to a linear system of equations with $\approx PMlk_0$ unknowns and bandwidth $4M$. Its solution by a direct solver would require $16PM^3lk_0$ operations. Alternatively, one could apply a Numerov method directly to Eq. (2.1) and use a bandsolver. Under the same accuracy constraints this comparison gives that the coupled mode technique is preferable in terms of operations if

$$100PM^2N + 1.5PM^3N + 16PM^3lk_0 < 4P^3M^3lk_0.$$

Clearly, coupled modes are more efficient than a finite difference solution of (2.1) if N is less than the number of wavelengths over the length of the duct. Thus the decisive criterion is to see if the condition (2.13) can be met on a grid whose density on average exceeds the wavelength. This question must be probed by some extra eigenvalue computations in which we try to select $\Delta\xi_i$ as large as possible without impairing the conditions (2.13).

The scale (2.12) plays an important role in subsequent calculations to simplify the system (2.6). Both cost and storage requirements will be proportional to N.

For the data representation of the coefficients (2.7) we use a complete cubic spline interpolant with the breakpoints (2.12) that is twice continuously differentiable in $[0, l]$ [8]. Derivatives, whenever needed, are obtained by differentia-

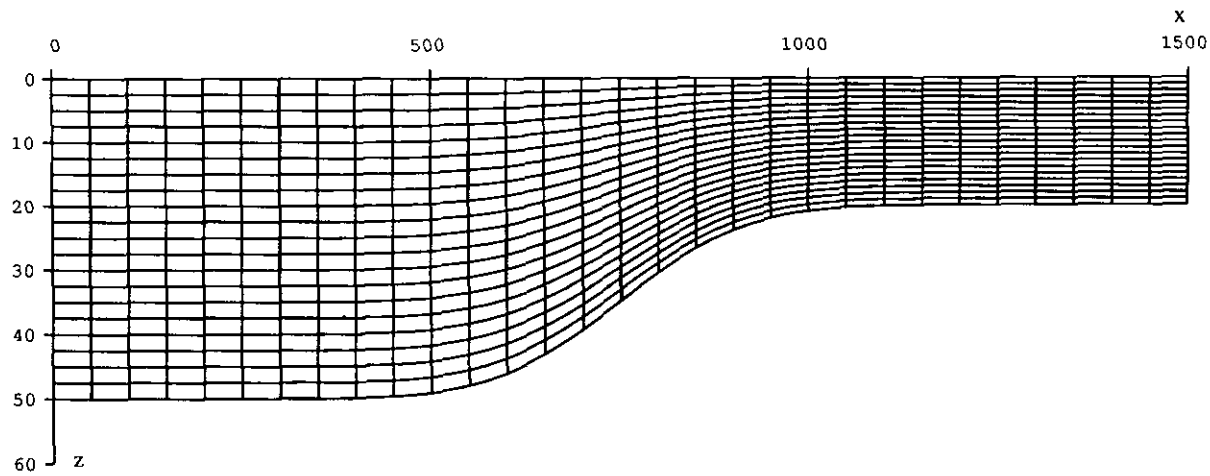

FIG. 2. The geometry of the duct in example (2.16). Some coordinate lines of the transformation (2.3) are also shown. The axes are drawn with different scale units which make the grid look skew.

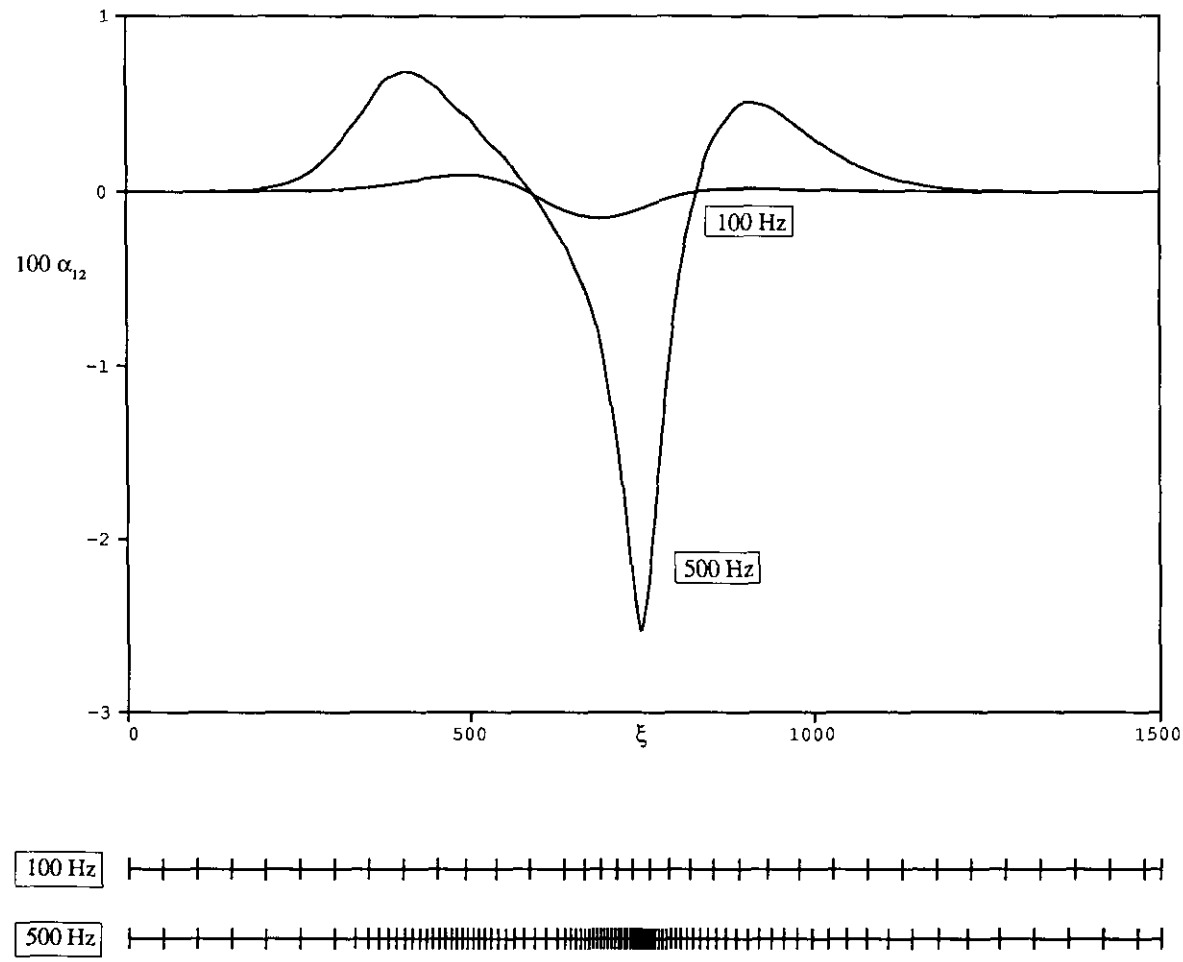

FIG. 3. The graph of the coupling coefficient α_{12} for example (2.16) for $f = 100$ Hz and $f = 500$ Hz. The distribution of the grid points (2.12) is also shown.

tion of the cubic spline. The interpolation error of the jth derivative, $j=0, 1, 2$, is of the order $(\Delta\xi_i)^{4-j}$. The coefficients of the cubic spline are computed only once and stored so that evaluation at any point can be done rapidly.

To be able to employ the theory of [17, 18] we shall rewrite (2.6) as a system of first-order equations. In the standard way we put

$$u_m = \phi_m, \qquad v_m = \phi'_m, \qquad m = 1, 2, ..., M,$$

for which (2.6) and (2.10) go over into

$$\begin{bmatrix} u_m \\ v_m \end{bmatrix}' = \sum_{n=1}^{M} A_{mn} \begin{bmatrix} u_n \\ v_n \end{bmatrix}, \qquad 0 \leqslant \xi \leqslant l, \tag{2.14}$$

$$\begin{aligned} k_{0m}^{-1} v_m + u_m &= 2u_{0m}^{+}, && \xi = 0, \\ k_{1m}^{-1} v_m - u_m &= 0, && \xi = l, \end{aligned} \tag{2.15}$$

for $m = 1, 2, ..., M$. Here the 2×2 matrices A_{mn} are given by

$$A_{mm} = \begin{bmatrix} 0 & 1 \\ \tilde{\lambda}_m & 0 \end{bmatrix}, \qquad A_{mn} = \begin{bmatrix} 0 & 0 \\ \beta_{mn} & \alpha_{mn} \end{bmatrix}, \qquad n \neq m.$$

In what follows we shall use the norms

$$\|a\| = \max_{0 \leqslant \xi \leqslant l} |a(\xi)|, \qquad \|a\|_1 = \int_0^l |a(\xi)|\, d\xi$$

for complex-valued functions $a(\xi)$.

We shall illustrate our results for a duct with a homogeneous fluid (water) and a range-dependent geometry (Fig. 2):

$$\begin{aligned} &\gamma_1 = (x, 0), && 0 \leqslant x \leqslant 1500 \text{ m}, \\ &\gamma_2 = (x, h(x)), && \text{-"-} \quad , \\ &h(x) = 50 - d_0 \int_0^x \exp(-(4(s-750)/750)^2)\, ds, \\ &d_0 = 4/(25\pi^{1/2}), \\ &\rho = 10^3 \text{ kg/m}^3 \\ &k = 2\pi f/1500 \text{ m}^{-1}, && f = 100 \text{ or } 500 \text{ Hz} \\ &p = 0 \quad \text{on } \gamma_1, && p_n = 0 \quad \text{on } \gamma_2, \\ &M = 10, && u_{0m}^{+} = 1, \quad m = 1, 2, ..., M. \end{aligned} \tag{2.16}$$

Often one is interested in solving stationary problems for a number of frequencies f. The effect of inhomogeneities of the medium and the geometry is frequency dependent and the numerical resolution must be adapted accordingly. This point is exemplified by Fig. 3 which shows the coupling coefficient α_{12} (multiplied by 100) for $f = 100$ Hz and $f = 500$ Hz. The number of gridpoints (2.12) as determined by (2.13) is $N = 35$ and $N = 93$.

3. REDUCTION OF MODE COUPLINGS

For weak mode couplings the most natural way of solving the system (2.14) is by iterations

$$\begin{bmatrix} u_m^j \\ v_m^j \end{bmatrix}' - A_{mm} \begin{bmatrix} u_m^j \\ v_m^j \end{bmatrix} = \sum_{n \neq m} A_{mn} \begin{bmatrix} u_n^{j-1} \\ v_n^{j-1} \end{bmatrix}, \qquad 0 \leqslant \xi \leqslant l, \tag{3.1}$$

$m = 1, 2, ..., M$, $j \geqslant 1$, and with $u_m^0 = v_m^0 \equiv 0$. The boundary conditions for u_m^j, v_m^j are the same as for u_m, v_m, that is, (2.15). Each iteration with (3.1) requires the solution of M separate 2×2 systems. With a finite difference method the cost would be $\approx 30ML$ operations where $\Delta\xi = l/L$ is a suitable stepsize for the integration. In comparison a direct solver for a finite difference solution of the system (2.14), (2.15) requires $\approx 72M^3L$ operations. Therefore it is usually profitable to use (3.1), provided it converges.

It is tempting to use only one iteration in (3.1) because it frees us from the computation of the coefficients A_{mn}, $m \neq n$. This is the adiabatic approximation which is widely used in underwater acoustics [10, 23]. However, it is impossible to predict when it is accurate. The mode couplings depend in a nonlinear way on all the parameters involved and, even though they were known, it is difficult to judge their effect. For example, the error $e = \max_m \|u_m - u_m^1\|$ of the adiabatic solution of example (2.16) is $e = 0.15$ and $e = 1.6$ for $f = 100$ and $f = 500$ Hz, respectively. Therefore we intend to solve the system (2.14), (2.15) in full.

In order to speed up the rate of convergence of (3.1) we shall make the system (2.14) more block diagonally dominant by introducing new variables $\hat{u}_m, \hat{v}_m$ by transformations of Riccati type. Despite this effort a block diagonal iteration operator might not suffice to obtain convergence. Sometimes it must be enlarged by the inclusion of strong intermode couplings. Then the iteration scheme takes the form

$$\begin{bmatrix} \hat{u}_m^j \\ \hat{v}_m^j \end{bmatrix}' - \sum_{n \in m(n)} A_{mn} \begin{bmatrix} \hat{u}_n^j \\ \hat{v}_n^j \end{bmatrix} = \sum_{n \notin m(n)} A_{mn} \begin{bmatrix} \hat{u}_n^{j-1} \\ \hat{v}_n^{j-1} \end{bmatrix}, \qquad 0 \leqslant \xi \leqslant l, \tag{3.2}$$

where $n(m) \subset \{1, 2, ..., M\}$ is a set of indices such that the left-hand side of (3.2) constitutes a number of separate subsystems. Next we describe how to arrive at a suitable scheme (3.2) in an automatic way.

To begin, we single out a 4×4 subsystem from (2.14) which comprises only the direct coupling between the mth and nth modes, $m < n$. In 2×2 block form it can be written

$$\begin{bmatrix} \mathbf{u}_m \\ \mathbf{u}_n \end{bmatrix}' = \begin{bmatrix} A_{mm} & A_{mn} \\ A_{nm} & A_{nn} \end{bmatrix} \begin{bmatrix} \mathbf{u}_m \\ \mathbf{u}_n \end{bmatrix}, \qquad 0 \leqslant \xi \leqslant l, \tag{3.3}$$

where $\mathbf{u}_m = [u_m, v_m]^T$. This system can be considered as two coupled oscillators designated by m and n. We want to find an upper bound of the influence of the oscillator n on m represented by the matrix A_{mn}. Let $\mathbf{u}_m^0, \mathbf{u}_n^0$ denote the solution of (3.3) with $A_{mn} = A_{nm} = 0$ and the boundary conditions (2.15). A first approximation to $\mathbf{u}_m$ is

$$\mathbf{u}_m(\xi) \approx \mathbf{u}_m^0(\xi) + \int_0^l G^m(\xi, t)\, A_{mn} \mathbf{u}_n^0(t)\, dt,$$

where $G^m = [G_{i,j}^m]$, $i, j = 1, 2$, is the Green's function of oscillator m. A dimensionless measure of the strength of the perturbation A_{mn} is then given by

$$\begin{aligned}\mathscr{L}(A_{mn}) = \|G_{11}^m\| \{\|a_{11}\|_1 + \|\lambda_n^{1/2}\| \cdot \|a_{12}\|_1\} \\ + \|G_{12}^m\| \{\|a_{21}\|_1 + \|\lambda_n^{1/2}\| \cdot \|a_{22}\|_1\},\end{aligned} \quad (3.4)$$

where a_{ij} are the elements of A_{mn}. The computation of the Green's function will be described later. The functional (3.4) can be evaluated and if found too large, say $> M^{-1}$, we try to suppress A_{mn} and A_{nm} by a transformation,

$$\begin{aligned}&\begin{bmatrix}\mathbf{u}_m \\ \mathbf{u}_n\end{bmatrix} = \hat{T}_{mn}\begin{bmatrix}\hat{\mathbf{u}}_m \\ \hat{\mathbf{u}}_n\end{bmatrix}, \qquad 0 \leqslant \xi \leqslant l, \\ &\hat{T}_{mn} = \begin{bmatrix}I + RS & R \\ S & I\end{bmatrix}, \\ &\hat{T}_{mn}^{-1} = \begin{bmatrix}I & -R \\ S & I + SR\end{bmatrix}.\end{aligned} \quad (3.5)$$

Here R and S are 2×2 matrices satisfying the equations

$$\begin{aligned}A_{mm}R - RA_{nn} - RA_{nm}R + A_{mn} = 0, \\ (A_{nn} + A_{nm}R)\, S - S(A_{mm} - RA_{nm}) + A_{nm} = 0\end{aligned} \quad (3.6)$$

for $0 \leqslant \xi \leqslant l$. Substituting (3.5) into (3.3) using (3.6) gives

$$\begin{bmatrix}\hat{\mathbf{u}}_m \\ \hat{\mathbf{u}}_n\end{bmatrix}' = \begin{bmatrix}\hat{A}_{mm} & \hat{A}_{mn} \\ \hat{A}_{nm} & \hat{A}_{nn}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{u}}_m \\ \hat{\mathbf{u}}_n\end{bmatrix}, \qquad 0 \leqslant \xi \leqslant l, \quad (3.7)$$

where

$$\begin{aligned}&\hat{A}_{mm} = A_{mm} - RA_{nm} - R'S, \qquad \hat{A}_{mn} = -R' \\ &\hat{A}_{nm} = -S' + SR'S, \qquad \hat{A}_{nn} = A_{nn} + A_{nm}R + SR'.\end{aligned} \quad (3.8)$$

Equations (3.6) are recognized as the Riccati equations for (3.3), except for the omission of $-R'$ and $-S' + SR'S$ on the left-hand side. Any solution of the Riccati equations would decouple (3.3) entirely by the transformation (3.5). However, the differential Riccati equations are numerically intractable in contrast to the algebraic equations (3.6). The solution of interest of (3.6) is essentially determined by its linear part which is nonsingular due to the condition $\lambda_m \neq \lambda_n$ [18]. Clearly (3.7) is an improvement over (3.3) if

$$\mathscr{L}(\hat{A}_{mn}) < \mathscr{L}(A_{mn}) \qquad \text{and} \qquad \mathscr{L}(\hat{A}_{nm}) < \mathscr{L}(A_{nm}), \quad (3.9)$$

where $\mathscr{L}$ is given by (3.4). Let us for a moment see when the transformation (3.5) is usuable. By ignoring the nonlinear terms of (3.6) we find that

$$\begin{aligned}&R \approx -(\lambda_m - \lambda_n)^{-1}\begin{bmatrix}\alpha_{mn} & \beta_{mn} \\ \lambda_n\beta_{mn} & \alpha_{mn}\end{bmatrix}, \\ &S \approx (\lambda_m - \lambda_n)^{-1}\begin{bmatrix}\alpha_{nm} & \beta_{nm} \\ \lambda_m\beta_{nm} & \alpha_{nm}\end{bmatrix}.\end{aligned} \quad (3.10)$$

Let $l_m = 2\pi\, |\lambda_m|^{-1/2}$ and $l_n = (1 + \delta)\, l_m$ denote the local wavelengths of the problem (3.3). Then an evaluation of the criterion (3.9) shows that (3.7) is more block diagonal than (3.3) if the relative variation of all coefficients of (3.3) over a distance $\approx l_m\, \delta^{-1}$ is small. The efficiency of (3.5) in reducing the coupling increases by a larger eigenvalue separation and a weaker variation of the coefficients. There is no need for a precise formulation of these conditions because the criterion (3.9) can be used operationally. We just try (3.5) and solve (3.6) by Newton's method starting with (3.10). If there is convergence we check the criterion (3.9) and the decision whether to accept (3.5) is all automatic.

For the full system (2.14) we perform a sequence of transformations (3.5) by pairing of 2×2 blocks as in (3.3). We start at the corner $m = 1$, $n = M$, which exhibits the largest eigenvalue separation. Then we sweep over block diagonals $n = m + j, j = M - 1, ..., 1$, towards the main diagonal. Once (3.9) fails, say for block (m, n), then further attempts on row m are considered wasteful. Moreover, if (3.4) is small enough for a particular (m, n) this block is bypassed. When the transformation (3.5) operates on the full system (2.14) it will affect all blocks on the rows and columns with block indices m and n. For example, the blocks of the nth column are replaced by

$$A_{kn} := A_{kn} + A_{km}R, \qquad k = 1, 2, ..., M, \qquad k \neq n, m,$$

and the perturbation is small for R small. Similar expressions hold in the other cases.

The transformations (3.5) must be computed and stored at all grid points (2.12). The cost for one full sweep over all blocks is $\approx 16NM^3$ operations.

When the above reduction process is completed we inspect the transformed system, now represented by blocks $\hat{A}_{mn}$, and decide on a suitable form of the iteration operator on the left-hand side of (3.2). Tentatively we have used the following recipe. In (3.2) we require that $n \in m(n)$ if $n = m$ or if $\mathscr{L}(\hat{A}_{mn}) > 1$. In this way one arrives at a banded iteration

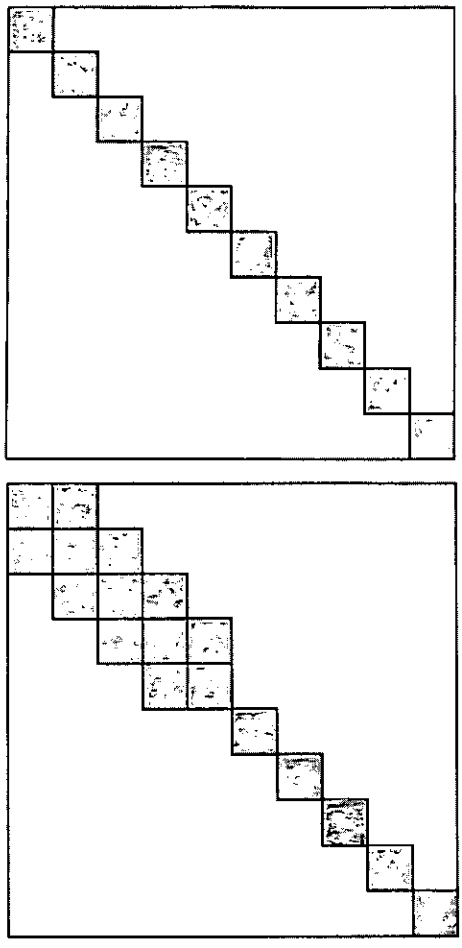

FIG. 4. The structure of the iteration operator (3.2) for example (2.16) for $f = 100$ Hz (top) and $f = 500$ Hz (bottom).

operator which in favorable cases is partitioned into a number of separate subsystems. At best one obtains a block diagonal system as in (3.1). At worst the transformations (3.5) do not work at all and one is left with a fully coupled system (2.14). For example, if the above reduction process is applied to example (2.16) the structure of the iteration operator in (3.2) is represented by the shaded blocks in Fig. 4.

4. PARTITION INTO ONE-WAY EQUATIONS

The iteration operator of the scheme (3.2) consists of a number of banded subsystems of the form

$$\begin{bmatrix} u_m \\ v_m \end{bmatrix}' = \sum_{|m-n| \le b_s} A_{mn} \begin{bmatrix} u_n \\ v_n \end{bmatrix}, \qquad 0 \le \xi \le l, \tag{4.1}$$

for $m = 1, 2, ..., M_s$. Such a subsystem may comprise all modes ($M_s = M$), merely a 2×2 block ($M_s = 1$), or anything intermediate. We retain the notation (2.14) although it is understood that we are dealing with a modified system that was the result of the above reduction process.

Sometimes the system (4.1) can be simplified by division into two separate subsystems of equal size which represent right- and leftgoing waves. Solving two subsystems instead of a full one reduces the computational cost by a factor four. Still more important, such a partition converts the boundary value problem for (4.1) into two initial-value problems, which are cheaper to solve.

Again we shall apply transformations of Riccati type to make a split into forward and backward components. The conditions to succeed are more favorable here because the eigenvalue separation between the two blocks is usually larger than the one within each block. In particular, this is pronounced for high frequencies. Sometimes this enables a block decomposition even though all mode couplings within each block must be retained. As before we shall not aim at a complete factorization. Instead, we hope that the remaining coupling between the forward and backward block is so weak that a full solution is obtained within a few iterations between the blocks.

The division into two blocks will be based on the corresponding one for each of the diagonal blocks A_{mm}, $m = 1, ..., M_s$. This is a natural approach here because it was assumed that the system (2.14), and thereby (4.1), decouples into separate 2×2 systems of constant coefficients around the endpoints. If one of the 2×2 diagonal systems fails a decomposition in some part of $[0, l]$, then we shall leave the system (4.1) as it stands. This is the most troublesome case and it is left for future study. As in [7] one could use local eigenvectors of the full system. We shall do so for a single 2×2 system but the generalization to large systems is so complicated that its use is open to doubt.

To begin with we want to compute the general solution of the 2×2 system

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}' = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \qquad 0 \le \xi \le l. \tag{4.2}$$

Of course, once this is done, a specific solution satisfying any initial or boundary conditions is easily found. Moreover, using two linearly independent solutions one can construct the Green's function which is needed for the evaluation of (3.4).

As a prototype for (4.2) we shall take the case

$$a_{11} = a_{22} \equiv 0, \qquad a_{12} \equiv 1, \qquad a_{21} = \lambda(\xi) \tag{4.3}$$

which is equivalent to the scalar equation

$$u'' = \lambda(\xi)\, u, \qquad 0 \le \xi \le l.$$

For oscillatory problems one often has $\lambda(\xi) \approx \lambda_0 < 0$, while the number of wavelengths $\tilde{l} = \frac{1}{2} l\, |\lambda_0|\, \pi^{-1}$ over $[0, l]$ is very

large. The computational cost for a finite difference method increases more than linearly by l [4]. Under certain conditions we want to show that one can solve (4.2) at the expense of a few CPU seconds and this is rather independent of the size of l. This is made possible by the use of the WKB-approximation. It amounts to a transformation of (4.2) into almost diagonal form. This is very efficient because it can be accomplished on a scale larger than the wavelength. However, WKB is usually applicable only in part of $[0, l]$ and elsewhere it must be supplemented by a finite difference scheme or the like. The solutions are then tied together at the connection points by continuity conditions. To ensure a smooth transition all solutions must be produced under strict error control. The idea is already used in [9] and we shall develop it further here.

As a first step in diagonalizing [4.2] we introduce a transformation based on the eigenvectors of $[a_{ij}]$. Let $[\xi_0, \xi_1]$ be a subinterval in which the eigenvalues $\sigma_{1,2}$ of $[a_{ij}]$ are separated. Then, with

$$\begin{aligned} &\mathbf{u} = \tilde{T}(\xi)\,\tilde{\mathbf{u}}, \qquad \tilde{T} = \begin{bmatrix} 1 & 1 \\ \tilde{\sigma}_1 & \tilde{\sigma}_2 \end{bmatrix}, \\ &\tilde{T}^{-1} = d^{-1}\begin{bmatrix} \tilde{\sigma}_2 & -1 \\ -\tilde{\sigma}_1 & 1 \end{bmatrix}, \\ &d = (\sigma_2 - \sigma_1)/a_{12}, \qquad \tilde{\sigma}_i = (\sigma_i - a_{11})/a_{12}, \\ &\sigma_{1,2} = (a_{11}+a_{22}) \pm (a_{12}a_{21} + \tfrac{1}{4}(a_{11}-a_{22})^2)^{1/2}, \end{aligned} \tag{4.4}$$

we obtain

$$\begin{aligned} &\begin{bmatrix} \tilde{u}_1 \\ \tilde{u}_2 \end{bmatrix}' = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}\begin{bmatrix} \tilde{u}_1 \\ \tilde{u}_2 \end{bmatrix}, \qquad \xi_0 \leqslant \xi \leqslant \xi_1, \\ &b_{11} = \sigma_1 + \tilde{\sigma}_1' d^{-1}, \qquad b_{12} = \tilde{\sigma}_2' d^{-1}, \\ &b_{21} = -\tilde{\sigma}_1' d^{-1}, \qquad b_{22} = \sigma_2 - \tilde{\sigma}_2' d^{-1}. \end{aligned} \tag{4.5}$$

In the case (4.3) the omission of b_{12} and b_{21} results in the classical WKB-solution

$$u \approx |\lambda|^{-1/4} \exp\left(\pm \int^{\xi} \sqrt{\lambda(t)}\, dt \right). \tag{4.6}$$

Next we try to suppress b_{12}, b_{21} by setting

$$\begin{aligned} &\tilde{\mathbf{u}} = \hat{T}(\xi)\,\hat{\mathbf{u}}, \qquad \hat{T} = \begin{bmatrix} 1+rs & r \\ s & 1 \end{bmatrix}, \\ &\hat{T}^{-1} = \begin{bmatrix} 1 & -r \\ -s & 1+rs \end{bmatrix}, \end{aligned} \tag{4.7}$$

where r and s are the solution of

$$\begin{aligned} &(b_{11}-b_{22})\,r + b_{12} - r^2 b_{21} = 0, \qquad \xi_0 \leqslant \xi \leqslant \xi_1, \\ &(b_{22} - b_{11} + 2b_{21}r)\,s + b_{21} = 0. \end{aligned} \tag{4.8}$$

Then (4.5) goes over into

$$\begin{aligned} &\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \end{bmatrix}' = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \end{bmatrix}, \qquad \xi_0 \leqslant \xi \leqslant \xi_1, \\ &c_{11} = b_{11} - rb_{21} - r's, \qquad c_{12} = -r', \\ &c_{22} = b_{22} + rb_{21} + sr', \qquad c_{21} = -s' + r's^2. \end{aligned} \tag{4.9}$$

The transformation (4.7) and Eqs. (4.8) are the scalar analogue of (3.6), (3.7). By substitution of the case (4.3) one finds that the off-diagonal elements of (4.9) are smaller than those of (4.5) if the dimensionless parameters

$$|\lambda'\lambda^{-3/2}| \qquad \text{and} \qquad |\lambda''/(\lambda^{1/2}\lambda')|$$

are small. For $\lambda < 0$ this is the well-known condition for the validity of the WKB-approximation (4.6), namely that the relative variation of λ and λ' over one wavelength must be small. We shall now estimate the error when c_{12} and c_{21} are omitted in (4.9). Although small, their influence might be appreciable [5]. The estimate will be based on the following lemma [17].

LEMMA 4.1. *Consider the scalar initial value problem*

$$\begin{aligned} &u' = c_{11}(\xi)\,u + f(\xi), \qquad \xi_0 \leqslant \xi \leqslant \xi_1, \\ &y(\xi_0) = 0, \end{aligned} \tag{4.10}$$

and assume that there is a constant α, $\alpha \leqslant 0$, *such that*

$$\operatorname{Re} \int_t^x c_{11}\, d\xi \leqslant \alpha(x-t), \qquad \xi_0 \leqslant t \leqslant x \leqslant \xi_1. \tag{4.11}$$

Then the solution of (4.10) *satisfies the estimates*

$$\|u\| \leqslant \|f\|_1 \tag{4.12}$$

and

$$\|u\| \leqslant |\alpha|^{-1}\,\|f\|, \qquad \textit{for} \quad \alpha \neq 0, \tag{4.13}$$

The same estimates hold in the case that an end condition $y(\xi_1) = 0$ *is given, provided that* $\alpha \geqslant 0$.

The estimate (4.13) is usually sharper if the damping is strong, that is, $\alpha(\xi_1 - \xi_0) \ll -1$, while only (4.12) remains for problems with no damping at all ($\alpha = 0$). In any case, for $\alpha \neq 0$, one could evaluate both bounds and select the smallest one.

Consider now (4.9) with the boundary conditions

$$\hat{u}_1(\xi_0) = A, \qquad \hat{u}_2(\xi_1) = B. \tag{4.14}$$

Then the solution $\hat{\mathbf{u}}_d$ of the diagonal part of the system (4.9) is given by

$$\hat{u}_{1d} = A \exp\left(\int_{\xi_0}^{\xi} c_{11}\, ds\right), \qquad \hat{u}_{2d} = B \exp\left(\int_{\xi_1}^{\xi} c_{22}\, ds\right). \tag{4.15}$$

An upper bound of the relative error of $\hat{\mathbf{u}} - \hat{\mathbf{u}}_d$ is given by

THEOREM 4.1. *Consider the system* (4.9), (4.14) *and assume that* $\alpha \leqslant 0$, *where*

$$\alpha = \max\{\max_{[\xi_0,\xi_1]} \operatorname{Re} c_{11}, -\min_{[\xi_0,\xi_1]} \operatorname{Re} c_{22}\}. \tag{4.16}$$

Then the error $\mathbf{e} = \hat{\mathbf{u}} - \hat{\mathbf{u}}_d$ *is bounded by*

$$\begin{aligned} \|e_1\| &\leqslant \gamma(1-\gamma^2)^{-1}(\gamma\,\|\hat{u}_{1d}\| + \|\hat{u}_{2d}\|), \\ \|e_2\| &\leqslant \gamma(1-\gamma^2)^{-1}(\|\hat{u}_{1d}\| + \gamma\,\|\hat{u}_{2d}\|), \end{aligned} \tag{4.17}$$

provided $\gamma < 1$, *where*

$$\gamma = \max\{\|c_{12}\|_1, \|c_{21}\|_1\}. \tag{4.18}$$

If $\alpha < 0$, *then* (4.17) *also holds with*

$$\gamma = |\alpha|^{-1} \max\{\|c_{12}\|, \|c_{21}\|\}.$$

Proof. The proof follows by a straightforward application of Lemma 4.1.

The relative error is essentially bounded by γ when γ is small. Now by (4.9), (4.18) we have

$$\gamma = \max\left\{\int_{\xi_0}^{\xi_1} |r'|\, d\xi, \int_{\xi_0}^{\xi_1} |s' - r's^2|\, d\xi\right\}, \tag{4.19}$$

where r and s are the solutions of Eqs. (4.8). Therefore we can control the error of $\hat{\mathbf{u}}_d$ by selecting an interval $[\xi_0, \xi_1]$ so that

$$\gamma \leqslant \varepsilon, \tag{4.20}$$

where ε is the desired error tolerance. For the prototype problem (4.3) we obtain

$$|r| \approx |s| \approx \tfrac{1}{8}\,|\lambda'\lambda^{-3/2}|, \tag{4.21}$$

provided this is so small that the nonlinear terms of (4.8) are harmless. Now if r and s are monotone in $[\xi_0, \xi_1]$ it follows from (4.19) that

$$\gamma \approx |r(\xi_1)| + |r(\xi_0)| \tag{4.22}$$

and the error estimate is directly related to the dimensionless quantity (4.21). For example, in the simple case

$$\lambda = \omega^2 a \xi^m, \qquad m = 1, 2, \ldots,$$

the criterion

$$\tfrac{1}{4}|\lambda'\lambda^{-3/2}| \leqslant \varepsilon$$

is fulfilled outside the interval $[\xi_1, \xi_2]$, where

$$\xi_2 = -\xi_1 = (m/(4\varepsilon\,|\omega|\,\sqrt{|a|}))^{2/(2+m)}.$$

The length of the interval $[\xi_1, \xi_2]$ in terms of the local wavelength at the endpoints is given by $\hat{l} = m(4\pi\varepsilon)^{-1}$. Thus it is independent of the frequency ω and for $\varepsilon = 0.01$, $m = 1,2$, $\hat{l} \approx 10$. In the general case we actually find a maximal interval $[\xi_0, \xi_1]$ so that (4.20) holds at last approximately. In solving this problem one could use (4.22) as guidance. Elsewhere, say $[\xi_1, \xi_2]$, where (4.20) is violated, the system (4.1) must be solved numerically. In the case (4.3) this happens, for example, around points where λ goes through zero. Often one can expect that the general solution exhibits both exponentially growing and decaying components in some part of $[\xi_1, \xi_2]$. In order to avoid ill-conditioning the problem must be treated numerically as a boundary value problem. We propose Numerov's fourth-order accurate scheme

$$\begin{aligned} &(1 - \tfrac{1}{2}hA_{v+1} + \tfrac{1}{12}h^2(A_{v+1}^2 + A'_{v+1}))\,\mathbf{u}_{v+1} \\ &\quad = (1 + \tfrac{1}{2}hA_v + \tfrac{1}{12}h^2(A_v^2 + A'_v))\,\mathbf{u}_v, \\ &\quad v = 0, 1, \ldots, J-1, \end{aligned} \tag{4.23}$$

where $A_v = [a_{ij}(x_v)]$, $x_v = \xi_1 + vh$, $h = (\xi_2 - \xi_1)/J$. Two linearly independent solutions $\mathbf{u}_A$ and $\mathbf{u}_B$ are now determined by the constraints

$$\begin{aligned} &\hat{u}_{1A}(\xi_1) = 1, \qquad \hat{u}_{2A}(\xi_2) = 0, \qquad \hat{\mathbf{u}} = T^{-1}\mathbf{u}, \\ &\hat{u}_{1B}(\xi_1) = 0, \qquad \hat{u}_{2B}(\xi_2) = 1, \end{aligned} \tag{4.24}$$

where $T = \tilde{T}\hat{T}$ is the transformation (4.4), (4.7). It implies that incoming waves of unit amplitude are specified at the endpoints. The pentadiagonal system (4.23), (4.24) takes $\approx 25J$ operations to solve by Gauss elimination. In practice, for $\varepsilon = 0.01$ it is sufficient to take $J \approx 100$. The actual error of the numerical solution is controlled by Richardson extrapolation.

The numerical scheme (4.23), (4.24) can be combined with the WKB-approach (4.15) as follows. Divide the interval $[0, l]$ into K nonoverlapping subintervals $[\xi_{k-1}, \xi_k]$, $k = 1, 2, \ldots, K$, $\xi_0 = 0$, $\xi_K = l$, such that the criterion (4.20) is fulfilled in every other subinterval. The first and the last one

will always satisfy (4.20) because $[a_{ij}]$ is constant around the endpoints $\xi = 0, l$. Then for odd intervals we use the solution (4.15), while even intervals are handled by the scheme (4.23), (4.24). Let A_k, B_k denote the excitation coefficients in each subinterval. A global solution is now assembled by setting $A_K = 1$ and $B_K = 0$. We then proceed to the left and determine A_k, B_k step by step. These calculations are explicit owing to the boundary conditions (4.24). To avoid over- and underflow on the computer only the exponents of (4.15) and the logarithm of A_k, B_k should be computed. Finally we normalize by A_1 so that this solution is the response of a unit amplitude wave coming from the left. If $A_1 \approx 0$ there is resonance or near resonance. Similarly, another solution, linearly independent from the first one, is obtained by letting a wave arrive from the right. These two solutions are suitable for the construction of the Green's function corresponding to the boundary conditions (2.15).

The error control guarantees that the relative error is at most ε. By experience $\varepsilon = 0.01$ seems to be a good compromise between computational economy and accuracy. If ε is chosen too small then the WKB-solution is less usuable and the interval for the numerical scheme (4.23) increases and so does the cost. As an illustration consider the problem

$$\begin{aligned} &u'' = \lambda(\xi)\, u, && 0 \leqslant \xi \leqslant 1000, \\ &\lambda^{-1/2} u' + u = 2, && \xi = 0, \\ &\lambda^{-1/2} u' - u = 0, && \xi = 1000, \end{aligned} \tag{4.25}$$

where

$$\lambda(\xi) = 25 f(\xi),$$

$$f(\xi) = -1 + \exp(-t_1^2) + (25\pi)^{-1} \int_0^{\xi} \exp(-t_2^2)\, dt,$$

$$t_1 = (\xi - 400)/100, \qquad t_2 = 4(t - 750)/250;$$

see Fig. 5.

There are two turning points, $\xi \approx 400$ and $\xi \approx 775$. The solution of (4.25) around these points is shown in Fig. 6. The intervals in which the solution has been computed by the scheme (4.23) are also indicated in Fig. 6. The CPU time for $\varepsilon = 0.01$ was 1.5 s. If the scheme (4.23) is used over the whole interval [0, 1000] then $\approx 30{,}000$ points are required for two digits of accuracy. Alternatively, if $\lambda(\xi)$ is approximated by a stepwise constant function with an analytic solution in each subinterval, then ≈ 3000 steps are necessary with a uniform partition.

A decomposition of the full system (4.1) into two separate subsystems for forward and backward waves can now be based on the corresponding partition for the 2×2 case (4.2). This is possible if every 2×2 system extracted from the diagonal blocks of (4.1) admits a factorization throughout $[0, l]$. Then we introduce a transformation $\mathbf{u} = TP\hat{\mathbf{u}}$, where T is a block diagonal matrix formed by the transformations (4.4), (4.7); P is a permutation matrix which reorders $\mathbf{u}$ such

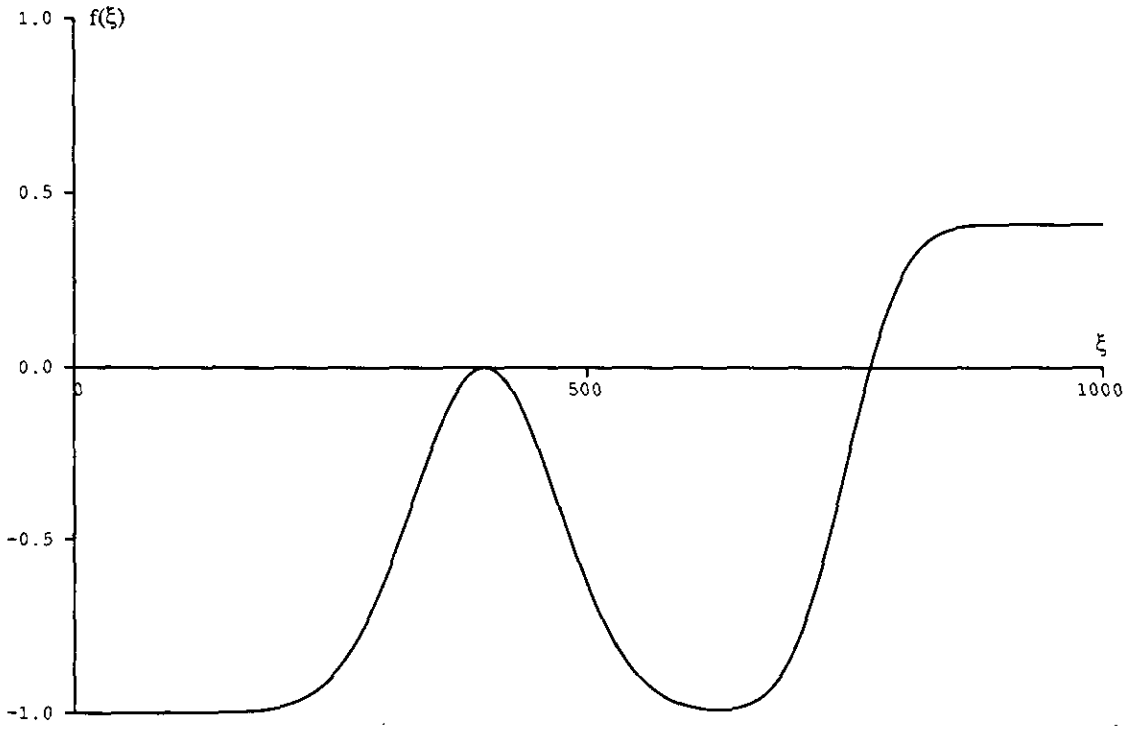

FIG. 5. The graph of $f(\xi)$ used in (4.25).

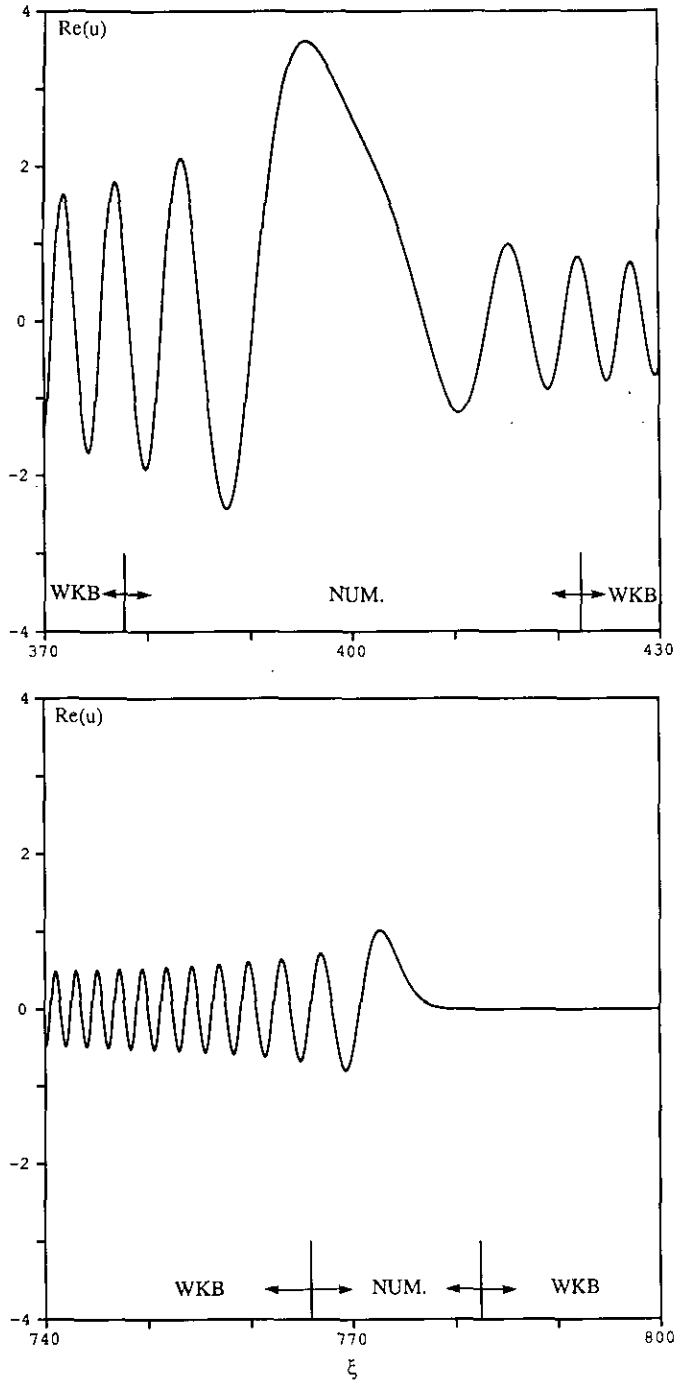

FIG. 6. The graph of the solution of (4.25) around the turning points $\xi \approx 400$ (top) and $\xi \approx 775$ (bottom).

that forward components come first. Then the system (4.1) takes the form

$$\hat{u}' = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix} \hat{u}, \qquad 0 \leq \xi \leq l,$$

where $\hat{A}_{ij}$ are $M_s \times M_s$ band matrices. The eigenvalues of $\hat{A}_{11}$ and $\hat{A}_{22}$ differ essentially only in sign and their separation is usually much larger than the one within each $\hat{A}_{ii}$. This makes it possible to suppress $\hat{A}_{ij}$, $i \neq j$. The simplest way is to apply a sequence of transformations (4.7) for 2×2 systems (4.5) which now is obtained by extracting 2×2 principal matrices with one entry from each $\hat{A}_{ij}$. The remainder of $\hat{A}_{ij}$, $i \neq j$, after one sweep over its band is put on the right-hand side of the iteration scheme (3.2). A complete annihilation of $\hat{A}_{ij}$, $i \neq j$, would require repeated sweeps which is hardly worth the effort. The final result, for example, (2.16) is shown in Fig. 7.

As a result, in each iteration for $f = 100$ Hz we need to solve 12 scalar initial value problems and four boundary value problems for 2×2 systems. For $f = 500$ Hz we end up with 12 subsystems of one-way equations. Two of them are 5×5 while the others are scalar. The total CPU time to generate the grid (2.12), set up the system (2.14) by eigenvalue computations, make the decomposition (3.2), and obtain convergence by (3.2) was 248 s for $f = 100$ Hz. This is used as a reference in the comparison of different methods

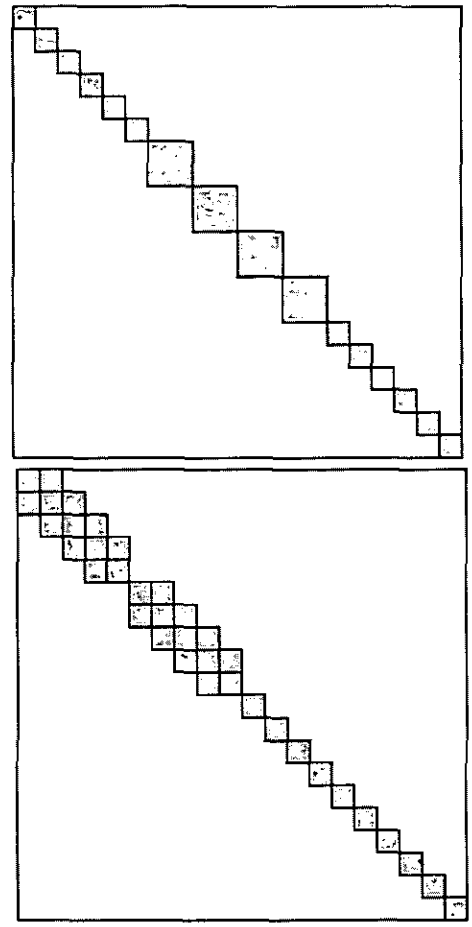

FIG. 7. The structure of the iteration operator (3.2) for example (2.16) for $f = 100$ Hz (top) and $f = 500$ Hz (bottom) after partition into one-way equations. The larger blocks for $f = 100$ Hz are 2×2 and cannot be split. The smaller ones denote a single component. This figure should be compared with Fig. 4.

TABLE I

Relative CPU Times for the Solution of the Problem (2.16) for $f = 100$ Hz and $f = 500$ Hz

f	100	500
A	1	9
B	1.4	16
C	8	32
D	54	268

Note. A, the iteration scheme (3.2) partitioned as in Fig. 7; B, the iteration scheme (3.1); C, the finite difference scheme (4.23) applied to the full system (2.14); D, a finite difference solution of (2.1). The figures for methods C and D are estimates based on an operations count for a bandsolver.

in Table I. The spatial resolution was $\Delta\xi = l/L$ with $L = 4000$ and $L = 20{,}000$ for $f = 100$ Hz and $f = 500$ Hz, respectively.

It is apparent that the iteration schemes (3.1) or (3.2) are more efficient than a direct solution of the finite-difference equations for (2.14) or (2.1). The gain in using the more sophisticated scheme (3.2) over the simple one (3.1) is less for $f = 100$ Hz than for $f = 500$ Hz. For $f = 100$ Hz mode couplings are relatively weak and the iteration operator of (3.1) is dominant enough to achieve convergence in a few iterations. Then the extra cost of increasing the diagonal dominance and the partitioning into one-way equations hardly pays off. However, for $f = 500$ Hz the iterations (3.1) barely converge and it is likely they would fail for larger f. By enlarging the iteration operator as illustrated in Fig. 4 (bottom) one obtains rapid convergence. Yet this iteration scheme would be expensive without the split of the 10×10 system (each block in Fig. 4 represents two unknowns) into two 5×5 systems of one-way equations as shown in Fig. 7 (bottom). This measure counts for the major part of the time savings in this case.

5. CONCLUDING REMARKS

The purpose of this paper is to extend the classical method of separation of variables for the reduced wave equation (normal-mode theory) to cases which are almost separable in the sense that the axial scale of variation of the transversal eigensystem is larger than the characteristic wavelength of the problem. This condition is local and allows a large departure from separability over a distance of many wavelengths. The nonseparability leads to a fully coupled system of ordinary differential equations for the modes. This system can be simplified by computations that need not be done on a scale of wavelength. First, one can increase the diagonal dominance by transformations of Riccati type. This decoupling process is always applicable for modes with a large eigenvalue separation. In this way one arrives at an iteration scheme which is banded and partitioned into a number of separate subsystems. Second, each subsystem can sometimes be split into two halves and thereby convert a boundary value problem into two initial-value problems. The success of the above operations depends on the degree of separability, a question that must be probed numerically. Therefore it is necessary to introduce computational criteria like (3.9) to make the whole decoupling process automatic.

As remarked above we have not attempted to exhaust all possibilities to make the coupled equations numerically tractable by computations on a relatively large spatial scale. However, further extensions must be weighted against increased complexity of the algorithm.

For a three-dimensional duct the only difference is that the eigenvalue problem (2.5) is now defined over a two-dimensional cross section. Again, if almost separable conditions hold over the cross section this eigenvalue problem could be handled by the technique of separation of variables along the same ideas as above. Then the entire solution procedure is essentially reduced to the solution of a number of one-dimensional systems, each of which contains one or a few unknowns.

A fast iterative solver for a finite-difference scheme for (2.1) could be designed similar to the continuous case. An eigenfunction expansion for the discrete solution leads to an almost diagonal system for which a suitable iteration operator could be identified as above. However it seems more natural to apply the separation ansatz (1.1) directly on (2.1).

REFERENCES

1. L. Abrahamsson, *J. Comput. Appl. Math.* **34**, 305 (1991).
2. L. Abrahamsson, Tech. Rpt. D20170-2.7, Nat. Def. Res. Est., Sweden, 1991 (unpublished).
3. A. Bayliss, C. I. Goldstein, and E. Turkel, *J. Comput. Phys.* **51**, 443 (1983).
4. A. Bayliss, C. I. Goldstein, and E. Turkel, *J. Comput. Phys.* **59**, 396 (1985).
5. R. Bellman, *Stability Theory of Differential Equations* (Mc. Graw-Hill, New York, 1953), p. 42.
6. R. Bellman and R. Vasudevan, *Wave Propagation, an Invariant Imbedding Approach* (Reidel, Dordrecht, 1986).
7. I. B. Bernstein, L. Brookshaw, and P. A. Fox, *J. Comput. Phys.* **98**, 269 (1992).
8. C. de Boor, *A Practical Guide to Splines* (Springer-Verlag, Berlin, 1978).
9. J. R. Brannan, G. P. Forney, and R. F. Henrick, *J. Comput. Phys.* **66**, 21 (1986).
10. L. Brekhovskikh and Yu. Lysanov, *Fundamentals of Ocean Acoustics* (Springer-Verlag, Berlin, 1982).
11. Y.-C. Cho, *J. Acoust. Soc. Am.* **67**, 1421 (1980).
12. R. B. Evans, *J. Acoust. Soc. Am.* **74**, 188 (1983).
13. W. Eversman, *J. Sound Vib.* **47**, 515 (1976).

14. G. J. Fix and S. P. Marin, *J. Comput. Phys.* **28**, 253 (1978).
15. I. Karasalo and L. Westerling, Tech. Rpt. D 20138-2.2, Nat. Def. Res. Est., Sweden, 1988 (unpublished).
16. H.-O. Kreiss, *Math. Comput.* **26**, 605 (1972).
17. H.-O. Kreiss, *SIAM J. Numer. Anal.* **16**, 980 (1979).
18. H.-O. Kreiss, N. K. Nichols, and D. L. Brown, *SIAM J. Numer. Anal.* **23**, 325 (1986).
19. P. M. van Loon, *Continuous Decoupling Transformations for Linear Boundary Value Problems*, CWI Tracts (Math. Centrum, Amsterdam, 1988).
20. M. L. Munjal, *Acoustics of Ducts and Mufflers* (Wiley, New York, 1987).
21. W. Möhring, *J. Acoust. Soc. Am.* **64**, 1186 (1978).
22. B. N. Parlett, *The Symmetric Eigenvalue Problem* (Prentice–Hall, Englewood Cliffs, NJ, 1980).
23. A. D. Pierce, *J. Acoust. Soc. Am.* **37**, 19 (1965).