ITERATIVE SOLUTION OF THE HELMHOLTZ EQUATION BY A SECOND-ORDER METHOD*

KURT OTTO† AND ELISABETH LARSSON‡

Report CS-TR-3727
UMIACS-TR-96-95
December 1996

Abstract. The numerical solution of the Helmholtz equation subject to nonlocal radiation boundary conditions is studied. The specific problem is the propagation of hydroacoustic waves in a two-dimensional curvilinear duct. The problem is discretized with a second-order accurate finite-difference method, resulting in a linear system of equations. To solve the system of equations, a preconditioned Krylov subspace method is employed. The preconditioner is based on fast transforms, and yields a direct fast Helmholtz solver for rectangular domains. Numerical experiments for curved ducts demonstrate that the rate of convergence is high. Compared with band Gaussian elimination the preconditioned iterative method shows a significant gain in both storage requirement and arithmetic complexity.

* This research was supported by the U. S. National Science Foundation under grant ASC-8958544 and by the Swedish National Board for Industrial and Technical Development (NUTEK).

† Department of Scientific Computing, Uppsala University, Box 120, S-751 04 Uppsala, Sweden (kurt@tdb.uu.se). Part of this work was performed during a postdoctoral visit at the Dept. of Computer Science, Univ. of Maryland, College Park, MD.

‡ Department of Scientific Computing, Uppsala University, Box 120, S-751 04 Uppsala, Sweden (bette@tdb.uu.se).

1. Introduction. The Helmholtz equation arises in many physical applications, e.g., scattering problems in electromagnetics and acoustics [Ernst94], [AbKr94]. In realistic applications, a wide range of wavenumbers is often of interest. For a finite element (or finite-difference) discretization of the two-dimensional Helmholtz equation, it is necessary that the number of grid points grows faster than quadratically in the wavenumber in order to maintain a given accuracy [BaGoTu85a], [IhlBa97]. Thus, for high wavenumbers, the discretized Helmholtz equation "leads to a huge linear system of equations" [AbKr94]. Due to the large bandwidth, the storage requirement renders Gaussian elimination prohibitive. To handle high wavenumbers and large domains for the Helmholtz equation in duct acoustics, Abrahamsson and Kreiss [AbKr94] devised a special iteration technique related to separation of variables. However, the effectiveness of the method relies on the degree of separability of the problem. Another way to address the computational difficulties for the discretized Helmholtz equation is to design iterative methods. Bayliss et al. [BaGoTu83] used a preconditioned conjugate gradient method applied to the normal equations for a finite element discretization [BaGuTu82]. Due to the ill-conditioning of the normal equations, the unpreconditioned algorithm suffered from extremely slow convergence. The convergence rate was substantially improved through preconditioners based on symmetric successive overrelaxation [BaGoTu83]; or a multigrid V-cycle [BaGoTu85b], [Gold82]; only for the Laplacian part of the Helmholtz operator. Recently, the iterative quasi-minimal residual algorithm has been applied to capacitance matrix methods for exterior Helmholtz problems [Ernst94].

The objective of this paper is to develop a technique for solving the Helmholtz equation with an iterative method. In order to be a viable method, it should exploit the sparsity of the discretization matrix in an efficient way, converge rapidly, and be competitive with Gaussian elimination in regard to the total arithmetic complexity. Our approach is to apply a preconditioned Krylov subspace method [FrGoNa92] directly to the discretized equations. Typically and especially for high wavenumbers, the discretization matrix is large, complex, indefinite, and ill-conditioned. As a result, standard preconditioning techniques like diagonal scaling and incomplete LU decomposition are likely to do poorly. Instead we construct preconditioners based on fast transforms, see [Otto96] and the survey in [ChanNg96]. In order to get a highly structured matrix, facilitating the design of the preconditioner, a finite-difference method is used for the discretization. For the same reason, special attentention is given to the choice of radiation boundary conditions. A finite element method would be more flexible for complicated geometries, but also less amenable to fast transform-based preconditioners. This is particularly noticeable for higher orders of approximation, where some of the degrees of freedom typically are *not* node values.

The paper is organized as follows. In §2 the governing equations, the boundary conditions, and the finite-difference discretization of a Helmholtz problem are derived. The specific problem is the propagation of hydroacoustic waves in a curvilinear duct. The same technique would easily carry over to, e.g., an electromagnetic waveguide. Issues concerning the preconditioner are treated in §3. Section 4 is devoted to computational aspects with an emphasis on resolution criteria, i.e., relations between the wavenumber, the grid size, and the desired accuracy. Finally, numerical experiments are presented in §5 followed by conclusions.

2. The model problem. In this section the theory needed to determine the system of equations for the model problem is discussed.

2.1. Notation. The quantity I_m denotes the identity matrix of order m. The square matrices $\mathrm{diag}_{j,m}(\beta_j)$ and $\mathrm{trid}_{j,m}(\alpha_j, \beta_j, \gamma_j)$ are defined in the following way:

$$\mathrm{diag}_{j,m}(\beta_j) = \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_m \end{pmatrix},$$

$$\mathrm{trid}_{j,m}(\alpha_j, \beta_j, \gamma_j) = \begin{pmatrix} \beta_1 & \gamma_1 & & & \\ \alpha_2 & \beta_2 & \gamma_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \alpha_{m-1} & \beta_{m-1} & \gamma_{m-1} \\ & & & \alpha_m & \beta_m \end{pmatrix}.$$

2.2. Governing equations. We study the propagation of time-harmonic sound waves under water. Neglecting sound absorption and assuming that the fluid is homogeneous, the waves are governed by the Helmholtz equation

$$-\frac{\partial^2 u}{\partial x_1^2} - \frac{\partial^2 u}{\partial x_2^2} - \kappa^2 u = 0, \tag{1}$$

where $u(x_1, x_2)$ is the phasor of the acoustic pressure $\Re e(u(x_1, x_2)e^{-i2\pi f t})$. The wavenumber is given by $\kappa = 2\pi f/c$, where f is the frequency, and $c = 1500$ m/s is the sound speed. For heterogeneous media, the sound speed and consequently the wavenumber would depend on the space coordinates.

We consider a physical domain

$$\begin{cases} x_1 = x_1(\xi_1, \xi_2) \\ x_2 = x_2(\xi_1, \xi_2) \end{cases}$$

that can be mapped onto the unit square

$$\begin{cases} 0 \le \xi_1 \le 1 \\ 0 \le \xi_2 \le 1 \end{cases}$$

via an orthogonal transformation. Equation (1) is then transformed into

$$-\frac{\partial}{\partial \xi_1}\left(a \frac{\partial u}{\partial \xi_1}\right) - \frac{\partial}{\partial \xi_2}\left(a^{-1} \frac{\partial u}{\partial \xi_2}\right) - eu = 0, \tag{2}$$

where the metric coefficients a and e are given by

$$a = \sqrt{\frac{\left(\frac{\partial x_1}{\partial \xi_2}\right)^2 + \left(\frac{\partial x_2}{\partial \xi_2}\right)^2}{\left(\frac{\partial x_1}{\partial \xi_1}\right)^2 + \left(\frac{\partial x_2}{\partial \xi_1}\right)^2}},$$

$$e = \kappa^2 \sqrt{\left(\left(\frac{\partial x_1}{\partial \xi_1}\right)^2 + \left(\frac{\partial x_2}{\partial \xi_1}\right)^2\right)\left(\left(\frac{\partial x_1}{\partial \xi_2}\right)^2 + \left(\frac{\partial x_2}{\partial \xi_2}\right)^2\right)}.$$

2.3. Boundary conditions. We now choose the physical domain to be a two-dimensional duct, see the shaded area in Fig. 1.

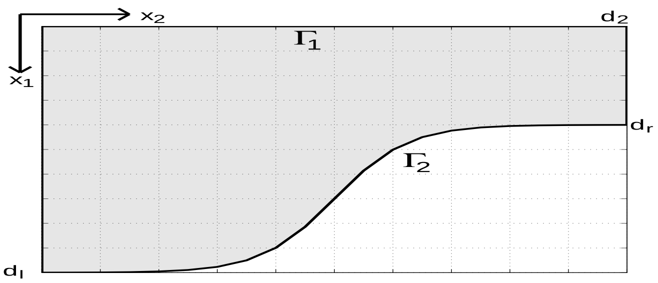

FIG. 1. *The physical domain.*

For the problem to be well-posed, conditions are needed on all four boundaries. Our model problem is partly fixed by letting the physical boundary Γ_1 be a soft wall (air), whereas Γ_2 is a rigid wall (rock). The sound field is generated by a source along $x_2 = 0$, specified by a source term $g(\xi_1) \equiv g(x_1(\xi_1, 0))$. At $x_2 = d_2$ an artificial far-zone boundary has been introduced. Originally, the domain is semi-infinite ($d_2 \to \infty$), but for computational reasons it is truncated by assigning d_2 some finite value. For the soft wall Γ_1, the boundary condition is $u = 0$ (pressure release). This leads to

$$u(0, \xi_2) = 0, \quad 0 \leq \xi_2 \leq 1, \tag{3}$$

in the computational domain. Since the bottom Γ_2 is rigid, a condition on the normal derivative is imposed:

$$\frac{\partial u}{\partial n} = 0, \quad (x_1, x_2) \in \Gamma_2.$$

Due to the orthogonal transformation this becomes

$$\frac{\partial u}{\partial \xi_1}(1, \xi_2) = 0, \quad 0 \leq \xi_2 \leq 1. \tag{4}$$

For the radiation conditions at the near- and far-zone boundaries, Dirichlet–to–Neumann (DtN) maps [KeGi89] are employed. The main reason for choosing nonlocal DtN maps, instead of the local radiation conditions described in [BaGuTu82], is that discretized DtN maps are more apt to preconditioning by fast transforms. Our design of radiation conditions follows the principles outlined in [FixMa78], where a variational formulation of DtN conditions was derived for an axially symmetric duct parametrized by cylindrical coordinates. Boundary conditions based on DtN maps require the boundary in question to be a separable coordinate surface. Moreover, for the radiation condition in [FixMa78], it is implicitly assumed that the duct could be extended beyond the artificial boundary by parallel straight walls. This is a so-called anechoical termination [AbKr94]. Since the wavenumber κ is independent of ξ_2, the above prerequisites are fulfilled by requiring the duct to be flat *only* in an infinitesimal neighborhood of $x_2 = d_2$ and $x_2 = 0$. In the present application, the wavenumber is actually a constant. Thus, without any significant loss of accuracy we can use the slightly more restrictive assumption that, in the vicinity of $x_2 = 0$, there is a local transformation

$$\begin{cases} x_1 &= \xi_1 d_\ell, \quad 0 \le \xi_1 \le 1, \\ x_2 &= \xi_2 d_2. \end{cases}$$

Substituting this into (2), together with (3) and (4), yields

$$\begin{cases} -\frac{d_2}{d_\ell}\frac{\partial^2 u}{\partial \xi_1^2} - \frac{d_\ell}{d_2}\frac{\partial^2 u}{\partial \xi_2^2} - \kappa^2 d_\ell d_2 u = 0, \\ u(0,\xi_2) = 0, \\ \frac{\partial u}{\partial \xi_1}(1,\xi_2) = 0. \end{cases} \tag{5}$$

The condition at the near-zone boundary is based on the fact that the solution for

$$0 \le \xi_1 \le 1, \quad \xi_2 \le 0,$$

can be obtained through separation of variables, i.e., $u(\xi_1,\xi_2) = \psi(\xi_1)\varphi(\xi_2)$. Solving (5) with this ansatz gives

$$\begin{aligned} \psi_m(\xi_1) &= \sqrt{2}\sin((m-\tfrac{1}{2})\pi\xi_1), & m = 1,2,\ldots, \\ \varphi_m(\xi_2) &= A_m \exp(i\sqrt{-\lambda_m}\xi_2) + B_m \exp(-i\sqrt{-\lambda_m}\xi_2), & m = 1,2,\ldots, \end{aligned}$$

where

$$\lambda_m = \big((m-\tfrac{1}{2})\pi d_2/d_\ell\big)^2 - (\kappa d_2)^2.$$

The eigenfunctions $\{\psi_m(\xi_1)\}_{m=1}^\infty$ are orthonormal with respect to the scalar product

$$\langle f, g\rangle \equiv \int_0^1 \bar{f}(\xi_1) g(\xi_1) d\xi_1.$$

The general solution to the eigenproblem (5) becomes

$$u(\xi_1,\xi_2) = \sum_{m=1}^\infty A_m \psi_m(\xi_1)\exp(i\sqrt{-\lambda_m}\xi_2) + B_m\psi_m(\xi_1)\exp(-i\sqrt{-\lambda_m}\xi_2). \tag{6}$$

For mode indices below the cutoff limit, i.e.,

$$m \le \mu_\ell = \left\lfloor \frac{\kappa d_\ell}{\pi} + \frac{1}{2} \right\rfloor, \tag{7}$$

the eigenvalues λ_m become negative, yielding propagating modes. If λ_m were positive, we would get evanescent modes. Analogously to the motivation in [FixMa78], the influence of the evanescent modes is neglible, especially on the far field. Thus, an appropriate way to truncate the series in (6) is to retain only the terms with mode indices $m \le \mu_\ell$. The situation is somewhat different for a purely exterior Helmholtz problem, where an appropriate truncation of DtN maps is a more delicate matter [GrKe95].

The A_m-terms in (6) correspond to rightgoing waves, and the B_m-terms correspond to leftgoing waves. In our model we have a source at the left boundary. We will treat the rightgoing waves as originating from a "truncated" point source positioned at depth $\xi_1 = \delta_s$ by letting

$$A_m = \langle \psi_m(\xi_1), g(\xi_1)\rangle = \psi_m(\delta_s), \quad m = 1,\ldots,\mu_\ell. \tag{8}$$

Note that leftgoing waves are feasible in order to handle possible reflections from the curved bottom. Inserting (7) and (8) into (6) yields

$$u(\xi_1,\xi_2) = \sum_{m=1}^{\mu_\ell} \psi_m(\delta_s)\psi_m(\xi_1)\exp(i\sqrt{-\lambda_m}\xi_2) + B_m\psi_m(\xi_1)\exp(-i\sqrt{-\lambda_m}\xi_2). \tag{9}$$

The coefficients B_m are determined from the solution by exploiting the orthonormality of the functions $\psi_m(\xi_1)$. From (9) we get

$$B_m = \langle\psi_m(\xi_1), u(\xi_1,0)\rangle - \psi_m(\delta_s). \tag{10}$$

The nonlocal boundary condition at $\xi_2 = 0$ is obtained by differentiating (9) with respect to ξ_2 and using (10). Thus,

$$\begin{aligned} -\frac{\partial u}{\partial \xi_2}(\xi_1,0) &- i\sum_{m=1}^{\mu_\ell}\sqrt{-\lambda_m}\langle\psi_m(\xi_1), u(\xi_1,0)\rangle\psi_m(\xi_1) \\ &= -i\sum_{m=1}^{\mu_\ell} 2\sqrt{-\lambda_m}\psi_m(\delta_s)\psi_m(\xi_1). \end{aligned} \tag{11}$$

The boundary condition at $\xi_2 = 1$ is derived in a similar way. Due to the anechoical termination of the duct, there are no reflections, i.e., only rightgoing waves:

$$u(\xi_1,\xi_2) = \sum_{m=1}^{\mu_r} A_m\psi_m(\xi_1)\exp\left(i\sqrt{-\lambda_m}(\xi_2-1)\right), \tag{12}$$

where

$$\begin{aligned} \lambda_m &= \left((m-\tfrac{1}{2})\pi d_2/d_r\right)^2 - (\kappa d_2)^2, \\ \mu_r &= \left\lfloor \frac{\kappa d_r}{\pi} + \frac{1}{2} \right\rfloor. \end{aligned}$$

The coefficients A_m are determined by

$$A_m = \langle\psi_m(\xi_1), u(\xi_1,1)\rangle. \tag{13}$$

Differentiation of (12) and insertion of (13) gives the condition for the far-zone boundary:

$$\frac{\partial u}{\partial \xi_2}(\xi_1,1) - i\sum_{m=1}^{\mu_r}\sqrt{-\lambda_m}\langle\psi_m(\xi_1), u(\xi_1,1)\rangle\psi_m(\xi_1) = 0. \tag{14}$$

2.4. Discretization. Now when the analytical problem is defined, we design the numerical method. Introduce a uniform grid as

$$\begin{cases} \xi_{1,j} = jh_1, & j = 0,\dots,m_1+1, \\ \xi_{2,k} = \left(k-\tfrac{3}{2}\right)h_2, & k = 1,\dots,m_2, \end{cases}$$

where

$$h_1 = \frac{1}{m_1+\frac{1}{2}}, \qquad h_2 = \frac{1}{m_2-2}.$$

Let $u_{j,k}$ denote the approximate solution at the point $(\xi_{1,j}, \xi_{2,k})$. We use centered difference operators to obtain second-order accuracy. Equation (2) is approximated with

$$
\begin{aligned}
&- h_1^{-2}\big(a_{j+\frac{1}{2},k}(u_{j+1,k} - u_{j,k}) - a_{j-\frac{1}{2},k}(u_{j,k} - u_{j-1,k})\big) \\
&- h_2^{-2}\big(a^{-1}_{j,k+\frac{1}{2}}(u_{j,k+1} - u_{j,k}) - a^{-1}_{j,k-\frac{1}{2}}(u_{j,k} - u_{j,k-1})\big) - e_{j,k}u_{j,k} = 0
\end{aligned}
\tag{15}
$$

for inner points $k = 2, \ldots, m_2 - 1$ and $j = 1, \ldots, m_1$. The boundary conditions (3) and (4) become

$$u_{0,k} = 0, \quad k = 1, \ldots, m_2, \tag{16}$$

$$u_{m_1+1,k} = u_{m_1,k}, \quad k = 1, \ldots, m_2. \tag{17}$$

For the other two boundaries matters are more complicated. We discretize the modal expansions involving the eigenfunctions $\psi_m(\xi_1)$ by evaluating them on the ξ_1-grid, i.e.,

$$\psi_m \equiv \{\psi_m(\xi_{1,j})\}_{j=1}^{m_1} = \left\{\sqrt{2}\sin\left((m - \tfrac{1}{2})\pi j h_1\right)\right\}_{j=1}^{m_1}, \quad m = 1, \ldots, m_1, \tag{18}$$

and by approximating the integrals with a second-order accurate combination of the composite trapezoid rule and the rectangle rule. Our specific choice of ξ_1-grid and quadrature rule makes the column vectors (18) orthonormal with respect to the discrete scalar product

$$\langle \psi_m, \psi_n \rangle \equiv \psi_m^* h_1 \psi_n \approx \int_0^1 \bar{\psi}_m(\xi_1)\psi_n(\xi_1)d\xi_1.$$

Moreover, a second-order accurate finite-difference discretization of the eigenproblem yields *exactly* the same eigenvectors as (18). The resulting discretization of conditions (11) and (14) is

$$h_2^{-1}(u_1 - u_2) - i\sum_{m=1}^{\mu_\ell} \sqrt{-\lambda_m}\psi_m\psi_m^* h_1 \tfrac{1}{2}(u_1 + u_2) = -i\sum_{m=1}^{\mu_\ell} 2\sqrt{-\lambda_m}\psi_m(\delta_s)\psi_m, \tag{19}$$

$$h_2^{-1}(u_{m_2} - u_{m_2-1}) - i\sum_{m=1}^{\mu_r} \sqrt{-\lambda_m}\psi_m\psi_m^* h_1 \tfrac{1}{2}(u_{m_2-1} + u_{m_2}) = 0, \tag{20}$$

where

$$u_k \equiv (u_{1,k} \cdots u_{m_1,k})^T.$$

This can be written

$$(I_{m_1} - C_\ell)u_1 + (-I_{m_1} - C_\ell)u_2 = g_1,$$

$$(-I_{m_1} - C_r)u_{m_2-1} + (I_{m_1} - C_r)u_{m_2} = 0,$$

where

$$C_\ell = ih_2\sum_{m=1}^{\mu_\ell} \sqrt{-\lambda_m}\psi_m\psi_m^* \frac{h_1}{2}, \qquad C_r = ih_2\sum_{m=1}^{\mu_r} \sqrt{-\lambda_m}\psi_m\psi_m^* \frac{h_1}{2}, \tag{21}$$

$$g_1 = -ih_2 \sum_{m=1}^{\mu_\ell} 2\sqrt{-\lambda_m}\psi_m(\delta_s)\psi_m.$$

Notice that λ_m depends on the depth and is different for the left and right boundaries.

Applying (15) to inner grid points, using (19) and (20), and eliminating the boundary values defined by (16) and (17) gives the complete system of equations

$$Bu = g,$$

$$u \equiv \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{m_2} \end{pmatrix}, \qquad g = \begin{pmatrix} g_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad B = \mathrm{trid}_{k,m_2}(B_{k,-1}, B_{k,0}, B_{k,1}),$$

where

$$\begin{array}{c} B_{1,0} = I_{m_1} - C_\ell, \quad B_{1,1} = -I_{m_1} - C_\ell, \\ B_{m_2,-1} = -I_{m_1} - C_r, \quad B_{m_2,0} = I_{m_1} - C_r, \end{array} \tag{22}$$

and for $k = 2, \ldots, m_2 - 1$:

$$\begin{array}{rcl} B_{k,-1} & = & \mathrm{diag}_{j,m_1}(-\gamma a^{-1}_{j,k-\frac{1}{2}}), \\ B_{k,0} & = & \mathrm{trid}_{j,m_1}(-a_{j-\frac{1}{2},k}, \alpha_{j,k}, -a_{j+\frac{1}{2},k}), \\ B_{k,1} & = & \mathrm{diag}_{j,m_1}(-\gamma a^{-1}_{j,k+\frac{1}{2}}), \end{array}$$

where

$$\begin{array}{rcl} \alpha_{1,k} & = & a_{\frac{1}{2},k} + a_{\frac{3}{2},k} + \gamma(a^{-1}_{1,k-\frac{1}{2}} + a^{-1}_{1,k+\frac{1}{2}}) - h_1^2 e_{1,k}, \\ \alpha_{j,k} & = & a_{j-\frac{1}{2},k} + a_{j+\frac{1}{2},k} + \gamma(a^{-1}_{j,k-\frac{1}{2}} + a^{-1}_{j,k+\frac{1}{2}}) - h_1^2 e_{j,k}, \quad j = 2, \ldots, m_1 - 1, \\ \alpha_{m_1,k} & = & a_{m_1-\frac{1}{2},k} + \gamma(a^{-1}_{m_1,k-\frac{1}{2}} + a^{-1}_{m_1,k+\frac{1}{2}}) - h_1^2 e_{m_1,k}, \\ \gamma & = & \frac{h_1^2}{h_2^2}. \end{array}$$

By some minor modifications, the discretization could accommodate other combinations of boundary conditions on the boundaries Γ_1 and Γ_2. For Dirichlet conditions at Γ_1 and Γ_2, a suitable grid for ξ_1 would be

$$\xi_{1,j} = jh_1, \quad j = 0, \ldots, m_1 + 1, \quad h_1 = \frac{1}{m_1 + 1}.$$

The resulting alteration of the matrix B would solely be

$$\alpha_{m_1,k} = a_{m_1-\frac{1}{2},k} + a_{m_1+\frac{1}{2},k} + \gamma(a^{-1}_{m_1,k-\frac{1}{2}} + a^{-1}_{m_1,k+\frac{1}{2}}) - h_1^2 e_{m_1,k}.$$

For Neumann conditions at Γ_1 and Γ_2, a convenient choice of ξ_1-grid would be

$$\xi_{1,j} = (j - \tfrac{1}{2})h_1, \quad j = 0, \ldots, m_1 + 1, \quad h_1 = \frac{1}{m_1}.$$

This would cause $\alpha_{1,k}$ to change into

$$\alpha_{1,k} = a_{\frac{3}{2},k} + \gamma(a^{-1}_{1,k-\frac{1}{2}} + a^{-1}_{1,k+\frac{1}{2}}) - h_1^2 e_{1,k},$$

but leaving the rest of B intact.

3. Preconditioning. We employ a Krylov subspace method to solve the system of equations. For simplicity and robustness, we choose the restarted generalized minimal residual (GMRES(ℓ)) algorithm [SaadSch86], where ℓ is the restarting length. For the iterative method to be competitive, an effective preconditioner is needed. Otherwise the cost of computing the solution would be too high. After preconditioning, the original system $Bu = g$ is transformed into $M^{-1}Bu = M^{-1}g$. We construct a preconditioner that preserves the block structure of B, thus exploiting sparsity. Moreover, it should be possible to form and apply the preconditioner at low arithmetic costs. To meet these demands, we use a preconditioner [Otto96] based on fast trigonometric transforms [VLoan92], [BaSw91]. The main idea in the design is to approximate the matrix B with a preconditioner having the same block structure, and where all the blocks have the same prescribed eigenvectors. These eigenvectors depend on the boundary conditions, but are chosen so that the corresponding similarity transformation is associated with a fast transform.

For the discretization matrix B in §2.4, a Dirichlet condition was imposed on Γ_1 and a Neumann condition on Γ_2. Hence, a suitable choice for the unitary eigenvector matrix is

$$Q \equiv [q_1, \dots, q_{m_1}], \qquad q_m = \sqrt{h_1}\psi_m,$$

which is connected to a slightly modified [Otto96] sine transform-II [VLoan92]. Form a preconditioner

$$M = \mathrm{trid}_{k,m_2}(M_{k,-1}, M_{k,0}, M_{k,1}),$$

the blocks of which are diagonalized by Q, i.e.,

$$M_{k,r} \equiv Q\Lambda_{k,r}Q^*, \tag{23}$$

where $\Lambda_{k,r}$, $r = -1, 0, 1$, are diagonal matrices. There are several possible choices for $\Lambda_{k,r}$. The specific choice

$$\Lambda_{k,r} = \mathrm{diag}(Q^* B_{k,r} Q)$$

minimizes $\|B_{k,r} - M_{k,r}\|_F$ for matrices of type (23), and it also minimizes $\|B - M\|_F$. Observe that the blocks defined by (21) can be rewritten as linear combinations of outer products $q_m q_m^*$. This means that the matrix blocks (22) corresponding to the left and right boundaries will be diagonalized by Q. In fact, for a duct with a flat bottom, all the blocks in B would be diagonalized by Q, yielding $M = B$ [Otto96]. Hence, the operator M^{-1} is a direct fast Helmholtz solver for rectangular domains. For a duct with a curved bottom, blocks corresponding to inner grid lines will not be completely diagonalized. However, when the domain is moderately curved, the preconditioner presumably acts like a viable convergence accelerator.

For the Dirichlet–Dirichlet and Neumann–Neumann boundary conditions discussed in §2.4, the eigenvector matrices would rather be chosen as those associated with the sine and cosine transforms, respectively. The preconditioners thus arising would also yield direct fast solvers for rectangular domains, see [Otto96].

For each iteration, the computation $x = M^{-1}y$ has to be performed. Due to the structure of the blocks of M, it holds that

$$\Lambda \equiv (I_{m_2} \otimes Q^*)M(I_{m_2} \otimes Q) = \mathrm{trid}_{k,m_2}(\Lambda_{k,-1}, \Lambda_{k,0}, \Lambda_{k,1}),$$

leading to

$$M^{-1} = (I_{m_2} \otimes Q)\Lambda^{-1}(I_{m_2} \otimes Q^*).$$

The computation $x = M^{-1}y$ can now be done in three steps.

1. $v = (I_{m_2} \otimes Q^*)y$
2. solve $\Lambda z = v$
3. $x = (I_{m_2} \otimes Q)z$

Step 2 consists of solving a block tridiagonal system, where each block is diagonal. By permuting the unknowns, we get m_1 independent tridiagonal systems of order m_2. Steps 1 and 3 consist of m_2 sine transforms-II and inverse sine transforms-II of length m_1. We can utilize fast Fourier transform methods [BaSw91] for computing these transforms [Otto96].

4. Computational issues.

4.1. Resolution. Since the solutions to the Helmholtz equation are waves, it is evident that the grid size h must follow the wavenumber κ in order to achieve a given accuracy. A naïve approach would be to use a fixed number of grid points per wavelength, i.e., keeping κh constant. Bayliss et al. [BaGoTu85a] established that such a resolution criterion is insufficient. Instead they presented estimates predicting that the L_2 norm of the error behaves like $\mathcal{O}(\kappa^{p+1}h^p)$ for a pth-order finite element discretization. Similar estimates have been rigorously proved [IhlBa97] for a one-dimensional model problem with Dirichlet–Robin boundary conditions. The estimates are in accordance with results conjected from numerical experiments [ThoPin94]. The objective of this section is to specify convenient resolution criteria, for the finite-difference discretization in §2.4, resembling those in [BaGoTu85a].

The analysis is based on a one-dimensional counterpart of (2), i.e.,

$$-\frac{d^2v}{d\xi_2^2} - (\kappa d_2)^2 v = 0, \quad 0 < \xi_2 < 1, \tag{24}$$

with Robin boundary conditions

$$\begin{aligned} -\frac{dv}{d\xi_2}(0) - i\kappa d_2 v(0) &= -2i\kappa d_2 A, \\ \frac{dv}{d\xi_2}(1) - i\kappa d_2 v(1) &= 0 \end{aligned} \tag{25}$$

replacing (11) and (14). Note that for a one-dimensional problem, the Sommerfeld condition (25) is exact inasmuch as it allows only rightgoing waves. Applying the same discretization as in §2.4 to (24) results in

$$-(v_{k+1} - 2v_k + v_{k-1}) - (\kappa d_2)^2 h_2^2 v_k = 0, \qquad k = 2, \ldots, m_2 - 1, \tag{26}$$

$$-(v_2 - v_1) - i\kappa d_2 \frac{h_2}{2}(v_1 + v_2) = -2i\kappa d_2 h_2 A, \tag{27}$$

$$(v_{m_2} - v_{m_2-1}) - i\kappa d_2 \frac{h_2}{2}(v_{m_2-1} + v_{m_2}) = 0 \tag{28}$$

for the finite-difference approximation $v_k \approx v(\xi_{2,k})$. The difference equation (26) has the following characteristic equation

$$r^2 - (2 - (\kappa d_2)^2 h_2^2)r + 1 = 0$$

with roots denoted by r_1 and r_2. The root

$$r_1 = 1 - \frac{(\kappa d_2)^2 h_2^2}{2} + \sqrt{-(\kappa d_2)^2 h_2^2 + \frac{(\kappa d_2)^4 h_2^4}{4}}$$

corresponds to the rightgoing mode; whereas the remaining root

$$r_2 = 1 - \frac{(\kappa d_2)^2 h_2^2}{2} - \sqrt{-(\kappa d_2)^2 h_2^2 + \frac{(\kappa d_2)^4 h_2^4}{4}}$$

is associated with the leftgoing mode. Thus, the solution to (26) is

$$v_k = C_1 r_1^{k-\frac{3}{2}} + C_2 r_2^{k-\frac{3}{2}},$$

where the coefficients C_1 and C_2 are determined from (27) and (28), yielding

$$C_1 = \left(1 + \mathcal{O}((\kappa d_2)^2 h_2^2)\right) A, \qquad C_2 = \mathcal{O}((\kappa d_2)^2 h_2^2 A).$$

Combining this with a Taylor expansion of $r_1^{k-\frac{3}{2}}$ leads to

$$v_k = A \exp\left(i\kappa d_2 \xi_{2,k} + \frac{i(\kappa d_2)^3 h_2^2}{24}\xi_{2,k} + \mathcal{O}((\kappa d_2)^4 h_2^3)\right) + \mathcal{O}((\kappa d_2)^2 h_2^2 A).$$

Comparing this with the true solution to (24), i.e.,

$$v(\xi_{2,k}) = A \exp(i\kappa d_2 \xi_{2,k}),$$

we conclude that the leading phase error of magnitude $\frac{(\kappa d_2)^3 h_2^2}{24}\xi_{2,k}$ grows linearly in ξ_2. Furthermore, a reasonable resolution criterion is

$$\frac{(\kappa d_2)^3 h_2^2}{24} = \tau,$$

where τ is a given tolerance. Notice that for this resolution we obtain

$$v_k = A \exp\left(i\kappa d_2 \xi_{2,k} + i\tau\xi_{2,k} + \mathcal{O}(\tau^{\frac{3}{2}}(\kappa d_2)^{-\frac{1}{2}})\right) + \mathcal{O}(\tau(\kappa d_2)^{-1}A).$$

When κd_2 is sufficiently large and τ is less than one, the terms $\mathcal{O}(\tau(\kappa d_2)^{-1}A)$ and $\mathcal{O}(\tau^{\frac{3}{2}}(\kappa d_2)^{-\frac{1}{2}})$, representing artificial reflections and amplitude errors, are neglible compared with the phase error $\tau\xi_{2,k}$. Under these circumstances, the phase error is a measure of the pointwise relative error. Extensive numerical experiments, comparing the numerical solution v_k with the true solution $v(\xi_{2,k})$, corroborate that the phase error prediction above is sharp.

Thus, for the two-dimensional problem in §2.4, we are led to the following resolution in the ξ_2-direction:

$$h_2 \leftarrow \frac{(24\tau)^{\frac{1}{2}}}{(\kappa d_2)^{\frac{3}{2}}}, \qquad m_2 \leftarrow \left\lceil \frac{1}{h_2} + 2 \right\rceil. \tag{29}$$

For the ξ_1-direction, the choice of resolution is more subtle. Lacking a more sophisticated analysis, a rescaling of condition (29) is advocated:

$$\begin{aligned} d_1 &\leftarrow \max(d_\ell, d_r), \\ h_1 &\leftarrow \frac{(24\tau d_1/d_2)^{\frac{1}{2}}}{(\kappa d_1)^{\frac{3}{2}}}, \qquad m_1 \leftarrow \left\lceil \frac{1}{h_1} - \frac{1}{2} \right\rceil. \end{aligned} \tag{30}$$

4.2. Complexity. In this section we discuss the efficiency of our method regarding memory requirement and arithmetic complexity. Note that only the highest order terms will be considered, and that the number of arithmetic operations will be normalized by the number of unknowns $m_1 m_2$. A complex addition will be counted as two arithmetic operations, a complex multiplication as six arithmetic operations, and a complex division as eleven arithmetic operations.

In order to determine the arithmetic complexity, we must specify how the initial approximation and the stopping criterion are computed. As an initial approximation we use the preconditioned right-hand side $M^{-1}g$, which is advantageous if $M^{-1}B$ is close to the identity matrix. We have imposed the following stopping criterion

$$\frac{\|M^{-1}(g - Bu^{(i)})\|_2}{\|M^{-1}g\|_2} < \epsilon$$

with tolerance $\epsilon = 10^{-4}$.

The arithmetic work can be divided into initialization and iteration. The initial part consists of forming [Otto96] the preconditioner and factorizing the tridiagonal systems at a cost of $a_{pf} = 20\frac{\hat{m}}{m_1}\log_2 \hat{m} + 139$, where $\hat{m} = 2^{\lceil \log_2(2m_1+1)\rceil + 1}$. The computation of the initial approximation is done with a preconditioner solve that requires $a_{ps} = 20\frac{\hat{m}}{m_1}\log_2 \hat{m} + 117$ arithmetic operations per unknown. The iterative method also goes through some initial steps. The cost for these is $a_{in} = 2a_{ps} + a_m + 10$, where $a_m = 40$ is the work required for a matrix–vector product $y = Bx$. Accordingly, the total arithmetic cost for the initialization becomes $a_{init} = a_{pf} + a_{ps} + a_{in}$.

The work for one iteration of GMRES(ℓ) is taken as the average over a complete cycle of ℓ iterations, and is given by $a_{it} = a_m + a_{ps} + 8\ell + 44$.

If we let n_{it} be the number of iterations required for convergence, then the total work for solving $M^{-1}Bu = M^{-1}g$ with the GMRES(ℓ) method is $a_{init} + n_{it}a_{it}$.

The memory requirement for our method is $m_m + m_p + m_{it}$; where $m_m = 7m_1m_2$ is the number of memory positions needed for the coefficient matrix, the right-hand side, and the solution; $m_p = 8m_1m_2 + 4\hat{m}m_2$ is the number of memory positions used by the preconditioner; and $m_{it} = 2(\ell+1)m_1m_2$ denotes the storage requirement for the iterative method. Note that a complex value is considered to take up two memory positions.

In Table 1 our method is compared with band Gaussian elimination, which is the standard solution technique. The storage requirements have been normalized by the number of unknowns.

Table 1
Comparison of GMRES(ℓ) *and band Gaussian elimination.*

	arithmetic complexity	memory requirement
band GE	$8m_1^2 + 27m_1$	$4m_1 + 4$
GMRES(ℓ)	$80\frac{\hat{m}}{m_1}\log_2 \hat{m} + 540$ $+ n_{it}(20\frac{\hat{m}}{m_1}\log_2 \hat{m} + 8\ell + 201)$	$2\ell + 17 + 4\frac{\hat{m}}{m_1}$

5. Numerical experiments. In this section the results from some numerical experiments are presented. In all experiments, the systems of equations have been solved using the GMRES(ℓ) method combined with the preconditioner defined in §3. The orthogonal grid is generated by a code based on the method described in [Abra91]. The implementations are made in Fortran 90, utilizing 64 bit precision for the grid generation, and 32 bit precision for the iterative method and the preconditioner. The numerical experiments were performed on a DEC AlphaServer 8200 EV5/300.* The geometry of the duct, i.e., the bottom profile is defined by the following functions:

$$\begin{cases} x_1(\theta) = d_\ell + (d_r - d_\ell)\frac{\tanh(s(\theta-\delta_c))-\tanh(-s\delta_c)}{\tanh(s(1-\delta_c))-\tanh(-s\delta_c)} \\ x_2(\theta) = \theta d_2 \end{cases}, \quad 0 \le \theta \le 1,$$

where

$$\delta_c = 0.5, \qquad s = \frac{4}{\min(\delta_c, 1-\delta_c)}.$$

By this choice the depth varies smoothly from d_ℓ at the left boundary to d_r at the right boundary. The parameter δ_c determines the center of the slope, whereas s controls the steepness. By increasing s, the slope steepens and the bottom flattens out at the ends. The relative source depth δ_s is set to 0.5 in all experiments. We use resolution criteria (29) and (30) with a phase error tolerance $\tau = 8\%$.

It would be interesting to investigate the arithmetic speedup for the preconditioned GMRES(ℓ) method compared with plain GMRES(ℓ), but the latter does not converge in a reasonable number of iterations. However, the effectiveness of the preconditioner is indicated when comparing unpreconditioned and preconditioned spectra. The spectra for a small problem are shown in Figs. 2 and 3.

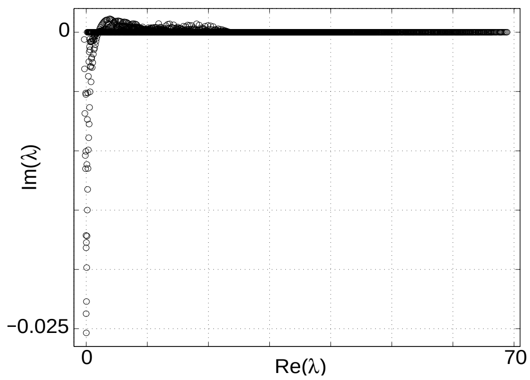

FIG. 2. *The spectrum of B for $d_2 = 300$, $d_\ell = 50$, $d_r = 20$, and $f = 25$.*

* The actual computer is part of the Yggdrasil computing facilities at the Dept. of Scientific Computing, Uppsala Univ.

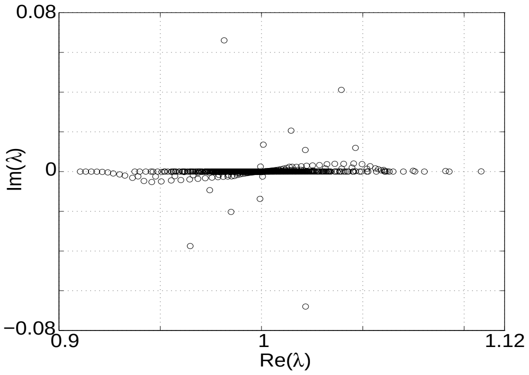

FIG. 3. *The spectrum of* $M^{-1}B$ *for* $d_2 = 300$, $d_\ell = 50$, $d_r = 20$, *and* $f = 25$.

The preconditioned spectrum exhibits a high degree of clustering around one, which is favorable for Krylov subspace methods [Axel94], [Axel88].

Since the preconditioner coincides with the discretization matrix for the model problem in a duct with a flat bottom, it is to be expected that the rate of convergence will be affected by the geometry. When the bottom of the duct gets more curved, the preconditioner is not as good an approximation of B. Figure 4 shows how the geometry influences the number of iterations for GMRES(6). Notice that here the number of iterations decreases when the problem size increases (and the duct gets less curved).

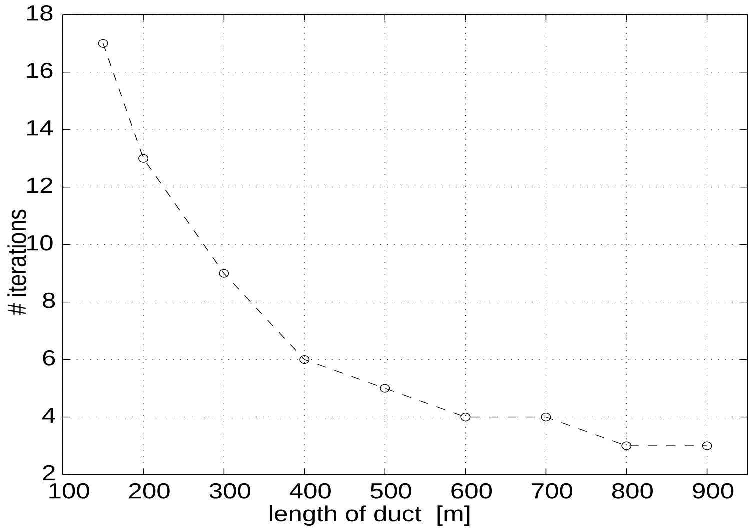

FIG. 4. *Number of iterations for ducts where the length* d_2 *is varied. All the other parameters are held constant,* $d_\ell = 50$, $d_r = 20$, *and* $f = 100$.

Another interesting issue is how the frequency affects the number of iterations. This is demonstrated in Fig. 5 for a duct of medium steepness and frequencies in the low–to–intermediate range.

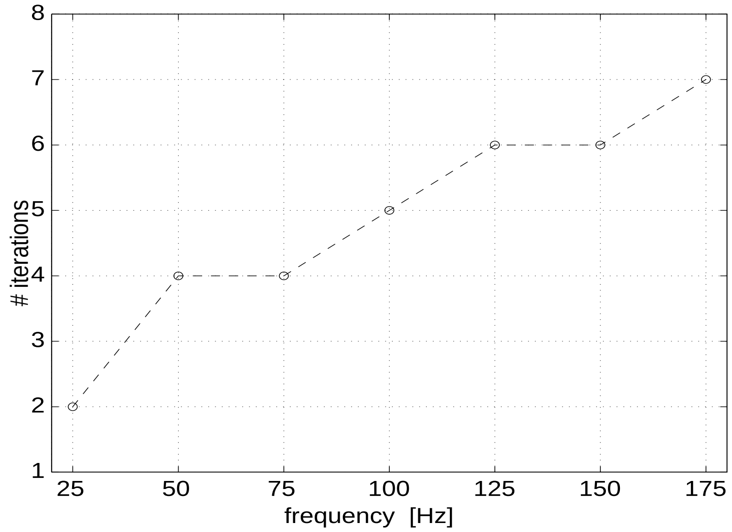

FIG. 5. *Number of iterations for different frequencies,* $d_2 = 500$, $d_\ell = 50$ *and* $d_r = 20$.

In Figs. 6 and 7, the results from comparative experiments are shown. The number of unknowns depends cubically on the frequency and ranges from 7452 to 2563902. It is clear that our method is more efficient than band Gaussian elimination both regarding arithmetic complexity and memory requirement for all problem sizes considered. Furthermore, the relative gain increases as the frequency increases.

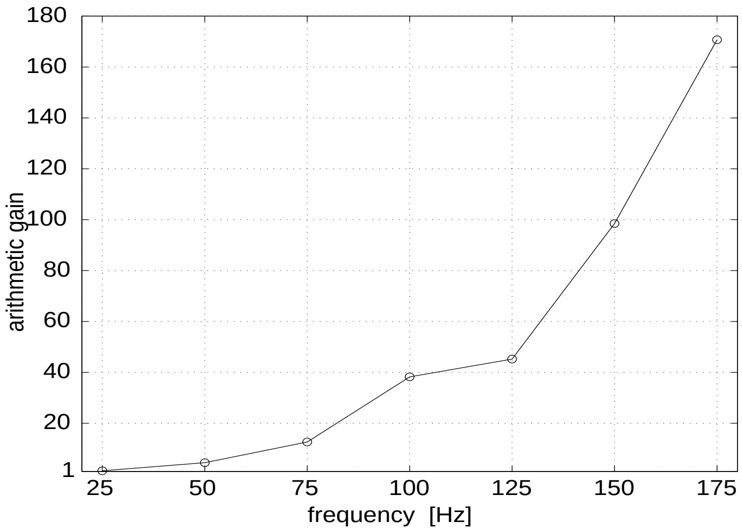

FIG. 6. *Arithmetic gain for* GMRES(6) *compared with band Gaussian elimination.*

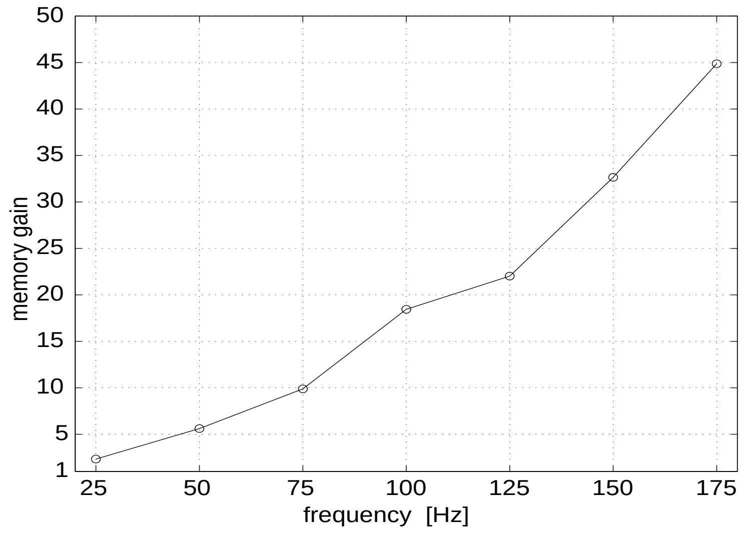

FIG. 7. *Memory gain for* GMRES(6) *compared with band Gaussian elimination.*

Finally, we display the solutions for two different frequencies. We have chosen rather low frequencies, because those solutions are easier to visualize.

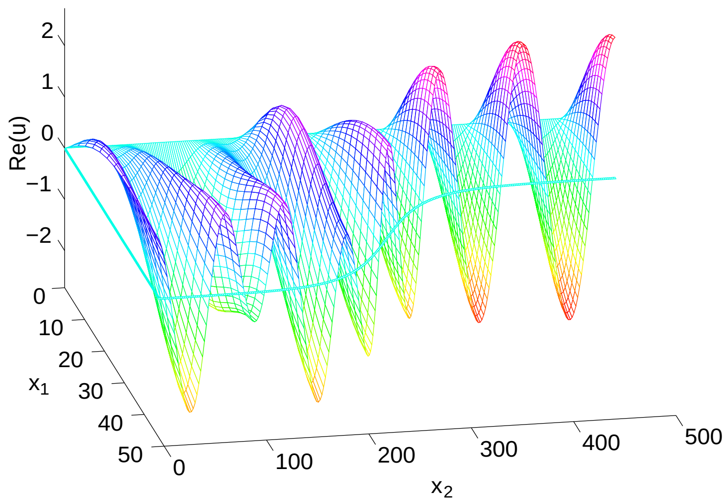

FIG. 8. *The solution for $f = 25$, $d_2 = 500$, $d_\ell = 50$ and $d_r = 20$. The contour of the duct is also depicted.*

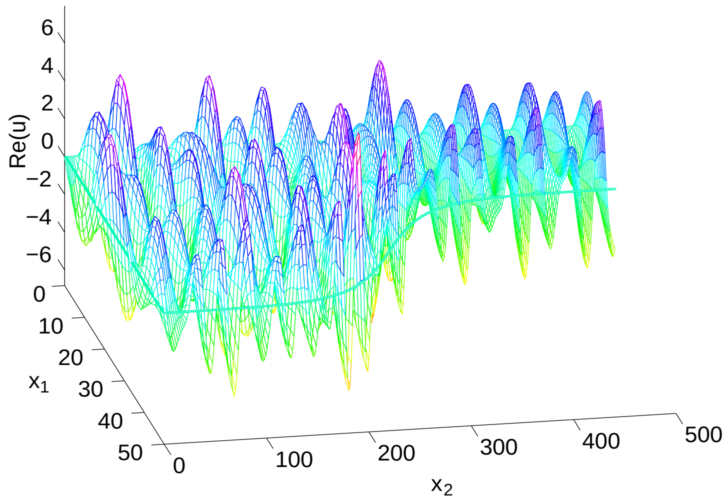

FIG. 9. *The solution for $f = 75$, $d_2 = 500$, $d_\ell = 50$ and $d_r = 20$. The contour of the duct is also depicted.*

Note that, for the lower frequency, only one wave mode is transmitted in the narrow part of the duct. For the higher frequency, several modes are transmitted and interfere.

6. Conclusions. We have applied a preconditioned GMRES(ℓ) algorithm to a second-order finite-difference discretization of the two-dimensional Helmholtz equation subject to Dirichlet, Neumann, and DtN boundary conditions. The preconditioner is based on fast transforms, and results in a direct fast Helmholtz solver for rectangular domains. The memory requirement for the preconditioned method is linear in the number of unknowns. Thus, the sparsity of the original discretization matrix is efficiently exploited. Numerical experiments, for a hydroacoustic wave propagation problem, show that the preconditioned iterative method yields a significant gain both in storage requirement and arithmetic complexity, when it is compared with band Gaussian elimination. Especially, the relative gain increases when the wavenumber is raised. Moreover, the number of iterations required for convergence grows moderately (or even decreases) as the number of unknowns increases.

In order to suppress the phase error, the number of unknowns has to grow cubically in the wavenumber due to the second-order accurate discretization. Thus, for high wavenumbers, the *discretization* is less tractable from a computational point of view. The memory requirement might be a bottle-neck. To mitigate this adverse effect, high-order discretizations will be investigated in a forthcoming paper. Another pertinent concern is to perform a more rigorous phase error analysis. Further directions of research will also entail applications to heterogeneous media, e.g., cases where the sound speed depends on the depth due to temperature gradients, changes in hydrostatic pressure, and variable salinity.

Acknowledgments. The authors would like to thank Dr. Leif Abrahamsson for supplying the grid generation code. The first author also expresses his gratitude to Prof. Howard Elman, who invited the author to a postdoctoral visit at the Dept. of Computer Science, Univ. of Maryland, where this research was completed.

REFERENCES

[Abra91] L. ABRAHAMSSON, *Orthogonal grid generation for two-dimensional ducts*, J. Comput. Appl. Math., 34 (1991), pp. 305–314.

[AbKr94] L. ABRAHAMSSON AND H.-O. KREISS, *Numerical solution of the coupled mode equations in duct acoustics*, J. Comput. Phys., 111 (1994), pp. 1–14.

[Axel88] O. AXELSSON, *A restarted version of a generalized preconditioned conjugate gradient method*, Comm. Appl. Numer. Methods, 4 (1988), pp. 521–530.

[Axel94] ———, *Iterative Solution Methods*, Cambridge University Press, New York, 1994.

[BaSw91] D. H. BAILEY AND P. N. SWARZTRAUBER, *The fractional Fourier transform and applications*, SIAM Rev., 33 (1991), pp. 389–404.

[BaGoTu83] A. BAYLISS, C. I. GOLDSTEIN, AND E. TURKEL, *An iterative method for the Helmholtz equation*, J. Comput. Phys., 49 (1983), pp. 443–457.

[BaGoTu85a] ———, *On accuracy conditions for the numerical computation of waves*, J. Comput. Phys., 59 (1985), pp. 396–404.

[BaGoTu85b] ———, *The numerical solution of the Helmholtz equation for wave propagation problems in underwater acoustics*, Comput. Math. Appl., 11 (1985), pp. 655–665.

[BaGuTu82] A. BAYLISS, M. GUNZBURGER, AND E. TURKEL, *Boundary conditions for the numerical solution of elliptic equations in exterior regions*, SIAM J. Appl. Math., 42 (1982), pp. 430–451.

[ChanNg96] R. H. CHAN AND M. K. NG, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.

[Ernst94] O. G. ERNST, *Fast Numerical Solution of Exterior Helmholtz Problems with Radiation Boundary Condition by Imbedding*, Ph.D. thesis, Dept. of Computer Science, Stanford Univ., Stanford, CA, 1994.

[FixMa78] G. J. FIX AND S. P. MARIN, *Variational methods for underwater acoustic problems*, J. Comput. Phys., 28 (1978), pp. 253–270.

[FrGoNa92] R. W. FREUND, G. H. GOLUB, AND N. M. NACHTIGAL, *Iterative solution of linear systems*, Acta Numerica, 1 (1992), pp. 57–100.

[Gold82] C. I. GOLDSTEIN, *A finite element method for solving Helmholtz type equations in waveguides and other unbounded domains*, Math. Comp., 39 (1982), pp. 309–324.

[GrKe95] M. J. GROTE AND J. B. KELLER, *On nonreflecting boundary conditions*, J. Comput. Phys., 122 (1995), pp. 231–243.

[IhlBa97] F. IHLENBURG AND I. BABUŠKA, *Finite element solution of the Helmholtz equation with high wave number. Part* II: *The h-p version of the FEM*, SIAM J. Numer. Anal., 34 (1997), to appear.

[KeGi89] J. B. KELLER AND D. GIVOLI, *Exact non-reflecting boundary conditions*, J. Comput. Phys., 82 (1989), pp. 172–192.

[Otto96] K. OTTO, *A unifying framework for preconditioners based on fast transforms*, Report No. 187, Dept. of Scientific Computing, Uppsala Univ., Uppsala, Sweden, 1996.

[SaadSch86] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[ThoPin94] L. L. THOMPSON AND P. M. PINSKY, *Complex wavenumber Fourier analysis of the p-version finite element method*, Comput. Mech., 13 (1994), pp. 255–275.

[VLoan92] C. F. VAN LOAN, *Computational Frameworks for the Fast Fourier Transform*, SIAM, Philadelphia, PA, 1992.