Journal of Computational Physics **182,** 27–46 (2002)
doi:10.1006/jcph.2002.7148

Computational Aspects of the Ultra-Weak Variational Formulation

Tomi Huttunen,* Peter Monk,† and Jari P. Kaipio*

**Department of Applied Physics, University of Kuopio, P.O. Box 1627, 70211 Kuopio, Finland; and †Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19711*
E-mail: jari.kaipio@uku.fi

Received December 18, 2001; revised July 7, 2002

The ultra-weak variational formulation (UWVF) approach has been proposed as an effective method for solving Helmholtz problems with high wave numbers. However, for coarse meshes the method can suffer from instability. In this paper we consider computational aspects of the ultra-weak variational formulation for the inhomogeneous Helmholtz problem. We introduce a method to improve the UWVF scheme and we compare iterative solvers for the resulting linear system. Computations for the acoustic transmission problem in 2D show that the new approach enables Helmholtz problems to be solved on a relatively coarse mesh for a wide range of wave numbers. © 2002 Elsevier Science (USA)

1. INTRODUCTION

The inhomogeneous Helmholtz equation arises in many physical problems. The modeling of time-harmonic acoustic and electromagnetic fields in heterogenous media are widely known examples. For long wavelengths these problems can be approximated using low-order finite element or finite difference methods. As the wavelength decreases these methods become increasingly expensive due to the requirement that there must be sufficiently many points per wavelength to obtain a reliable solution (ten grid points per wavelength is often mentioned as a rule of thumb). In addition, numerical pollution due to the accumulation of phase error forces the use of even more grid points per wavelength to maintain accuracy at a desired level [7]. In many applications this leads to intolerable computational complexity.

To avoid the problems associated with lower order finite elements, a variety of techniques have been proposed. Modifications of the basic finite element method include, for example, higher order methods [8], least-squares finite elements [6, 11, 15], and partition of unity methods (PUM) [1]. The PUM make it possible to include *a priori* information about the solution in the approximation subspace. Compared to standard finite elements this has been shown to give considerable reduction in computational complexity [9].

A common feature of finite element methods with special shape functions, such as PUM, is the need for numerical quadratures in the computation of the associated integrals. For bases that consist of oscillatory functions this requires higher order quadratures [9] or special integration techniques [12]. In addition, conditioning problems sometimes require a regularization-type approach to stabilize the problem [12].

Another approach is to approximate the global solution of the Helmholtz equation by a family of solutions of the Helmholtz equation in each element and enforce continuity as far as possible across element boundaries via the numerical scheme. One obvious method is to minimize the least-squares difference in the jumps of the solution and its normal derivative across element edges by minimizing a least-squares functional; see, e.g., [10, 14]. In [10] the method was analyzed using plane wave and Bessel function bases. Both bases provided efficient means to obtain good accuracy. However, the plane wave basis was recommended due to the simplicity of evaluating integrals. Conditioning problems were noted as the number of basis functions per element increased.

The ultra-weak variational formulation (UWVF) is another approach to using discontinuous local solutions of the Helmholtz equation on each element. In this approach, which was proposed and analyzed in [3–5], integration by parts is used to derive a variational formulation that weakly enforces appropriate continuity conditions between elements via impedance boundary conditions. As in the least-squares method a family of local solutions of the Helmholtz equation is used to construct the approximation space. However, unlike the least-squares method, the final equations satisfied by the discrete solution are given by the Galerkin procedure rather than by the more ad hoc least-squares approach. However, on the theoretical level the least-squares method is better understood than the UWVF in that global convergence can be proved.

In principle there are many possible choices for the local approximation functions on each element in the UWVF. However, only plane waves have been used so far, and based on the theoretical studies in [10], it seems unlikely that Bessel function bases would offer much improvement when approximating smooth solutions. Hence, as discussed further in Section 3, we use plane waves in this paper.

An advantage of the use of the plane waves is that in most cases integrals occuring in the resulting matrix system can be evaluated in closed form. As a drawback, ill-conditioning of the problem has been reported when fine meshes or large dimensional bases are used [3]. However, it is shown in numerical examples that the method can produce accurate results when the element size is twice the wavelength.

In this paper we investigate the UWVF from the computational point of view. We show that the conditioning problem is particularly severe when the UWVF is applied to inhomogeneous problems or when unstructured meshes with varying element sizes are used. We propose the use of a basis with a nonuniform number of basis functions per element as a feasible method for improving the conditioning of the UWVF. Numerical examples show that the proper choice of basis enables us to use very large geometric elements, sometimes five times the size of the wavelength. In addition, we compare the Richardson and stabilized Bi-Conjugate Gradient iterative methods for solving the resulting linear system.

The paper is organized as follows. In Section 2 we give a short review of the ultra-weak variational formulation. In Section 3 we summarize the discrete form of the method. The computational scheme for choosing the bases and solving the linear system is described in Section 4. In Section 5 we give numerical examples of the method applied to a high-frequency acoustic transmission problem.

Now let us describe the problem considered in this paper. Let Ω be a domain in $\mathbb{R}^2$ with the smooth boundary Γ and outward unit normal ν. The inhomogeneous Helmholtz problem for the field u is defined as

$$\nabla \cdot \left(\frac{1}{\rho}\nabla u\right) + \frac{\kappa^2}{\rho} u = 0 \quad \text{in } \Omega, \tag{1}$$

$$\left(\frac{1}{\rho}\frac{\partial u}{\partial \nu} - i\sigma u\right) = Q\left(-\frac{1}{\rho}\frac{\partial u}{\partial \nu} - i\sigma u\right) + g \quad \text{on } \Gamma, \tag{2}$$

where $\kappa = \kappa(x_1, x_2) \in \mathbb{C}$ with $\operatorname{Im}(\kappa) \geq 0$, $\operatorname{Re}(\kappa) > 0$ is the wave number, and $|Q| \leq 1$, $Q \in \mathbb{C}$. The parameters $\rho = \rho(x_1, x_2)$ and σ are real and positive. The source term on Γ is denoted by g.

2. THE ULTRA-WEAK VARIATIONAL FORMULATION OF THE HELMHOLTZ EQUATION

In this section we outline the UWVF for the inhomogeneous Helmholtz problem (1)–(2). Let us partition the domain Ω into a collection of disjoint finite elements $\{\Omega_k\}_{k=1}^{K}$. In this report each element Ω_k is a triangle except near curved interfaces or boundaries where Ω_k can be a curvilinear triangle. In principle it is possible to mix triangles and quadrilaterals in the same mesh, but we have not studied this idea here. Let Σ_{kj} denote the edge between element Ω_k and element Ω_j, and let ν_k denote the outward unit normal on $\partial\Omega_k$. The exterior edges are denoted by Γ_k; see Fig. 1.

The coefficients ρ and κ are assumed to be piecewise constants, so that $\rho_k \equiv \rho|_{\Omega_k}$ and $\kappa_k \equiv \kappa|_{\Omega_k}$. Problem (1) and (2) can now be decomposed into subproblems for each element Ω_k, $k = 1, \ldots, K$,

$$\Delta u_k + \kappa_k^2 u_k = 0 \quad \text{in } \Omega_k \tag{3}$$

$$u_k = u_j \quad \text{on } \Sigma_{kj} \tag{4}$$

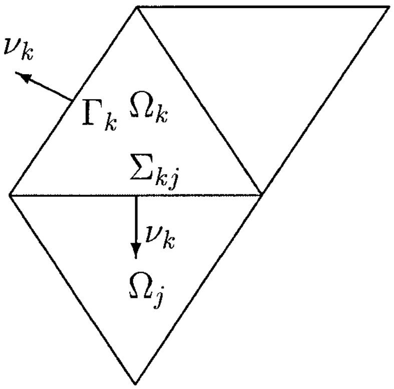


FIG. 1. A part of the mesh. The interface between elements Ω_k and Ω_j is Σ_{kj}. The outward unit normal on the boundary $\partial\Omega_k$ is ν_k. Furthermore, if the element is on the exterior boundary, the corresponding part of $\partial\Omega_k$ is denoted by Γ_k.

$$\frac{1}{\rho_k}\frac{\partial u_k}{\partial \nu_k} = -\frac{1}{\rho_j}\frac{\partial u_j}{\partial \nu_j} \quad \text{on } \Sigma_{kj} \tag{5}$$

$$\left(\frac{1}{\rho_k}\frac{\partial u_k}{\partial \nu_k} - i\sigma_k u_k\right) = Q\left(-\frac{1}{\rho_k}\frac{\partial u_k}{\partial \nu_k} - i\sigma_k u_k\right) + g \quad \text{on } \Gamma_k, \tag{6}$$

where $u_k = u|_{\Omega_k}$. The transmission conditions (4) and (5) on the interface Σ_{kj} can be written in the coupled form [2]

$$\frac{1}{\rho_k}\frac{\partial u_k}{\partial \nu_k} - i\sigma u_k = -\frac{1}{\rho_j}\frac{\partial u_j}{\partial \nu_j} - i\sigma u_j, \quad \text{and} \quad \frac{1}{\rho_k}\frac{\partial u_k}{\partial \nu_k} + i\sigma u_k = -\frac{1}{\rho_j}\frac{\partial u_j}{\partial \nu_j} + i\sigma u_j, \tag{7}$$

where σ is an appropriate real-valued parameter that is defined on the element boundary $\partial\Omega_k$. Since σ must have the same dimensions as κ/ρ [2] we have used

$$\sigma = \frac{1}{2}\left(\frac{\operatorname{Re}(\kappa_k)}{\rho_k} + \frac{\operatorname{Re}(\kappa_j)}{\rho_j}\right) \quad \text{on } \Sigma_{kj}, \tag{8}$$

which is the mean value of $\operatorname{Re}(\kappa)/\rho$ on the interface Σ_{kj}. On the exterior boundary Γ the choice of the parameter σ depends on the boundary condition.

Let us now define a new function

$$\chi_k = \left(\left(-\frac{1}{\rho_k}\frac{\partial}{\partial \nu_k} - i\sigma\right)u_k\right)\Bigg|_{\partial\Omega_k}, \quad 1 \le k \le K. \tag{9}$$

From (3), (6), and (7) and integration by parts it follows that χ_k satisfies [3–5]

$$\begin{aligned}
&\sum_{k=1}^{K}\int_{\partial\Omega_k}\frac{1}{\sigma}\chi_k\overline{\left(-\frac{1}{\rho_k}\frac{\partial}{\partial \nu_k} - i\sigma\right)v_k} - \sum_{k=1}^{K}\sum_{j=1}^{K}\int_{\Sigma_{kj}}\frac{1}{\sigma}\chi_j\overline{\left(\frac{1}{\rho_k}\frac{\partial}{\partial \nu_k} - i\sigma\right)v_k}\\
&\quad + \sum_{k=1}^{K}\int_{\Gamma_k}\frac{Q}{\sigma}\chi_k\overline{\left(\frac{1}{\rho_k}\frac{\partial}{\partial \nu_k} - i\sigma\right)v_k}\\
&= \sum_{k=1}^{K}\int_{\Gamma_k}\frac{1}{\sigma}g\overline{\left(\frac{1}{\rho_k}\frac{\partial}{\partial \nu_k} - i\sigma\right)v_k}
\end{aligned} \tag{10}$$

for all piecewise smooth test functions v_k that are solutions of the adjoint Helmholtz equation

$$\Delta\bar{v}_k + \kappa_k^2\bar{v}_k = 0, \quad \text{in } \Omega_k, \tag{11}$$

where the bars stand for complex conjugate.

We now rewrite (10) to facilitate our discussion of the discrete problem. Let us define an operator

$$F_k : L^2(\partial\Omega_k) \to L^2(\partial\Omega_k) \tag{12}$$

such that if $y_k \in L^2(\partial\Omega_k)$ then $F_k(y_k) \in L^2(\partial\Omega_k)$ is given by

$$F_k(y_k) = \left(\frac{1}{\rho_k}\frac{\partial}{\partial \nu_k} - i\sigma\right)v_k \quad \text{on } \partial\Omega_k, \tag{13}$$

where $v_k \in H^1(\Omega_k)$ satisfies (11) and

$$\left(-\frac{1}{\rho_k}\frac{\partial}{\partial \nu_k} - i\sigma\right)v_k = y_k \quad \text{on } \partial\Omega_k. \tag{14}$$

Using F_k we see that (10) may be rewritten as the problem of finding $\chi_k \in L^2(\partial\Omega_k)$, $k = 1, 2, \ldots, K$ such that

$$\sum_{k=1}^{K}\int_{\partial\Omega_k}\frac{1}{\sigma}\chi_k\bar{y}_k - \sum_{k=1}^{K}\sum_{j=1}^{k}\int_{\Sigma_{kj}}\frac{1}{\sigma}\chi_j\overline{F_k(y_k)} + \sum_{k=1}^{K}\int_{\Gamma_k}\frac{Q}{\sigma}\chi_k\overline{F_k(y_k)} = \sum_{k=1}^{K}\int_{\Gamma_k}\frac{1}{\sigma}g\overline{F_k(y_k)} \tag{15}$$

for all $y_k \in L^2(\partial\Omega_k)$, $k = 1, 2, \ldots K$. Equation (15) is called the ultra-weak variational formulation of the inhomogeneous Helmholtz problem (1) and (2). This formulation makes clear that the unknown functions χ_k are computed on $\partial\Omega_k$ using as test functions y_k that are also functions on $\partial\Omega_k$. Thus the UWVF generates a direct approximation to the field u and $\partial u/\partial\nu_k$ on the skeleton of the mesh $\partial\Omega_k$, $k = 1, \ldots, K$. To compute u away from the skeleton involves a local postprocessing step. In the discrete case this will be discussed in the following section. Note that a knowledge of F_k, $k = 1, \ldots, K$ is required.

3. THE DISCRETE PROBLEM

Following [3, 4], we use a Galerkin approach to the discretization of the UWVF (15). We need to discretize the spaces $L^2(\partial\Omega_k)$, $k = 1, \ldots, K$ for functions appearing in (15). In principle any choice of complete family in $L^2(\partial\Omega_k)$ could work (for example, piecewise constant finite elements on $\partial\Omega_k$). However, to implement (10) we must be able to compute $F_k(y_k^a)$ for discrete functions y_k^a. This would be difficult using a piecewise constant basis. With this in mind Cessenat and Despres [3, 4] suggest the following strategy for constructing a discretization of $L^2(\partial\Omega_k)$ that makes the computation of F_k trivial. For each Ω_k a finite family of functions $\varphi_{k,\ell}$, $\ell = 1, \ldots, p_k$ is chosen which satisfy Eq. (11) so

$$\Delta\bar{\varphi}_{k,\ell} + \kappa_k^2\bar{\varphi}_{k,\ell} = 0 \quad \text{on } \Omega_k \tag{16}$$

and $\varphi_{k,\ell} = 0$ on $\Omega\backslash\bar{\Omega}_k$. Then the discrete space approximating $L^2(\partial\Omega_k)$ consists of all functions y_k^a such that

$$y_k^a = \sum_{\ell=1}^{p_k} y_{k,\ell}\left(-\frac{1}{\rho_k}\frac{\partial}{\partial\nu_k} - i\sigma\right)\varphi_{k,\ell} \quad k = 1, 2, \ldots, K, \tag{17}$$

where $\{y_{k,\ell}\}_{\ell=1}^{p_k}$ are arbitrary constants. Similarly,

$$\chi_k^a = \sum_{\ell=1}^{p_k} \chi_{k,\ell}\left(-\frac{1}{\rho_k}\frac{\partial}{\partial\nu_k} - i\sigma\right)\varphi_{k,\ell}, \tag{18}$$

where the expansion coefficients $\{\chi_{k,\ell}\}_{\ell=1}^{p_k}$ are the unknown functions we wish to compute.

Of course $F_k(y_k^a)$ is easy to compute since

$$F_k\left(y_k^a\right) = \sum_{\ell=1}^{p_k} y_{k,\ell}\left(\frac{1}{\rho_k}\frac{\partial}{\partial \nu_k} - i\sigma\right)\varphi_{k,\ell}. \tag{19}$$

The discrete UWVF is then obtained by replacing χ_k by χ_k^a, and y_k by y_k^a, in (15).

There are still a number of possible choices for the functions $\varphi_{k,\ell}$. Obviously we want $\{\varphi_{k,\ell}\}_{\ell=1}^{\infty}$ to be a complete family of solutions in the sense that any function in $L^2(\Omega)$ can be approximated to any desired accuracy by a function of the form (17) provided p_k is chosen large enough. For example, we could choose

$$\varphi_{k,\ell}(x) = J_{\ell-1}(\bar{\kappa}_k|x - x_k|)e^{i(\ell-1)\theta}, \quad 1 \le \ell \le p_k, \tag{20}$$

where $x \in \mathbb{R}^2$, $x_k \in \Omega_k$, and $J_{\ell-1}$ is the Bessel function of first kind and order $\ell - 1$. For the least-squares problem this basis did not offer a significant advantage over the basis we chose next [10]. In addition the integrals in (15) must be computed by quadrature.

Thus we turn to the choice advocated by Cessenat and Despres [3, 4] of the plane wave basis given by

$$\varphi_{k,\ell} = \begin{cases} \exp(i\bar{\kappa}_k d_{k,\ell}\cdot x) & \text{in } \Omega_k \\ 0 & \text{elsewhere,} \end{cases}$$

where $d_{k,\ell}$ is a unit vector giving the direction of propagation of the wave. The wave plane basis for the element Ω_k can be constructed using angularly equispaced directions

$$d_{k,\ell} = \left(\cos\left(2\pi\frac{\ell-1}{p_k}\right), \sin\left(2\pi\frac{\ell-1}{p_k}\right)\right). \tag{21}$$

The choice of equally spaced directions is not required by the UWVF. It is possible that another choice of directions would reduce the number of required directions if some *a priori* information about the solution is available, but this topic is not studied here. Instead, we focus on allowing the number of directions p_k to vary between elements.

In the Galerkin approach the test function $v_{k,\ell}$ is chosen from the basis functions so that successively $v_{k,\ell} = \varphi_{k,\ell}$, $1 \le \ell \le p_k$ and $1 \le k \le K$. Then the discrete form of the UWVF can be written as the matrix equation [3]

$$(D - C)X = b, \tag{22}$$

where $X = (\chi_{1,1}, \ldots, \chi_{1p_1}, \chi_{2,1}, \ldots)^T$. The entries in the Hermitian block diagonal matrix D are

$$D_k^{\ell,m} = \int_{\partial\Omega_k} \frac{1}{\sigma}\left(-\frac{1}{\rho_k}\frac{\partial\varphi_{k,m}}{\partial\nu_k} - i\sigma\varphi_{k,m}\right)\overline{\left(-\frac{1}{\rho_k}\frac{\partial\varphi_{k,\ell}}{\partial\nu_k} - i\sigma\varphi_{k,\ell}\right)}, \tag{23}$$

where the subscript of D refers to the block and the superscript shows the element in the block. The matrix C is also sparse and has a block structure. The entries in C are given by

$$\begin{aligned} C_{k,j}^{\ell,m} &= \int_{\Sigma_{kj}} \frac{1}{\sigma}\left(\frac{1}{\rho_j}\frac{\partial \varphi_{j,m}}{\partial \nu_k} - i\sigma\varphi_{j,m}\right)\overline{\left(\frac{1}{\rho_k}\frac{\partial \varphi_{k,\ell}}{\partial \nu_k} - i\sigma\varphi_{k,\ell}\right)} \\ &\quad + \int_{\Gamma_k} \frac{Q}{\sigma}\left(-\frac{1}{\rho_k}\frac{\partial \varphi_{k,m}}{\partial \nu_k} - i\sigma\varphi_{k,m}\right)\overline{\left(\frac{1}{\rho_k}\frac{\partial \varphi_{k,\ell}}{\partial \nu_k} - i\sigma\varphi_{k,\ell}\right)}. \end{aligned} \tag{24}$$

The entries for the right-hand side of the system are

$$b_{k,\ell} = \int_{\Gamma_k} \frac{1}{\sigma} g \overline{\left(\frac{1}{\rho_k}\frac{\partial \varphi_{k,\ell}}{\partial \nu_k} - i\sigma\varphi_{k,\ell}\right)}. \tag{25}$$

If the edges of the elements are straight the integrals above can be evaluated in closed form. For details see [3, 4]. On curved element edges the integrals must be computed numerically. We note that it is vital to use curved elements because large errors can occur from approximating curved boundaries by multiwavelength-sized elements (this is another way in which our implementation differs from the original implementation in [3]).

For numerical stability it is suggested [4] that Eq. (22) be solved in the form

$$(I - D^{-1}C)X = D^{-1}b. \tag{26}$$

This preconditioned approach requires inversion of the matrix D. Due to the block diagonal strucure of D the inversion can be done element-wise for each D_k separately. Using knowledge of the conditioning of the blocks we can improve stability of the resulting matrix system (26). This is discussed in the next section.

Provided that κ_k is real in Ω_k, the solution of the problem (1)–(2) can be approximated by

$$u^a|_{\Omega_k} = \sum_{\ell=1}^{p_k} \chi_{k,\ell}\varphi_{k,\ell}. \tag{27}$$

This is a direct consequence of Eqs. (3), (9), (11), and (18), together with the uniqueness of the solution of the Helmholtz equation. On elements where κ_k is not real, a further local problem must be solved element by element.

4. COMPUTATIONAL PROCEDURE

The solution of the problem can be carried out in three steps. First, the matrix D is computed. Although it is possible to fix the number of functions in the basis on each element beforehand, we allow changes in the number of basis functions per element during the building of the matrix. Hence, we reduce the severity of the stability problems that were reported in [3]. When the matrix D is computed, it is Cholesky factorized for later use in solving the matrix equation (26).

In the second step, after the number of functions in the basis on each on each element is chosen, the matrix C can be computed. In the third step, the matrix system (26) is solved using an appropriate direct or iterative matrix solver. The last step is the most time consuming. Therefore, the choice of the solver is an important issue.

4.1. Invertibility of the Matrix D

The block diagonal structure of the matrix D allows the separate factorization of the blocks D_k. The conditioning of the matrix block D_k depends on a variety of factors, such as the element size h_k and the number of bases p_k in that element. It was shown in [3] that the condition number of D_k for $p_k \geq 4$ is bounded from below by $Ch_k^{-2\text{int}(p_k/2)+2}$, where C is a positive constant and int(a) refers to the integer part of a. Numerical simulations show that the conditioning of the matrix block D_k also depends on the wave number κ_k; see Section 5.1.

Obviously, we can control the condition number of the matrix D by controlling the element size h_k and the number of bases p_k. These parameters should be chosen so that stable inversion for all blocks D_k is possible. Although we could vary the element size during mesh generation, we focus on controlling the number of based for a fixed mesh. This approach is justified because in many applications it is desirable to solve the problem with many wave numbers using the same mesh. On the other hand, we know from the least-squares method that the easiest way to improve the accuracy is to use more basis functions rather than a finer mesh, provided the solution to be computed is smooth [10]. Motivated by those considerations, we also investigate what size elements are allowed in the UWVF to obtain a tolerable accuracy. The main difficulty in the use of large elements is the need for a high-dimensional local basis which causes ill-conditioning of the blocks D_k for other elements if used uniformly regardless of element size.

4.2. Choosing the Basis

The simplest possibility is to use a fixed number of basis functions (i.e., a fixed number of directions for the plane waves) in all elements. However, due to the variability in the wave number and element size within the computation domain, this may result in severely ill-conditioned blocks leading to instability of the solution. In this paper we propose a scheme in which the number of bases is chosen dynamically during computation of the matrix D.

An appropriate criterion to characterize the stability of the inversion is the L_1-condition number

$$\text{Cond}(D_k) = \|D_k\|_1 \left\|D_k^{-1}\right\|_1. \tag{28}$$

The method we use is based on the sequential computation of the blocks D_k and estimation of the condition number. We start by setting the highest allowed value for the condition number and fixing the initial number of functions in the basis on each element. Then, we proceed element by element, building the block and estimating the condition number for the current basis. Depending on the condition number, we can reduce or increase the number of functions in the local basis, recompute the block, and reestimate the condition number. When the appropriate number of functions in the basis for the element is found, the Cholesky-factorized block is saved and the same procedure is repeated for the next element. As the outcome, we get the Cholesky-factorized matrix D and the number of basis functions for each element.

Various criteria can be used to choose the admissible number of bases. For example, one can choose the highest dimensional basis for which the condition number is below a predetermined limit. Alternatively, an initial guess can be a relatively high dimensional uniform basis which is known to generate ill-conditioned blocks. The dimension can be reduced only for the elements with the worst conditioning. Computation time is naturally dependent on the method and the initial guess for the basis. However, the basis is independent of the boundary data, and therefore for a single frequency and mesh it must be computed only once.

4.3. Iterative Algorithms

Problem (26) can be solved using a variety of techniques. Due to the large size of the problem an interative solver is preferred. In [3] the Richardson algorithm was used. Algorithm 1 shows a pseudo code for the method.

```
ALGORITHM 1.
Set $\epsilon > 0$
$\beta_0 = \text{rand}(0.5; 1 - \epsilon)$
$\hat{b} = D^{-1}b$
$X_0 = \beta_0\hat{b}$
for $i = 1, 2, 3, \ldots$
    $X_i = \beta_{i-1}\hat{b} + [(1 - \beta_{i-1})I + \beta_{i-1}D^{-1}C]X_{i-1}$
    if $X_i$ accurate enough; quit;
    $\beta_i = \text{rand}(0.5; 1 - \epsilon)$
end.
```

By rand(a, b) we denote a uniformly distributed random number between a and b. The behavior of the method for the UWVF problem is analyzed in detail in [3, 4].

In this paper we compare the Richardson scheme with another iterative solver, namely the stabilized Bi-Conjugate Gradient (Bi-CGStab) [16]. This variant of the conjugate gradient has been shown to be an efficient and smoothly convergent method for solving high-dimensional linear systems, see, e.g., [13, 16, 17]. It is applicable to nonHermitian matrices as encountered here (although the reduction of the residual will not necessarily be monotone). Algorithm 2 describes steps in the Bi-CGStab for the system (26).

```
ALGORITHM 2.
$X_0$ is an initial guess; $r_0 = D^{-1}b - (I - D^{-1}C)X_0$
$\hat{r}_0 = r_0$;
$\rho_0 = \alpha = \omega_0 = 1$;
$v_0 = p_0 = 0$;
for $i = 1, 2, 3, \ldots$
    $\rho_i = (\hat{r}_0, r_{i-1})$; $\beta = (\rho_i/\rho_{i-1})(\alpha/\omega_{i-1})$;
    $p_i = r_{i-1} + \beta(p_{i-1} - \omega_{i-1}v_{i-1})$;
    $v_i = (I - D^{-1}C)p_i$;
    $\alpha = \rho_i/(\hat{r}_0, v_i)$;
    $s = r_{i-1} - \alpha v_i$;
    $t = (I - D^{-1}C)s$;
    $\omega_i = (t, s)/(t, t)$;
    $X_i = X_{i-1} + \alpha p_i + \omega_i s$;
    $r_i = s - \omega_i t$;
    if $X_i$ accurate enough; quit;
end.
```

From the computational point of view it must be noted that the Bi-CGStab requires two multiplications by $I - D^{-1}C$ and four vector inner products on each iteration. Only one corresponding matrix–vector multiplication is needed in the Richardson algorithm.

5. NUMERICAL EXAMPLES

As an example of the inhomogeneous Helmholtz problem we consider acoustic scattering from the obstacle Ω_1 with different properties than those in the surrounding medium Ω_2. The concentric circular domains Ω_1 and Ω_2 have radii $a = 5.0$ cm and $R = 10$ cm, respectively (Fig. 2). The aim is to assess the behavior of the error of the UWVF approximation rather than to emulate any particular physical problem. The simple geometry allows us to compute an accurate approximation for the problem (1) and (2) using truncated Fourier series.

The acoustic pressure in $\Omega_1 \cup \Omega_2$ is now denoted by u. Let the speed of sounds be $c_1 = 3000$ m/s and $c_2 = 1500$ m/s in Ω_1 and in Ω_2, respectively. The corresponding densities are $\rho_1 = 2000$ kg/m^3 and $\rho_2 = 1000$ kg/m^3. The wave numbers are now $\kappa_1 = 2\pi f/c_1$ and $\kappa_2 = 2\pi f/c_2$, where f is the frequency of the sound field. The values for the physical parameters c and ρ are typical for biological tissues.

On the exterior boundary Γ we have

$$\frac{1}{\rho_2}\frac{\partial u_2}{\partial \nu} - i\kappa_2 u_2 = \frac{1}{\rho_2}\frac{\partial u^{\text{in}}}{\partial \nu} - i\kappa_2 u^{\text{in}}, \tag{29}$$

where u^{in} is the incident wave. The boundary condition (29) is obtained from the general form (2) by choosing $Q = 0$, $\sigma = \kappa_2$, and

$$g = \frac{1}{\rho_2}\frac{\partial u^{\text{in}}}{\partial \nu} - i\kappa_2 u^{\text{in}} \quad \text{on } \Gamma. \tag{30}$$

As the incident field we use a point source

$$u^{\text{in}} = \frac{i}{4} H_0^{(1)}(\kappa_2 |x - x_0|) \tag{31}$$

located at x_0 which is 1.0 cm outside the exterior boundary. $H_0^{(1)}$ is the zero-order Hankel function of the first kind. In many physical problems the sound source can be constructed from a combination of point sources.

The boundary condition (29) is the lowest order absorbing boundary condition for the scattered part of the pressure field u. Although there are more accurate absorbing boundary

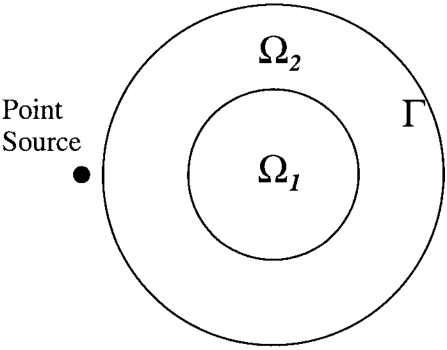


FIG. 2. The geometry of the model problem.

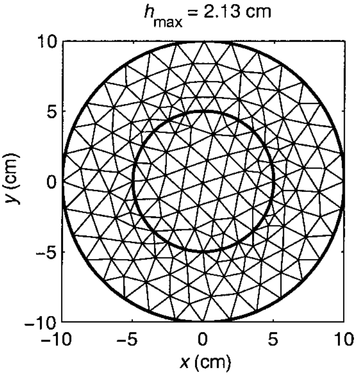


FIG. 3. One of the meshes used in the computations, consisting of 334 elements and 184 vertices.

conditions, this one is chosen because we can derive the exact solution for the problem which enables comparison with numerical results. Our error analysis compares the UWVF approximation to the exact series solution of (1)–(2) with the above choice of data.

To increase accuracy of the solution we allow curved boundaries on the boundaries $\Omega_1 \cap \Omega_2$ and Γ. The integrals (23)–(25) on those boundaries were computed with the 21-point Gauss–Legendre quadrature.

We used three meshes with $h_{\max} = 2.13$ cm, $h_{\max} = 1.21$ cm, and $h_{\max} = 0.68$ cm, where $h_{\max}$ is the maximum length of the element edges in the mesh. The coarsest mesh is shown in Fig. 3. We consider ultrasound frequencies spanning 100 to 500 kHz. Then, wavelengths $\lambda = 2\pi/\kappa$ vary between 0.6 and 3.0 cm in the domain Ω_1 and between 0.3 and 1.5 cm in the domain Ω_2.

5.1. Results for a Fixed Mesh

We start our study of the behavior of the error, the condition number, and the performance of the iterative solvers using the coarsest mesh with $h_{\max} = 2.13$ cm. Due to the mesh density requirements, this mesh is useful for the standard finite elements with frequencies up to 7 kHz in the model problem (assuming that 10 grid points per wavelength are needed). We show that the UWVF is capable of generating useful results even when the frequency is 450 kHz, which corresponds to about six wavelengths per element. However, high wave numbers require the use of the nonuniform basis. We start by comparing the accuracy and conditioning of the UWVF with the uniform and nonuniform bases.

The analytical solution and two UWVF approximations of the problem for the frequency $f = 250$ kHz are shown in Fig. 4. The UWVF approximations are computed in the mesh of Fig. 3.

The effect of variability of the element size and the wave number on the conditioning of the matrix blocks D_k is shown in Fig. 5. The uniform basis leads to severe conditioning problems in the domain Ω_1 where the wave number is low. The highest values are in the smallest elements. However, for the nonuniform number of functions in the basis we can keep the condition numbers low and still reach the same accuracy with an almost equal number of degrees of freedom. The condition number of the blocks D_k in this example is limited to 10^4. Table I compares the accuracy and the condition numbers for the problem with uniform and nonuniform bases.

TABLE I
The Comparison of Uniform and Nonuniform Bases for $f = 250$ kHz

	Number of bases	Relative error	Max(Cond(D_k))	Number of degrees of freedom
Uniform basis	21	0.1349	$5.3 \cdot 10^{11}$	7014
Nonuniform basis	11 . . . 29	0.1260	$1.0 \cdot 10^{4}$	6956

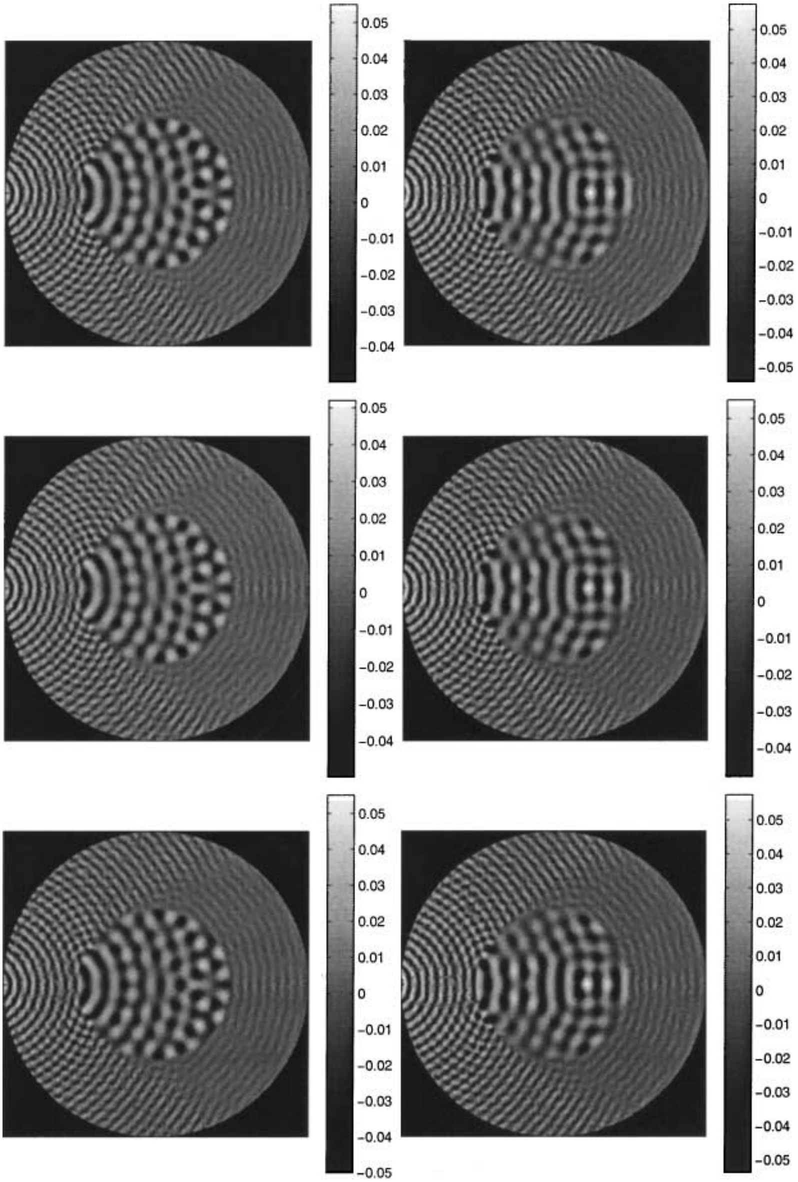


FIG. 4. Top: The analytical solution of the problem with $f = 250$ kHz. Middle: The UWVF approximation using the uniform basis with $p_k = 21$. Bottom: The UWVF approximation using the nonuniform basis with $p_k = 11, \ldots, 29$. The real parts are shown on the left and on the right are the imaginary parts. The UWVF approximations are computed in the mesh with $h_{\max} = 2.13$ cm.

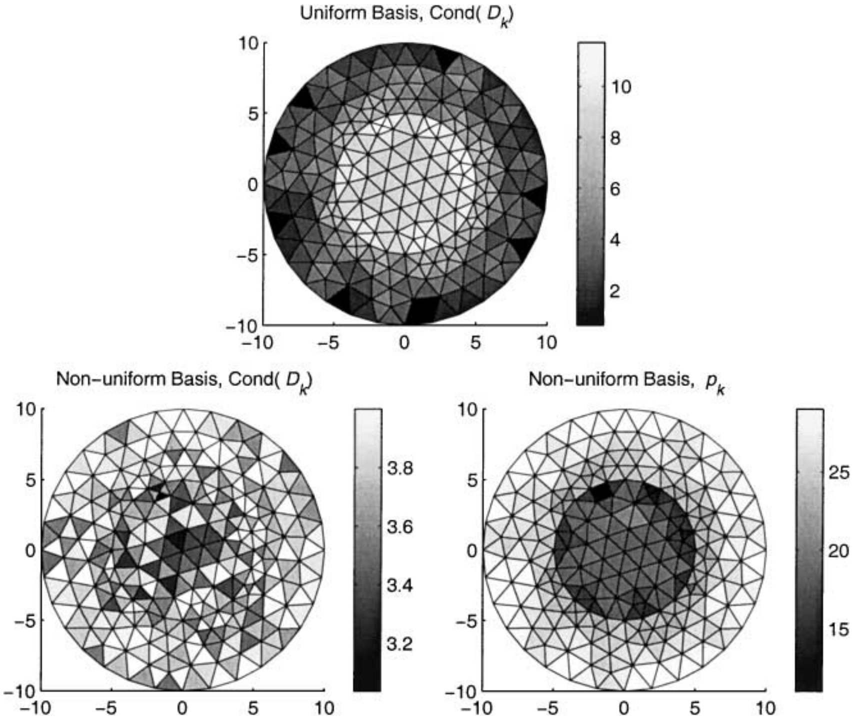


FIG. 5. Top: Color of an element represents the base 10 logarithm of the condition number of the matrix block D_k corresponding to the case of the uniform basis with $p = 21$ and $f = 250$ kHz. The condition number can be seen to be high in the domain of the lower wave number and especially in the small elements. To improve stability of the problem, the condition numbers of the blocks D_k are limited below 10^4. Bottom left: Base 10 logarithms of the condition numbers are shown for the nonuniform basis. Bottom right: The highest number of bases for each element Ω_k for which the condition number of D_k is below 10^4. The condition numbers for the elements are now between 10^3 and 10^4 while the number of basis functions per element varies from 11 to 29.

In this paper we have used the largest number of basis functions per element that give a condition number below the predetermined limit. The initial guess was the uniform basis with five plane waves. Although the blocks D_k had to be computed several times, due to faster convergence of the iterative algorithms, the total computation time did not increase. In the simulations of Fig. 4 the computation time for the uniform basis was 117 s, while using the nonuniform basis reduced the time to 89 s. The computations were done using a Pentium III PC with a 600-MHz processor and 1 GB of RAM. The code is written in Fortran90.

Naturally, a better choice of the initial basis would significantly reduce the computation time used to determine the bases.

We end this section by studying the dependence of the condition number of the matrix block D_k on the wave number κ_k. Using a fixed mesh and a fixed number of basis functions we compute the condition number for the frequency span $f = 100, \ldots, 500$ kHz. The condition numbers for two arbitrarily chosen elements are shown in Fig. 6. The condition numbers are graphed as a function of the ratio λ_k / h_k, where λ_k is the wavelength in the element Ω_k, and h_k is the length of the longest edge of the element. The results confirm that the condition number increases as the wavelength decreases, as noted in previous simulations.

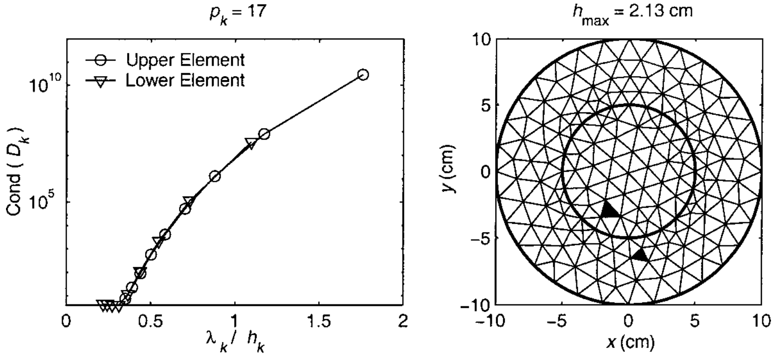


FIG. 6. Left: The condition number of the matrix block D_k is shown as a function of the elements per wavelength for two arbitrarily chosen elements. Right: The results are computed for the colored elements.

5.1.1. *The Error and Conditioning of* D_k

In this section we investigate the accuracy of the UWVF approximation and the conditioning of the matrix blocks D_k. The relative errors and the largest condition numbers of the blocks D_k for frequencies $f = 100$ kHz, $f = 250$ kHz, and $f = 400$ kHz are shown in Fig. 7. The number of grid points per wavelength is measured as the minimum ratio $\min(\lambda_k / h_k)$. The ratios corresponding to the above frequencies are $\min(\lambda_k / h_k) = 0.70$, 0.28, and 0.18, respectively. The results indicate that it is possible to use fairly coarse meshes, i.e., the element size is several times the wavelength.

Note that in the high-frequency case the choice of a uniform basis results in an unacceptable condition number before the error decreases to 49%. However, with the nonuniform

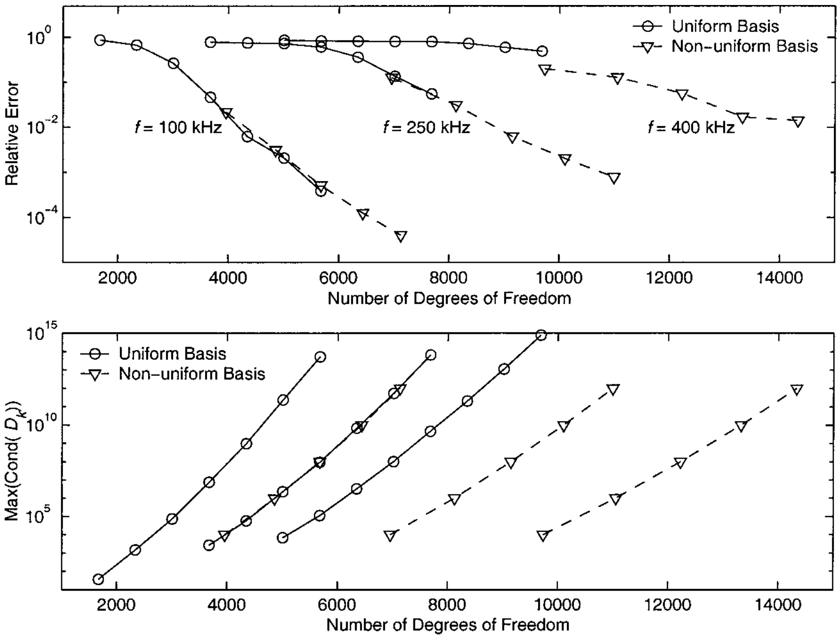


FIG. 7. Top: The figure shows the relative discrete L_2 error for three different frequencies against the number of degrees of freedom. Bottom: Corresponding maximum condition numbers of the blocks D_k. All results are computed in the same mesh with $h_{\max} = 2.13$ cm.

basis we are able to obtain 1.4% error for the highest frequency $f = 400$ kHz. On the other hand, with the lower frequencies the nonuniform basis approach provided a means for preconditioning the resulting matrix system. This topic is discussed in the following section together with iterative solvers.

5.1.2. *Preconditioning and Performance of the Iterative Solvers*

We have shown that an appropriate choice of basis can improve the stability of the inversion of the blocks D_k. To determine stability of the matrix equation (26) one needs to study conditioning of the operator $I - D^{-1}C$. We present the L_1-condition numbers for various operators for the frequency $f = 100$ kHz in Fig. 8. The plots correspond to the $f = 100$ kHz simulations in Fig. 7.

The results suggest that the maximum condition number of D_k characterizes the conditioning of the operator $I - D^{-1}C$. Also, note the superiority of the form $I - D^{-1}C$ compared to the nonpreconditioned equation $D - C$.

The convergence of the iterative solvers is studied by observing the norms

$$\text{Residual} = \frac{\|D^{-1}b - (I - D^{-1}C)X\|_2}{\|D^{-1}b\|_2}; \tag{32}$$

$$\text{Change in } X_i = \frac{\|X_i - X_{i-1}\|_2}{\|X_i\|_2}. \tag{33}$$

The residual is computed in the Bi-CGStab as $\|r_i\|_2/\|D^{-1}b\|_2$. The iterations are terminated when the residuals get below 10^{-6}.

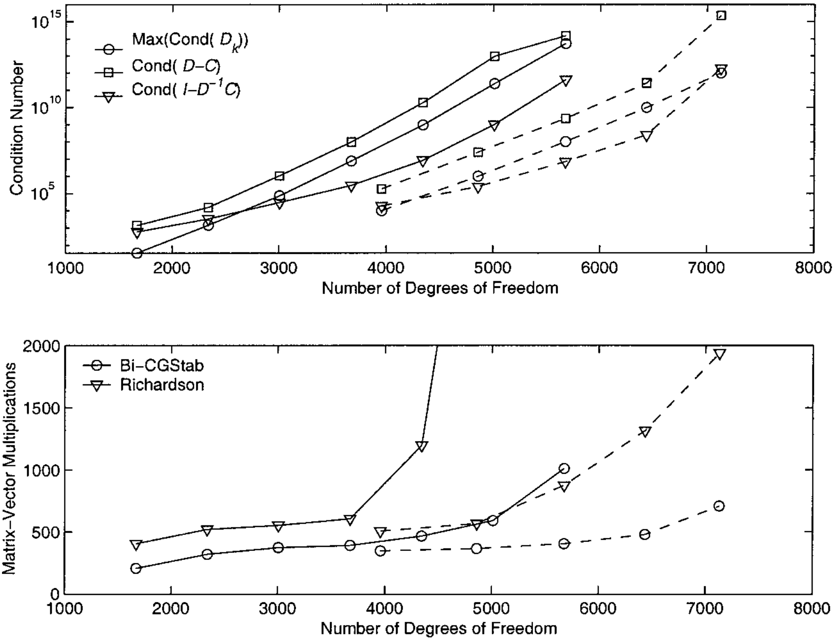
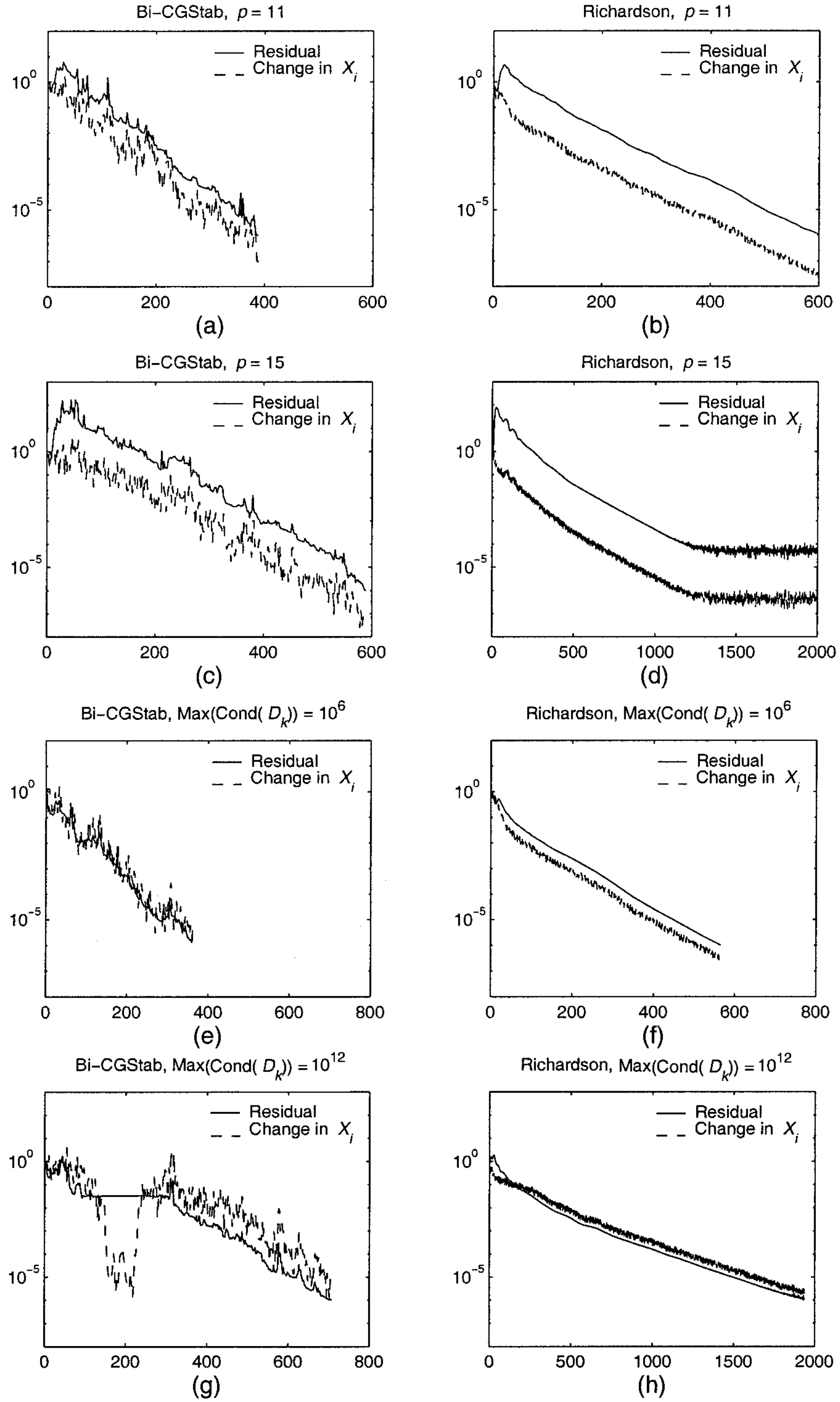


FIG. 8. Top: The figure shows conditioning of the UWVF matrices in the case of a uniform basis (solid lines) and of a nonuniform basis (dashed line) plotted against the number of degrees of freedom for $f = 100$ kHz and $h_{\max} = 2.13$ cm. Bottom: In the figure are the number of matrix–vector multiplications required to reach the termination criterion.



(a) (b) (c) (d) (e) (f) (g) (h)

FIG. 9. Convergence of the iterative solvers for $f = 100$ kHz and $h_{\max} = 2.13$ cm. The norms for the Bi-CGStab are in the left column and for the Richardson in the right column. We present the residuals for the uniform basis with $p = 11$ (a)–(b) and $p = 15$ (c)–(d). Convergence with the nonuniform basis is shown when the condition number of D_k is limited to 10^6 (e)–(f) and 10^{12} (g)–(h).

The iterative solvers are compared by counting the matrix–vector multiplications $y = (I - D^{-1}C)x = x - D^{-1}Cx$ needed to achieve the termination criterion. We remember that in addition to the matrix–vector operations four vector dot products are needed in the Bi-CGStab. However, the computational effort required for that is only a fraction of that needed for the matrix–vector operation.

The Bi-CGStab converges faster than the Richardson algorithm; see Fig. 8. The Bi-CGStab also reached the stopping criterion in all cases, whereas the Richardson stagnated in the case of the largest dimensional uniform basis.

The residuals as a function of iteration number for some simulations are presented in Fig. 9. In the same problem the Richardson algorithm failed to reach the termination criterion. Table II summarizes the simulations of Fig. 9. In all examples the convergence of the Richardson algorithm was smoother than that of the Bi-CGStab.

Finally, we show the relative error and variation in the number of bases as a function of the number of grid points per wavelength in Fig. 10. The results suggest that it is possible to obtain fairly accurate results on very coarse meshes. We get results with an error of 9% although we have over six wavelengths per element (at frequencies up to 450 kHz in the test problem). For higher frequencies ill-conditioning spoiled the results even though a nonuniform basis was used. We point out that the number of degrees of freedom needed to reach about 1% error for 450 kHz is only 13,200. That is orders of magnitude lower than that required in the piecewise linear finite element approach for the same problem with corresponding accuracy.

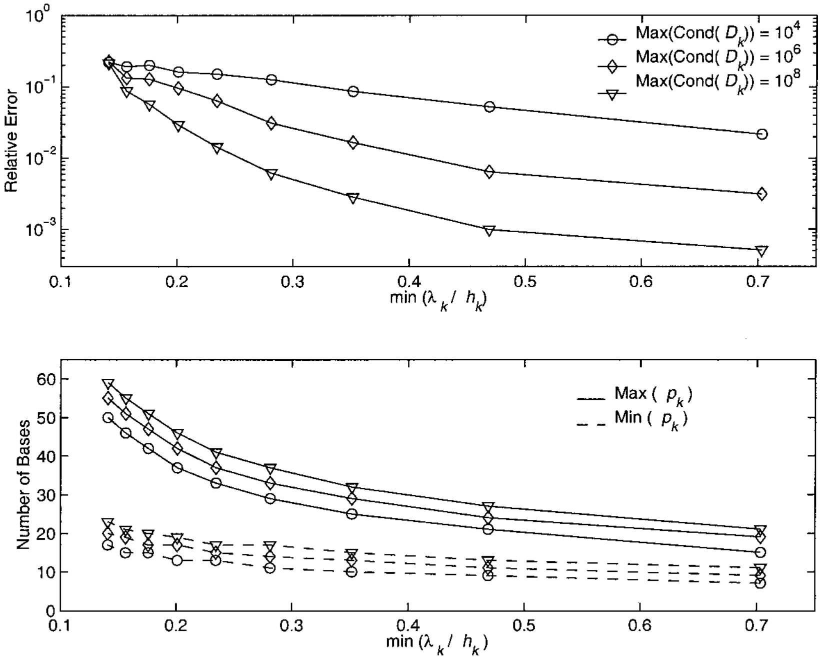


FIG. 10. Top: The relative error shown as a function of the elements per wavelength. We used different criteria for choosing the basis, i.e., the maximum condition numbers of the matrix blocks D_k were limited below 10^4, 10^6, and 10^8. We have used the highest number of bases that gives the condition number below the limits. Bottom: The figure shows corresponding maximum and minimum numbers of bases in each simulation. The largest variation in bases occurs when the element size is large compared to the wavelength. The results are computed in the mesh with $h_{\max} = 2.13$ cm and the frequency spanning 100 to 500 kHz.

TABLE II
A Summary of the Simulations of Fig. 9

	Uniform basis		Nonuniform basis	
Number of bases	11	15	9...19	15...27
Number of degrees of freedom	3674	5010	4859	7128
Max(Cond(D_k))	$7.7 \cdot 10^6$	$2.3 \cdot 10^{11}$	$9.9 \cdot 10^5$	$9.9 \cdot 10^{11}$
Cond($D - C$)	$9.7 \cdot 10^7$	$9.3 \cdot 10^{12}$	$2.4 \cdot 10^7$	$2.2 \cdot 10^{15}$
Cond($I - D^{-1}C$)	$2.9 \cdot 10^5$	$9.9 \cdot 10^8$	$2.5 \cdot 10^5$	$1.8 \cdot 10^{12}$
Matrix–Vector multiplications, Bi-CGStab	388	588	362	706
Matrix–Vector multiplications, Richardson	604	Stagnation	565	1939
Relative error	$4.63 \cdot 10^{-2}$	$2.07 \cdot 10^{-3}$	$3.13 \cdot 10^{-3}$	$4.05 \cdot 10^{-5}$

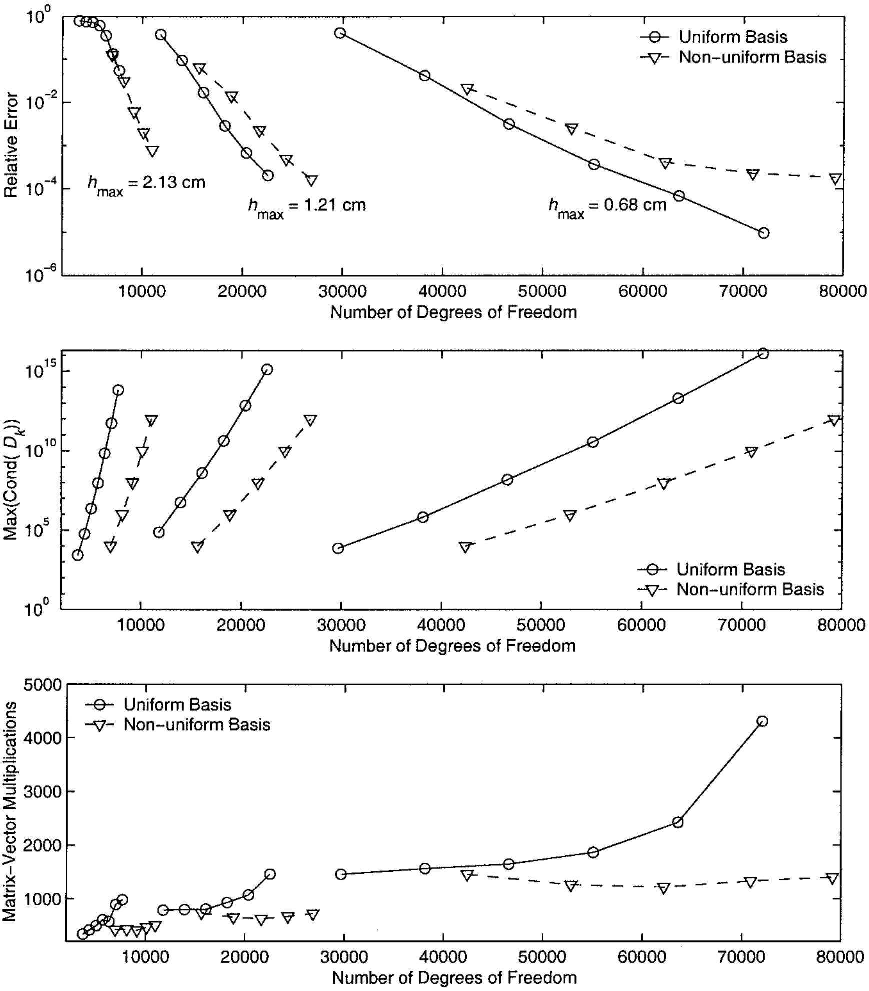


FIG. 11. Top: The figure represents the relative error against the number of degrees of freedom for different meshes for $f = 250$ kHz. Middle: The maximum condition numbers of D_k corresponding to the errors above. Bottom: The number of matrix–vector multiplications needed in the stabilized Bi-CG solver to reach the termination criterion.

For a fixed mesh the largest variation in the number of basis functions per element occurs for the shortest wavelengths. On the other hand, when the wavelegth increases the variation between elements decreases and it may be reasonable to use a uniform basis.

5.2. The Results under Mesh Refinement

We show the effect of mesh refinement on accuracy, the conditioning of D_k, and the convergence of iterations in Fig. 11. From the results it is obvious that an improvement in accuracy can be obtained with less effort by increasing the dimension of the basis rather than refining the mesh. Although the nonuniform basis did not improve the accuracy for the finer meshes an advantage is obtained in better stability and therefore faster convergence of the iterative solver.

We present the number of the matrix–vector operations for the Bi-CGStab method only since more operations were needed with the Richardson method. In addition, the convergence of the Richardson method stagnated in some of the most ill-conditioned cases.

6. CONCLUSIONS

We have shown that the use of nonuniform plane wave bases in the ultra-weak variational formulation improves its applicability to inhomogeneous Helmholtz problems with varying element sizes. The method proposed in this paper was based on the preconditioning of blocks in the resulting matrix equation. This led to variable dimension bases on different elements.

The results indicate that it is possible to use very large elements compared to the wavelength; in some simulations up to six wavelengths per elements size. This makes it possible to solve high-wave-number problems with coarse meshes and with a relatively low computational effort. We also showed that the benefit from the nonuniform basis approach is most significant when large elements are used.

In addition, we compared the Richardson and the stabilized Bi-Conjugate Gradient methods for solving the resulting linear system. The Richardson iteration converged more smoothly but stagnated in some cases. Fewer matrix–vector multiplications were needed in the Bi-CGStab to reach the same termination criterion. Using a nonuniform basis improved the convergence of both methods.

In this paper the number of basis functions was chosen by approximating the condition number of the blocks in the resulting matrix system. The number of basis functions per element was changed if the condition number was far from the predetermined value. This approach required the matrix blocks to be computed several times. A useful improvement would be a method to estimate the condition number for the blocks as a function of the number of bases based on the material parameters and the geometry of the elements.

REFERENCES

1. I. Babuska and J. M. Melenk, The partition of unity method, *Int. J. Numer. Meth. Eng.* **40**, 727 (1997).
2. J. D. Benamou and B. Despres, A domain decomposition method for the Helmholtz equation and related optimal control problems, *J. Comput. Phys.* **136**, 68 (1997).
3. O. Cessenat and B. Despres, Application of an ultra weak variational formulation of elliptic PDEs to the two-dimension Helmholtz problem, *SIAM J. Numer. Anal.* **35**(1), 255 (1998).

4. O. Cessenat, *Application d'une Nouvelle Formulation Variationelle des Equations d'Ondes Harmonigues, Problemes de Helmholtz 2D et de Maxwell 3D*, Ph.D. thesis (Paris IX Dauphine, 1996).
5. B. Despres, Sur une formulation variationnelle de type ultra-faible, *Comptes Rendus de l Academic des Sciences Series I* **318**, 939 (1994).
6. I. Harari and T. J. R. Hughes, Galerkin/least-squares finite element methods for the reduced wave equation with non-reflecting boundary conditions in unbounded domains, *Comput. Meth. Appl. Mech. Eng.* **98**, 411 (1992).
7. F. Ihlenburg and I. Babuska, Finite element solution of the Helmholtz equation with high wave number part I: The h-version of the FEM, *Comput. Math. Appl.* **30**(9), 9 (1995).
8. F. Ihlenburg and I. Babuska, 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**(1), 315 (1997).
9. O. Laghrouche and P. Bettess, Short wave modelling using special finite elements, *J. Comput. Acous.* **8**(1), 189 (2000).
10. P. Monk and D. Wang, A least squares method for the Helmholtz equation, *Comput. Meth. Appl. Mech. Eng.* **175**, 121 (1999).
11. A. A. Oberai and P. M. Pinsky, A residual-based finite element method for the Helmholtz equation, *Int. J. Numer. Meth. Eng.* **49**(3), 399 (2000).
12. P. Ortiz and E. Sanchez, An improved partition of the unity finite element model for diffraction problems, *Int. J. Numer. Meth. Eng.* **50**, 2727 (2001).
13. G. L. G. Sleijpen and D. R. Fokkema, BiCGstab(ℓ) for linear equations involving unsymmetric matrices with complex spectrum, *Electron. Trans. Numer. Anal.* **1**, 33 (1993).
14. M. Stojek, Least-squares Trefftz-type elements for the Helmholtz equation, *Int. J. Numer. Meth. Eng.* **41**(5), 831 (1998).
15. L. L. Thompson and P. M. Pinsky, A Galerkin least-squares finite element method for the two-dimensional Helmholtz equation, *Int. J. Numer. Meth. Eng.* **38**, 371 (1995).
16. H. A. Van Der Vorst, Bi-CGStab: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems, *SIAM J. Sci. Stat. Comput.* **13**(2), 631–644 (1992).
17. R. Weiss, *Parameter-Free Iterative Linear Solvers* (Akademie Verlag, Berlin, 1996).