INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 11, Number 3, Pages 587–607

A WEIGHTED VARIATIONAL FORMULATION BASED ON PLANE WAVE BASIS FOR DISCRETIZATION OF HELMHOLTZ EQUATIONS

QIYA HU AND LONG YUAN

Abstract. In this paper we are concerned with numerical methods for solving Helmholtz equations. We propose a new variant of the Variational Theory of Complex Rays (VTCR) method introduced in [15, 16]. The approximate solution generated by the new variant has higher accuracy than that generated by the original VTCR method. Moreover, the accuracy of the resulting approximate solution can be further increased by adding two suitable positive relaxation parameters into the new variational formula. Besides, a simple domain decomposition preconditioner is introduced for the system generated by the proposed variational formula. Numerical results confirm the efficiency of the new method.

Key words. Helmholtz equations, wave basis functions, variational formulation, error estimate, preconditioner, iteration counts.

1. Introduction

In recent years, the study of the vibrational behavior of mechanical systems has become a cornerstone of the design of industrial products and of the optimization of their performances. A key point in structural design is the modeling and calculation of the vibrational response of industrial structures. Strategies to analyze the structural and acoustical behavior of structures have been developed based on the finite element method [5, 17, 23, 1, 22], the boundary element method [10, 2, 4]. However, these methods are limited mainly to low-frequency problems and are either inaccurate or costly.

Today, there are also dedicated computational strategies for the resolution of medium-frequency problems, known as Trefftz methods [25], which differ from the traditional FEM and the BEM in the sense that the basis functions in Trefftz methods are chosen as some exact solutions of the governing differential equation without boundary condition. These approaches include the Ultra Weak Variational Formulation (UWVF) (see [3, 7]), the plane wave least-squares method [19], the plane wave discontinuous Galerkin methods (PWDG) (see [8, 11]), the discontinuous enrichment method [6] and the Variational Theory of Complex Rays (VTCR) introduced in [15, 16, 20] (see also [14] and [21]). An important advantage of these approaches is that they are capable of producing an approximate solution with high accuracy by using only a small number of DOFs. In this paper, we are interested in the development of the VTCR method.

The VTCR method has some similarity with the UWVF method. There are two basic ingredients in the both methods: a triangulation on the underlying domain and a set of wave basis functions in each element. But, two different kinds of unknown functions are chosen in the variational equations of these two methods.

Received by the editors May 8, 2013 and, in revised form August 22, 2013.
2000 *Mathematics Subject Classification.* 65N30, 65N55.
This research was supported by the Major Research Plan of Natural Science Foundation of China G91130015, the Key Project of Natural Science Foundation of China G11031006 and National Basic Research Program of China G2011309702.

For the VTCR method, the restrictions of the desired approximate solution on every elements are chosen as the unknown functions, and some weak continuity of the traces of the approximation across each local interface generated by the triangulation is imposed by a direct variational formulation involved the traces. For the UWVF method, the Robin boundary functions of the approximate solution on the boundaries of every elements are chosen as the unknown functions, the conjugation of each Robin boundary function has to be defined by introducing an additional mapping. Then, in the UWVF method, the same weak continuity of the traces is imposed by an indirect variational formulation involved the Robin boundary functions and their conjugations. The design of the UWVF method is based on the idea in the non-overlapping domain decomposition with Robin interface conditions, on contrary to this, the design of the VTCR method is only based on an intuitive idea to impose some weak continuity of the traces. It seems that the variational equation in the VTCR method is simpler than the one in the UWVF method, and the VTCR method is easier to implement than the UWVF method. By the way, the unknown functions in the PWDG method is also the restrictions of the desired approximate solution on every elements, and the PWDG method was derived by using the techniques in the DG method.

In this paper, we present a new variant of the VTCR method. In the variational equation of the original VTCR method, two different kinds of traces of the test function and the trial function are used in each integral on the common interface between two neighboring elements. The design of our method is also based on an intuitive idea to impose some weak continuity of the traces, but we change that variational equation such that the same kinds of traces of the test function and the trial function are used in each interface integral. We find that the approximate solution generated by the new variant has higher accuracy than the one generated by the original VTCR method. More importantly, the accuracy of the new approximation can be improved further by adding two suitable relaxation parameters into the new variational equation. For convenience, the resulting variational formulation is called a weighted variational formulation (WVF). We prove a L^2 error estimate of the approximate solution generated by the discrete WVF method. Numerical experiments for both two-dimensional and three-dimensional problems show that the new WVF method is obviously superior to the original VTCR method, and is as good as the UWVF method in convergence (the new WVF method seems easier to implement than the UWVF method). Unlike the existing discretization methods for Helmholtz equations, the coefficient matrix of the algebraic system generated by the WVF method is Hermite positive definite, so the algebraic system is easier to solve.

To solve the algebraic system generated by the WVF method in an efficient manner, we construct a simple domain decomposition preconditioner for the coefficient matrix of the algebraic system. The numerical results indicate that the systems generated by the WVF method for Helmholtz equations can be solved rapidly by the preconditioned GMRES method with the proposed preconditioner.

The paper is organized as follows: In Section 2, we briefly review the Variational Theory of Complex Rays for Helmholtz equations. In Section 3, we present a new variant of the VTCR for Helmholtz equations, with two relaxation parameters. In Section 4, we describe discretization of the variational formulation and derive an error estimate of the resulting approximate solution. In section 5, we construct a domain decomposition preconditioner for the stiffness matrix associated with the

new variational formulation. In Section 6, we report some numerical results to confirm the effectiveness of the new method.

2. Variational Theory of Complex Rays for Helmholtz equations

The purpose of this section is to recall the basic principles of the VTCR modeling methodology for the resolution of Helmholtz's equation (see [20] for more details). At first the original problem to be solved is defined. Then the variational formulation is presented in details.

2.1. The reference problem. Let Ω, which denotes an acoustic cavity in applications, be a bounded and connected Lipschitz domain in $\mathbb{R}^l, (l = 2, 3)$. Consider Helmholtz equations which is formalized, normalizing the wave's velocity to 1, by

$$\left\{\begin{array}{ll} -\Delta u - \omega^2 u = f & \text{in} \quad \Omega, \\ (\partial_{\mathbf{n}} + i\omega)u = t(-\partial_{\mathbf{n}} + i\omega)u + g & \text{on} \quad \gamma, \\ |t| < 1, t \in \mathbb{C}. \end{array}\right. \tag{1}$$

The outer normal derivative is referred to by $\partial_{\mathbf{n}}$, and the angular frequency is denoted by ω.

The following classical result can be found in [3].

Theorem 2.1. *Let Ω be an open bounded set, and γ be its boundary assuming it is of class C^1 nearly everywhere. Let $f \in L^2(\Omega)$ and $g \in L^2(\Omega)$. We let $\zeta = \frac{1-t}{1+t}$ and assume t to be constant, $|t| < 1$ (then $\Re(\zeta) > 0$). Then, there exists a unique $u \in H^1(\Omega)$ satisfying*

$$\left\{\begin{array}{c} \forall v \in H^1(\Omega) \\ \displaystyle\int_\Omega \nabla u \cdot \nabla \bar{v} - \omega^2 \int_\Omega u\bar{v} + i\omega\zeta \int_\gamma u\bar{v} = \int_\Omega f\bar{v} + \frac{1}{1+t}\int_\gamma g\bar{v}, \end{array}\right. \tag{2}$$

or equivalently

$$\left\{\begin{array}{ll} -\Delta u - \omega^2 u = f & \textit{in} \quad \Omega, \\ (\partial_{\mathbf{n}} + i\omega)u = t(-\partial_{\mathbf{n}} + i\omega)u + g & \textit{on} \quad \gamma. \end{array}\right. \tag{3}$$

where $\overline{\diamond}$ designate the complex conjugate of the complex quantity $\diamond$.

□

2.2. The variational formulation of the problem. Let Ω be divided into a partition in the sense that

$$\overline{\Omega} = \bigcup_{k=1}^{N} \overline{\Omega}_k, \quad \Omega_k \bigcap \Omega_j = \emptyset \quad \text{for } k \neq j.$$

In practice, the partition is a mesh of domain, and the sets $\{\Omega_k\}$ are the elements. Let $\mathcal{T}_h$ denote the triangulation associated with the elements $\{\Omega_k\}$, where h is the size of the triangulation. Define

$$\begin{array}{c} \Gamma_{kj} = \partial\Omega_k \bigcap \partial\Omega_j \quad \text{for } k \neq j, \\ \gamma_k = \overline{\Omega}_k \bigcap \partial\Omega \quad (k = 1, \cdots, N), \\ \gamma = \displaystyle\bigcup_{k=1}^{N} \gamma_k, \quad \Gamma = \bigcup_{k=1}^{N} \partial\Omega_k. \end{array} \tag{4}$$

For simplicity, we consider only the case $t = 0$ in the rest of this paper, and we can directly generalize it to the other case with $t \neq 0$. Set $u|_{\Omega_k} = u_k$ $(k = 1, \cdots, N)$. Then the reference problem to be solved consists in finding the local acoustic pressures $u_k \in H^1(\Omega_k)$ such that

$$(5)\quad \begin{cases} -\Delta u_k - \omega^2 u_k = 0 & \text{in} \quad \Omega_k, \\ (\partial_{\mathbf{n}} + i\omega)u_k = g & \text{on} \quad \gamma \ (if\ \partial\Omega_k \cap \gamma \neq \emptyset) \end{cases} \qquad (k = 1, 2, \ldots, N),$$

and

$$(6)\quad \begin{cases} u_k - u_j = 0 & \text{over} \quad \Gamma_{kj}, \\ \partial_{\mathbf{n}_k} u_k + \partial_{\mathbf{n}_j} u_j = 0 & \text{over} \quad \Gamma_{kj} \end{cases} \qquad (k \neq j;\ k, j = 1, 2, \cdots, N).$$

The first equation of (5) is the homogeneous Helmholtz equation, where the wave's velocity equals 1, so we call ω the wave number. The second equation of (5) and the equation (6) are related to the boundary condition of the problem and the continuity conditions at the interface between the subcavities Ω_k and Ω_j.

A weak form of the reference problem introduced in Section 2.1 can be derived using a variational formulation introduced in [15]. This formulation verifies the boundary conditions (5) and (6) in a weak sense.

Let $V(\Omega_k)$ denote the space of the functions which verify Helmholtz's homogeneous equation (5) on the cavity Ω_k:

$$(7)\qquad V(\Omega_k) = \{v_k \in H^1(\Omega_k);\ \Delta v_k + \omega^2 v_k = 0\}.$$

Define

$$V(\mathcal{T}_h) = \prod_{k=1}^{N} V(\Omega_k),$$

with the natural scalar product

$$(u, v)_V = \sum_{k=1}^{N} \int_{\Omega_k} u_k \cdot \overline{v}_k \ d\mathbf{x},\ \forall u, v \in V(\mathcal{T}_h).$$

In the original VTCR method (see, for example, [21]), the variational problem of (5) and (6) can be expressed as follows : find $u \in V(\mathcal{T}_h)$ such that

$$(8)\quad \begin{aligned} Re\Bigg\{ \sum_{k=1}^{N} \int_{\gamma_k} \frac{1}{2}\Bigg(((\partial_{\mathbf{n}} + i\omega)u_k - g) \cdot \overline{\frac{-1}{\omega}\partial_{\mathbf{n}} v_k} + \overline{i((\partial_{\mathbf{n}} + i\omega)u_k - g)} \cdot v_k \Bigg) ds \\ + \frac{1}{2} \sum_{j \neq k} \int_{\Gamma_{kj}} \Bigg((u_k - u_j) \cdot \overline{i(\partial_{\mathbf{n}_k} v_k - \partial_{\mathbf{n}_j} v_j)} \\ + \overline{i(\partial_{\mathbf{n}_k} u_k + \partial_{\mathbf{n}_j} u_j)} \cdot (v_k + v_j) \Bigg) ds \Bigg\} = 0, \forall v \in V(\mathcal{T}_h), \end{aligned}$$

which is equivalent to

$$(9)\quad \begin{aligned} Re\Bigg\{ \sum_{k=1}^{N} \int_{\gamma_k} \Bigg(\frac{1}{\omega}((\partial_{\mathbf{n}} + i\omega)u_k - g) \cdot \overline{\partial_{\mathbf{n}} v_k} + i\ \overline{((\partial_{\mathbf{n}} + i\omega)u_k - g)} \cdot v_k \Bigg) ds \\ + i \sum_{j \neq k} \int_{\Gamma_{kj}} \Bigg((u_k - u_j) \cdot \overline{(\partial_{\mathbf{n}_k} v_k - \partial_{\mathbf{n}_j} v_j)} \\ + \overline{(\partial_{\mathbf{n}_k} u_k + \partial_{\mathbf{n}_j} u_j)} \cdot (v_k + v_j) \Bigg) ds \Bigg\} = 0, \forall v \in V(\mathcal{T}_h), \end{aligned}$$

where $Re\{\diamond\}$ designate the real part of the complex quantity $\diamond$.

When the space $V(\mathcal{T}_h)$ is replaced with a suitable finite dimensional subspace of $V(\mathcal{T}_h)$ (see [20] and Section 5 for the details), then a discrete version of (9) can be derived. The VTCR method is easier to understand and implement. However, our numerical experiments indicate that the approximate solutions generated by the discrete version have lower accuracy than that generated by the UWVF method introduced by [3]. In particular, numerical results show that the VTCR method is not available for the singularity problems. Moreover, the algebraic system induced from this discrete problem seems difficult to solve when the mesh sizes become small. In order to investigate the reasons, we need to analyze the structure of the integral (which appears in the second sum of (9)):

$$\int_{\Gamma_{kj}} (u_k - u_j) \cdot \overline{(\partial_{\mathbf{n}_k} v_k - \partial_{\mathbf{n}_j} v_j)} ds.$$

Note that $(u_k - u_j)|_{\Gamma_{kj}}$ is $0-order$ trace but $(\partial_{\mathbf{n}_k} v_k - \partial_{\mathbf{n}_j} v_j)|_{\Gamma_{kj}}$ is $1-order$ trace. This means that two different kinds of traces are used in the integral. From the mathematical viewpoint, the functions $(u_k - u_j)|_{\Gamma_{kj}}$ and $(\partial_{\mathbf{n}_k} v_k - \partial_{\mathbf{n}_j} v_j)|_{\Gamma_{kj}}$ belong two different spaces, namely, $(u_k - u_j)|_{\Gamma_{kj}} \in H^{\delta-\frac{1}{2}}(\Gamma_{kj})$ but $(\partial_{\mathbf{n}_k} v_k - \partial_{\mathbf{n}_j} v_j)|_{\Gamma_{kj}} \in H^{\delta-\frac{3}{2}}(\Gamma_{kj})$ when $u_k \in H^{\delta}(\Omega_k)$ and $u_j \in H^{\delta}(\Omega_j)$ with $\delta > \frac{3}{2}$. Thus the continuity of u defined by (9) across the local interface Γ_{kj} must hold in $H^{\delta-\frac{1}{2}}(\Gamma_{kj})$, which is too strong (since $\delta > \frac{3}{2}$).

Based on the above observation, we propose a new variant of (9) in the next section.

3. A new variational formulation for Helmholtz equations

Our main idea is to use the same kind of trace in each interface integral. For ease of understanding, we would like to derive the variational formula from a minimization problem.[1]

Let α and β be two given positive real numbers. Corresponding to the boundary condition in (5) and the interface continuity condition (6), we define the functional

$$J(v) = \sum_{k=1}^{N} \int_{\gamma_k} |(\partial_{\mathbf{n}} + i\omega) v_k - g|^2 ds$$
$$+ \sum_{j \neq k} \left(\alpha \int_{\Gamma_{kj}} |v_k - v_j|^2 ds + \beta \int_{\Gamma_{kj}} |\partial_{\mathbf{n}_k} v_k + \partial_{\mathbf{n}_j} v_j|^2 ds \right), \ v \in V(\mathcal{T}_h).$$

It is clear that $J(v) \geq 0$. Consider the minimization problem: find $u \in V(\mathcal{T}_h)$ such that

$$J(u) = \min_{v \in V(\mathcal{T}_h)} J(v) \tag{10}$$

If u is the solution of the problem (1) (with $t = 0$), i.e., $u \in V(\mathcal{T}_h)$ satisfies the boundary condition in (5) and the interface continuity condition (6), then we have $J(u) = 0$, which implies that u is also the solution of the minimization problem (10). The introduction of the two relaxation parameters α and β in the above

[1]The authors found that a similar variational formula had been proposed in [19] when the current article was prepared to publish, but the authors can not make essential revision to the introduction of the article.

functional is based on the following motive: when the wave number ω is large, the analytic solution u of (1) becomes high oscillating, and so the jump

$$|\partial_{\mathbf{n}_k} u_k + \partial_{\mathbf{n}_j} u_j|$$

may be not so small as the jump

$$|u_k - u_j|.$$

In particular, when the analytic solution u is singular, the jump $|\partial_{\mathbf{n}_k} u_k + \partial_{\mathbf{n}_j} u_j|$ may not be small on Γ_{kj}. We can imagine that

$$\alpha \int_{\Gamma_{kj}} |u_k - u_j|^2 ds \approx \beta \int_{\Gamma_{kj}} |\partial_{\mathbf{n}_k} u_k + \partial_{\mathbf{n}_j} u_j|^2 ds \to 0, \ \forall v \in V(\mathcal{T}_h).$$

Then, for a large ω, we can choose the relaxation parameters α and β as $\alpha \gg \beta$, such that

$$|u_k - u_j| \ll |\partial_{\mathbf{n}_k} u_k + \partial_{\mathbf{n}_j} u_j| \to 0 \text{ on } \Gamma_{kj}.$$

The variational problem associated with the minimization problem (10) can be expressed as follows: find $u \in V(\mathcal{T}_h)$ such that

$$\begin{aligned} &\sum_{k=1}^{N} \int_{\gamma_k} ((\partial_{\mathbf{n}} + i\omega)u_k - g) \cdot \overline{(\partial_{\mathbf{n}} + i\omega)v_k} ds + \sum_{j \neq k} \Bigg(\alpha \int_{\Gamma_{kj}} (u_k - u_j) \cdot \overline{(v_k - v_j)} ds \\ &\quad + \beta \int_{\Gamma_{kj}} (\partial_{\mathbf{n}_k} u_k + \partial_{\mathbf{n}_j} u_j) \cdot \overline{(\partial_{\mathbf{n}_k} v_k + \partial_{\mathbf{n}_j} v_j)} ds \Bigg) = 0, \ \forall v \in V(\mathcal{T}_h). \end{aligned} \tag{11}$$

In the numerical experiments made in Section 6, we will choose $\alpha = \omega^2$ and $\beta = 1$ (for the case with smooth solution), or $\alpha = \omega$ and $\beta = \frac{1}{\omega}$ (for the case with singular solution). We will find that the effectiveness of the method can be significantly improved by such choice of α and β. For convenience, the variational formulation (11) is called a *weighted variational formulation* (WVF) for the problem defined by (5) and (6).

Define the sesquilinear form $a(\cdot, \cdot)$ by

$$\begin{aligned} a(u, v) = &\sum_{k=1}^{N} \int_{\gamma_k} ((\partial_{\mathbf{n}} + i\omega)u_k) \cdot \overline{(\partial_{\mathbf{n}} + i\omega)v_k} ds \\ &+ \sum_{j \neq k} \Bigg(\alpha \int_{\Gamma_{kj}} (u_k - u_j) \cdot \overline{(v_k - v_j)} ds \\ &+ \beta \int_{\Gamma_{kj}} (\partial_{\mathbf{n}_k} u_k + \partial_{\mathbf{n}_j} u_j) \cdot \overline{(\partial_{\mathbf{n}_k} v_k + \partial_{\mathbf{n}_j} v_j)} ds \Bigg), \ \forall v \in V(\mathcal{T}_h), \end{aligned} \tag{12}$$

and $\xi \in V(\mathcal{T}_h)$, via the Riesz representation theorem, by

$$(\xi, \ v)_V = \sum_{k=1}^{N} \int_{\gamma_k} g \cdot \overline{(\partial_{\mathbf{n}} + i\omega)v_k} ds \ \forall v \in V(\mathcal{T}_h). \tag{13}$$

Then (11) can be written as:

$$\begin{cases} \textit{Find } u \in V(\mathcal{T}_h), \textit{s.t.} \\ a(u, v) = (\xi, \ v)_V, \quad \forall v \in V(\mathcal{T}_h). \end{cases} \tag{14}$$

Theorem 3.1. *Let $u \in V(\mathcal{T}_h)$. For $k = 1, \cdots, N$, assume that $u_k \in H^{1+\delta_k}(\Omega_k)$ with $\delta_k > \frac{1}{2}$ such that $\partial_{\mathbf{n}_k} u_k \in L^2(\partial\Omega_k)$. Then the reference problem (5) and (6) is equivalent to the new variational problem (14).*

Proof. It is clear that the solution of the problem (5)-(6) is also the solution of the variational problem (14). Therefore one needs only to verify the uniqueness of solution of the problem (14).

The verification is standard. Let us consider two solutions $u = (u_1, \cdots, u_N), u' = (u'_1, \cdots, u'_N)$ of the variational problem (11), and let $\tilde{u} = (\tilde{u}_1, \cdots, \tilde{u}_N)$ denote the difference between the two solutions. Because of (11), these two solutions must verify the following equation:

$$\begin{aligned} &\sum_{k=1}^{N} \int_{\gamma_k} ((\partial_{\mathbf{n}} + i\omega)\tilde{u}_k \cdot \overline{(\partial_{\mathbf{n}} + i\omega)v_k} ds + \sum_{j\neq k} \Big(\alpha \int_{\Gamma_{kj}} (\tilde{u}_k - \tilde{u}_j) \cdot \overline{(v_k - v_j)} ds \\ &+ \beta \int_{\Gamma_{kj}} (\partial_{\mathbf{n}_k} \tilde{u}_k + \partial_{\mathbf{n}_j} \tilde{u}_j) \cdot \overline{(\partial_{\mathbf{n}_k} v_k + \partial_{\mathbf{n}_j} v_j)} ds \Big) = 0, \ \forall v \in V(\mathcal{T}_h). \end{aligned} \tag{15}$$

Taking $v = \tilde{u}$, the equation (15) simplifies to:

$$\begin{aligned} &\sum_{k=1}^{N} \int_{\gamma_k} (\partial_{\mathbf{n}} + i\omega)\tilde{u}_k \cdot \overline{(\partial_{\mathbf{n}} + i\omega)v_k} ds + \sum_{j\neq k} \Big(\alpha \int_{\Gamma_{kj}} (\tilde{u}_k - \tilde{u}_j) \cdot \overline{(\tilde{u}_k - \tilde{u}_j)} ds \\ &+ \beta \int_{\Gamma_{kj}} (\partial_{\mathbf{n}_k} \tilde{u}_k + \partial_{\mathbf{n}_j} \tilde{u}_j) \cdot \overline{(\partial_{\mathbf{n}_k} \tilde{u}_k + \partial_{\mathbf{n}_j} \tilde{u}_j)} ds \Big) = 0. \end{aligned} \tag{16}$$

Namely,

$$\begin{aligned} &\sum_{k=1}^{N} \int_{\gamma_k} |(\partial_{\mathbf{n}} + i\omega)\tilde{u}_k|^2 ds + \sum_{j\neq k} \Big(\alpha \int_{\Gamma_{kj}} |\tilde{u}_k - \tilde{u}_j|^2 ds \\ &+ \beta \int_{\Gamma_{kj}} |\partial_{\mathbf{n}_k} \tilde{u}_k + \partial_{\mathbf{n}_j} \tilde{u}_j|^2 ds \Big) = 0. \end{aligned} \tag{17}$$

Note that $\alpha, \beta > 0$, the above equality implies that

$$\int_{\gamma_k} |(\partial_{\mathbf{n}} + i\omega)\tilde{u}_k|^2 ds = 0, \ \int_{\Gamma_{kj}} |\tilde{u}_k - \tilde{u}_j|^2 ds = 0, \ \int_{\Gamma_{kj}} |\partial_{\mathbf{n}_k} \tilde{u}_k + \partial_{\mathbf{n}_j} \tilde{u}_j|^2 ds = 0.$$

These show that the function $\tilde{u}$ satisfies the interface continuity (6) and verifies the initial Helmholtz reference problem (5) (note that $\tilde{u} \in V(\mathcal{T}_h)$) with the homogeneous boundary condition. Therefore $\tilde{u}$ vanishes on Ω, which proves the uniqueness of solution of (14).

□

4. Discretization of the variational formulation

In this section, we consider a discretization of the variational problem (14). The discretization is based on a finite dimensional space $V_p(\mathcal{T}_h) \subset V(\mathcal{T}_h)$. We first give the exact definition of such a space $V_p(\mathcal{T}_h)$.

4.1. The basis functions of $V_p(\mathcal{T}_h)$. In each element Ω_k, we introduce a finite number of functions y_{kl} $(l = 1, 2, \cdots, p)$ supported in Ω_k and that are independent solutions of the homogeneous Helmholtz equation (without boundary condition) in the element Ω_k $(k = 1, 2, \cdots, N)$.

To simplify, we consider some constant number p of basis functions for all elements Ω_k. Particularly, in this paper we will choose y_{kl} as the wave shape functions on Ω_k, which satisfy

$$\begin{cases} y_{kl}(\mathbf{x}) = e^{i\omega(\mathbf{x}\cdot\boldsymbol{\alpha}_l)}, \ \mathbf{x} \in \Omega_k, \\ \boldsymbol{\alpha}_l \cdot \boldsymbol{\alpha}_l = 1, \\ l \neq s \rightarrow \boldsymbol{\alpha}_l \neq \boldsymbol{\alpha}_s, \end{cases} \tag{18}$$

where $\boldsymbol{\alpha}_l$ $(l = 1, \cdots, p)$ are unit wave propagation directions to be specified later.

The basis functions of $V_p(\mathcal{T}_h)$ can be defined as

$$\phi_{kl}(\mathbf{x}) = \begin{cases} y_{kl}(\mathbf{x}), \ \mathbf{x} \in \Omega_k, \\ 0, \ \mathbf{x} \in \Omega_j \text{ satisfying } j \neq k \end{cases} \quad (k, j = 1, \cdots, N; \ l = 1, \cdots, p). \tag{19}$$

Thus the space $V(\mathcal{T}_h)$ is discretized by the subspace

$$V_p(\mathcal{T}_h) = span\Big\{\phi_{kl} : \ k = 1, \cdots, N; \ l = 1, \cdots, p\Big\}. \tag{20}$$

During numerical simulations, the directions of the wave vectors of these wave functions, for two-dimensional problems, are uniformly distributed as follows:

$$\boldsymbol{\alpha}_l = \begin{pmatrix} (cos(2\pi(l-1)/p) \\ sin(2\pi(l-1)/p)) \end{pmatrix} \ (l = 1, \cdots, p).$$

For three-dimensional problems, the following formula proposed in [24] can be used to generate the wave propagation derections

$$\boldsymbol{\alpha}_{j1,j2,j3} = \frac{\hat{\boldsymbol{\alpha}}_{j1,j2,j3}}{||\hat{\boldsymbol{\alpha}}_{j1,j2,j3}||}, \hat{\boldsymbol{\alpha}}_{j1,j2,j3} = \begin{pmatrix} tan((2j_1/n_t - 1)\pi/4) \\ tan((2j_2/n_t - 1)\pi/4) \\ tan((2j_3/n_t - 1)\pi/4) \end{pmatrix}$$

where n_t is a given positive integer and $j_1, j_2, j_3 = 0, \cdots, n_t$ are positive integers chosen so that at least one of j_1, j_2, or j_3 is equal to zero or to n_t. Using this construction algorithm, the number of directions p becomes equal to $6n_t^2 + 2$. For example, choosing $n_t = 2, n_t = 3$, and $n_t = 4$ leads to 26, 56 and 98 wave functions, respectively.

4.2. The discrete problem and the algebraic form of (14). Let $V_p(\mathcal{T}_h)$ be defined in the last subsection. Then the discrete variational problem associated with (14) can be described as follows:

$$\begin{cases} Find \ u_h \in V_p(\mathcal{T}_h), \ s.t. \\ a(u_h, v_h) = (\xi, \ v_h)_V, \ \forall v_h \in V_p(\mathcal{T}_h). \end{cases} \tag{21}$$

After solving (21), the approximated solutions of Helmholtz equations (1) are obtained directly since the unknown u_h are defined on the elements $\{\Omega_k\}$, instead of their boundaries $\partial\Omega_k$ (as in UWVF). Moreover, the structure of the sesquiinear form $a(\cdot, \cdot)$ is very simple, so the method seem easier to implement than the UWVF method.

Let $\mathcal{A}$ be the stiffness matrix associated with the sesquiinear form $a(\cdot, \cdot)$ and the space $V_p(\mathcal{T}_h)$, and let b denote the vector associated with the scalar product $(\xi, \ v_h)_V$. Namely, the entries of the matrix $\mathcal{A}$ are computed by $a_{k,j}^{l,m} = a(\phi_{jm}, \phi_{kl})$; and the complements of the vector b are defined as $b_{k,l} = (\xi, \ \psi_{kl})_V$. Then the discretized problem (21) leads to the algebraic system below:

$$\mathcal{A}X = b, \tag{22}$$

where $X = (x_{11}, x_{12}, \cdots, x_{1p}, x_{21}, \cdots, x_{2p}, \cdots, x_{N1}, \cdots, x_{Np})^t \in \mathbb{C}^{pN}$ is the unknown vector. From the definition of the bilinear form $a(\cdot,\cdot)$, we know that the matrix $\mathcal{A}$ is Hermite positive definite, so the system (22) is easier to solve than that generated by the other existing methods.

In general the system (22) is solved by some iterative method, for example, the preconditioned GMRES method. Then we need to construct an efficient preconditioner $\mathcal{B}$ for the matrix $\mathcal{A}$, and use GMRES method to solve the equivalent system

$$\mathcal{B}^{-1}\mathcal{A}X = \mathcal{B}^{-1}b. \tag{23}$$

4.3. Error estimates of the approximate solution defined by (21). Before construct the preconditioner $\mathcal{B}$, we derive estimates of the error $u - u_h$ in this subsection, where u and u_h are defined by (14) and (21) respectively.

As in [11], for a given a domain $D \subset \mathbb{R}^l, (l = 2, 3)$, let $||\cdot||_{s,\omega,D}$ be the $\omega-$weighted Sobolev norm defined by

$$||v||_{s,\omega,D} = \sum_{j=0}^{s} \omega^{2(s-j)}|v|^2_{j,D}.$$

The following lemma is a direct consequence of Theorem 5.2 and Theorem 5.3 in [18].

Lemma 4.1. [18] *Let $m \geq 2$ be an integer and set $p = 2m + 1$ (for 2d case) or $p = (m+1)^2$ (for 3d case). Assume that $u \in C^{m+1}(\Omega_k)$ for each element Ω_k. Then there is a function $w_h \in V_p(\mathcal{T}_h)$ such that*

$$||u - w_h||_{\omega,l,\Omega_k} \leq Ch^{m+1-l}||u||_{\omega,m+1,\Omega_k} \ (k = 1, \cdots, N), \tag{24}$$

for $l = 0, 1, 2$.

□

It is clear that $a(v, v) \geq 0$. Moreover, from the proof of Theorem 3.1, we can see that $a(v, v) = 0$ for $v \in V(\mathcal{T}_h)$ if only if $v = 0$. Thus $a(\cdot,\cdot)$ is a norm on $V(\mathcal{T}_h)$. For ease of notation, this norm is denoted by $||\cdot||_V$.

The following lemma can be obtained as Corollary 3.8 in [11].

Lemma 4.2. *There is a constant C independent of h and ω such that*

$$||u - u_h||^2_{0,\Omega} \leq C(h\omega + h^{-1}\omega^{-1})\omega^{-1}||u - u_h||^2_V \tag{25}$$

□

Theorem 4.1. *Let the assumptions in the above lemma be satisfied. Assume that $\alpha = \omega^2$, $\beta = 1$ and $\omega h \leq c_0$ for a constant c_0. Then*

$$||u - u_h||_V \leq Ch^{m-\frac{1}{2}}(\sum_{k=1}^{N} ||u||^2_{\omega,m+1,\Omega_k})^{\frac{1}{2}} \tag{26}$$

and

$$||u - u_h||_{0,\Omega} \leq C(1 + (h\omega)^{-1})h^m(\sum_{k=1}^{N} ||u||^2_{\omega,m+1,\Omega_k})^{\frac{1}{2}}, \tag{27}$$

where C is independent of h, ω, but may be dependent of p (see [11] for the details).

Proof. Let w_h be defined by Lemma 4.1. Then $u_h - w_h \in V_p(\mathcal{T}_h)$. Note that $V_p(\mathcal{T}_h) \subset V(\mathcal{T}_h)$, we have by the definition of u and u_h

$$a(u - u_h, u_h - w_h) = 0.$$

Then we get by Cauchy-Schwarz inequality

$$||u-u_h||_V^2 = a(u-u_h, u-w_h) \le ||u-u_h||_V \cdot ||u-w_h||_V,$$

which implies that

$$||u-u_h||_V \le ||u-w_h||_V. \tag{28}$$

It suffices to estimate $||u-w_h||_V$. For ease of notation, set $\varepsilon_h = u - w_h$. By the definition of the norm $||\cdot||_V$, we get

$$\begin{aligned} ||\varepsilon_h||_V^2 &= \int_\gamma |i\omega\varepsilon_h + \frac{\partial \varepsilon_h}{\partial \mathbf{n}}|^2 ds \\ &+ \sum_{\Gamma_{kj}} \left\{ \omega^2 \int_{\Gamma_{kj}} |\varepsilon_{h,k} - \varepsilon_{h,j}|^2 ds + \int_{\Gamma_{kj}} |\frac{\partial \varepsilon_{h,k}}{\partial \mathbf{n}_k} - \frac{\partial \varepsilon_{h,j}}{\partial \mathbf{n}_j}|^2 ds \right\} \\ &\le \sum_{r=1}^{N} \left\{ \omega^2 \int_{\partial\Omega_r} |\varepsilon_{h,r}|^2 ds + \int_{\partial\Omega_r} |\frac{\partial \varepsilon_{h,r}}{\partial \mathbf{n}_r}|^2 ds \right\}, \end{aligned} \tag{29}$$

where $\varepsilon_{h,r} = \varepsilon_h|_{\Omega_r}$ $(r = k, j)$. In an analogous way with the proof of Lemma 3.10 in [11], we can prove, by the trace theorems and Lemma 4.1, that

$$\omega^2 \int_{\partial\Omega_r} |\varepsilon_{h,r}|^2 ds + \int_{\partial\Omega_r} |\frac{\partial \varepsilon_{h,r}}{\partial \mathbf{n}_r}|^2 ds \le Ch^{2m-1} ||u||^2_{\omega,m+1,\Omega_k}.$$

Substituting the above inequality into (29) and combing (28), yields (26). Furthermore, the inequality (27) can be derived by Lemma 4.2.

□

5. A domain decomposition preconditioner $\mathcal{B}$

In this section, we construct a preconditioner $\mathcal{B}$ based on the non-overlapping domain decomposition method.

5.1. A space decomposition of $V_p(\mathcal{T}_h)$. We first coarsen the triangulation $\{\Omega_k\}$ as follows: let Ω be decomposed into the union of $D_1, D_2, \cdots, D_{n_0}$ such that D_r is just the union of several elements in $\{\Omega_k\}$ and satisfies

$$D_r \bigcap D_l = \emptyset \quad for \quad r \ne l.$$

Then

$$\Omega = \bigcup_{r=1}^{n_0} D_r$$

is a non-overlapping domain decomposition of Ω. For convenience, we use $\mathcal{T}_h^r$ to denote the restriction of the triangulation $\mathcal{T}_h$ on the subdomain D_r $(r = 1, \cdots, n_0)$. Let $\{\phi_{kl}\}$ be the basis functions defined in Subsection 4.1. For $r = 1, \cdots, n_0$, define

$$V_p(\mathcal{T}_h^r) = span\left\{ \phi_{kl} : \ supp\ \phi_{kl} \subset D_r \right\}.$$

Let d denote the size of the subdomains $D_1, \cdots, D_{n_0}$, and let $\mathcal{T}_d$ denote the triangulation associated with the subdomains $D_1, D_2, \cdots, D_{n_0}$. For $r = 1, \cdots, n_0$, set $y_{rl}^d(\mathbf{x}) = e^{i\omega(\mathbf{x}\cdot\overrightarrow{\alpha_l})}$ $(\mathbf{x} \in D_r; l = 1, 2, \cdots, p)$. Define

$$\tilde{\phi}_{rl} = \begin{cases} y_{rl}^d, \text{ on } \Omega_k \text{ satisfying } \Omega_k \subset D_r, \\ 0, \text{ on } \Omega_k \text{ satisfying } \Omega_k \nsubseteq D_r \end{cases} \quad (r = 1, \cdots, n_0;\ l = 1, \cdots, p), \tag{30}$$

which are the basis functions of the space

$$\tilde{V}_p(\mathcal{T}_d) = span\Big\{\tilde{\phi}_{rl} : \ r = 1, \cdots, n_0; \ l = 1, \cdots, p\Big\}.$$

Then we have the space decomposition

$$V_p(\mathcal{T}_h) = \sum_{r=1}^{n_0} V_p(\mathcal{T}_h^r) + \tilde{V}_p(\mathcal{T}_d), \tag{31}$$

where the *coarse* space $\tilde{V}_p(\mathcal{T}_d)$, which is associated with the *coarse* triangulation $\mathcal{T}_d$, has the same structure with the original space $V_p(\mathcal{T}_h)$.

5.2. The preconditioner. Based on the space decomposition (31), we can construct the desired preconditioner by the general framework. For the purpose of implementation, we would like to describe the preconditioner in the algebraic form.

For $r = 1, \cdots, n_0$, let $\mathcal{A}_r$ denote the stiffness matrix induced from the sesquiinear form $a(\cdot, \cdot)$ on the subspace $V_p(\mathcal{T}_h^r)$. When the order of the basis functions $\{\phi_{kl}\}$ are arranged in a suitable manner, the original stiffness matrix $\mathcal{A}$ can be written as a block matrix with the diagonal sub-matrices $\mathcal{A}_1, \cdots, \mathcal{A}_{n_0}$. Set

$$\mathcal{D} = diag(\mathcal{A}_1, \cdots, \mathcal{A}_{n_0}).$$

In the following we define the coarse solver. To this end, we need to define a transformation matrix. With the basis functions defined in the last subsection, we define the transformation matrix $\mathcal{C}_d$ by

$$\begin{aligned}&(\tilde{\phi}_{11} \cdots \tilde{\phi}_{1p} \ \tilde{\phi}_{21} \cdots \tilde{\phi}_{2p} \cdots \tilde{\phi}_{n_0 1} \cdots \tilde{\phi}_{n_0 p})^t = \\ &\mathcal{C}_d(\phi_{11} \cdots \phi_{1p} \ \phi_{21} \cdots \phi_{2p} \cdots \phi_{N1} \cdots \phi_{Np})^t,\end{aligned} \tag{32}$$

where $\{\phi_{kl}\}$ are the basis functions of $V_p(\mathcal{T}_h)$ (see (19)). Let $\mathcal{A}_d$ denote the stiffness matrix induced from the sesquiinear form $a(\cdot, \cdot)$ on the coarse subspace $\tilde{V}_p(\mathcal{T}_d)$. As usual the matrix $\mathcal{A}_d$ is called a *coarse solver*. It is easy to verify that $\mathcal{A}_d = \mathcal{C}_d \mathcal{A} \mathcal{C}_d^t$.

Then the preconditioner associated with the space decomposition (31) is defined as

$$\mathcal{B}^{-1} = \mathcal{D}^{-1} + \mathcal{C}_d^t \mathcal{A}_d^{-1} \mathcal{C}_d, \tag{33}$$

which is the desired preconditioner for the original stiffness matrix $\mathcal{A}$. Since both $\mathcal{A}_r$ and $\mathcal{A}_d$ in general have much lower orders than the original stiffness matrix $\mathcal{A}$, the implementation of the action of $\mathcal{B}^{-1}$ is much cheaper than that of $\mathcal{A}^{-1}$.

6. Numerical experiments

In this section we report some numerical results to compare accuracies of the approximate solutions generated by the weighted variational formulation (11) (WVF), the original variational formulation (9) (VTCR) and the ultra weak variational formulation (UWVF). In order to compare the accuracies of the approximations reliably, we first solve the resulting systems by the direct method. Besides, we apply the preconditioned GMRES method with the preconditioners $\mathcal{B}$ and the simple GMRES method to solve the system (22) to illustrate the efficiency of the new preconditioner $\mathcal{B}$.

In the examples tested in this section, we adopt a uniform triangulation $\mathcal{T}_h$ for the domain Ω as follows: Ω is divided into some small cubes (for three-dimensional case), rectangles or triangles (for two-dimensional case) with the same size, where h denotes the length of the longest edge of the elements.

In order to determine the coarse solver $\mathcal{A}_d$ described in Subsection 5.2, we define subdomain D_r $(r = 1, \cdots, n_0)$ as follows: each subdomain (coarse element) D_r is a cube (for three-dimensional case) or rectangle (for two-dimensional case), which is just the union of several elements, and every subdomains D_r have the same size. Let d denote the length of the longest edge of the subdomains D_r. In general we can define $d \approx \sqrt{h}$ in the domain decomposition method. Here, for simplicity, we set $d = 4h$, namely, each subdomain D_r $(r = 1, \cdots, n_0)$ is the union of $4 \times 4 \times 4$ (4×4 for two-dimensional case) fine mesh elements.

To measure the accuracy of the numerical solution u_h, we introduce the following relative error:

$$\text{err.} = \frac{||u_{ex} - u_h||_{L^2(\Omega)}}{||u_{ex}||_{L^2(\Omega)}} .$$

We use the above relative L^2 error to measure the accuracy of the numerical solution u_h.

The stopping criterion in the iterative algorithms is that the relative L^2-norm ϵ of the residual of the iterative approximation satisfies $\epsilon < 1.0e-8$ for 2-D case and $\epsilon < 1.0e-6$ for 3-D cases (we choose initial guess $X_0 = 0$ in the iteration), and the maximum number of iteration steps $maxit$ satisfies $maxit = 5000$. Moreover, N_{iter} and T_{sol} represents the iteration numbers and computing time for solving the algebraic system respectively.

6.1. Wave propagation in a duct with rigid walls. The first model problem is the following Helmholtz equations for the acoustic pressure u and associated boundary conditions (see [12]):

$$\begin{aligned} &\Delta u + \omega^2 u = 0 \quad \text{in} \quad \Omega, \\ &\frac{\partial u}{\partial \mathbf{n}} + i\omega u = g \quad \text{on} \quad \partial\Omega, \end{aligned} \tag{34}$$

where $\Omega = [0, 2] \times [0, 1]$, and $g = (\frac{\partial}{\partial \mathbf{n}} + i\omega)u_{ex}$.

The exact solution to the problem can be obtained in the closed form as

$$u_{ex}(x, y) = \cos(k\pi y)(A_1 e^{-i\omega_x x} + A_2 e^{i\omega_x x})$$

where $\omega_x = \sqrt{\omega^2 - (k\pi)^2}$, and coefficients A_1 and A_2 satisfy the equation

$$\begin{pmatrix} \omega_x & -\omega_x \\ (\omega - \omega_x)e^{-2i\omega_x} & (\omega + \omega_x)e^{2i\omega_x} \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = \begin{pmatrix} -i \\ 0 \end{pmatrix} \tag{35}$$

The solution respectively represents propagating modes and evanescent modes when the mode number k is below the cut-off value $k \leqslant k_{\text{cut-off}} = \frac{\omega}{\pi}$ and up the cut-off value $k > k_{\text{cut-off}}$. For completeness, we compute approximate solutions for both the highest propagating mode and the lowest evanescent modes in the following tests.

We first consider the simplest choice of the parameters α and β in the new WVF method: $\alpha = \beta = 1$. The following table 1 give a comparison of error estimates of the approximations generated by the WVF method and the original VTCR method. The results listed in the above table indicate that the new WVF method can generate better approximations than the original VTCR even if we simply set $\alpha = \beta = 1$. But the simplest choice is not our interest, and the tests in the rest of the paper are made for the case $\alpha \neq \beta$, which may generate higher accuracy approximations.

As we pointed out in Section 3, we choose $\alpha = \omega^2$ and $\beta = 1$ in the variational equation (11) for this case with smooth solution. In order to determine the

TABLE 1. Errors of the approximations for the case $\omega = 20, p = 12$.

Methods		WVF ($\alpha = 1, \beta = 1$)	VTCR
h	k	err.	err.
$\frac{1}{12}$	6	8.59e-6	2.82e-5
	7	1.80e-4	2.06e-4
$\frac{1}{24}$	6	1.09e-7	2.78e-7
	7	2.71e-6	4.36e-6

number of plane wave basis functions per element, we numerically investigate the p-convergence (Figure 1 left) of three methods for the case of $\omega = 40, \omega h = 1, k = 13$ and the h-convergence (Figure 1 right) of the WVF method for the case of $\omega = 40, k = 13$. These plots in Figure 1 left highlight two different regimes of p for both the WVF method and the UWVF method when increasing p: a pre-asymptotic region with slow convergence and a post-asymptotic region of faster convergence. Moreover, for the smaller p, the accuracy for the three methods are not satisfactory. So in order to reach the high accuracy of the approximations with a suitable size of the problem, and give a fair comparison of the h-convergence of three methods, the number p of basis functions in each element is set to 12 in this subsection.

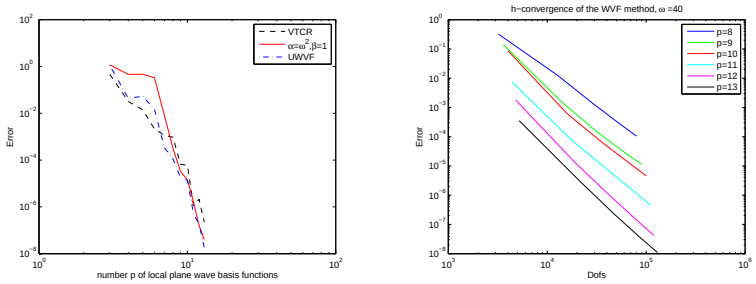

FIGURE 1. (left) p-convergence for the solution u_{ex} plotted against $p \in \{3, \cdots, 13\}$. (right) h-convergence of the WVF method.

Table 2, Table 3 and Figure 2 below show the h-convergence of the approximations generated by the method WVF, VTCR and UWVF, where the resulting systems (9) and (11) are solved by the direct method.

TABLE 2. Errors of the approximations for the case $\omega = 20$.

Methods		WVF ($\alpha = \omega^2, \beta = 1$)	VTCR	UWVF
h	k	err.	err.	err.
$\frac{1}{12}$	6	2.05e-6	2.82e-5	2.40e-6
	7	1.90e-5	2.06e-4	2.08e-5
$\frac{1}{16}$	6	3.35e-7	4.53e-6	3.91e-7
	7	3.14e-6	4.96e-5	3.47e-6
$\frac{1}{24}$	6	2.77e-8	2.78e-7	3.20e-8
	7	2.70e-7	4.34e-6	2.81e-7

TABLE 3. Errors of the approximations for the case $\omega = 40$.

Methods		WVF ($\alpha = \omega^2, \beta = 1$)	VTCR	UWVF
h	k	err.	err.	err.
$\frac{1}{20}$	12	6.35e-6	9.95e-5	7.53e-6
	13	1.27e-5	3.02e-4	1.42e-5
$\frac{1}{32}$	12	3.19e-7	5.33e-6	3.75e-7
	13	6.31e-7	1.46e-5	7.17e-7
$\frac{1}{40}$	12	8.03e-8	4.73e-7	9.41e-8
	13	1.59e-7	2.12e-6	1.79e-7

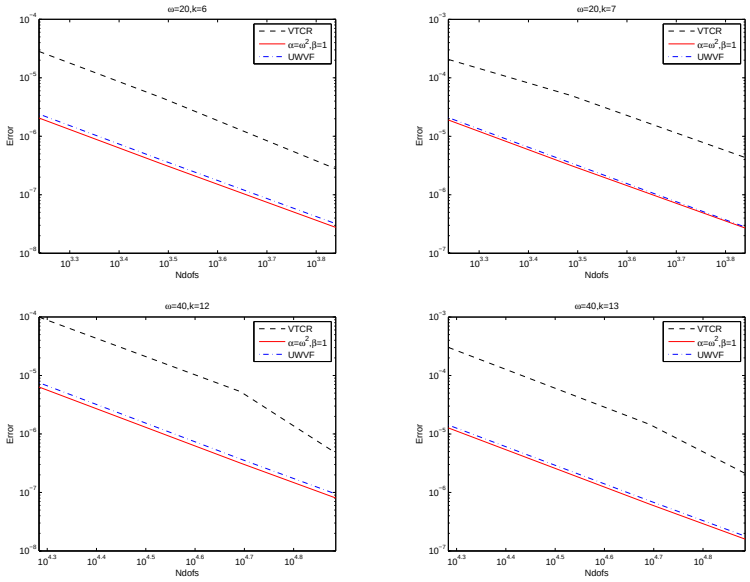

FIGURE 2. A comparison of the three strategies on the accuracy.

The results listed in the tables 2-3 indicate that the approximation generated by the new VWF method possesses much higher accuracy than the one generated by the original VTCR method, and has slightly higher accuracy than that generated by the UWVF method. Moreover, Figure 2 tells us that the new method is more stable with respect to h than the original method.

For simplicity, in the following we call the preconditioned GMRES method with the preconditioner $\mathcal{B}$ as the PGMRES method, and the simple GMRES method as the GMRES method, respectively. Tables 4-5 below give the iteration numbers, computing time and convergence of the PGMRES method and the simple GMRES method for solving the system generated by the WVF method ($\alpha = \omega^2, \beta = 1, p = 12$).

It can be seen, from the tables 4 - 5, that the iteration numbers and computing time of the PGMRES method are much more smaller than those of the GMRES method. Besides, in the two tables the iteration numbers and computing time of the PGMRES method increases more slowly when h decreases than that of the GMRES method. All these show that the proposed preconditioner $\mathcal{B}$ is very effective. It is interesting that the PGMRES method is less effective to the system generated by

TABLE 4. A comparison between two iterative algorithms for $\omega = 20$.

		PGMRES			GMRES		
h	k	N_{iter}	$T_{sol}(sec)$	err.	N_{iter}	$T_{sol}(sec)$	err.
$\frac{1}{12}$	6	73	6.08	2.06e-6	3169	2.46e+3	6.23e-6
	7	71	4.28	1.91e-5	3796	3.59e+3	2.68e-5
$\frac{1}{24}$	6	125	18.50	6.25e-8	5000+	1.22e+4	3.59e-5
	7	126	26.63	4.19e-7	5000+	1.87e+4	1.72e-4

TABLE 5. A comparison between two iterative algorithms for $\omega = 40$.

		PGMRES			GMRES		
h	k	N_{iter}	$T_{sol}(sec)$	err.	N_{iter}	$T_{sol}(sec)$	err.
$\frac{1}{20}$	12	69	6.59	6.35e-6	5000+	1.16e+4	1.37e-5
	13	67	7.01	1.27e-5	5000+	1.10e+4	6.01e-5
$\frac{1}{40}$	12	111	46.08	8.98e-8	5000+	5.14e+4	1.37e-4
	13	110	50.04	2.91e-7	5000+	6.09e+4	1.21e-4

the original VTCR method, which implies that the system generated by the VTCR method is indeed difficult to solve (when h decreases).

In most applications, the wave number ω may be large, so we would like to investigate how the errors of the approximate solutions depend on the wave number ω. Figure 3 highlight a comparison of the three strategies on the accuracy for fixed $\omega h = 1$ and variable ω. It is clear that three methods offer the same trend of error for fixed $\omega h = 1$ and increasing ω. But the WVF method provides the higher accuracy than other methods.

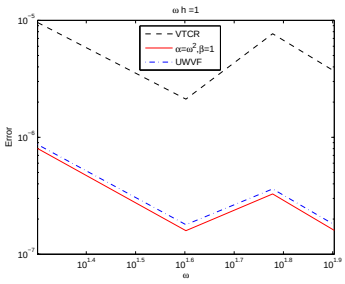

FIGURE 3. A comparison of of the three strategies on the accuracy for fixed $\omega h = 1$ and variable ω.

6.2. A smooth homogeneous problem in 3D. The second model problem is the following Helmholtz equations:

$$\Delta u + \omega^2 u = 0 \quad \text{in} \quad \Omega,$$
$$\frac{\partial u}{\partial \mathbf{n}} + i\omega u = g \quad \text{over} \quad \partial\Omega,$$

where $\Omega = [0,1] \times [0,1] \times [0,1]$, and $g = i\omega(1 + \vec{v_0} \cdot \mathbf{n})e^{i\omega \vec{v_0} \cdot \vec{x}}$.

The exact solution of the problem can be obtained in the closed form as

$$u_{ex}(\vec{x}) = e^{i\omega \vec{v_0} \cdot \vec{x}},$$

where $\vec{v_1} = (tan(-\pi/10), 0, tan(\pi/5))^t, \vec{v_0} = \vec{v_1}/||\vec{v_1}||_2$.

As in the last subsection, we also choose $\alpha = \omega^2$ and $\beta = 1$ in the variational equation (11) for this case. We numerically investigate the p-convergence of three methods for the case of $\omega = 10$ in order to determine the number of basis function in each element, see Figure 4. We observe that in order to reach the high accuracy, the suitable size of the problem and give a fair comparison of the h-convergence of three methods, the number p of basis functions in each element is set to 26 from this subsection.

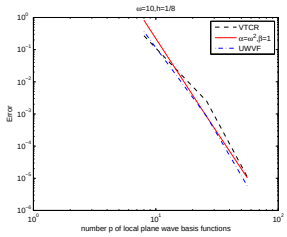

FIGURE 4. p-convergence for the solution u_{ex} plotted against $p \in \{8, 26, 58\}$.

Tables 6 - 7 and Figure 5 below give comparisons to the errors of the approximations generated by the three methods. Similarly to 2-D test, the corresponding systems (9) and (11) are solved by the direct method.

TABLE 6. Errors of the approximations for the case $\omega = 10$.

Methods	WVF $(\alpha = \omega^2, \beta = 1)$	VTCR	UWVF
h	err.	err.	err.
$\frac{1}{8}$	8.27e-4	2.50e-3	8.46e-4
$\frac{1}{12}$	1.42e-4	6.00e-4	1.43e-4
$\frac{1}{16}$	4.10e-5	1.89e-4	4.12e-5

TABLE 7. Errors of the approximations for the case $\omega = 20$.

Methods	WVF $(\alpha = \omega^2, \beta = 1)$	VTCR	UWVF
h	err.	err.	err.
$\frac{1}{20}$	3.58e-4	1.10e-3	2.73e-4
$\frac{1}{24}$	1.43e-4	5.65e-4	1.22e-4
$\frac{1}{28}$	6.88e-5	2.99e-4	6.23e-5

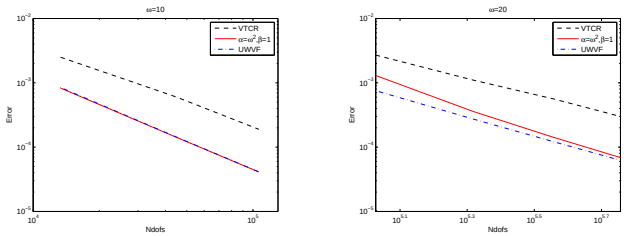

FIGURE 5. A comparison of the three strategies on the accuracy.

Similarly to the above 2-D test, the results listed in Tables 6 - 7 indicate that the approximation generated by the WVF method possesses much higher accuracy than the one generated by the original VTCR method, and has almost the same accuracy with the one generated by the UWVF method. Moreover, Fig.2 tells us that the new method is also more stable with respect to h than the original method.

Table 8 below lists iteration numbers, computing time and convergence of the PGMRES method and the simple GMRES method for solving the system generated by the WVF method ($\alpha = \omega^2, \beta = 1,\ p = 26$). It can be seen from Table 8 that

TABLE 8. A comparison between two iterative algorithms.

		PGMRES			GMRES		
ω	h	N_{iter}	$T_{sol}(sec)$	err.	N_{iter}	$T_{sol}(sec)$	err.
10	$\frac{1}{8}$	40	5.20	8.27e-4	943	3.64e+2	8.25e-4
	$\frac{1}{16}$	64	73.16	4.10e-5	2945	4.35e+4	9.69e-5
20	$\frac{1}{12}$	78	35.01	7.90e-3	1441	2.76e+3	7.90e-3
	$\frac{1}{24}$	56	1.50e+2	1.44e-4	5000+	6.35e+5	2.01e-4

the iteration number and computing time of the preconditioned GMRES method with the new preconditioner $\mathcal{B}$ increases more slowly when h decreases than that of the simple GMRES method. Both the iteration counts and the computing times in Table 8 indicate that the proposed preconditioner $\mathcal{B}$ is very effective.

The following Fig.6 describes real part and imaginary part of the exact solution and the numerical solution in the plane $z = 0.5$, respectively. It can be found from these figures that the approximate solutions computed by using the WVF method almost coincide with the analytic solution.

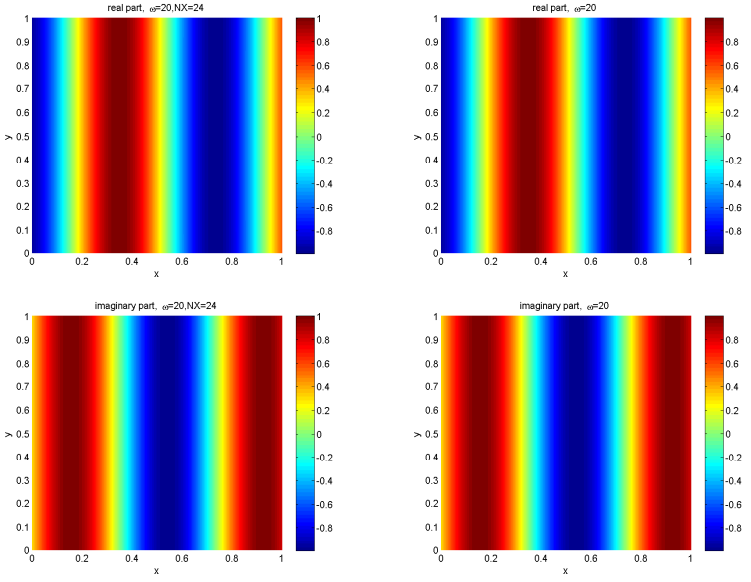

FIGURE 6. The first row is the real part of the exact solution and the numerical solution. The second row is the imaginary part of the exact solution and the numerical solution.

.

6.3. A point source problem. The following test problem consists of a point source and associated boundary conditions for homogeneous Helmholtz equations (see [13]):

$$u(r, r_0) = \frac{1}{4\pi} \frac{e^{i\omega|r-r_0|}}{|r-r_0|} \text{ in } \Omega,$$
$$\frac{\partial u}{\partial \mathbf{n}} + i\omega u = g \quad \text{over} \quad \partial\Omega,$$

in a cubic computational domain $\Omega = [-1,1] \times [-1,1] \times [-1,1]$ centred at the origin. The location of the source is off-centred at $r_0 = (1,1,1)$ and $r = (x,y,z)$ is an observation point. Note that the analytic solution of the Helmholtz equation with such boundary condition has a singularity at $r = r_0$.

Since the analytic solution is singular, the jump $\partial_{\mathbf{n}_k} u_k + \partial_{\mathbf{n}_j} u_j$ may not be small on Γ_{kj}. Thus, we choose $\alpha = \omega$ and $\beta = \frac{1}{\omega}$ in the variational equation (11). Table 9, Table 10 and Figure 7 below give comparisons to the errors of the approximations generated by the WVF method, the VTCR method and the UWVF method respectively, where the resulting systems are solved by the direct method.

TABLE 9. Errors of the approximations for the case $\omega = 10$.

Methods	WVF $(\alpha = \omega, \beta = \omega^{-1})$	VTCR	UWVF
h	err.	err.	err.
$\frac{1}{6}$	1.14e-1	1.54e+1	1.16e-1
$\frac{1}{8}$	1.11e-1	3.22e+1	1.14e-1
$\frac{1}{12}$	1.10e-1	—	1.13e-1

TABLE 10. Errors of the approximations for the case $\omega = 20$.

Methods	WVF $(\alpha = \omega, \beta = \omega^{-1})$	VTCR	UWVF
h	err.	err.	err.
$\frac{1}{8}$	1.58e-1	2.06e+1	9.24e-2
$\frac{1}{12}$	8.53e-2	—	8.48e-2
$\frac{1}{16}$	6.18e-2	—	8.33e-2

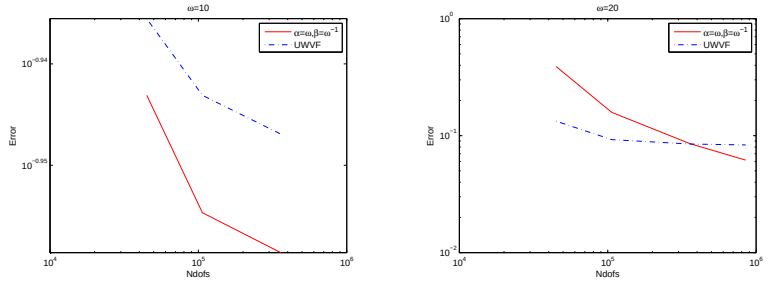

FIGURE 7. A comparison of the three strategies on the accuracy.

The results listed in Tables 9-10 indicate that, although the VTCR method is completely invalid to solve such point source problem, the approximation generated

by the WVF method can also achieve a certain accuracy for such singularity problem, and possesses slightly higher accuracy than the one generated by the UWVF method when h decreases. Since the analytic solution $u(r, r_0)$ has singularity, the approximation generated by the new method does not converges uniformly with respect to h.

Table 11 below lists iteration numbers, computing time and convergence of the PGMRES method and the simple GMRES method for solving the system generated by the WVF method ($\alpha = \omega, \beta = \omega^{-1}$, $p = 26$).

TABLE 11. A comparison between two iterative algorithms.

		PGMRES			GMRES		
ω	h	N_{iter}	$T_{sol}(sec)$	err.	N_{iter}	$T_{sol}(sec)$	err.
10	$\frac{1}{4}$	58	6.21	1.28e-1	1424	1.28e+3	1.28e-1
	$\frac{1}{8}$	38	27.52	1.12e-1	5000+	2.24e+5	1.13e-1
20	$\frac{1}{6}$	77	30.57	3.90e-1	847	1.71e+3	3.90e-1
	$\frac{1}{12}$	81	2.71e+2	8.71e-2	5000+	8.57e+5	9.12e-2

Similarly to the smooth 3-D test, we can see from the above table that the proposed preconditioner $\mathcal{B}$ is very effective.

The following Figure 8 describes real part and imaginary part of the exact solution and the numerical solution in the plane $z = 0.5$, respectively. It can be found from these figures that the approximate solutions computed by using the WVF method almost coincide with the analytic solution.

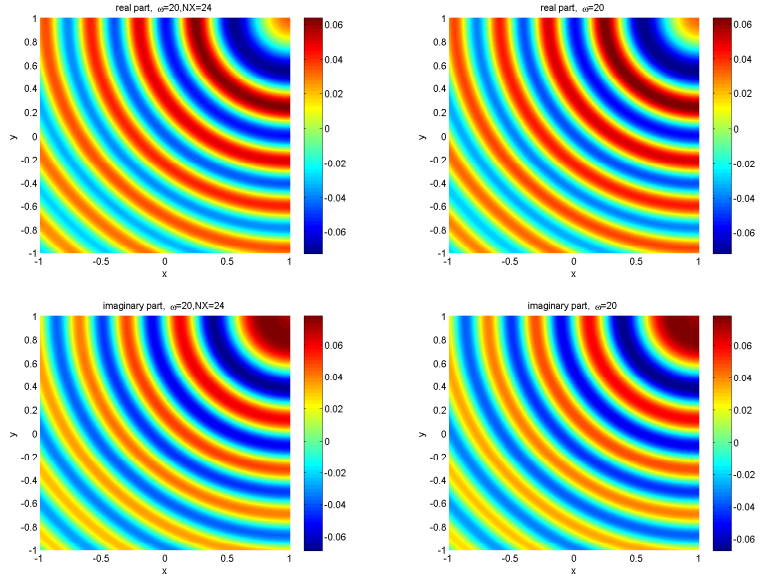

FIGURE 8. The first row is the real part of the exact solution and the numerical solution. The second row is the imaginary part of the exact solution and the numerical solution.

.

7. Conclusion

In this paper we have put forward a weighted variational formulation (WVF) for both two-dimensional and three-dimensional Helmholtz equations, inspired by the variational theory of complex rays (VTCR, see [16]). We proved a L^2 error estimate of the approximate solution generated by the discrete WVF method. The all numerical examples show that the approximations generated by the WVF method possess much higher accuracies than that generated by the original VTCR method, and have almost the same accuracies with that generated by the UWVF method (note that the WVF method is easier to implement than the UWVF method), provided that the weighes α and β are chosen in a suitable manner. Moreover, the new method is more stable with respect to mesh size h than the original VTCR method. Besides, we have introduced a simple domain decomposition preconditioner for the system arising from the WVF method for Helmholtz equations. We found that the preconditioned GMRES method with such preconditioner is an effective method for solving such system, in both two-dimensional and three-dimensional cases.

References

[1] I. Babuska, F. Ihlenburg, E. Paik and S. Sauter, A generalized finite element method for solving the helmholtz equation in two dimensions with minimal pollution, Comput. Meth. Appl. Mech. Engng., 128(1995), pp. 325-359.

[2] M. Bonnet, S. Chaillat and J. Semblat, A multi-level fast multipole BEM for 3-D elastodynamics in the frequency domain, Computer Methods in applied Mechanics and Engineering, 197(2008), pp. 4233-4249.

[3] O. Cessenat and B. Despres, Application of an ultra weak variational formulation of elliptic pdes to the two-dimensional helmholtz problem, SIAM J. Numer. Anal. 35(1998), No.1, pp.255-299.

[4] Z. Chen, W. Kreuzer, H. Waubke, A burton-miller formulation of the boundary element method for baffle problems in acoustics and the BEM/FEM coupling, Engineering analysis with boundary elements, 35(2011), No. 3, pp. 279-288.

[5] A. Deraemaeker, I. Babuska and P. Bouillard, Dispersion and pollution of the FEM solution for the Helmholtz equation in one, two and three dimensions, Int. J. Numer. Meth. Engng., 46(1999), pp. 471-499.

[6] C. Farhat, I. Harari and L. Franca, The discontinuous enrichment method, Comput. Meth. Appl. Mech. Engng., 190(2001), pp.6455-6479.

[7] P. Gamallo, R. Astley, A comparison of two Trefftz-type method: the ultra weak variational formulation and the least-squares method, for solving shortwave 2-D Helmholtz problems, Int. J. Numer. Meth. Engng., 71(2007), Issue 4, pp. 406-432.

[8] C. Gittelson, R. Hiptmair and I. Perugia, Plane wave discontinuous Galerkin methods: Analysis of the h-version, ESAIM: M2AN Math. Model. Numer. Anal., 43(2009), pp. 297-331.

[9] P. Grisvard, Elliptic problems in nonsmooth domains, Monogr. Stud. Math. 24, Pitman, Boston, 1985.

[10] I. Harari and T. Hughes, A cost comparison of boundary element and finite element methods for problems of time-harmonic acoustics, Comput. Meth. Appl. Mech. Engng., 97 (1992), pp.77-102.

[11] R. Hiptmair, A. Moiola, and I. Perugia, Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the p-version, Tech. Rep. 2009-05, SAM-ETH zürich, 2009.

[12] T. Huttunen, P. Gamallo and R. Astley, Comparison of two wave element methods for the Helmholtz problem, Commun. Numer. Meth. Engng., 25(2009), pp. 35-52.

[13] T. Huttunen, J. Kaipio and P. Monk. The perfectly matched layer for the ultra weak variational formulation of the 3D Helmholtz equation, Int. J. Numer. Meth. Engng., 61(2004), pp. 1072-1092.

[14] L. Kovalevsky, P. Ladevèze, H. Riou, The Fourier version of the variational theory of complex rays for medium-frequency acoustics. Computer Methods in Applied Mechanics and Engineering, 225-228(2012), No. 0, pp. 142-153.

[15] P. Ladevèze, A new computational approach for structure vibrations in the medium frequency range, Comptes Rendus Académie des Sciences Paris. 322(IIb) (1996), pp.849-856.
[16] P. Ladevèze, L. Arnaud, P. Rouch and C. Blanzé, The variational theory of complex rays for the calculation of medium-frequency vibrations, Engng. Comput., 18(2001), pp. 193-214.
[17] A. Legay, An extended finite element method approach for structural-acoustics problems involving immersed structures at arbitrary positions, Int. J. Meth. Engng, 93(2013), No. 4, pp. 376-399.
[18] A. Moiola, R. Hiptmair and I. Perugia, Plane wave approximation of homogeneous Helmholtz solutions, Z. Angew. Math. Phys. 62(2011), 809-837.
[19] P. Monk and D. Wang, A least-squares method for the helmholtz equation, Comput. Meth. Appl. Mech. Engng., 175(1999), pp.121-136.
[20] H. Riou, P. Ladevèze, B. Sourcis, The multiscale VTCR approach applied to acoustics problems, J. Comput. Acous., 16(2008), No. 4, pp. 487-505.
[21] H. Riou, P. Ladevèze, B. Sourcis, B. Faverjon and L. Kovalevsky, An adaptive numerical strategy for the medium-frequency analysis of Helmholtz's problem, J. Comput. Acous., 20(2012), No. 1, DOI: 10.1142/S0218396X11004481.
[22] R. Rumpler, J. Deue, P. Goransson, A modal-based reduction method for suound absorbing porous materials in poro-acoustic finite element models, J. Acoust. Soc. Am, 132(2012), No. 5, pp. 3162-3179.
[23] C. Soize, Reduced models in the medium frequency range for the general dissipative structural dynamic systems, Euro. J. Mech. A/Solids, 17(1998), pp. 657-685.
[24] R. Tezaur, C. Farhat, Three-dimensional directional discontinuous Galerkin elements with plane waves and Lagrange multipliers for the solution of mid-frequency Helmholtz problems, Int. J. Numer. Meth. Engng., 66(2006), pp. 796-815.
[25] E. Trefftz, Ein gegenstuck zum ritzschen verfahren, Sec. Inte. Cong. Appl. Mech., (1926), pp.131-137.

LSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

E-mail: hqy@lsec.cc.ac.cn and yuanlong@lsec.cc.ac.cn