Subspace Iteration Search Method for Generalized Eigenvalue Problems with Sparse Complex Unsymmetric Matrices in Finite-Element Analysis of Waveguides

Javier Arroyo and Juan Zapata, *Member, IEEE*

***Abstract*— In this paper, a numerical method for the robust computation of the number of eigenvalues within a closed contour of a generalized complex eigenvalue problem is presented. As a result of this computation, it is possible to perform a systematic search for the eigenvalues, ensuring that no eigenvalues are forgotten, and to optimize their calculation. Application is made to the finite-element modal analysis of inhomogeneous waveguides.**

***Index Terms*— Algorithms, eigenvalues/eigenfunctions, finite-element methods, nonhomogeneously loaded waveguides, poles, zeros.**

I. INTRODUCTION

A LARGE class of engineering problems can be formulated in such a way that the solution is numerically obtained from a generalized matrix eigenvalue problem. The matrices involved are usually large and sparse and can be complex and nonsymmetric in the most general case. A sparse matrix solver for the efficient solution of this class of completely arbitrary generalized eigenvalue problems has been described in [1]–[3], in the context of using the finite-element method for the modal characterization of dielectric waveguides with symmetry of translation. The solver is based upon the inverse subspace iteration method, in which a whole subspace is (inversely) iterated to simultaneously yield a (small) set of eigenvalues and eigenvectors.

The generalized eigenvalue problem arising from the finite-element analysis of a waveguide has the form

$$AX = BX\Lambda \tag{1}$$

or, equivalently,

$$(A-\mu B)X = BX\Gamma \tag{2}$$

where A, B are non-Hermitian matrices, Γ and Λ are diagonal matrices of eigenvalues, μ is a desired shift, and X is the matrix of right eigenvectors. All these matrices and the shift will, in general, be complex. The shift μ is introduced in order to force the convergence of the algorithm toward a cluster of desired eigenvalues centered at μ in the complex plane. The eigenvalues of (1) and (2) are related by

$$\Lambda = \Gamma + \mu I. \tag{3}$$

Let the dimension of the matrices be $N \times N$, p the number of eigenvectors we want to obtain, and q the dimension of the subspace of vectors to be iterated, with $p < q \ll N$. It has been suggested in [5] the heuristic choice

$$q = \min(2p, p+8). \tag{4}$$

The iterations start with the selection of a set of q trial vectors X_0 and proceed with the computation of the subsequent vectors using the recursion

$$(A-\mu B)X_{s+1} = BX_s \tag{5}$$

followed by a suitable B-normalization. The order of the problem is reduced by carrying out a Rayleigh–Ritz analysis. Defining the transformations

$$\begin{aligned} A_{s+1} &= X_{s+1}^T(A-\mu B)X_{s+1} \\ B_{s+1} &= X_{s+1}^T BX_{s+1} \end{aligned} \tag{6}$$

the dense eigenvalue problem of order q is formed and solved

$$A_{s+1}\Phi_{s+1} = B_{s+1}\Phi_{s+1}\Gamma_{s+1}. \tag{7}$$

The estimation of the eigenvectors is then

$$X = X_{s+1}\Phi_{s+1} \tag{8}$$

and the convergence is achieved when the p-shifted eigenvalues of smallest absolute value verify

$$\begin{gathered} \frac{|\gamma_{s+1}^i - \gamma_s^i|}{|\gamma_{s+1}^i|} \le \varepsilon, \qquad i = 1, \cdots, p \\ |\gamma_{s+1}^1| \le |\gamma_{s+1}^2| \le \cdots \le |\gamma_{s+1}^p| \le |\gamma_{s+1}^{p+1}| \le \cdots \le |\gamma_{s+1}^q|. \end{gathered} \tag{9}$$

The convergence rate is governed by the quotient

$$\frac{|\gamma^p|}{\min\limits_{q<i\le N}|\gamma^i|} \tag{10}$$

which justifies the election of a subspace dimension higher than the actual number of desired eigenvalues.

As it stands, the method can be very powerful, provided that one is able to make the correct decisions as to what values should be given to the different parameters involved; the number of desired eigenvalues p especially.

The interest of the authors in the solver is the same interest that led to its development, i.e., the finite-element computation of the propagation constants of the modes of

Manuscript received September 19, 1997; revised January 26, 1998. This work was supported by CICYT, Spain, under Contract TIC95-0137-C02-01.

The authors are with the Departamento de Electromagnetismo y Teoría de Circuitos, Universidad Politécnica de Madrid, 28040 Ciudad Universitaria S/N, Madrid, Spain.

Publisher Item Identifier S 0018-9480(98)05508-2.

closed waveguides with symmetry of translation, filled with lossy, inhomogeneous, anisotropic, or bi-anisotropic media. The following discussion will focus on this specific application of the solver, but, in general, is valid.

A pure finite-element method can tackle problems where (e.g., due to their geometrical complexity) no analytical method is applicable, but conversely it provides no analytical insight of the propagation phenomenon or the distribution of the modal spectrum. Consequently, there is an enormous practical difference between the cases when one is trying to check their results against previously existing data from when one is conducting an unprecedented investigation. In the former case, it is quite straightforward to choose convenient complex shifts and the right dimension for the iterated subspace each time. Besides, the total number of eigenvalues to be found is obvious from the available graph. However, in the latter case, one has to perform a *true search* for the eigenvalues, involving not only their location, but their number as well, which is unknown in advance. The standard theorems cannot be invoked since the matrices hold no special properties of definiteness or symmetry. This is in contrast with lossless three-dimensional (3-D) eigenvalue problems in which it is indeed possible to make use of the properties of Sturm sequences to efficiently compute this number for closed segments of the real axis [4].

The uncertainty about the number of eigenvalues which must be calculated is twofold: uncertainty about the total number of eigenvalues within a certain large region or with some specific properties (e.g., the number of modes below cutoff or with an attenuation constant less than a given threshold) and uncertainty about the number of eigenvalues in a region surrounding a complex shift, i.e., uncertainty about the choice of p. Under these circumstances, there is the potential risk of inadvertently missing some modes of the waveguide, something which would rend the analysis useless for certain practical purposes, such as the use of the numerical modes for mode-matching applications. Additionally, the convergence of the iterations may fail if one attempts to compute an inappropriate number of eigenvectors. Moreover, these difficulties are exacerbated by the use of functionals with extraneous nonphysical modes associated to the zero eigenvalue, which then reaches a multiplicity of hundreds for an average-size grid. It is the presence of this multiple eigenvalue that makes it harder to locate and compute those eigenvalues that lie very near the origin than those farther away from it. In general, multiple physical eigenvalues or clusters of very close eigenvalues may also be a source of trouble, as it becomes more difficult to ascertain that all of the eigenvalues have been found.

The aim of this paper is to describe a way of performing a systematic search for the eigenvalues in the complex plane, fixing the shortcomings of the subspace iteration method described in the preceding paragraph. To do this, it is essential to compute the number of eigenvalues enclosed in a given contour before their finding is attempted. The convergence of the eigenvalues depends on their distance from the shift, a fact that suggests that the contour be a circumference centered at the selected shift. The dimension of the subspace would be given accordingly to the known number of eigenvalues to be found or, in case where this was too high, the circle could be subdivided into smaller slightly overlapping circular regions where the procedure would be recurrently repeated. The method of computing the number of eigenvalues lying within a circle of given radius which we present is quite straightforward from the theoretical point of view, but it turns out to be preferable to other methods which can be found in the mathematical literature, as shall be explained.

II. Computation of the Number of Eigenvalues

Let λ_0 be a complex number. λ_0 is an eigenvalue of (1), if and only if

$$f(\lambda)|_{\lambda=\lambda_0} = |A - \lambda_0 B| = 0 \tag{11}$$

where $f(\lambda)$ is the characteristic polynomial of degree N defined as the determinant $f(\lambda) = |A - \lambda B|$. Let C be a closed contour on the λ plane, such that there are no zeros of $f(\lambda)$ on the contour. The number of zeros ν in the interior of C is equal to the contour integral

$$\nu = \frac{1}{2\pi j} \oint_C \frac{f'(\lambda)}{f(\lambda)}\, d\lambda. \tag{12}$$

The mapping $\omega = f(\lambda)$ maps the contour C into the closed curve $\overline{C}$. Equation (12) can be put as

$$\begin{aligned}\nu &= \frac{1}{2\pi j} \oint_C \frac{f'(\lambda)}{f(\lambda)}\, d\lambda \\ &= \frac{1}{2\pi j} \oint_C \frac{d[\ln f(\lambda)]}{d\lambda}\, d\lambda \\ &= \frac{1}{2\pi j} \oint_{\overline{C}} d[\ln \omega]\end{aligned} \tag{13}$$

where $\ln\, \omega = \ln |\omega| + j \text{ phase } \omega$, which immediately leads to the well-known result that ν is simply the number of times that $\overline{C}$ encircles the origin.

In theory, there are several ways of computing ν. In [6], a thorough discussion of the subject of locating the zeros of an arbitrary analytic function can be found, including the computation of its number as a preliminary step. A successful direct application of the methods expounded in the reference can be found, for example, in [7]. However, on account of the peculiarities of the function $f(\lambda)$, all of the proposed methods based upon (12) have to be discarded, for the following reasons.

1) Only the function and not its derivative is available. Thus, the direct computation of the contour integral (12) has to be ruled out. Another possibility would be to estimate the derivative directly from the computed values of f in the integration points of some quadrature rule, as explained in [6]. However, this method does not work because of 2).
2) The value of $f(\lambda)$ cannot always be represented using either REAL or even DOUBLE PRECISION arithmetic. In practical cases, it is perfectly possible to find absolute values as low as $1.E - 3000$ or as high as $1.E + 2000$, e.g., see 3).

3) Even worse, the dynamic range of the values can easily outrun the dynamic range of a REAL and, in some cases, a DOUBLE PRECISION number. That means that normalized quantities may be almost as unwieldy as the unnormalized ones. Moreover, even if normalization is possible, these extremely strong variations cause the numerical integration to give absolutely disastrous results for any reasonable or feasible number of integration points.

The numerical experiments that we have carried out show that the only practical way of tackling this problem is through the direct computation of the number of times that the closed contour $\overline{C}$ encircles the origin of the ω plane. To perform this computation, only the phase of the determinants is relevant. Assume that the contour C can be described as a continuous function of a real parameter θ

$$C \equiv \{\lambda \in \mathbb{C}/\lambda = \lambda(\theta), \theta \in \mathbb{R}, \theta_a \le \theta \le \theta_b\} \tag{14}$$

with the condition $\lambda(\theta_a) = \lambda(\theta_b)$. For example, if $C \equiv \{\lambda/|\lambda - \mu| = r\}$ then $\lambda(\theta) = \mu + re^{j\theta}, 0 \le \theta \le 2\pi$. Now, if θ varies continuously from θ_a to θ_b, the phase of $f(\theta) = f(\lambda(\theta))$ will vary continuously from $\mathrm{phase}[f(\theta_a)]$ to $\mathrm{phase}[f(\theta_b)]$ and the number of loops around the origin is

$$\begin{aligned}\nu &= \frac{1}{2\pi}(\mathrm{phase}[f(\theta_b)] - \mathrm{phase}[f(\theta_a)]) \\ &= \frac{1}{2\pi}(\mathrm{phase}[\omega_b] - \mathrm{phase}[\omega_a]).\end{aligned} \tag{15}$$

The difficulty with (15) is that we cannot compute the continuous function $\mathrm{phase}[\omega]$ directly. Instead, we have

$$\mathrm{phase}[\omega] = \mathrm{atan2}(\mathrm{imag}(\omega), \mathrm{real}(\omega)) + 2\pi n \tag{16}$$

where the function $\mathrm{atan2}$ is the standard FORTRAN intrinsic function, returning an argument between $-\pi$ and π; the integer n has to be increased or decreased by one each time the ω-contour cuts the negative real axis in the counterclockwise direction or the clockwise direction, respectively, compensating for the jump discontinuity of the argument returned by the $\mathrm{atan2}$ function.

If we sample the values ω_i, ω_{i+1} of the function for two "close" points $\lambda_i = \lambda(\theta_i)$, $\lambda_{i+1} = \lambda(\theta_{i+1})$ on the contour C the correction n_{i+1} should be given the value which makes the phase difference smaller, leading to the "unwrapping" rule [6]

$$\left.\begin{aligned}\mathrm{phase}[\omega_i] &= \mathrm{atan2}(\omega_i) + 2\pi\hat{n}_i \\ \mathrm{phase}[\omega_{i+1}] &= \mathrm{atan2}(\omega_{i+1}) + 2\pi\hat{n}_{i+1}\end{aligned}\right\} \Rightarrow$$

$$\hat{n}_{i+1} = \begin{cases}\hat{n}_i + 1, & \mathrm{atan2}(\omega_{i+1}) - \mathrm{atan2}(\omega_i) \le -\pi \\ \hat{n}_i, & -\pi < \mathrm{atan2}(\omega_{i+1}) - \mathrm{atan2}(\omega_i) \le \pi \\ \hat{n}_i - 1, & \pi < \mathrm{atan2}(\omega_{i+1}) - \mathrm{atan2}(\omega_i)\end{cases} \tag{17}$$

where $\hat{n}_i$ and $\hat{n}_{i+1}$ are the estimations of the true factors. The two sample points can be defined to be close precisely if (17) gives the correct answer for $\mathrm{phase}[\omega_{i+1}]$, provided $\hat{n}_i$ is correct ($\hat{n}_i = n_i$). The unwrapping rule will work correctly when the straight-line segment connecting the couple of consecutive points (ω_i, ω_{i+1}) cuts the negative real axis the same number of times (i.e., one or not all) as the corresponding curved segment of the contour $\overline{C}$ [6]. Using this rule, a first basic algorithm for the computation of the number of zeros would be as follows.

1) Sample the contour C at the points

$$\lambda_i = \lambda(\theta_i), \theta_i = \theta_a + i\Delta\theta, \Delta\theta = \frac{(\theta_b - \theta_a)}{N_p}, \qquad i = 0, \cdots, N_p.$$

2) Compute the arguments $\mathrm{atan2}(\omega_i) = \mathrm{atan2}(|A - \lambda_i B|)$. The arguments are computed as $\Sigma_{i=1}^{N}\ \mathrm{atan2}(u^i_{jj})$ (adequately shifted to the interval $(-\pi, \pi]$) where the $u^{i\prime}_{jj}s$ are the diagonal elements of the matrix U^i resulting from the LU factorization of the matrix $A - \lambda_i B$: $L^i U^i = A - \lambda_i B$.
3) Compute the phases $\mathrm{phase}[\omega_i]$ using (17).
4) Compute the number of zeros using (15).

The bulk of the numerical burden of this algorithm is attached to performing the LU factorizations of 2). The computation time grows rapidly with the dimension of the matrices and it is essential to keep the number of computed factorizations as low as possible. The minimum number of sample points is given by the number below which the unwrapping rule starts to fail.

Obviously, a major shortcoming of the algorithm is that the necessary number of samples is unknown. One could progressively refine the discretization of the contour, computing the estimated number of zeros for each discretization, and accepting the result when the convergence had apparently been attained. However, this approach is neither efficient nor reliable. Clearly, a uniform sampling scheme is inadequate for this problem because the phase variations (the "bandwidth" of the phase) can be arbitrarily fast since there is always the possibility of finding several simple or multiple zeros arbitrarily near the contour C. In theory, if a sample point λ_i happens to lie sufficiently near a too-close zero, the matrix $A - \lambda_i B$ will become numerically singular and the LU factorization will fail, signaling the presence of the zero, but this is a random and improbable event that will seldom take place; most often the phase will go undersampled and an incorrect number of zeros will be computed, even though most of the contour will have been vastly oversampled.

It is clear from the above discussion that the optimum sampling scheme would be a variable step-size sampling scheme. The step size should vary on a sample by sample basis, reflecting the changes of the derivative of the phase. The steps should be biggest where the phase is varying more predictably (i.e., approximately linearly with respect to the parameter θ) and smallest where the phase is rapidly changing its rate of growth. The algorithm that we propose is based on the prediction of the value $\mathrm{phase}[\omega_{i+1}]$ from the previously estimated values $\mathrm{phase}[\omega_j], j = i - k, \cdots, i$, where k is the order of the predictor. If the predicted value $\hat{p}[\omega_{i+1}]$ and the computed value differ only tolerably, the estimation is accepted and a new step size is calculated for the next sample. If, instead, the difference is greater than a tolerance, the iteration is repeated with a smaller step size.

The computation of the number of zeros would proceed with the following steps:

```
Step 1. compute $\mathrm{atan2}[\omega_i], i = 0, \cdots, k$, with $\Delta\theta = \Delta\theta_{\min}$;
Step 2. use (17) to compute $\mathrm{phase}[\omega_i], i = 0, \cdots, k$
loop $i = k, \cdots$;
Step 3. $\theta_{i+1} = \theta_i + \Delta\theta$
        if $\theta_{i+1} > \theta_b$ then $\theta_{i+1} = \theta_b; \Delta\theta = \theta_b - \theta_i$ end if;
Step 4. compute the prediction $\hat{p}[\omega_{i+1}]$ from
        $\mathrm{phase}[\omega_j], j = i - k, \cdots, i$;
Step 5. compute $\mathrm{atan2}[\omega_{i+1}]$ and $\mathrm{phase}[\omega_{i+1}]$ using (17),
        with the prediction $\hat{p}[\omega_{i+1}]$ as the reference;
Step 6. compare the error $\varepsilon_{i+1} = |\mathrm{phase}[\omega_{i+1}] - \hat{p}[\omega_{i+1}]|$
        with the tolerances and branch the execution accordingly:

    if $\varepsilon_{i+1} \leq \varepsilon_{\mathrm{acc}}$ then
      /* error is OK   */
      if $\theta_{i+1} = \theta_b$ then
          /* successful end */
          $\nu = \frac{1}{2\pi}(\mathrm{phase}[\omega_{i+1}] - \mathrm{phase}[\omega_0])$
          return
      else
          /* size-step for new sample point */
          $\Delta\theta = \Delta\theta * s\left(\frac{\varepsilon_{\mathrm{acc}}}{\varepsilon_{i+1}}\right)$
          $i = i + 1$
      end if
    else if $\Delta\theta > \Delta\theta_{\min}$ then
      if $\varepsilon_{i+1} \leq \tilde{\varepsilon}_{\mathrm{acc}}$ then
          /* error model still OK:
          reduce size-step to bring error
          below the acceptable tolerance */
          $\Delta\theta = \Delta\theta * \tilde{s}\left(\frac{\varepsilon_{\mathrm{acc}}}{\varepsilon_{i+1}}\right)$
      else
          /* the error model is not working:
          abrupt decrease of step-size */
          $\Delta\theta = \max\left(\frac{\Delta\theta}{m}, \Delta_{\min}\right)$
          re-sample the previously accepted $k$
          points before the last
              with new step-size:
              compute $\mathrm{atan2}(\omega_{i-j})$ with
             $\theta_{i-j} = \theta_i - j\Delta\theta, j = 1, \cdots, k$
              compute $\mathrm{phase}[\omega_{i-j}]$ with the new
             $\hat{p}[\omega_{i-j}], j = 1, \cdots, k$ as reference
      end if
    else
      /* step-size has become too small:
      restart iterations from the beginning
      for a new contour $C$ */
              choose new parameters for the new
              contour
              go to 1)
      end if
    end loop.
```

Some of the steps of this variable step-size algorithm deserve further discussion. The prediction $\hat{p}[\omega_{i+1}]$ is simply the value of the interpolating polynomial that fits the previous $k+1$ estimated phases. Since the polynomial is used for extrapolation purposes rather than interpolation, its order cannot be too high, because the error can grow very fast outside the interval $[\theta_{i-k}, \theta_i]$ for high-order polynomials; a good choice is a third-order polynomial. Besides, the lower the order of the polynomial, the more affordable the resetting of the predictor and the error model becomes. The phase reference for the unwrapping is the predicted value rather than the previous estimated phase. This makes it possible to skip over whole loops around the origin in a single step whenever the phase variation is smooth enough for the prediction error to keep within the acceptable bounds. It is especially advantageous when the contour C encloses a very high number of zeros, particularly when the multiple null eigenvalue $\lambda = 0$ (a multiplicity ranging the hundreds) of certain functionals is "searched," for the reasons that will be explained in the context of the practical application of the method.

One aspect that greatly affects the performance of the algorithm is the sound choice of the tolerances $\varepsilon_{\mathrm{acc}'}\tilde{\varepsilon}_{\mathrm{acc}}$ and the minimum step $\Delta\theta_{\min}$. The values of $\varepsilon_{\mathrm{acc}}$ and $\Delta\theta_{\min}$ have to be chosen regarding the level of numerical noise of the computed arguments; the rule is that one should not try to predict the phase with greater accuracy than the actual accuracy of the arguments computed from the LU factorization, neither should the step size be so small that the discrete nature of the underlying arithmetic becomes evident. The upper bounds of the parameters should attend to the need of preserving the robustness of the algorithm while minimizing the number of computed determinants. The objective of the algorithm is to resolve the ambiguity of the arguments returned by the $\mathrm{atan2}$ function. The prediction errors are unimportant from the standpoint of the accuracy of the reconstructed phase as long as the ambiguity is correctly resolved; besides, these errors are not cumulative because actually the true values of the phase and not the predicted ones are used for the subsequent predictions.

Obviously, the magnitude of the error is an indicator of the performance of the predictor, which can be thought to be in one of the following states regarding the tracking of the phase.

Locked: The error model is working well. The calculated step size $\Delta\theta = \Delta\theta * s(\varepsilon_{\mathrm{acc}}/\varepsilon_{i+1})$ from the error model allows an accurate prediction.

Partially locked: The calculated step size has resulted in an inaccurate prediction, although the ambiguity of the phase is actually correctly resolved. The error is small enough to estimate a new smaller step size $\Delta\theta = \Delta\theta * \tilde{s}(\varepsilon_{\mathrm{acc}}/\varepsilon_{i+1})$ which should bring the predictor back to the locked state again.

Unlocked: The error is unacceptable, which indicates that the error model is not working, due to an abrupt variation, compared with the present scale of sampling, of the rate of change of the phase. An

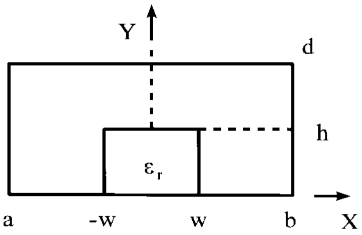

Fig. 1. Cross section of a shielded dielectric image guide.

attempt to force the predictor back to the locked state is made by resetting the filter, resampling backward the phase function with a strongly reduced step size $\Delta\theta = \max((\Delta\theta/m), \Delta_{\min})$.

The calculation of each new step size is based on the assumption that the dependence of the prediction error on the step size is of the form

$$\varepsilon(\theta; \Delta\theta) \approx c(\theta)(\Delta\theta)^{k+1} \tag{18}$$

from which we have

$$s\left(\frac{\varepsilon_{\rm acc}}{\varepsilon_{i+1}}\right) = \min\left(c_s\left(\frac{\varepsilon_{\rm acc}}{\varepsilon_{i+1}}\right)^{1/k+1}, s_{\max}\right) \geq c_s$$
$$\tilde{s}\left(\frac{\varepsilon_{\rm acc}}{\varepsilon_{i+1}}\right) = c_s\left(\frac{\varepsilon_{\rm acc}}{\varepsilon_{i+1}}\right)^{1/\tilde{k}} \leq c_s, \qquad \tilde{k} \leq k+1 \tag{19}$$

where c_s is a safety factor slightly smaller than unity and $s_{\max}$ bounds the relative increment of the step size to prevent excessive differences between adjacent points which deteriorate the effective order of the predictor.

III. Numerical Results

To demonstrate the validity of the method, we will search for the constants of propagation of the modes of the shielded dielectric image guide [1], [11] (see Fig. 1) using two different vectorial finite-element formulations. The first one [8] has the square of the propagation constant γ as the eigenvalue ($\lambda = \gamma^2$) and the discretized magnetic (or electric) field $\overline{H}_t + \hat{z}h_z$ as the eigenvector, where h_z is the scaled longitudinal component H_z, $\gamma h_z = H_z$. The expression from which the generalized eigenproblem is derived is

$$\iint_\Omega \left[(j\omega\varepsilon)^{-1}\nabla_t \times \overline{\varpi}_t \cdot \nabla_t \times \overline{H}_t + j\omega\mu\overline{\varpi}_t \cdot \overline{H}_t\right] d\Omega$$
$$= \gamma^2 \iint_\Omega \left[(j\omega\varepsilon)^{-1}(\overline{\varpi}_t + \nabla_t \varpi_z) \cdot (\overline{H}_t + \nabla_t h_z) + j\omega\mu\varpi_z h_z\right] d\Omega \tag{20}$$

where Ω is the transversal section of the waveguide ε, μ are the permittivity and permeability of the medium, and $\overline{\varpi} = \overline{\varpi}_t + \hat{z}\varpi'_z = \overline{\varpi}_t - \hat{z}\gamma\varpi_z$ is any test function in the same admissible function space as the trial function $\overline{h} = \overline{H}_t + \hat{z}h_z$; the waveguide walls are of either perfect electrical conductor (PEC), perfect magnetic conductor (PMC), or a combination of both. The null eigenvalue is a solution of the generalized eigenvalue problem and its multiplicity equals the number of free longitudinal components of the discretized field. The second formulation [9] has the propagation constant as the eigenvalue ($\lambda = \gamma$) and the eigenvector corresponds to the transversal components of both the electric and magnetic field $\overline{E}_t$, $\overline{H}_t$. For a waveguide inhomogeneously filled with isotropic media, it is straightforward to derive an equivalent Galerkin method

$$\iint_\Omega \left[(j\omega\varepsilon)^{-1}\nabla_t \times \overline{\varpi}_h \cdot \nabla_t \times \overline{H}_t + j\omega\mu\overline{\varpi}_h \cdot \overline{H}_t\right] d\Omega$$
$$= \gamma \iint_\Omega \overline{\varpi}_h \cdot \hat{z} \times \overline{E}_t \, d\Omega$$
$$\iint_\Omega \left[(j\omega\mu)^{-1}\nabla_t \times \overline{\varpi}_e \cdot \nabla_t \times \overline{E}_t + j\omega\varepsilon\overline{\varpi}_e \cdot \overline{E}_t\right] d\Omega$$
$$= \gamma \iint_\Omega -\overline{\varpi}_e \cdot \hat{z} \times \overline{H}_t \, d\Omega \tag{21}$$

where $\overline{\varpi}_h, \overline{\varpi}_e$ are magnetic and electric test functions, respectively, Ω is the section of the waveguide, and the boundaries are either PEC, PMC, or both. No extraneous solutions are associated to the null eigenvalue in this case, but both the forward-traveling and backward-traveling modes appear as distinct solutions with opposite sign eigenvalues. Both formulations have been implemented using mixed-order covariant elements [10]. For a given mesh of fixed number and location of the elements, the formulation in $\lambda = \gamma^2$ gives problems with a lower number of degrees of freedom and, consequently, the computation of each determinant is faster.

It should be pointed out that the example has been chosen because of the availability of well-known results for validation purposes. Although the matrices involved are not as general as the formulation assumes, no advantage has been gained from this fact, and the completely general sparse complex LU factorization and eigensolver have been used. The algorithm has been proven equally effective in every practical problem to which we have applied it.

In these examples, we will deal exclusively with regions of circular shape. Consequently, the computed number of eigenvalues within the circumference directly becomes the specified number of eigenvalues p to be found while the shift is the center of the circumference

$$C \equiv \{\lambda / |\lambda - \mu| = r\} = \{\lambda(\theta) = \mu + re^{j\theta}, 0 \leq \theta \leq 2\pi\}. \tag{22}$$

The first example consists in computing the multiplicity of the null eigenvalue of the formulation with $\lambda = \gamma^2$ (which will be referred to as "formulation I") for a given mesh. In all the following examples, the dimensions of the waveguide are $b = -a = d = 7.899$ mm, $w = 3.45$ mm, and $h = 3.2$ mm [1], the relative permittivity of the dielectric is $\varepsilon_r = 9$, the frequency is $f = 12$ GHz. The mesh consists of 45 elements, the number of degrees of freedom is 597, and the number of free longitudinal components is 209. It takes approximately 1.3 s to compute each determinant on a Pentium-133 PC. We choose a small radius $r = 10^{-4}$, but one as small as

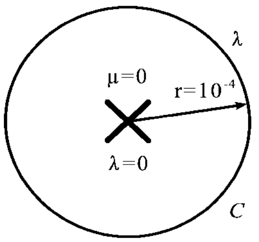

Fig. 2. Multiple null eigenvalue $\lambda = 0$ (formulation I) in the λ plane. A contour C of a very small radius encloses only this eigenvalue, except for the different cutoff frequencies of the modes.

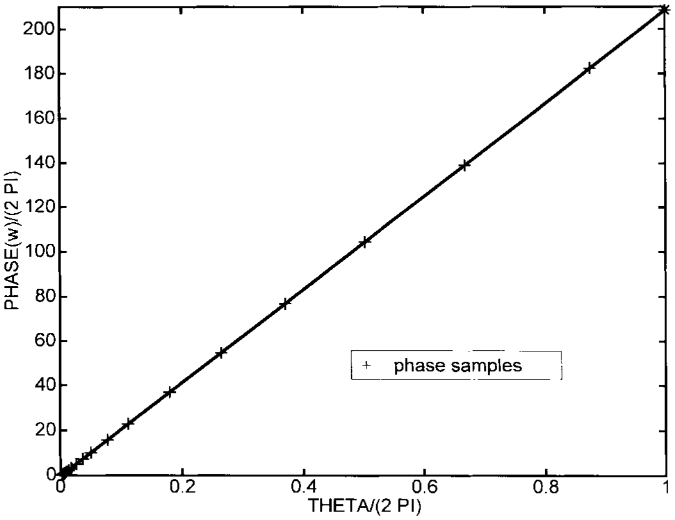

Fig. 3. Normalized reconstructed phase (loops) for the contour of Fig. 2. The number of zeros is 209, while the number of samples is only 34.

$r = 5.10^{-6}$ is equally practical (see Fig. 2). The number of eigenvalues within the contour coincides with the multiplicity of the null eigenvalue, except when the frequency approaches the cutoff frequency of one of the modes of the guide. In this case ($f = 12$ GHz), no mode is near cutoff, but it would suffice to repeat the calculation at a slightly different frequency to validate the result. The reconstructed phase is represented in Fig. 3, normalized to directly give the number of loops around the origin or zeros, which is $\nu_{\text{non-phys}} = 209$; the number of samples is 34 (it would take at least 209*2 = 418 equidistant samples for the simple unwrapping rule (17) to give the correct answer). Since the shift coincides with the multiple zero and the remaining zeros are very distant compared to the radius of the circle, the phase is an almost perfectly linear function of the parameter θ. This fact makes the computation especially efficient and tens of loops are skipped from sample to sample once the step has grown from its starting minimum value, which initially ensures that the ambiguity of the phase is correctly resolved. The increase of the sampling step can be observed in detail in Fig. 4.

Once the multiplicity of the null eigenvalue is known, the number of physical eigenvalues of absolute value less than a value r is simply

$$\nu_{\text{phys}}(r) = \nu(r) - \nu_{\text{non-phys}}. \tag{23}$$

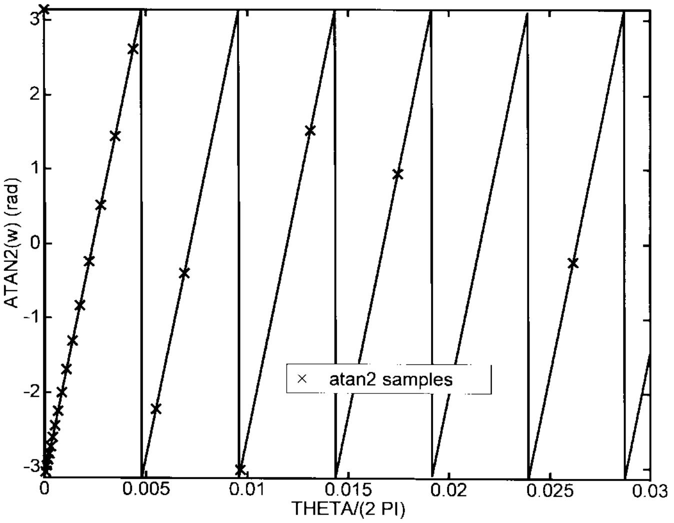

Fig. 4. Argument $\text{atan2}(\omega)$ and actual samples $\text{atan2}(\omega_i)$ for the contour of Fig. 2. Only the first loop is unambiguously sampled.

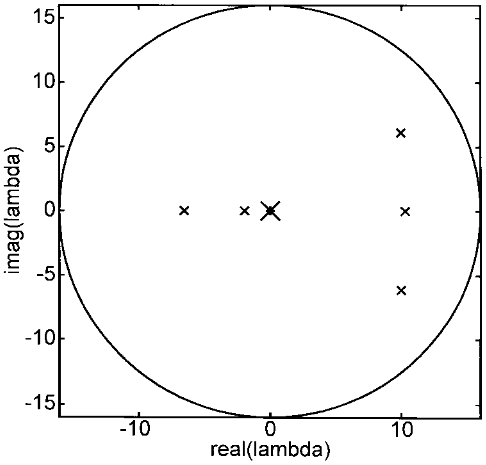

Fig. 5. Location of the eigenvalues $\lambda_{0i} = \gamma_i^2/|\lambda_{0i}| < 16$ (formulation I: $f = 12$ GHz; $\nu_{\text{non-phys}} = 209$).

Next, we will attempt to find all the physical eigenvalues lying within a circumference of radius $r = 16$ for a frequency f of 12 GHz (see Fig. 5). The search will be carried out in a systematic way. The first step is to compute $\nu(16)$ (see Fig. 6); the result is $\nu(16) = 214$ and, from (23), we have $\nu_{\text{phys}}(16) = 214 - 209 = 5$, i.e., there are five physical eigenvalues to be found. Next, the circle is covered by nine overlapping smaller circles in the way suggested by Lehmer and adopted in [6] as: 1) a concentric circle of radius $r/2$ and 2) eight circles of radius $5r/12$ regularly spaced around the remaining annulus (the distance of their centers from the original center is $(3/2)r\sqrt{1-(1/\sqrt{2})}$).

In this case (see Fig. 7), we compute $\nu(r/2) = \nu(8) = 211; \nu_{\text{phys}}(8) = 2$. At this stage, one of these eigenvalues plus the three within the annulus and a new outer eigenvalue are successfully found.

In Figs. 8–10, the details concerning circle 2 can be observed. The total number of samples is 117, with 13 acceptable prediction errors and two unacceptable ones. The errors are

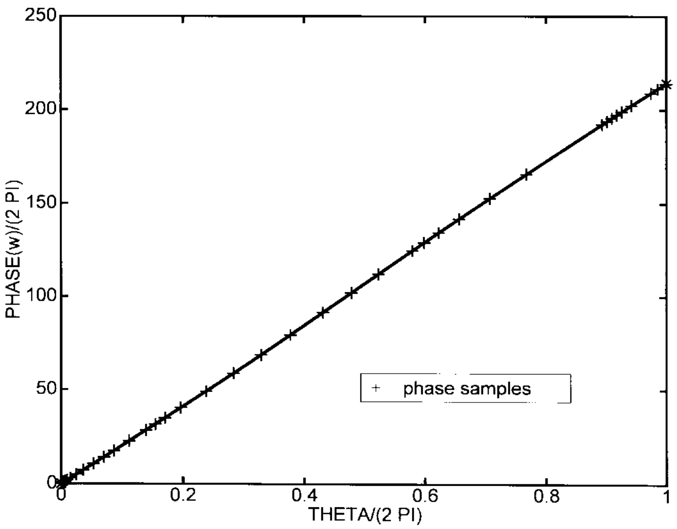

Fig. 6. Normalized reconstructed phase (loops) for the contour of Fig. 5. $\nu(16) = 214$. The number of samples (total)/acceptable errors/unacceptable errors is 64/6/1.

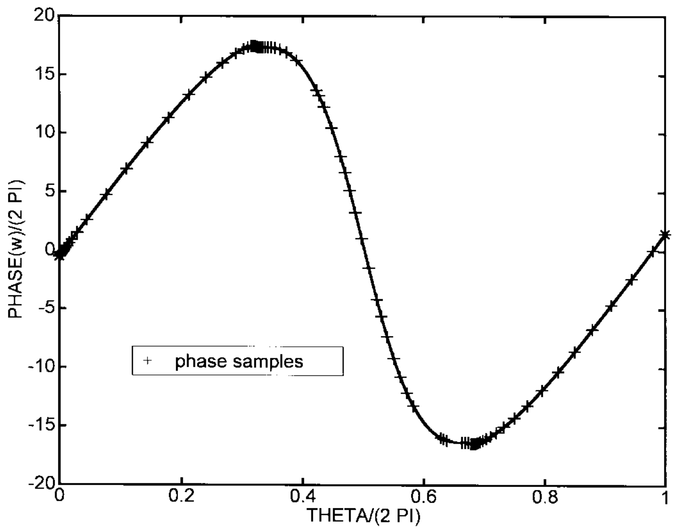

Fig. 9. Normalized reconstructed phase (loops) for the contour of Fig. 8. $\nu = 2$. The number of samples (total)/acceptable errors/unacceptable errors is 117/13/2.

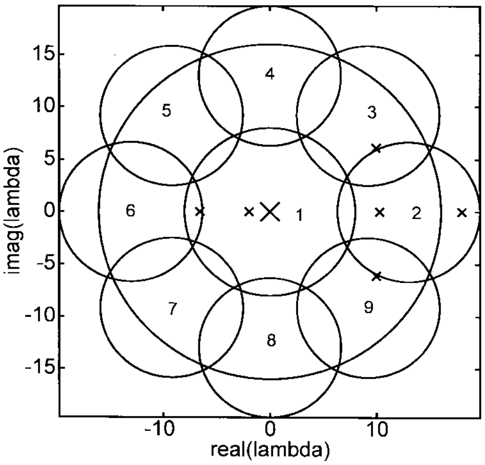

Fig. 7. Lehmer's subdivision of the circle of Fig. 5. Four out of the five eigenvalues within the contour and a new outer eigenvalue are successfully found.

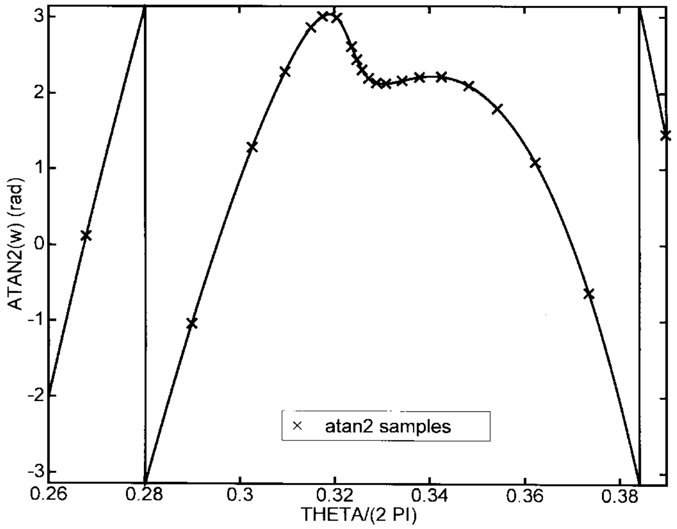

Fig. 10. Argument $\mathrm{atan2}(\omega)$ and actual samples $\mathrm{atan2}(\omega_i)$ for the contour of Fig. 8 (detailed view). The hump of the function corresponds to the presence of an outer complex zero lying near the contour.

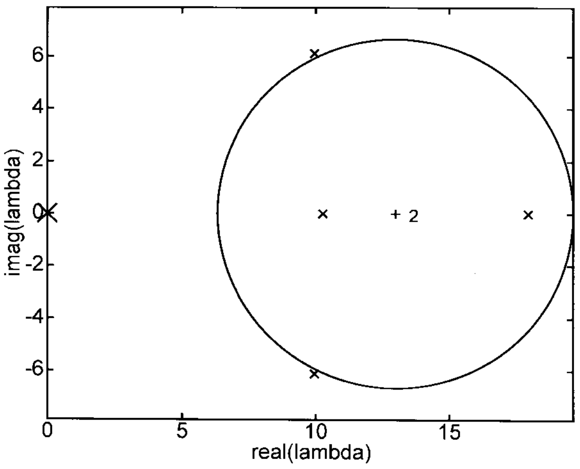

Fig. 8. Lehmer's subdivision of the circle of Fig. 5. Location of the relevant zeros for circle 2. $\mu = 12.988\,71, r = 6.666\,667$.

due to the presence of the two complex conjugate zeros lying very near the contour, which provoke a sharp variation of the derivative of the phase in its vicinity, and the strong overall fluctuation of the phase because of the presence of the multiple null eigenvalue displaced from the center of the circle. The subdivision could proceed recursively to find the single remaining zero, but it is more straightforward to try out different shifts around the origin and set $p = 1$ directly.

We will search for the same modes ($f = 12$ GHz) using the second formulation ("formulation II"). The mesh is the same as the one used with formulation I (45 elements), but now the number of degrees of freedom is 776 and it takes approximately 1.7 s to compute each determinant. We have $\lambda_{0i} = \pm\gamma_i$ and the radius of the circle must be $r = \sqrt{16} = 4$. The location of the zeros and different Lehmer's contours can be seen in Fig. 11. The need for a subdivision of the original circle depends on the ability of the solver to handle a subspace of dimension $q = 2p = 20$. As in the previous example, two additional outer eigenvalues are found in the

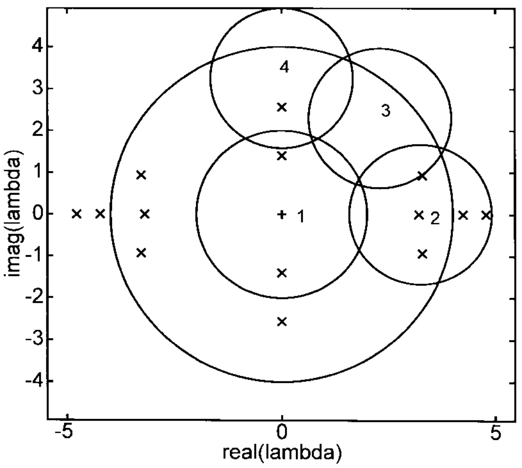

Fig. 11. Location of the eigenvalues $\lambda_{0i} = \pm\gamma_i/|\lambda_{0i}| < 4$ (formulation II: $f = 12$ GHz; $\nu_{\text{non-phys}} = 0$) and Lehmer's contours employed for their computation (the rest are unnecessary due to the symmetry of the location of the zeros).

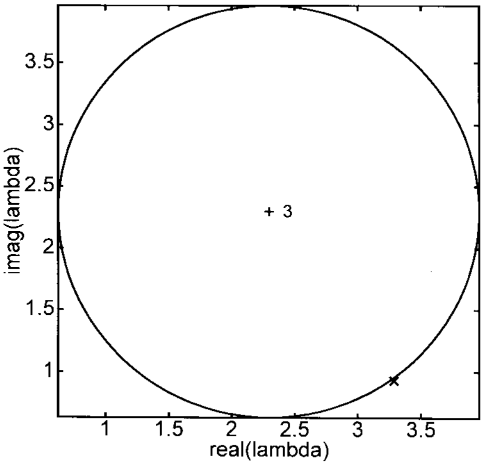

Fig. 12. Lehmer's subdivision of the circle of Fig. 11. Location of the relevant zeros for circle 3. $\mu = 2.2961(1 + j); r = 1.6667$.

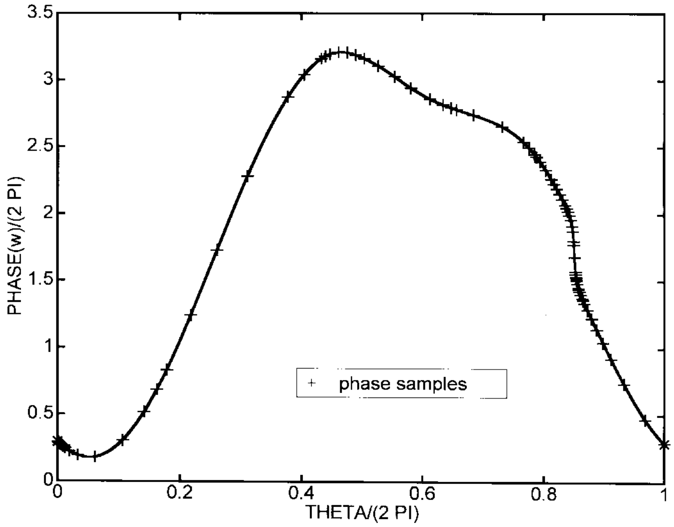

Fig. 13. Normalized reconstructed phase (loops) for the contour of Fig. 12. $\nu = 0$. The number of samples (total)/acceptable errors/unacceptable errors is 94/7/4.

process of the subdivision and, again, there are some zeros lying very near the contour of some of the circles, which stresses the importance of ensuring the robustness of the routine that computes the number of zeros. In Fig. 12, it can be seen in circle 3 with an outer zero bordering the contour; the normalized reconstructed phase can be observed in Fig. 13. The total number of samples/acceptable errors/unacceptable errors is 94/7/4. The rest of the contours of Fig. 11 pose no special difficulty.

Regarding the number and the location of the eigenvalues, the results are consistent with those shown in the references. Of course, it is hardly necessary to mention that the results are fully self-consistent (i.e., that the predicted and the computed number of eigenvalues are identical in every case, which asserts the robustness of the predictor).

IV. Conclusion

A method has been presented which solves the shortcomings of the otherwise powerful inverse subspace iteration method for the solution of generalized complex eigenvalue problems. The reasons that justify the need for this method have been explained. An algorithm for the computation of the number of eigenvalues within a closed contour is given. The aim of the algorithm is to minimize the numerical burden and yet to ensure the robustness of the calculation. It has been shown that the presence of an identically null eigenvalue of high multiplicity affects the way in which the search for eigenvalues of small absolute value has to be conducted. Several numerical examples have been analyzed which show the effectiveness of the proposed method and illustrate the possibility of performing a systematic search for the eigenvalues of a problem. The recursive and exhaustive nature of the proposed method of search guarantees that all the modes lying within a finite-size closed contour be found.

References

[1] F. A. Fernández, J. B. Davies, S. Zhu, and Y. Lu, "Sparse matrix eigenvalue solver for finite element solution of dielectric waveguides," *Electron. Lett.*, vol. 27, no. 20, pp. 1824–1826, Sept. 26, 1991.

[2] Y. Lu, S. Zhu, and F. A. Fernández, "The efficient solution of large sparse nonsymmetric and complex eigensystems by subspace iteration," *IEEE Trans. Magn.*, vol. 30, pp. 3582–3585, Sept. 1994.

[3] F. A. Fernández and Y. Lu, *Microwave and Optical Waveguide Analysis by the Finite Element Method.* Taunton, Somerset, U.K.: Educational Research Studies Press Ltd., 1996, ch. 7, pp. 117–134.

[4] J. R. Bauer and G. C. Lizalek, "Microwave filter analysis using a new 3-D finite-element modal frequency method," *IEEE Trans. Microwave Theory Tech.*, vol. 45, pp. 810–818, May 1997.

[5] K. Bathe, *Finite Element Procedures in Engineering Analysis.* Englewood Cliffs, NJ: Prentice-Hall, 1982.

[6] L. M. Delves and J. N. Lyness, "A numerical method for locating the zeros of an analytic function," *Math. Comput.*, vol. 21, pp. 543–560, 1967.

[7] F. Mesa, "Estudio de las características de propagación electromagnética en líneas multiconductoras de configuración planar inmersas en medios bianisótropos estratificados," Ph.D. dissertation, Facultad de Física, Universidad de Sevilla, Seville, Spain, 1991, ch. 3.

[8] J.-F. Lee, "Finite-element analysis of lossy dielectric waveguides," *IEEE Trans. Microwave Theory Tech.*, vol. 42, pp. 1025–1031, June 1994.

[9] E. W. Lucas and T. P. Fontana, "Vector finite-element implementation of the variational Et–Ht generalized eigenmode formulation," in *IEEE AP Symp. Dig.*, Seattle, WA, June 1994, pp. 1764–1767.

[10] R. Miniowitz and J. P. Webb, "Covariant-projection quadrilateral elements for the analysis of waveguides with sharp edges," *IEEE Trans. Microwave Theory Tech.*, vol. 39, pp. 501–505, Mar. 1991.

[11] J. Strube and F. Arndt, "Rigorous hybrid-mode analysis of the transition from rectangular waveguide to shielded dielectric image guide," *IEEE Trans. Microwave Theory Tech.*, vol. 33, pp. 391–401, May 1985.

Javier Arroyo received the Ingeniero de Telecomunicación degree from the Universidad Politécnica de Madrid, Madrid, Spain, in 1995, and is currently working toward the Ph.D. degree in the Departamento de Electromagnetismo y Teoríia de Circuitos.

His current interest is in the numerical modeling of passive radiators.

Juan Zapata (M'93) received the Ingeniero de Telecomunicación degree and the Ph.D. degree from the Universidad Politécnica de Madrid, Madrid, Spain, in 1970 and 1974, respectively.

Since 1970, he has been with the Departamento de Electromagnetismo y Teoría de Circuitos, Universidad Politécnica de Madrid, where he became an Assistant Professor in 1970, Associate Professor in 1975, and Professor in 1983. He has been engaged in research on microwave active circuits and interactions of electromagnetic fields with biological tissues. His current research interest includes computer-aided design for microwave passive circuits and numerical methods in electromagnetism.