

*Journal of*  
***Mechanics of***  
***Materials and Structures***

*Special issue*

***Ninth Pan American Congress  
of Applied Mechanics***

***Volume 2, N° 8***

***October 2007***

# JOURNAL OF MECHANICS OF MATERIALS AND STRUCTURES

<http://www.jomms.org>

EDITOR-IN-CHIEF Charles R. Steele  
ASSOCIATE EDITOR Marie-Louise Steele  
Division of Mechanics and Computation  
Stanford University  
Stanford, CA 94305  
USA

## BOARD OF EDITORS

D. BIGONI University of Trento, Italy  
H. D. BUI École Polytechnique, France  
J. P. CARTER University of Sydney, Australia  
R. M. CHRISTENSEN Stanford University, U.S.A.  
G. M. L. GLADWELL University of Waterloo, Canada  
D. H. HODGES Georgia Institute of Technology, U.S.A.  
J. HUTCHINSON Harvard University, U.S.A.  
C. HWU National Cheng Kung University, R.O. China  
IWONA JASIUK University of Illinois at Urbana-Champaign  
B. L. KARIHALOO University of Wales, U.K.  
Y. Y. KIM Seoul National University, Republic of Korea  
Z. MROZ Academy of Science, Poland  
D. PAMPLONA Universidade Católica do Rio de Janeiro, Brazil  
M. B. RUBIN Technion, Haifa, Israel  
Y. SHINDO Tohoku University, Japan  
A. N. SHUPIKOV Ukrainian Academy of Sciences, Ukraine  
T. TARNAI University Budapest, Hungary  
F. Y. M. WAN University of California, Irvine, U.S.A.  
P. WRIGGERS Universität Hannover, Germany  
W. YANG Tsinghua University, P.R. China  
F. ZIEGLER Technische Universität Wien, Austria

## PRODUCTION

PAULO NEY DE SOUZA Production Manager  
SHEILA NEWBERY Senior Production Editor  
SILVIO LEVY Scientific Editor

---

See inside back cover or <http://www.jomms.org> for submission guidelines.

---

Regular subscription rate: \$500 a year.

Subscriptions, requests for back issues, and changes of address should be sent to Mathematical Sciences Publishers, 798 Evans Hall, Department of Mathematics, University of California, Berkeley, CA 94720-3840.

---

©Copyright 2008. Journal of Mechanics of Materials and Structures. All rights reserved.

 mathematical sciences publishers

## PREFACE

JEFFREY W. EISCHEN AND GUILLERMO MONSIVAIS

The aim of the sponsors of the 9th Pan American Congress of Applied Mechanics (PACAM IX) was to promote progress in the broad field of mechanics by (1) exposing mature engineers and scientists, as well as advanced graduate students, to new research findings, techniques, and problems, and (2) providing opportunities for personal interactions through formal presentations and informal conversations. The meetings are traditionally held every two years in a Latin American venue, at a time when few other conferences are scheduled. The previous Congresses were held in Rio de Janeiro, Brazil (1989); Valparaíso, Chile (1991); São Paulo, Brazil (1993); Buenos Aires, Argentina (1995); San Juan, Puerto Rico (1997); Rio de Janeiro, Brazil (1999); Temuco, Chile (2002); and Havana, Cuba (2004).

PACAM IX was held at the Fiesta Americana Hotel in Mérida, Mexico, from January 2–6, 2006, and was cosponsored by the American Academy of Mechanics; the US National Science Foundation; the Air Force Office of Scientific Research; Universidad Nacional Autónoma de México's Instituto de Ingeniería, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Centro de Ciencias Físicas, Coordinación de la Investigación Científica, Facultad de Ingeniería, Instituto de Física; and Universidad Autónoma de Yucatán's Facultad de Ingeniería, Facultad de Matemáticas; and CINVESTAV, Mérida, Yucatán. Approximately 75 attendees from 20 countries enjoyed a very productive and collegial meeting in Mérida.

Following the PACAM IX meeting the organizers invited authors of selected talks to submit full-length articles on the matter of their presentation. These papers were then subjected to the normal peer review and editorial process of the *Journal of Mechanics of Materials and Structures*. This special issue is the collection of the accepted papers.

Finally, we thank the Editor-in-Chief and Associate Editor of JoMMS, Charles and Marie-Louise Steele, for giving us the great opportunity to organize this special issue. We thank the contributing authors for their excellent papers and also the anonymous reviewers, who helped immensely in shaping this special issue.

July 2007

JEFFREY W. EISCHEN: [eischen@ncsu.edu](mailto:eischen@ncsu.edu)

Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, NC 27695-7910, United States

GUILLERMO MONSIVAIS: [monsi@fisica.unam.mx](mailto:monsi@fisica.unam.mx)

Instituto de Física, Universidad Nacional Autónoma de México, Apdo. Postal 20-364, México 01000 DF, 01000 México, D.F., Mexico

## OPTIMIZATION OF A SATELLITE WITH COMPOSITE MATERIALS

JORGE A. C. AMBRÓSIO, MARIA AUGUSTA NETO AND ROGÉRIO PEREIRA LEAL

The design of complex flexible multibody systems for industrial applications requires not only the use of powerful methodologies for the system analysis, but also the ability to analyze potential designs and to decide on the merits of each one of them. This paper presents a methodology using optimization procedures to find the optimal layouts of fiber composite structure components in multibody systems. The goal of the optimization process is to minimize structural deformation and to fulfill a set of multidisciplinary constraints. These methodologies rely on the efficient and accurate calculation of the system sensitivities to support the optimization algorithms. In this work a general formulation for the computation of the first order analytic sensitivities based on the direct differentiation method is used. The direct method for sensitivity calculation is obtained by direct differentiation of the equations defining the response of the structure with respect to the design variables. The equations of motion and the sensitivities of the flexible multibody system are solved simultaneously and, therefore, the accelerations and velocities of the system, and the sensitivities of the accelerations and velocities, are integrated in time using a multistep multiorder integration algorithm. Different models for the flexible components of the system, using beam and plate elements, are also considered. Finally, the methodology proposed here is applied to the optimization of the unfolding of a complex satellite made of composite plates and beams. The ply orientations of lamination are the continuous design variables. The potential difficulties in the optimization of composite flexible multibody systems are highlighted in the discussion of the results obtained.

### 1. Introduction

*Modeling* refers to the tools used in the construction of models of individual and coupled components of technical systems. The simplest models for multibody systems assume rigid body components while more complex models require the description of the components' flexibility. The finite element-based strategies used to represent the components' flexibility in multibody systems is a well accepted and widely used method. For systems in which the bodies are made of standard materials, there is a wide variety of finite elements that may be used, but when bodies are made of composite materials, the model flexibility often necessitates expensive finite element models with an inherent growth in complexity. Models of systems involving multibody dynamics methodologies also require a complete knowledge of the arrangement of the system components, which is achieved by the definition of kinematic joints, the introduction of models for external forces and the incorporation of the equilibrium equations of other disciplines [Heckmann et al. 2005; Møller et al. 2005; Bottasso et al. 2006]. Regardless of each particular type of joint used, the mathematical description of the restrictions involving only rigid bodies are the simplest to obtain. The presence of flexible bodies tends to increase the complexity of the description, and methods for simplifying the description are required [Lehner and Eberhard 2006; Hardeman et al.

---

*Keywords:* flexible multibody dynamics, sensitivity analysis, automatic differentiation, large rotations, floating frame.

2006]. However, the concept of virtual bodies provides a general framework for developing general kinematic joints for flexible multibody systems with minimal effort [Ambrósio 2003].

Analyses of rigid mechanical systems are the simplest and the least expensive, regardless of model. Flexible systems, in which the bodies only experience small elastic deformations, have higher computational costs. For these systems it is common to use mode component synthesis to reduce the number of generalized elastic coordinates and, consequently, the equations of motion are written in terms of modal coordinates [Nikravesh and Lin 2005; Gonçalves and Ambrósio 2005; Lehner and Eberhard 2006]. However, when the system components experience nonlinear deformations, the use of reduction methods is not possible, in general, and the finite element nodal coordinates are the generalized coordinates used [Ambrósio 1996; Dmitrochenko et al. 2006; Gerstmayr and Schöberl 2006; Vetyukov et al. 2006]. Furthermore, the analysis of these systems is more complex and, usually, computationally more expensive than the analysis of flexible systems with bodies that experience linear deformations.

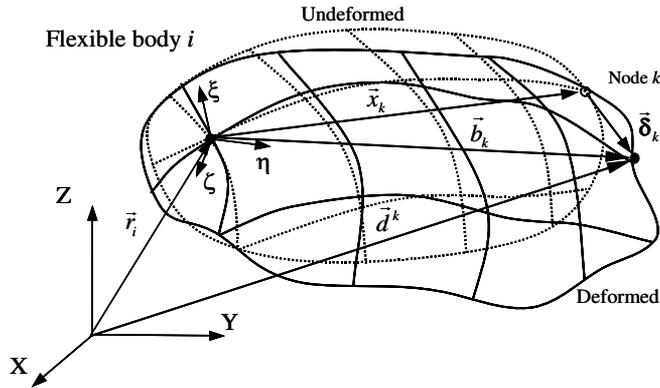
In terms of the optimization complexity, the most complex and expensive problems are global or integer optimization problems with a large number of design variables. The simplest and cheapest problems to solve are continuous local problems with a small number of design variables [Venkataraman and Haftka 1999; Venkataraman and Haftka 2002]. Stochastic optimization algorithms, like simulated annealing methods or genetic algorithms, offer a way to perform global optimization, but they usually require several hundreds or even thousands of expensive simulation runs [He and Mcphee 2005; Kübler et al. 2005]. Eberhard and co-workers used a stochastic evolution strategy in combination with parallel computing in order to reduce the computation times while maintaining the inherent robustness [Eberhard et al. 2003]. Deterministic optimization algorithms, on the other hand, have a tendency to reach local minima, not necessarily the global optimum [Eberhard et al. 1999]. When supported by efficient calculation of the system sensitivities, these deterministic optimization algorithms often converge rapidly towards a local minimum with smaller computation times than other optimization approaches.

In this work, a general approach for sensitivity analysis of rigid-flexible multibody systems with composite materials based on the automatic differentiation method is used. The direct differentiation of the system equations of motion is obtained by the ADIFOR program [Bischof et al. 1992]. The dynamic equations and the time derivatives of the sensitivities are all integrated at the same time, thus the control of the time integration errors becomes more effective. The simultaneous integration of the equations is even more important when a variable step size or variable order integration algorithm is used, as is generally the case in multibody dynamic systems.

The optimization of the multibody composite components is performed by taking the ply orientations of lamination as continuous design variables. The multibody dynamic and sensitivities analysis code is linked with general optimization algorithms included in the package DOT/DOC [Vanderplaats 1992].

## 2. The multibody analysis methodology

**2.1. Multibody equations of motion.** The location of a rigid body is defined by the position of a body-fixed reference frame,  $\xi\eta\zeta$ , and its orientation with respect to an inertial frame,  $XYZ$ , as shown in Figure 1. The position and the orientation of the rigid body  $i$  is defined by the translation coordinates  $r_i$  and the rotational coordinates  $p_i$ . These coordinates are grouped in the vector  $q_{ri} = [r_i^T p_i^T]^T$ . The coordinate vector of the complete flexible system is designated by  $q = [q_r^T u'^T]^T$ , which is composed



**Figure 1.** Flexible body with its body fixed coordinate system.

of the coordinate vector of the individual bodies and the elastic coordinates of the flexible bodies  $\mathbf{u}'_i$ , generally the nodal coordinates of the finite element mesh measured with respect to the body-fixed coordinate system or the modal coordinates when a mode component synthesis method is used to represent the deformation of the flexible body.

For a multibody system, a set of constraint equations associated to the kinematic joints that restrict the relative motion between the bodies is defined as [Ambrósio and Gonçalves 2001]:

$$\Phi(\mathbf{q}, t) \equiv \mathbf{0}, \quad (1)$$

where  $t$  refers to the kinematic constraints that depend on time. The constraints equations are added to the equilibrium equations using Lagrange multipliers

$$\mathbf{M}\ddot{\mathbf{q}} + \Phi_q^T \lambda = \mathbf{g} + \mathbf{s} - \mathbf{K}\mathbf{q}, \quad (2)$$

where  $\mathbf{M}$  is the system mass matrix,  $\mathbf{K}$  is the extended stiffness matrix of the system,  $\mathbf{g}$  is a vector of external applied forces and  $\mathbf{s}$  is the vector of the forces that depend on the square of the system velocities. Equation (2) includes  $n$  unknown accelerations and  $m$  unknown Lagrange multipliers associated with the algebraic constraint equations, but it only has  $n$  equations. The second time derivatives of the constraint equations provide the extra set of  $m$  equations necessary to support the solution of Equation (2). These acceleration constraint equations are

$$\ddot{\Phi}(\ddot{\mathbf{q}}, \dot{\mathbf{q}}, \mathbf{q}, t) \equiv \Phi_q \ddot{\mathbf{q}} - \gamma = \mathbf{0}. \quad (3)$$

Therefore, the complete system of equations that needs to be solved for a flexible multibody system is given by [Ambrósio and Gonçalves 2001]

$$\begin{bmatrix} \mathbf{M}_r & \mathbf{M}_{rf} & \Phi_{q_r}^T \\ \mathbf{M}_{fr} & \mathbf{M}_{ff} & \Phi_{q_f}^T \\ \Phi_{q_r} & \Phi_{q_f} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{q}}_r \\ \ddot{\mathbf{u}}' \\ \lambda \end{Bmatrix} = \begin{Bmatrix} \mathbf{g}_r \\ \mathbf{g}_f \\ \gamma \end{Bmatrix} - \begin{Bmatrix} \mathbf{s}_r \\ \mathbf{s}_f \\ \mathbf{0} \end{Bmatrix} - \begin{Bmatrix} \mathbf{0} \\ \mathbf{K}_{ff} \mathbf{u}' \\ \mathbf{0} \end{Bmatrix}, \quad (4)$$

where  $\mathbf{K}_{ff}$  is the standard finite element stiffness matrix. The Jacobian matrix  $\Phi_q^T$  and the right-hand-side vector  $\boldsymbol{\gamma}$  of Equations (3) and (4) depend on the type of kinematic constraints used. The system equation matrix shows a large number of null elements and submatrix blocks of fixed size. The Markowitz sparse matrix solver is employed here to solve the system of equations defined by Equation (4) [Duff et al. 1986; Ambrósio 2003; Liu et al. 2007].

The equations of motion for the flexible multibody systems represented by (4) require a large number of coordinates to describe complex models. However, using component mode synthesis, the flexible body is described by a sum of selected modes of vibration as

$$\mathbf{u}' = \mathbf{X}\mathbf{w}, \tag{5}$$

where vector  $\mathbf{w}$  represents the contributions of the vibration modes towards the nodal displacements and  $\mathbf{X}$  is the modal matrix. Due to the reference conditions, the modes of vibration used here are constrained modes and due to the assumption of linear elastic deformations the modal matrix is invariant. The reduced equations of motion for the flexible body are [Ambrósio and Gonçalves 2001]

$$\begin{bmatrix} \mathbf{M}_r & \mathbf{M}_{rf}\mathbf{X} & \Phi_{q_r}^T \\ \mathbf{X}^T\mathbf{M}_{fr} & \mathbf{I} & \mathbf{X}^T\Phi_{q_f}^T \\ \Phi_{q_r} & \Phi_{q_f}\mathbf{X} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{q}}_r \\ \dot{\mathbf{w}} \\ \boldsymbol{\lambda} \end{Bmatrix} = \begin{Bmatrix} \mathbf{g}_r \\ \mathbf{X}^T\mathbf{g}_f \\ \boldsymbol{\gamma} \end{Bmatrix} - \begin{Bmatrix} \mathbf{s}_r \\ \mathbf{X}^T\mathbf{s}_f \\ \mathbf{0} \end{Bmatrix} - \begin{Bmatrix} \mathbf{0} \\ \boldsymbol{\Lambda}\mathbf{w} \\ \mathbf{0} \end{Bmatrix}, \tag{6}$$

where  $\boldsymbol{\Lambda}$  is a diagonal matrix with the squares of the natural frequencies associated with the modes of vibration selected. The number of elastic coordinates in Equation (6) is equal to the number of vibration modes selected. For a more detailed discussion on the selection of the modes used, the interested reader is referred to [Cavin and Dusto 1977; Yoo and Haug 1986; Pereira and Proença 1991].

**2.2. Flexible bodies made of composite materials.** In this work the composite finite element used for the study of laminated plates is based on the Mindlin–Reissner plate theory, where only  $C^\circ$  continuity is required for the approximation of the kinematic variables. At the element level and in local coordinates, the element stiffness matrix is given by [Neto et al. 2004]

$$\mathbf{K}_{ff}^{(e)} = \int_0^1 \int_0^{1-\eta} \begin{bmatrix} \mathbf{B}_m^T \mathbf{D}_m \mathbf{B}_m & \mathbf{B}_m^T \mathbf{D}_{mb} \mathbf{B}_b & \mathbf{0} \\ \mathbf{B}_b^T \mathbf{D}_{bm} \mathbf{B}_m & \mathbf{B}_b^T \mathbf{D}_b \mathbf{B}_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_s^T \mathbf{D}_s \mathbf{B}_s \end{bmatrix} |J| d\xi d\eta \tag{7}$$

which in a more compact form is written as

$$\mathbf{K}_{ff}^{(e)} = \int_0^1 \int_0^{1-\eta} (\mathbf{B}^T \mathbf{D} \mathbf{B})^{(e)} |J| d\xi d\eta. \tag{8}$$

The strain-displacement matrix is denoted by  $\mathbf{B}$  while  $\mathbf{D}$  is the elasticity matrix and  $|J|$  is the determinant of the Jacobian matrix. The subscripts  $m$ ,  $b$  and  $s$  stand for membrane, bending and shear. Because each layer may have different properties, the elasticity matrix  $\mathbf{D}$  is evaluated as a summation carried out over the thickness of all the layers. Therefore, equivalent single layer theories produce equivalent

stiffness matrices as weighted averages of the individual layer stiffness through the thickness. These matrices are dependent on each layer orientation, and are given by

$$\begin{aligned} (\mathbf{D}_m, \mathbf{D}_b, \mathbf{D}_{mb}, \mathbf{D}_s) &= \sum_{k=1}^n (\mathbf{D}_m, \mathbf{D}_b, \mathbf{D}_{mb}, \mathbf{D}_s)_k \\ &= \sum_{k=1}^n (\mathbf{C}_{3 \times 3}^1 H_1, \mathbf{C}_{3 \times 3}^1 H_2, \mathbf{C}_{3 \times 3}^1 H_3, \mathbf{C}_{2 \times 2}^2 H)_k \end{aligned} \quad (9)$$

with

$$H_n = \int_{h_{l-1}}^{h_l} (x_3^{n-1}) dz = \frac{1}{n} (h_{l+1}^n - h_l^n), \quad (10)$$

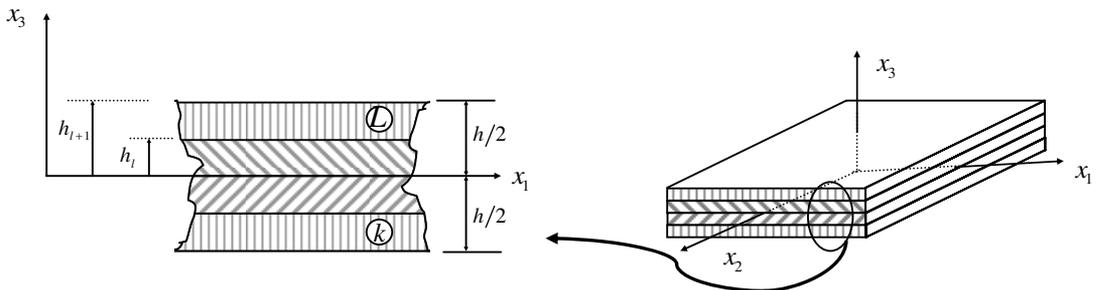
where  $h_i$  is defined in Figure 2. The axis  $x_3$  is positive upward from the mid-plane of the plate. The  $L$ th layer is located between the points  $x_3 = h_l$  and  $x_3 = h_{l+1}$  in the direction of the thickness.

At the element level and in local coordinates, the consistent mass matrix is given by

$$\mathbf{M}_{ff}^{(e)} = \int_0^1 \int_0^{1-\eta} \rho^{(e)} (\mathbf{S}^T \mathbf{m} \mathbf{S})^{(e)} |\mathbf{J}| d\xi d\eta, \quad (11)$$

where  $\mathbf{m}$  is a matrix that contains the inertial terms, and  $\rho$  represents the specific mass of the element. Before the mass matrix given by Equation (11), is used in what follows, a procedure to obtain a diagonal mass matrix is applied [Cook 1987].

The description of some of the flexible bodies of the multibody systems requires the use of composite plates, discretized by triangular finite elements. The finite element is based in the theory described and has six degrees of freedom per node:  $u_1^o, u_2^o, u_3^o, \phi_1, \phi_2$  and  $\phi_3$ . In the finite element mesh of some of the flexible bodies of the system composite beam elements are also used. For the sake of conciseness, none of these elements is described here, but for details on the formulations of the different composite finite elements, the interested reader is referred to [Neto et al. 2004].



**Figure 2.** Coordinate system and layer numbering used for a typical laminated plate.

### 3. Sensitivity analysis of the multibody system

The optimization algorithms used in this work require not only the evaluation of the functional values of the behavior functions but also their sensitivities with respect to the design variables. The calculation of these sensitivities can be carried out analytically or numerically. In this work only the analytical sensitivities are obtained by using automatic differentiation.

**3.1. Sensitivity of the equation of motion.** For a rigid-flexible multibody system, the equations of motion in terms of modal coordinates are given by Equation (6). The sensitivities of the system accelerations and Lagrange multipliers with respect to the design variables are obtained by differentiating Equation (6) with respect to the design variables  $\mathbf{b}$ :

$$\begin{bmatrix} \mathbf{M}_r & \mathbf{M}_{rf} \mathbf{X} & \Phi_{q_r}^T \\ \mathbf{X}^T \mathbf{M}_{fr} & \mathbf{I} & \mathbf{X}^T \Phi_{q_f}^T \\ \Phi_{q_r} & \Phi_{q_f} \mathbf{X} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{q}}_{rb} \\ \ddot{\mathbf{w}}_b \\ \lambda_b \end{Bmatrix} = \begin{Bmatrix} \mathbf{Q}_b \\ \mathbf{R}_b \\ \gamma_b \end{Bmatrix}, \quad (12)$$

where  $(\cdot)_b$  denotes the sensitivity of quantity  $(\cdot)$  with respect to  $\mathbf{b}$ . The sensitivities of the right-hand-side of the equation  $\mathbf{Q}_b$ ,  $\mathbf{R}_b$  and  $\gamma_b$  are

$$\begin{aligned} \mathbf{Q}_b &= \frac{\partial}{\partial \mathbf{q}_r} (\mathbf{g}_r - \mathbf{S}_r - \mathbf{M}_r \ddot{\mathbf{q}}_r - \mathbf{M}_{rf} \mathbf{X} \ddot{\mathbf{w}} - \Phi_{q_r}^T \lambda) \mathbf{q}_{rb} + \frac{\partial}{\partial \dot{\mathbf{q}}_r} (\mathbf{g}_r - \mathbf{S}_r) \dot{\mathbf{q}}_{rb} + \frac{\partial}{\partial \dot{\mathbf{w}}} (\mathbf{g}_r - \mathbf{S}_r) \dot{\mathbf{w}}_b \\ &+ \frac{\partial}{\partial \mathbf{w}} (\mathbf{g}_r - \mathbf{S}_r - \mathbf{M}_r \ddot{\mathbf{q}}_r - \mathbf{M}_{rf} \mathbf{X} \ddot{\mathbf{w}} - \Phi_{q_r}^T \lambda) \mathbf{w}_b + \frac{\partial}{\partial \mathbf{b}} (\mathbf{g}_r - \mathbf{S}_r - \mathbf{M}_r \ddot{\mathbf{q}}_r - \mathbf{M}_{rf} \mathbf{X} \ddot{\mathbf{w}} - \Phi_{q_r}^T \lambda); \quad (13) \end{aligned}$$

$$\begin{aligned} \mathbf{R}_b &= \frac{\partial}{\partial \mathbf{q}_r} \left( \mathbf{X}^T \mathbf{g}_f - \mathbf{X}^T \mathbf{S}_f - \mathbf{X}^T \mathbf{K}_{ff} \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{M}_{fr} \ddot{\mathbf{q}}_r - \mathbf{X}^T \mathbf{M}_{ff} \mathbf{X} \ddot{\mathbf{w}} - \mathbf{X}^T \Phi_{q_f}^T \lambda \right) \mathbf{q}_{rb} \\ &+ \frac{\partial}{\partial \mathbf{w}} \left( \mathbf{X}^T \mathbf{g}_f - \mathbf{X}^T \mathbf{S}_f - \mathbf{X}^T \mathbf{K}_{ff} \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{M}_{fr} \ddot{\mathbf{q}}_r - \mathbf{X}^T \mathbf{M}_{ff} \mathbf{X} \ddot{\mathbf{w}} - \mathbf{X}^T \Phi_{q_f}^T \lambda \right) \mathbf{w}_b \\ &+ \frac{\partial}{\partial \mathbf{b}} \left( \mathbf{X}^T \mathbf{g}_f - \mathbf{X}^T \mathbf{S}_f - \mathbf{X}^T \mathbf{K}_{ff} \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{M}_{fr} \ddot{\mathbf{q}}_r - \mathbf{X}^T \mathbf{M}_{ff} \mathbf{X} \ddot{\mathbf{w}} - \mathbf{X}^T \Phi_{q_f}^T \lambda \right) \\ &+ \frac{\partial}{\partial \dot{\mathbf{q}}_r} (\mathbf{X}^T \mathbf{g}_f - \mathbf{X}^T \mathbf{S}_f) \dot{\mathbf{q}}_{rb} + \frac{\partial}{\partial \dot{\mathbf{w}}} (\mathbf{X}^T \mathbf{g}_f - \mathbf{X}^T \mathbf{S}_f) \dot{\mathbf{w}}_b; \quad (14) \end{aligned}$$

$$\begin{aligned} \gamma_b &= \frac{\partial}{\partial \mathbf{q}_r} (\gamma - \Phi_{q_r} \ddot{\mathbf{q}}_r - \Phi_{q_f} \mathbf{X} \ddot{\mathbf{w}}) \mathbf{q}_{rb} + \frac{\partial}{\partial \mathbf{w}} (\gamma - \Phi_{q_r} \ddot{\mathbf{q}}_r - \Phi_{q_f} \mathbf{X} \ddot{\mathbf{w}}) \mathbf{w}_b \\ &+ \frac{\partial}{\partial \mathbf{b}} (\gamma - \Phi_{q_r} \ddot{\mathbf{q}}_r - \Phi_{q_f} \mathbf{X} \ddot{\mathbf{w}}) + \frac{\partial \gamma}{\partial \dot{\mathbf{q}}_r} \dot{\mathbf{q}}_{rb} + \frac{\partial \gamma}{\partial \dot{\mathbf{w}}} \dot{\mathbf{w}}_b. \quad (15) \end{aligned}$$

After solving the linear system of Equations (12) to obtain the sensitivities  $\ddot{\mathbf{q}}_{rb}$ ,  $\ddot{\mathbf{w}}_b$  and  $\lambda_b$  the state variables' sensitivities are obtained by direct integration of  $\dot{\mathbf{q}}_{rb}$ ,  $\dot{\mathbf{w}}_b$ ,  $\ddot{\mathbf{q}}_{rb}$  and  $\ddot{\mathbf{w}}_b$ . The process is started

with the initial conditions given by:

$$\begin{cases} \mathbf{q}_{r_b}(t_0) = \mathbf{q}_{r_b}^0, \\ \mathbf{w}_b(t_0) = \mathbf{w}_b^0, \\ \dot{\mathbf{q}}_{r_b}(t_0) = \dot{\mathbf{q}}_{r_b}^0, \\ \dot{\mathbf{w}}_b(t_0) = \dot{\mathbf{w}}_b^0. \end{cases} \quad (16)$$

Generally, the initial conditions for the sensitivities expressed in Equation (16) are assumed to be null. Note also that the leading matrix of (6) is equal to the leading matrix of Equation (12). Generally, the factorized matrix used to obtain the solution of the equation of motion does not have to be calculated again when the sensitivities system of equations need to be solved. However, because an automatic differentiation tool is used [Bischof et al. 1996], the subroutine that computes the solution of the system equations of motion is differentiated in order to obtain the sensitivity of the solution vector. The differentiated version of the subroutine is not only used to compute the sensitivities solution vector, but also to evaluate the derivative of the algorithm by which the solution is computed. The system accelerations ( $\ddot{\mathbf{q}}_r, \ddot{\mathbf{w}}$ ) and the sensitivity solution vector of Equation (6), ( $\ddot{\mathbf{q}}_{r_b}, \ddot{\mathbf{w}}_b$ ), are obtained simultaneously.

Due to the coordinate reduction, which uses component mode synthesis, the nodal displacements of the flexible body are described by Equation (5). The sensitivity of the nodal displacement is obtained by computing the derivative of this equation with respect the design variables written as

$$\frac{d\mathbf{u}'}{d\mathbf{b}} = \frac{\partial \mathbf{X}}{\partial \mathbf{b}} \mathbf{w} + \mathbf{X} \frac{\partial \mathbf{w}}{\partial \mathbf{b}} = \mathbf{X}_b \mathbf{w} + \mathbf{X} \mathbf{w}_b, \quad (17)$$

where  $\mathbf{X}_b$  are the sensitivities of the eigenmodes. The relation expressed in Equation (17) transforms the modal sensitivities to nodal sensitivities. Haftka and Gürdal [1992] suggests evaluating this transformation by the fixed-mode approach, in which the derivatives of vibration modes are neglected, or by the updated-mode approach, where the derivatives of vibration modes are accounted for. The fixed-mode approach is computationally less expensive but the updated-mode approach can occasionally be more accurate. The right-hand side of Equation (12) also depends on the sensitivities of the eigenmodes. Therefore, the same approach is used in the computation of the derivatives of the modal forces and in the derivatives of the modal stiffness matrix. The modal stiffness matrix derivative is computed in the updated-mode approach by

$$\frac{\partial}{\partial \mathbf{b}} (\mathbf{X}^T \mathbf{K}_{ff} \mathbf{X}) = \frac{\partial \mathbf{X}^T}{\partial \mathbf{b}} \mathbf{K}_{ff} \mathbf{X} + \mathbf{X}^T \frac{\partial \mathbf{K}_{ff}}{\partial \mathbf{b}} \mathbf{X} + \mathbf{X}^T \mathbf{K}_{ff} \frac{\partial \mathbf{X}}{\partial \mathbf{b}}, \quad (18)$$

while in the fixed-mode approach, it is obtained as

$$\frac{\partial}{\partial \mathbf{b}} (\mathbf{X}^T \mathbf{K}_{ff} \mathbf{X}) = \mathbf{X}^T \frac{\partial \mathbf{K}_{ff}}{\partial \mathbf{b}} \mathbf{X}. \quad (19)$$

The computation of the sensitivities of the eigenmodes is done using the Nelson scheme in the case of distinct eigenvalues. However, when repeated eigenvalues are a possibility, Ojavo's method is used [Dailey 1989].

**3.2. Derivative of the element stiffness matrix.** In this work, the design variables used for the laminate optimization problem are the fiber angles of each lamina that make up the laminate, denoted by vector  $\theta$ . Therefore, the derivative of the stiffness matrix of the composite flexible body with respect to the layers' orientations has to be accounted for. At the element level, in local coordinates, the stiffness matrix is given by Equation (8). In this equation, only the matrix  $D$  depends on the design variables. Thus, the sensitivity of this equation is given by

$$\frac{\partial \mathbf{K}_{ff}^{(e)}}{\partial \mathbf{b}} = \int_0^1 \int_0^{1-\eta} \left( \mathbf{B}^T \frac{\partial \mathbf{D}}{\partial \mathbf{b}} \mathbf{B} \right)^{(e)} |J| d\xi d\eta. \tag{20}$$

The elasticity matrix  $D$  depends on the submatrices  $D_m, D_b, D_{mb}$  and  $D_s$ , which are defined by Equation (9). The partial derivative of Equation (9) with respect to the design variables vector is

$$\begin{aligned} (D_m, D_b, D_{mb}, D_s)_b &= \left( \sum_{k=1}^n (D_m, D_b, D_{mb}, D_s)_k \right)_b \\ &= \sum_{k=1}^n (C_{3 \times 3b}^1 H_1, C_{3 \times 3b}^1 H_2, C_{3 \times 3b}^1 H_3, C_{2 \times 2b}^1 H)_k \end{aligned} \tag{21}$$

with

$$(\mathbf{C}_b)_k = \left( \frac{\partial \mathbf{T}^T}{\partial \mathbf{b}} \bar{\mathbf{C}} \mathbf{T} + \mathbf{T}^T \frac{\partial \bar{\mathbf{C}}}{\partial \mathbf{b}} \mathbf{T} + \mathbf{T}^T \bar{\mathbf{C}} \frac{\partial \mathbf{T}}{\partial \mathbf{b}} \right)_k. \tag{22}$$

In Equation (22)  $(\bar{\mathbf{C}}_b)_k$  is the sensitivity of the material matrix of elastic coefficients for the layer  $k$  expressed in the local body frame, and  $(\partial \mathbf{T} / \partial \mathbf{b})_k$  is the sensitivity of the transformation matrix relative to the design variables. Matrix  $\mathbf{T}$  represents the transformation between the local body frame and the material coordinate systems for layer  $k$ . The element mass matrix does not depend on the design variables therefore the partial derivative of this matrix with respect the design variables is null.

### 4. Optimization criteria

The different optimization problems in multibody systems lead, in general, to different criteria functions and design constraints. The objective functions most widely used in multibody problems are of one of two types: maximum or minimum values and the integral type. Consider a general multibody response defined by function  $f_0(\mathbf{b}, \mathbf{z}, \boldsymbol{\lambda}, t)$ , which is dependent on time and on the state and design variables. In multibody systems, all the terms present in the equations of motion may be functions of the design parameters. In a compact form the problem objective functions are given by Chang and Nikravesh [1985]:

$$\Psi_i = \Psi_i(\mathbf{b}, \mathbf{z}, \boldsymbol{\lambda}, t), \quad i = 0, \dots, n, \tag{23}$$

where the state vector  $\mathbf{z}$  includes the coordinates, velocities and accelerations. The variables of the state vector may depend on time and on the design variables. Therefore, the dependency of the state variables on the design variables and time is explicitly written as

$$\mathbf{z}(\mathbf{b}, t) = (\mathbf{q}(\mathbf{b}, t), \dot{\mathbf{q}}(\mathbf{b}, t), \ddot{\mathbf{q}}(\mathbf{b}, t)). \tag{24}$$

The dependencies of the state variables on the design variables are explicitly taken into account by the automatic differentiation tool that uses the chain rule to calculate the sensitivities.

**4.1. Mini-max optimization problem.** The min-max optimization problem, for the time interval between  $t_i$  and  $t_e$  is stated as

$$\text{minimize } \Psi_0^{\max} = \max f_0(\mathbf{b}, \mathbf{z}, \boldsymbol{\lambda}, t), \quad t_i \leq t \leq t_e, \quad (25)$$

where the problem consists in the minimization of the maximum value of a specific function during a given time interval. The use of the maximum value of a time dependent function response as the objective function makes it a more difficult problem to solve. This type of objective function may appear, for instance, when the minimization of the maximum value of acceleration or force in a given point of a body is required during dynamic analysis. In this optimization problem two situations can occur:

- (1) The instant in which the function is at the maximum value is unique and perfectly defined. In this case, during the optimization process the instant  $t_m$  is not dependent on the design variables, and therefore the objective function (25) can be replaced by a simpler objective function as

$$\text{minimize } \Psi_0^{\max} = \max f_0(\mathbf{b}, \mathbf{z}(t_m), \boldsymbol{\lambda}(t_m)). \quad (26)$$

- (2) The instant in which the function is at the maximum value, varies during the optimization process. One form of dealing with this problem is to introduce an extra design variable and make the objective function equal to the value of that variable [Haftka and Gürdal 1992; Kim and Choi 1996]:

$$\text{minimize } \Psi_0 = b_{n+1} \quad (27)$$

with the additional time-dependent constraint

$$\Psi_{n+1} = f_0(\mathbf{b}, \mathbf{z}, \boldsymbol{\lambda}, t) - b_{n+1} \leq 0, \quad t_i \leq t \leq t_e. \quad (28)$$

The constraint given by Equation (28), when added to the total number of constraints ensures, that the dynamic response is below the maximum value defined by the auxiliary variable  $b_{n+1}$ . This approach poses some difficulties for the search direction in the optimization algorithm and can lead to small steps in the line search method, or even to a stall of the process. To overcome these difficulties, Kim and Choi [1996] proposed to handle directly the maximum value point only in the optimization process.

**4.2. Minimization of an integral type criteria.** The integral type objective function may be used to represent mean values of the response over time, accumulated values, or other special criteria. For a response  $f_0(\mathbf{b}, \mathbf{z}, \boldsymbol{\lambda}, t)$  of the dynamic system, the objective function is [Eberhard et al. 2003]

$$\Psi_0 = G_0(\mathbf{b}, \mathbf{z}_{t_e}, \boldsymbol{\lambda}_{t_e}, t_e) + \int_{t_i}^{t_e} f_0(\mathbf{b}, \mathbf{z}, \boldsymbol{\lambda}, t) dt, \quad (29)$$

where  $f_0(\mathbf{b}, \mathbf{z}, \boldsymbol{\lambda}, t)$  depends on the dynamic behavior during the complete time interval  $[t_i, t_e]$ , while  $G_0$  considers only the final state. This type of objective function is most common in vehicle design. Comfort or injury criteria are defined by integral type functions and often are used in the optimization process.

**4.3. Time-dependent constraints.** Mathematical programming algorithms generally cannot deal with parametric constraints such as

$$\Psi_i = f_i(\mathbf{b}, \mathbf{z}(t), \boldsymbol{\lambda}(t), t) \leq c, \quad t_i \leq t \leq t_e, \quad (30)$$

or even with constraints such as the one described by Equation (28). Such constraints have to be reformulated to remove their time dependency. During the simulation the function value can only be obtained for discrete time points. The most straightforward way to remove the time dependency of the original constraint is to discretize the time interval into time points. Then, the original constraint represented by Equation (30) is replaced by  $n_{tp}$  constraints written as [Haug and Arora 1979]:

$$\Psi_i = f_i(\mathbf{b}, \mathbf{z}(t_k), \boldsymbol{\lambda}(t_k), t_k) \leq c, \quad k = 1, \dots, n_{tp}. \quad (31)$$

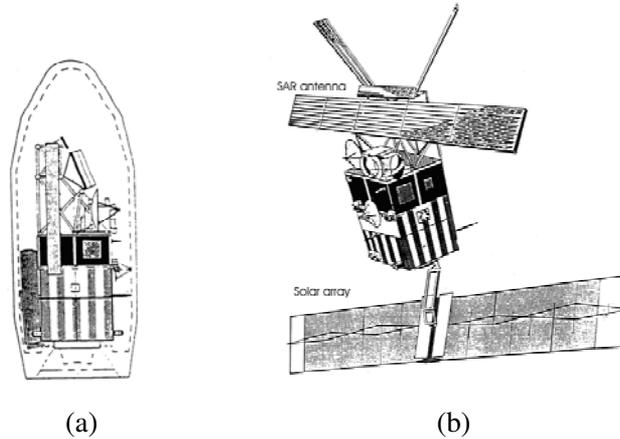
The distribution of the time points has to be sufficiently dense to avoid large constraint violations between two adjacent time points [Hsieh and Arora 1984]. Thus, discretizing time-dependent constraints can significantly increase the number of constraints, and thereby the cost of optimization [Haftka and Gürdal 1992]. In order to reduce the number of constraints, a first alternative consists of replacing the original constraints by an equivalent integrated constraint, which averages the severity of the constraint over the time interval. Hsieh and Arora [1984] showed that from an optimization theory point of view, the constraints described by Equation (30) and equivalent integral constraints are different. In fact, an equivalent integral constraint represents the behavior of the time dependent constraint  $f_i(\mathbf{b}, \mathbf{z}(t), \boldsymbol{\lambda}(t), t)$  on the complete time domain by a single value  $\Psi_i^e$ , leading to a loss of information. As a consequence, equivalent constraints tend to blur the design trends [Haftka and Gürdal 1992]. Hsieh and Arora [1984] and Grandhi et al. [1986] propose an alternative procedure that consists of exchanging the initial constraint given by (30) for a set of constraints of the type of Equation (31), in which  $n_{tp}$  is replaced by  $n_{ctp}$ , with  $n_{ctp} < n_{tp}$  being  $n_{ctp}$  the number of critical time points. These critical points are related with the existence of local maxima or minima of the function.

## 5. Optimization algorithms

In dynamic problems the evaluation of the system dynamic behavior requires the numerical integration of the equation of motion. The time dependency of this system makes these optimization problems more complex and requires that special techniques be used in the solution process. Both deterministic and stochastic optimization methods can be applied. Eberhard et al. [2003] has successfully used a stochastic evolution strategy in combination with a parallel computing environment to reduce computation time. However, in this work the Modified Method of Feasible Directions is used, which is a deterministic optimization method implemented in the DOT optimization routines library [Vanderplaats 1992]. In order to calculate the gradients, the direct differentiation method is used, the sensitivities being obtained by the automatic differentiation program ADIFOR [Bischof et al. 1996].

## 6. Optimization of a satellite unfolding process

The proposed methodology is demonstrated through the optimization of a complex multibody system made of composite material. The technical system modeled within this application consists in the unfolding process of a satellite antenna, the Synthetic Aperture Radar (SAR) antenna, which is a part of



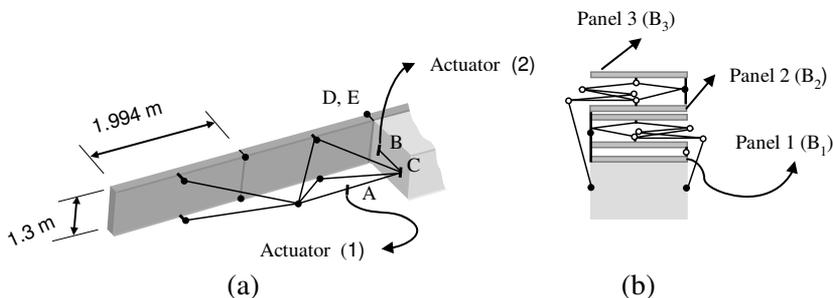
**Figure 3.** The SAR antenna in the (a) folded and (b) unfolded configurations.

the European research satellite ERS-1. The model of this antenna has been the object of different studies in several studies in multibody dynamics being first proposed by Hiller [1983] and Anantharaman and Hiller [1991].

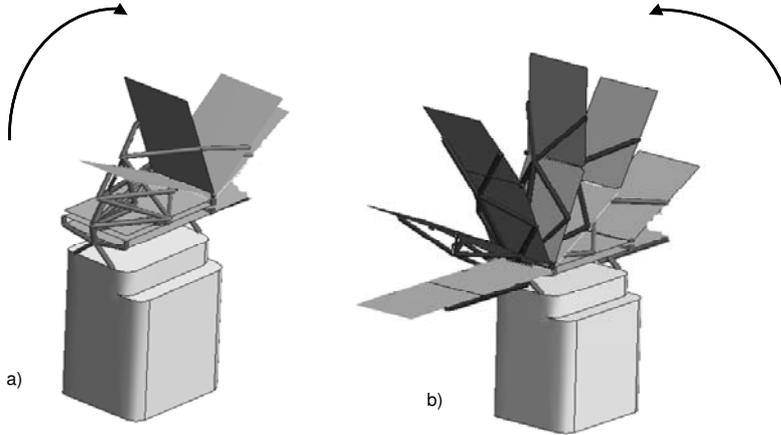
**6.1. Description of the SAR antenna.** The folding antenna shown in Figure 3 is achieved through a relatively complex spatial mechanism. Both the solar array and the SAR antenna of the ERS-1 satellite have the same configuration and share the same kinematic features. During transportation the antenna and the solar array are folded, as shown in Figure 3a, in order to occupy as small a space as possible. After unfolding, the mechanical components take the configuration represented in Figure 3b.

The SAR antenna consists of two identical subsystems, each with three coupled planar four-bar links that unfold two panels on each side. The central panel is attached to the main body of the satellite. Each unfolding system has two degrees of freedom, driven individually by actuators located in the joints A and B, shown in Figure 4.

The unfolding process consists of two phases, schematically represented in Figure 5. In the first phase the panel 3 is rolled out, about an axis normal to the main body, by a rotational spring-damper-actuator in joint A, while the panel 2 is held down by locking joints D and E, as shown in Figure 5a. The second phase begins with joint A locked, the panels 2 and 3 being swung out to the final position by a rotational



**Figure 4.** The SAR antenna: (a) one half unfolded state; (b) folded antenna.

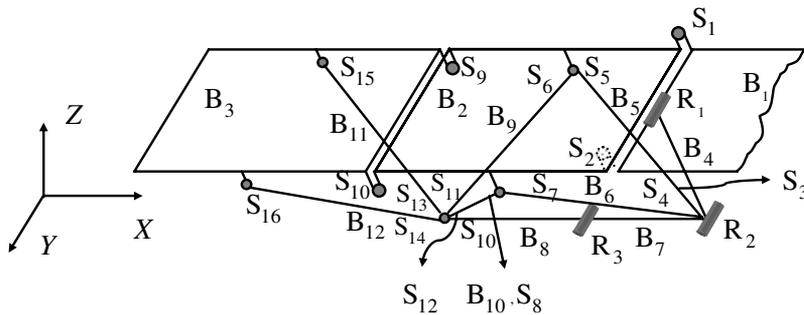


**Figure 5.** Unfolding process of the SAR antenna: (a) first phase; (b) second phase.

spring-damped-actuator in joint B, as observed in Figure 5b. The second half of the antenna, which has been omitted in Figures 4 and 5, is unfolded in the same way as the first half shown here. When the complete antenna is deployed all five panels are aligned in the final configuration.

The model used for one half of the folding antenna, schematically depicted Figure 6, is composed of 12 bodies ( $B_1$  a  $B_{12}$ ), 16 spherical joints ( $S_1$  a  $S_{16}$ ) and 3 revolute joints ( $R_1$ ,  $R_2$ ,  $R_3$ ). The central panel is attached to the satellite, defined as body  $B_1$ , which has mass and inertias much higher than the remaining bodies.

In the first phase of the unfolding antenna a rotational spring-damper-actuator is applied to the revolute joint  $R_3$ . For the second phase, the revolute joint  $R_3$  is locked and the system is moved to the next equilibrium position by a spring-damper-actuator positioned in joint  $R_1$ . Each panel is 1.994 m long by 1.3 m wide and has a thickness of 2 mm. The linkage between the panels and the four-bar linkage mechanism is assured by a set of supports, six in body 2 and four in body 3. All truss members have a uniform circular cross-section [Neto 2005].



**Figure 6.** Multibody model of the SAR antenna.

	1st Layer	2nd Layer	3rd Layer	4th Layer
Lay-up 1	0°	0°	0°	0°
Lay-up 2	0°	90°	90°	0°
Thickness (m)	0.0005	0.0005	0.0005	0.0005

**Table 1.** Characteristics of the two lay-ups considered for the composite panels.

**6.2. First phase of the antenna unfolding process.** The material used in the different components of the antenna is a carbon reinforced plastic IM6/SC1081 where the matrix is made of Epoxy SC1081 and the fibers are made of Carbon IM6. Note that the material model used here is not necessarily that of the real satellite antenna, as the characteristics of the material are not publicly available. The properties of the composite material, for a single layer with an orientation of 0° relative to the  $X$  axis are:  $E_1 = 177$  GPa;  $E_2 = 10.8$  GPa;  $G_{12} = G_{13} = 7.6$  GPa;  $G_{23} = 8.504$  GPa;  $\nu_{12} = 0.27$ ; with a specific mass of 1600 Kg/m<sup>3</sup>. Two different laminates with four layers in each, described in Table 1, are considered as potential design solutions.

In flexible multibody models the use of all the nodal degrees of freedom, resulting from the model of the complex system, as generalized coordinates is not viable. The application of the modal superposition technique in this kind of problem, characterized by linear elastic deformations, can be done without compromising accuracy. By using of a small set of the modes of vibration associated to the lower frequencies it is possible to reproduce the structural deformations of the panels with a small number of generalized elastic coordinates.

The modes of vibration for all flexible bodies in the antenna are obtained by performing a modal analysis of each one of the flexible bodies independently. The structural attachment conditions used in the eigenproblem are the same as those used to fix the body coordinate system, that is, the node in the center of mass is fixed to the body fixed frame. In this manner the free rigid body modes are removed.

In Tables 2 and 3 the 14 lowest frequencies are presented for panels 2 and 3 with composite material lay-ups 1 and 2, respectively. The modes corresponding to the two lower frequencies are almost rigid modes, resulting from the flexibility around a fixed node. However, these modes also represent deformation of the panels and cannot be neglected.

The actuator that is applied in revolute joint  $R_3$ , to initiate the satellite unfolding process, is modeled as a nonlinear spring and damper actuator. The spring-damper-actuator is described by piecewise-linear characteristics given by:

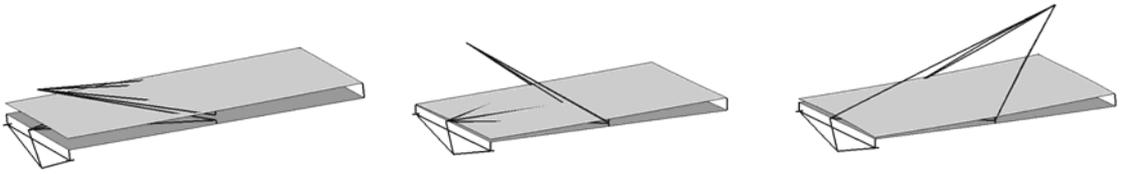
$$M(\theta, \dot{\theta}) = c\dot{\theta} + \begin{cases} 0.10 + 9.00(3.12 - \theta) & 3.08 < \theta \leq 3.12 \\ 0.45 + 60.41(3.08 - \theta) & 3.02 < \theta \leq 3.08 \\ 4.03 - 5.19(3.02 - \theta) & 2.63 < \theta \leq 3.02 \\ 2.00 & 0.20 < \theta \leq 2.63 \\ 10.00\theta & -0.20 \leq \theta \leq 0.20 \\ -2.00 & -0.20 > \theta, \end{cases} \quad (32)$$

Mode	Panel 2 Frequency [Hz]	Panel 3 Frequency [Hz]
1	0.990	0.992
2	1.457	1.460
3	1.677	1.681
4	1.746	1.749
5	4.000	4.001
6	4.609	4.620
7	6.099	6.118
8	6.814	6.850
9	8.538	8.564
10	8.578	8.583
11	12.434	12.451
12	12.828	12.833
13	14.354	14.404
14	14.415	14.485

**Table 2.** First 14 natural modes of vibration for panels 2 and 3 with the composite material lay-up 1.

Mode	Panel 2 Frequency [Hz]	Panel 3 Frequency [Hz]
1	1.311	1.313
2	1.563	1.566
3	1.694	1.699
4	2.334	2.336
5	4.719	4.719
6	5.755	5.770
7	6.220	6.242
8	7.388	7.427
9	12.832	12.844
10	12.953	12.971
11	13.626	13.685
12	13.829	13.873
13	15.282	15.303
14	15.969	15.986

**Table 3.** First 14 natural modes of vibration for panels 2 and 3 with the composite material lay-up 2.



**Figure 7.** Configuration of the composite panels with the original damped spring-actuator.

where the damping coefficient used is  $c = 0.5 \text{ Nms}$ .

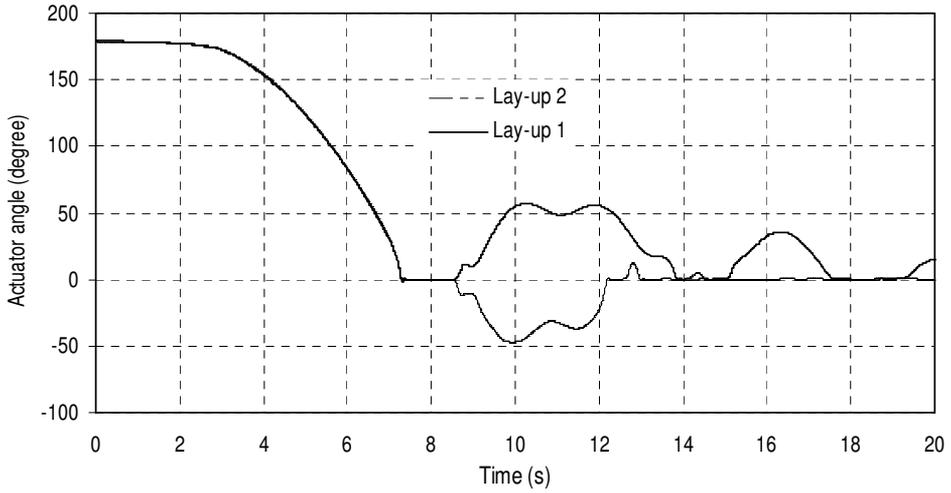
The actuation law presented here is different from that reported by [Anantharaman and Hiller \[1991\]](#), which was used to model the SAR antenna with panels made of isotropic material. In fact, when the actuation law used by Anantharaman and Hiller is used for the composite flexible models the satellite antenna is driven to a different equilibrium state than that obtained in the rigid model. The trusses connected to the actuator quickly reach their equilibrium, but panel 3 hardly moves because the unfolding trusses break through the panel, as represented in [Figure 7](#). This behavior is clearly unfeasible because contact between trusses and panels would take place, preventing such penetration from happening. Therefore, the reported results show that due to the deformations of the trusses the undesirable contacts between trusses and panels are possible if the high torques associated to the original actuator have been maintained. Consequently, the solution is to apply a ‘softer’ actuation law, in the sense of preventing such contact.

The problems associated with the unfolding of an isotropic flexible model due to the actuator deployment law have been identified by [Anantharaman and Hiller \[1991\]](#), and the solution found was to modify that actuation in order to prevent the wrong deployment mechanism, which is in essence similar to the solution adopted here. When using composite material models, the problem of the first phase of the unfolding process increases in importance not only because the bending of the panels is significant but also because torsional modes come in play. In [Figure 8](#) the variation of the actuator angle during the simulation period for the composite models is presented.

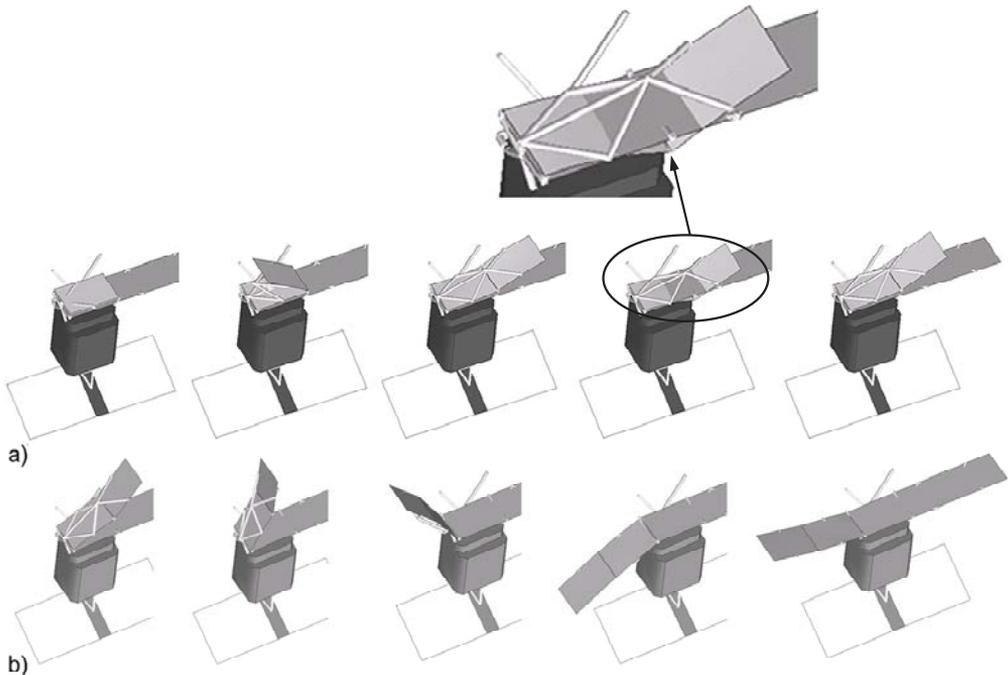
[Figure 8](#) shows that the two models lead to similar simulation results. However, it is observed that after the equilibrium positions are reached for both models, in the period from 7 to 8 s, the direction of rotation of the truss members of the panels made with the lay-up 1 is opposite to that of the same truss members of the model made with lay-up 2. This discrepancy can result from the difference between the vibration modes of the both models. In fact, the lay-up 1 has no layers with the  $90^\circ$  orientation, thus the stiffness of this model in the Y direction is smaller than that observed with lay-up 2. A similar difference in stiffness is also visible in the X direction of the lay-up 1.

When observing the fourth frame of the unfolding in [Figure 9a](#), it can be noticed that the flexible model of the satellite antenna predicts interference between panel 2 and panel 1, which is attached to the base satellite, when bodies  $B_7$  and  $B_8$  get aligned. This can be perceived as a flaw in the design of the unfolding process of the satellite that requires being fixed. If not detected, the interference would cause impact between the panels eventually leading to their failure.

**6.3. Optimization of the SAR antenna.** In this section the multibody model of the SAR antenna is used within the framework of an optimization problem. The flexibility of the panels of the SAR antenna is fundamental for the functional requirements of the antenna. The use of stiffer panels can improve the



**Figure 8.** Actuator angle during the first phase of deployment for different composite material lay-ups.



**Figure 9.** Configuration of the antenna unfolded process: a) first phase b) second phase.

Panels	Design Variable	Lower Bound	Initial Value	Upper Bound
2 = 3	$\theta_1/\theta_2/\theta_2/\theta_1$	$-90^\circ/-90^\circ/-90^\circ/-90^\circ$	$55^\circ/-55^\circ/-55^\circ/55^\circ$	$+90^\circ/+90^\circ/+90^\circ/+90^\circ$

**Table 4.** Design variables for panels' layers' orientations in the satellite optimization.

time needed to unfold the antenna, during the first phase of the unfolding process, allowing for the use of a stiffer actuator on revolute joint  $R_3$ . Furthermore, the stiffness increase of the panels, in particular of panel 2, prevents the interference detected between panels in the first part of the unfolding process.

The multibody model of the flexible antenna for the first part of the unfolding process is composed of two flexible panels. Therefore, the antenna deformation energy of the panels for instant  $t_n$  is defined as

$$\begin{aligned}
 2U_m(\mathbf{w}_i, t_n) &= \sum_{i=2}^3 \mathbf{w}_i^T \mathbf{X}_i^T \mathbf{K}_{ff}^i \mathbf{X}_i \mathbf{w}_i \\
 &= \sum_{i=2}^3 \mathbf{w}_i^T \mathbf{\Lambda} \mathbf{w}_i = 2(U_2(\mathbf{w}_2, t_n) + U_3(\mathbf{w}_2, t_n)),
 \end{aligned} \tag{33}$$

where the index  $m$  refers to the model used and index  $i$  refers to the body number of panels of the multibody model SAR antenna. Equation (33) indicates that the deformation energy of the multibody model of the SAR antenna is obtained as a summation of the deformation energy of the two panels of the model. Then the function  $f_0 = 2U_m$  is used to optimize the SAR antenna.

The goal defined by Equation (29) represents an area defined by the curve of function  $f_0 = 2U_m$  during the simulation period  $t_i = 0 \text{ s} \leq t \leq t_e = 3 \text{ s}$ . The minimum value of the area may be achieved with a peak value of the maximum deformation energy of each panel that exceeds acceptable limits. In order to avoid this situation, the maximum value of the deformation energy, in each panel, is constrained to be

$$\Psi_i(\boldsymbol{\theta}) \leq c_i; \quad i = 2, 3. \tag{34}$$

The values  $c_i$  are defined as being the maximum values of deformation energy, in each panel, observed in the initial design. Therefore, in the initial design the optimization algorithm has two active constraints.

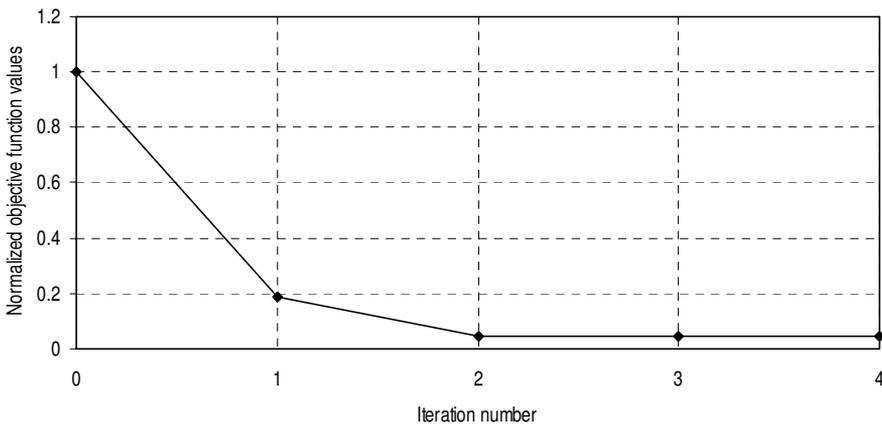
All material models considered herein are symmetric laminates with the number of layers being fixed. The simulation scenario considered is restricted to the first three seconds of the unfolding process, identified as the critical period. Two design variables are used in the optimization process, corresponding to the orientation of the layers that make up the laminate used to model panels 2 and 3. The initial design of laminate used in the panels is defined in Table 4. The optimization method used is the Modified Method of Feasible Directions (MMFD), as available in DOT [Vanderplaats 1992]. The analytic sensitivities computed by the direct differentiation method are used to compute the gradients required by the optimizer.

In Table 5 the optimization results are presented for the flexible multibody of the antenna. In Figure 10 the evolution of the objective function for the antenna flexible multibody model is showed, the progress of the design variables being shown in Figure 11. Figure 12 shows the actuator angle history during the first phase of the unfolding antenna for the original and optimum designs.

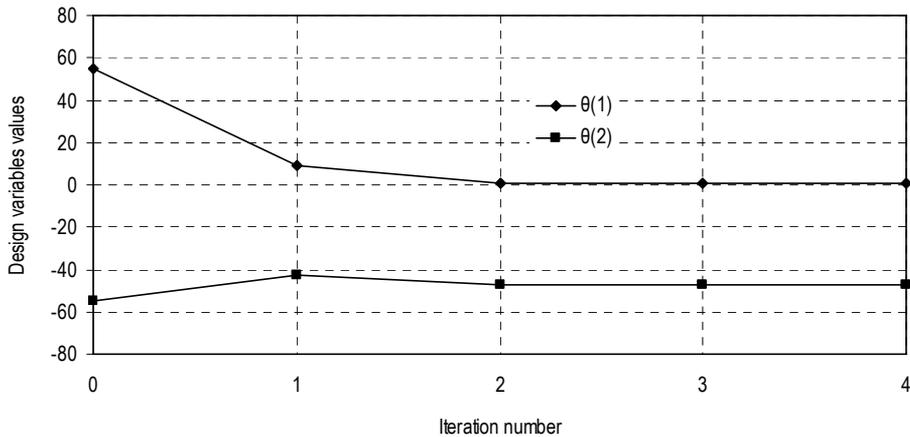
	Panel 2 (MFD)	Panel 3 (MFD)
Optimum <i>Layer orientations</i>	1.06°/−47°/−47°/1.06°	
Initial <i>objective function</i>	0.0219814	
Optimum <i>objective function</i>	0.00097180	
Reduction of <i>objective function</i>	95.6%	
Number of Constraints	2	
Number of Design Variables	2	
Active Constraints	0	
Active Side Constraints	0	
Function Calls	14	
Gradient Calls	4	
Number of Iterations	4	

**Table 5.** Summary of the optimization results of the satellite on the second optimization scenario.

In [Figure 10](#) it is possible to verify that the optimization procedure converges very fast to the optimum solution, reducing the deformation energy on the order of 95%. The largest variation in the design variables observed is associated with the outside layers of the laminate, as depicted in [Figure 11](#). The deformation energies of panels 2 and 3 are compared for the initial and the optimized models of the panels 2 and 3 in [Figures 13](#) and [14](#), respectively. By observing the initial and optimized deformation energy of panels 2 and 3 it is possible to conclude that the major contribution to the reduction of the deformation energy is verified in panel 2.



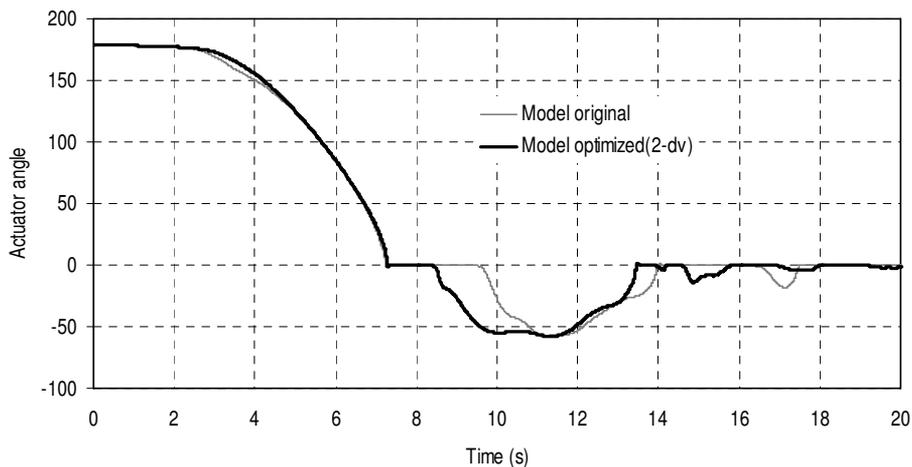
**Figure 10.** Evolution of the objective function during the optimization process.



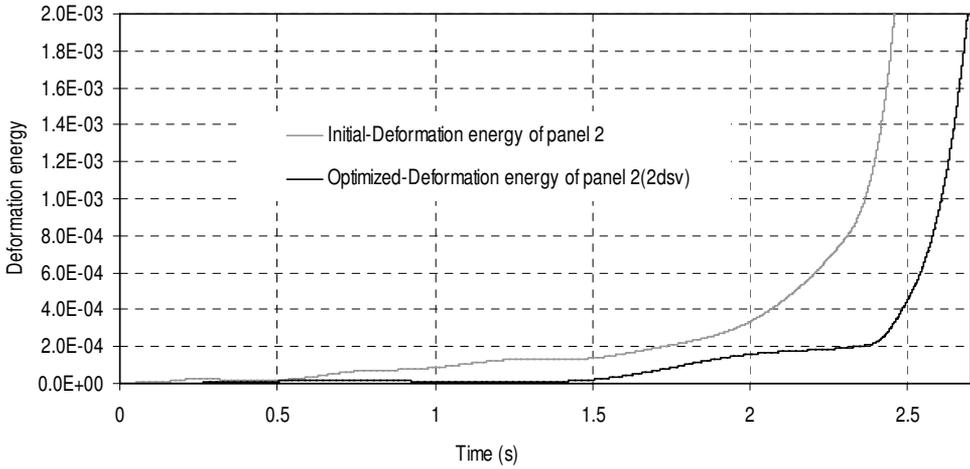
**Figure 11.** Evolution of the design variables during the optimization process.

## 7. Conclusions

A general method for the design optimization of flexible multibody systems made of composite materials has been presented in this work and demonstrated by an application to the design of the unfolding process of a satellite antenna. First, the correct choice of the optimization methods and the optimal problem definition is more complex when the nonlinear dynamic response of the systems is involved. Furthermore, the need to use analytic sensitivities instead of numerical sensitivities requires that expeditious methods of obtaining these are implemented in order to allow for the definition of more general objective functions, constraints and design variables. This has been achieved in this work by using an automatic differentiation tool to obtain the gradients required by the optimizer. Finally, the optimization of the nonlinear dynamic systems in general, and of the flexible multibody systems in particular, often present time-dependent



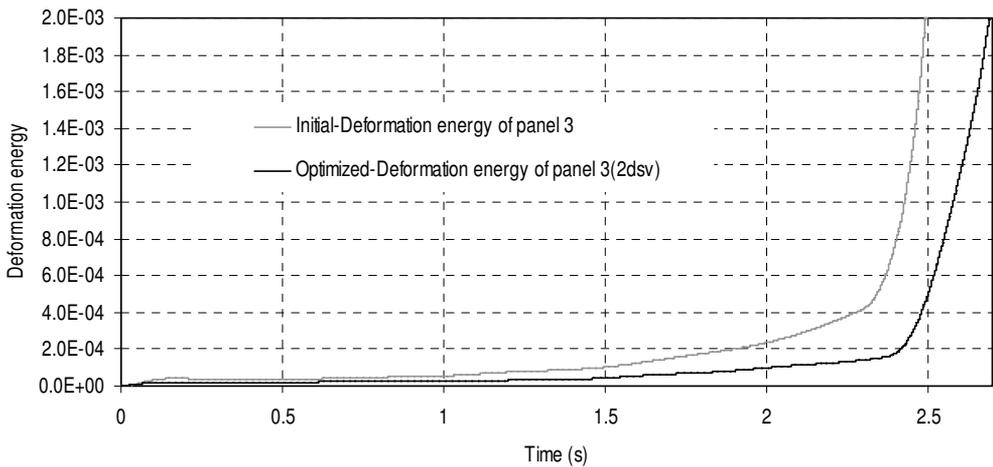
**Figure 12.** Actuator angle for the initial and optimum laminate of the antenna panels.



**Figure 13.** Deformation energy for the initial and optimum model of panel 2.

constraints that are difficult to tackle with common procedures. The use of min-max optimization is a form of handling most optimization problems of this type.

The application of the methodology developed for a complex system was demonstrated by considering the multibody model of the SAR antenna. The optimization method was applied to minimize the deformation energy of the SAR antenna panels. To get a stiffer antenna, the optimization problem was formulated as minimization problems of the deformation energy of each panel. The design variables of the optimization problem were the fiber orientations of the layers that form the lamination used to model the material properties of the panels. The design problem considers the case of finding optimum symmetric lamination with four layers only. In the optimization scenario two design variables were used to define the optimum lamination on both panels. The results of this application demonstrate that not



**Figure 14.** Deformation energy for the initial and optimum model of panel 3.

only are there feasible designs for the antenna in which interference between panels is avoided but also that the control over of the deformation energy of the antenna was possible. In the process it was shown that feasible designs for the actuation law during the deployment are obtained.

### Acknowledgements

We would like to thank Professors Manfred Hiller and Andrés Kecskeméthy, of the University of Duisburg, Germany, for providing the necessary data to model the SAR antenna and for useful discussions.

### References

- [Ambrósio 1996] J. Ambrósio, “Dynamics of structures undergoing gross motion and nonlinear deformations: a multibody approach”, *Comput. Struct.* **59**:6 (1996), 1001–1012.
- [Ambrósio 2003] J. Ambrósio, “Efficient kinematic joint descriptions for flexible multibody systems experiencing linear and non-linear deformations”, *Int. J. Numer. Methods Eng.* **56**:12 (2003), 1771–1793.
- [Ambrósio and Gonçalves 2001] J. Ambrósio and J. Gonçalves, “Complex flexible multibody systems with application to vehicle dynamics”, *Multibody Syst. Dyn.* **6**:2 (2001), 163–182.
- [Anantharaman and Hiller 1991] M. Anantharaman and M. Hiller, “Numerical simulation of mechanical systems using methods for differential-algebraic equations”, *Int. J. Numer. Methods Eng.* **32**:8 (1991), 1531–1542.
- [Bischof et al. 1992] C. Bischof, A. Carle, G. Corliss, A. Griewank, and P. Hovland, “ADIFOR: generating derivative codes from fortran programs”, *Sci. Program.* **1**:1 (1992), 11–29.
- [Bischof et al. 1996] C. Bischof, P. Khademi, A. Mauer, and A. Carle, “ADIFOR 2.0: automatic differentiation of Fortran 77 programs”, *IEEE Comput. Sci. Eng.* **3**:3 (1996), 18–32.
- [Bottasso et al. 2006] C. Bottasso, A. Croce, B. Savini, W. Sirchi, and L. Trainelli, “Aero-servo-elastic modeling and control of wind turbines using finite-element multibody procedures”, *Multibody Syst. Dyn.* **16**:3 (2006), 291–308.
- [Cavin and Dusto 1977] R. Cavin and A. Dusto, “Hamilton’s principle: finite-element methods and flexible body dynamics”, *AIAA J.* **15**:12 (1977), 1684–1690.
- [Chang and Nikravesh 1985] C. Chang and P. Nikravesh, “Optimal design of mechanical systems with constraint violation stabilization method”, *J. Mech. Transm. (Trans. ASME)* **107** (1985), 493–498.
- [Cook 1987] R. Cook, *Concepts and applications of finite element analysis*, 2nd ed., Wiley, New York, 1987.
- [Dailey 1989] R. L. Dailey, “Eigenvector derivatives with repeated eigenvalues”, *AIAA J.* **27**:4 (1989), 486–491.
- [Dmitrochenko et al. 2006] O. Dmitrochenko, W.-S. Yoo, and D. Pogorelov, “Helicoseir as shape of a rotating string I): 2D theory and simulation using ANCF”, *Multibody Syst. Dyn.* **15**:2 (2006), 135–158.
- [Duff et al. 1986] I. Duff, A. Erisman, and J. Reid, *Direct methods for sparse matrices*, Clarendon Press, Oxford [Oxfordshire], 1986.
- [Eberhard et al. 1999] P. Eberhard, W. Schiehlen, and D. Bestle, “Some advantages of stochastic methods in multicriteria optimization of multibody systems”, *Arch. Appl. Mech.* **69**:8 (1999), 543–554.
- [Eberhard et al. 2003] P. Eberhard, F. Dignath, and L. Kübler, “Parallel evolutionary optimization of multibody systems with application to railway dynamics”, *Multibody Syst. Dyn.* **9**:2 (2003), 143–164.
- [Gerstmayr and Schöberl 2006] J. Gerstmayr and J. Schöberl, “A 3D finite element method for flexible multibody systems”, *Multibody Syst. Dyn.* **15**:4 (2006), 305–320.
- [Gonçalves and Ambrósio 2005] J. Gonçalves and J. Ambrósio, “Road vehicle modeling requirements for optimization of ride and handling”, *Multibody Syst. Dyn.* **13**:1 (2005), 3–23.
- [Grandhi et al. 1986] R. Grandhi, R. Haftka, and L. Watson, “Design-oriented identification of critical times in transient response”, *AIAA J.* **24**:4 (1986), 649–656.

- [Haftka and Gürdal 1992] R. Haftka and Z. Gürdal, *Elements of structural optimization*, 3rd rev. exp. ed., Kluwer Academic Publishers, Dordrecht, 1992.
- [Hardeman et al. 2006] T. Hardeman, R. G. K. M. Aarts, and J. B. Jonker, “A finite element formulation for dynamic parameter identification of robot manipulators”, *Multibody Syst. Dyn.* **16**:1 (2006), 21–35.
- [Haug and Arora 1979] E. Haug and J. Arora, *Applied optimal design: mechanical and structural systems*, Wiley, New York, 1979.
- [He and Mcphee 2005] Y. He and J. Mcphee, “Multidisciplinary optimization of multibody systems with application to the design of rail vehicles”, *Multibody Syst. Dyn.* **14**:2 (2005), 111–135.
- [Heckmann et al. 2005] A. Heckmann, M. Arnold, and O. Vaculík, “A modal multifield approach for an extended flexible body description in multibody dynamics”, *Multibody Syst. Dyn.* **13**:3 (2005), 299–322.
- [Hiller 1983] M. Hiller, *Mechanische systeme*, Springer-Verlag, Berlin, Germany, 1983.
- [Hsieh and Arora 1984] C. Hsieh and J. Arora, “Design sensitivity analysis and optimization of dynamic response”, *Comput. Methods Appl. Mech. Eng.* **43**:2 (1984), 195–219.
- [Kim and Choi 1996] M. Kim and D. Choi, “A new approach to the min-max dynamic response optimization”, pp. 65–72 in *IUTAM-symposium on optimization of mechanical systems* (Stuttgart, 1996), edited by D. Bestle and W. Schiehlen, Kluwer, Dordrecht, 1996.
- [Kübler et al. 2005] L. Kübler, C. Henninger, and P. Eberhard, “Multi-criteria optimization of a hexapod machine”, *Multibody Syst. Dyn.* **14**:3–4 (2005), 225–250.
- [Lehner and Eberhard 2006] M. Lehner and P. Eberhard, “On the use of moment-matching to build reduced order models in flexible multibody dynamics”, *Multibody Syst. Dyn.* **16**:2 (2006), 191–211.
- [Liu et al. 2007] J.-F. Liu, J. Yang, and K. Abdel-Malek, “Dynamics analysis of linear elastic planar mechanisms”, *Multibody Syst. Dyn.* **17**:1 (2007), 1–25.
- [Møller et al. 2005] H. Møller, E. Lund, J. Ambrósio, and J. Gonçalves, “Simulation of fluid loaded flexible multibody systems”, *Multibody Syst. Dyn.* **13**:1 (2005), 113–128.
- [Neto 2005] M. Neto, *Optimization of flexible multibody systems with composite materials*, Ph. D. Thesis, Mechanical Engineering Department of Coimbra University, Coimbra, Portugal, 2005.
- [Neto et al. 2004] M. Neto, J. Ambrósio, and R. Leal, “Flexible multibody systems models using composite materials components”, *Multibody Syst. Dyn.* **12**:4 (2004), 385–405.
- [Nikravesh and Lin 2005] P. Nikravesh and Y.-S. Lin, “Use of principal axes as the floating reference frame for a moving deformable body”, *Multibody Syst. Dyn.* **13**:2 (2005), 211–231.
- [Pereira and Proença 1991] M. Pereira and P. Proença, “Dynamic analysis of spatial flexible multibody systems using joint co-ordinates”, *Int. J. Numer. Methods Eng.* **32**:8 (1991), 1799–1812.
- [Vanderplaats 1992] G. Vanderplaats, “DOT-Design Optimization Tools, Version 3.0”, VMA Engineering, Colorado Springs, CO, 1992.
- [Venkataraman and Haftka 1999] S. Venkataraman and R. Haftka, “Optimization of composite panels: a review”, in *Proc. of the 14th annual technical conference of the american society of composites*, Technomic Publishing, Dayton, OH, Sep. 27–29 1999.
- [Venkataraman and Haftka 2002] S. Venkataraman and R. Haftka, “Structural optimization: what has Moore’s law done for us?”, in *43rd AIAA/ASME/ASCE/AHS/ASC structures structural dynamics and materials conference*, Denver, CO, 22–25 April 2002.
- [Vetyukov et al. 2006] Y. Vetyukov, J. Gerstmayr, and H. Irschik, “Modeling spatial motion of 3D deformable multibody systems with nonlinearities”, *Multibody Syst. Dyn.* **15**:1 (2006), 67–84.
- [Yoo and Haug 1986] W. Yoo and E. Haug, “Dynamics of flexible mechanical systems using vibration and static correction modes”, *J. Mech. Transm. (Trans. ASME)* **108** (1986), 315–322.

JORGE A. C. AMBRÓSIO: [jorge@dem.ist.utl.pt](mailto:jorge@dem.ist.utl.pt)

*Instituto de Engenharia Mecânica — Instituto Superior Técnico, Technical University of Lisbon, Av. Rovisco Pais, 1041-001, Lisbon, Portugal*

MARIA AUGUSTA NETO: [augusta.neto@dem.uc.pt](mailto:augusta.neto@dem.uc.pt)

*Departamento de Engenharia Mecânica, Faculdade de Ciência e Tecnologia da Universidade de Coimbra (Polo II), 3020 Coimbra, Portugal*

ROGÉRIO PEREIRA LEAL: [rogerio.leal@dem.uc.pt](mailto:rogerio.leal@dem.uc.pt)

*Departamento de Engenharia Mecânica, Faculdade de Ciência e Tecnologia da Universidade de Coimbra (Polo II), 3020 Coimbra, Portugal*

# GEOMETRIC ANALYSIS OF THE DYNAMICS OF A DOUBLE PENDULUM

JAN AWREJCEWICZ AND DARIUSZ SENDKOWSKI

In this paper we make use of Riemannian geometry to analyze the dynamics of a simple low dimensional system with constraints, namely a double physical pendulum. The dynamics are analyzed by means of the Jacobi–Levi–Civita equation and its solutions. We show that this geometrical approach is in qualitative agreement with the classical techniques devoted to the study of dynamical systems.

## 1. Introduction

The classical approach to analysis of Hamiltonian systems has been widely applied, providing a classical explanation of the onset of chaos in these systems. In addition to the classical techniques for analyzing Hamiltonian systems, the geometric approach plays an important role. The geometric approach is based on the relation between Riemannian geometry and Hamiltonian dynamics, but is distinct from the geometric formulation of Hamiltonian mechanics in terms of symplectic geometry. This technique has been successfully applied [Cerruti-Sola and Pettini 1995; 1996, Casetti et al. 1996; Di Bari and Cipriani 1998; Casetti et al. 2000], especially to systems with many degrees of freedom. It has also been widely applied in general relativity [Szydłowski 2000] and to low dynamical systems with a nondiagonal metric tensor [Awrejcewicz et al. 2006]. It is believed that the geometric approach can provide an alternative to the classical explanation for the onset of chaos in Hamiltonian systems, which involves the homoclinic intersections [Lichtenberg and Lieberman 1992]. In the geometric approach to Hamiltonian dynamics, the analysis of dynamical trajectories and behavior of a system is cast into the analysis of a geodesic flow in a corresponding Riemannian space. The main tool of this approach is the so-called Jacobi–Levi–Civita (JLC) equation [do Carmo 1992; Di Bari and Cipriani 1998]. In general, the JLC equation is a system of second-order differential equations with respect to a geodesic length, and it describes the evolution of a tangent vector (so-called Jacobi vector) along the geodesic. Although there are many dynamical systems that can be described in this manner, there are some that can not, namely systems with velocity-dependent potentials. However, this kind of dynamical system can be analyzed by means of the Finslerian geometry [Di Bari and Cipriani 1998]. In this paper, we confine ourselves to conservative Hamiltonian systems, which can be geometrized within the Riemannian geometry approach. The main idea of this approach is to make use of the fact that Hamilton’s least action principle,

$$\delta \int_{t_1}^{t_2} L(q^i, \dot{q}^i, t) dt = 0,$$

---

*Keywords:* pendulum, chaos, Riemannian geometry.

This work has been supported (2005–2007) by Poland’s Ministry of Science and Higher Education (Grant No. 4 T07A 034 29).

can be connected with the condition of minimizing the arc-length functional in the Riemannian space between two points  $A, B$ . The condition has the form

$$\delta \int_A^B ds = 0.$$

The point is that motion of a Hamiltonian system can be viewed as the motion of a single virtual particle along a geodesic in a suitable Riemannian space  $\mathcal{Q}$ . From the above condition one can obtain the geodesics equation, which has the following form in local coordinates [do Carmo 1992]

$$\frac{d^2 q^i}{ds^2} + \Gamma_{jk}^i \frac{dq^j}{ds} \frac{dq^k}{ds} = 0. \quad (1)$$

The Riemannian space is endowed with a metric tensor, which is obtained from the dynamics of the analyzed system. In order to make use of the geometric approach we must choose a Riemannian manifold and the metric tensor. There are several choices for a Riemannian manifold and metric tensor: a space-time configuration manifold and the Eisenhart metric [Szydłowski 1998], a configuration manifold and the Jacobi metric [Casetti et al. 1996; Cerruti-Sola and Pettini 1996], etc. In this paper we choose a configuration space of an analyzed system for a Riemannian manifold. Hence, the metric tensor is the Jacobi metric  $g$ , which is connected to the dynamics by the following relationship [Casetti et al. 2000]

$$g_{ij} = 2W a_{ij}(q), \quad W \equiv E - V,$$

where  $E$  is a total energy and  $V$  is potential energy. The matrix  $a$  is a kinetic energy matrix (we use the Einstein summation convention):

$$L = \frac{1}{2} \dot{q}^i a_{ij} \dot{q}^j - V. \quad (2)$$

This relationship follows from the Maupertuis principle, which gives

$$ds = 2W dt.$$

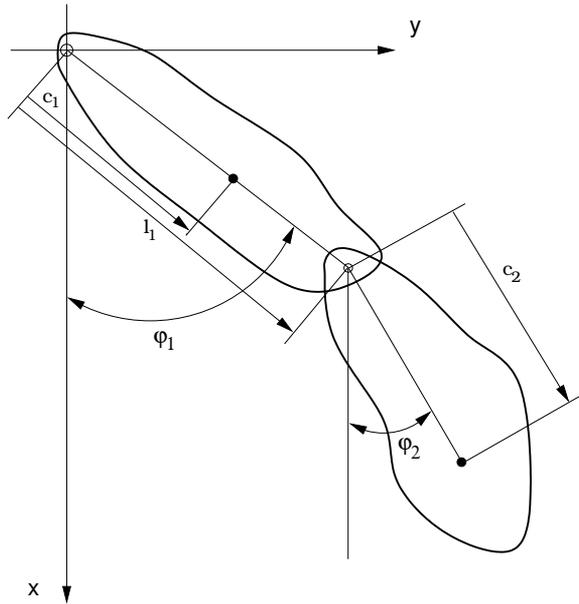
The main tool of the geometric approach, namely the JLC equation in a local coordinate system, has the following form [do Carmo 1992; Casetti et al. 2000]

$$\frac{\delta^2 J^n}{\delta s^2} + J^i \frac{dq^j}{ds} \frac{dq^k}{ds} R^n_{kij} = 0, \quad n = 1, 2, \dots, \dim \mathcal{Q}, \quad (3)$$

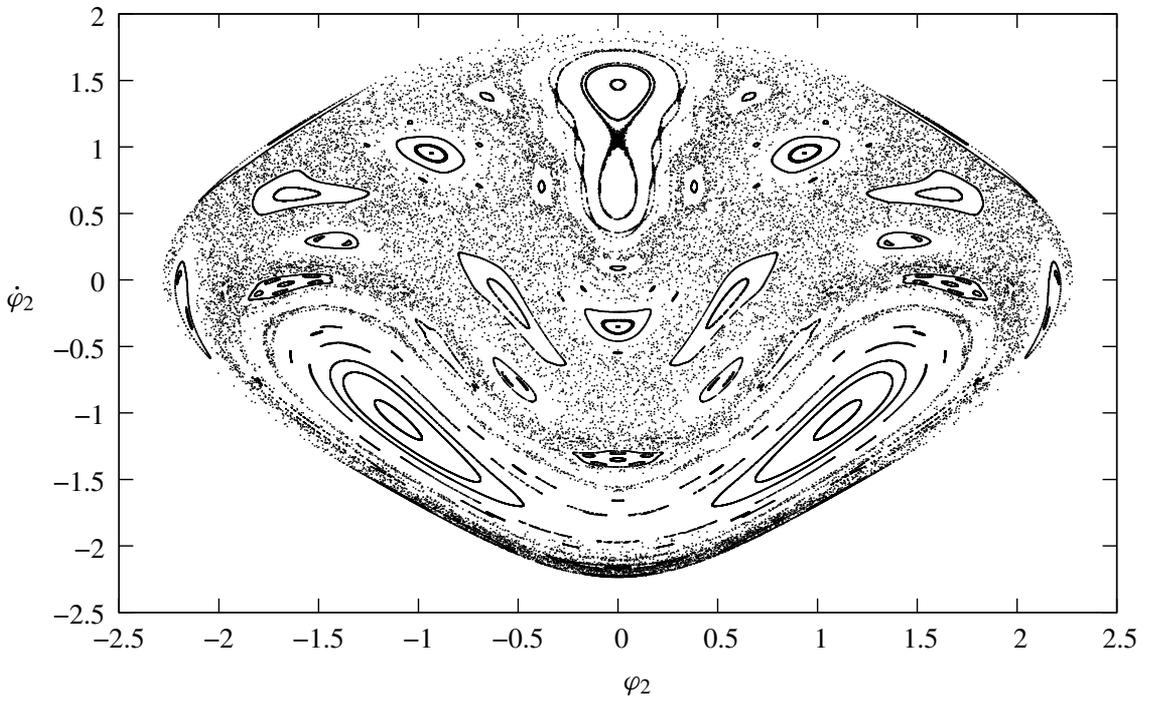
where  $q^j$  satisfy the geodesics Equation (1),  $J^n$  are components of the Jacobi vector,  $R^n_{kij}$  are components of the Riemann curvature tensor, and

$$\frac{\delta J^n}{\delta s} = \frac{dJ^n}{ds} + \Gamma_{kl}^n J^n \frac{dq^l}{ds}$$

are so-called absolute derivatives. The above equation has a similar form to the tangent dynamics equation, which is used to evaluate Lapunov's exponents. In fact, Equation (3) takes exactly the same form in the case of the Eisenhart metric [Casetti et al. 1996]. This means that there is a connection between the JLC equation and the tangent dynamics equation. Moreover, it is possible to find Lapunov's exponents using the Riemannian geometry approach. This has been done only for systems with many degrees of freedom and diagonal metric tensor [Casetti et al. 2000].



**Figure 1.** Double physical pendulum.



**Figure 2.** Poincaré section for  $\mathcal{E}=1.1$ .

Because we are interested in systems of only two degrees of freedom, the Riemannian space  $\mathcal{Q}$  is two-dimensional. This implies that we have only one nonzero component  $R_{2121}$  of the Riemann curvature tensor [Nakahara 1990; do Carmo 1992]. In this case, the JLC equation (3) takes the form

$$\frac{d^2\Psi}{ds^2} + \frac{R_{2121}}{\det g} \Psi = 0,$$

where  $\Psi$  is a normal component of the Jacobi vector relative to the geodesic. The tangent component of the Jacobi vector evolves only linearly in a geodesic length, so it does not contribute to the character of the solution [Di Bari and Cipriani 1998]. Next, making use of the fact that

$$\mathcal{R} = \frac{2R_{2121}}{\det g}, \tag{4}$$

we obtain a single differential equation which carries information about the system behavior

$$\frac{d^2\Psi}{ds^2} + \frac{1}{2}\mathcal{R}\Psi = 0, \tag{5}$$

where  $\mathcal{R}$  is the scalar curvature which, in general, is not periodic in  $\tau$ . At this point, we can see where a possible explanation of the onset of chaos in Hamiltonian system lies. The component  $\Psi$  of the Jacobi vector represents a distance between two nearby geodesics, which in turn represent trajectories of the analyzed system. The solutions of Equation (5) can exhibit exponential growth due to parametric excitations in the scalar curvature. Hence, this formulation and description of Hamiltonian dynamics gives us a qualitatively different explanation of the onset of chaos as a parametric instability of geodesics [Cerruti-Sola and Pettini 1996].

In order to solve Equation (5), we need to transform it into a differential equation with respect to the real time,  $t$ . Taking into account Equation (2) we find

$$\ddot{\Psi} - \frac{\dot{W}}{W}\dot{\Psi} + 2\mathcal{R}W^2\Psi = 0.$$

The above equation can be easily transformed into another form by means of the following substitution [Cerruti-Sola and Pettini 1996]

$$\Psi = J\sqrt{W},$$

which gives

$$\ddot{J} + \Omega(\tau)J = 0, \tag{6}$$

where

$$\Omega(\tau) \equiv \frac{1}{2} \left( \frac{\ddot{W}}{W} - \frac{1}{2} \left( \frac{\dot{W}}{W} \right)^2 + 4\mathcal{R}W^2 \right).$$

It should be emphasized here that  $\Omega$  is not, in general, periodic in  $\tau$ -time. Although  $\Omega$  is written as a function of  $\tau$ , it does not depend on  $\tau$  explicitly. In fact, it depends on a particular trajectory of the system.

### 2. The pendulum

In this paper we analyze a mechanical system with constraints, namely a double physical pendulum. The dynamics of the pendulum are described by the following lagrangian  $L$

$$L = \frac{1}{2} (m_1 c_1^2 + J_1 + m_2 l_1^2) \dot{\varphi}_1^2 + \frac{1}{2} (m_2 c_2^2 + J_2) \dot{\varphi}_2^2 + m_2 c_2 l_1 \dot{\varphi}_1 \dot{\varphi}_2 \cos (\varphi_1 - \varphi_2) - V (\varphi_1, \varphi_2),$$

where

$$V (\varphi_1, \varphi_2) = g (m_2 l_1 + m_2 c_2 + m_1 c_1) - g (m_1 c_1 + m_2 l_1) \cos \varphi_1 - m_2 g c_2 \cos \varphi_2,$$

$m_1$  and  $m_2$  are masses,  $J_1$  and  $J_2$  are moments of inertia, and  $c_1$  and  $c_2$  are the positions of centers of masses of the first and second link, respectively (see [Figure 1](#)). In order to cast the above lagrangian into a nondimensional form, we introduce the following scaling

$$\begin{aligned} \tau &\equiv t \sqrt{\frac{m_1 g c_1 + m_2 g l_1}{J_1 + m_1 c_1^2 + m_2 l_1^2}}, & \beta &\equiv \frac{J_2 + m_2 c_2^2}{J_1 + m_1 c_1^2 + m_2 l_1^2} \\ \kappa &\equiv \frac{m_2 c_2 l_1}{J_1 + m_1 c_1^2 + m_2 l_1^2}, & \mu &\equiv \frac{m_2 c_2}{m_1 c_1 + m_2 l_1}. \end{aligned}$$

Hence, the lagrangian takes the nondimensional form

$$\begin{aligned} L &= \frac{1}{2} \dot{\varphi}_1^2 + \frac{\beta}{2} \dot{\varphi}_2^2 + \kappa \dot{\varphi}_1 \dot{\varphi}_2 \cos \phi - 1 - \mu + \cos \varphi_1 + \mu \cos \varphi_2 \\ &= \frac{1}{2} (\dot{\varphi}_1 \dot{\varphi}_2) \mathbf{a} \begin{pmatrix} \dot{\varphi}_1 \\ \dot{\varphi}_2 \end{pmatrix} - 1 - \mu + \cos \varphi_1 + \mu \cos \varphi_2, \quad \phi \equiv \varphi_1 - \varphi_2, \end{aligned}$$

where

$$\mathbf{a} = \begin{pmatrix} 1 & \kappa \cos \phi \\ \kappa \cos \phi & \beta \end{pmatrix},$$

The dot over  $\varphi$  denotes  $\tau$  derivative. Using the Euler–Lagrange equations we obtain the equations of motion

$$\begin{cases} \ddot{\varphi}_1 = \frac{-\kappa \sin \phi (\kappa \cos \phi \dot{\varphi}_1^2 + \beta \dot{\varphi}_2^2) - \beta \sin \varphi_1 + \kappa \mu \sin \varphi_2 \cos \phi}{\beta - \kappa^2 \cos^2 \phi}, \\ \ddot{\varphi}_2 = \frac{\kappa \sin \phi (\dot{\varphi}_1^2 + \kappa \cos \phi \dot{\varphi}_2^2) - \mu \sin \varphi_2 + \kappa \sin \varphi_1 \cos \phi}{\beta - \kappa^2 \cos^2 \phi}. \end{cases}$$

### 3. Geometrization

Let us consider the Jacobi metric  $\mathbf{g}$  of the physical pendulum

$$\mathbf{g} = 2^{\mathcal{W}} \mathbf{a} = 2^{\mathcal{W}} \begin{pmatrix} 1 & \kappa \cos \phi \\ \kappa \cos \phi & \beta \end{pmatrix}, \quad \mathcal{W} \equiv \mathcal{E} - 1 - \mu + \cos \varphi_1 + \mu \cos \varphi_2.$$

Next, we find the connection coefficients  $\Gamma_{jk}^i$ :

$$\Gamma_{11}^1 = \frac{1}{2^{\circ}W \det \mathbf{a}} (2\kappa^2 \sin \varphi_1 \cos^2 \phi + {}^{\circ}W \kappa^2 \sin (2\phi) - \mu \kappa \sin \varphi_2 \cos \phi - \beta \sin \varphi_1),$$

$$\Gamma_{22}^2 = \frac{1}{2^{\circ}W \det \mathbf{a}} (2\mu \kappa^2 \sin \varphi_2 \cos^2 \phi - {}^{\circ}W \kappa^2 \sin (2\phi) - \beta \kappa \sin \varphi_1 \cos \phi - \beta \mu \sin \varphi_2),$$

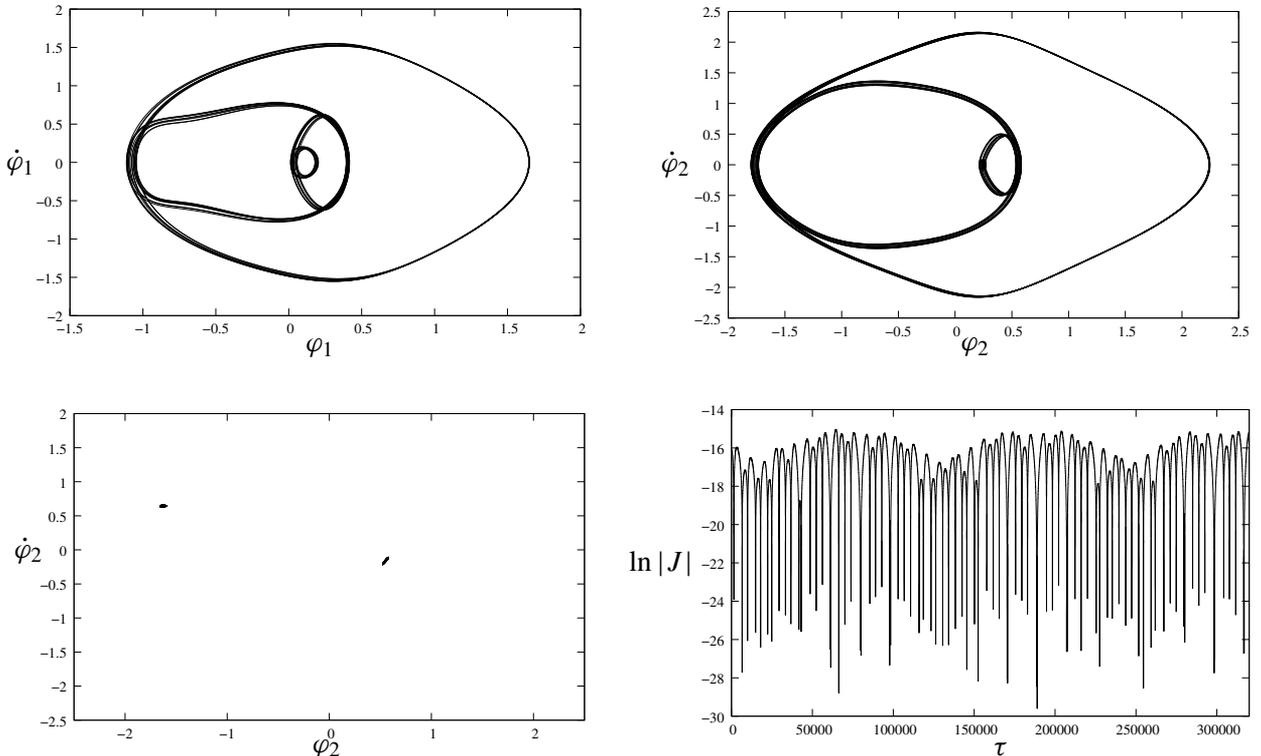
$$\Gamma_{11}^2 = \frac{1}{2^{\circ}W \det \mathbf{a}} (\mu \sin \varphi_2 - \kappa \sin \varphi_1 \cos \phi - 2^{\circ}W \kappa \sin \phi),$$

$$\Gamma_{22}^1 = \frac{\beta}{2^{\circ}W \det \mathbf{a}} (\beta \sin \varphi_1 - \mu \kappa \sin \varphi_2 \cos \phi + 2^{\circ}W \kappa \sin \phi),$$

$$\Gamma_{12}^2 = \frac{1}{2^{\circ}W \det \mathbf{a}} (\mu \kappa \sin \varphi_2 \cos \phi - \beta \sin \varphi_1),$$

$$\Gamma_{12}^1 = \frac{\beta}{2^{\circ}W \det \mathbf{a}} (\kappa \sin \varphi_1 \cos \phi - \mu \sin \varphi_2).$$

In a two-dimensional space there is only one nonzero component of the Riemann curvature tensor, namely,



**Figure 3.** Initial conditions:  $\varphi_2 = -1.63, \dot{\varphi}_2 = 0.63$ .

$$R_{2121} = \mu \cos \varphi_2 + 2^{\circ}W\kappa \cos \phi + \beta \cos \varphi_1 + \frac{1}{\mathfrak{W}} (\kappa \sin \varphi_1 \cos \phi - \mu \sin \varphi_2)^2 + \frac{\sin^2 \varphi_1 \det \mathfrak{a}}{\mathfrak{W}} - \frac{\kappa \sin \phi}{\det \mathfrak{a}} (\beta \kappa \sin \varphi_1 \cos \phi - \mu \kappa \sin \varphi_2 \cos \phi - \beta \mu \sin \varphi_2 + \beta \sin \varphi_1) - \frac{2^{\circ}W\kappa^3 \sin^2 \phi \cos \phi}{\det \mathfrak{a}}.$$

Making use of Equation (4) we find the scalar curvature:

$$\mathcal{R} = \frac{\kappa \cos(\phi)}{\mathfrak{W} \det \mathfrak{a}} - \frac{\kappa^3 \sin^2 \phi \cos \phi}{\mathfrak{W} \det^2 \mathfrak{a}} + \frac{\mu \cos \varphi_2 + \beta \cos \varphi_1}{2^{\circ}W^2 \det \mathfrak{a}} - \frac{\kappa \sin \phi (\beta \kappa \sin \varphi_1 \cos \phi - \mu \kappa \sin \varphi_2 \cos \phi - \beta \mu \sin \varphi_2 + \beta \sin \varphi_1)}{2^{\circ}W^2 \det^2 \mathfrak{a}} + \frac{\sin^2 \varphi_1}{2^{\circ}W^3} + \frac{(\kappa \sin \varphi_1 \cos \phi - \mu \sin \varphi_2)^2}{2^{\circ}W^3 \det \mathfrak{a}}.$$

Finally, inserting the obtained scalar curvature into Equation (6) we get the JLC equation for the physical pendulum.

### 4. Numerical simulations

The equations of motion have been numerically solved by means of the symplectic algorithm of Strömer-Verlet [Hairer et al. 2006], whilst the JLC equation (6) has been solved by the Dormand-Prince 5(4)

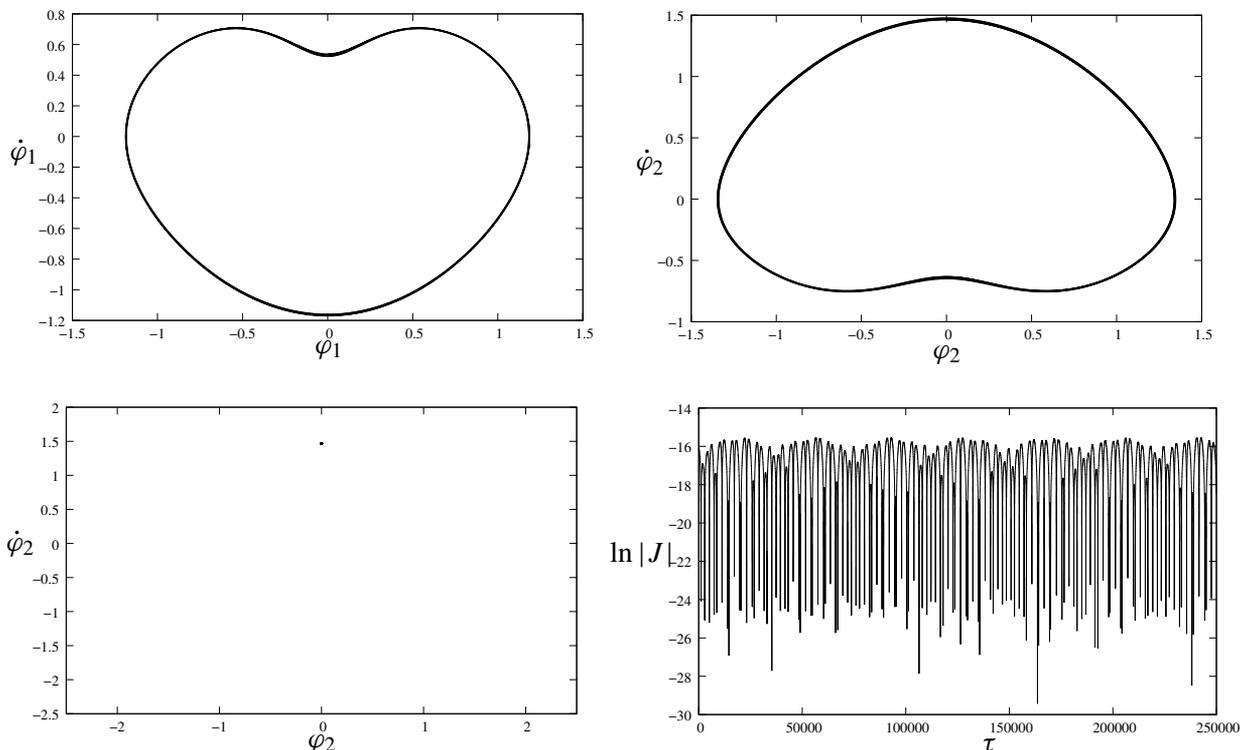
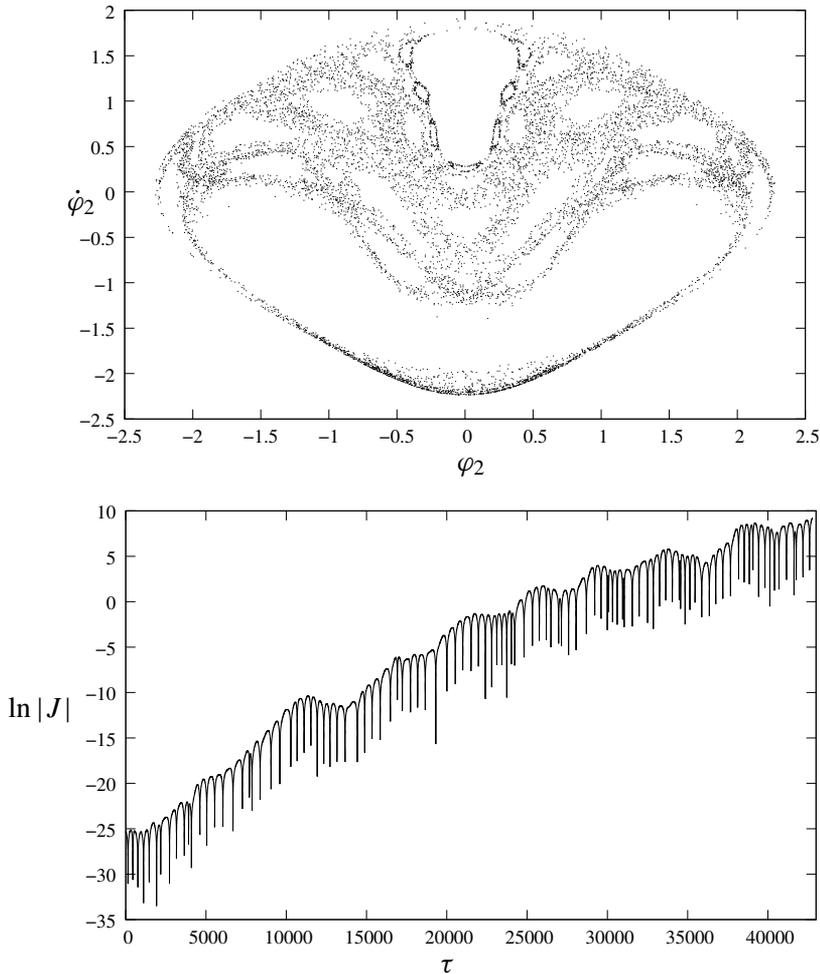


Figure 4. Initial conditions:  $\varphi_2 = 0, \dot{\varphi}_2 = 1.46$ .



**Figure 5.** Initial conditions:  $\varphi_2 = -1.93$ ,  $\dot{\varphi}_2 = -0.23$ .

algorithm with variable time-step size and the energy correction. Numerical simulation parameters were given the following values:  $\beta = 0.6$ ,  $\kappa = 0.4$ ,  $\mu = 0.66667$ . The simulation was performed for the total energy  $\mathcal{E} = 1.1$ . Below, we present the Poincaré section, in which one can observe chaotic regions as well as islands of regular behavior. Thus, we can analyze the system's behavior on the same energy level. The numerical results include three cases, namely two of them (Figures [Figure 3](#), and [Figure 4](#)) from regions of regular behavior and the last ([Figure 5](#)) one from the chaotic region. The initial conditions of the regular behavior cases have been taken from the interior of the regular islands, so that trajectories stay in regular regions regardless of numerical errors. The presented figures include two projections of the phase trajectories (only in the case of regular behavior), the corresponding Poincaré section of a particular trajectory, and the graph, which presents the evolution rate of a solution of the JLC equation. One can easily observe that in the case of regular trajectories (Figures [Figure 3](#), and [Figure 4](#)) the evolution of the Jacobi vector along the geodesic is bounded. However, in [Figure 5](#) we can observe the unbounded evolution of the Jacobi vector, which means that two nearby geodesics originating from

the neighborhood of the initial condition move away from each other and hence the distance between them grows exponentially. This is caused by the parametric resonance occurring in the JLC Equation (5).

## 5. Concluding remarks

We have applied the Riemannian approach to a low dimensional system with constraints, and have shown that the geometric approach gives results that are in qualitative agreement with those obtained from the classical approach. The existence of constraints is manifest in the metric tensor, which has a nondiagonal form in this case. Although the obtained results show that there is an agreement between classical and geometric approaches, a more thorough analysis is needed. The aim of this approach is to make use of the Riemannian geometry tools to gain information about a system's behavior without referring to the geodesic evolution. The geometric approach has already been applied to systems that have no constraints and many degrees of freedom [Di Bari and Cipriani 1998; Casetti et al. 2000]. However, systems with few degrees of freedom and constraints are more difficult to analyze in this manner. Nevertheless, the obtained results are very promising and enable us to work out a more analytical way to analyze such systems within the geometric approach.

## References

- [Awrejcewicz et al. 2006] J. Awrejcewicz, D. Sendkowski, and M. Kazmierczak, “Geometrical approach to the swinging pendulum dynamics”, *Comput. Struct.* **84**:24–25 (2006), 1577–1583.
- [do Carmo 1992] M. P. do Carmo, *Riemannian geometry*, Birkhäuser, Boston, 1992. Translated from the second Portuguese edition by Francis Flaherty. [MR 92i:53001](#) [Zbl 0752.53001](#)
- [Casetti et al. 1996] L. Casetti, C. Clementi, and M. Pettini, “Riemannian theory of Hamiltonian chaos and Lyapunov exponents”, *Phys. Rev. E* **54**:6 (1996), 5969–5984. [MR 97i:58134](#)
- [Casetti et al. 2000] L. Casetti, M. Pettini, and E. G. D. Cohen, “Geometric approach to Hamiltonian dynamics and statistical mechanics”, *Phys. Rep.* **337**:3 (2000), 237–341. [MR 2002a:82004](#)
- [Cerruti-Sola and Pettini 1995] M. Cerruti-Sola and M. Pettini, “Geometric description of chaos in self-gravitating systems”, *Phys. Rev. E* **51**:1 (1995), 53–64. [MR 97a:58115](#)
- [Cerruti-Sola and Pettini 1996] M. Cerruti-Sola and M. Pettini, “Geometric description of chaos in two-degrees-of-freedom Hamiltonian systems”, *Phys. Rev. E* **53**:1 (1996), 179–188.
- [Di Bari and Cipriani 1998] M. Di Bari and P. Cipriani, “Geometry and chaos on Riemann and Finsler manifolds”, *Planet. Space Sci.* **46**:11–12 (1998), 1543–1555.
- [Hairer et al. 2006] E. Hairer, C. Lubich, and G. Wanner, *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations*, 2nd ed., Springer Series in Computational Mathematics **31**, Springer, Berlin, 2006.
- [Lichtenberg and Lieberman 1992] A. J. Lichtenberg and M. A. Lieberman, *Regular and chaotic dynamics*, 2nd ed., Applied Mathematical Sciences **38**, Springer, New York, 1992. [MR 93c:58071](#) [Zbl 0748.70001](#)
- [Nakahara 1990] M. Nakahara, *Geometry, topology and physics*, Adam Hilger, Bristol, 1990. [MR 91j:58003](#) [Zbl 0764.53001](#)
- [Szydłowski 1998] M. Szydłowski, “The Eisenhart geometry as an alternative description of dynamics in terms of geodesics”, *Gen. Relat. Gravit.* **30**:6 (1998), 887–914. [MR 99b:58184](#) [Zbl 0997.37034](#)
- [Szydłowski 2000] M. Szydłowski, “The general relativity dynamics in the Eisenhart geometry”, *Chaos Solitons Fract.* **11**:5 (2000), 685–695. [MR 2001a:83014](#) [Zbl 0952.83008](#)

Received 7 Jul 2006. Accepted 20 Apr 2007.

JAN AWREJCEWICZ: [awrejcew@p.lodz.pl](mailto:awrejcew@p.lodz.pl)

*Technical University of Łódź, Department of Automatics and Biomechanics, Stefanowskiego St. 1/15, Łódź, 90-924, Poland*

DARIUSZ SENDKOWSKI: [dsend@p.lodz.pl](mailto:dsend@p.lodz.pl)

*Technical University of Łódź, Department of Automatics and Biomechanics, Stefanowskiego St. 1/15, Łódź, 90-924, Poland*

# NUMERICAL SIMULATION OF GRANULAR MATERIALS IN A ROTATING TUMBLER

HORACIO TAPIA-MCCLUNG

Using a simple numerical model based on particle dynamics, we perform two-dimensional simulations of granular materials inside a rotating tumbler. Still in its first stages of development, the model can quantitatively reproduce some of the general behavior observed in these systems, both for monodisperse and binary mixtures. Work is currently being done to validate the model and obtain results that are directly comparable to experiments.

## 1. Introduction

Until recently, two special cases of granular segregation have received the most attention: systems in which particles have the same size but differ in density (D-systems) and those in which particles have the same density but different size (S-systems). In both cases, the mechanisms that give rise to segregation are still not well understood. A more challenging case is that in which particles of different sizes and densities are present [Jain et al. 2005]. This last situation is of practical importance since many of the granular flows encountered in nature and industry involve particles of different sizes, shapes, and densities.

It is known that particles of higher density or smaller size segregate to the core of the granular bed while particles of lower density or larger size segregate to the outer edges and to the flowing layer. In mixed systems, it would be expected that the mechanisms that give rise to segregation compete to decrease or increase the final mixed state. Much of the experimental and numerical work already done to study segregation in a rotating tumbler has focused on bidisperse particle systems in which only one of the particle properties is varied (S and D systems), and only recently has a first step towards characterizing regimes of segregation for systems with mixed particles been done experimentally [Jain et al. 2005]. In this study it was found that when the different mechanisms all contribute to segregation, a *classical* behavior is observed, and when they oppose one another, a transition from a core composed of dense particles to a core composed of small particles occurs and mixing can be achieved if the denser particles are also bigger and if the ratio of particle size is greater than the ratio of particle density.

On the numerical side, granular flow has been approached by means of continuum-based models [Khakhar et al. 1997; Jain et al. 2005] and particle dynamics simulations [Dury and Ristow 1997], but in both cases, only the size or the density of the particles were different. In this work we present the preliminary results of a numerical study on granular materials placed in a two-dimensional rotating drum based on a particle dynamics simulation. Our final objective is to determine regimes of mixing and segregation and their dependence on system parameters such as particle properties, filling level of the

---

*Keywords:* granular materials, rotating drum, numerical simulations.

drum and angular velocity of rotation of the drum. Numerical simulations allow us to track the evolution of quantities that are not accessible in real experiments and to scan a wide range of system parameters in order to gain knowledge of the fundamental processes responsible for the complex collective behavior that is observed in granular materials.

## 2. The system

A rotating tumbler is probably one of the simplest and most common devices used to mix particles. In its most basic form, it is a cylindrical container that rotates about its axis of symmetry. Particles (or fluids) placed inside of it undergo different flow regimes that depend on many factors, the most evident one being the rotational velocity of the tumbler; but perhaps just as important as the angular velocity is the friction of the particles with the inside walls and with themselves.

In the case of granular materials, as the tumbler rotates, particles are transported inside the drum mostly by friction with the inner wall. For very small rotational velocities, the system of particles behaves as a rigid body. As the velocity is increased, individual avalanches begin to appear on the surface layer, carrying particles from one side of the rotating drum to the other. Further increase in the angular velocity above a critical value results in a continuous flow on the surface layer, and for even higher values of rotational velocity, the system reaches a centrifugal regime.

The choice of a rotating drum as a study case presents several advantages: rich behavior can be observed and, under appropriate conditions, the processes taking place in such a device can be considered almost two-dimensional. Industrial applications are wide and can provide a direct comparison between experimental and numerical work.

## 3. The numerical model

We model single grains by idealized spherical particles (discs in two dimensions) of radius  $r_i^0$  and mass  $m_i$  that interact only in pairs during collisions. The time evolution of each grain is obtained by numerically solving the coupled equations of motion using the velocity-explicit Verlet algorithm for temporal integration of second-order differential equations. This algorithm approximates the solution using a nonzero time step  $dt$  to update the positions at time  $t_{n+1} = (n + 1)dt$  from the knowledge of the position, the velocity and the forces acting on the particle at time  $t_n = ndt$ . The velocity is updated by using the information from the previous time step (at time  $t_n$ ) and the forces acting on the particle at the current time. Further details can be seen in [Andersen 1983; Allen and Tildesley 1989; Tapia-McClung and Grønbech-Jensen 2005]. To use this numerical method, the forces acting on a particle have to be known in advance in order to update the positions and velocities. Because of the nature of granular materials and the complexity of their collective behavior, it is difficult to know what the forces acting on each individual grain are [Schaefer et al. 1996]. In a numerical simulation we need a model simple enough to be treated computationally, that is able to account for the observed behavior of the system. This is particularly true for granular materials which are highly dissipative systems. In this work we have chosen a simple and numerically tractable model that accounts for the basic interactions between particles in granular media and that has been successfully used to study granular flows in different configurations [Dury and Ristow 1997; Hirshfeld and Rapaport 1997]. According to Newton's second law, the equation

of motion for a single grain looks like

$$m_i \ddot{\mathbf{r}}_i + \alpha_i \dot{\mathbf{r}}_{ij} = -\nabla_i e(r_{ij}) - \sum_{i \neq j} \beta_{ij} u_{ij}(r_{ij}) \dot{\mathbf{r}}_{ij} + \mu_{it} u_{it}(r_{it}) (\Delta \bar{\mathbf{v}}_R \times d\hat{\mathbf{r}}) - m_i \bar{\mathbf{g}},$$

where

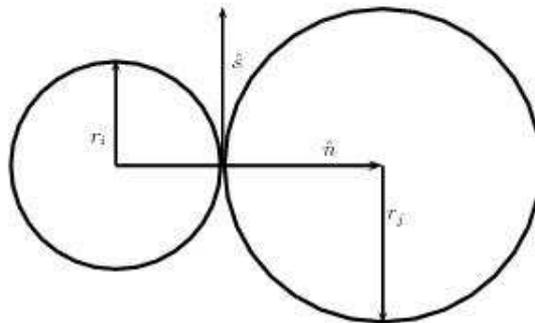
$$e(r_{ij}) = \sum_{i > j} u_{ij}(r_{ij}),$$

$$u_{ij}(r_{ij}) = 4\epsilon_{ij} \left( \left( \frac{\sigma_{ij}}{r_{ij}^0 - r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}^0 - r_{ij}} \right)^6 \right) + \epsilon_{ij}$$

is the potential energy between pairs of grains, given in terms of the parameters  $\epsilon_{ij}$ , which is a measure of the strength of the interaction between grains and which provides a unit for the energy of the interaction,  $\sigma_{ij}$ , the *softness* of the grains, and  $r_{ij}$ , the relative separation of particles. The repulsive force in the normal direction along the line joining the centers of mass of the particles is thus given by  $-\nabla_i e(r_{ij})$ . This force depends only on the relative separation of the particles,  $r_{ij}$ , and prevents them from overlapping. The analytical expression for the potential energy is that of the common Lennard-Jones potential widely used in Molecular Dynamics simulations [Allen and Tildesley 1989]. This potential provides a smooth force function (required by the numerical method) with the desired physical properties for the system: strongly repulsive for small separations of the particles and zero for separation between particles that are larger than a cut-off, since the particles do not interact when they are far apart from each other.

The terms proportional to the relative velocity in the equations of motion,  $\dot{\mathbf{r}}_{ij}$ , correspond to a dissipative force that accounts for the inelastic collisions, and which is determined by the constant parameter  $\beta_{ij}$ . The force in the shear direction is connected to the normal force by the Coulomb laws of friction [Schaefer et al. 1996]. A diagram of the forces acting on a pair of grains is shown in Figure 1. It is worth noticing that, in three dimensions, the shear force is not uniquely defined and we must require that the tangential force lie in the plane of the relative velocity.

The last two terms in the equations of motion correspond to the interactions between the particles and the tumbler’s wall. We use  $\mu_{it}$  as the parameter for the frictional forces between the particles and the



**Figure 1.** Normal and tangential directions during particle contacts.

interior wall, while the force resulting from the angular velocity is given by  $\Delta \bar{v}_R \times d\hat{r}$ , where  $\Delta \bar{v}_R$  is the relative velocity between the particle and the tumbler, and where  $d\hat{r} = \frac{\bar{r}_i - \bar{r}_t}{|\bar{r}_i - \bar{r}_t|}$  is a unit vector in the direction determined by the center of the particle and the center of the tumbler.

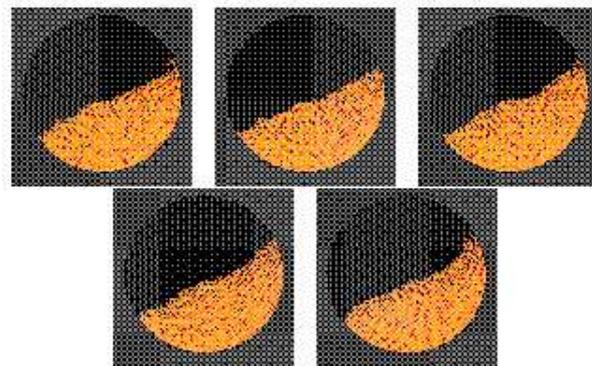
#### 4. Preliminary results

We are currently performing simulations for monodisperse systems of particles in order to establish an appropriate correspondence between the model and the experimental measurements. For these simulations, we fix the tumbler's radius and the total number of particles used. It should be noticed that with this choice, varying the size of the grains also changes the filling level of the tumbler and therefore the dynamics since, clearly, a tumbler that is 10% filled will show different behavior than one that is 90% filled. We also fix the model parameters  $\alpha_i$ ,  $\beta_{ij}$  and  $\mu_{it}$  and study the system as the rotational velocity of the tumbler is varied. Figure 2 shows the results of simulations using 1500 particles of size  $r_i^0 = 1.0$  (in arbitrary units) inside a tumbler of radius  $R = 52.5$  with  $\alpha_i = \beta_{ij} = \mu_{it} = 1.0$  for rotational velocities between 2.0 rpm and 2.4 rpm. The s-shape that can be seen in the surface layer has been observed experimentally, and is commonly seen before a transition in the flow regime occurs [Ding et al. 2002].

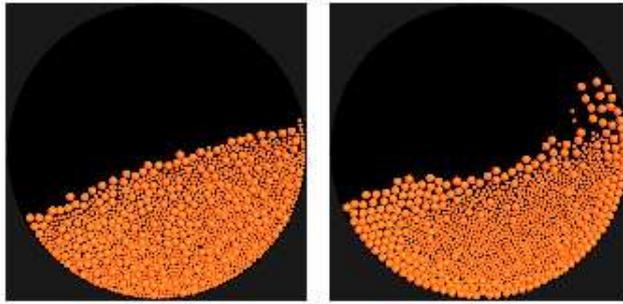
A first step towards successfully applying the model discussed here to the study of mixing and segregation in granular materials is shown in Figure 3. In this case, the simulation is done with 850 particles of size  $r_i^0 = r^0 = 0.1$  in a drum of radius  $R = 35$ . Both types of particles have the same mass density and the rotational speed is  $\Omega = 2.0$  rpm. The figure at the left is a snapshot of the simulation at the initial stage and the one at the right shows a snapshot taken after only a few rotations of the tumbler. The images show that the system is reaching a segregation state in which the smallest particles are accumulating around the cores, as should be expected.

#### 5. Discussion

Although this work is in a development stage, the simple model presented here is enough to observe some of the general features of granular particles inside a rotating tumbler, not only for monodisperse systems but also for the classical cases of D- and S-systems. Even in the simplest cases of monodisperse



**Figure 2.** Numerical simulations for mono disperse particles. Rotational speeds from left to right are:  $\Omega = 2.0, 2.1, 2.2, 2.3$  and  $2.4$  rpm.



**Figure 3.** Snapshot of a numerical simulation of a binary mixture. The final segregated state can be observed on the right.

systems, our simulations have shown features and behavior that are richer than the originally expected ones. A phase diagram of the flow transitions as the rotational speed increases is being constructed while monitoring physical quantities like the average energy, the average torque on the tumbler due to the particles, etc.

Another work currently in progress is the study of the dynamics of elongated grains inside the rotating tumbler. In [Tapia-McClung and Grønbech-Jensen 2005] an efficient numerical algorithm was developed to constrain particles in a linear geometry. This allows us to numerically construct elongated grains that are composed of several individual particles and no longer have to be considered ideal grains represented by spherical particles, since we are able to construct grains with irregular shapes.

Even where we have not mentioned it, three-dimensional simulations are also being considered. The computational effort required to extend the calculations from two to three dimensions is relatively simple, and so it is contemplated as a future goal of this research.

### Acknowledgements

The author would like to thank Prof. Niels Grønbech-Jensen for his help in the implementation of the simulations and for the computational resources kindly provided. This work has been funded by the Mexican Council of Science and Technology, CONACyT.

### References

- [Allen and Tildesley 1989] M. P. Allen and D. J. Tildesley, *Computer simulation of liquids*, Oxford University Press, Oxford, 1989.
- [Andersen 1983] H. C. Andersen, “Rattle: a ”velocity” version of the shake algorithm for molecular dynamics calculations”, *J. Comput. Phys.* **52**:1 (1983), 24–34.
- [Ding et al. 2002] Y. Ding, R. Forster, J. Seville, and D. Parker, “Granular motion in rotating drums: bed turnover time and slumpin-rolling transition”, *Powder Technol.* **124**:1-2 (2002), 18–27.
- [Dury and Ristow 1997] C. M. Dury and G. H. Ristow, “Radial segregation in a two-dimensional rotating drum”, *J. Phys. I France* **7**:5 (1997), 737–745.
- [Hirshfeld and Rapaport 1997] D. Hirshfeld and D. Rapaport, “Molecular dynamics studies of grain segregation in sheared flow”, *Phys. Rev. E* **56**:2 (1997), 2012–2018.

- [Jain et al. 2005] N. Jain, J. M. Ottino, and R. M. Lueptow, “Regimes of segregation and mixing in combined size and density granular systems: an experimental study”, *Granul. Matter* **7**:2-3 (2005), 69–81.
- [Khakhar et al. 1997] D. Khakhar, J. McCarthy, and J. Ottino, “Radial segregation of granular mixtures in rotating cylinders”, *Phys. Fluids* **9**:12 (1997), 3600–3614.
- [Schaefer et al. 1996] J. Schaefer, S. Dippel, and D. Wolf, “Force schemes in simulations of granular materials”, *J. Phys. I* **6**:1 (1996), 5–20.
- [Tapia-McClung and Grønbech-Jensen 2005] H. Tapia-McClung and N. Grønbech-Jensen, “Non-iterative and exact method for constraining particles in a linear geometry”, *J. Polym. Sci. Pol. Phys.* **43**:8 (2005), 911–916.

Received 1 Aug 2006. Revised 28 Mar 2007. Accepted 20 Apr 2007.

HORACIO TAPIA-MCCLUNG: [hotapia@cabrillo.edu](mailto:hotapia@cabrillo.edu)

Department of Applied Science, University of California, Davis, One Shields Ave, Davis, CA 95616-8254, United States

## DENSITY MEASUREMENTS IN A SUPERSONIC JET

CATALINA ELIZABETH STERN, JOSÉ MANUEL ALVARADO AND CESAR AGUILAR

We use a nonintrusive optical technique for heterodyne detection of the light scattered elastically by the molecules of a moving transparent gas, a phenomenon known as Rayleigh scattering. It can be shown that the signal that comes out of the photodetector is proportional to the spatial Fourier transform as a function of time of the density fluctuations, for a wave vector given by the optical set-up. This is the only technique we are aware of that can study density fluctuations *inside* a flow.

In this paper we present results obtained from a supersonic axisymmetric air jet. The signal that comes out of the photodetector is processed, and the power spectrum calculated. In the spectrum, density fluctuations of two different origins can be identified: acoustic, that is, those that propagate at the speed of sound and are related to pressure variations, and entropic, those that have constant pressure and are convected by the flow. At certain locations we have found an additional peak related to the interaction between the flow and the shock structure. Furthermore, Rayleigh scattering can be used to visualize the shock structure of the flow. We provide supporting images for our results.

### 1. Introduction

The original objectives of this work were to localize sound sources in a supersonic jet, relate the production of sound with flow phenomena and determine the acoustic radiation pattern inside and outside the flow. We use a Rayleigh scattering technique that can measure density fluctuations inside the flow for a given wavevector. Our jet is very turbulent and sound sources are not localized. However, we have been able to visualize the shock structure, allowing us to relate the spectrum at each location and at each angle to the compression and expansion waves in the supersonic region of the jet. We have been able to determine the direction of propagation of sound waves inside and outside the jet including the mixing layer. We have also found an unexpected peak in the spectrum that is related to the interaction between the flow and the shock structure.

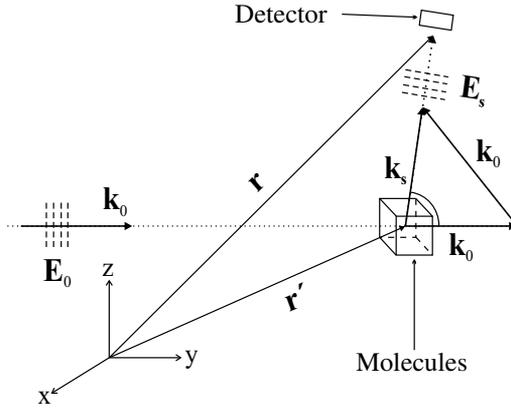
### 2. Background

**2.1. Acoustic emission.** There are several theories that try to explain acoustic emission in a jet. Some of them are based on the interactions between large-scale structures in the flow: pairing, cut and connect, and annihilation. In general, it is accepted that large-scale structure interactions produce sound that propagates at large angles, while small structures produce sound that propagates at small angles. There are other possible sources of emission such as the interaction of the flow with the shock waves and the feedback of acoustic waves that reenter the flow.

---

*Keywords:* aeroacoustics, Rayleigh scattering, supersonic flow.

This work was supported by UNAM through projects DGAPA IN107599, IN104102 and IN116206.



**Figure 1.** Molecules scatter light in all directions. Selecting detector orientation, we specify scattering angle and size of fluctuations to be studied.

Traditionally, experimental studies on acoustic waves have been performed by placing many microphones in the far field and correlating these measurements with events measured inside the flow. Not only is there a problem trying to determine sources from far field measurements as the solution of the inverse problem is not unique, but this method fails to take into account certain phenomena such as the diffraction of the acoustic waves by the mixing layer.

**2.2. Rayleigh scattering.** The elastic scattering of an electromagnetic wave of wavelength  $\lambda_0$  by a neutral particle of dimensions smaller than the wavelength is known as Rayleigh scattering. In a static transparent gas, light is scattered homogeneously, and the scattered field is constant. If the gas is in motion or with strong density variations, the characteristics of the scattered light reflect the characteristics of the structure and motion of the gas. In the far field, the light scattered by one molecule is given by

$$\vec{E}_{S0} = r_0^R \frac{e^{ik_0 r}}{r} \{ \vec{r} \times \vec{E}_0(\vec{r}') \times \vec{r} \},$$

where  $r_0^R$  is the Rayleigh scattering cross section. Figure 1 shows the wave vectors of the incident and scattered light. The total scattered field can be obtained from the integral

$$\vec{E}_S = \vec{E}_{S0}(\vec{r}, t) \int_{V_s} d^3 r' n(\vec{r}', t) e^{i\vec{k}_\Delta \cdot \vec{r}} = \vec{E}_{S0}(\vec{r}, t) n(\vec{k}_\Delta, t),$$

where  $n(\vec{r}', t)$  is the distribution of molecules in the scattering volume  $V_s$ ,  $n(\vec{k}_\Delta, t)$  is the spatial Fourier transform of the density function and  $\vec{E}_{S0}(\vec{r}, t)$  is the field scattered by one molecule. The scattered field has information about the motions of the molecules in the scattering volume through the spatial Fourier transform of the density.

**2.3. Density fluctuations.** When the speed of the flow is close to Mach1, compressibility becomes important and the equations that describe the flow are more complicated than for the incompressible case. However, if we consider small oscillations about the equilibrium, the equations can be linearized. Monin and Yaglom [1987] have shown that if we write the equations of motion in terms of the vorticity  $\Omega$ , the

divergence  $D$  of the velocity, the entropy  $S$  and the pressure  $P$ , all possible motions can be described by three noninteracting modes:

$$\frac{d\Omega(t)}{dt} = 0, \quad \frac{dS(t)}{dt} = 0, \quad \frac{d^2D(t)}{dt^2} + a_0^2k^2D(t) = 0, \quad \frac{d^2P(t)}{dt^2} + a_0^2k^2P(t) = 0.$$

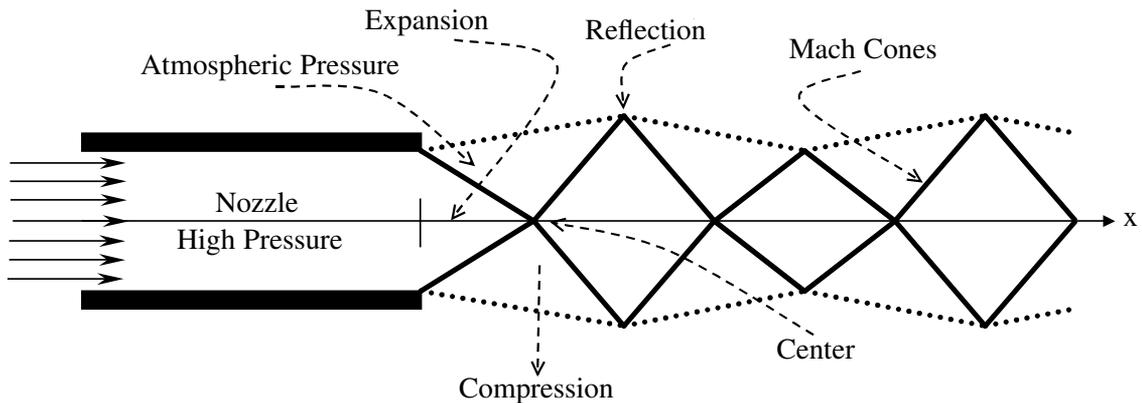
The incompressible vorticity mode and the entropy mode are stationary or move at constant speed. The acoustic or potential mode is related to pressure fluctuations that propagate at the speed of sound as we can see from the wave equation. We can expect to measure the two modes related to the compressible part of the flow: entropic and acoustic.

**2.4. Structure of a supersonic jet.** Figure 2 shows the structure of a supersonic jet. The discontinuity at the edge of the nozzle produces a perturbation that propagates at the speed of sound. Each new perturbation catches on the previous one because the speed of the flow is supersonic. The addition of these perturbations creates a shock, that is, a conic region of very high density. Starting with an expansion, a stationary pattern of shocks is formed in the supersonic region of the jet. As the speed decays, the flow becomes subsonic and the shocks disappear.

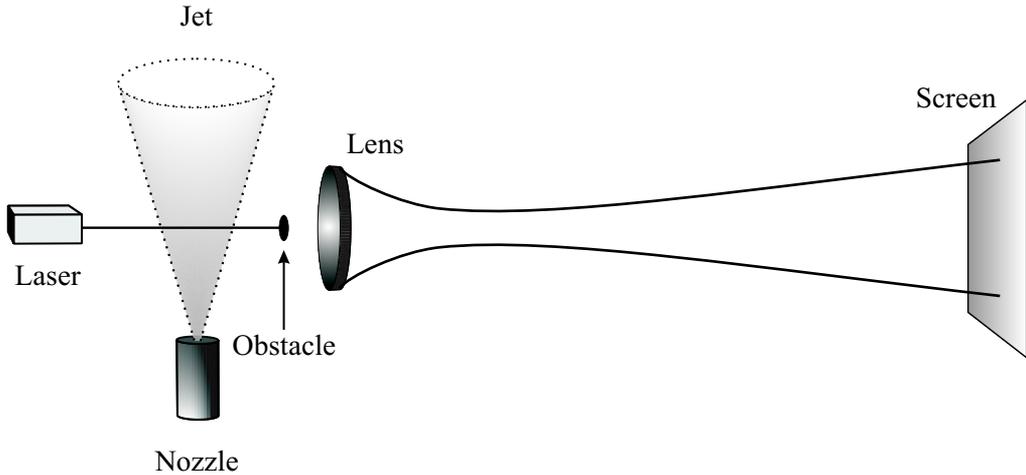
### 3. Experimental setup and techniques

**3.1. Visualization.** Figure 3 shows the experimental setup used for visualization. All the light scattered at small angles is collected by a lens and sent to a screen where the flow pattern can be visualized [Azpeitia 2004].

**3.2. Heterodyne detection of the scattered light.** The amplitude of the scattered light is extremely small and cannot be measured by a common diode. To solve this problem we mix, on the surface of the photodetector, the scattered light with a well known beam of light called the local oscillator. The frequency of the local oscillator is different from the frequency of the incident beam. This technique is known as heterodyning. Figure 4 shows the experimental setup.

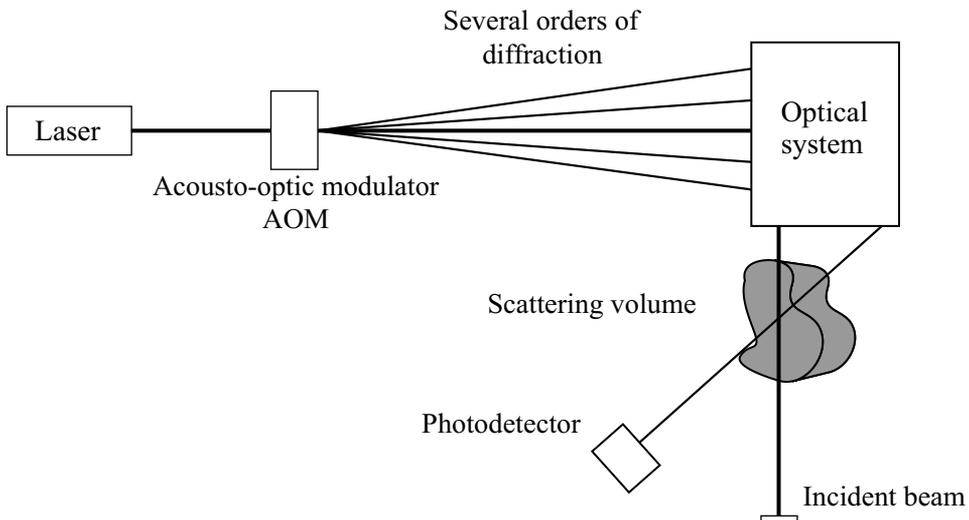


**Figure 2.** Discontinuity at nozzle end creates perturbation that gives a stationary shock pattern for supersonic flow.



**Figure 3.** Laser light is sent through the jet. Central part of beam is blocked, while small angle scattering is collected on the screen.

The beam that comes out of the laser is sent into an acoustic modulator. The modulator acts as a Bragg cell and several orders of diffraction come out. Order zero goes through without being deviated; we refer to this beam as incident or primary. Order one is diffracted at a particular angle, is less intense and is displaced in frequency by 110 MHz. We will refer to this beam as the local oscillator. Both beams are manipulated so that they cross at angle  $\theta$  in the volume to be studied. The local oscillator is sent directly to a photodetector, while the main beam is blocked just after the scattering. The photodetector then sees the part of the incident field scattered at the angle  $q$  and the local oscillator. The scattering angle determines the wavenumber of the fluctuations through  $k_{\Delta} = 2k_0 \sin(\theta/2)$ , where  $\vec{k}_0$  is the wavevector



**Figure 4.** Light sent to acousto-optic modulator comes out as several beams displaced in frequency. One is used as local oscillator and mixed with scattered light at detector.

of the incident field,  $\vec{k}_\Delta$  is the wavevector of the density fluctuations, and  $k_0, k_\Delta$  are their respective magnitudes.

The photodetector is sensitive to the intensity of the incident light, so the current it produces is proportional to the square of the electric field incident on its surface. The current of the photodiode is then proportional to

$$(\vec{E}_S + \vec{E}_{OL})^2 = |\vec{E}_S|^2 + |\vec{E}_{OL}|^2 + 2\vec{E}_S \cdot \vec{E}_{OL}.$$

The first two terms are constant. In particular, the first is too small and the second is of no interest. The third term gives a current that oscillates at the frequency that is the difference of the frequencies of the two electric fields. It contains the information we are interested in, and is modulated by the amplitude of the local oscillator. The current proportional to this term is known as the heterodyne current.

It can be shown [Stern and Grésillon 1983; Aguilar 2003] that the spectral density of the heterodyne current  $I(\omega)$  produced by all the scatterers is of the form

$$I(\omega) = \frac{1}{8\pi k_0^2} \left( \frac{\eta e}{\hbar \omega_0} \right)^2 n_0 (r_0^R)^2 \frac{\epsilon_0}{\mu_0} (\vec{E}_S \cdot \vec{E}_{OL})^2 \int d^3k |W(\vec{k}_\Delta - \vec{k})|^2 \times [S(\vec{k}, \omega - \omega_\Delta) + S(\vec{k}, \omega + \omega_\Delta)],$$

where  $\eta$  is the efficiency of the detector,  $n_0$  the mean density,  $W$  is related to the Gaussian profiles of the beams and  $S(\vec{k}_\Delta, \omega)$  is the form factor defined by

$$S(\vec{k}_\Delta, \omega) = \frac{|n(\vec{k}_\Delta, \omega)|^2}{n_0 V}.$$

To obtain the spectral density, the signal from the photodetector is either sent directly to a spectrum analyzer or acquired with a computer and treated with periodgrams that have a higher spectral resolution than the analyzer and the traditional Fourier transform.

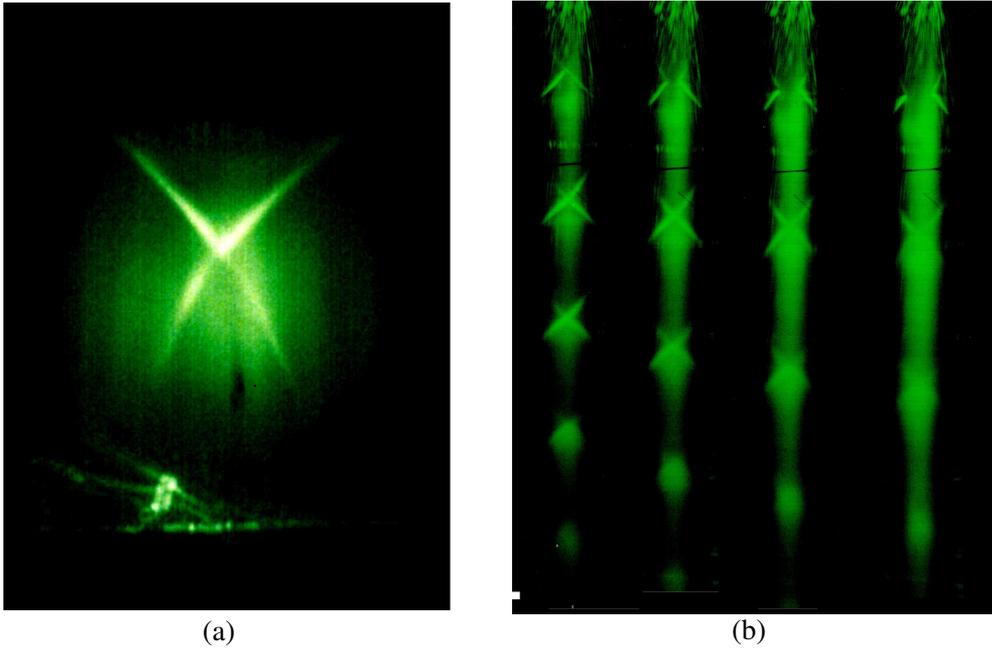
The diameter of our nozzle is 0.8 mm. It is mounted on a rotating and translating traverse in such a way that density fluctuations can be studied inside and outside the jet and in all directions. The technique described is sensitive to the wavevector of the fluctuations. Therefore, we can determine the direction of propagation of acoustic waves inside the flow, including the mixing layer.

## 4. Results

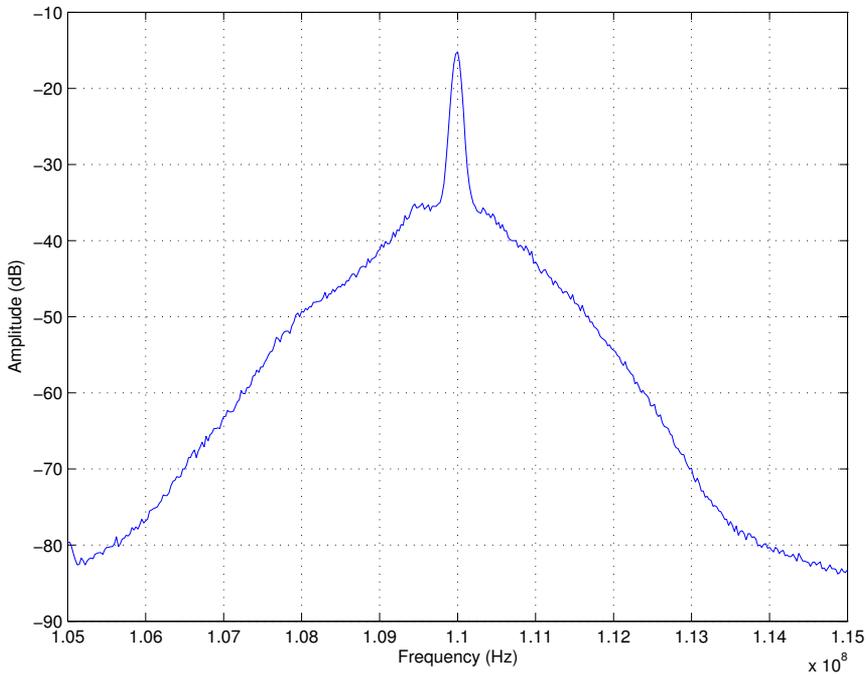
**4.1. Visualization.** When we visualize the near region of the flow we can see the shocks created by the discontinuity at the nozzle exit. Figure 5(a) shows the first shock after the nozzle.

If we use a cylindrical lens, we can create a sheet of light and observe a series of shocks along the jet. Figure 5(b) is a superposition of four images showing four jets with different exit velocities. The exit pressure is increasing from top to bottom. It can be observed that as the pressure increases, the crossover region where the expansion and the compression meet becomes flat.

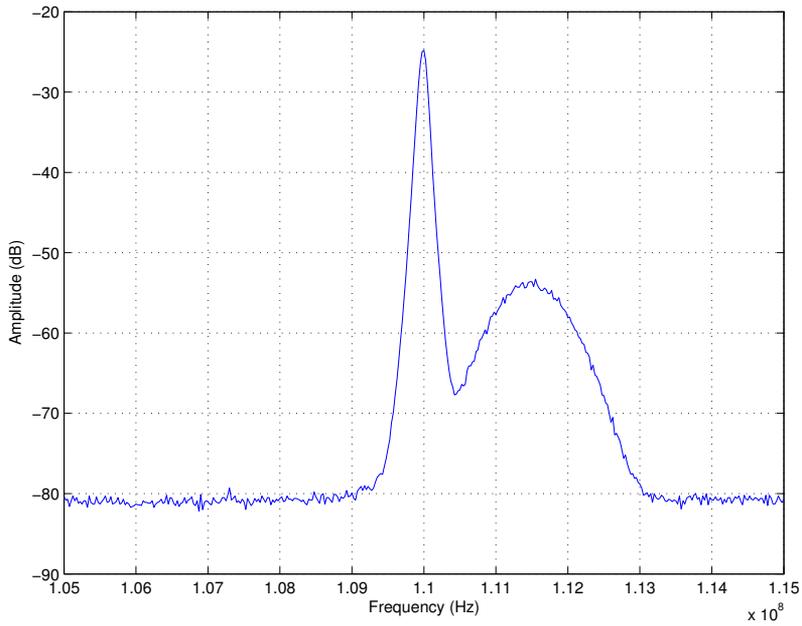
**4.2. Heterodyne detection.** Figure 6 shows the spectral density obtained with the spectrum analyzer for fluctuations that propagate perpendicular to the flow at a point in the centerline of the jet. The peak is so broad that it is hard to differentiate between entropic and acoustic fluctuations. Figure 7 shows the spectrum for fluctuations perpendicular to the flow outside the flow. As expected, only acoustic fluctuations are detected in this region. The frequency of the center of the peak divided by  $k_\Delta$  corresponds to the speed of sound.



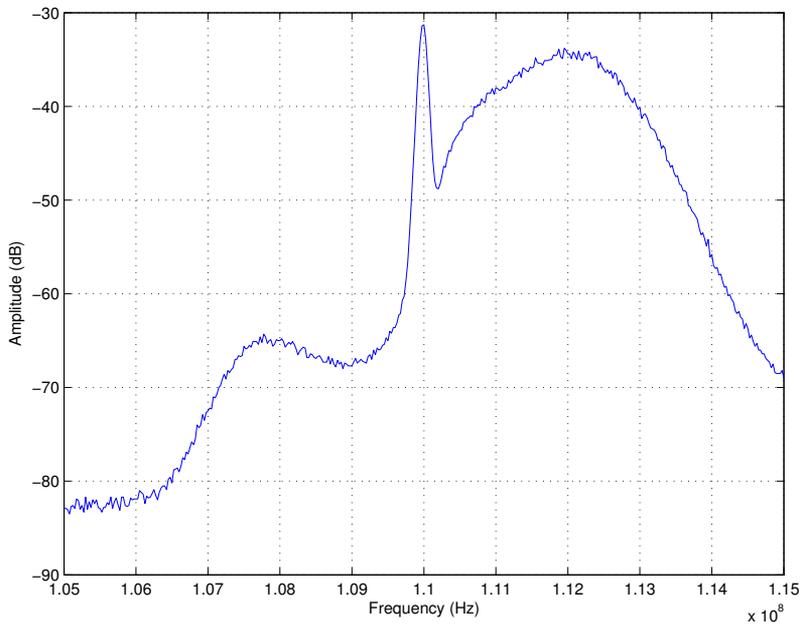
**Figure 5.** (a) First shock after nozzle; flow is upward. (b) Four jets visualized one at a time with a sheet of light; flow is downward. Several shocks are observed.



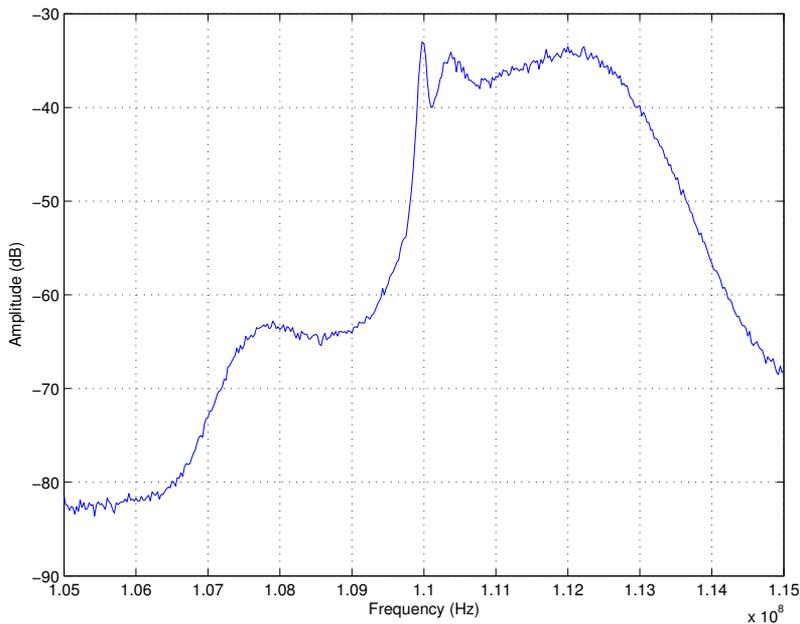
**Figure 6.** Fluctuations perpendicular to the jet on the axis.



**Figure 7.** Fluctuations perpendicular to the flow outside the jet.



**Figure 8.** Fluctuations parallel to the flow on the jet axis.



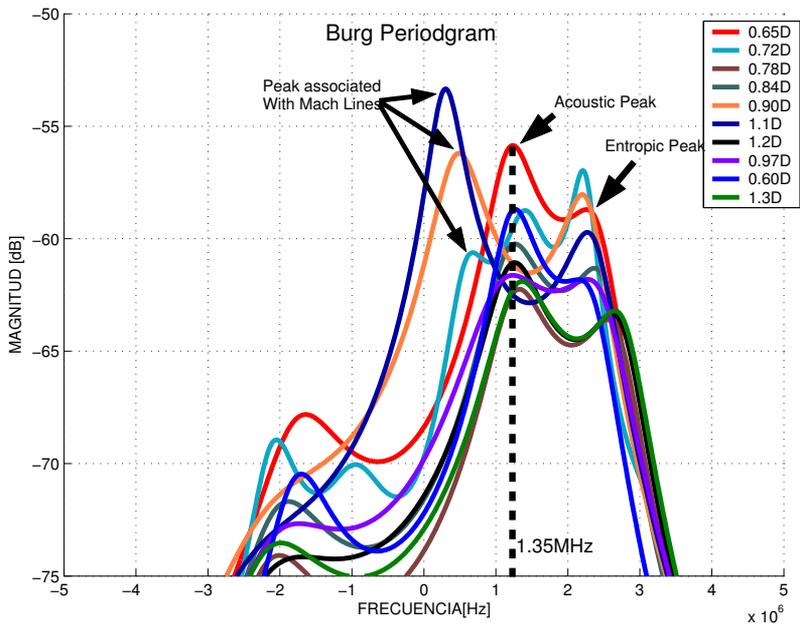
**Figure 9.** New low-frequency peak appears at some locations along jet center.

The spectral density in [Figure 8](#) corresponds to fluctuations at a point on the centerline of the jet that propagate parallel to the flow. It can be seen that the entropic part of the peak is shifted because the scattering molecules are convected with the flow. Through the change of frequency of the entropic peak, that is, the Doppler shift, we can determine the local mean speed of the flow. An acoustic peak, due to a reflection on the optical setup can be observed on the left hand side of the spectrum. The spectrum in [Figure 9](#) shows that at certain locations along the axis an additional low frequency peak can be observed.

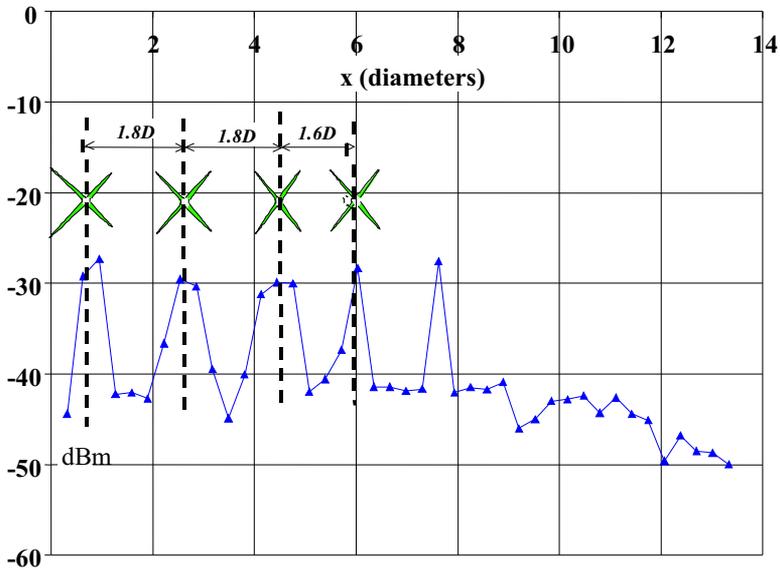
In all the figures shown above, the spectral densities were obtained through a spectrum analyzer. The resolution in frequency is quite poor. To ameliorate these results, we have deheterodyned the signal to shift the reference frequency to zero, acquired it with a computer and treated it with parametric periodgrams of the Burg type [[Alvarado 2004](#)].

[Figure 10](#) shows the spectral density at various locations along the centerline for fluctuations perpendicular to the flow. Three peaks are visible. The acoustic peak is always at the same location. The entropic peak changes with the local speed of the flow, and the new peak appears and disappears along the centerline and changes slightly its frequency. It is interesting to note that when the new peak has its highest amplitude, the acoustic peak disappears and vice versa.

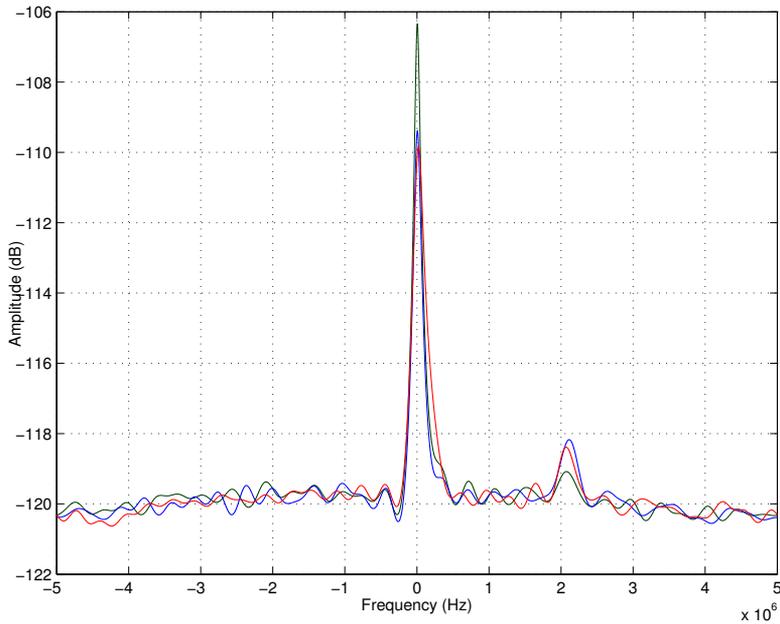
By comparing the photographs obtained with the visualization and the graph of the amplitude of the new peak as a function of position as seen in [Figure 11](#), we have been able to determine that the regions of maximum amplitude for the low frequency peak correspond to the crossover between expansion and compression in the shocks. The  $x$  coordinate is given in multiples of the nozzle diameter. We are convinced that the peak is related to the interaction of the flow with the shocks. A more detailed study of the origin of this peak is underway.



**Figure 10.** Spectral densities calculated by means of Burgs parametric periodograms at several locations along jet center.



**Figure 11.** Comparison of new peak amplitude with jet shock structure.



**Figure 12.** Spectral densities outside the flow at same location for different wavevector directions. Maximum amplitude corresponds to propagation direction of acoustic wave.

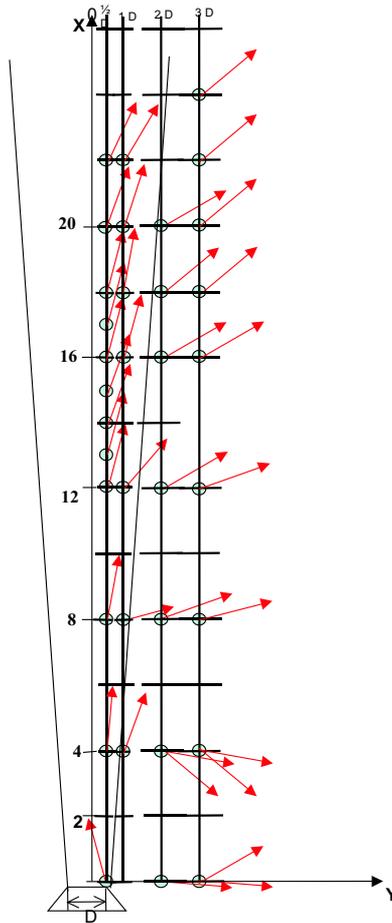
The signal of the photodetector depends on the wave vector and is thus sensitive not only to the wavenumber, but also to the direction of propagation of the fluctuations. Figure 12 shows how the amplitude changes when measurements are taken at the same point (outside the flow), for the same wavenumber but different direction of propagation.

It can be observed that the amplitude changes with the direction. If we consider that the maximum amplitude corresponds to the direction of propagation of the acoustic wave, we can determine the acoustic radiation pattern of the jet inside and outside the flow. Figure 13 shows a preliminary radiation pattern obtained by seeking, at each point in the jet, the direction where the acoustic peak has the highest amplitude.

## 5. Conclusions

We have shown that heterodyne detection can measure density fluctuations in the flow and differentiate among three different phenomena: acoustic waves propagating at the speed of sound, entropy fluctuations that are convected by the flow, and low frequency fluctuations that appear close to the shock structure. The mean local velocity of the jet can also be measured through the Doppler shift of the spectrum. The signal is sensitive to the direction of propagation of the fluctuations, so the technique can also be used to determine the acoustic radiation pattern inside and outside of the jet, and eventually to localize the acoustic sources.

Moreover, Rayleigh scattering can be used to visualize certain aspects of the shock structure. To better understand the origin of the low frequency peak we are planning two kinds of experiments. It is well known that a certain noise known as screech is produced by the interaction of the flow with shocks. The



**Figure 13.** Preliminary acoustic radiation pattern inside and outside the flow.

Strouhal number of the screech is well documented. We aim to detect the screech outside the flow with a transducer and correlate it with our measurements.

Furthermore, we aim to find all the positions in the flow where the peak appears, as well as the direction of maximum amplitude to understand how the flow behaves outside the centerline.

### References

- [Aguilar 2003] C. Aguilar, *Detection of acoustic waves in a supersonic jet using Rayleigh scattering*, Bachelor's thesis, Department of Physics, School of Science, UNAM, Mexico City, 2003.
- [Alvarado 2004] M. Alvarado, "Spectral analysis of signals from Rayleigh scattering experiment", Master's thesis, School of Engineering, UNAM, Mexico City, 2004.
- [Azpeitia 2004] C. Azpeitia, *Use of Rayleigh scattering to localize acoustic sources in a supersonic jet*, Bachelor's thesis, Department of Physics, School of Science, UNAM, Mexico City, 2004.
- [Monin and Yaglom 1987] A. S. Monin and A. M. Yaglom, *Statistical fluid mechanics*, MIT Press, Cambridge, MA, 1987.

[Stern and Grésillon 1983] C. Stern and D. Grésillon, “Fluctuations de densité dans la turbulence d’un jet: observation par diffusion Rayleigh et détection heterodyne”, *J. Phys.* **44** (1983), 1325–1335.

Received 23 Apr 2007. Accepted 8 May 2007.

CATALINA ELIZABETH STERN: [catalina@graef.fciencias.unam.mx](mailto:catalina@graef.fciencias.unam.mx)

*Laboratorio de Acústica, Facultad de Ciencias, Ciudad Universitaria, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Col. Copilco el Bajo, Del. Coyoacán, Distrito Federal 04510, Mexico*

JOSÉ MANUEL ALVARADO: [manuel@graef.fciencias.unam.mx](mailto:manuel@graef.fciencias.unam.mx)

*Laboratorio de Acústica, Facultad de Ciencias, Ciudad Universitaria, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Col. Copilco el Bajo, Del. Coyoacán, Distrito Federal 04510, Mexico*

CESAR AGUILAR: [cae@graef.fciencias.unam.mx](mailto:cae@graef.fciencias.unam.mx)

*Laboratorio de Acústica, Facultad de Ciencias, Ciudad Universitaria, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Col. Copilco el Bajo, Del. Coyoacán, Distrito Federal 04510, Mexico*

## FREQUENCY AND SPATIAL RESPONSE OF BASILAR MEMBRANE VIBRATION IN A THREE-DIMENSIONAL GERBIL COCHLEAR MODEL

YONGJIN YOON, SUNIL PURIA AND CHARLES R. STEELE

The cochlea of the inner ear presents severe difficulties for measurement and computation, and controversy exists on virtually every issue. However, the first in vivo measurement of the spatial distribution of elastic response for a fixed frequency is now available. This work compares experimental results and those from calculations with a three-dimensional model. This is a standard model that consists of a long, fluid-filled box with a partition, a portion of which is the elastic BM (basilar membrane). The BM velocity at a fixed point as a function of frequency and the spatial response for a fixed frequency are calculated. The model includes the three-dimensional viscous fluid and the pectinate zone of the elastic orthotropic BM with the gerbil dimensional and material property variation along its length. The radial BM thickness variation is, however, replaced by an equivalent constant thickness. The active process is represented by adding the motility of the OHCs (outer hair cells) to the passive model with a feed-forward approximation of the organ of Corti (OC). Asymptotic and numerical methods combined with Fourier series expansions are used to provide a fast and efficient iterative procedure that requires about one second on a desktop computer for obtaining the BM response for a given frequency. Our three-dimensional model results show the following agreement with the experimental measurements in various situations: (i) for map of place of maximum response to frequency — excellent; (ii) for the response at a fixed point as a function of frequency — excellent for amplitude, poor for phase; (iii) for the spatial distribution for fixed frequency — fair for amplitude and excellent for phase. The discrepancies in (ii) and (iii) remain to be clarified.

### 1. Introduction

The cochlea is a snail-shaped, fluid-filled duct that is divided along its longitudinal direction by the compliant basilar membrane (BM), upon which is located the organ of Corti (OC) containing all sensory cells. The fluid and compliant structures within the cochlea are set in motion in response to sound input at the stapes, and the detection of this motion by inner hair cells (IHCs) initiates hearing through afferent auditory nerve firing transmitted to the auditory cortex. In this study, the mechanical behavior of the cochlea, especially BM velocity, was simulated with a physiologically based, three-dimensional cochlear model. Model results were compared with in vivo cochlear experimental data in the characteristic frequency-to-place (CF-to-place) map and BM velocity magnitude and phase for the frequency and spatial distribution. Access for in vivo measurement in the cochlea is severely restricted and difficult. The first in vivo measurement of the spatial distribution of elastic response for a fixed frequency by [Ren \[2002\]](#) provides the motivation for the present study.

---

*Keywords:* cochlear model, mechanical response, basilar membrane velocity, outer hair cell, gerbil.  
This work was funded by HFSP Grant No. RGP0051.

Numerous mathematical models describe the biomechanical activity in the cochlea. Models extend the passive cochlear model with the inclusion of the motions of OC, particularly the active behavior of the outer hair cells (OHC), beginning with simplified one-dimensional models with negative damping [de Boer 1983]. Higher-dimensional active models have also been developed. Two-dimensional finite difference models were constructed by using a feedback law [Neely 1985; 1993]. Numerically intense three-dimensional finite-element models had been developed with the inclusion of varying details and complexities of the OC, but the fluid was still modeled as inviscid [Kolston and Ashmore 1996; Böhnke and Arnold 1998]. Models include the activity in the OC as a feed-forward mechanism from the longitudinal tilt of the OHCs. Two-dimensional [Geisler and Sang 1995] and three-dimensional models with the active feed-forward mechanism have been employed [Steele et al. 1993; Steele and Lim 1999; Lim and Steele 2002].

The present study uses the physiologically based, linear three-dimensional feed-forward model for gerbil anatomy. The model uses a combination of the asymptotic phase integral method that is commonly known as WKB (Wentzel–Kramers–Brillouin) method and the fourth order Runge–Kutta (RK4) numerical forward integration. This hybrid approach provides significantly faster computations than the finite difference or finite element methods and more accuracy than the WKB alone [Lim and Steele 2002].

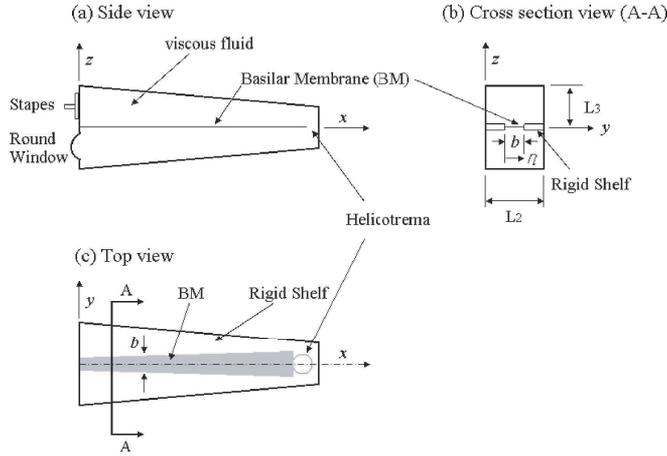
The present model is as simple as possible while still representing the essential features of the BM response. Included in the model are the variation of the dimensions and material properties along the cochlear duct and three-dimensional viscous fluid effects. Only one degree of freedom of the partition, the flexing of the pectinate zone of the orthotropic BM, is considered. The spiral coiling of the cochlea is neglected, as it has been shown to have little effect on the model response [Loh 1983; Steele and Taber 1979]. The simulation results obtained from this active model successfully demonstrate various aspects of *in vivo* measurements. Since it is difficult to understand the dynamic response of a structure from measurements alone, particularly when the measurements are restricted to a few locations along the cochlea, a reliable model and calculation procedure plays an important role.

## 2. Mathematical methods

**2.1. Passive model.** The physical cochlea consists of a rigid bony housing containing two coiled, fluid-filled ducts, separated by the cochlear partition. The model is based on these physiologic features of the cochlea. A schematic drawing of the model with the side, cross-section, and top view is shown in Figure 1. The detailed derivations and features for the passive mechanics were described in a previous study [Lim and Steele 2002]. Briefly, the three-dimensional fluid equations are integrated over the cross-section to obtain the relation between the volume flow and the fluid pressure at the BM. Then the fluid impedance is matched to the elastic BM impedance, which yields the second order reduced wave equation

$$G_{,xx} + n(x, \omega)^2 G = 0, \quad (1)$$

in which  $n(x, \omega)$  is the local wave number, determined from an *eikonal* equation dependent on distance  $x$  and frequency  $\omega$ . The eikonal equation is complex valued, dependent on the Fourier series expansions on the cross-section of the three-dimensional viscous fluid. An iterative solution for  $n$  is obtained at each  $x$  for fixed  $\omega$ . The form of the  $x$ -dependence is not assumed *a priori*, but comes from the solution of



**Figure 1.** Schematic drawing of the passive cochlear model geometric layout. Distances parameterized in Cartesian coordinates  $\{x, y, z\}$  representing distance from stapes, distance across scala width, and height above partition, respectively. (a) Side, (b) cross-section (A-A), and (c) top views of cochlear model.

**Equation (1).** The dependent variable  $G(x)$  provides the potential  $\Phi(x)$  for the fluid,

$$\Phi(x) = \frac{G(x)n}{T_0 \sinh(nL_3)},$$

where  $T_0(x)$  is the Fourier coefficient for the 0-th component of scalar potential for fluid displacement, and  $L_3$  is the height of fluid chamber.

The function  $G(x)$  is obtained from Equation (1) by using the well-known WKB asymptotic solution in the short wavelength region ( $n$  large) and the RK4 forward integration in the long wavelength region ( $n$  small). The boundary conditions of matching the volume displacement at the stapes and zero pressure at the helicotrema are taken into account with forward and reverse traveling waves.

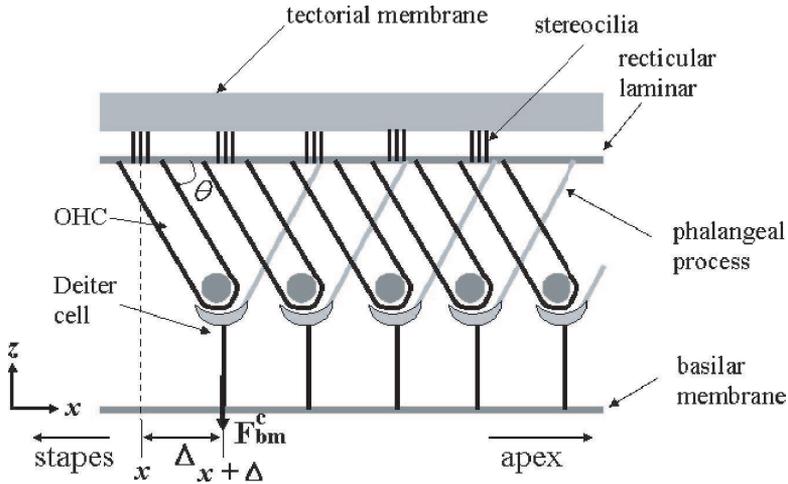
**2.2. Feed-forward active model.** The active elements in the cochlea are the OHCs which behave as piezoelectric actuators. In this model, the force applied by the OHCs on the BM partition is assumed to be proportional to the total force acting on the BM. The total force at the pectinate zone (PZ) results from the fluid force difference across the two scala and forces resulting from the OHCs motility,

$$F_{PZ} = 2F_{BM}^f + F_{BM}^C. \tag{2}$$

The OHC force acting at  $x + \Delta$  is proportional to the BM displacement sensed at  $x$  by the effect of the OHC longitudinal tilt as in Figure 2, or

$$F_{BM}^C(x + \Delta) = \alpha(x)F_{PZ}(x), \tag{3}$$

where  $\alpha$  is the feed-forward gain factor and  $\Delta$  is the longitudinal distance between the apex and base of the OHC, which depends on the length of the OHC  $l_{OHC}$  and its angle with respect to the longitudinal direction  $\theta$ , via  $\Delta = l_{OHC} \cos \theta$ . Combining Equations (2) and (3) provides a modification of the eikonal equation keeping the same form of the equation for  $G$  Equation (1) [Lim and Steele 2002].



**Figure 2.** Schematic of longitudinal view of OC, showing longitudinal tilt of OHC. Longitudinal distance between base and apex of OHC is defined as  $\Delta$ .

### 3. Results and discussion

The cochlear model is used to calculate the response of a gerbil cochlea. The material property values in Table 1 were taken from a number of sources [Smith 1968; Lim and Steele 2002; Miller 1985; Steele et al. 1995; Karavitaki 2002] and the dimensions in Table 2 were from the anatomical measurements for gerbil cochlea [Sokolich et al. 1976; Greenwood 1990; Dannhof et al. 1991; Cohen et al. 1992; Edge et al. 1998; Thorne et al. 1999].

The model is meshed into 12000 sections along the gerbil cochlea length of 12 mm. Forty terms are used in the Fourier expansion across the width of the cross-section. Running on an Intel Pentium IX 3.40 GHz processor, the average time taken for a single harmonic excitation calculation is about one second. This method is a fast and efficient solution compared to a full-scale finite element model. We

Basilar membrane	$E_{11} = 1.0 \times 10^{-3}$ GPa $E_{22} = 1.0$ GPa $E_{12} = 0.0$ GPa $\rho_p = 1.0 \times 10^3$ kg/m <sup>3</sup> $\nu = 0.5$
Scala fluid	$\rho_f = 1.0 \times 10^3$ kg/m <sup>3</sup> $\mu = 0.7 \times 10^{-3}$ Pa s
Outer hair cell	$\theta = 60^\circ$ and $80^\circ$ $\alpha = 0.17$

**Table 1.** Gross properties in cochlear model.

$x$ (mm)	$b$ (mm)	$h$ (mm)	$f$	$L_2, L_3$ (mm)	$l_{OHC}$ ( $\mu\text{m}$ )
0		0.0210	0.030	1.000	25.0
1.5		0.0175			
2.9	0.162			0.750	
3.5		0.0131			
5.0				0.480	
5.9		0.0088			
7.2	0.190	0.0073		0.370	
8.4					
9.0		0.0055		0.340	
10.2	0.205	0.0044			
12.0		0.0031	0.007	0.310	65.0

**Table 2.** Property variations in gerbil cochlear model:  $b$ ,  $h$  and  $f$  are width, thickness and fiber density of plate, respectively;  $L_2, L_3$  are width and height of fluid chamber;  $l_{OHC}$  is OHC length. The thickness  $h$  is not anatomical but an equivalent value to give proper tuning.

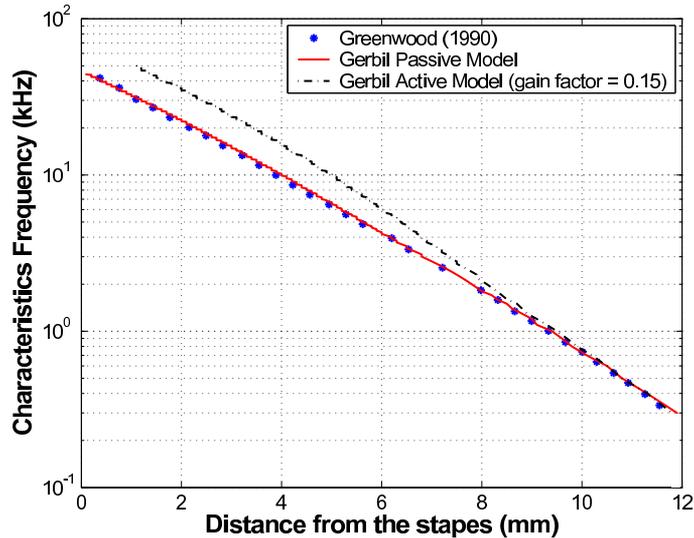
note that the computation time indicated by Parthasarathi et al. [2000] is measured in hours of computing time for the linear solution for a single frequency.

The results include CF-to-place map for the gerbil cochlea, BM velocity frequency response, and BM velocity spatial response. The modeling results for the gerbil cochlea are compared with recent in vivo experiment measurements in the cochlea.

**3.1. CF-to-place map.** CF versus distance from the stapes along the gerbil cochlear (CF range: 0.3 kHz–50 kHz) is shown in Figure 3. The gerbil CF-to-place map [Sokolich et al. 1976; Greenwood 1990] was measured with cochlear-microphonic recording. The maps from the passive model and measurement are in excellent agreement; see Figure 3. Near the stapes (0–4 mm from the stapes), the active model shows 4.5 dB CF shift, whereas CF shift disappears near the helicotrema. Due to the lower wave number for the low input frequency, which has a peak near the apical region of the cochlea, feed-forward gain from the active model shows less gain near the helicotrema.

**3.2. Frequency response of BM velocity.** The gerbil cochlear BM velocity magnitude and phase for 4.2 mm from the base ( $CF = 9.5$  kHz) relative to the stapes displacement are computed over a range of excitation frequencies up to 18 kHz; see Figure 4. Results from the model are compared with the gerbil experimental data [Ren and Nuttall 2001]. The passive model shows quantitatively very good agreement with data which are measured at a high stimulus level (100 dB SPL at the ear canal).

Karavitaki [2002] gives the angle of tilt of gerbil OHC as  $\theta = 84^\circ$ , closer to perpendicular to the basilar membrane; see Figure 2. We calculated the gain from OHC for two cases; a nominal mammalian value of  $\theta = 60^\circ$  and  $\theta = 80^\circ$ . The active model shows fairly good agreement with data at low stimulus level (30 dB SPL at the ear canal) with 27 dB gain for either  $\theta = 60^\circ$  with feed-forward gain factor  $\alpha = 0.15$  (solid-red in Figure 4) or  $\theta = 80^\circ$  with forward gain factor  $\alpha = 0.28$  (solid-brown). So only a slightly



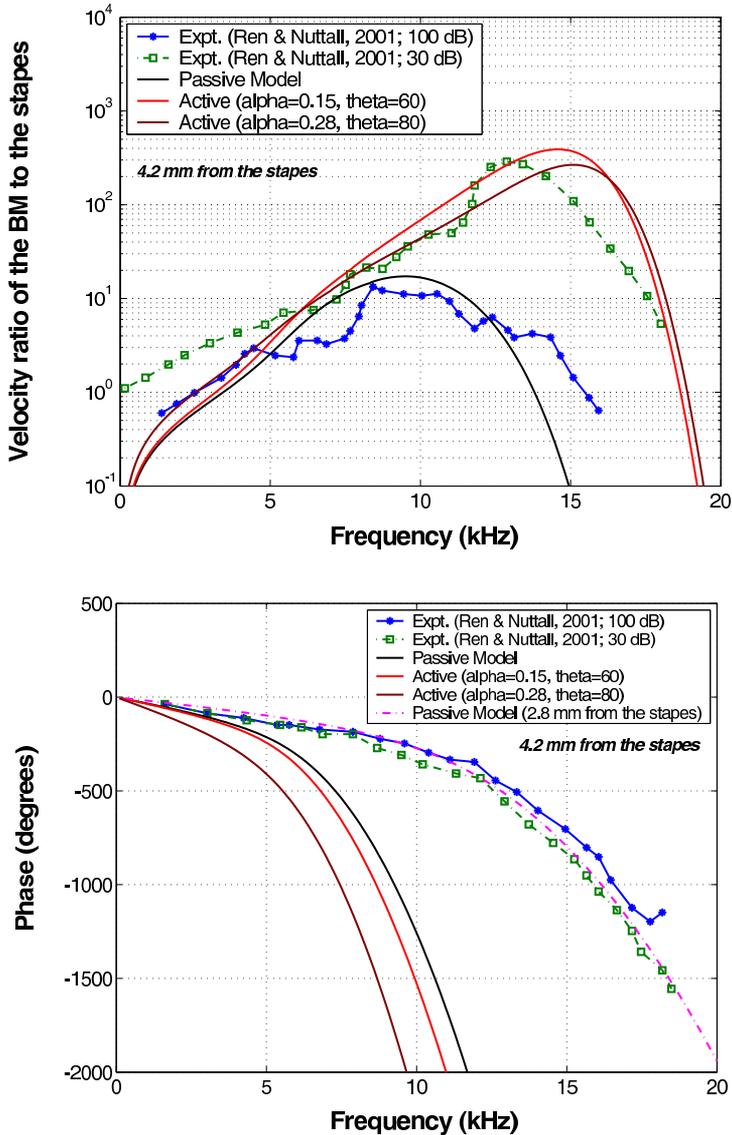
**Figure 3.** CF versus position for passive cochlear model (solid) compared to measurements (\*) and active cochlear model (dashed-dot). Present three-dimensional model represents cochlear CF-to-place map of gerbil [Sokolich et al. 1976; Greenwood 1990] over 0.3–50 kHz range spanning a length of 12 mm.

higher gain, still in the physiologically reasonable range, is needed for the OHC nearly perpendicular to the BM.

In the relative BM velocity magnitude plot (top of Figure 4), CF place shifts from 9.5 kHz (passive model) to 15 kHz (active model), which is  $3/5$  octave higher. In the animal measurement CF is also near 9.5 kHz for the high level passive case. For the low level active case, CF place shifts to about 13 kHz, which is only about  $2/5$  octave higher. So the model appears to overestimate the CF for the active case.

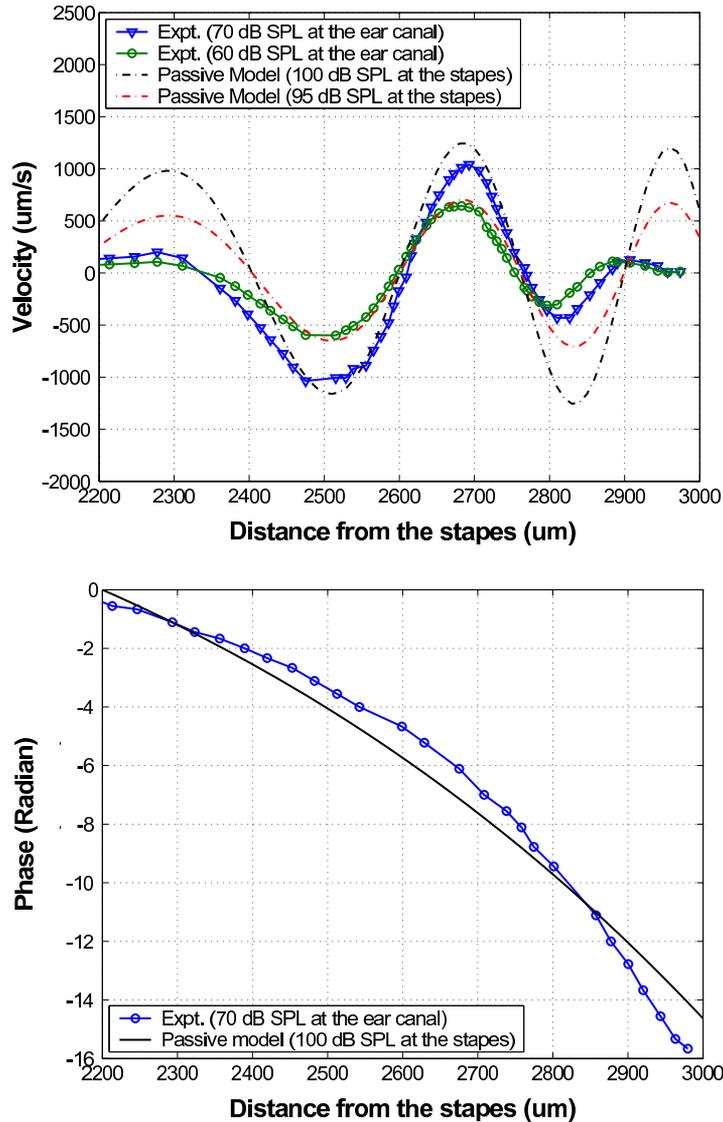
In the model, the phase is normalized to the volume flow rate at  $x = 0$ , as the stapes is assumed to be a piston at the end of the fluid chamber. As shown in the bottom of Figure 4, the phase of the response obtained from the model shows a larger roll-off with frequency than the experimental measurements. In the region of the low frequency input, below 4 kHz, the BM velocity phases both from the model and measurement are similar. However, after 4 kHz, the phase of BM velocity from the model shows a larger roll-off than the phase from the data, which means over fluctuation in the model above 4 kHz excitation frequency range. To match the phase from the model to the measurement, the phase at 2.8 mm from the stapes (CF of 15 kHz for the passive case) gives the same phase accumulation (magenta dashed-dot line in the bottom of Figure 4) as the data. The actual position of the stapes in the cochlea extends over a small portion of the basal end of the scala vestibuli, which may result in this discrepancy in the phase.

**3.3. Spatial response of BM velocity.** Ren [2002] measured the waveform of cochlear partition vibration along the cochlear partition from the gerbil cochlea in vivo by using a scanning laser interferometer. In this measurements, he could successfully obtain a *snapshot*, that is, the instantaneous waveform of the cochlear partition vibration, for 16 kHz tones in a longitudinal region of the BM (2200–3000  $\mu\text{m}$  from the stapes). In Figure 5, instantaneous waveforms and BM velocity phase for the passive case are presented.



**Figure 4.** BM velocity relative to the stapes: (top) magnitude and (bottom) corresponding phase for the gerbil cochlea at 4.2 mm from the base ( $CF = 9.5$  kHz). For active model, 0.15 feed-forward gain factor  $\alpha$  for  $\theta = 60^\circ$  (solid-red) and 0.28 feed-forward gain factor for  $\theta = 80^\circ$  (solid-brown) was used. Experimental data included for comparison [Ren and Nuttall 2001].

Envelope (magnitude) of the waveform from the model shows less sharp than envelope of the waveform from the measurements; see top of Figure 5. However, the peak place from the model is identical to the measurement (near  $2550 \mu\text{m}$  from the stapes). The snapshot and phase of the waveform from the model shows good agreement with experimental measurement. Around 16 kHz (CF at  $2550 \mu\text{m}$  from the



**Figure 5.** Longitudinal patterns of the instantaneous waveform (top) and phase (bottom) of the BM velocity from passive model and measurements [Ren 2002] near the CF location ( $\sim 2550 \mu\text{m}$ ). Data (70 dB and 60 dB SPL at the ear canal) collected with 16 kHz tones for sensitive gerbil cochlea. Considering 30 dB middle ear gain [Olson 1998], 100 dB and 95 dB SPL at the stapes are used as input pressure in the model, respectively.

stapes), instantaneous waveforms in the model also show energy propagation along the BM, supporting the existence of the cochlear traveling wave on the BM, which was observed in the *in vivo* measurements; see top of Figure 5. As shown in the bottom of Figure 5, BM velocity phase for the passive model shows around 5 radians delay at the CF for 16 kHz and the total phase delay is around 18 radians ( $\sim 6\pi$ ) over the observed range as in the measurement. Unlike the total phase change of the traveling wave in human

cadavers, which is about  $3\pi$  radians [Békésy 1960], the phase delay in the gerbil model and measurement is as much as  $6\pi$  radians over a spatial range of  $1000\ \mu\text{m}$  from the stapes.

#### 4. Conclusions

Measurements of the spatial distribution of the response of the BM at a fixed frequency [Ren 2002] and the frequency distribution at a fixed point [Ren and Nuttall 2001] offer an unusual opportunity for validation of model calculations. The macromechanical cochlear model is a simplified box with three-dimensional fluid and geometry from the gerbil anatomy. The BM properties are physical, with orthotropic elastic properties and no fictitious mass or damping. Hence there are no parameters to adjust to fit experimental results.

The comparison of results from the model and experiment is promising, but not fully satisfactory. Using a single set of anatomically based parameters, the model predicts several significant features of cochlear response. The CF-to-place map in the passive model, frequency and spatial responses of BM velocity were in close agreement with those observed in animal measurement. The feed-forward linear active model, the most speculative feature of the framework presented, showed excellent agreement with experimental data in the BM relative velocity magnitude. However, the calculated phase shows excellent agreement for the spatial distribution for fixed frequency, but a much larger roll-off for the frequency dependence at a fixed point. In contrast, the calculated amplitude shows excellent agreement for the fixed point and not quite as sharp a roll-off for the spatial distribution. These limitations in our current model could be resolved by including more detailed structures of the organ of Corti to the current model [Steele and Puria 2005].

#### References

- [Békésy 1960] G. Békésy, *Experiments in hearing*, McGraw-Hill series in psychology, McGraw-Hill, New York, 1960.
- [de Boer 1983] E. de Boer, “On active and passive cochlear models-towards a generalized analysis”, *J. Acoust. Soc. Am.* **73**:2 (1983), 574–576.
- [Böhnke and Arnold 1998] F. Böhnke and W. Arnold, “Nonlinear mechanics of the organ of Corti caused by Deiters cells”, *IEEE T. Bio-med. Eng.* **45**:10 (1998), 1227–1233.
- [Cohen et al. 1992] Y. E. Cohen, C. K. Bacon, and J. C. Saunders, “Middle ear development, III : morphometric changes in the conducting apparatus of the Mongolian gerbil”, *Hearing Res.* **62**:2 (1992), 187–193.
- [Dannhof et al. 1991] B. J. Dannhof, B. Roth, and V. Bruns, “Length of hair cells as a measure of frequency representation in the mammalian inner ear”, *Naturwissenschaften* **78**:12 (1991), 570–573.
- [Edge et al. 1998] R. M. Edge, B. N. Evans, M. Pearce, R. C. P., X. Hu, and P. S. Dallos, “Morphology of the unfixed cochlea”, *Hearing Res.* **124**:1-2 (1998), 1–16.
- [Geisler and Sang 1995] C. D. Geisler and C. Sang, “A cochlear model using feed-forward outer-hair-cell forces”, *Hearing Res.* **86**:1-2 (1995), 132–146.
- [Greenwood 1990] D. D. Greenwood, “A cochlear frequency-position function for several species-29 years later”, *J. Acoust. Soc. Am.* **87**:6 (1990), 2592–2605.
- [Karavitaki 2002] K. D. Karavitaki, *Measurements and models of electrically-evoked motion in the gerbil organ of Corti*, Ph.D. Thesis, MIT, 2002, Available at <http://hdl.handle.net/1721.1/8087>.
- [Kolston and Ashmore 1996] P. J. Kolston and J. F. Ashmore, “Finite element micromechanical modeling of the cochlea in three dimensions”, *J. Acoust. Soc. Am.* **99**:1 (1996), 455–467.

- [Lim and Steele 2002] K. M. Lim and C. R. Steele, “A three-dimensional nonlinear active cochlear model analyzed by the WKB-numeric method”, *Hearing Res.* **170**:1-2 (2002), 190–205.
- [Loh 1983] C. H. Loh, “Multiple scale analysis of the spirally coiled cochlea”, *J. Acoust. Soc. Am.* **74**:1 (1983), 95–103.
- [Miller 1985] C. E. Miller, “Structural implications of basilar membrane compliance measurements”, *J. Acoust. Soc. Am.* **77** (1985), 1465–1474.
- [Neely 1985] S. T. Neely, “Mathematical modeling of cochlear mechanics”, *J. Acoust. Soc. Am.* **78**:1 (1985), 345–352.
- [Neely 1993] S. T. Neely, “A model of cochlear mechanics with outer hair cell motility”, *J. Acoust. Soc. Am.* **94**:1 (1993), 137–146.
- [Olson 1998] E. S. Olson, “Observing middle and inner ear mechanics with novel intracochlear pressure sensors”, *J. Acoust. Soc. Am.* **103**:6 (1998), 3445–3463.
- [Parthasarathi et al. 2000] A. A. Parthasarathi, K. Grosh, and A. L. Nuttall, “Three-dimensional numerical modeling for global cochlear dynamics”, *J. Acoust. Soc. Am.* **107**:1 (2000), 474–485.
- [Ren 2002] T. Ren, “Longitudinal pattern of basilar membrane vibration in the sensitive cochlea”, *PNAS* **99**:26 (2002), 17101–17106.
- [Ren and Nuttall 2001] T. Ren and A. L. Nuttall, “Basilar membrane vibration in the basal turn of the sensitive gerbil cochlea”, *Hearing Res.* **151**:1-2 (2001), 48–60.
- [Smith 1968] C. A. Smith, “Ultrastructure of the organ of Corti”, *Adv. Sci.* **24**:122 (1968), 419–433.
- [Sokolich et al. 1976] W. G. Sokolich, R. P. Hamernik, J. J. Zwisioccki, and R. A. Schmiedt, “Inferred response polarities of cochlear hair cells”, *J. Acoust. Soc. Am.* **59**:4 (1976), 963–974.
- [Steele and Lim 1999] C. R. Steele and K. M. Lim, “Cochlear model with three-dimensional fluid, inner sulcus and feed-forward mechanism”, *Audiol. Neuro-Otol.* **4** (1999), 197–203.
- [Steele and Puria 2005] C. R. Steele and S. Puria, “Force on inner hair cell cilia”, *Int. J. Solids Struct.* **42**:21-22 (2005), 5887–5904.
- [Steele and Taber 1979] C. R. Steele and L. A. Taber, “Comparison of WKB calculations and experimental results for three-dimensional cochlear models”, *J. Acoust. Soc. Am.* **65**:4 (1979), 1007–1018.
- [Steele et al. 1993] C. R. Steele, G. Baker, J. Tolomeo, and D. Zetes, “Electro-mechanical models of the outer hair cell”, pp. 207–215 in *Proc. int. symp. biophysics of hair cell sensory systems*, edited by H. Duifhuis et al., World Scientific, Singapore, 1993.
- [Steele et al. 1995] C. R. Steele, G. Baker, J. Tolomeo, and D. Zetes, “Cochlear mechanics”, pp. 505–516 in *The biomedical engineering handbook*, edited by Z. D. Bronzino, CRC press, 1995.
- [Thorne et al. 1999] M. Thorne, A. N. Salt, J. E. DeMott, M. M. Henson, O. W. Henson Jr., and S. L. Gewalt, “Cochlear fluid space dimensions for six species derived from reconstructions of three-dimensional magnetic resonance images”, *Laryngoscope* **109**:10 (1999), 1661–1668.

Received 20 Jul 2006. Revised 17 Apr 2007. Accepted 20 Apr 2007.

YONGJIN YOON: [yongjiny@stanford.edu](mailto:yongjiny@stanford.edu)

*Mechanics and Computation Division, Stanford University, 496 Lomita Mall, Durand Building, Stanford, CA 94305-4035, United States*

SUNIL PURIA: [puria@stanford.edu](mailto:puria@stanford.edu)

*Department of Otolaryngology — Head and Neck Surgery, Stanford University, Stanford, CA 94305, United States*

CHARLES R. STEELE: [chasst@stanford.edu](mailto:chasst@stanford.edu)

*Mechanics and Computation Division, Stanford University, 496 Lomita Mall, Durand Building, Stanford, CA 94305-4035, United States*

## FLUSHING OF THE PORT OF ENSENADA USING A SIBEO WAVE-DRIVEN SEAWATER PUMP

STEVEN PETER CZITROM, CESAR CORONADO AND ISMAEL NUÑEZ

A SIBEO wave-driven seawater pump is proposed to inject clean and oxygen-rich seawater from outside the port of Ensenada, Baja California, to promote flushing in the more stagnant sections of the harbor. Results from a simple two-dimensional numerical model of the port hydrodynamics shed light on how the tides cannot on their own adequately flush the system. A three-dimensional model, which includes thermal stratification, illustrates how the pumped seawater ventilation can spread throughout the harbor via a density channel set up by the seasonal thermocline.

### 1. Introduction

Many human coastal settlements use the adjacent ocean to dispose of domestic and industrial refuse. Substantial growth of these settlements has resulted in an increased concentration of pollutants in the ocean, sometimes reaching levels that are dangerous for human habitation and the ecosystem health. This problem is further exacerbated in semienclosed coastal water bodies such as ports with breakwaters [Fischer et al. 1979].

The port of Ensenada, to the north of the Baja California Peninsula, Mexico, has witnessed brisk activity since it was established in the 19th century, and is today an important hub of development. The fishing, manufacture and tourist industries, among others, have grown substantially in support of social and economic development, increasing living standards of the local and state populations. In the last few decades growth has witnessed an explosion in size and diversity. This activity, however, has not been without cost to the port's ecosystem where domestic and industrial refuse have been dumped. Since the construction of breakwaters to protect shipping, a large section of the port has become increasingly isolated and stagnant, making it more vulnerable to the accumulation of pollutants.

There are essentially two ways to diminish the concentration of pollutants in a body of water. The first is to restrict the flow of contaminant by diminishing its input and/or providing treatment. The second is to increase the flow of unpolluted water through the system to encourage the expulsion of the accumulated contaminants. A combination of both measures is probably the most adequate allowing ventilation to be achieved in less time. Note that the added flushing should not be taken as a free ticket to increase the discharge of pollutants. This combination of solutions must take into account that it is not healthy in the long run to take the adjacent ocean as a universal and inexhaustible digester, into which one can pour endless quantities of contaminants [Fischer et al. 1979]. This, unfortunately, has been common practice throughout time in most parts of the world.

In the case of the port of Ensenada, in the last decade or so sewage treatment plants and a more strict enforcement of antipollution legislation have substantially diminished the input of contaminants

*Keywords:* wave energy, flushing of stagnant coastal water bodies.

to the harbor. High levels of pollution remain, however, despite the flushing action of the tides [Orozco and Gutiérrez 1983; Delgadillo and Orozco 1987; 1989; Segovia and Rivera 1988; Portillo and Lizárraga 1997; Macías et al. 1997]. An additional flow of clean and oxygen-rich water from the adjacent ocean into the stagnant and contaminated sections of the port is quite likely to promote a more effective ventilation. This flow would dilute the polluted waters and, as they are displaced towards sectors with a greater circulation, spread ventilation to a growing area.

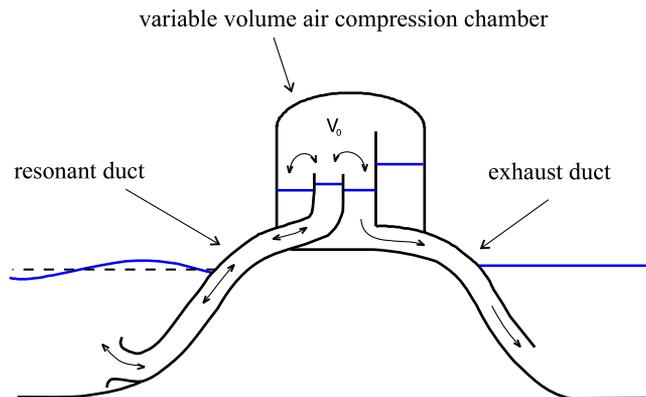
In this paper the application of a wave energy driven seawater pump is discussed as a means of delivering clean and oxygen-rich seawater into a stagnant polluted marine area to promote its ventilation. Hydrodynamic numerical models of the water circulation in the harbor of Ensenada are used to shed light on why significant pollution remains in the water and sediments of the northern section of the port, despite the flushing action of the tides, and by which mechanisms the flow of clean and oxygen-rich water from a wave-driven seawater pump can help ventilate the Ensenada harbor.

## 2. The SIBEO wave-driven seawater pump

Starting in the late 1980's, research has been conducted at the Instituto de Ciencias del Mar y Limnología of the National University of Mexico (UNAM) to develop technology which uses wave energy for pumping seawater. Czitrom [1997] proposed a wave-driven seawater pump (SIBEO<sup>1</sup>) in which a mechanical oscillator, composed of an air spring flanked by two water masses in ducts, is excited by the waves.

A schematic diagram of the SIBEO can be seen in Figure 1. The pump is primed by a partial vacuum that brings water up from the ocean and the receiving body of water to a working level in the compression chamber. The variable pressure signal induced by the waves at the resonant duct-mouth drives an oscillating flow which spills water in the compression chamber with each passing wave. The spilt water gathers in the chamber and descends by gravity to the receiving body of water via the exhaust duct; see also [Carey and Meratla 1976].

The system operates optimally at resonance when the frequency of the driving waves coincides with the SIBEO natural frequency of oscillation. A condition of resonance can be maintained, in an evolving



**Figure 1.** Schematic diagram of the SIBEO wave-driven seawater pump.

<sup>1</sup>SIBEO is an acronym for the Spanish Sistema de Bombeo por Energía de Oleaje.

wave field, by means of a variable volume compression chamber which adjusts the hardness of the mechanical oscillator air spring. The SIBEO natural frequency of oscillation can thus be matched to that of the most energetic driving waves at all times.

In practice, the wave field is composed of various frequencies with particular energies associated to each. Tuning to the highest energy frequency can be carried out automatically under control of a programmed microchip that samples the wave field and adjusts the volume of air required for resonance at the appropriate frequency. This tuning device, which can be used in other oscillating water column wave-energy conversion devices (see, for example, [Falnes and McIver 1985]), was patented through the National University of Mexico [Czitrom 2002].

Extensive theoretical and experimental studies back the SIBEO development. The pump equations were derived by applying the Bernoulli equation to streamlines in the resonant and exhaust ducts and adding terms for the losses due to viscosity, vortex formation and radiation damping [Czitrom 1997; Czitrom et al. 2000a]. A numerical model of the SIBEO, which solves the pump equations, reproduces 1:25 scale wave tank test data remarkably well [Czitrom et al. 2000b]. A SIBEO prototype was temporarily installed and field tested on the coast of Oaxaca, Mexico, with the help of a fisherman's cooperative [Czitrom 1996; 1997].

### 3. The port of Ensenada, Baja California

A general disposition of the port of Ensenada can be seen in Figure 2. At first glance it is apparent that the corner at the base of the main breakwater is one of the sections of the port most isolated from the adjacent ocean. The natural location for the SIBEO is near the breakwater base, where it is highly exposed to the incoming Pacific Ocean waves, and can have a greater impact over one of the more stagnant sections of the harbor.



**Figure 2.** Aerial view of the port of Ensenada, Baja California. North is at approximately one o'clock on this figure.

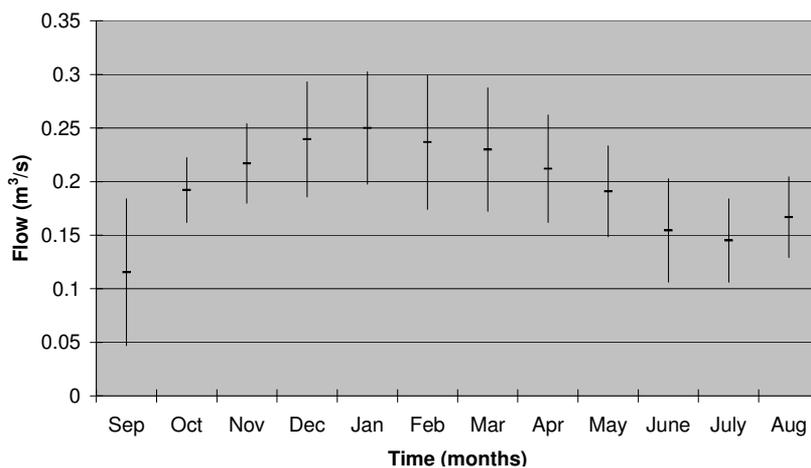
The pump numerical model was used to estimate the flow which would be generated throughout the year by the SIBEO with a 1.4 m diameter resonant duct (Figure 3). Maximum and minimum flows for each month were computed with the extreme values of wave amplitude and period observed in that lapse of time. An average yearly flow of some 200 liters/second can be expected from the SIBEO, varying between 50 and 300 l/s, mainly due to changes in the wave size.

A very crude indication of the effect a 200 l/s flow might have on the harbor is the time it takes to inject an equivalent volume of water. The port of Ensenada is approximately 1.9 km long, 0.8 km wide and 10 m deep with equivalent volume  $\sim 1.5 \times 10^7 \text{ m}^3$ , so that it would take 2.4 years to inject an equivalent volume of water using a single SIBEO. This figure suggests that, in the first few months, the ventilating effect would only be noticeable close to the location of the exhaust point. As an example, a volume of water equivalent to that contained in a quadrant of 500 m radius, at the base of the breakwater, would be injected by the SIBEO in somewhat less than 4 months.

By contrast with the SIBEO, the tides input a much greater volume of water through the navigation channel. The  $M_2$  constituent at Ensenada is the most significant with a range of about 1 m so that, given the area of the port, an average flow of  $70 \text{ m}^3/\text{s}$  enters during the 6 hours of the flood tide. This input is 350 times greater than that of the SIBEO making it clear that questions must be answered concerning the SIBEO effectiveness against that of the tides in their capacity to flush the port. In the first instance it is necessary to understand why tidal flushing has such a reduced ventilating effect on the contaminated waters and sediments of the northern section of the port. In the second we must clarify how a much smaller but focused flow from the SIBEO pump might more effectively achieve the desired ventilation.

#### 4. Port hydrodynamics

In order to answer these questions, two numerical models of the tidal and wind-driven hydrodynamics of the port of Ensenada were implemented. A two dimensional single layer version of the Hamburg Shelf Ocean Model (HAMSOM), a semiimplicit model known for its simplicity and proven performance

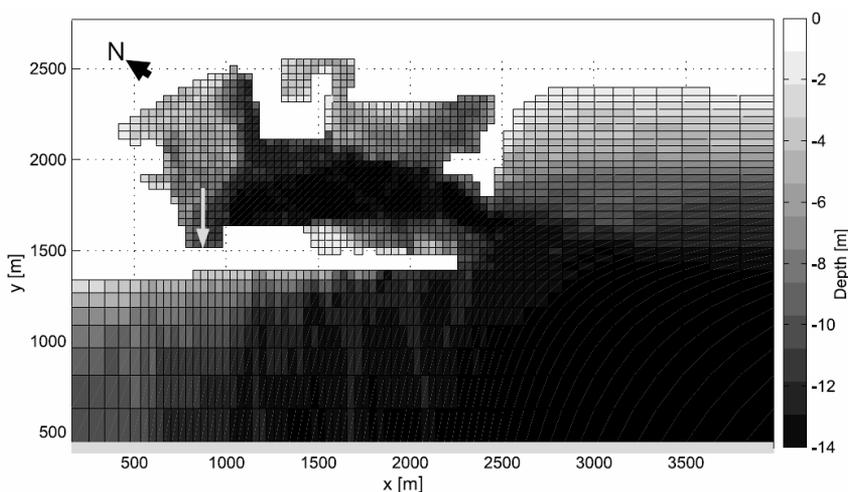


**Figure 3.** Estimated SIBEO flow, using wave data measured at the breakwater from 1986 to 1987 [Martínez Díaz de León et al. 1989].

[Backhaus 1983; 1985; Huang 1995] was chosen to provide relatively straightforward first estimates. The HAMSOM model can be easily set up to include external flows such as that of the SIBEO and has been applied with success to places such as the North Sea [Backhaus 1985], the delta of the Colorado River [Carbajal et al. 1997], and a coastal lagoon with river discharge on the west coast of Mexico [Núñez Riboni 2000], among others. A full account of the implementation of this model to the port of Ensenada can be found in [Czitrom et al. 2003].

Three-dimensional modeling has become a practical way of simulating circulation and the thermal-haline field in coastal lagoons [Ramírez and Imberger 2002; Balas and Özhan 2002], and estuaries [Cheng et al. 1993; Cheng and Casulli 2002]. The three dimensional Estuary and Lake Computer Model (ELCOM), which uses a semiimplicit finite difference solution scheme and includes thermodynamic effects, was developed by Hodges et al. [2000]. ELCOM can reproduce the first-order three-dimensional baroclinic physical response of an estuary to environmental and tidal forcing on a coarse grid with efficient CPU usage. The model has been recently applied to predict internal wave propagation in Lake Kinneret in Israel [Hodges et al. 2000]. Laval et al. [2003] improved the scalar and momentum mixing scheme used in ELCOM and successfully reproduced internal wave motions in Lake Kinneret. A full description of the application of the ELCOM model to the port of Ensenada, including calibration procedures and comparison to field measurements, can be found in [Coronado 2003; Coronado et al. 2007]. Results of this application are used here to examine the mechanisms by which the SIBEO flow can spread its ventilating effects throughout the harbor at times when the water column is stratified.

The Ensenada port bathymetry can be seen in Figure 4. The simulation domain for the two-dimensional HAMSOM model was cut off at the port entrance navigation channel while a section external to the port was included in the three-dimensional simulations to account for the influence of the stratified water column there. The harbor has a surface area of approximately 1.51 km<sup>2</sup> and opens into Todos Santos Bay. It is protected by a 1640 m long breakwater and a 570 m long El Gallo jetty. Semidiurnal tides, with



**Figure 4.** Bathymetry of the port of Ensenada [Coronado 2003]. A grey arrow points to the proposed location for the SIBEO exhaust.

a maximum range of 2 m, propagate from Todos Santos Bay via a 350 m wide channel entrance. The bathymetry of the harbor is characterized by a 13 m deep navigation channel, which runs parallel to the breakwater.

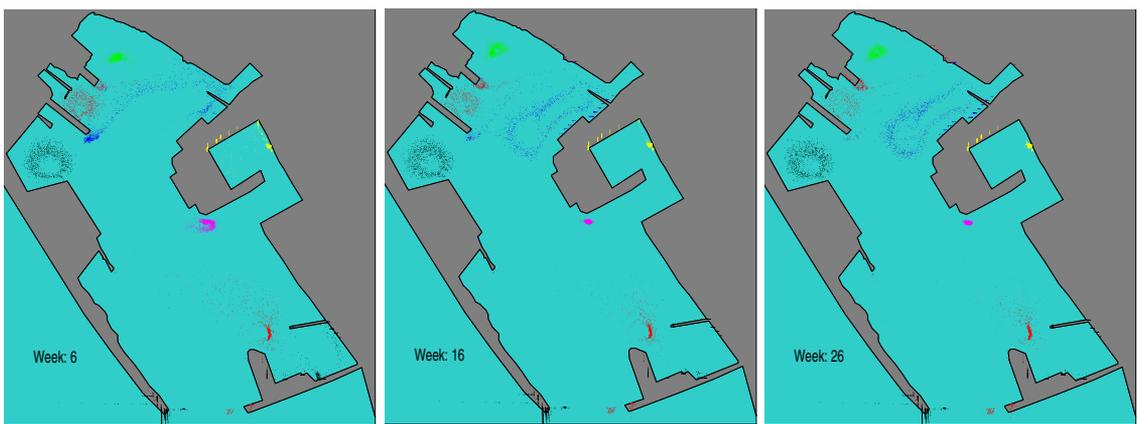
### 5. Two-dimensional model results

In order to visualize the dispersion of inert particles in the harbor, and thus shed light on the flushing characteristics of the port, the modeled two-dimensional circulation was used to simulate the trajectory of labeled water parcels released within the harbor during a period of 6 months. Clusters of 1000 color labeled water parcels were released at various points within the port and traced from one time step to the next using the velocity fields derived from the hydrodynamic model.

The erratic movement of water particles which occurs in turbulent flow causes them to disperse in spreading trajectories. Turbulent dispersion was simulated by introducing random variations in the particle velocity at each computational step. The lagrangian trajectories, which include advection as well as turbulent diffusion, can help identify the regions where particles are trapped in eddies and can thus be used to find the best location and flow intensity for the SIBEO to adequately flush the port.

Figure 5 shows the distribution of labeled water parcels after 6, 16 and 26 weeks of dispersion simulation for the case of the  $M_2$  tides without the SIBEO discharge in the port. With the exception of a couple of stagnation points, it is clear that most of the particles from the southern section of the harbor eventually reach the navigation channel through which they exit the system to the adjacent ocean. In the northern section the particles remain gyrating in a series of closed eddies from which they cannot escape. It appears that the tides have a flushing effect restricted to the southern section of the harbor while driving closed circulation patterns in the north which in effect trap the particles released there. This result seems to explain why the tides are not capable of renewing seawater in the more stagnant northern section of the Ensenada harbor, which thus remains contaminated despite the flushing action of the tides.

Figure 6 shows the distribution of particles after 6, 16 and 26 weeks of dispersion simulation by the  $M_2$  tides, with a  $0.6 \text{ m}^3/\text{s}$  flow from 3 SIBEOs pumped into the north-west corner of the port, and an



**Figure 5.** Dispersion of color labeled water parcels in the Ensenada Harbor by the  $M_2$  tides after 6, 16 and 26 weeks of simulation.

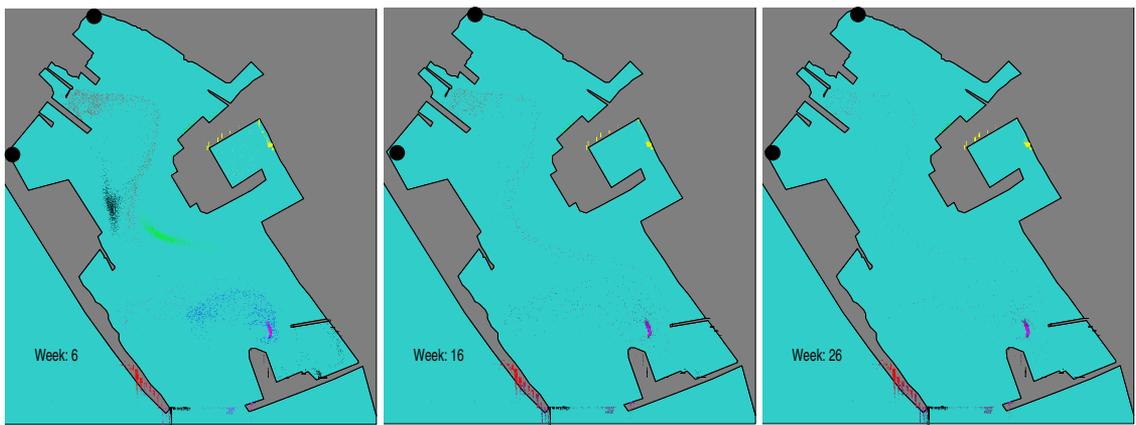
additional  $0.4 \text{ m}^3/\text{s}$  flow from 2 SIBEOs to the north; the SIBEO discharge points are marked with dots. Similar to the previous case, the southern portion of the harbor is adequately flushed by the tides. In the northern section, the corners at which the SIBEO pumps discharge are swept clean of particles by the injected water. The sequence of images suggests that a combined flow of  $1 \text{ m}^3/\text{s}$  at the two corners is adequate to flush the northern section of the port, displacing the particles southward, where the influence of the tides can eventually expel them through the navigation channel. The  $1 \text{ m}^3/\text{s}$  flow seems sufficient to alter the closed residual circulation eddies generated by the  $M_2$  tides in the north.

### 6. Three-dimensional model results

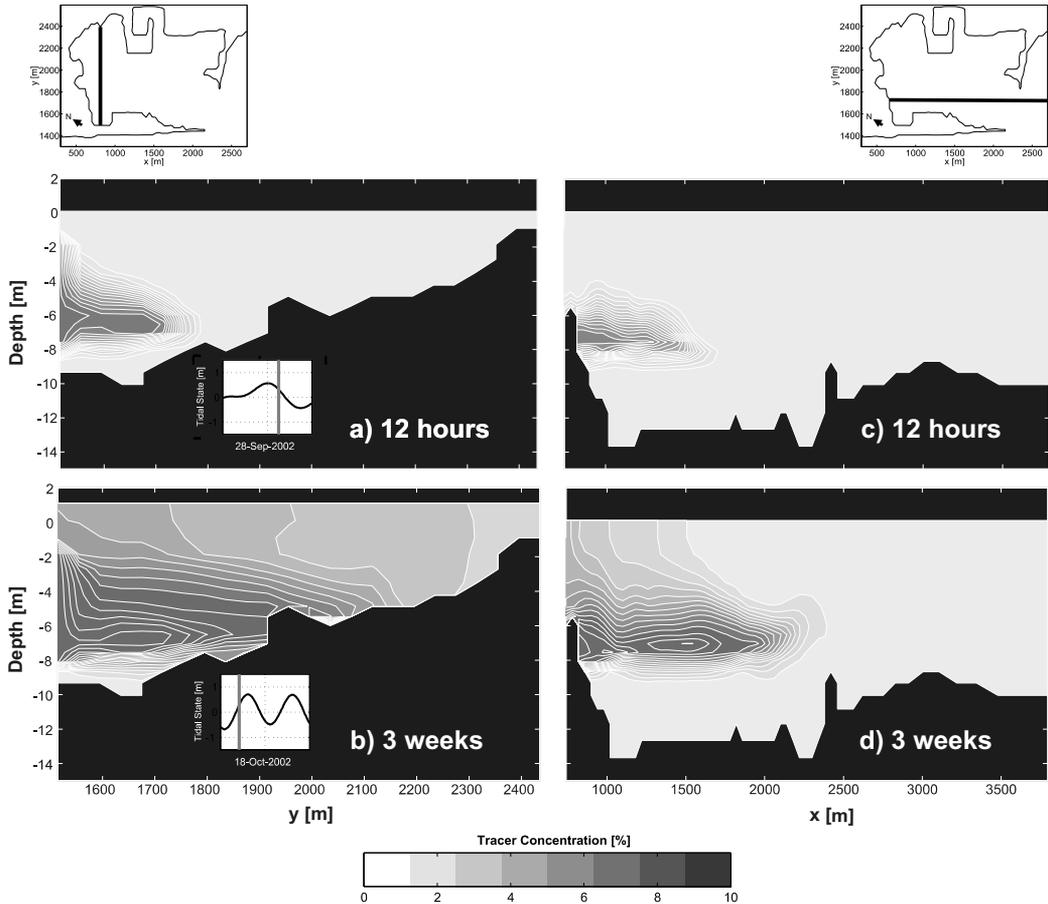
The two-dimensional model results provide reasonable answers to the questions posed in Section 3 when the water column is vertically mixed. During the spring and summer months, however, heating at the surface stratifies the water column, and a seasonal thermocline develops in the port and the continental shelf around. At this time, circulation and mixing processes are best described using a three-dimensional approach, for which the ELCOM model is most adequate.

In Figure 7 the SIBEO water concentration in vertical sections along and perpendicular to the main navigational channel are shown after 12 hours' and three weeks' simulations, respectively. The  $1 \text{ m}^3/\text{s}$  SIBEO flow was input at the surface near the breakwater base. It is apparent that the pumped water from the neighboring ocean first sinks to its density level within the port and then spreads along the thermocline. After 12 hours, the effect of the pumped water is noticeable only near the discharge point, reaching most of the port after a few weeks.

In Figure 8 the depth averaged modeled SIBEO water concentration throughout the harbor is shown at various time intervals up to 4 weeks. It is clear that in the first few hours the effect of the SIBEO water is noticeable near the discharge point at the base of the breakwater. After one week, however, the ventilating effect of the SIBEO water reaches most of the port, while in successive weeks, water external



**Figure 6.** Distribution of color labeled water parcels after 6, 16 and 26 weeks of dispersion simulation by the  $M_2$  tides. SIBEO flows of  $0.6 \text{ m}^3/\text{s}$  and  $0.4 \text{ m}^3/\text{s}$  were injected to the north west and due north corners of the harbor, respectively, at the positions marked with circles.

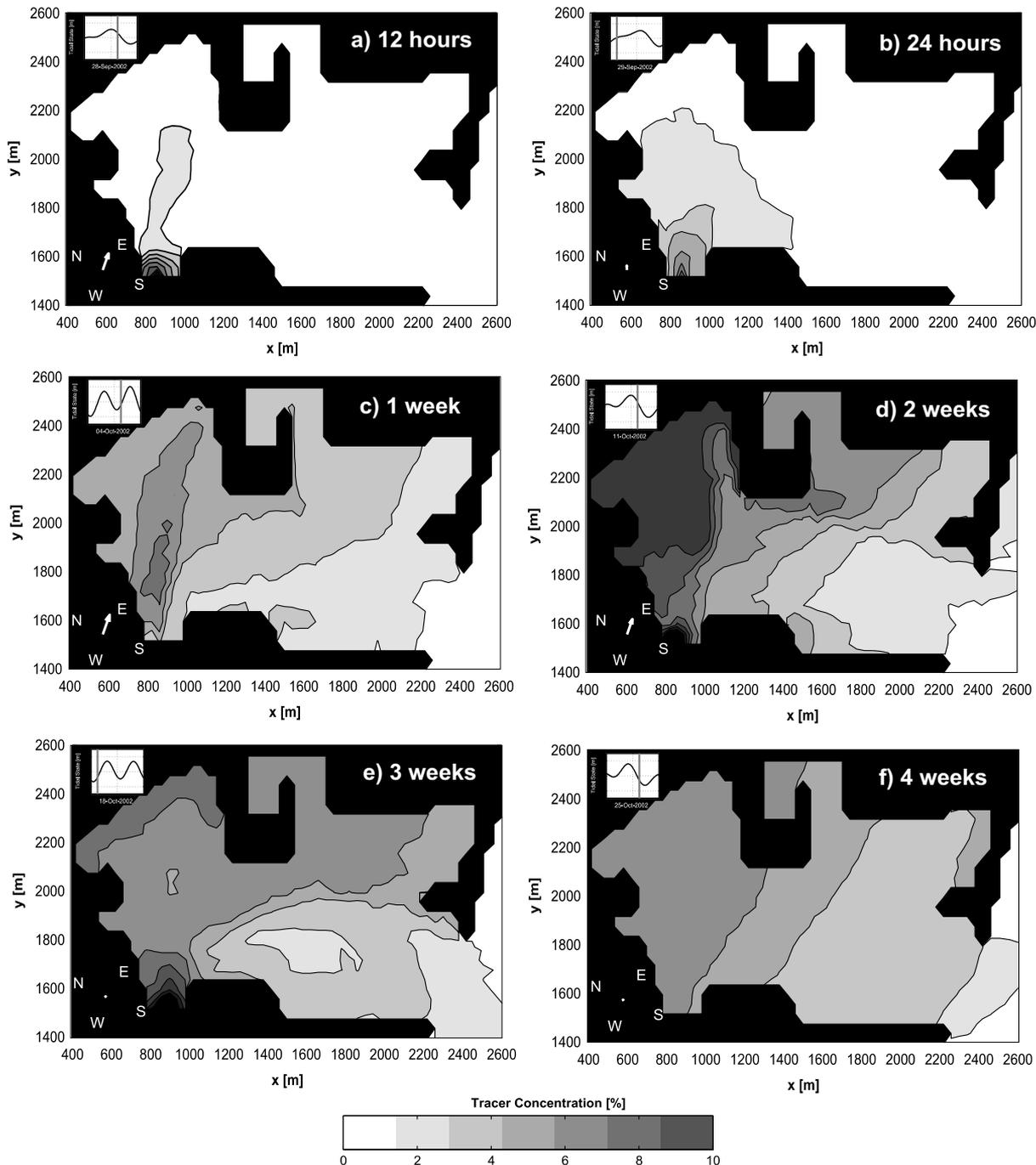


**Figure 7.** Snapshots of vertical distribution of the SIBEO water tracer along a transect at the head of the harbor (left panels) and along the main navigational channel (right panels). (a) and (c) show the tracer distribution after 12 hours of simulation, (b) and (d) after 3 weeks. Tide state is indicated in (a) and (b). Note that the horizontal scales in left panels differ from the right ones. Discharge location is at the left of each panel.

to the port spreads in patterns which are less noticeably linked to the SIBEO exhaust position. In the last panel, higher concentrations appear in the northern section of the port, which is in effect the most polluted and where the ventilating effect of clean and oxygen-rich water is most beneficial.

## 7. Conclusions

At times when the water column is vertically mixed, closed eddy circulation patterns driven by the tides in the northern section of the port of Ensenada inhibit flushing of the stagnant polluted waters there. A proposed  $1 \text{ m}^3/\text{s}$  flow of clean and oxygen-rich seawater from outside the port, forced into this section by SIBEO wave-driven seawater pumps, would alter these patterns enough to flush the contaminated waters southward, where the tides are able to expel them from the port.



**Figure 8.** Snapshots of the depth-averaged distributions of the SIBEO water tracer after (a) 12 hours of simulation, (b) 24 hours, (c) 1 week, (d) 2 weeks, (e) 3 weeks and (f) 4 weeks. Tidal state and wind vector at snapshot instant are in each panel.

At times in summer when the water column is stratified due to heat input at the surface, the clean and oxygen-rich water from outside the port sinks to its density level, spreading the beneficial ventilating effect throughout the port via the density channel formed by the pycnocline.

### Acknowledgements

This work was supported by the DOF, CICESE, the CONACYT (Project 33354T), the Fondo Sectorial de Investigación Ambiental SEMARNAT-CONACYT (Project 2002-C01-0016), CONACYT Grant U47899-F and the Centre for Water Research of The University of Western Australia. Meteorological data were supplied by the Dirección General de Investigación y Desarrollo, Estación de Investigación Oceanográfica of Ensenada, Secretaría de Marina. We are grateful to the authorities of the Administración Portuaria Integral de Ensenada for their support with the field measurements. Thanks are also due to Dr. Andrew Brooker and Dr. Ben R. Hodges for general support with ELCOM. Part of this paper appears in the B.Sc. Thesis of C. Coronado, supervised by Dr. Isabel Ramírez, Dr. Rafael Hernández and Dr. Carlos Torres. Their generous support is gratefully acknowledged.

### References

- [Backhaus 1983] J. O. Backhaus, "A semiimplicit scheme for the shallow water equations for applications to shelf sea modelling", *Cont. Shelf. Res.* **2** (1983), 243–254.
- [Backhaus 1985] J. O. Backhaus, "A three-dimensional model for the simulation of the shelf sea dynamic", *Deutsche hydrographische Zeitschrift* **38** (1985), 165–254.
- [Balas and Özhan 2002] L. Balas and E. Özhan, "Three-dimensional modelling of stratified coastal waters", *Estuar. Coast. Shelf. S.* **54** (2002), 75–87.
- [Carbajal et al. 1997] N. Carbajal, A. Souza, and R. Durazo, "A numerical study of the ex-ROFI of the Colorado river", *J. Marine Syst.* **12** (1997), 17–33.
- [Carey and Meratla 1976] D. J. Carey and Z. Meratla, British Patent 1,572,086, 1976. granted 1980.
- [Cheng and Casulli 2002] R. Cheng and V. Casulli, "Evaluation of the UnTRIM model for 3D tidal circulation", pp. 628–641 in *Estuarine and coastal modelling*, edited by M. Spaulding, ASCE, 2002.
- [Cheng et al. 1993] R. Cheng, V. Casulli, and J. Gartner, "Tidal, residual, intertidal mudat (TRIM) model and its applications to San Francisco Bay, California", *Estuar. Coast. Shelf. S.* **36** (1993), 235–280.
- [Coronado 2003] C. Coronado, "Aplicación de un modelo hidrodinámico tridimensional en el puerto de ensenada", B. Sc. Thesis, Facultad de Ciencias Marinas, Baja California, México, 2003. UABC, 74.
- [Coronado et al. 2007] C. Coronado, S. Czitrom, and J. Imberger, "Three-dimensional modelling of the effect of a wave-driven seawater pump inflow into the Port of Ensenada, Mexico", 2007. In preparation.
- [Czitrom 1996] S. P. R. Czitrom, "Sea water pumping by resonance I", pp. 366–370 in *Proceedings of the second european wave power conference*, United Kingdom, 1996. ISBN 92-827-7492-9.
- [Czitrom 1997] S. P. R. Czitrom, "Wave energy-driven resonant sea-water pump", *J. Offshore Mech. Arct. Eng. (Trans. ASME)* **119** (1997), 191–195.
- [Czitrom 2002] S. P. R. Czitrom, "Sintonizador para sistemas de extracción de energía de oleaje que operan por resonancia", Instituto de Ciencias del Mar y Limnología, 2002. Mexican patent No. 210329.
- [Czitrom et al. 2000a] S. P. R. Czitrom, R. Godoy, P. E., P. Pérez, and R. Peralta-Fabi, "Hydrodynamics of an oscillating water column seawater pump, part I: theoretical aspects", *Ocean Eng.* **27**:11 (2000a), 1181–1198.
- [Czitrom et al. 2000b] S. P. R. Czitrom, R. Godoy, E. Prado, A. Olvera, and C. Stern, "Hydrodynamics of an oscillating water column seawater pump, part II: tuning to monochromatic waves", *Ocean Eng.* **27**:11 (2000b), 1199–1219.

- [Czitrom et al. 2003] S. Czitrom, I. N. Núñez, and I. Ramírez, “Innovative uses of wave power: environmental management of the port of Ensenada”, *Mar. Technol. Soc. J.* **36**:4 (2003), 74–84.
- [Delgadillo and Orozco 1987] H. y. Delgadillo and B. Orozco, “Bacterias patógenas en bahía de todos santos”, *Cienc. Mar.* **13**:3 (1987), 31–38.
- [Delgadillo and Orozco 1989] H. y. Delgadillo and B. Orozco, “Contaminación fecal en sedimentos superficiales de la bahía de todos santos, Baja California”, *Cienc. Mar.* **15**:1 (1989), 47–62.
- [Falnes and McIver 1985] J. Falnes and P. McIver, “Surface wave interactions with systems of oscillating bodies and pressure distributions”, *Appl. Ocean. Res.* **7**:4 (1985), 225–234.
- [Fischer et al. 1979] H. B. Fischer, E. J. List, R. C. Y. Koh, J. Imberger, and N. Brooks, *Mixing in inland and coastal waters*, Academic Press Inc., 1979.
- [Hodges et al. 2000] B. Hodges, J. Imberger, A. Saggio, and K. Winters, “Modelling basinscale internal waves in a stratified lake”, *Limnol. Oceanogr.* **45**:7 (2000), 1603–1620.
- [Huang 1995] D. Huang, “Modelling studies of barotropic and baroclinic dynamics in the Bohai sea”, Technical report, Beriche aus dem Zentrum für Meeres- und Klimaforschung, 1995.
- [Laval et al. 2003] B. Laval, J. Imberger, B. Hodges, and R. Stocker, “Modelling circulation in lakes: Spatial and temporal variations”, *Limnol. Oceanogr.* **48**:3 (2003), 983–994.
- [Martínez Díaz de León et al. 1989] A. Martínez Díaz de León, C. y. Nava Button, and F. J. Ocampo Torres, “Estadística del oleaje en la Bahía de Todos Santos, Baja California, de Septiembre de 1986 a Agosto de 1987”, *Cienc. Mar.* **15**:3 (1989), 1–20.
- [Macías et al. 1997] V. Macías, J. Macías, and J. Villaescusa, “Organotin compounds in marine water and sediments from the Port of Ensenada, Baja California”, *Cienc. Mar.* **23**:3 (1997), 377–394.
- [Orozco and Gutiérrez 1983] M. V. y. Orozco and G. E. A. Gutiérrez, “Contaminación fecal costera en la zona del puerto de Ensenada, Baja California”, *Cienc. Mar.* **9**:1 (1983), 27–34.
- [Portillo and Lizárraga 1997] A. Portillo and M. Lizárraga, “Detección de *Vibrio cholerae* 01 en diferentes hábitats de la Bahía de Todos Santos, Baja California”, *Cienc. Mar.* **23** (1997), 435–447.
- [Ramírez and Imberger 2002] I. Ramírez and J. Imberger, “The numerical simulation of the hydrodynamics of Barbamarco Lagoon”, *Appl. Numer. Math.* **40** (2002), 273–289.
- [Núñez Riboni 2000] I.-D. Núñez Riboni, “Dinámica y procesos dispersivos en el complejo lagunar bahía de Altata / Ensenada del pabellón, Sinaloa”, Tesis de Maestría, Unidad Académica Mazatlán del Instituto de Ciencias del Mar y Limnología, Universidad Nacional Autónoma de México, 2000.
- [Segovia and Rivera 1988] Z. y. Segovia and D. Rivera, “Efectos de desechos orgánicos en las zonas adyacentes a los efluentes de bahía de todos santos”, *Cienc. Mar.* **14**:4 (1988), 101–116.

Received 17 Aug 2006. Revised 17 Apr 2007. Accepted 20 Apr 2007.

STEVEN PETER CZITROM: [czitrom@mar.icmyl.unam.mx](mailto:czitrom@mar.icmyl.unam.mx)

Instituto de Ciencias del Mar y Limnología, Universidad Nacional Autónoma de México, Circuito Exterior S/N, Ciudad Universitaria, 04510 México D.F., Mexico

CESAR CORONADO: [coronado@cicese.mx](mailto:coronado@cicese.mx)

Centro de Investigación Científica y de Educación Superior de Ensenada, Km 107 Carretera Tijuana-Ensenada, 22860 Ensenada, Baja California, Mexico

ISMAEL NUÑEZ: [Ismael.Nunez-Riboni@awi.de](mailto:Ismael.Nunez-Riboni@awi.de)

Alfred-Wegener Institut für Polar und Meeresforschung, Bussestrasse 24, D-27570, Bremerhaven, Germany

# SYMMETRY ANALYSIS OF EXTREME AREAL POISSON'S RATIO IN ANISOTROPIC CRYSTALS

LEWIS WHEELER AND CLIFF YI GUO

Poisson's ratio is defined as the negative of the ratio of the transverse strain to the longitudinal strain in response to a longitudinal uniaxial stress. In the presence of anisotropy, this means that the ratio depends on two directions. With a view to assessing crystals that exhibit directions for which the ratio is negative, we resort to a transverse average to eliminate one directional variable and at the same time to arrive at a measure that poses a challenge to achieving significant negative values. The areal Poisson ratio coincides with the Poisson ratio for an isotropic material. We determine the stationary directions of the areal Poisson ratio for all crystal symmetry classes. The directions represented by invariant stationary points—those that hold independently of the material—we identify and explain class-by-class in terms of the axes of symmetry for the class. It is shown that for cubic crystals, positive definiteness of the strain energy requires that the areal Poisson ratio lie between  $-1$  and  $1/2$ , as it does for isotropy. We conclude that the areal Poisson ratio for the classes of lower symmetry are not restricted.

## 1. Introduction

Over the last two decades there has been increasing interest in finding, creating, and understanding material structures that exhibit a negative Poisson's ratio describing materials that are referred to as *auxetic*, a term attributed to [Evans et al. \[1991\]](#). While much of the work has focused on microstructures, there is an abundance of crystal structures that possess a negative ratio values for specific directions due to their anisotropic nature. The knowledge that a crystal may possess a negative Poisson's ratio is by no means recent. [Love \[1927\]](#) mentions a pyrite that yields a value near  $-1/7$ . Moreover, auxeticity in crystals is not uncommon, since nearly 69 of cubic elemental metals have a negative Poisson's ratio when the stressed axis lies along the  $[110]$  direction [\[Baughman et al. 1998\]](#). [Ting and Barnett \[2005\]](#) derived simple necessary and sufficient conditions on elastic compliances to identify if any given material of cubic or hexagonal symmetry is completely auxetic or nonauxetic. Further examples of auxetic behavior in crystals of cubic, hexagonal, and monoclinic symmetry are discussed in [\[Tokmakova 2005\]](#) with the aid of stereographic projections.

The meaning of the Poisson's ratio in the presence of anisotropy raises questions that are not apparent in the isotropic case. Not only does the ratio depend upon the choice of a direction for the longitudinal strain, but all directions at right angles to it for the transverse strain component. This transverse variation is apt to yield offsetting ratios [\[Baughman et al. 1998\]](#), a negative value for one transverse direction and a positive value for another, that diminish or negate the auxetic effect. [Guo and Wheeler \[2006\]](#)

---

*Keywords:* auxetic, areal Poisson's ratio, crystal anisotropy.

The authors wish to acknowledge the support of the Texas Institute for Intelligent Bio-Nano Materials and Structures for Aerospace Vehicles, funded by NASA Cooperative Agreement No. NCC-1-02038.

introduced an *areal* Poisson's ratio that is relatively simple in form and serves to measure the offsetting effects. The search for auxetic directions leads naturally to the search for the direction of the minimum areal Poisson's ratio, and more broadly to an examination of directions that yield stationary values. Of special interest, as we demonstrate, are stationary directions that are related only to the symmetry of the material and bear a simple relation to the crystallographic directions. For each crystal class, we find the stationary directions of the areal Poisson's ratio, examine their extremal nature, and graphically illustrate them for a particular crystal within the class.

The effect of crystal symmetry on the elastic constants of crystals is covered thoroughly in [Nye 1957; Ting 1996]. Cazzani and Rovati [2003; 2005] examine the directionality and extrema of Young's modulus in crystals of cubic, transversely isotropic and tetragonal symmetry. Ting and Chen [2005] proved that for all of the seven crystal classes, the Poisson's ratio can have an arbitrarily large positive or negative value under the constraint of positive definiteness of the strain energy density. In contrast, for the cubic crystal class we conclude here that the areal Poisson's ratio must lie within bounds. For the remaining crystal classes, there are no bounds on the areal Poisson's ratio.

In this paper, we investigate the directional variation of the areal Poisson's ratio for all nine crystal classes. Stationary directions that are independent of the material are called *invariant* stationary points. The directions represented by invariant stationary points are related to the axes of symmetry belonging to the particular crystal class. Where sensible, both the invariant and material dependent stationary directions are found, and their extremal nature is discussed. Based on the values of the areal Poisson's ratio at stationary directions and positive definiteness of the compliance tensor, we analyze the boundedness of the areal Poisson's ratio for each crystal class.

## 2. Preliminaries

We denote by  $\mathbb{C}$  the linear operator on the linear space of all symmetric 2-tensors that accounts for the elastic properties in the linear theory of anisotropic elastic solids. The elasticity operator  $\mathbb{C}$  and its adjoint  $\mathbb{C}^*$ , are related by

$$\langle \mathbf{A}, \mathbb{C}[\mathbf{B}] \rangle = \langle \mathbb{C}^*[\mathbf{A}], \mathbf{B} \rangle,$$

under the inner product

$$\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr } \mathbf{A} \mathbf{B}^T.$$

Here, the elasticity operator  $\mathbb{C}$  is required to be self adjoint,  $\mathbb{C} = \mathbb{C}^*$ , in other words to possess the major symmetry, so that

$$\langle \mathbf{A}, \mathbb{C}[\mathbf{B}] \rangle = \langle \mathbf{B}, \mathbb{C}[\mathbf{A}] \rangle.$$

Let  $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$  denote a right-handed orthonormal frame, for short a *cartesian* frame. Define  $\mathbf{E}_{ij}$  as

$$\mathbf{E}_{ij} = \text{sym}(\mathbf{e}_i \otimes \mathbf{e}_j).$$

The set  $\{\mathbf{E}_{ij}\}$  is an orthogonal basis for the linear space of 2-tensors. These basis elements  $\mathbf{E}_{ij}$  though orthogonal are not normalized, but rather obey

$$\langle \mathbf{E}_{ij}, \mathbf{E}_{kl} \rangle = \frac{1}{2} (\delta_{ik} \delta_{jl} + \delta_{il} \delta_{jk}), \quad (1)$$

which implies

$$|\mathbf{E}_{ij}|^2 = \begin{cases} 1, & i = j, \\ \frac{1}{2}, & i \neq j. \end{cases}$$

The components of  $\mathbb{C}$  in the frame  $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$  are given by Gurtin [1972]:

$$C_{ijkl} = \langle \mathbf{E}_{ij}, \mathbb{C}[\mathbf{E}_{kl}] \rangle. \tag{2}$$

These components are simultaneously the components of the operator  $\mathbb{C}$  and the fourth-order tensor associated with  $\mathbb{C}$ .

The components  $I_{ijkl}$  of the identity  $\mathbb{I}$  are given by the right side of Equation (1),

$$I_{ijkl} = \frac{1}{2} (\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}). \tag{3}$$

We assume for the remainder of this presentation that  $\mathbb{C}$  is positive definite. Thus,  $\mathbb{C}$  has an inverse, the compliance operator, denoted by  $\mathbb{S}$ , that, like  $\mathbb{C}$ , is self-adjoint and positive definite.

The reduced forms of the matrix of elastic constants that appear in [Nye 1957] and [Gurtin 1972] represent these constants in a preferred frame, which we denote by  $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$  to distinguish it from the generic frame  $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ . Remarkably, this frame may be taken as orthonormal. Here, we frequently refer to the frame  $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$  as a *crystallographic* frame. The crystallographic counterparts of the basis elements  $\mathbf{E}_{ij}$  are denoted by  $\mathbf{A}_{ij}$ .

The Voigt compliances  $s_{ij}$  and the corresponding crystallographic tensor components  $S_{ijkl}$  are related through [Nye 1957],

$$(s_{ij}) = \begin{pmatrix} s_{11} & s_{12} & s_{13} & s_{14} & s_{15} & s_{16} \\ & s_{22} & s_{23} & s_{24} & s_{25} & s_{26} \\ & & s_{33} & s_{34} & s_{35} & s_{36} \\ & & & s_{44} & s_{45} & s_{46} \\ & & & & s_{55} & s_{56} \\ & & & & & s_{66} \end{pmatrix} = \begin{pmatrix} S_{1111} & S_{1122} & S_{1133} & 2S_{1123} & 2S_{1131} & 2S_{1112} \\ & S_{2222} & S_{2233} & 2S_{2223} & 2S_{2231} & 2S_{2212} \\ & & S_{3333} & 2S_{3323} & 2S_{3331} & 2S_{3312} \\ & & & 4S_{2323} & 4S_{2331} & 4S_{2312} \\ & & & & 4S_{3131} & 4S_{3112} \\ & & & & & 4S_{1212} \end{pmatrix}. \tag{4}$$

### 3. Definition of the Poisson's ratio and areal Poisson's ratio for an anisotropic crystal

Consider a unit uniaxial stress

$$\boldsymbol{\tau} = \mathbf{l} \otimes \mathbf{l}, \quad |\mathbf{l}| = 1$$

in the direction  $\mathbf{l}$ . The longitudinal strain  $\boldsymbol{\varepsilon}(\mathbf{l})$  is given by

$$\boldsymbol{\varepsilon}(\mathbf{l}) = \mathbf{l} \bullet \boldsymbol{\varepsilon} \mathbf{l} = \mathbf{l} \bullet \mathbb{S}[\mathbf{l} \otimes \mathbf{l}] \mathbf{l} = \langle \mathbf{l} \otimes \mathbf{l}, \mathbb{S}[\mathbf{l} \otimes \mathbf{l}] \rangle. \tag{5}$$

Let  $\mathbf{t}$  be a given direction perpendicular to  $\mathbf{l}$ , that is,  $\mathbf{l} \bullet \mathbf{t} = \mathbf{0}$ ,  $|\mathbf{t}| = 1$ . The strain  $\boldsymbol{\varepsilon}(\mathbf{t})$  in the transverse direction  $\mathbf{t}$  is given by

$$\boldsymbol{\varepsilon}(\mathbf{t}) = \mathbf{t} \bullet \boldsymbol{\varepsilon} \mathbf{t} = \mathbf{t} \bullet \mathbb{S}[\mathbf{l} \otimes \mathbf{l}] \mathbf{t} = \langle \mathbf{t} \otimes \mathbf{t}, \mathbb{S}[\mathbf{l} \otimes \mathbf{l}] \rangle. \tag{6}$$

The Poisson’s ratio corresponding to the longitudinal direction  $\mathbf{l}$  and the transverse direction  $\mathbf{t}$  is defined as

$$\nu(\mathbf{l}, \mathbf{t}) = -\frac{\varepsilon(\mathbf{t})}{\varepsilon(\mathbf{l})},$$

and in view of Equations (5) and (6) is expressed in terms of the compliance in the form

$$\nu(\mathbf{l}, \mathbf{t}) = -\frac{\langle \mathbf{t} \otimes \mathbf{t}, \mathbb{S}[\mathbf{l} \otimes \mathbf{l}] \rangle}{\langle \mathbf{l} \otimes \mathbf{l}, \mathbb{S}[\mathbf{l} \otimes \mathbf{l}] \rangle}. \tag{7}$$

For given orthogonal unit vectors  $\mathbf{l}$  and  $\mathbf{t}$ , the ratio is determined by the elastic properties of the crystal. We note that for  $\mathbb{S}$  positive definite, the denominator is positive, so the sign of  $\nu$  is determined by the numerator. The areal Poisson’s ratio is defined by

$$\widehat{\nu}(\mathbf{l}) = \frac{1}{2\pi} \int_0^{2\pi} \nu(\mathbf{l}, \mathbf{t}(\alpha)) d\alpha.$$

It is readily seen that this averaging reduces to finding the average of  $\mathbf{t} \otimes \mathbf{t}$ , with the result

$$\widehat{\nu}(\mathbf{l}) = -\frac{\langle \langle \mathbf{t} \otimes \mathbf{t} \rangle, \mathbb{S}[\mathbf{l} \otimes \mathbf{l}] \rangle}{\langle \mathbf{l} \otimes \mathbf{l}, \mathbb{S}[\mathbf{l} \otimes \mathbf{l}] \rangle}$$

where

$$\langle \mathbf{t} \otimes \mathbf{t} \rangle = \frac{1}{2}(\mathbf{I} - \mathbf{l} \otimes \mathbf{l}). \tag{8}$$

Therefore,

$$\widehat{\nu}(\mathbf{l}) = \frac{1}{2} \left( 1 - \frac{\text{tr } \mathbb{S}[\mathbf{l} \otimes \mathbf{l}]}{\langle \mathbf{l} \otimes \mathbf{l}, \mathbb{S}[\mathbf{l} \otimes \mathbf{l}] \rangle} \right). \tag{9}$$

Of course,  $\widehat{\nu}$  reduces to the Poisson’s ratio if  $\mathbb{S}$  is isotropic.

The direction  $\mathbf{l}$  of the stressed axis can be expressed in spherical coordinates,

$$\mathbf{l} = \cos \theta \sin \phi \mathbf{a}_1 + \sin \theta \sin \phi \mathbf{a}_2 + \cos \phi \mathbf{a}_3,$$

where  $0 \leq \phi \leq \pi$ ,  $0 \leq \theta < 2\pi$ . Thus, the areal Poisson’s ratio can be expressed in terms of the polar angles  $\phi$  and  $\theta$  through

$$\widehat{\nu}(\mathbf{l}) = \widehat{\nu}(\phi, \theta) = \frac{1}{2} \left( 1 - \frac{\text{tr } \mathbb{S}[\mathbf{l}(\phi, \theta) \otimes \mathbf{l}(\phi, \theta)]}{\langle \mathbf{l}(\phi, \theta) \otimes \mathbf{l}(\phi, \theta), \mathbb{S}[\mathbf{l}(\phi, \theta) \otimes \mathbf{l}(\phi, \theta)] \rangle} \right).$$

To identify the directions  $\mathbf{l}$  for which the areal Poisson’s ratio attains extreme values, we begin by examining stationary directions, those for which

$$\begin{cases} \widehat{\nu}_\phi = \frac{\partial \widehat{\nu}(\phi, \theta)}{\partial \phi} = 0, \\ \widehat{\nu}_\theta = \frac{\partial \widehat{\nu}(\phi, \theta)}{\partial \theta} = 0, \end{cases} \tag{10}$$

which we also at times refer to as stationary “points”. With the aid of the matrix

$$J = \begin{pmatrix} \widehat{v}_{\phi\phi} & \widehat{v}_{\phi\theta} \\ \widehat{v}_{\phi\theta} & \widehat{v}_{\theta\theta} \end{pmatrix}, \tag{11}$$

we are able to further analyze the stationary points. If  $J$  is nonsingular, we can determine the extremal nature of a stationary point. If the matrix is sign definite, there is an extreme point—a minimum if positive definite, a maximum if negative definite. For  $J$  nonsingular but indefinite, there is a saddle point. If  $J$  is singular, additional analysis is required.

#### 4. Poisson's ratio for the isotropic case

For an isotropic medium, the elasticity tensor may be expressed in spectral form as

$$\mathbb{C} = 3k \frac{1}{3} \mathbf{I} \otimes \mathbf{I} + 2\mu (\mathbb{I} - \frac{1}{3} \mathbf{I} \otimes \mathbf{I}), \tag{12}$$

where  $k$  denotes the bulk modulus and  $\mu$  stands for the shear modulus. The principal values of  $\mathbb{C}$  are  $3k$  and  $2\mu$ . They are coefficients of orthogonal projections of rank 1 and rank 5, respectively. Hence,  $\mathbb{S} = \mathbb{C}^{-1}$  is given by

$$\mathbb{S} = \frac{1}{3k} \frac{1}{3} \mathbf{I} \otimes \mathbf{I} + \frac{1}{2\mu} (\mathbb{I} - \frac{1}{3} \mathbf{I} \otimes \mathbf{I}).$$

Therefore, and by Equation (7), one finds

$$\nu = \frac{1}{2} \left( \frac{3k - 2\mu}{3k + \mu} \right). \tag{13}$$

Similarly, Equation (9) furnishes

$$\widehat{\nu} = \frac{1}{2} \left( \frac{3k - 2\mu}{3k + \mu} \right),$$

and we see that *the Poisson's ratio and its areal counterpart reduce to the same elastic constant if the material is isotropic*. In passing, we mention that in view of Equation (12), positive definiteness is equivalent to

$$k > 0, \mu > 0, \tag{14}$$

and by Equation (13), furnish the well-known restriction on the Poisson's ratio for isotropic materials,

$$-1 < \nu < \frac{1}{2}.$$

Such bounds do not hold for the Poisson's ratio for the crystal classes, as demonstrated in [Ting and Chen 2005]. We examine the corresponding question for the areal Poisson's ratio in what follows.

**4.1. Cubic materials.** For a crystal of cubic symmetry, the Voigt compliance matrix takes the form

$$(s_{ij}) = \begin{pmatrix} s_{11} & s_{12} & s_{12} & 0 & 0 & 0 \\ & s_{11} & s_{12} & 0 & 0 & 0 \\ & & s_{11} & 0 & 0 & 0 \\ & & & s_{44} & 0 & 0 \\ & & & & s_{44} & 0 \\ & & & & & s_{44} \end{pmatrix}. \tag{15}$$

In terms of the Voigt compliances, positive definiteness is equivalent to (see [Nye 1957])

$$s_{11} > 0, s_{44} > 0, -\frac{1}{2}s_{11} < s_{12} < s_{11}. \tag{16}$$

The areal Poisson’s ratio can be expressed in spherical coordinates as:

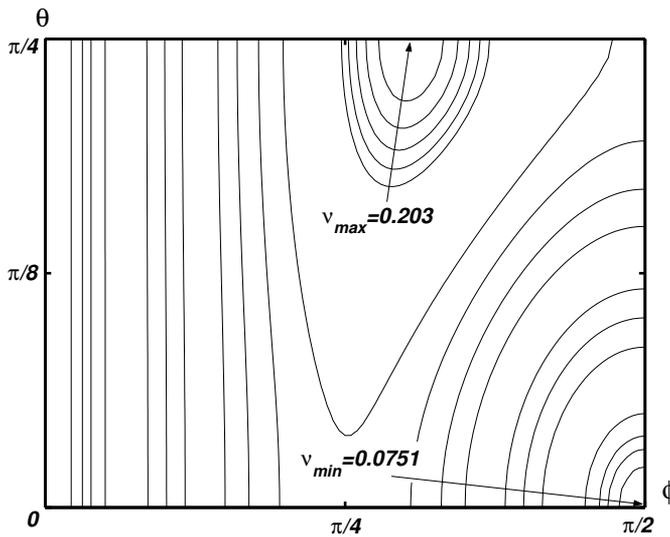
$$2\widehat{v}(\phi, \theta) = 1 - \frac{(S_{1111} + 2S_{1122})}{S_{1122} + 2S_{1212} + (S_{1111} - S_{1122} - 2S_{1212})[(\sin^4 \theta + \cos^4 \theta) \sin^4 \phi + \cos^4 \phi]}. \tag{17}$$

From this expression, we find

$$\widehat{v}(\phi, \theta) = \widehat{v}(\pi - \phi, \theta) = \widehat{v}\left(\phi, \frac{\pi}{2} + \theta\right) = \widehat{v}\left(\phi, \frac{\pi}{2} - \theta\right),$$

a manifestation of the symmetry associated with the class of crystals of cubic symmetry.

So we can limit the scope to  $0 \leq \phi \leq \pi/2, 0 \leq \theta \leq \pi/4$  without loss of generality. A contour plot of the areal Poisson’s ratio for the cubic material *pyrite* is shown in [Figure 1](#). At room temperature, the independent elastic compliance components are  $s_{11} = 2.652 \text{ (TPa)}^{-1}, s_{12} = -0.199 \text{ (TPa)}^{-1}, s_{44} = 9.141 \text{ (TPa)}^{-1}$  [Simmons and Wang 1971]. By [Equation \(17\)](#), the angles  $(\theta, \phi)$  for the stationary directions of cubic



**Figure 1.** Contours for the areal Poisson’s ratio for a pyrite of cubic symmetry.

materials obey

$$\begin{cases} 0 = \frac{2 \sin 2\phi [\sin^2 \phi (\sin^4 \theta + \cos^4 \theta) - \cos^2 \phi] (S_{1111} + 2S_{1122}) \beta}{\{S_{1122} + 2S_{1212} + \beta [(\sin^4 \theta + \cos^4 \theta) \sin^4 \phi + \cos^4 \phi]\}^2}, \\ 0 = \frac{-\sin^4 \phi \sin 4\theta (S_{1111} + 2S_{1122}) \beta}{\{S_{1122} + 2S_{1212} + \beta [(\sin^4 \theta + \cos^4 \theta) \sin^4 \phi + \cos^4 \phi]\}^2}. \end{cases} \quad (18)$$

The factor

$$\beta \stackrel{\text{def}}{=} S_{1111} - S_{1122} - 2S_{1212} \quad (19)$$

does not vanish unless the material is isotropic and the factor  $S_{1111} + 2S_{1122}$  is positive owing to positive definiteness. Hence, the stationary points of the areal Poisson's ratio are given by

$$\begin{cases} \phi = 0, \text{ and } \phi = \frac{\pi}{2}, \theta = 0, \\ \phi = \frac{\pi}{2}, \theta = \frac{\pi}{4}, \text{ and } \phi = \frac{\pi}{4}, \theta = 0, \\ \theta = \frac{\pi}{4}, (\cos \phi)^2 = \frac{1}{3}. \end{cases} \quad (20)$$

The stationary points in the first line of Equation (20) lie along the [100] direction, those in the second line lie along the [110] direction, and the last line describes stationary points along the [111] direction. The directions represented by these stationary points thus lie respectively on a four-fold axis, a two-fold axis and a three-fold axis of symmetry for cubic crystals. One important fact is that these directions do not depend upon the compliances. For a crystal of cubic symmetry, the directions of the extreme areal Poisson's ratio must coincide with the direction of a lattice vector, face diagonal, or body diagonal.

To determine the nature of a stationary point, whether it is a local extremum or a saddle point, we examine the value of the areal Poisson's ratio and its second derivatives at these points. For a stationary point lying along the [100] direction,

$$\begin{cases} \widehat{v} = -S_{1122}/S_{1111}, \\ \widehat{v}_{\phi\phi} = \widehat{v}_{\theta\theta} = -2(S_{1111} + 2S_{1122})\beta/S_{1111}^2, \\ \widehat{v}_{\phi\theta} = 0. \end{cases}$$

Assuming that the material is not isotropic, so that  $\beta \neq 0$ , the matrix  $J$  defined by Equation (11) assumes diagonal form. The eigenvalues,  $\widehat{v}_{\phi\phi}$  and  $\widehat{v}_{\theta\theta}$ , are positive or negative accordingly as  $\beta$  is negative or positive. In conclusion for the [100] direction, a stationary point is a minimum or maximum accordingly as  $\beta \leq 0$ . For a stationary point on the [110] direction,

$$\begin{cases} \widehat{v} = \frac{1}{2} \left[ 1 - \frac{2(S_{1111} + 2S_{1122})}{(S_{1111} + S_{1122} + 2S_{1212})} \right], \\ \widehat{v}_{\phi\phi} = -\frac{4(S_{1111} + 2S_{1122})\beta}{(S_{1111} + S_{1122} + 2S_{1212})^2}, \\ \widehat{v}_{\theta\theta} = \frac{8(S_{1111} + 2S_{1122})\beta}{(S_{1111} + S_{1122} + 2S_{1212})^2}, \\ \widehat{v}_{\phi\theta} = 0. \end{cases}$$

The matrix Equation (11) is again diagonal, but the eigenvalues are of opposite sign for  $\beta \neq 0$ , indicative of a saddle point. The stationary values of the areal Poisson's ratio associated with a face diagonal direction are neither a local minimum nor a local maximum. If a stationary point lies along the [111] direction,

$$\begin{cases} \widehat{\nu} = \frac{1}{2} \left[ 1 - \frac{3(S_{1111} + 2S_{1122})}{(S_{1111} + 2S_{1122} + 4S_{1212})} \right], \\ \widehat{\nu}_{\phi\phi} = \frac{12(S_{1111} + 2S_{1122})\beta}{(S_{1111} + 2S_{1122} + 4S_{1212})^2}, \\ \widehat{\nu}_{\theta\theta} = \frac{8(S_{1111} + 2S_{1122})\beta}{(S_{1111} + 2S_{1122} + 4S_{1212})^2}, \\ \widehat{\nu}_{\phi\theta} = 0. \end{cases}$$

Hence, the matrix Equation (11) is once again in diagonal form. Its eigenvalues are positive if  $\beta > 0$ , yielding a relative minimum, whereas a relative maximum is present if  $\beta < 0$ .

The global minimum and maximum are obtained by comparing the values of the areal Poisson's ratio at the stationary points. For the pyrite described earlier,  $\beta = -4.372$ . Thus the [111] direction locates the maximum value,  $\widehat{\nu}_{\max} = 0.203$ , whereas the [100] direction is associated with the minimum value,  $\widehat{\nu}_{\min} = 0.075$ . In Figure 1, the extreme points are easily identified by the closed contours.

In summary, for a cubic crystal that is *not* isotropic, we find

$$\beta > 0, \begin{cases} \widehat{\nu}_{\max} = -\frac{S_{1122}}{S_{1111}} < \frac{1}{2}, \text{ along } \mathbf{a}_1, \\ \widehat{\nu}_{\min} = \frac{1}{2} \left[ 1 - \frac{3(S_{1111} + 2S_{1122})}{S_{1111} + 2S_{1122} + 4S_{1212}} \right] > -1, \text{ along } \mathbf{q}, \end{cases}$$

$$\beta < 0, \begin{cases} \widehat{\nu}_{\max} = \frac{1}{2} \left[ 1 - \frac{3(S_{1111} + 2S_{1122})}{S_{1111} + 2S_{1122} + 4S_{1212}} \right] < \frac{1}{2}, \text{ along } \mathbf{q}, \\ \widehat{\nu}_{\min} = -\frac{S_{1122}}{S_{1111}} > -1, \text{ along } \mathbf{a}_1, \end{cases}$$

where

$$\mathbf{q} = \frac{1}{\sqrt{3}}(\mathbf{a}_1 + \mathbf{a}_2 + \mathbf{a}_3).$$

With the aid of these results, and by means of definiteness conditions (16), we conclude that for the case of cubic symmetry,

$$-1 < \widehat{\nu} < \frac{1}{2},$$

the same as for an isotropic medium. If  $s_{11} + 2s_{12} \rightarrow 0^+$ , both  $\widehat{\nu}_{\min}$  and  $\widehat{\nu}_{\max}$  approach the upper bound  $1/2$  and the material behaves like an isotropic medium. There are many cubic materials for which  $s_{12}/s_{11}$  is near  $-1/2$ , for example, gold ( $-0.462$ ),  $\gamma$  - Fe ( $-0.440$ ), lead ( $-0.459$ ),  $\text{Cu}_{2.7}\text{AlMn}_{0.3}$  ( $-0.475$ ). When the shear compliance  $s_{44}$  is much larger than  $s_{11}$  and  $s_{12}$ ,  $\beta$  is negative, and the areal Poisson's ratio assumes its maximum value ( $1/2$ ) along the  $\mathbf{q}$  direction. But if  $s_{44}$  is much smaller than  $s_{11}$  and  $|s_{12}|$ , then the areal Poisson's ratio can assume either the maximum or the minimum value along  $\mathbf{q}$ ,

depending on the relative values of  $s_{12}$  and  $s_{11}$ , and approach the limits  $-1$  for  $s_{11} \gg -2s_{12}$  or  $1/2$  for  $s_{11} \simeq -2s_{12}$ . Moreover, we see that positive  $s_{12}$  yields a negative areal ratio. Measured values of this constant are recorded for many cubic materials, including the ones just mentioned, in [Landolt and Bornstein 1992], and the scarcity of cubic materials possessing a positive  $s_{12}$  is readily apparent.

**4.2. Hexagonal crystal.** The Voigt compliance matrix for the hexagonal class assumes the form [Nye 1957]

$$(s_{ij}) = \begin{pmatrix} s_{11} & s_{12} & s_{13} & 0 & 0 & 0 \\ & s_{11} & s_{13} & 0 & 0 & 0 \\ & & s_{33} & 0 & 0 & 0 \\ & & & s_{44} & 0 & 0 \\ & & & & s_{44} & 0 \\ & & & & & 2(s_{11} - s_{12}) \end{pmatrix}$$

in an orientation frame  $\{a_1, a_2, a_3\}$  There are thus five independent elastic compliance constants for material possessing hexagonal symmetry. Positive definiteness is equivalent to [Nye 1957]

$$\begin{cases} s_{11} > 0, s_{33} > 0, s_{44} > 0, s_{11} + |s_{12}| > 0, \\ s_{33}s_{11} > s_{13}^2, s_{33}(s_{11} + s_{12}) > 2s_{13}^2. \end{cases} \tag{21}$$

In terms of spherical angles, the areal ratio takes the form

$$\begin{aligned} \widehat{\nu}(\phi, \theta) = & [(\cos 4\phi - 1)(S_{1111} - 4S_{1313} + S_{3333}) - 8(\sin \phi)^2 S_{1122} \\ & - 2(5 + 2\cos 2\phi + \cos 4\phi) S_{1133}] / \{16[(\sin \phi)^4 S_{1111} \\ & + (\cos \phi)^4 S_{3333} + 2(\sin \phi)^2 (\cos \phi)^2 (S_{1133} + 2S_{1313})\}, \end{aligned}$$

from which we conclude that the areal Poisson's ratio is independent of  $\theta$ . We further conclude that

$$\widehat{\nu}(\phi, \theta) = \widehat{\nu}(\phi) = \widehat{\nu}(\pi - \phi),$$

so we can limit the range of  $\phi$  to  $0 \leq \phi \leq \pi/2$ . A plot of the areal Poisson's ratio for hexagonal crystalline *graphite* is shown in Figure 2. The room temperature compliances are [Landolt and Bornstein 1992]:

$$\begin{aligned} s_{11} &= 0.98 \text{ (TPa)}^{-1}, & s_{12} &= -0.16 \text{ (TPa)}^{-1}, \\ s_{13} &= -0.33 \text{ (TPa)}^{-1}, & s_{33} &= 27.5 \text{ (TPa)}^{-1}, \\ & & s_{44} &= 250 \text{ (TPa)}^{-1} \end{aligned}$$

The stationary condition for a hexagonal crystal is

$$\begin{aligned} 0 = & \sin 2\phi \{16(\sin \phi)^4 S_{1111}^2 - 16(\cos \phi)^4 S_{3333}^2 - 2[16(\sin \phi)^4 S_{1122} \\ & - (6 + 24\cos 2\phi + 2\cos 4\phi) S_{1133}] (S_{1133} + 2S_{1313}) + 4(\cos \phi)^2 [(2\cos 2\phi - 6)(S_{1122} + S_{1133}) \\ & + 16(\cos \phi)^2 S_{1313}] S_{3333} + 4S_{1111} [4(\sin \phi)^4 S_{1122} + (10 + 6\cos 2\phi)(\sin \phi)^2 S_{1133} - 16(\sin \phi)^4 S_{1313} \\ & - 4\cos 2\phi S_{3333}]\} / \{32[(\sin \phi)^4 S_{1111} + (\cos \phi)^4 S_{3333} + 2(\sin \phi)^2 (\cos \phi)^2 (S_{1133} + 2S_{1313})]^2\}, \tag{22} \end{aligned}$$

from which we conclude that the stationary points of  $\widehat{\nu}$  are given by

$$\begin{cases} \phi = 0, \text{ and } \phi = \frac{\pi}{2}, \\ \phi_S = \arcsin\left(\sqrt{\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}}\right), \end{cases} \quad (23)$$

where the subscript  $S$  indicates that  $\phi$  depends upon the compliance constants, and  $a, b, c$  are given by

$$a = [S_{1111}^2 + S_{1111} S_{1122} - 3S_{1111} S_{1133} - 2S_{1122} S_{1133} + 2S_{1133}^2 + S_{1122} S_{3333} + S_{1133} S_{3333} - S_{3333}^2 + 4S_{1313} (-S_{1111} - S_{1122} + S_{1133} + S_{3333})],$$

$$b = 2(2S_{1111} S_{1133} - 4S_{1133}^2 - 8S_{1133} S_{1313} + S_{1111} S_{3333} - 4S_{1313} S_{3333} + S_{3333}^2),$$

$$c = (4S_{1133}^2 + 8S_{1133} S_{1313} - S_{1111} S_{3333} - S_{1122} S_{3333} - S_{1133} S_{3333} + 4S_{1313} S_{3333} - S_{3333}^2).$$

The stationary points  $\phi = 0$  and  $\phi = \pi/2$  are *invariant stationary points*. The direction represented by the stationary point  $\phi = 0$  coincides with the unique six-fold rotation symmetry axis, and the direction represented by stationary point  $\phi = \pi/2$  lies in the reflection symmetry plane along an axis of two-fold symmetry. The stationary point  $\phi = \phi_S$ , which depends on the elastic compliance constants, lies between  $\phi = 0$  and  $\phi = \pi/2$ . The three stationary points for the hexagonal material graphite, indicated in Figure 2, have the values  $\widehat{\nu}_{\max} = 0.433$  at  $\phi = \phi_S$ ,  $\widehat{\nu}_{\min} = 0.0121$  at  $\phi = 0$  and  $\widehat{\nu} = 0.254$  at  $\phi = \pi/2$ .

Consider  $\widehat{\nu}$  at the stationary points

$$\widehat{\nu}(0) = -S_{1133}/S_{3333}, \quad (24)$$

$$\widehat{\nu}\left(\frac{\pi}{2}\right) = -(S_{1122} + S_{1133}) / (2S_{1111}), \quad (25)$$

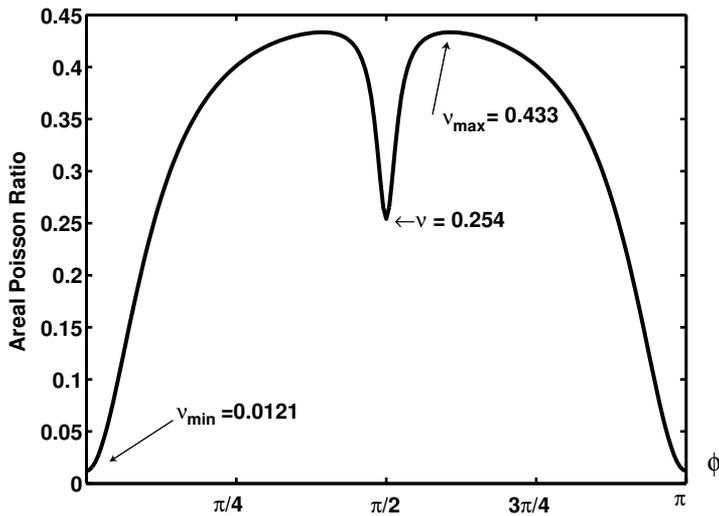


Figure 2. Areal Poisson’s ratio for graphite (hexagonal).

$$\widehat{\nu}(\phi_S) = \frac{-[\sin^4 \phi_S (S_{1111} - 2S_{1133} - 4S_{1313} + S_{3333}) + 2S_{1133} - \sin^2 \phi_S (S_{1111} - S_{1122} - S_{1133} - 4S_{1313} + S_{3333})]}{2[\sin^4 \phi_S (S_{1111} + S_{3333} - 2S_{1133} - 4S_{1313}) + 2 \sin^2 \phi_S (S_{1133} + 2S_{1313} - S_{3333}) + S_{3333}]}$$

Whether the areal Poisson's ratio is a local minimum or maximum at  $\phi = 0, \pi/2$ ,  $\phi_S$  depends on the elastic compliance constants. The global extreme values of the areal Poisson's ratio can be found by comparing the values at the stationary points. Without violating the positive definite conditions (21),  $S_{1122}, S_{1133}$  can be expressed in terms of  $S_{1111}, S_{3333}$ .

$$S_{1122} = pS_{1111}, \quad -1 < p < 1,$$

$$S_{1133} = q\sqrt{S_{1111}S_{3333}}, \quad -1 < q < 1.$$

The formulae (24) and (25) can be rewritten as

$$\begin{cases} \widehat{\nu}(0) = -q\sqrt{\frac{S_{1111}}{S_{3333}}}, \\ \widehat{\nu}\left(\frac{\pi}{2}\right) = -\frac{1}{2}\left(p + q\sqrt{\frac{S_{3333}}{S_{1111}}}\right). \end{cases}$$

The parameters  $p, q$  are bounded by  $\pm 1$ . The ratio  $\chi = S_{3333}/S_{1111}$  can take on any positive value without violating the positive definite conditions (21). For  $q < 0$ , the limits as  $\chi \rightarrow \pm\infty$  are

$$\begin{cases} \chi \rightarrow \infty, & \widehat{\nu}(0) \rightarrow 0^+, & \widehat{\nu}\left(\frac{\pi}{2}\right) \rightarrow \infty, \\ \chi \rightarrow 0, & \widehat{\nu}(0) \rightarrow \infty, & \widehat{\nu}\left(\frac{\pi}{2}\right) \rightarrow -\frac{1}{2}p. \end{cases}$$

and for  $q > 0$ ,

$$\begin{cases} \chi \rightarrow \infty, & \widehat{\nu}(0) \rightarrow 0^-, & \widehat{\nu}\left(\frac{\pi}{2}\right) \rightarrow -\infty, \\ \chi \rightarrow 0, & \widehat{\nu}(0) \rightarrow -\infty, & \widehat{\nu}\left(\frac{\pi}{2}\right) \rightarrow -\frac{1}{2}p. \end{cases}$$

This means that there is neither an upper bound nor a lower bound for the areal Poisson's ratio of a hexagonal crystal. As the case of  $s_{12}$  for cubic material,  $s_{13}$  is negative for all hexagonal materials in [Landolt and Bornstein 1992]. This interesting fact can be investigated in future research.

**4.3. Tetragonal (six constants).** For tetragonal crystal material with symmetry  $4mm, \bar{4}2, 422, 4/mmm$ , the Voigt compliance matrix  $s_{ij}$  takes the form

$$(s_{ij}) = \begin{pmatrix} s_{11} & s_{12} & s_{13} & 0 & 0 & 0 \\ & s_{11} & s_{13} & 0 & 0 & 0 \\ & & s_{33} & 0 & 0 & 0 \\ & & & s_{44} & 0 & 0 \\ & & & & s_{44} & 0 \\ & & & & & s_{66} \end{pmatrix}.$$

indicating six independent elastic compliance constants. Positive definiteness is equivalent to [Nye 1957]

$$\begin{cases} s_{11} > 0, & s_{33} > 0, & s_{44} > 0, & s_{66} > 0, \\ s_{11} > \pm s_{12}, & s_{33}s_{11} > s_{13}^2, & s_{33}(s_{11} + s_{12}) > 2s_{13}^2. \end{cases} \quad (26)$$

In terms of spherical angles, the areal Poisson's ratio takes the form

$$\widehat{\nu}(\phi, \theta) = \frac{1}{2} \frac{2 \sin^2 \phi (S_{1111} + S_{1122}) + (3 + \cos 2\phi) S_{1133} + 2 \cos^2 \phi S_{3333}}{\sin^4 \phi [(3 + \cos 4\theta) S_{1111} + 2 \sin^2 2\theta (S_{1122} + 2S_{1212})] + 2 \sin^2 2\phi (S_{1133} + 2S_{1313}) + 4 \cos^4 \phi S_{3333}}.$$

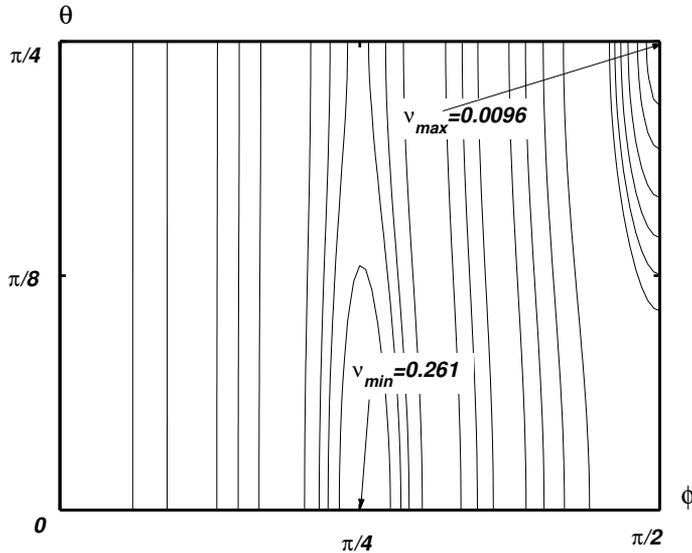
From this expression, we find that the restrictions imposed by the symmetry are

$$\widehat{\nu}(\phi, \theta) = \widehat{\nu}(\pi - \phi, \theta) = \widehat{\nu}\left(\phi, \frac{\pi}{2} + \theta\right) = \widehat{\nu}\left(\phi, \frac{\pi}{2} - \theta\right),$$

so we can limit the scope to  $0 \leq \phi \leq \frac{\pi}{2}$ ,  $0 \leq \theta \leq \frac{\pi}{4}$  without loss of generality. A contour plot for  $\alpha$ -cristobalite is shown in [Figure 3](#). The room temperature compliances are [[Yeganeh-Heari et al. 1992](#)]:

$$\begin{aligned} s_{11} &= 17.0 \text{ (TPa)}^{-1}, & s_{12} &= -0.965 \text{ (TPa)}^{-1}, \\ s_{13} &= 1.67 \text{ (TPa)}^{-1}, & s_{33} &= 23.9 \text{ (TPa)}^{-1}, \\ s_{44} &= 14.9 \text{ (TPa)}^{-1}, & s_{66} &= 38.9 \text{ (TPa)}^{-1}. \end{aligned}$$

From the stationary conditions (10), the invariant stationary points of the areal Poisson's ratio are



**Figure 3.** Areal Poisson's ratio for tetragonal material:  $\alpha$ -cristobalite.

$$\left\{ \begin{array}{l} \phi = 0, \text{ and } \phi = \frac{\pi}{2}, \theta = 0, \\ \phi = \frac{\pi}{2}, \theta = \frac{\pi}{4}, \\ \phi_{S1} = \arcsin\left(\sqrt{\frac{-b_1 \pm \sqrt{b_1^2 - 4a_1c_1}}{2a_1}}\right), \theta = 0, \\ \phi_{S2} = \arcsin\left(\sqrt{\frac{-b_2 \pm \sqrt{b_2^2 - 4a_2c_2}}{2a_2}}\right), \theta = \frac{\pi}{4}, \end{array} \right.$$

where

$$\begin{aligned} a_1 &= [S_{1111}^2 + S_{1111}S_{1122} - 3S_{1111}S_{1133} - 2S_{1122}S_{1133} + 2S_{1133}^2 + S_{1122}S_{3333} + S_{1133}S_{3333} - S_{3333}^2 \\ &\quad + 4S_{1313}(-S_{1111} - S_{1122} + S_{1133} + S_{3333})], \\ b_1 &= 2(2S_{1111}S_{1133} - 4S_{1133}^2 - 8S_{1133}S_{1313} + S_{1111}S_{3333} - 4S_{1313}S_{3333} + S_{3333}^2), \\ c_1 &= (4S_{1133}^2 + 8S_{1133}S_{1313} - S_{1111}S_{3333} - S_{1122}S_{3333} - S_{1133}S_{3333} + 4S_{1313}S_{3333} - S_{3333}^2), \end{aligned}$$

and

$$\begin{aligned} a_2 &= 2[S_{1111}^2 + 2S_{1111}S_{1122} + S_{1122}^2 - 5S_{1111}S_{1133} - 5S_{1122}S_{1133} + 4S_{1133}^2 + 2S_{1212}(S_{1111} + S_{1122} - S_{1133} - S_{3333}) \\ &\quad + 8S_{1313}(-S_{1111} - S_{1122} + S_{1133} + S_{3333}) + S_{3333}(S_{1111} + S_{1122} + 2S_{1133} - 2S_{3333})], \\ b_2 &= 4[2S_{1133}(S_{1111} + S_{1122} - 4S_{1133} + 2S_{1212} - 8S_{1313}) + S_{3333}(S_{1111} + S_{1122} + 2S_{1212} - 8S_{1313} + 2S_{3333})], \\ c_2 &= 4(4S_{1133}^2 + 8S_{1133}S_{1313} - S_{1111}S_{3333} - S_{1122}S_{3333} - S_{1133}S_{3333} + 4S_{1313}S_{3333} - S_{3333}^2). \end{aligned}$$

The stationary points  $(\phi, \theta) = (0, \theta), (\pi/2, 0), (\pi/2, \pi/4)$  are invariant stationary points. The directions represented by invariant stationary points are thus respectively on a unique four-fold axis, a two-fold axis and another two-fold axis of rotation symmetry for tetragonal crystals. The stationary points  $(\phi_{S1}, 0), (\phi_{S2}, \pi/4)$  depend on the elastic compliances. The global extreme values of the areal Poisson's ratio can be obtained by comparing the values at above stationary points. Consider the values of the areal Poisson's ratio at the stationary points:

$$\left\{ \begin{array}{l} \hat{\nu}(0, \theta) = -S_{1133}/S_{3333}, \\ \hat{\nu}(\frac{\pi}{2}, 0) = -(S_{1122} + S_{1133})/2S_{1111}, \\ \hat{\nu}(\frac{\pi}{2}, \frac{\pi}{4}) = (-\beta - 2S_{1122} - 2S_{1133})/2(S_{1111} + S_{1122} + 2S_{1212}); \end{array} \right.$$

$$\left\{ \begin{array}{l} \hat{\nu}(\phi_{S1}, 0) = \frac{1}{2} - [\sin^2 \phi_{S1} (S_{1111} + S_{1122} - S_{1133} - S_{3333}) \\ \quad + 2S_{1133} + S_{3333}]/\{2[\sin^4 \phi_{S1} (S_{1111} + S_{3333} - 2S_{1133} - 4S_{1313}) \\ \quad + 2 \sin^2 \phi_{S1} (S_{1133} + 2S_{1313} - S_{3333}) + S_{3333}]\}, \\ \hat{\nu}(\phi_{S2}, \frac{\pi}{4}) = \frac{1}{2} - [2 \sin^2 \phi_{S2} (S_{1111} + S_{1122} - S_{1133} - S_{3333}) \\ \quad + 4S_{1133} + 2S_{3333}]/\{\sin^4 \phi_{S2}[2S_{1111} + 2(S_{1122} + 2S_{1212}) - 8(S_{1133} + 2S_{1313}) + 4S_{3333}] \\ \quad + 8 \sin^2 \phi_{S2} (S_{1133} + 2S_{1313} - S_{3333}) + 4S_{3333}\}. \end{array} \right.$$

For the tetragonal material  $\alpha$ -cristobalite, the  $(\pi/2, \pi/4)$  direction locates the maximum value,  $\widehat{\nu}_{\max} = 9.6 \times 10^{-4}$  and the  $(\phi_{S1}, 0)$  direction is associated with the minimum value,  $\widehat{\nu}_{\min} = -0.261$ . In [Figure 3](#), the extreme points are easily identified by the closed contours. While many crystal materials can have a negative Poisson’s ratio in a particular direction,  $\alpha$ -cristobalite is one of the few materials that also yield a negative areal Poisson’s ratio.

Proceeding as we did in the hexagonal case, from positive definiteness [\(26\)](#), we obtain

$$\begin{cases} \widehat{\nu}(0) = -q\sqrt{\frac{S_{1111}}{S_{3333}}}, \\ \widehat{\nu}(\frac{\pi}{2}) = -\frac{1}{2}\left(p + q\sqrt{\frac{S_{3333}}{S_{1111}}}\right). \end{cases}$$

The ratio  $\chi = S_{3333}/S_{1111}$  can be any positive value. If we set  $q \neq 0$ , the areal Poisson’s ratio is not bounded for tetragonal crystal material with symmetry  $4mm, \bar{4}2m, 422, 4/mmm$  either.

**4.4. Tetragonal (seven constants).** For crystal material with tetragonal symmetry  $4, \bar{4}, 4/m$ , the Voigt compliance matrix  $s_{ij}$  takes the form

$$(s_{ij}) = \begin{pmatrix} s_{11} & s_{12} & s_{13} & 0 & 0 & s_{16} \\ & s_{11} & s_{13} & 0 & 0 & -s_{16} \\ & & s_{33} & 0 & 0 & 0 \\ & & & s_{44} & 0 & 0 \\ & & & & s_{44} & 0 \\ & & & & & s_{66} \end{pmatrix},$$

which shows six independent elastic compliance constants. In addition to the inequalities in [Equation \(26\)](#) the positive definite of strain energy requires

$$2s_{16}^2 - (s_{11} - s_{12})s_{66} > 0.$$

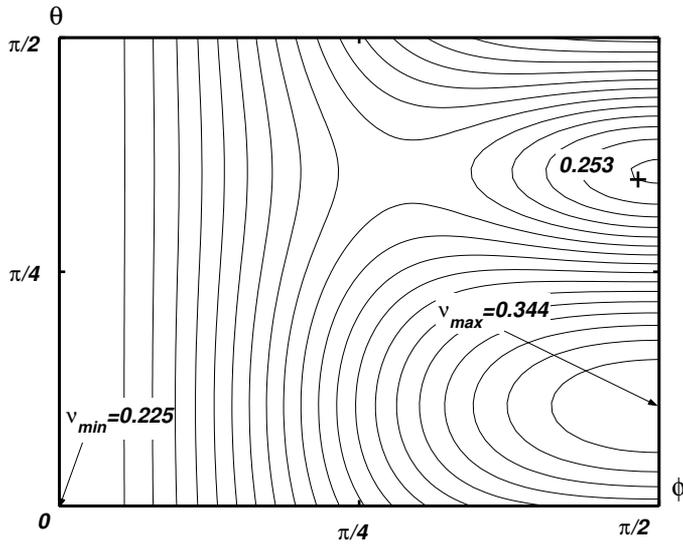
In terms of spherical angles, the areal Poisson’s ratio takes the form

$$\widehat{\nu}(\phi, \theta) = \frac{1}{2} \frac{2 \sin^2 \phi (S_{1111} + S_{1122}) + (3 + \cos 2\phi) S_{1133} + 2 \cos^2 \phi S_{3333}}{\sin^4 \phi [(3 + \cos 4\theta) S_{1111} + 4 \sin 4\theta S_{1112} + 2 \sin^2 2\theta (S_{1122} + 2S_{1212})] \sin^2 2\phi (S_{1133} + 2S_{1313}) + 4 \cos^4 \phi S_{3333}}.$$

From this expression, we can find that the relationships imposed by tetragonal  $4, \bar{4}, 4/m$  symmetry are

$$\widehat{\nu}(\phi, \theta) = \widehat{\nu}(\pi - \phi, \theta) = \widehat{\nu}\left(\phi, \frac{\pi}{2} + \theta\right),$$

so we can limit the scope to  $0 \leq \phi \leq \frac{\pi}{2}, 0 \leq \theta \leq \frac{\pi}{2}$  without loss of generality. Contours of the areal Poisson’s ratio are plotted for the tetragonal material calcium molybdate ( $+Z = -Z$ ) in [Figure 4](#). At



**Figure 4.** Areal Poisson ratio for calcium molybdate (+Z = - Z).

room temperature, the compliances are [Landolt and Bornstein 1992]:

$$\begin{aligned}
 s_{11} &= 9.90 \text{ (TPa)}^{-1}, & s_{12} &= -4.2 \text{ (TPa)}^{-1}, \\
 s_{13} &= -2.1 \text{ (TPa)}^{-1}, & s_{16} &= 4.2 \text{ (TPa)}^{-1}, \\
 s_{33} &= 9.48 \text{ (TPa)}^{-1}, & s_{44} &= 27.1 \text{ (TPa)}^{-1}, \\
 & & s_{66} &= 24.4 \text{ (TPa)}^{-1}.
 \end{aligned}$$

From the stationary conditions (10), the stationary points of the areal Poisson's ratio are

$$\begin{cases} \phi = 0, \\ \phi = \frac{\pi}{2}, \theta_{S1} = \frac{1}{4} \arctan(4S_{112}/\beta), \\ \phi = \frac{\pi}{2}, \theta_{S2} = \theta_{S1} + \frac{\pi}{4}. \end{cases}$$

The stationary point  $\phi = 0$  is the only invariant stationary point. The direction represented by this stationary point is the unique four-fold ( $C_4$ ) rotation symmetry axis. The stationary points  $(\frac{\pi}{2}, \theta_{S1}), (\frac{\pi}{2}, \theta_{S2})$  depend on the elastic compliance constants. The global extreme values of the areal Poisson's ratio can be obtained by comparing values at the stationary points. We find

$$\hat{v}(0, \theta) = -S_{1133}/S_{3333},$$

$$\widehat{v}\left(\frac{\pi}{2}, \theta_{S1}\right) = \frac{1}{2} - \frac{2(S_{1111} + S_{1122}) + 2S_{1133}}{3S_{1111} + S_{1122} + 2S_{1212} + 4\sin 4\theta_{S1}S_{1112} + \cos 4\theta_{S1}(S_{1111} - S_{1122} - 2S_{1212})},$$

$$\widehat{v}\left(\frac{\pi}{2}, \theta_{S2}\right) = \frac{1}{2} - \frac{2(S_{1111} + S_{1122}) + 2S_{1133}}{3S_{1111} + S_{1122} + 2S_{1212} - 4\sin 4\theta_{S2}S_{1112} - \cos 4\theta_{S2}(S_{1111} - S_{1122} - 2S_{1212})}.$$

For tetragonal material Calcium Molybdate (+Z = -Z), the  $(\frac{\pi}{2}, \theta_{S1})$  direction locates the maximum value,  $\widehat{v}_{\max} = 0.344$  and the  $(0, \theta)$  direction is associated with the minimum value,  $\widehat{v}_{\min} = 0.225$  as illustrated in Figure 4. The stationary point  $(\frac{\pi}{2}, \theta_{S2})$  is also a local extreme point since it is circumscribed by contours.

Similar to the hexagonal case, from the positive-definiteness conditions (26), we obtain

$$\widehat{v}(0) = -q\sqrt{\frac{S_{1111}}{S_{3333}}}.$$

The ratio  $\chi = S_{3333}/S_{1111}$  may assume any positive value. For  $q \neq 0$ , the areal Poisson’s ratio for tetragonal crystal material with symmetry  $4, \bar{4}, 4/m$ , like those before and those to follow, is unbounded.

**4.5. Trigonal crystal (six constants).** For crystal material with trigonal symmetry  $32, \bar{3}m, 3m$ , the Voigt compliance matrix  $s_{ij}$  appears as

$$(s_{ij}) = \begin{pmatrix} s_{11} & s_{12} & s_{13} & s_{14} & 0 & 0 \\ & s_{11} & s_{13} & -s_{14} & 0 & 0 \\ & & s_{33} & 0 & 0 & 0 \\ & & & s_{44} & 0 & 0 \\ & & & & s_{44} & s_{14} \\ & & & & & 2(s_{11} - s_{12}) \end{pmatrix},$$

indicating six independent elastic compliance constants. In addition to (21), positive definiteness requires

$$s_{44}s_{11} > s_{14}^2, (s_{11} - s_{12})s_{44} > 2s_{14}^2, (s_{11} - s_{12}) > \frac{s_{14}}{2}. \tag{27}$$

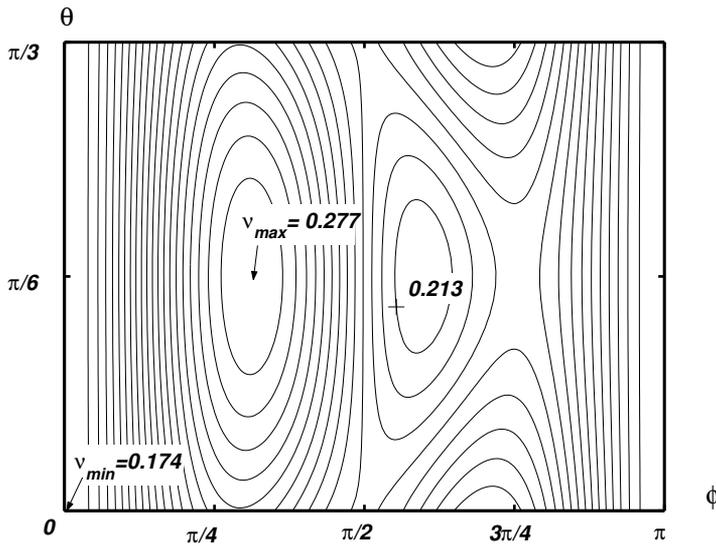
In terms of spherical angles, the areal ratio takes the form

$$\widehat{v}(\phi, \theta) = \frac{\frac{1}{2} - [2\sin^2\phi(S_{1111} + S_{1122}) + 2\cos^2\phi S_{3333} + (3 + \cos 2\phi)S_{1133}]}{\{2[2\sin^4\phi S_{1111} + 8\sin^3\phi \cos\phi \sin 3\theta S_{1123} + \sin^2 2\phi(S_{1133} + 2S_{1313}) + 2\cos^4\phi S_{3333}]\}}.$$

Hence, the relationships imposed by trigonal symmetry are

$$\widehat{v}(\phi, \theta) = \widehat{v}\left(\phi, \theta + \frac{2\pi}{3}\right) = \widehat{v}\left(\pi - \phi, \theta + \frac{\pi}{3}\right).$$

Thus, we may limit the ranges to  $0 \leq \phi \leq \pi, 0 \leq \theta \leq \frac{\pi}{3}$  without loss of generality. Contours of the areal Poisson’s ratio for the trigonal material aluminum oxide are plotted in Figure 5. The room temperature



**Figure 5.** Areal Poisson's ratio for trigonal material: aluminum oxide.

elastic compliance constants are [Landolt and Bornstein 1992]:

$$\begin{aligned}
 s_{11} &= 2.35 \text{ (TPa)}^{-1}, & s_{12} &= -0.69 \text{ (TPa)}^{-1}, \\
 s_{13} &= -0.38 \text{ (TPa)}^{-1}, & s_{14} &= 0.47 \text{ (TPa)}^{-1}, \\
 s_{33} &= 2.18 \text{ (TPa)}^{-1}, & s_{44} &= 7.0 \text{ (TPa)}^{-1}.
 \end{aligned}$$

By Equation (10), the stationary points of the areal Poisson's ratio are:

$$\begin{cases}
 \phi = 0, \\
 \phi = \frac{\pi}{2}, \theta = 0, \frac{\pi}{3}, \\
 \theta = \frac{\pi}{6}, \phi = \phi_{S1}, \phi_{S2},
 \end{cases}$$

where  $\phi_{S1}, \phi_{S2}$  satisfy the condition:

$$\begin{aligned}
 0 = & \{ (S_{1111} + S_{1122} - S_{1133} - S_{3333}) [2 \sin^4 \phi S_{1111} + 8 \cos \phi \sin^3 \phi S_{1123} \\
 & + \sin^2 2\phi (S_{1133} + 2S_{1313}) + 2 \cos^4 \phi S_{3333}] - 2 [2 \sin^2 \phi (S_{1111} + S_{1122}) \\
 & + (3 + \cos 2\phi) S_{1133} + 2 \cos^2 \phi S_{3333}] \times [\sin 4\phi (S_{1133} + 2S_{1313}) \\
 & + 2 \sin 2\phi \sin^2 \phi S_{1111} + \sin 3\phi S_{1123} - \cos^3 \phi S_{3333}] \}.
 \end{aligned}$$

The stationary points  $(\phi, \theta) = (0, \theta), (\pi/2, 0), (\pi/2, \pi/3)$  are invariant stationary points. The direction for the stationary point  $\phi = 0$  is the unique three-fold ( $C_3$ ) rotation symmetry axes, and the directions corresponding to the stationary points  $(\frac{\pi}{2}, 0), (\pi/2, \pi/3)$  are two-fold ( $C_2$ ) rotation symmetry axes, while the stationary points  $(\phi_{S1}, \pi/6), (\phi_{S2}, \pi/6)$  depend on the elastic compliance constants. The

global extrema may be analyzed by comparing the values at the stationary points. For  $\phi = 0$ ,

$$\begin{cases} \widehat{v}(0, \theta) = -\frac{S_{1133}}{S_{3333}}, \\ \widehat{v}_{\phi\phi}(0, \theta) = \frac{4S_{1133}(S_{1133}+2S_{1313})-S_{3333}(S_{1111}+S_{3333}+S_{1122}+S_{1133}-4S_{1313})}{S_{3333}^2}, \\ \widehat{v}_{\theta\theta}(0, \theta) = \widehat{v}_{\phi\theta}(0, \theta) = 0. \end{cases}$$

The values represent, at  $\phi = 0$ , a local minimum for  $\widehat{v}_{\phi\phi} > 0$ , and a local maximum for  $\widehat{v}_{\phi\phi} < 0$ . For  $(\phi, \theta) = (\pi/2, 0), (\pi/2, \pi/3)$ , we have

$$\begin{cases} \widehat{v}(\frac{\pi}{2}, 0) = \widehat{v}(\frac{\pi}{2}, \frac{\pi}{3}) = -\frac{S_{1122}+S_{1133}}{2S_{1111}}, \\ \widehat{v}_{\phi\phi}(\frac{\pi}{2}, 0) = \frac{2(S_{1122}+S_{1133})(S_{1133}+2S_{1313})-S_{1111}(S_{1111}+S_{3333}+S_{1122}+S_{1133}-4S_{1313})}{S_{1111}^2}, \\ \widehat{v}_{\phi\phi}(\frac{\pi}{2}, \frac{\pi}{3}) = \widehat{v}_{\phi\phi}(\frac{\pi}{2}, 0), \\ \widehat{v}_{\theta\theta}(\frac{\pi}{2}, 0) = \widehat{v}_{\theta\theta}(\frac{\pi}{2}, \frac{\pi}{3}) = 0, \\ \widehat{v}_{\phi\theta}(\frac{\pi}{2}, 0) = -6S_{1123}(S_{1111}+S_{1122}+S_{1133})/S_{1111}^2, \\ \widehat{v}_{\phi\theta}(\frac{\pi}{2}, \frac{\pi}{3}) = -\widehat{v}_{\phi\theta}(\frac{\pi}{2}, 0). \end{cases}$$

Hence, at  $(\pi/2, 0), (\pi/2, \pi/3)$ ,

$$\det(J) = -\frac{36S_{1123}^2(S_{1111}+S_{1122}+S_{1133})^2}{S_{1111}^4} < 0.$$

Thus the values of the areal Poisson’s ratio at  $(\pi/2, 0), (\pi/2, \pi/3)$  furnish neither a local minimum nor a local maximum. The global extreme values are achieved at  $\phi = 0$  and the material dependent stationary points  $(\phi_{S1}, \pi/6), (\phi_{S2}, \pi/6)$ . This is illustrated in [Figure 5](#) for the trigonal material aluminum oxide, where the  $(\phi_{S1}, \pi/6)$  direction locates the maximum value,  $\widehat{v}_{\max} = 0.277$  and the  $(0, \theta)$  direction is associated with the minimum value,  $\widehat{v}_{\min} = 0.174$ . The stationary point  $(\phi_{S2}, \pi/6)$  is also a local extreme point.

Without violating the definiteness conditions (27), we may write

$$\widehat{v}(0) = -q\sqrt{\frac{S_{1111}}{S_{3333}}}.$$

The ratio  $\chi = S_{3333}/S_{1111}$  is free to assume any positive value. If we take  $q \neq 0$ , we see that the areal Poisson’s ratio is not bounded for trigonal crystal material with symmetry  $32, \bar{3}m, 3m$ .

**4.6. Trigonal crystal (seven constants).** For a crystal with trigonal  $3, \bar{3}$  symmetry, the Voigt compliance matrix takes the form

$$(s_{ij}) = \begin{pmatrix} s_{11} & s_{12} & s_{13} & s_{14} & s_{15} & 0 \\ & s_{11} & s_{13} & -s_{14} & -s_{15} & 0 \\ & & s_{33} & 0 & 0 & 0 \\ & & & s_{44} & 0 & -s_{15} \\ & & & & s_{44} & s_{14} \\ & & & & & 2(s_{11} - s_{12}) \end{pmatrix},$$

which indicates the presence of seven independent elastic compliance constants.

In addition to the constraints in Equation (21), positive definiteness requires

$$\begin{cases} s_{44}s_{11} > s_{14}^2, & s_{44}s_{11} > s_{15}^2, & (s_{11} - s_{12})s_{44} > 2s_{14}^2, \\ (s_{11} - s_{12})s_{44} > \frac{s_{14}^2}{2}, & (s_{11} - s_{12})s_{44} > \frac{s_{15}^2}{2}. \end{cases} \quad (28)$$

In terms of spherical angles, the areal Poisson's ratio reads as

$$\hat{\nu}(\phi, \theta) = \frac{1}{2} \frac{2 \sin^2 \phi [(S_{1111} + S_{1122}) + (3 + \cos 2\phi) S_{1133} + 2 \cos^2 \phi S_{3333}]}{2[2 \sin^4 \phi S_{1111} + 8 \sin^3 \phi \cos \phi (\cos 3\theta S_{1113} + \sin 3\theta S_{1123}) + \sin^2 2\phi (S_{1133} + 2S_{1313}) + 2 \cos^4 \phi S_{3333}]}$$

From this expression, we see that the relationships imposed by trigonal symmetry are:

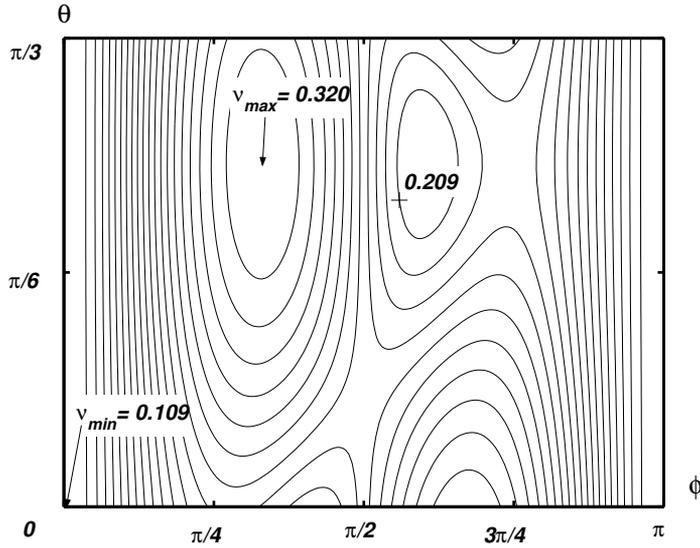
$$\hat{\nu}(\phi, \theta) = \hat{\nu}\left(\phi, \theta + \frac{2\pi}{3}\right) = \hat{\nu}\left(\pi - \phi, \theta + \frac{\pi}{3}\right).$$

Thus, we may limit the ranges to  $0 \leq \phi \leq \pi, 0 \leq \theta \leq \pi/3$  without loss of generality.

Contours for the trigonal material MgSiO<sub>3</sub> ilmenite are shown in Figure 6. At room temperature, the independent elastic compliance constants are reported to be [Weidner and Ito 1985]:

$$\begin{aligned} s_{11} &= 2.604 \text{ (TPa)}^{-1}, & s_{12} &= -0.976 \text{ (TPa)}^{-1}, \\ s_{13} &= -0.298 \text{ (TPa)}^{-1}, & s_{14} &= 0.911 \text{ (TPa)}^{-1}, \\ s_{15} &= -0.810 \text{ (TPa)}^{-1}, & s_{33} &= 2.727 \text{ (TPa)}^{-1}, \\ & & s_{44} &= 10.265 \text{ (TPa)}^{-1}. \end{aligned}$$

The first derivatives of the areal Poisson's ratio are:



**Figure 6.** Areal Poisson's ratio for MgSiO<sub>3</sub> ilmenite.

$$\begin{aligned} \widehat{v}_\phi = & -\{2 \sin \phi [2 \sin^2 \phi (S_{1111} + S_{1122}) + (3 + \cos 2\phi) S_{1133} + 2 \cos^2 \phi S_{3333}] \\ & \times [\cos \phi \sin^2 \phi S_{1111} - \cos^3 \phi S_{3333} + \sin 3\phi (\cos 3\theta S_{1113} + \sin 3\theta S_{1123}) \\ & + \cos \phi \cos 2\phi (S_{1133} + 2S_{1313})] - \sin 2\phi (S_{1111} + S_{1122} - S_{1133} - S_{3333}) \\ & \times [\sin^4 \phi S_{1111} + \cos^4 \phi S_{3333} + 4 \sin^3 \phi \cos \phi (\cos 3\theta S_{1113} + \sin 3\theta S_{1123}) \\ & + 2 \sin^2 \phi \cos^2 \phi (S_{1133} + 2S_{1313})]\} / \{2[\sin^4 \phi S_{1111} + \cos^4 \phi S_{3333} \\ & + 4 \sin^3 \phi \cos \phi (\cos 3\theta S_{1113} + \sin 3\theta S_{1123}) \\ & + 2 \sin^2 \phi \cos^2 \phi (S_{1133} + 2S_{1313})\}^2, \end{aligned}$$

$$\begin{aligned} \widehat{v}_\theta = & -\{3 \sin^3 \phi \cos \phi (\sin 3\theta S_{1113} - \cos 3\theta S_{1123}) [2 \sin^2 \phi (S_{1111} + S_{1122}) \\ & + (3 + \cos 2\phi) S_{1133} + 2 \cos^2 \phi S_{3333}]\} / [\sin^4 \phi S_{1111} \\ & + 4 \sin^3 \phi \cos \phi (\cos 3\theta S_{1113} + \sin 3\theta S_{1123}) \\ & + 2 \sin^2 \phi \cos^2 \phi (S_{1133} + 2S_{1313}) + \cos^4 \phi S_{3333}]^2. \end{aligned}$$

Thus, the stationary points are:

$$\begin{cases} \phi = 0, \\ \phi = \frac{\pi}{2}, \theta = \theta_{S1}, \\ \theta = \theta_{S2}, \phi = \phi_{S1}, \phi_{S2}, \end{cases}$$

where  $\theta_{S1}$  obeys  $\cos 3\theta S_{113} + \sin 3\theta S_{123} = 0$ ,  $\theta_{S2}$  satisfies  $\sin 3\theta S_{113} - \cos 3\theta S_{123} = 0$ ,  $\phi_{S1}$  is governed by

$$0 = [2 \sin^2 \phi (S_{1111} + S_{1122}) + (3 + \cos 2\phi) S_{1133} + 2 \cos^2 \phi S_{3333}] \\ \times [\cos \phi \sin^2 \phi S_{1111} + \sin 3\phi \sqrt{(S_{1113}^2 + S_{1123}^2)} + \cos \phi \cos 2\phi (S_{1133} + 2S_{1313}) - \cos^3 \phi S_{3333}] \\ - \cos \phi (S_{1111} + S_{1122} - S_{1133} - S_{3333}) \times [\sin^4 \phi S_{1111} + 4 \sin^3 \phi \cos \phi \sqrt{(S_{1113}^2 + S_{1123}^2)} \\ + 2 \sin^2 \phi \cos^2 \phi (S_{1133} + 2S_{1313}) + \cos^4 \phi S_{3333}],$$

and  $\phi_{S2}$  satisfies

$$0 = [2 \sin^2 \phi (S_{1111} + S_{1122}) + (3 + \cos 2\phi) S_{1133} + 2 \cos^2 \phi S_{3333}] \\ \times [\cos \phi \sin^2 \phi S_{1111} - \sin 3\phi \sqrt{(S_{1113}^2 + S_{1123}^2)} + \cos \phi \cos 2\phi (S_{1133} + 2S_{1313}) - \cos^3 \phi S_{3333}] \\ - \cos \phi (S_{1111} + S_{1122} - S_{1133} - S_{3333}) \times [\sin^4 \phi S_{1111} - 4 \sin^3 \phi \cos \phi \sqrt{(S_{1113}^2 + S_{1123}^2)} \\ + 2 \sin^2 \phi \cos^2 \phi (S_{1133} + 2S_{1313}) + \cos^4 \phi S_{3333}].$$

The values  $(\phi, \theta) = (0, \theta)$  furnish the only invariant stationary points. The direction represented by stationary point  $\phi = 0$  is the unique three-fold ( $C_3$ ) rotation symmetry axes, while  $(\pi/2, \theta_{S1})$ ,  $(\phi_{S1}, \theta_{S2})$ ,  $(\phi_{S2}, \theta_{S2})$  depend on the elastic compliance constants. The global extreme values of the areal Poisson's ratio are obtained by comparing the values at above stationary points. Thus

$$\begin{cases} \widehat{v}(0, \theta) = -\frac{S_{1133}}{S_{3333}}, \\ \widehat{v}_{\phi\phi}(0, \theta) = \frac{4S_{1133}(S_{1133} + 2S_{1313}) - S_{3333}(S_{1111} + S_{3333} + S_{1122} + S_{1133} - 4S_{1313})}{S_{3333}^2}, \\ \widehat{v}_{\theta\theta}(0, \theta) = \widehat{v}_{\phi\theta}(0, \theta) = 0. \end{cases}$$

The areal Poisson's ratio has a local minimum if  $\widehat{v}_{\phi\phi} > 0$ , a local maximum if  $\widehat{v}_{\phi\phi} < 0$ . Further,

$$\begin{cases} \widehat{v}(\frac{\pi}{2}, \theta_{S1}) = -\frac{S_{1122} + S_{1133}}{2S_{1111}}, \\ \widehat{v}_{\phi\phi}(\frac{\pi}{2}, \theta_{S1}) = \frac{2(S_{1122} + S_{1133})(S_{1133} + 2S_{1313}) - S_{1111}(S_{1111} + S_{3333} + S_{1122} + S_{1133} - 4S_{1313})}{S_{1111}^2}, \\ \widehat{v}_{\theta\theta}(\frac{\pi}{2}, \theta_{S1}) = 0, \\ \widehat{v}_{\phi\theta}(\frac{\pi}{2}, \theta_{S1}) = -\frac{6\sqrt{S_{1123}^2 + S_{1113}^2}(S_{1111} + S_{1122} + S_{1133})}{S_{1111}^2}. \end{cases}$$

This yields

$$\det(J) = -\frac{36(S_{1123}^2 + S_{1113}^2)(S_{1111} + S_{1122} + S_{1133})^2}{S_{1111}^4} < 0.$$

Thus the value of the areal Poisson's ratio at the invariant stationary point  $(\pi/2, \theta_{S1})$  is neither a local minimum nor a local maximum. The global extreme values are achieved at  $\phi = 0$  and the material dependent stationary points. These conclusions are demonstrated in [Figure 6](#) for trigonal material  $MgSiO_3$

ilmenite, where the  $(\phi_{S1}, \theta_{S1})$  direction locates the maximum value  $\widehat{v}_{\max} = 0.320$  and the  $(0, \theta)$  direction is associated with the minimum value,  $\widehat{v}_{\min} = 0.109$ .

Without violating (28), we may write

$$\widehat{v}(0) = -q \sqrt{\frac{S_{1111}}{S_{3333}}}.$$

The ratio  $\chi = S_{3333} / S_{1111}$  is free to assume arbitrary positive values. For  $q \neq 0$ , positive definiteness fails to impose bounds on  $\widehat{v}$  for trigonal crystals with symmetry  $3, \bar{3}$ .

**4.7. Orthorhombic.** For an orthorhombic crystal, the Voigt compliance matrix  $s_{ij}$  takes the form

$$(s_{ij}) = \begin{pmatrix} s_{11} & s_{12} & s_{13} & 0 & 0 & 0 \\ & s_{22} & s_{23} & 0 & 0 & 0 \\ & & s_{33} & 0 & 0 & 0 \\ & & & s_{44} & 0 & 0 \\ & & & & s_{55} & 0 \\ & & & & & s_{66} \end{pmatrix}.$$

There are nine independent elastic compliance constants. Positive definiteness imposes the requirements

$$\begin{cases} s_{11} > 0, & s_{22} > 0, & s_{33} > 0, & s_{44} > 0, & s_{55} > 0, & s_{66} > 0, \\ s_{11}s_{22} > s_{12}^2, & s_{33}s_{11} > s_{13}^2, & s_{33}s_{22} > s_{23}^2, \\ s_{11}(s_{33}s_{22} - s_{23}^2) - s_{12}^2s_{33} + 2s_{12}s_{13}s_{23} - s_{13}^2s_{22} > 0. \end{cases} \tag{29}$$

The areal Poisson’s ratio can be expressed in spherical coordinates as

$$\begin{aligned} \widehat{v}(\phi, \theta) = & \frac{1}{2} - \frac{1}{2} \left\{ \sin^2 \phi [S_{1122} + \cos^2 \theta (S_{1111} + S_{1133}) + \sin^2 \theta (S_{2222} + S_{2233})] \right. \\ & + \cos^2 \phi (S_{1133} + S_{2233} + S_{3333}) \left. \right\} / \left\{ \sin^4 \phi (\cos^4 \theta S_{1111} + \sin^4 \theta S_{2222} + 2 \sin^2 \theta \cos^2 \theta S_{1122}) \right. \\ & + \cos^4 \phi S_{3333} + \sin^2 \phi [2 \cos^2 \phi \cos^2 \theta S_{1133} + 4 \sin^2 \theta \cos^2 \theta \sin^2 \phi S_{1212} \\ & \left. + 2 \cos^2 \phi (2 \cos^2 \theta S_{1313} + \sin^2 \theta (S_{2233} + 2S_{2323})) \right\}. \end{aligned}$$

From this expression, we find that orthorhombic symmetry requires

$$\widehat{v}(\phi, \theta) = \widehat{v}(\pi - \phi, \theta) = \widehat{v}(\phi, \pi + \theta) = \widehat{v}(\phi, \pi - \theta).$$

Therefore, we may limit the ranges to  $0 \leq \phi \leq \pi/2, 0 \leq \theta \leq \pi/2$  without loss of generality. Contours for the orthorhombic material acenaphthene are shown in Figure 7. The elastic compliance matrix (at room

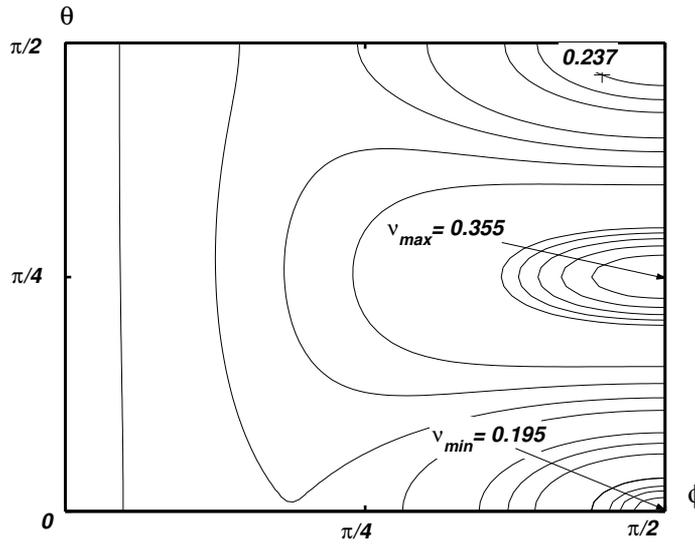


Figure 7. Areal Poisson's ratio for acenaphthene.

temperature), expressed in term of  $(\text{TPa})^{-1}$  is [Simmons and Wang 1971]

$$(s_{ij}) = \begin{pmatrix} 81.438 & -3.125 & -28.605 & 0 & 0 & 0 \\ & 93.354 & -37.298 & 0 & 0 & 0 \\ & & 115.385 & 0 & 0 & 0 \\ & & & 377.358 & 0 & 0 \\ & & & & 344.828 & 0 \\ & & & & & 540.540 \end{pmatrix}.$$

By (10), the stationary points of the areal Poisson's ratio are

$$\begin{cases} \phi = 0, \\ \phi = \frac{\pi}{2}, \theta = 0, \frac{\pi}{2}, \\ \phi = \frac{\pi}{2}, \theta = \theta_S, \end{cases}$$

where  $\theta_S$  satisfies the condition

$$0 = \sin 2\theta [\cos^4 \theta S_{1111} + 2 \sin^2 \theta \cos^2 \theta (S_{1122} + 2S_{1212}) + \sin^4 \theta S_{2222}] \times (S_{1111} + S_{1133} - S_{2222} - S_{2233}) \\ - \sin 2\theta [2 \cos^2 \theta S_{1111} - 2 \cos 2\theta (S_{1122} + 2S_{1212}) - 2 \sin^2 \theta S_{2222}] \\ \times [S_{1122} + \cos^2 \theta (S_{1111} + S_{1133}) + \sin^2 \theta (S_{2222} + S_{2233})].$$

The stationary points  $(\phi, \theta) = (0, \theta), (\pi/2, 0), (\pi/2, \pi/2)$  are invariant with respect to the elastic constants. The directions represented by these stationary points are the two-fold ( $C_2$ ) rotation symmetry axes. The stationary point  $(\pi/2, \theta_S)$  depends on the elastic compliance constants. The global extreme

values of areal Poisson’s ratio can be obtained by comparing the values at above stationary points,

$$\begin{cases} \widehat{\nu}(0, \theta) = -\frac{S_{1133} + S_{2233}}{2S_{3333}}, \\ \widehat{\nu}\left(\frac{\pi}{2}, 0\right) = -\frac{S_{1122} + S_{1133}}{2S_{1111}}, \\ \widehat{\nu}\left(\frac{\pi}{2}, \frac{\pi}{2}\right) = -\frac{S_{1122} + S_{2233}}{2S_{2222}}, \end{cases} \tag{30}$$

$$\widehat{\nu}\left(\frac{\pi}{2}, \theta_S\right) = \frac{1}{2} - \frac{1}{2} \frac{S_{1122} + \cos^2 \theta_S (S_{1111} + S_{1133}) + \sin^2 \theta_S (S_{2222} + S_{2233})}{[\cos^4 \theta_S S_{1111} + 2 \sin^2 \theta_S \cos^2 \theta_S (S_{1122} + 2S_{1212}) + \sin^4 \theta_S S_{2222}]}.$$

For the orthorhombic material acenaphthene, the  $(\pi/2, \theta_S)$  direction locates the maximum value  $\widehat{\nu}_{\max} = 0.355$  and the  $(\pi/2, 0)$  direction is associated with the minimum value,  $\widehat{\nu}_{\min} = 0.195$  as indicated in Figure 7. From the definiteness conditions Equation (29),  $S_{1122}, S_{1133}, S_{2233}$  can be expressed in terms of  $S_{1111}, S_{2222}, S_{3333}$ :

$$\begin{aligned} S_{1122} &= r\sqrt{S_{1111}S_{2222}}, & -1 < r < 1, \\ S_{1133} &= q\sqrt{S_{1111}S_{3333}}, & -1 < q < 1, \\ S_{2233} &= w\sqrt{S_{2222}S_{3333}}, & -1 < w < 1. \end{aligned}$$

The expressions in Equation (30) give way to

$$\begin{cases} \widehat{\nu}(0, \theta) = -\frac{1}{2} \left( q\sqrt{\frac{S_{1111}}{S_{3333}}} + w\sqrt{\frac{S_{2222}}{S_{3333}}} \right), \\ \widehat{\nu}\left(\frac{\pi}{2}, 0\right) = -\frac{1}{2} \left( r\sqrt{\frac{S_{2222}}{S_{1111}}} + q\sqrt{\frac{S_{3333}}{S_{1111}}} \right), \\ \widehat{\nu}\left(\frac{\pi}{2}, \frac{\pi}{2}\right) = -\frac{1}{2} \left( r\sqrt{\frac{S_{1111}}{S_{2222}}} + w\sqrt{\frac{S_{3333}}{S_{2222}}} \right). \end{cases}$$

Consider  $\widehat{\nu}(0, \theta)$ . Since the ratios  $S_{3333}/S_{1111}$  and  $S_{3333}/S_{2222}$  may take on any positive value without violating the definiteness conditions (29), it is not bounded. A similar argument can be made for  $\widehat{\nu}(\pi/2, 0)$  and  $\widehat{\nu}(\pi/2, \pi/2)$ . Thus the areal Poisson’s ratio is not bounded for an orthorhombic crystal.

**4.8. Monoclinic.** The Voigt compliance matrix  $s_{ij}$  for monoclinic materials takes the form

$$(s_{ij}) = \begin{pmatrix} s_{11} & s_{12} & s_{13} & 0 & 0 & s_{16} \\ & s_{22} & s_{23} & 0 & 0 & s_{26} \\ & & s_{33} & 0 & 0 & s_{36} \\ & & & s_{44} & s_{45} & 0 \\ & & & & s_{55} & 0 \\ & & & & & s_{66} \end{pmatrix},$$

involving thirteen independent elastic constants. In addition to the conditions in Equation (29), definiteness of the strain energy requires

$$\begin{cases} s_{44}s_{55} > s_{45}^2, & s_{33}s_{66} > s_{36}^2, \\ s_{22}s_{66} > s_{26}^2, & s_{11}s_{66} > s_{16}^2. \end{cases} \tag{31}$$

The areal Poisson's ratio can be expressed in spherical coordinates as

$$\begin{aligned} \widehat{\nu}(\phi, \theta) = & \frac{1}{2} - \frac{1}{2} \left\{ \sin^2 \phi [S_{1122} + \cos^2 \theta (S_{1111} + S_{1133}) + \sin^2 \theta (S_{2222} + S_{2233})] \right. \\ & + \cos^2 \phi (S_{1133} + S_{2233} + S_{3333}) \left. \right\} / \left\{ \sin^4 \phi (\cos^4 \theta S_{1111} + \sin^4 \theta S_{2222} + 2 \sin^2 \theta \cos^2 \theta S_{1122}) \right. \\ & + \cos^4 \phi S_{3333} + \sin^2 \phi [2 \cos^2 \phi \cos^2 \theta S_{1133} + 4 \sin^2 \theta \cos^2 \theta \sin^2 \phi S_{1212} \\ & \left. + 2 \cos^2 \phi (2 \cos^2 \theta S_{1313} + \sin^2 \theta S_{2233} + 2 \sin^2 \theta S_{2323}) \right\}. \end{aligned}$$

From this expression, we see that the relationships imposed by monoclinic symmetry are

$$\widehat{\nu}(\phi, \theta) = \widehat{\nu}(\pi - \phi, \theta) = \widehat{\nu}(\phi, \pi + \theta).$$

Accordingly, we may limit the ranges of the spherical angles to  $0 \leq \phi \leq \frac{\pi}{2}$ ,  $0 \leq \theta \leq \pi$ . A contour plot of the areal Poisson's ratio for the monoclinic material feldspar (plagioclase—29 AN) is shown in Figure 8. The room temperature elastic compliance matrix (TPa)<sup>-1</sup> is [Simmons and Wang 1971]:

$$(s_{ij}) = \begin{pmatrix} 15.460 & -3.403 & -3.739 & 0 & 0 & 1.333 \\ & 7.786 & -0.852 & 0 & 0 & 0.266 \\ & & 9.526 & 0 & 0 & 4.390 \\ & & & 54.157 & 1.737 & 0 \\ & & & & 29.210 & 0 \\ & & & & & 34.861 \end{pmatrix}$$

By Equation (10), the stationary points are

$$\begin{cases} \phi = 0, \\ \phi = \frac{\pi}{2}, \theta = \theta_{S1}, \\ \phi = \phi_S, \theta = \theta_{S2}, \end{cases}$$

where  $\theta_{S1}$  satisfy

$$\begin{aligned} 0 = & - \left\{ \cos^4 \theta S_{1111} + \sin \theta [4 \cos^3 \theta S_{1112} + \sin^3 \theta S_{2222} + \sin 2\theta (\cos \theta (S_{1122} + 2S_{1212}) + 2 \sin 2\theta S_{2212})] \right\} \\ & \times \left[ \sin 2\theta (-S_{1111} - S_{1133} + S_{2222} + S_{2233}) + 2 \cos 2\theta (S_{1112} + S_{2212} + S_{3312}) \right] + \frac{1}{2} \left[ 4 (\cos 2\theta + \cos 4\theta) S_{1112} \right. \\ & \left. + 2 \sin 4\theta (S_{1122} + 2S_{1212}) + 8 \sin \theta (-\cos^3 \theta S_{1111} + \sin 3\theta S_{2212} + \cos \theta \sin^2 \theta S_{2222}) \right] \\ & \times \left\{ \cos^2 \theta S_{1111} + \sin 2\theta S_{1112} + S_{1122} + \cos^2 \theta S_{1133} + \sin \theta \left[ \sin \theta (S_{2222} + S_{2233}) + 2 \cos 2\theta (S_{2212} + S_{3312}) \right] \right\}. \end{aligned}$$

The point  $\phi = 0$  is the only invariant stationary point. The direction represented by this stationary point is the two-fold ( $C_2$ ) rotation symmetry axis. The stationary points  $(\pi/2, \theta_{S1})$ ,  $(\phi_S, \theta_{S2})$  depend on the

elastic compliance constants. The global extreme values of the areal Poisson’s ratio may be obtained by comparing the values at the above stationary points.

Let us consider the values of the areal Poisson’s ratio at stationary points

$$\begin{cases} \widehat{\nu}(0, \theta) = -\frac{S_{1133} + S_{2233}}{2S_{3333}}, \\ \widehat{\nu}\left(\frac{\pi}{2}, \theta\right) = \frac{1}{2} - \frac{1}{2} \frac{\cos^2 \theta (S_{1111} + S_{1133}) + \sin^2 \theta (S_{2222} + S_{2233}) + S_{1122} + 2 \sin \theta \cos \theta (S_{1112} + S_{2212} + S_{3312})}{\cos^4 \theta S_{1111} + \sin^4 \theta S_{2222} + 2 \sin^2 \theta \cos^2 \theta (S_{1122} + 2S_{1212}) + 4 \sin \theta \cos^3 \theta S_{1112} + 4 \sin^3 \theta \cos \theta S_{2212}}. \end{cases}$$

For the monoclinic material feldspar (plagioclase — 29 AN), the  $(\pi/2, \theta_{S1})$  direction locates the maximum value,  $\widehat{\nu}_{max} = 0.399$  and the  $(\phi_S, \theta_{S2})$  direction is associated with the minimum value,  $\widehat{\nu}_{min} = 0.168$  in Figure 8.

Similar to the orthorhombic case, we obtain

$$\widehat{\nu}(0, \theta) = -\frac{1}{2} \left( q \sqrt{\frac{S_{1111}}{S_{3333}}} + w \sqrt{\frac{S_{2222}}{S_{3333}}} \right).$$

Since the ratios  $S_{3333}/S_{1111}$  and  $S_{3333}/S_{2222}$  may be arbitrarily small or large while remaining positive, the areal Poisson’s ratio is thus not bounded for monoclinic crystal materials.

**4.9. Triclinic.** The Voigt compliance matrix  $s_{ij}$  is shown in Equation (4). There are twenty one independent elastic constants. The elastic compliance matrix must obey the positive definiteness conditions (29) and (31).

The areal Poisson’s ratio are restricted only by inversion center symmetry

$$\widehat{\nu}(\phi, \theta) = \widehat{\nu}(\pi - \phi, \pi + \theta),$$

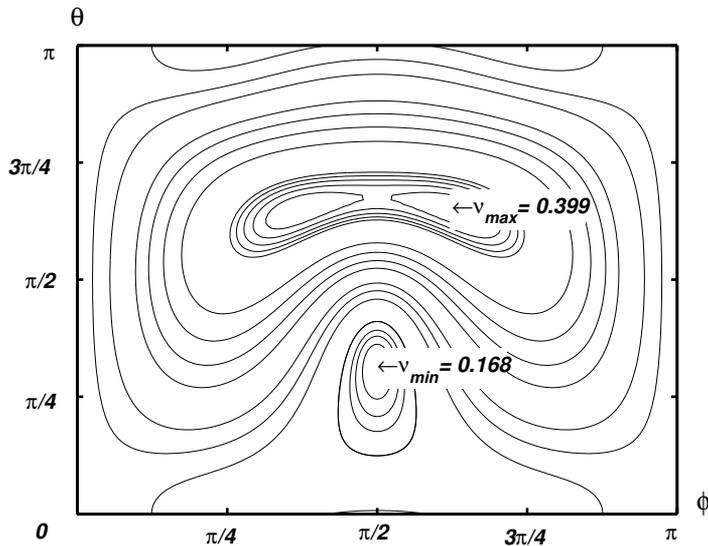


Figure 8. Areal Poisson’s ratio for monoclinic material: feldspar.

so we can limit the scope to  $0 \leq \phi \leq \pi$ ,  $0 \leq \theta \leq \pi$  without loss of generality. Contours are shown for the triclinic material copper sulfate pentahydrate in Figure 9. At room temperature, the elastic compliance matrix in term of  $(\text{TPa})^{-1}$  is [Krishnan et al. 1971]

$$(s_{ij}) = \begin{pmatrix} 28.61 & -9.67 & -9.77 & 2.39 & 0.45 & 9.83 \\ & 49.26 & -25.21 & -6.24 & 2.26 & -8.01 \\ & & 39.16 & 6.92 & 1.94 & 3.26 \\ & & & 60.0 & -4.32 & -0.76 \\ & & & & 88.04 & 23.46 \\ & & & & & 110.64 \end{pmatrix}.$$

By investigating the stationary conditions (10), we find no invariant stationary point for triclinic materials. All stationary points depend on the elastic compliance constants. For the triclinic material copper sulfate pentahydrate, the maximum value,  $\hat{v}_{\max} = 0.456$  and minimum value  $\hat{v}_{\min} = 0.250$  are shown in Figure 9.

Let's look at the values of the areal Poisson's ratio at  $\phi = 0$ .

$$\hat{v}(0, \theta) = -(S_{1133} + S_{2233}) / 2S_{3333}$$

Similar to the orthorhombic case, we obtain

$$\hat{v}(0, \theta) = -\frac{1}{2} \left( q \sqrt{\frac{S_{1111}}{S_{3333}}} + w \sqrt{\frac{S_{2222}}{S_{3333}}} \right).$$

Since the ratios  $S_{3333} / S_{1111}$  and  $S_{3333} / S_{2222}$  can be arbitrary small or large positive value, the areal Poisson's ratio is not bounded for triclinic crystal material.

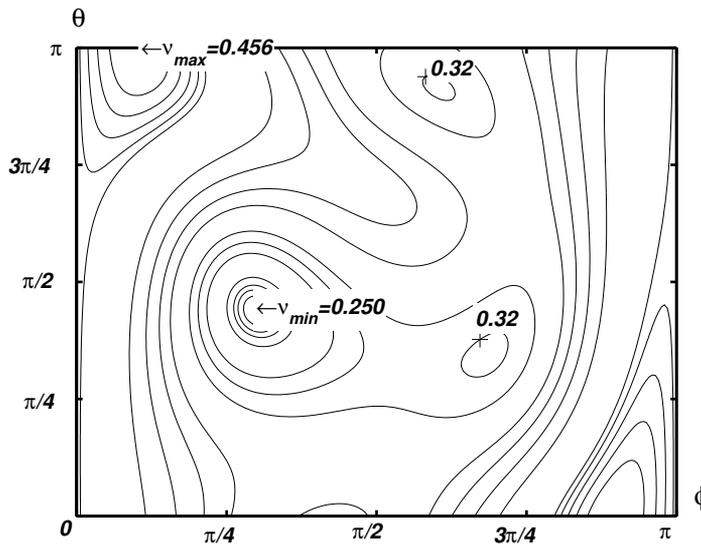


Figure 9. Areal Poisson's ratio for copper sulphate pentahydrate.

Symmetry Class	1	2	3
Cubic	$(0, \theta), (\pi/2, 0)$	$(\pi/2, \pi/4), (\pi/4, 0)$	$(\arctan \sqrt{3}/3, \pi/4)$
Hexagonal	$(0, \theta)$	$(\pi/2, \theta)$	none
Tetragonal ( $4mm, \bar{4}2m, 422, 4/mmm$ )	$(0, \theta)$	$(\pi/2, 0)$	$(\pi/2, \pi/4)$
Tetragonal ( $4, \bar{4}, 4/m$ )	$(0, \theta)$	none	none
Trigonal ( $32, \bar{3}m, 3m$ )	$(0, \theta)$	$(\pi/2, 0)$	$(\pi/2, \pi/3)$
Trigonal ( $3, \bar{3}$ )	$(0, \theta)$	none	none
Orthorhombic	$(0, \theta)$	$(\pi/2, 0)$	$(\pi/2, \pi/2)$
Monoclinic	$(0, \theta)$	none	none
Triclinic	none	none	none

**Table 1.** Invariant stationary points  $(\phi, \theta)$  of all crystal symmetry classes.

## 5. Summary

We determine the stationary points of the areal Poisson's ratio for all crystal classes, and illustrate them graphically. The directions of *invariant* stationary points are related directly to the symmetry of the crystal class, but do not depend upon the elastic constants of the particular material at hand. The invariant stationary directions are summarized in [Table 1](#), apart from points that are trivially related to these by symmetry.

For crystals of low symmetry, at least one of the global extreme values occurs on the direction of an invariant stationary point. To find the remaining global extreme, we have to consider both invariant and material dependent stationary points. It is also shown that the areal Poisson's ratio for cubic crystal is bounded between  $-1$  and  $1/2$ , just as the case for isotropic material. But the areal Poisson's ratio the remaining eight lower symmetry crystal classes can have arbitrarily large positive or negative values without violating the positive definiteness of strain energy density.

## References

- [Baughman et al. 1998] R. H. Baughman, J. M. Shacklette, A. A. Zakhidov, and S. Stafstrom, "Negative Poisson's ratios as a common feature of cubic metals", *Nature* **392** (1998), 362–365.
- [Cazzani and Rovati 2003] A. Cazzani and M. Rovati, "Extrema of Young's modulus for cubic and transversely isotropic solids", *Int. J. Solids Struct.* **40** (2003), 1713–1744.
- [Cazzani and Rovati 2005] A. Cazzani and M. Rovati, "Extrema of Young's modulus for elastic solids with tetragonal symmetry", *Int. J. Solids Struct.* **42** (2005), 5057–5096.
- [Evans et al. 1991] K. Evans, M. Nkanasah, I. Hutchinson, and S. Rogers, "Molecular network design", *Nature* **353** (1991), 124.
- [Guo and Wheeler 2006] C. Y. Guo and L. T. Wheeler, "Extreme Poisson's ratios and related elastic crystal properties", *J. Mech. Phys. Solids* **54** (2006), 690–707.
- [Gurtin 1972] M. E. Gurtin, "The linear theory of elasticity", pp. 1–296 in *Handbuch der physik*, vol. Volume VIa/2, Springer-Verlag, 1972.

- [Krishnan et al. 1971] R. S. Krishnan, V. Radha, and E. S. R. Gopal, "Elastic constants of triclinic copper sulphate pentahydrate crystals", *J. Phys. D: Appl. Phys.* **4** (1971), 171–173.
- [Landolt and Bornstein 1992] H. H. Landolt and R. Bornstein, *Numerical data and functional relationships in science and technology*, vol. III/29/a, Second and higher order elastic constants, Springer-Verlag, Berlin, 1992.
- [Love 1927] A. E. H. Love, "A treatise on the mathematical theory of elasticity", Dover Reprint of the Cambridge University Press, 1927. 4th ed.
- [Nye 1957] J. F. Nye, *Physical properties of crystals*, Oxford University Press, 1957. reprinted 2000.
- [Simmons and Wang 1971] G. Simmons and H. Wang, *Single crystal elastic constants and calculated aggregate properties: a handbook*, 2nd edition ed., MIT Press, 1971.
- [Ting 1996] T. C. T. Ting, *Anisotropic elasticity: theory and application*, Oxford University Press, New York, 1996.
- [Ting and Barnett 2005] T. C. T. Ting and D. M. Barnett, "Negative Poisson's ratios in anisotropic linear elastic media", *J. Appl. Mech.* **72** (2005), 929–931.
- [Ting and Chen 2005] T. C. T. Ting and T. Chen, "Poisson's ratio for anisotropic materials can have no bounds", *Quart. J. Mech. Appl. Math.* **58**:1 (2005), 73–82.
- [Tokmakova 2005] S. Tokmakova, "Stereographic projections of Poisson's ratio in auxetic crystals", *phys stat. sol. (b)* **242**:3 (2005), 721–729.
- [Weidner and Ito 1985] D. J. Weidner and E. Ito, "Elasticity of MgSiO in the ilmenite phase", *Phys. Earth Planet. Int.* **40** (1985), 65–70.
- [Yeganeh-Heari et al. 1992] A. Yeganeh-Heari, D. J. Weidner, and J. B. Parise, "Elasticity of  $\alpha$ -cristobalite: a silicon dioxide with a negative Poisson's ratio", *Science* **257**:5070 (1992), 650–652.

Received 19 Jun 2006. Revised 12 Mar 2007. Accepted 20 Apr 2007.

LEWIS WHEELER: [lwheeler@uh.edu](mailto:lwheeler@uh.edu)

Department of Mechanical Engineering, N207 Engineering Building 1, University of Houston, Houston, TX 77204-4006, United States

CLIFF YI GUO: [gyi@uh.edu](mailto:gyi@uh.edu)

Department of Mechanical Engineering, N207 Engineering Building 1, University of Houston, Houston, TX 77204-4006, United States

## A THREE DIMENSIONAL CONTACT MODEL FOR SOIL-PIPE INTERACTION

NELLY PIEDAD RUBIO, DEANE ROEHL AND CELSO ROMANEL

One of the most common causes of collapse of pipelines crossing unstable slopes is the large deformation induced by landslides. This paper presents a numerical methodology based on the finite element method for the analysis of buried pipelines considering the nonlinear behavior of the soil-pipe interface. This problem is inherently complex since it involves the interaction between two different bodies (pipe and soil), and is affected by many elements such as material nonlinearities, local and global buckling, soil settlement, pipe upheaval, among others. An important aspect that should be considered in the study of buried pipes is the mechanical behavior along the interface between the structure and the soil. The contact problem, which includes both a normal and a tangential constitutive law, is formulated through a penalty method. The finite element model considers full three-dimensional geometry, elasto-plastic material behavior and accounts for the presence of large displacements and deformations.

### 1. Introduction

In Brazil transport of petroleum, gas and oil derivatives between refineries and the port tanking terminals that collect and export petroleum products is generally made through buried pipelines that cross the mountain range of Serra do Mar. These mountains run parallel to the Atlantic Coast and stand between the Brazilian plateau, where most of the largest cities are located, and the lower sea plains.

A major concern during design and performance monitoring of these buried structures is the potential occurrence of soil movements, usually triggered by heavy rainfalls in areas lacking protective forest covering or those that have recently experienced changes of landscape caused by excavations, cuts and embankments due to road constructions, new industrial developments, etc. In cases of pipeline damage the consequences may be quite severe in terms of economical losses, social and environmental impacts. For example, the rupture of an expansion gasket during oil pumping in the state of Paraná in 2000 provoked a leakage of more than a million gallons of crude oil, endangering fauna and flora in addition to interrupting the distribution of potable water to the population of nearby towns.

Many analytical and computational procedures for the investigation of the mechanics of soil-pipe interaction problems are presented in the literature. The available numerical solutions are generally based on the finite element method and consider models ranging from simple one-dimensional beam models [Zhou and Murray 1993; Zhou and Murray 1996; Lim et al. 2001] and two-dimensional analysis of buried galleries [Katona 1983], to shell models [Selvadurai and Pang 1988]. Numerical models based on the boundary element model have also been employed [Mandolini et al. 2001]. Moreover, many different material models have been adopted to represent soil behavior, the most popular of which are elastic and elasto-plastic models.

---

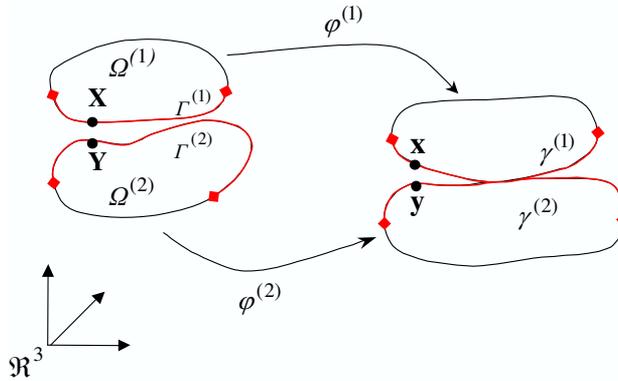
*Keywords:* soil-pipe interaction, frictional contact, penalty method, large deformation.

In the analysis of the behavior of buried pipes one very important aspect is the consideration of the interface behavior. This problem is inherently complex since it involves the interaction between two different bodies (pipe and soil) and is affected by many elements such as material nonlinearities, local and global buckling, soil settlement, pipe upheaval, among others. Various possible modes of deformation must be taken into account, including the stick and slip modes, for which normal stress remains compressive, as well as the debonding and rebonding modes, for which normal stress can reach zero. Models for the pipe-soil interface describe limiting cases such as perfect adhesion [Selvadurai and Pang 1988], elastic and inelastic springs for both transversal and longitudinal behavior [Zhou and Murray 1993], and continuum interface elements as in the pioneer works [Katona 1983; Desai et al. 1984]. More realistic continuum contact models including both normal and longitudinal contact forces models can be generically framed as optimization models, by which the contact constraints are introduced in the general equations of motion through a Lagrangian multiplier formulation and solved through mathematical programming algorithms. A long list of authors who have adopted this strategy includes [Simo et al. 1985; Kwak and Lee 1988; Lee et al. 1994; Laursen and Simo 1993; Ferreira and Roehl 2001]. Alternatively, the contact conditions are satisfied empirically through a penalty based formulation. Examples of this type of contact model are [Bathe and Chaudhary 1985; Peric and Owen 1992; Laursen 2002].

This paper presents a numerical methodology based on the finite element method for the analysis of buried pipelines considering the nonlinear behavior of the soil-pipe interface. The finite element model considers full three-dimensional geometry, elasto-plastic material behavior and accounts for the presence of large displacements and deformations. Both pipe and soil are modeled with hexahedral enhanced assumed strain elements. The numerical solution procedure is based on an incremental, iterative procedure, forming a sequence of nonlinear incremental problems solved by a Newton–Raphson scheme. The solution of boundary problems subject to the normal contact restrictions (impenetrability and compressive normal tractions at contact) and to the friction law (tangential constitutive law) is carried out here with a penalty formulation, by which the contact restrictions are approximated through an easy-to-implement procedure. The incremental evolution equations for the contact constitutive model are obtained through numerical integration with an implicit Euler algorithm. The element stiffness and contact matrices are obtained in the framework of a consistent linearization of the contact virtual work. Finally, application of the model to a slowly sliding slope with buried pipe is presented.

## 2. Continuum governing equations

**2.1. Equations of motion.** The formulation of the contact problem presented in this work is based on the work of Laursen and Simo [1993] and is reviewed here for the case of two deformable bodies  $\mathbf{B}^i$  for  $i = 1, 2$  in the space  $\mathfrak{R}^3$  as shown in Figure 1. We assume that the bodies are contact-free in the corresponding reference configurations  $\Omega^1$  and  $\Omega^2$  at time  $t = 0$ . The subsequent configurations indicated as  $\phi_t^1, \phi_t^2$  cause the two bodies to physically come into contact introducing interactive forces. The contact surfaces are represented by  $\Gamma^{(1)}$  and  $\Gamma^{(2)}$ , the so-called *slave* and *master* surfaces, respectively. The current surface location is given by  $\gamma^{(i)} = \phi_t^i(\Gamma^{(i)})$ . In the initial configuration, material points on  $\Gamma^{(1)}$  and  $\Gamma^{(2)}$  are represented by  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively; correspondingly, the current configuration is given by  $\mathbf{x} = \phi_t^{(1)}(\mathbf{X})$  and  $\mathbf{y} = \phi_t^{(2)}(\mathbf{Y})$ .



**Figure 1.** Body configurations at  $t = 0$  and general  $t$ .

Assuming a quasistatic response and considering a description of motion in the reference configuration, the classical equations of motion for each body  $i$  at time  $t$  are given by

$$\text{DIV} P_t^{(i)} + f_t^{(i)} = 0 \quad \text{in } \Omega^{(i)}, \quad P_t^{(i)} n_0^{(i)} = \bar{t}_t^{(i)} \quad \text{in } \Gamma_\sigma^{(i)}, \quad \varphi_t^{(i)} = \overline{\varphi_t^{(i)}} \quad \text{in } \Gamma_\varphi^{(i)}, \quad (1)$$

where  $P_t^{(i)}$  is the first Piola–Kirchhoff stress tensor,  $f_t^{(i)}$  is the prescribed body force,  $n_0^{(i)}$  is the outward normal in the reference configuration,  $\Gamma_\sigma^{(i)}$ ,  $\Gamma_\varphi^{(i)}$  are, respectively, the parts of  $\partial\Omega^{(i)}$  where the tractions  $\bar{t}_t^{(i)}$  and displacements  $\overline{\varphi_t^{(i)}}$  are given, and  $P_t^{(i)}$  is assumed to be given by a hyperelastic constitutive law.

**2.2. Frictional contact formulation.** For a pair of motions  $\phi^{(1)}(\cdot, t)$ ,  $\phi^{(2)}(\cdot, t)$ , the impenetrability restriction can be formulated for all points  $X \in \Gamma^{(1)}$  by first identifying a potential contact point  $Y_-(X, t)$  on the master surface according to the following closest point projection in the spatial configuration:

$$Y_-(X, t) = \arg \min_{Y \in \Gamma_c^{(2)}} \|\varphi^{(1)}(X, t) - \varphi^{(2)}(Y, t)\|.$$

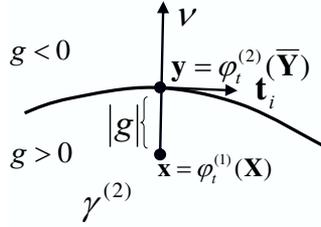
To formulate the contact conditions, a configuration-dependent differentiable distance function is introduced, which will be constrained to guarantee physical impenetrability.

For a pair  $X, Y_-$ , a gap function may be defined as  $g(X, t) = -\nu(\varphi^{(1)}(X, t) - \varphi^{(2)}(Y_-, t))$ , where  $\nu$  is the outward unit normal to the  $\gamma^{(2)}$  at  $y = \phi_t^{(2)}(Y)$  as illustrated in Figure 2. Then, the definition of  $g(X, t)$  is given in terms of the closest point projection of  $x = \phi_t^i(X)$  onto the opposing surface  $\gamma_c^{(2)}$ . The impenetrability condition is formulated as  $g(X, t) \leq 0$ .

Furthermore, the complementarity conditions are connected to the superficial contact force  $t^{(1)}(X, t) = P^{(1)}(X, t) \cdot n_0^{(1)}(X)$ , where  $P^{(1)}(X, t)$  is the first Piola–Kirchhoff tensor at  $X$  and  $n_0^{(1)}(X)$  is the outward normal at  $X$  in the reference configuration. This surface force may therefore be written as

$$t^{(1)}(X, t) = t_N(X, t)\nu + P_\nu t^{(1)}(X, t), \quad (2)$$

where  $P_\nu t^{(1)}$  is the projection of  $t^{(1)}$  onto the associated tangent plane, and  $t_N(X, t)$  represents the contact pressure at  $X$ .



**Figure 2.** Contact problem and unit outward normal.

The Kuhn–Tucker conditions for normal contact are given by

$$g(\mathbf{X}, t) \leq 0, \quad t_N(\mathbf{X}, t) \geq 0, \quad t_N(\mathbf{X}, t)g(\mathbf{X}, t) = 0, \quad t_N(\mathbf{X}, t)\dot{g}(\mathbf{X}, t) = 0. \quad (3)$$

The first three conditions reflect the impenetrability constraint, the compressive normal traction constraint, and the requirement that the pressure is nonzero only when contact takes place, that is, the gap function  $g = 0$ , respectively. The last requirement is the persistency condition used when considering frictional kinematics.

Once the impenetrability constraint (3)<sub>1</sub> induces a geometric structure through the gap function, an associated convective basis, adequate for definition of the frictional constraints, is necessary. Parameterizations for  $\Gamma^{(i)}$  and  $\gamma^{(i)}$  are adopted for body 2 (see Figure 3) according to the definition of a series of time indexed mappings  $\Psi_t^{(i)} : A^{(i)} \rightarrow \mathfrak{R}^{n-1}$ , with  $\Gamma^{(i)} = \Psi_0^{(i)}(A^{(i)})$ ,  $\gamma^{(i)} = \Psi_t^{(i)}(A^{(i)})$  and  $\Psi_t^{(i)} = \varphi_t^{(i)0} \Psi_0^{(i)}$ . The dimension of the contact surface  $\Gamma^{(i)}$  is one dimension lower than the number of spatial dimensions involved in the kinematic description. In the three-dimensional case, one point  $\xi \in A^{(2)}$  is given by  $\xi = (\xi^1, \xi^2)$ . Bases for  $\Gamma^{(2)}$  and  $\gamma^{(2)}$  are conveniently defined by partial derivatives with respect to these variables:

$$\mathbf{E}_\alpha(\xi) = \Psi_{0,\alpha}^{(2)}(\xi), \quad \mathbf{e}_\alpha(\xi) = \Psi_{t,\alpha}^{(2)}(\xi) = \mathbf{F}_t^{(2)}(\Psi_0^{(2)}(\xi))\mathbf{E}_\alpha(\xi), \quad \alpha = 1, 2.$$

In the above equations  $\mathbf{F}_t^{(2)}$  is the gradient deformation corresponding to  $\varphi^{(2)}$ . Subscript  $\alpha$  represents derivatives with respect to  $\xi^\alpha$ . For any point  $\mathbf{X} \in \Gamma^{(1)}$ , a point  $\mathbf{Y} \in \Gamma^{(2)}$  is assigned such that  $\mathbf{Y}_-$  is obtained through minimization.

Correlated points  $y_-$  and  $\xi_-$  in the spatial and parametric domains, respectively, are defined as

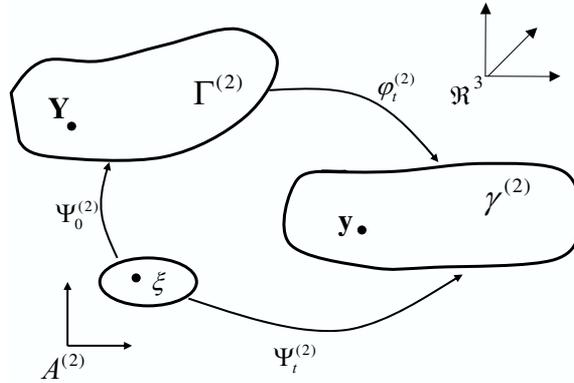
$$\mathbf{Y}_0(\mathbf{X}, t) = \psi_0^{(2)}(\xi_-(\mathbf{X}, t)), \quad \mathbf{y}_-(\mathbf{X}, t) = \psi_t^{(2)}(\xi_-(\mathbf{X}, t)).$$

Identification of  $\xi_-$  with point  $\mathbf{X}$  depends upon the motions of both bodies. The specific basis for  $\xi_-$  is

$$\mathbf{T}_\alpha = \mathbf{E}_\alpha(\xi_-), \quad \mathbf{t}_\alpha = \mathbf{e}_\alpha(\xi_-), \quad \alpha = 1, 2.$$

Tangent vectors  $\mathbf{T}_\alpha$  and  $\mathbf{t}_\alpha$  describe a convective basis at point  $\mathbf{X}$  relative to  $\Gamma^{(2)}$ . The normal vector is defined as  $\mathbf{v} = (\mathbf{t}_1 \times \mathbf{t}_2) / \|\mathbf{t}_1 \times \mathbf{t}_2\|$ .

According to the persistency condition, if  $\dot{g}(\mathbf{X}, t) = 0$ , the time rate of change of the relative position vector between  $\mathbf{x} = \phi^{(1)}$  and  $\mathbf{y}_- = \phi^{(2)}(\mathbf{Y}_-(\mathbf{X}, t), t)$  must be zero. The evaluation of this time derivative



**Figure 3.** Parameterizations for  $\Gamma^{(i)}$  and  $\gamma^{(i)}$ .

gives an important expression for the material relative velocity of  $X$ , namely,

$$V^{(1)}(X, t) - V^{(2)}(Y_-(X, t), t) = F_t^{(2)}(\Psi_0^{(2)}(\xi_-)) \frac{d}{dt}[\bar{Y}(X, t)].$$

In the above equation, the left side gives the relative material velocities of  $X$  and  $Y_-$ , thus physically representing the slip rate of  $X$  relative to the adjacent surface  $\gamma^{(2)} = \phi^{(2)}(\Gamma^{(2)})$ . The right hand side of this equation represents the geometry that is used in the definition of frictional evolution law.

$$V_T(X, t) := \frac{d}{dt}[Y_-(X, t)] = \dot{\xi}^\beta_-(X, t) T_\alpha.$$

Mathematically,  $V_T(X, t)$  represents the relative tangential velocity and, by the assumption of  $\dot{g}(X, t) = 0$  it contains no normal component. It is convenient to express  $V_T$  in a dual basis. One can define the dual basis vector, the metrics and the inverse metrics. The spatial counterpart of the material relative velocity  $V_T(X, t)$  is obtained through push forward transformation to the spatial frame. It and the frictional traction are expressed in the dual basis as

$$v_T^b(X, t) = M_{\alpha\beta} \dot{\xi}^\beta_-(X, t) t^\alpha, \quad t_T^b(X, t) := -P_v t^{(1)}(X, t) := t_{T\alpha}(X, t) t^\alpha.$$

Based on the description of slip velocity and traction, the Coulomb friction model is stated as

$$\Phi := \|t_T^b\| - \mu t_N \leq 0, \quad v_T^b - \zeta \frac{t_T^b}{\|t_T^b\|} = 0, \quad \zeta \geq 0, \quad \Phi \zeta = 0, \quad (4)$$

which are the friction law, relative tangential velocity, the irreversibility of slip, and the complementarity condition. In the above formulation  $\mu$  is the friction coefficient with hardening effects excluded,  $t_N$  and  $t_T$  are the normal and tangential contact forces,  $v_T$  is the relative tangential velocity. Frictional traction and velocity are expressed in dual basis, according to the large deformation theory. More details on the frictional contact formulation can be found in [Laursen and Simo 1993; Laursen 2002; Rubio et al. 2003].

**2.3. Formulation of the virtual work of contact.** We consider the approximate weak form of the global equilibrium equations. The test function  $\phi^{*(i)} : \Omega(i) \rightarrow R^3$  satisfies the condition  $\phi^{*(i)} = 0$  on  $\Gamma_\phi^{(i)}$ . Restrictions placed upon  $\phi^{*(i)}$  by the contact conditions are not imposed since such limitations are to

be removed using the penalty regularization introduced above. Multiplying Equations (1) by  $\varphi^{*(i)}$  and integrating by parts over  $\Omega^{(i)}$  we obtain the weak form of the equilibrium:

$$G^{(i)}(\varphi_t^{(i)}, \varphi^{*(i)}) := \int_{\Omega^{(i)}} \mathbf{P}_t^{(i)} \cdot \text{GRAD}[\varphi^{*(i)}] d\Omega^{(i)} - \int_{\Omega^{(i)}} \mathbf{f}_t^{(i)} \cdot \varphi^{*(i)} d\Omega^{(i)} - \int_{\Gamma_\sigma^{(i)}} \overline{\mathbf{t}}^{(i)} \cdot \varphi^{*(i)} d\Gamma_\sigma^{(i)},$$

$$G^{(i)}(\varphi_t^{(i)}, \varphi^{*(i)}) := \int_{\Gamma^{(i)}} \overline{\mathbf{t}}_t^{(i)} \cdot \varphi^{*(i)} d\Gamma^{(i)}.$$

The quantity  $G^{(i)}$  is the sum of the internal virtual work and the virtual work of the applied forces and tractions for body  $i$ . The balance of the virtual work of the contact forces acting on  $\Gamma^{(i)}$  is

$$G(\varphi_t, \varphi^*) := G^{(1)}(\varphi_t^{(1)}, \varphi^{*(2)}) + G^{(2)}(\varphi_t^{(2)}, \varphi^{*(2)}) = \int_{\Gamma^{(1)}} \overline{\mathbf{t}}_t^{(1)} \cdot \varphi^{*(1)} d\Gamma^{(1)} + \int_{\Gamma^{(2)}} \overline{\mathbf{t}}_t^{(2)} \cdot \varphi^{*(2)} d\Gamma^{(2)},$$

where  $\varphi_t$  is the collection of mappings  $\varphi_t^{(1)}$  and  $\varphi_t^{(2)}$  and so is  $\varphi^*$ . The contact contribution of the integral over  $\Gamma^{(1)}$  is

$$G(\varphi_t, \varphi^*) + G_c(\varphi_t, \varphi^*) = 0, \quad G_c(\varphi_t, \varphi^*) = - \int_{\Gamma^{(1)}} \mathbf{t}_t^{(1)}(\mathbf{X}) \cdot \{\varphi^{*(1)}(\mathbf{X}) - \varphi^{*(2)}[\overline{\mathbf{Y}}(\mathbf{X})]\} d\Gamma^{(1)}.$$

The statement of the contact virtual work is given by

$$G_c(\varphi_t, \varphi^*) = - \int_{\Gamma^{(1)}} [t_N \mathbf{v} - t_{T_\alpha} \boldsymbol{\tau}^\alpha] \cdot [\varphi^{*(1)}(\mathbf{X}) - \varphi^{*(2)}(\overline{\mathbf{Y}}(\mathbf{X}))] d\Gamma^{(1)} = \int_{\Gamma^{(1)}} [t_{N_t} \delta g - t_{T_{\alpha_t}} \delta \xi_-^\alpha] d\Gamma^{(1)}.$$

### 3. Numerical solution

**3.1. Penalty regularization of constraints.** The solution of boundary problems subject to restrictions such as those presented in Equation (2) for normal contact and in Equation (4) for the Coulomb friction laws is carried out here with a penalty formulation by which the restrictions are approximated through an easy-to-implement procedure. For normal contact, a normal penalty parameter  $\varepsilon_N$  is introduced in the definition of the constitutive relation of the normal force  $t_N = \varepsilon_N \langle g \rangle$ , where  $\langle \cdot \rangle$  denotes the positive part of the operand.

By introducing a tangential penalty  $\varepsilon_T$ , the regularization for the frictional response is expressed as

$$\Phi := \|\mathbf{t}_T^b\| - \mu t_N \leq 0, \quad \mathbf{v}_T^b - \zeta \frac{\mathbf{t}_T^b}{\|\mathbf{t}_T^b\|} = \frac{1}{\varepsilon_T} L_v \mathbf{t}_T^b, \quad \zeta \geq 0, \quad \Phi \zeta = 0,$$

where  $L_v \mathbf{t}_T^b := \dot{t}_{T_\alpha} \mathbf{t}^\alpha$  is the Lie derivative of the tangential force. The above regularization is exact only in the limit  $\varepsilon_N \rightarrow \infty$  and  $\varepsilon_T \rightarrow \infty$ , in which case the slip rate  $\zeta \mathbf{t}_T^b / \|\mathbf{t}_T^b\|$  equals the relative velocity  $\mathbf{v}_T^b$ . These relations are easy to incorporate in the virtual work principle and subsequently implement in a finite element procedure. For frictional problems, the tangential gap function is introduced as  $g_T^\alpha = \overline{\xi_{n+1}^\alpha} - \overline{\xi_n^\alpha}$ .

**3.2. Incremental finite element formulation.** The boundary value problem can be solved incrementally by considering a set of subintervals  $U_{n=0}^N [t_n, t_{n+1}]$ . The evolution equations for the constitutive model are obtained through numerical integration. Here we adopt an implicit Euler algorithm. In the framework

of a consistent linearization, the contact virtual work is defined according to

$$\Delta G^c(\varphi_t, \varphi^*) = \Delta \left\{ \int_{\Gamma_c^{(1)}} [t_N \delta g + t_{T_\alpha} \cdot \delta \bar{\xi}^\alpha] d\Gamma \right\} = \int_{\Gamma_c^{(1)}} [\Delta(t_N \delta g) + \Delta t_{T_\alpha} \delta \bar{\xi}^\alpha + t_{T_\alpha} \Delta \delta \bar{\xi}^\alpha] d\Gamma^{(1)}, \quad (5)$$

where  $t_N$  are contact pressures and  $t_T$  frictional tractions. The quantity  $\Delta(\delta g)$  is computed by linearizing  $\delta g$ , which is the linearized variation of the gap function,  $\delta \bar{\xi}$  is obtained by application of the orthogonality condition of the tangent vectors with the normal vector, and  $\Delta(\delta \bar{\xi})$  is obtained by computing the directional derivative of the orthogonality condition.

For finite element discretization of the domain, the contact virtual work expression in discrete form is

$$G_c(\varphi^h, \varphi^{*h}) = \int_{\Gamma^{(i)h}} [t_{Nt}^h \delta g^h + t_T^h \delta \bar{\xi}^{\alpha h}] d\Gamma^{(1)h},$$

where the discrete counterparts of  $\phi^{(i)}$  and  $\phi^{*(i)}$  are  $\phi^{(i)h}$  and  $\phi^{*(i)h}$ , defined over individual element surfaces as  $\phi_e^{(i)h}(\eta) = \sum N^a(\eta) \phi_a^{(i)}$  for  $a = 1, \dots, n_e$ . The term  $\phi_a^{(i)}$  is the nodal value of  $\phi^{(i)h}$ ,  $n_e$  is the number of nodes per element surface,  $N^a(\eta)$  is an isoparametric shape function for three-dimensional problems. Using the same scheme  $X_e^h(\eta) = \sum N^a(\eta) X_a$ .

Solution of the weak form of equilibrium is obtained here with the Newton–Raphson method, which requires linearization of Equation (5). For numerical integration, the linearized virtual contact work is

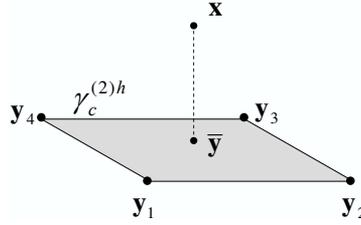
$$\Delta G_c(\varphi^h, \varphi^{*h}) \approx \sum_{j=1}^{n_{el}} \sum_{k=1}^{n_{int}} W^k j(\boldsymbol{\eta}^k) \left[ \Delta [t_N^h(\boldsymbol{\eta}^k) \delta g^h(\boldsymbol{\eta}^k)] + \Delta t_{T_\alpha}^h(\boldsymbol{\eta}^k) \delta \bar{\xi}^{\alpha h}(\boldsymbol{\eta}^k) + t_{T_\alpha}^h(\boldsymbol{\eta}^k) \Delta [\delta \bar{\xi}^{\alpha h}(\boldsymbol{\eta}^k)] \right],$$

where  $n_{int}$  is the number of integration points on each contact surface element  $\Gamma^{(1)h}$ ,  $W^k$  is the integration weight factor,  $\delta \boldsymbol{\Phi}_c^k$  is the vector of nodal displacement variations,  $\mathbf{R}_c^k$  is the residual vector, and the index  $k$  indicates the number of the integration point. The terms  $\delta g^h(\boldsymbol{\eta}^k)$  are the variations of  $g$  and the simplification of the variation  $\delta \bar{\xi}^{\alpha h}(\boldsymbol{\eta}^k)$  with the corresponding discrete fields. Expression (5) can now be written as

$$\Delta G_c(\varphi^h, \varphi^{*h}) = \sum_{j=1}^{n_{el}} \sum_{k=1}^{n_{int}} W^k j(\boldsymbol{\eta}^k) \delta \boldsymbol{\Phi}_c^k \cdot \mathbf{K}_c^k \Delta \boldsymbol{\Phi}_c^k. \quad (6)$$

In the above equation  $\mathbf{K}_c^k$  is the contact stiffness matrix. The linearized contact terms  $\Delta [t_N^h(\boldsymbol{\eta}^k) \delta g^h(\boldsymbol{\eta}^k)]$  and  $\Delta [\delta \bar{\xi}^{\alpha h}(\boldsymbol{\eta}^k)]$  are given by their corresponding discrete components. Vector  $\Delta \boldsymbol{\Phi}_c^k$  contains the nodal displacement values  $\mathbf{u}^h$ , which take part at contact. The term  $\Delta t_T^h(\boldsymbol{\eta}^k)$  is obtained by a classical plasticity return algorithm. In the present work a nodal quadrature is employed by the evaluation of Equation (6) as presented in detail in [Ferreira and Roehl 2001].

The finite element discretization is carried out with eight-node hybrid brick elements based on an enhanced assumed strain formulation in the framework of large strain J2 plasticity; for details see [Simo et al. 1985; Roehl and Ramm 1996]. In this case the node-to-surface contact involves five nodes as illustrated in Figure 4.



**Figure 4.** Slave node and master surface.

Accordingly,  $\delta\Phi$  contains the displacement variations of the contacting slave node  $\varphi^{*(1)}(X)$  and those of the four nodes on finite element on the master surface  $\gamma_e^{(2)h}(\varphi^{*(2)}(Y))$ :

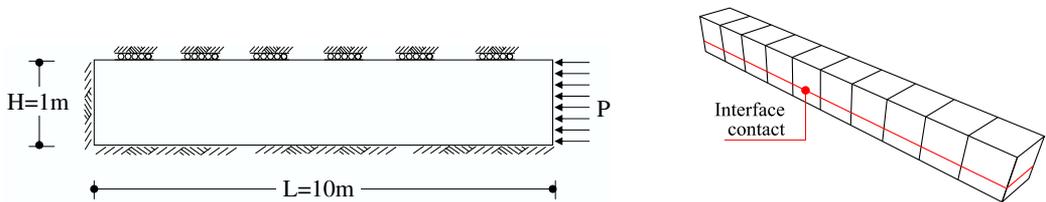
$$\delta\Phi = \begin{bmatrix} \varphi^{*(1)}(X) \\ \varphi^{*(2)}(Y_1) \\ \varphi^{*(2)}(Y_2) \\ \varphi^{*(2)}(Y_3) \\ \varphi^{*(2)}(Y_4) \end{bmatrix}, \quad \Delta\Phi = \begin{bmatrix} u^{(1)}(X) \\ u^{(2)}(Y_1) \\ u^{(2)}(Y_2) \\ u^{(2)}(Y_3) \\ u^{(2)}(Y_4) \end{bmatrix}.$$

### 4. Applications

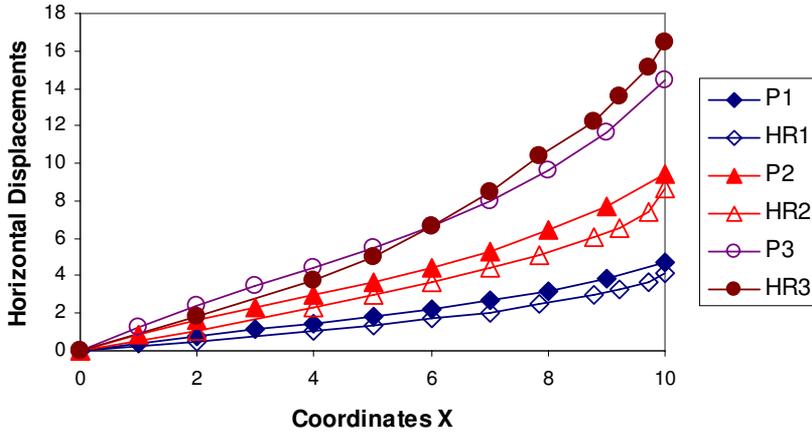
**4.1. Benchmark for soil-structure interface contact.** Figure 5 (left) illustrates a long elastic block (that is,  $L \gg H$ ) loaded in compression at one end and restrained at the other. The block is also restrained against compression in the  $x$  direction by the frictional contact model along its base. No strain is permitted in either the  $y$  or  $z$  direction. The block has length  $L = 10$  m with  $L/H = 10$ , Young’s modulus  $E = 1.0 \times 10^5$  kPa, Poisson’s coefficient  $\nu = 0.0$ , and the initial value of applied stress  $P = 100$  kPa.

For this analysis the penalty method was used with normal and tangential penalties equal to  $\varepsilon_N = 10^4$  and  $\varepsilon_T = 10^8$ , respectively. The Coulomb frictional law at the block-foundation interface has friction coefficient  $\mu = 0.5$ . The analysis was executed under loading control conditions. The system was modeled by a finite element mesh consisting of 20 eight-node hexahedral elements; see Figure 5 (right). The results for the horizontal displacements at the contact interface obtained in this analysis were compared with the results obtained in the numerical solutions developed for [Hird and Russell 1990] for different load levels as shown in Figure 6.

**4.2. Buried pipeline.** Figure 7 shows a buried steel pipe of diameter  $D_0 = 1$  m embedded at a depth  $2D_0$  with the following mechanical and geometrical properties: axial stiffness  $EA = 4.2 \times 10^5$  kN, flexural



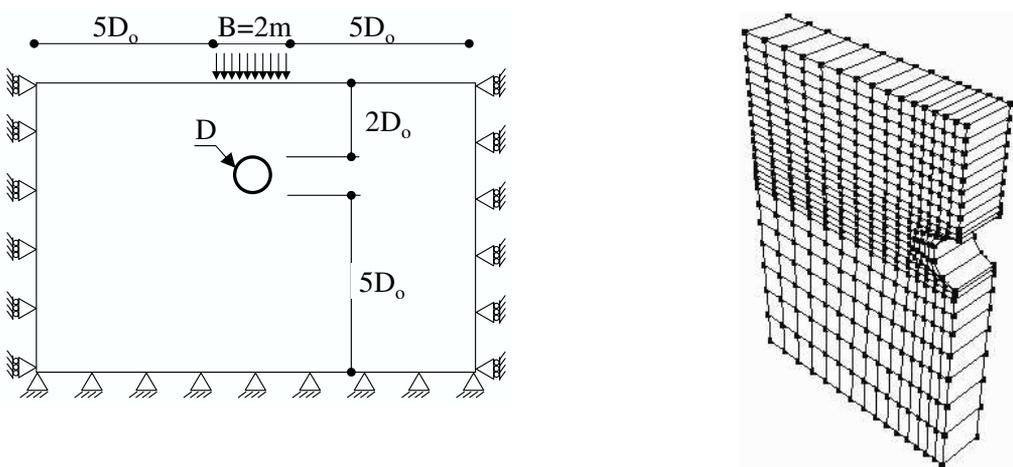
**Figure 5.** Definition of the long elastic block problem (left). Finite element mesh (right).



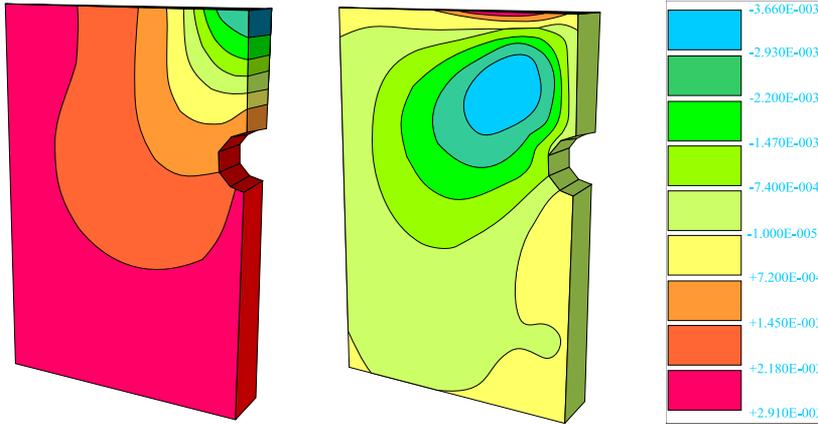
**Figure 6.** Horizontal displacements at contact interface; results of CARAT versus [Hird and Russell 1990].

stiffness  $EI = 0.1 \text{ kN}\cdot\text{m}^2$ , thickness  $t = 2 \text{ mm}$ , and Poisson’s ratio  $\nu_p = 0.3$ . The linear elastic soil layer has thickness  $H = 8 \text{ m}$ , Young’s modulus  $E = 2.7 \times 10^3 \text{ kPa}$ , and Poisson’s ratio  $\nu_s = 0.33$ . The soil layer is submitted to a strip load  $q = 100 \text{ kPa}$ , uniformly distributed over a length  $B = 2 \text{ m}$  in the  $xy$ -plane; see Figure 6.

Due to symmetry, only half of the soil-pipe system was modeled by a finite element mesh (Figure 7, right), consisting of 365 eight-node elements (brick8). The analysis was carried out under the assumption of plane strain conditions, by preventing axial displacements through the introduction of proper boundary constraints. The frictional coefficient was considered to be  $\mu = 0.5$ , and the penalty parameters were  $\epsilon_N = 10^4$  and  $\epsilon_T = 10^8$ . The analysis was executed under displacement control conditions. Figure 8 shows the horizontal and vertical field displacement of the soil-pipe system according to the frictional contact formulation.



**Figure 7.** Geometry of the soil-pipe system (left). Finite element mesh (right).



**Figure 8.** Vertical (left) and horizontal (right) displacement fields according to the frictional contact formulation.

**4.3. Soil-pipe interaction: three-dimensional model.** In this example the elastic behavior of the soil is considered with Young's modulus  $E = 50.0$  MPa and Poisson's coefficient  $\nu = 0.2$ . The pipeline assumes an elasto-plastic constitutive model based on the von Mises criteria with isotropic hardening. The yielding stress and the tangent modulus are  $S_y = 420$  MPa and  $E_T = 75000$  MPa, respectively. Pipe properties are listed in [Table 1](#).

The loading applied to the pipeline consists in an internal pressure equal to 9.0 MPa, transversal load of 1000.0 N/m, and the overburden ( $\gamma = 1.8$  KN/m<sup>3</sup>), according to [Table 2](#).

Due to symmetry, only half of the soil-pipe system is modeled by a finite element mesh (longitudinal direction), consisting of 622 eight-node hybrid brick elements (Hexa8-E3). Pipeline geometry and the finite element model are shown in [Figure 9](#). The frictional coefficient is considered to be  $\mu = 0.1$  and the penalty parameters are  $\varepsilon_N = 10^2$  and  $\varepsilon_T = 10^2$ . The analysis was carried out under load control conditions. The frictional contact problem formulation simulates the soil-pipe interface behavior.

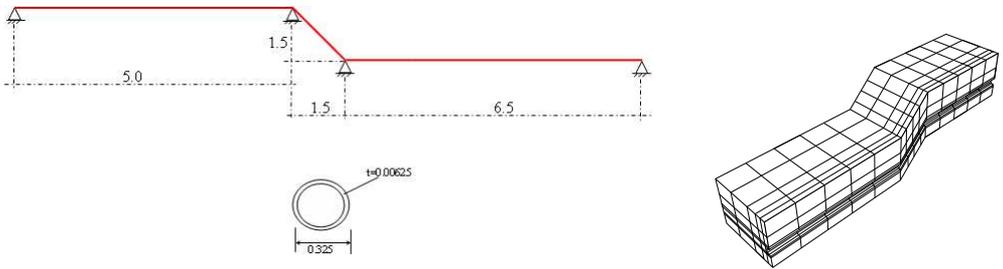
The internal pressure induces longitudinal stresses in the pipe due to Poisson's effect; see [Figure 10](#). These longitudinal stresses arise when the pipe is restricted at its ends and/or by the presence of longitudinal friction. We have verified that the pipe has achieved yielding; see [Figure 10](#).

Parameter	Value
$I_{zz}$ (m <sup>4</sup> )	$7.9516531 \times 10^{-5}$
$A$ (m <sup>2</sup> )	$6.2586416 \times 10^{-3}$
$De$ (m)	0.325
$Di$ (m)	0.3125
$t$ (m)	0.00625

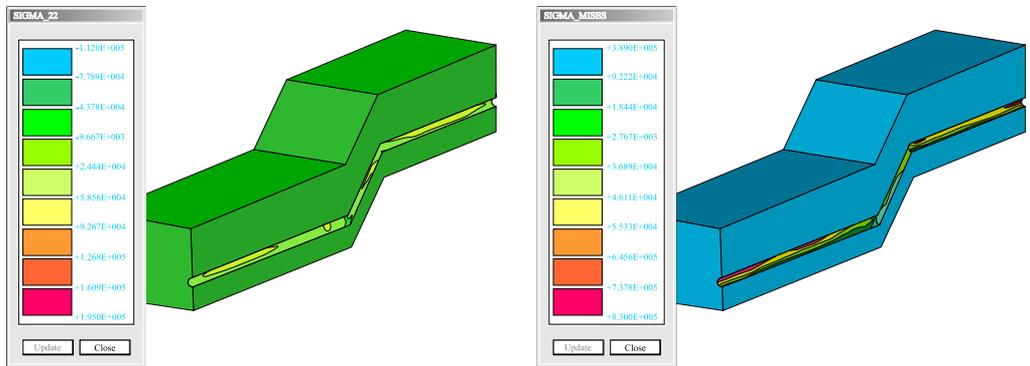
**Table 1.** Pipe properties.

Distance	Overburden	Additional Load	Internal Pressure
0 – 5.0 m	X	X	X
5.0 – 6.5 m	X	-	X
6.5 – 8.0 m	-	-	X
8.0 – 13.0 m	-	-	X

**Table 2.** Pipeline load.



**Figure 9.** The pipeline geometry in the yz-plane (left). Finite element mesh: 622 Hexa8-EAS elements (right).



**Figure 10.** Longitudinal stresses and von Mises stresses obtained with our model.

### 5. Conclusion

This work presents a finite element numerical model for the analysis of buried pipes. The solution of elasto-plastic contact problem includes the presence of large elasto-plastic strains. The contact conditions are imposed through a penalty formulation that has been proven quite effective in the cases studied, if the penalty parameters are adequately chosen. For problems in which the contacting bodies present stiffness of the same order of magnitude the choice of these parameters is not difficult. Large normal contact forces make the parameter calibration more troublesome.

This is not usually the case by pipe-soil systems. An application of the model to the problem of a buried pipe under close to site conditions illustrates the effectiveness of the soil-pipe interaction model for more realistic engineering problems.

## References

- [Bathe and Chaudhary 1985] K. J. Bathe and A. Chaudhary, “A solution method for planar and axisymmetric contact problems”, *Int. J. Numer. Methods Eng.* **21** (1985), 65–88.
- [Desai et al. 1984] C. S. Desai et al., “Thin-layer element for interfaces and joints”, *Int. J. Numer. Anal. Methods Geomech.* **8** (1984), 19–43.
- [Ferreira and Roehl 2001] K. I. Ferreira and D. Roehl, “Three dimensional elastoplastic contact analysis at large strains with enhanced assumed strain elements”, *Int. J. Solids Struct.* **38** (2001), 1855–1870.
- [Hird and Russell 1990] C. Hird and D. Russell, “A benchmark for soil-structure interface elements”, *Comput. Geotech.* **10** (1990), 139–147.
- [Katona 1983] M. Katona, “A simple contact-friction interface element with applications to buried culverts”, *Int. J. Numer. Anal. Methods Geomech.* **7** (1983), 371–384.
- [Kwak and Lee 1988] B. M. Kwak and S. S. A. Lee, “Complementarity problem formulation for two-dimensional frictional contact problems”, *Comput. Struct.* **28** (1988), 469–480.
- [Laursen 2002] T. A. Laursen, *Computational contact and impact mechanics*, Springer-Verlag, Berlin, 2002. Fundamentals of modeling interfacial phenomena in nonlinear finite element analysis. [MR 1902698 \(2003e:74050\)](#)
- [Laursen and Simo 1993] T. A. Laursen and J. C. Simo, “A continuum-based finite element formulation for the implicit solution of multibody, large deformation frictional contact problems”, *Int. J. Numer. Methods Eng.* **36** (1993), 3451–3485.
- [Lee et al. 1994] S. C. Lee, B. M. Kwak, and O. K. Kwon, “Analysis of incipient sliding contact by three-dimensional linear complementarity problem formulation”, *Comput. Struct.* **53** (1994), 695–708.
- [Lim et al. 2001] M. L. Lim, M. K. Kim, T. W. Kim, and J. W. Jang, “The behavior analysis of buried pipeline considering longitudinal permanent ground deformation”, in *Pipeline 2001: Advances in Pipelines Engineering & Construction* (San Diego, California), vol. 3, edited by J. P. Castronovo, 107, ASCE, 2001.
- [Mandolini et al. 2001] Mandolini et al., “Coupling of underground pipelines and slowly moving landslides by bem analysis”, *Comput. Model. Eng. Sci. CMES* **2:1** (2001), 39–48.
- [Peric and Owen 1992] D. Peric and R. J. Owen, “Computational model for 3D contact problems with friction based on the penalty method”, *Int. J. Numer. Methods Eng.* **35** (1992), 1289–1309.
- [Roehl and Ramm 1996] D. Roehl and E. Ramm, “Large elasto-plastic finite element analysis of solids and shells with the enhanced assumed strain concept”, *Int. J. Solids Struct.* **33:20–22** (1996), 3215–3237.
- [Rubio et al. 2003] N. Rubio, D. Roehl, and C. Romanel, “Design of buried pipes considering the reciprocal soil-structure interaction”, pp. 1279–1287 in *Pipelines 2003: New Pipeline Technologies, Security* (Baltimore, MD), edited by M. Najafi, ASCE, 2003.
- [Selvadurai and Pang 1988] A. P. S. Selvadurai and S. Pang, “Non-linear effects in soil-pipeline interaction in the ground subsidence zone”, pp. 1085–1094 in *Proc. 6th Intr. Conf. Numer. Methods in Geomechanics* (Innsbruck), vol. 2, A. A. Balkema, 1988.
- [Simo et al. 1985] J. C. Simo, P. Wriggers, and R. L. Taylor, “A perturbed Lagrangian formulation for the finite element solution of contact problems”, *Comput. Methods Appl. Mech. Eng.* **50** (1985), 163–180.
- [Zhou and Murray 1993] Z. Zhou and D. Murray, *Behavior of buried pipelines subjected to imposed deformations*, vol. Volume V, Pipeline Technology, ASME, 1993. OMAE.
- [Zhou and Murray 1996] Z. Zhou and D. Murray, “Pipeline beam models using stiffness property deformation relations”, *J. Transp. Eng.* **122:2** (1996), 164–172.

Received 2 Aug 2006. Accepted 20 Apr 2007.

NELLY PIEDAD RUBIO: [nrubio@civ.puc-rio.br](mailto:nrubio@civ.puc-rio.br)

*Civil Engineering Department, Pontifícia Universidade Católica do Rio de Janeiro, CEP 22453-900 Rio de Janeiro, Brazil*

DEANE ROEHL: [droehl@civ.puc-rio.br](mailto:droehl@civ.puc-rio.br)

*Civil Engineering Department, Pontifícia Universidade Católica do Rio de Janeiro, CEP 22453-900 Rio de Janeiro, Brazil*

CELSO ROMANEL: [romanel@civ.puc-rio.br](mailto:romanel@civ.puc-rio.br)

*Civil Engineering Department, Pontifícia Universidade Católica do Rio de Janeiro, CEP 22453-900 Rio de Janeiro, Brazil*

## CALCULATION OF INERTIAL PROPERTIES OF THE MALLEUS-INCUS COMPLEX FROM MICRO-CT IMAGING

JAE HOON SIM, SUNIL PURIA AND CHARLES R. STEELE

The middle ear bones are the smallest bones in the human body and are among the most complicated functionally. These bones are located within the temporal bone making them difficult to access and study. We use the micro-CT imaging modality to obtain quantitative inertial properties of the MIC (malleus-incus complex), which is a subcomponent of the middle ear. The principal moment of inertia of the malleus along the superior-inferior axis ( $17.3 \pm 2.3 \text{ mg/mm}^3$ ) is lower by about a factor of six in comparison to the anterior-posterior and lateral-medial axes. For the incus, the principal moment of inertia along the superior-inferior axis ( $35.3 \pm 6.9 \text{ mg/mm}^3$ ) is lower by about a factor of two than for the other two axes. With the two bones combined (MIC), the minimum principal moment of inertia ( $132.5 \pm 18.5 \text{ mg/mm}^3$ ) is still along the superior-inferior axis but is higher than for the individual bones. The superior-inferior axis inertia is lower by a factor of 1.3 than along the anterior-posterior axis and is lower by a factor 2 along the lateral-medial axis. Values for inertia of the MIC show significant individual differences in three human ears measured, suggesting that middle ear models should be based on individual anatomy. Imaging by micro-CT scanner is a nondestructive modality that provides three-dimensional volume information about middle ear bones at each stage of manipulation with resolution down to  $10 \mu\text{m}$ . In this work extraneous tissue is removed to obtain a sufficiently small specimen. However, advances in imaging hold promise that this capability will be available for in vivo measurements.

### 1. Introduction

The ossicular chain in the middle ear consists of the MIC and the stapes, which transfer vibrations of eardrum into fluid vibrations in the inner ear. This is a very important step in the hearing process. Because these bones are mobile in all three dimensions, the inertial properties are important for a biomechanical model of the middle ear. Inertial properties in the human middle ear bones have been studied [Kirikae 1960; Beer et al. 1996; Weistenhöfer and Hudde 1999], and those values have been widely used for models of the middle ear [Eiber and Freitag 2002; Gan et al. 2002; Koike et al. 2002].

Three-dimensional volume information for the ossicles is necessary to calculate the inertial properties. Such information has been obtained by microscopic surface measurements [Kirikae 1960; Beer et al. 1996; Weistenhöfer and Hudde 1999]. However, this method cannot yield information on the mass distribution inside bone and is not suitable for the complicated features of the middle ear bones. Alternatively, traditional histological methods are used, but they require several months for results. Another issue in modeling the middle ear is the fairly large difference in individual middle ear anatomy. Such a large difference does not allow a nominal middle ear model to be used for all ears. To construct the middle ear

---

*Keywords:* inertial properties, principal axes, ossicles, malleus-incus complex (MIC), middle ear, computed tomography (CT). This work was supported in part by a grant from the NIDCD of NIH (DC005960).

model based on individual anatomy, three-dimensional volume information should be obtained at each stage of manipulation of the specimen, and a nondestructive method is needed. Both microscopic surface measurement and the histological method are destructive.

The microscale X-ray computed tomography (micro-CT) scanner provides nondestructive imaging with resolution as small as  $10\ \mu\text{m}$  resolution. The first application of the micro-CT to obtain the general geometry of the middle ear bones [Decraemer et al. 2003; Lane et al. 2004; 2005] found more clarity than with MRI. The present effort extends the method to obtain the quantitative mechanical properties, so some detail of the procedure is given. While the steps involve well-known relations, assembling, generation and interpretation of the images require insight into the mechanics and materials.

## 2. Material and methods

**2.1. Temporal bone preparation.** Four human temporal bones from three different human cadavers were used (two right ears and one left ear). The ear canal was dissected, the cochlea was removed, and the middle ear cavity was dissected to reduce X-ray attenuation due to materials of no interest for present purposes. The reduction of specimen size also allows increased resolution. Since the focus was on the MIC, the eardrum and stapes were dissected by a surgical laser.

**2.2. Micro-CT scanning.** The vivaCT 40 micro-CT scanner developed by SCANCO Medical AG (see [www.scanco.ch](http://www.scanco.ch)) was used in this study. This machine permits control of the resolution, the photon energy, the intensity of the X-ray beam, and the integration time, all of which determine the visibility of the objects of interest.

Scans with higher resolution result in clearer images, especially on micron sized structures such as the ossicles. Figure 1 shows example slice images of the intact ear (left) and the isolated MIC preparation (right) when the resolution of  $12.5\ \mu\text{m}$  was obtained with a  $25.6\ \text{mm}$  diameter holder. This machine allows us to perform high resolution scans up to  $2048 \times 2048$  pixels per image, which correspond to resolutions of  $10.5\ \mu\text{m}$  for a  $21.5\ \text{mm}$  diameter holder. The best resolution of  $10.5\ \mu\text{m}$  could be obtained for our specimen by reducing its size to fit into the  $21.5\ \text{mm}$  diameter holder.

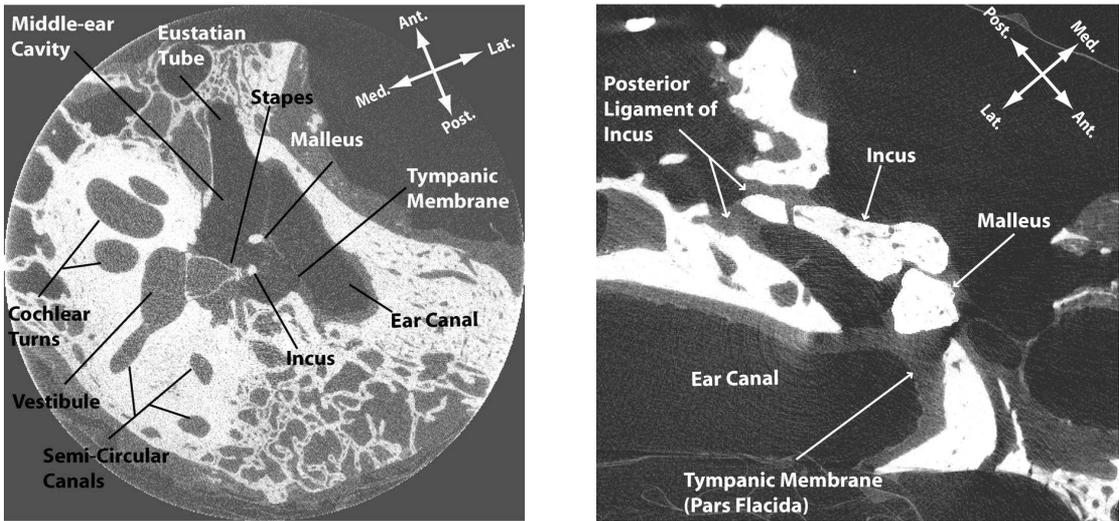
The transmitted photons from an X-ray source to a detector either interact with a particle of matter in their path or pass unaffected [Johns and Cunningham 1974]. The number of photons in the laser beam that are lost due to attenuation in a region of thickness  $\Delta x$  can be represented as

$$\Delta N = -\mu N \Delta x, \quad (1)$$

where  $N$  is the total number of impinging photons and  $\mu$  is a constant of proportionality known as the *linear attenuation coefficient* [Macovski 1983]. The final number of photons  $N_{\text{out}}$  after traversing an attenuation region of thickness  $x$  can be represented by the initial number of photons suppressed by an exponential decay term, the relationship known as Beer Lambert's Law:

$$N_{\text{out}} = N_{\text{in}} e^{-\mu x}. \quad (2)$$

The attenuation coefficient  $\mu$  depends on the photon energy of the beam and absorption characteristics of the elements as dictated by the quantum mechanical energy levels of the element involved during the absorption process. Since the absorption strength also depends on the mass of the material itself, attenuation coefficients are often characterized instead by the so-called mass attenuation coefficient  $\mu/\rho$



**Figure 1.** Micro-CT images of intact (left) and isolated MIC (right) from a human temporal bone preparation obtained from  $12.5 \mu\text{m}$  iso-volume scans with the 25.6 mm diameter holder.

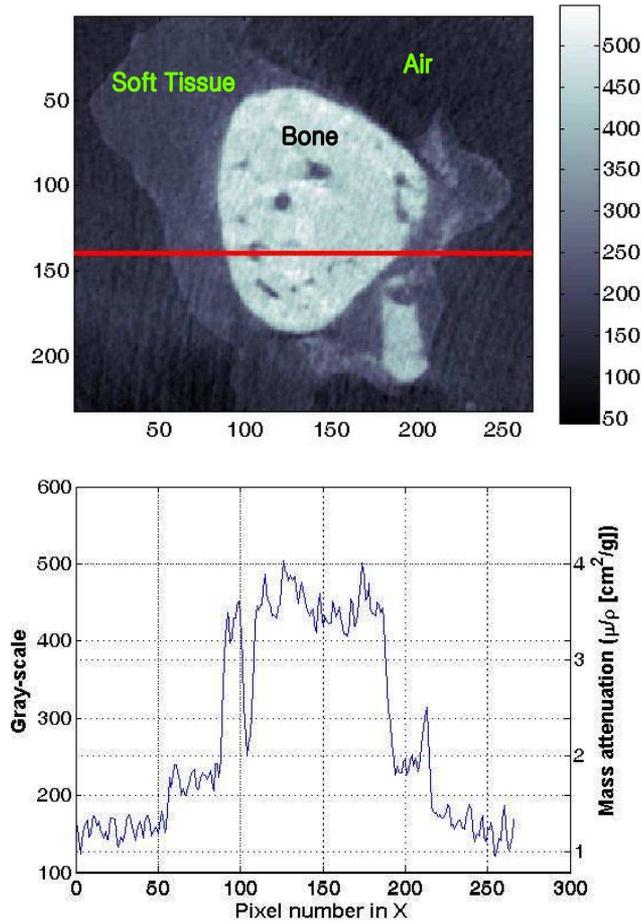
[Johns and Cunningham 1974; Macovski 1983]. Mass attenuation coefficients for body materials show relatively large differences in the lower photon energy regions, where the photoelectric effect is significant. At higher energies, where the attenuation is primarily due to Compton scatter, mass attenuation coefficients become the same for all biological tissues [Macovski 1983]. Even though lower photon energy provides a larger contrast ratio between biological tissues, it is limited by the nonlinear beam hardening artifacts [Brooks and Di Chiro 1976; Wang et al. 1996; Wang and Vannier 1998]. X-ray photons emitted from an X-ray source do not all have the same energy. As an X-ray beam traverses an object, photons within the lower energy spectrum are more readily absorbed and the portion of higher energy photons in the X-ray spectrum increases. Therefore, when high X-ray absorption structures are in the field of view, beam hardening effects are particularly pronounced since photoelectric absorption in bone is high due to the high calcium content.

The vivaCT 40 micro-CT scanner in this study allows 30, 55, or 70 keV as the diagnostic energy level. Because of large interruptions due to beam hardening in bony portions with the lower energy levels, 70 keV was selected as the photon energy, where bones are clearly distinguishable from the background.

The intensity at the detector  $I_d$  is given by

$$I_d(x, y) = \int I_o(E) \exp \left[ - \int \mu(x, E) \right] dE, \quad (3)$$

where  $I_o(E)$  is the incident X-ray beam intensity as a function of the energy per photon  $E$  and  $\mu(x, E)$  is the linear attenuation coefficient at each region [Macovski 1983; Ketcham and Carlson 2001]. The image clarity depends on the signal-to-noise ratio, which is directly affected by the X-ray intensity. Higher intensities improve the underlying counting statistics, but often require a larger focal spot, which results in degrading image sharpness [Ketcham and Carlson 2001]. The focal spot size of the X-ray tube

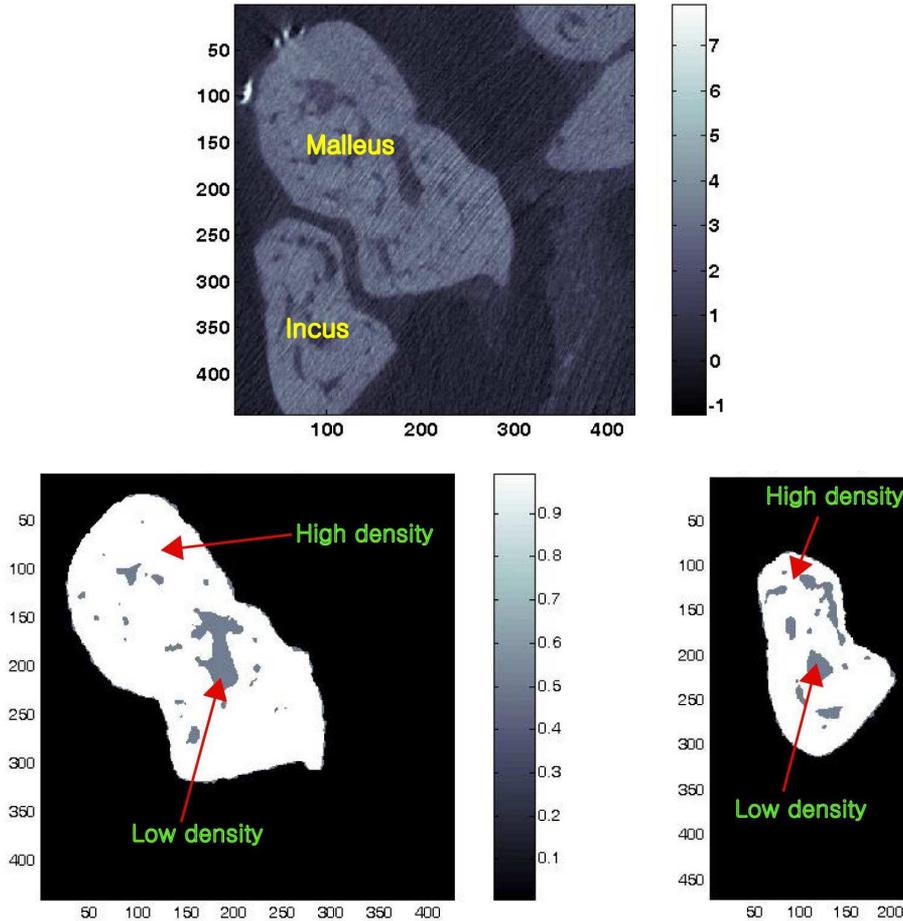


**Figure 2.** Top: slice image from micro-CT. Bone and soft tissue are distinguishable from surrounding air. Bottom: grayscale along the red line above. Grayscale value of bone has higher range (350 ~ 550) than for soft tissue (200 ~ 350) and air (below 200).

influences the unsharpness of the final image. Generally the smaller spot size is better for the image sharpness.

The micro-CT scanner used allows the maximum X-ray intensity of  $145 \mu\text{A}$ , where we could get good signal-to-noise ratio and good image clarity for the default integration time of 380 msec. The typical scan length was about 12 mm for scans in the superior-inferior direction and about 9 mm for scans in the anterior-posterior direction. These values in scan length correspond to approximately 1140 slices and 860 slices at the  $10.5 \mu\text{m}$  resolution, respectively.

**2.3. Three-dimensional volume reconstruction.** The three-dimensional volume reconstruction from a stack of slices consists of several steps. The first is to outline the object with contours in each slice image. For bone, with high contrast ratio relative to the surrounding soft tissue and air, contours are constructed semiautomatically. Once a contour that approximately matches the shape of the bone is hand-drawn, its shape is adapted to the nearest surface of the bone by a gauss segmentation algorithm. The algorithm is

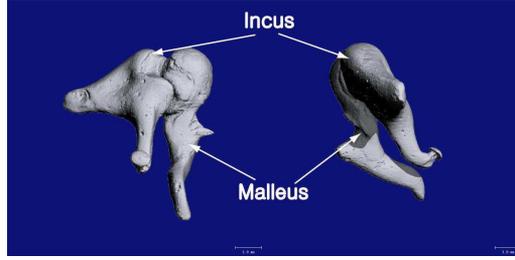


**Figure 3.** Top: slice image before segmentation. Bottom: segmented slice image of malleus (left) and incus (right). The different threshold values were applied for the low-density part (blood vessels) and high-density part (bone).

repeated until the region of interest (ROI) is judged to be adequately contoured. In essence, the contour *shrink wraps the ROI*. The contour is then copied to the next slice (iterating forward) or the previous slice (iterating backward), and the shrink wrapping procedure is repeated. Once the volume of interest (VOI) is separated from adjacent objects by contours, thresholds in grayscale are applied to identify *full voxel* and *empty voxel*, which correspond to the volume within threshold and out of threshold. Grayscale values in slice images make it easier to select the appropriate thresholds.

Bottom of [Figure 2](#) shows the grayscale values along the red line on top of [Figure 2](#), which were recalculated such that the maximum attenuation ( $\mu/\rho = 8 \text{ cm}^2/\text{g}$ ) and no attenuation ( $\mu/\rho = 0$ ) correspond to values of 1000 and 0. Grayscale values of 200–350 were set for the soft tissue range. A range above 350 is the range of bone and below 200 is the range for surrounding air.

[Figure 3](#) shows a slice image before (top) and after (bottom) contouring and applying threshold for the malleus (left) and incus (right) bones. After segmenting a stack of slices, they are combined to construct



**Figure 4.** Three-dimensional volume reconstruction of the malleus and incus bones (right ear).

the three-dimensional volume of the object. Figure 4 shows the reconstructed three-dimensional volume of the MIC bones.

**2.4. Calculation of inertial properties.** Portions of the malleus/incus bones are vascularized and thus contain lower-density blood vessels. Consequently, the entire bone cannot be treated as having uniform density; see bottom of Figure 3. The center of mass in the Cartesian coordinate system is calculated with the standard discretization

$$\bar{\mathbf{x}} \approx \frac{\sum_{i=1}^{N_L} \mathbf{x}_i \Delta m_L + \sum_{i=1}^{N_H} \mathbf{x}_i \Delta m_H}{\sum_{i=1}^{N_L} \Delta m_L + \sum_{i=1}^{N_H} \Delta m_H}, \quad (4)$$

while moments of inertia are calculated as

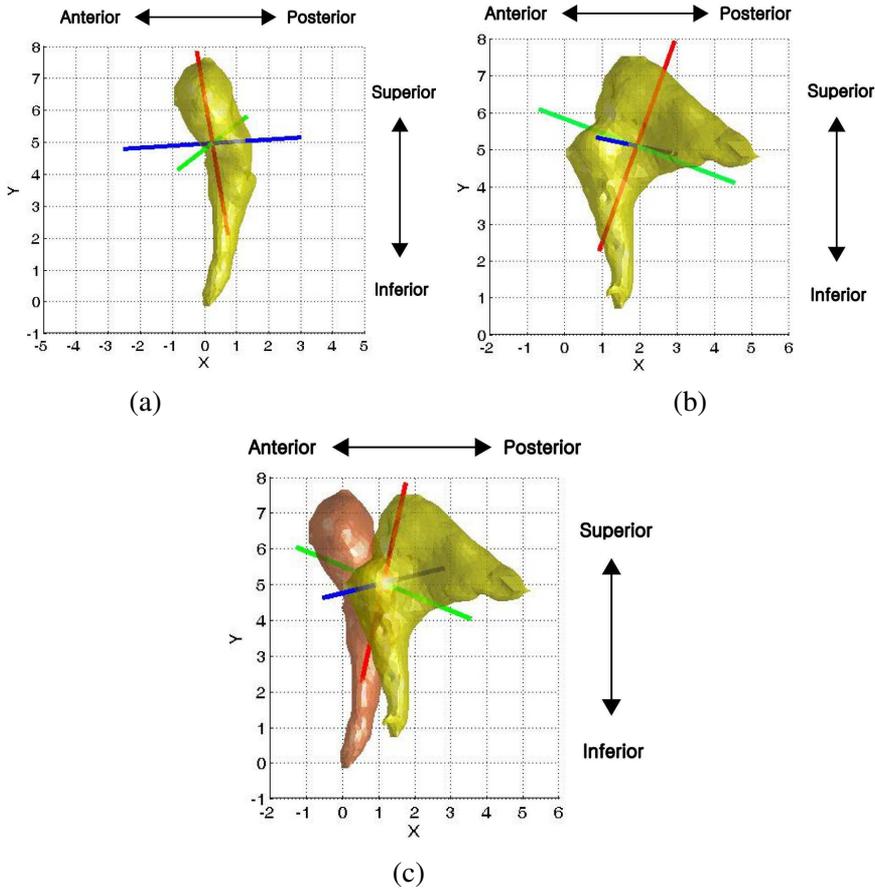
$$\begin{aligned} I_{xx} &\approx \sum_{i=1}^{N_L} (y_i^2 + z_i^2) \Delta m_L + \sum_{i=1}^{N_H} (y_i^2 + z_i^2) \Delta m_H, & I_{xy} &\approx - \sum_{i=1}^{N_L} x_i y_i \Delta m_L - \sum_{i=1}^{N_H} x_i y_i \Delta m_H, \\ I_{yy} &\approx \sum_{i=1}^{N_L} (x_i^2 + z_i^2) \Delta m_L + \sum_{i=1}^{N_H} (x_i^2 + z_i^2) \Delta m_H, & I_{yz} &\approx - \sum_{i=1}^{N_L} y_i z_i \Delta m_L - \sum_{i=1}^{N_H} y_i z_i \Delta m_H, \\ I_{zz} &\approx \sum_{i=1}^{N_L} (x_i^2 + y_i^2) \Delta m_L + \sum_{i=1}^{N_H} (x_i^2 + y_i^2) \Delta m_H, & I_{xz} &\approx - \sum_{i=1}^{N_L} x_i z_i \Delta m_L - \sum_{i=1}^{N_H} x_i z_i \Delta m_H. \end{aligned} \quad (5)$$

In the above equations,  $\Delta m_L$  is the mass of a *lower-density* voxel and  $\Delta m_H$  is the mass of a *higher-density* voxel. These can be calculated from the physically measured bone mass  $M$  and the number of lower-density and high-density voxels  $N_L$ ,  $N_H$ , respectively, with the assumption that the lower-density value  $\rho_L$  is just that of water

$$\Delta m_L = \rho_L \Delta v, \quad \Delta m_H = \rho_H \Delta v = \frac{M - N_L \Delta m_L}{N_H}, \quad (6)$$

where  $\rho_H$  indicates the higher-density value and  $\Delta v$  the volume of a single voxel.

Once moments of inertia are known in a given frame, the orientation of a second frame is calculated such that all products of inertia, that is, nondiagonal terms in inertia matrix given by the right side of Equation (5), are zero simultaneously. The principal directions of the second frame and corresponding



**Figure 5.** Principal axes of (a) malleus, (b) incus, and (c) MIC of (right) Ear 1. Red line denotes principal axis with minimum principal moment of inertia; blue line denotes maximum.

three principal moments of inertia are calculated from the eigenvalue problem as

$$[I]\{\omega\} = \alpha\{\omega\}, \tag{7}$$

where the three eigenvectors  $\{\omega\}$  provide the directions of the principal axes, and the three eigenvalues  $\alpha$  the corresponding principal moments of inertia.

### 3. Results

Figure 5 shows the principal axes of the malleus, incus, and MIC for Ear 1. In this figure, the intersection of the three axes is at the center of mass. The principal axis with the minimum moment of inertia is in nearly the same direction for the malleus, incus and the MIC (red lines in Figure 5), while the direction of the principal axis with the maximum moment of inertia is different (blue lines in Figure 5). The minimum moment of inertia occurs in the superior-inferior direction for the malleus, incus, and MIC.

The malleus has the maximum moment of inertia in the anterior-posterior direction, while incus and the MIC have the maximum moment of inertia in the medial-lateral direction.

Table 1 shows the tabulated mass, density, and the principal inertia measured and calculated from the three-dimensional volume of micro-CT images. Lower-density material within bone, which was measured to consist of 3 to 14% portion of the entire volume, make a relatively small contribution to the dynamic mechanical properties compared to the material of higher density.

Ear 3 has the largest mass and volume, while Ear 2 has the largest density for the malleus and the incus among three specimens. The malleus average density of  $2.39 \text{ mg/mm}^3$  is higher than the incus average density of  $2.15 \text{ mg/mm}^3$  by 11%. The difference between the malleus and incus density is distinguishably large for Ear 2, and only Ear 2 has a heavier malleus than incus.

The principal inertial values of the malleus are consistent with the particular feature of the malleus that length is large compared to the cross section dimensions, by more than a factor of 2. The malleus' moment of inertia along the principal axis of the superior-inferior direction of  $17.3 \pm 2.3 \text{ mg}\cdot\text{mm}^2$  is much smaller than the other two principal moments of inertia, which are similar, namely,  $106.1 \pm 10.9 \text{ mg}\cdot\text{mm}^2$  and  $100.6 \pm 10.1 \text{ mg}\cdot\text{mm}^2$ . The values of the principal moments of inertia of MIC are similar for Ears 1 and 2, while for Ear 3, which has much heavier bones than the other ears, these values are much larger, specifically 50% larger for the lateral-medial direction. Ear 1 has smaller principal moments of inertia for the malleus, but larger principal moments of inertia for the incus than Ear 2. Even though three ear samples showed a large diversity in the values of the principal moments of inertia, the ratio of the maximum moment of inertia to the minimum moment of inertia in MIC was about 2 for all three ear samples.

#### 4. Conclusion and discussion

Following previous work of [Decraemer et al. \[2003\]](#) and [Lane et al. \[2004; 2005\]](#), the micro-CT is found to be advantageous for the nondestructive investigation of the middle ear. The procedure for the determination of quantitative geometric and mechanical properties appears to be accurate. Inertial properties of the malleus-incus complex showed significant differences in three ear samples, and also some differences when compared to values found by other procedures by other authors [[Kirikae 1960](#); [Beer et al. 1996](#); [Weistenhöfer and Hudde 1999](#)]. For the densities of the malleus and the incus we obtained  $2.39 \pm 0.16 \text{ mg/mm}^3$  and  $2.15 \pm 0.07 \text{ mg/mm}^3$ , respectively, while [Kirikae \[1960\]](#) reported  $2.27\text{--}4.02 \text{ mg/mm}^3$  and  $1.48 \text{ mg/mm}^3$  as the corresponding values. Our principal inertial values  $132.5 \pm 18.5 \text{ mg}\cdot\text{mm}^2$ ,  $174.5 \pm 21.1 \text{ mg}\cdot\text{mm}^2$ , and  $259.4 \pm 34.2 \text{ mg}\cdot\text{mm}^2$  of the MIC were slightly larger than the corresponding values  $97.6 \text{ mg}\cdot\text{mm}^2$ ,  $165.0 \text{ mg}\cdot\text{mm}^2$ , and  $217.4 \text{ mg}\cdot\text{mm}^2$  obtained by [Weistenhöfer and Hudde \[1999\]](#). However, the present results are based on just three ears, so it appears likely that the present and previous values may be correct and indicate the actual variation that occurs in normal ears.

In ongoing work, the dynamic response of the middle ear bones is measured under various conditions with the objective of a better determination of the stiffness properties of ligament attachments. The results from the optimization procedure are sufficiently sensitive that it is important to have the correct inertial properties as input. The simple model for the middle ear consists of a rigid lever rotating about a fixed axis, to represent the ossicular chain, and a rigid piston to represent the eardrum.

Bone	Properties	Ear 1	Ear 2	Ear 3	Mean	SEM	
Malleus	Mass	25.9	29.8	35.1	30.3	2.7	
	Density	2.14	2.68	2.35	2.39	0.16	
	Principal moments of inertia	$n_{AP}^M$	88.3	103.9	126.0	106.1	10.9
		$n_{SI}^M$	13.6	16.7	21.5	17.3	2.3
$n_{LM}^M$		83.9	98.9	118.9	100.6	10.1	
Incus	Mass	29.4	27.8	38.7	32.0	3.4	
	Density	2.02	2.23	2.21	2.15	0.07	
	Principal moments of inertia	$n_{AP}^I$	57.4	48.6	72.6	59.5	7.0
		$n_{SI}^I$	31.8	25.5	48.6	35.3	6.9
$n_{LM}^I$		79.1	66.2	107.7	84.3	12.3	
MIC	Mass	55.3	57.6	73.8	62.2	5.8	
	Density	2.07	2.44	2.27	2.26	0.11	
	Principal moments of inertia	$n_{AP}^{MI}$	149.0	158.2	216.4	174.5	21.1
		$n_{SI}^{MI}$	114.9	113.2	169.4	132.5	18.5
$n_{LM}^{MI}$		223.9	226.4	327.8	259.4	34.2	

**Table 1.** Mass (in mg), density (in mg/mm<sup>3</sup>) and principal moments of inertia (in mg·mm<sup>2</sup>). SEM stands for standard error of mean.  $n_{AP}$ ,  $n_{SI}$ ,  $n_{LM}$  denote principal axes in the anterior-posterior, superior-inferior, and lateral-medial directions.

From many measurements and theoretical considerations, it is clear that such a model loses all credibility for frequencies above about 1 kHz. In particular, the ossicular chain has many modes of motion for high frequencies [Eiber and Freitag 2002]. It remains a puzzle how an efficient transfer of acoustic energy takes place with such modes. The present results provide necessary parameters for the analysis of the motion through the audio frequency range and the possibility for an answer to the puzzle.

## References

- [Beer et al. 1996] H. J. Beer, M. Borniz, J. Drescher, R. Schmidt, H. J. Hardtke, G. Hofmann, U. Vogel, T. Zahnert, and K. B. Hüttenbrink, “Finite element modeling of the human eardrum and application”, pp. 40–47 in *Proceeding of the international workshop on MEMRO*, 1996.
- [Brooks and Di Chiro 1976] R. A. Brooks and G. Di Chiro, “Beam hardening in x-ray reconstructive tomography”, *Phys. Med. Biol.* **21**:3 (1976), 390–398.
- [Decraemer et al. 2003] W. F. Decraemer, J. J. J. Dirckx, and W. R. Funnell, “Three-dimensional modelling of the middle-ear ossicular chain using a commercial high-resolution X-ray CT scanner”, *J. Assoc. Res. Otolaryngol.* **4**:2 (2003), 250–263.
- [Eiber and Freitag 2002] A. Eiber and H.-G. Freitag, “On simulation models in otology”, *Multibody Syst. Dyn.* **8**:2 (2002), 197–217.

- [Gan et al. 2002] R. Z. Gan, Q. Sun, R. K. J. Dyer, K.-H. Chang, and K. J. Dormer, “Three-dimensional modeling of middle ear biomechanics and its applications”, *Otol. Neurotol.* **23**:3 (2002), 271–280.
- [Johns and Cunningham 1974] H. E. J. Johns and J. R. Cunningham, *The physics of radiology*, 3rd ed., Thomas, Springfield, IL, 1974.
- [Ketcham and Carlson 2001] R. A. Ketcham and W. D. Carlson, “Acquisition, optimization and interpretation of x-ray computed tomographic imagery: applications to the geosciences”, *Comput. Geosci.* **27**:4 (2001), 381–400.
- [Kirikae 1960] J. Kirikae, *The middle ear*, University of Tokyo Press, Tokyo, 1960.
- [Koike et al. 2002] T. Koike, H. Wada, and T. Kobayashi, “Modeling of the human middle ear using the finite-element method”, *J. Acoust. Soc. Am.* **111**:3 (2002), 1306–1317.
- [Lane et al. 2004] J. I. Lane, R. J. Witte, C. L. W. Driscoll, J. J. Camp, and R. A. Robb, “Imaging microscopy of the middle and inner ear, I: CT microscopy”, *Clin. Anat.* **17**:8 (2004), 607–612.
- [Lane et al. 2005] J. I. Lane, R. J. Witte, O. W. Henson, C. L. W. Driscoll, J. Camp, and R. A. Robb, “Imaging microscopy of the middle and inner ear, II: MR microscopy”, *Clin. Anat.* **18**:6 (2005), 409–415.
- [Macovski 1983] A. Macovski, *Medical imaging systems*, Prentice-Hall, Upper Saddle River, NJ, 1983.
- [Wang and Vannier 1998] G. Wang and M. W. Vannier, “Computerized tomography”, in *Encyclopedia of Electrical and Electronics Engineering*, edited by J. G. Webster, John Wiley & Sons, New York, 1998.
- [Wang et al. 1996] G. Wang, D. L. Snyder, J. A. O’Sullivan, and M. W. Vannier, “Iterative deblurring for CT metal artifact reduction”, *IEEE T. Med. Imaging* **15**:5 (1996), 657–664.
- [Weistenhöfer and Hudde 1999] C. Weistenhöfer and H. Hudde, “Determination of the shape and inertia properties of the human auditory ossicles”, *Audiol. Neuro-Otol.* **4**:3-4 (1999), 192–196.

Received 18 Jul 2006. Revised 29 Mar 2007. Accepted 20 Apr 2007.

JAE HOON SIM: [jhsim@stanford.edu](mailto:jhsim@stanford.edu)

*Mechanics and Computation Division, Department of Mechanical Engineering, Stanford University, 496 Lomita Mall, Durand Building, Stanford, CA 94305-4035, United States*

and

*Palo Alto Veterans Administration, 3801 Miranda Avenue, Palo Alto, CA 94304, United States*

SUNIL PURIA: [puria@stanford.edu](mailto:puria@stanford.edu)

*Mechanics and Computation Division, Department of Mechanical Engineering, Stanford University, 496 Lomita Mall, Durand Building, Stanford, CA 94305-4035, United States*

and

*Department of Otolaryngology — Head and Neck Surgery, Stanford University, Stanford, CA 94305, United States*

and

*Palo Alto Veterans Administration, 3801 Miranda Avenue, Palo Alto, CA 94304, United States*

CHARLES R. STEELE: [chasst@stanford.edu](mailto:chasst@stanford.edu)

*Mechanics and Computation Division, Stanford University, 496 Lomita Mall, Durand Building, Stanford, CA 94305-4035, United States*

# ELECTROELASTIC INTENSIFICATION AND DOMAIN SWITCHING NEAR A PLANE STRAIN CRACK IN A RECTANGULAR PIEZOELECTRIC MATERIAL

YASUhide SHINDO, FUMIO NARITA AND FUMITOSHI SAITO

We study the effects of crack face boundary conditions and localized polarization switching on the piezoelectric fracture. This paper consists of two parts. In the first part, the electroelastic problem of an infinite piezoelectric material with a crack is formulated by means of integral transforms, and the exact solution is obtained. The electroelastic fields are expressed in closed form. The fracture mechanics parameters, such as energy release rate, are obtained for the permeable, impermeable and open crack models. In the second part, finite element analysis is carried out to study the crack behavior in a rectangular piezoelectric material by introducing a model for polarization switching in local areas of electroelastic field concentrations. A nonlinear behavior induced by localized polarization switching is observed between the fracture mechanics parameters and applied electric field.

## 1. Introduction

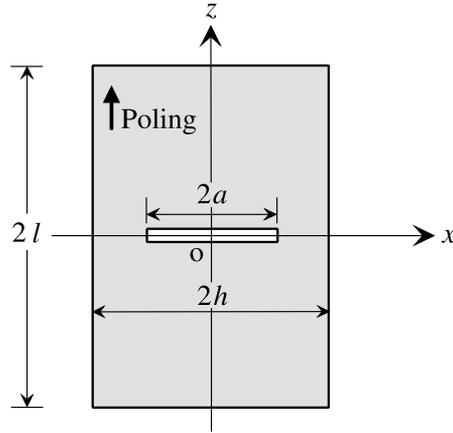
The fracture behavior of piezoelectric materials has received much attention in recent years. In the theoretical studies of the piezoelectric crack problems, there are two commonly used electrical boundary conditions across the crack face: (1) the permeable crack model and (2) the impermeable crack model. Theoretical analyses on cracked piezoelectric ceramics indicated that a negative energy release rate is produced for the impermeable crack model [Narita et al. 2003]. Furthermore, some experimental results show that the fracture loads are increased or decreased depending on the mechanical loading conditions (applied load or applied displacement) and direction of electric fields [Park and Sun 1995; Shindo et al. 2002; Narita et al. 2003; Shindo et al. 2005]. These experimentally observed phenomena contradict the results of the calculations using energy release rate for the impermeable crack model. Recently, some researchers [Xu and Rajapakse 2001; Wang and Mai 2003; Landis 2004; McMeeking 2004] used the open piezoelectric crack model [Hao and Shen 1994] and discussed the effect of electric fields on the fracture mechanics parameters such as energy release rate. Although the impermeable and open crack models may provide mathematical solutions of piezoelectric cracks, there is still a great deal of uncertainty in searching for fracture design parameters characterizing the electric failure.

The nonlinear effect caused by the polarization switching may affect the piezoelectric fracture behavior [Fu and Zhang 2000; Shindo et al. 2003]. In this investigation, the effects of crack face boundary conditions and localized polarization switching near the crack tip on the piezoelectric fracture mechanics parameters are studied by analyzing the plane strain electroelastic problem of a piezoelectric material with a crack. First, the crack problem of an infinite piezoelectric material is formulated by means of

---

*Keywords:* elasticity, finite element method, piezoelectric material, crack, energy release rate, polarization switching.

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan under the Grant-in-Aid for Scientific Research (B).



**Figure 1.** A rectangular piezoelectric material with a crack.

integral transforms and the solutions are obtained exactly. Electroelastic fields and energy release rate based on permeable, impermeable and open crack models are compared. Secondly, a finite element method incorporating the polarization switching mechanism is used to calculate the energy release rate in a rectangular piezoelectric material. The numerical results illustrate that the impermeable and open crack models can lead to significant errors regarding the effect of electric fields on piezoelectric crack propagation.

## 2. Statement of the problem and basic equations

A rectangular piezoelectric material of length  $2l$  and width  $2h$  contains a central crack of length  $2a$ , as shown in Figure 1. A set of Cartesian coordinates  $\{x, y, z\}$  is attached to the center of the crack normal to the  $z$ -axis. The piezoelectric material has symmetry properties of hexagonal crystal of 6 mm class with respect to the  $x, y, z$ -axes, and is under a state of plane strain. The material is loaded by mechanical displacement  $u_0$  with the electric field in the  $z$ -direction of the poling axis. Due to the symmetry of the problem, only the first quadrant with appropriate boundary conditions needs to be analyzed.

The constitutive equations can be written as

$$\sigma_{xx} = c_{11}u_{x,x} + c_{13}u_{z,z} - e_{31}E_z, \quad \sigma_{zx} = c_{44}(u_{x,z} + u_{z,x}) - e_{15}E_x, \quad \sigma_{zz} = c_{13}u_{x,x} + c_{33}u_{z,z} - e_{33}E_z, \quad (1)$$

$$D_x = e_{15}(u_{x,z} + u_{z,x}) + \epsilon_{11}E_x, \quad D_z = e_{31}u_{x,x} + e_{33}u_{z,z} + \epsilon_{33}E_z. \quad (2)$$

Here  $\sigma_{xx}, \sigma_{zz}, \sigma_{xz} = \sigma_{zx}$  are the components of stress tensor,  $D_x$  and  $D_z$  are the components of electric displacement vector,  $u_x$  and  $u_z$  are the components of displacement vector,  $E_x$  and  $E_z$  are the components of electric field intensity vector;  $c_{11}, c_{13}, c_{33}, c_{44}$  are the elastic stiffness constants measured in a constant electric field,  $\epsilon_{11}, \epsilon_{33}$  are the dielectric permittivities measured at constant strain, and  $e_{15}, e_{31}, e_{33}$  are the piezoelectric constants. A comma implies partial differentiation with respect to the coordinates. The electric field components are related to the electric potential  $\phi(x, z)$  via  $E_x = -\phi_{,x}$  and  $E_z = -\phi_{,z}$ . The

governing equations can be written as

$$\begin{aligned} c_{11}u_{x,xx} + c_{44}u_{x,zz} + (c_{13} + c_{44})u_{z,xz} + (e_{31} + e_{15})\phi_{,xz} &= 0, \\ c_{44}u_{z,xx} + c_{33}u_{z,zz} + (c_{13} + c_{44})u_{x,xz} + e_{15}\phi_{,xx} + e_{33}\phi_{,zz} &= 0, \\ (e_{31} + e_{15})u_{x,xz} + e_{15}u_{z,xx} + e_{33}u_{z,zz} - \epsilon_{11}\phi_{,xx} - \epsilon_{33}\phi_{,zz} &= 0. \end{aligned} \tag{3}$$

In a vacuum, the constitutive equations (2) and the governing equation (3)<sub>3</sub> become

$$D_x = \epsilon_0 E_x, \quad D_z = \epsilon_0 E_z, \quad \phi_{,xx} + \phi_{,zz} = 0, \tag{4}$$

where  $\epsilon_0 = 8.85 \times 10^{-12}$  C/Vm is the electric permittivity of the vacuum.

The crack face boundary and the loading conditions can be expressed in the form

$$\sigma_{zx}(x, 0) = 0 \quad (0 \leq x \leq h), \quad \sigma_{zz}(x, 0) = 0 \quad (0 \leq x < a), \quad u_z(x, 0) = 0 \quad (a \leq x \leq h) \tag{5}$$

$$E_x(x, 0) = E_x^c(x, 0) \quad (0 \leq x < a), \quad \phi(x, 0) = 0 \quad (a \leq x \leq h), \tag{6}$$

$$D_z(x, 0) = D_z^c(x, 0) \quad (0 \leq x < a), \tag{7}$$

$$u_z(x, l) = u_0, \quad (0 \leq x \leq h), \quad \phi(x, l) = \phi_0 \quad (0 \leq x \leq h). \tag{8}$$

where  $\phi_0$  is an applied electric potential and the superscript *c* stands for the electric field quantity in the void inside the crack. The electric potential is zero on the symmetry planes inside the crack and ahead of the crack, so the boundary conditions (6) reduce to  $\phi(x, 0) = 0$  for  $0 \leq x \leq h$ . Equations (6) and (7) are the permeable boundary conditions.

Applying the loading conditions (8), the stress  $\sigma_{zz}$  for the uncracked piezoelectric material is

$$\sigma_{zz}(x, z) = \sigma_0 - e_1 E_0, \quad \sigma_0 = \left( c_{33} - \frac{c_{13}^2}{c_{11}} \right) \frac{u_0}{l}, \quad E_0 = -\frac{\phi_0}{l}, \quad e_1 = e_{33} - \left( \frac{c_{13}}{c_{11}} \right) e_{31}. \tag{9}$$

The stress at  $z = l$  for the uncracked piezoelectric material is denoted by  $\sigma_l = \sigma_0 - e_1 E_0$ . Note that  $\sigma_0$  is the stress for a closed-circuit condition with the potential forced to remain zero (grounded) and depends only on the displacement at the edge  $z = l$ . When a uniform displacement  $u_0$  is applied and fixed at  $z = l$ , the stress  $\sigma_0$  will be uniform. On the other hand, when the stress  $\sigma_l$  is applied and fixed at  $z = l$ ,  $\sigma_l$  is left unchanged and the displacement  $u_0$  depends on  $E_0$ .

### 3. Cracked infinite piezoelectric material

In this section we consider the problem of an infinite piezoelectric material with a crack for  $l \rightarrow \infty$  and  $h \rightarrow \infty$ . The material is under applied uniform strain  $\epsilon_0$  and electric field  $E_0$  at infinity. The stress at infinity is denoted by  $\sigma_l = \sigma_0 - e_1 E_0$ , and Equation (9)<sub>2</sub> can be rewritten in terms of  $\epsilon_0$  as  $\sigma_0 = (c_{33} - c_{13}^2/c_{11})\epsilon_0$ . Fourier transforms are used to reduce the mixed boundary value problem to a pair of dual integral equations. The integral equations then can be solved exactly; see Appendix A. The energy release rate  $G$  for the permeable crack model may be expressed as

$$G = \frac{1}{2F^2} \left( -F \sum_{j=1}^3 \frac{d_j}{\gamma_j} + \sum_{k=1}^3 h_k d_k \sum_{j=1}^3 \frac{b_j d_j}{\gamma_j} \right) K_I^2, \tag{10}$$

where the stress intensity factor  $K_I$  is defined as  $K_I = \lim_{x \rightarrow a^+} [2\pi(x - a)]^{1/2} \sigma_{zz}(x, 0)$ . The stress intensity factor under applied strain and applied stress is given by, respectively,

$$K_I = \left\{ \left( c_{33} - \frac{c_{13}^2}{c_{11}} \right) \varepsilon_0 - e_1 E_0 \right\} (\pi a)^{1/2}, \quad K_I = \sigma_I (\pi a)^{1/2}. \quad (11)$$

Energy release rates for the impermeable and open crack models are discussed in [Appendices B and C](#), respectively.

#### 4. Cracked rectangular piezoelectric material

In this section the finite element computer program ANSYS is selected for the analysis of the configuration considered here. A nonlinear finite element model incorporating the polarization switching mechanisms with the energy release rate calculations is developed. Two criteria are used for this purpose: work done switching criterion, and internal energy density switching criterion.

The first criterion requires that a polarization switches when the combined electrical and mechanical work exceeds a critical value [\[Hwang et al. 1995\]](#)

$$\sigma_{xx} \Delta \varepsilon_{xx} + \sigma_{zz} \Delta \varepsilon_{zz} + 2\sigma_{zx} \Delta \varepsilon_{zx} + E_x \Delta P_x + E_z \Delta P_z = 2P^s E_c, \quad (12)$$

where  $\Delta \varepsilon_{xx}$ ,  $\Delta \varepsilon_{zz}$ ,  $\Delta \varepsilon_{zx}$  are the changes in the spontaneous strain  $\gamma^s$ ,  $\Delta P_x$ ,  $\Delta P_z$  are the changes in the spontaneous polarization  $P^s$ , and  $E_c$  is a coercive electric field. It is assumed that elastic and dielectric constants of the piezoelectric materials remain unchanged after  $180^\circ$  or  $90^\circ$  polarization switching occurs and only piezoelectric constants vary with switching. It is also assumed that for  $90^\circ$  switching there are two allowable directions of the poling in the coordinate system: in the positive and negative  $x$ -direction. The changes in spontaneous strains and polarizations for  $180^\circ$  switching can be expressed as

$$\Delta \varepsilon_{xx} = \Delta \varepsilon_{zz} = \Delta \varepsilon_{zx} = 0, \quad \Delta P_x = 0, \quad \Delta P_z = -2P^s.$$

For  $90^\circ$  switching in the  $xz$  plane, we have

$$\Delta \varepsilon_{xx} = \gamma^s, \quad \Delta \varepsilon_{zz} = -\gamma^s, \quad \Delta \varepsilon_{zx} = 0, \quad \Delta P_x = \pm P^s, \quad \Delta P_z = -P^s.$$

The polarization switching criterion based on internal energy density is defined as [\[Sun and Achuthan 2004\]](#)

$$U = U_c, \quad (13)$$

where  $U$  is the internal energy density and  $U_c$  is its critical value corresponding to the switching mode. The internal energy density associated with  $180^\circ$  and  $90^\circ$  switching, respectively, is

$$U = \frac{1}{2} D_z E_z, \quad U = \frac{1}{2} (\sigma_{xx} \varepsilon_{xx} + \sigma_{zz} \varepsilon_{zz} + 2\sigma_{zx} \varepsilon_{zx} + D_x E_x).$$

We assume that the critical value of internal energy density takes the form  $U_c = \frac{1}{2} \epsilon_{33}^T (E_c)^2$ , where  $\epsilon_{33}^T$  is the dielectric permittivity at constant stress.

Due to polarization switching, piezoelectric materials are often nonhomogeneous. The piezoelectric properties vary from one location to the other, and the variations are either continuous or discontinuous.

The energy release rate  $G$  can be obtained from the following crack tip integral [Shindo et al. 2005]:

$$G = \left( \int_{\Gamma_0} - \int_{\Gamma_p} \right) \{ H n_x - (\sigma_{xx} u_{x,x} + \sigma_{zx} u_{z,x}) n_x - (\sigma_{zx} u_{x,x} + \sigma_{zz} u_{z,x}) n_z + D_x E_x n_x + D_z E_x n_z \} d\Gamma,$$

where  $\Gamma_0$  is a small contour closing a crack tip,  $\Gamma_p$  is a path embracing that part of phase boundary which is enclosed by  $\Gamma_0$ , and  $n_x, n_z$  are the components of the outer unit normal vector. The electrical enthalpy density  $H$  is

$$H(u_x, u_z, E_x, E_z) = \frac{1}{2} (c_{11} u_{x,x}^2 + c_{33} u_{z,z}^2 + 2c_{13} u_{x,x} u_{z,z} + c_{44} (u_{x,z} + u_{z,x})^2) - \left[ \frac{1}{2} (\epsilon_{11} E_x^2 + \epsilon_{33} E_z^2) + e_{15} (u_{x,z} + u_{z,x}) E_x + (e_{31} u_{x,x} + e_{33} u_{z,z}) E_z \right].$$

Each element consists of many grains, and each grain is modeled as a uniformly polarized cell that contains a single domain. The model neglects the domain wall effects and interaction among different domains. In reality these effects matter, but the assumption does not affect the general conclusions drawn. The polarization of each grain initially aligns as closely as possible to the  $z$ -direction. Polarization switching is defined for each element in the material. The displacement  $u_0$  and electric potential  $\phi_0$  are applied at the edge  $0 \leq x \leq h, z = l$ , and the electroelastic fields of each element are computed from the finite element analysis. The switching criterion (12) or (13) is checked for every element to see if switching will occur. After all possible polarization switches have occurred, the piezoelectric tensor of each element is rotated to the new polarization direction. The electroelastic fields are recalculated, and the process is repeated until the solution converges. The macroscopic response of the material is determined by the finite element model, which is an aggregate of elements. The spontaneous polarization  $P^s$  and strain  $\gamma^s$  are assigned representative values of  $0.3 \text{ C/m}^2$  and  $0.004$ , respectively. Our previous experiments [Yoshida et al. 2003; Shindo et al. 2004; Narita et al. 2005] verified the accuracy of the above scheme, and showed that the results obtained are of general applicability. After polarization switching is predicted, J-integral paths are selected, which do not pass exactly through the singular point.

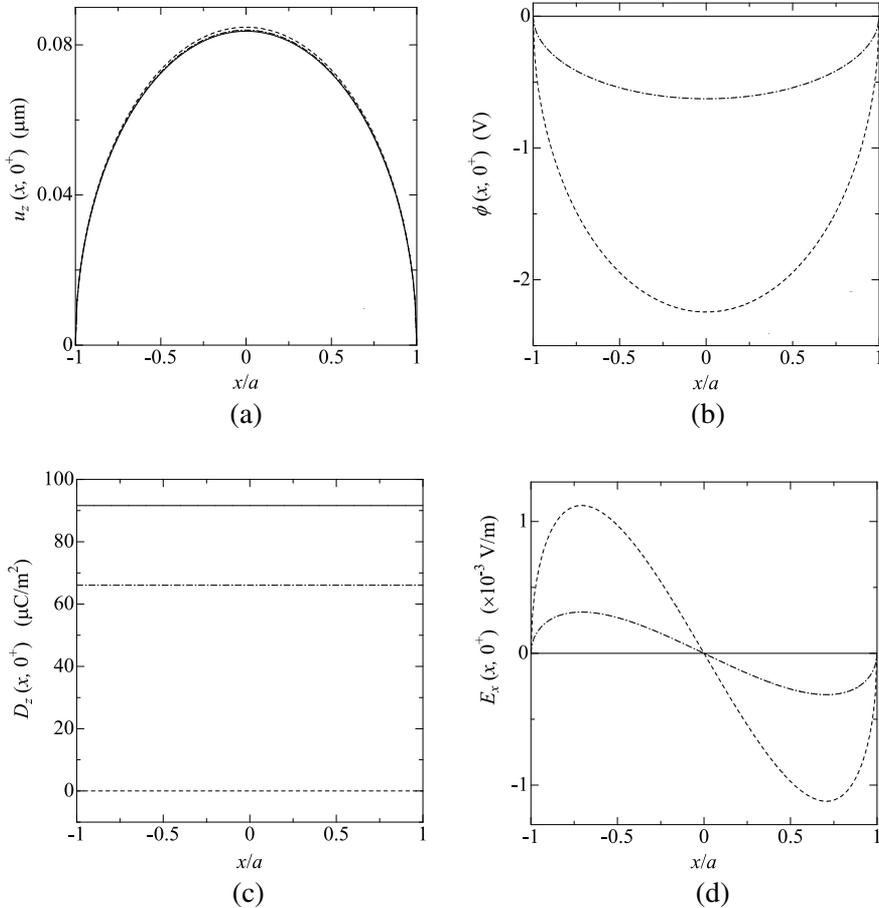
The calculations of the electroelastic fields and energy release rate for the open crack model are more complicated than for the permeable and impermeable crack models. The open crack model calculations start with  $\phi = 0$  on the crack surface [McMeeking 1999]. The crack opening displacement and charge density on the crack surface are estimated, and the resulting potential difference is applied to the crack surface. The electroelastic fields are again solved leading to new crack opening displacement and charge density on the crack surface. If this is accomplished, then the potential difference is applied once more to the crack surface. Such a procedure is repeated until the evolution of the objective solutions shows no improvements.

## 5. Numerical results and discussion

Numerical calculations have been carried out for commercially available piezoelectric ceramics C-91 (Fuji Ceramics, Japan). The material properties of C-91 are listed in Table 1, and the coercive electric field  $E_c$  is approximately  $0.35 \text{ MV/m}$ . Figure 2a shows the crack opening displacement  $u_z(x, 0^+)$  from the theoretical solutions for an infinite C-91 ( $l, h \rightarrow \infty$ ) with a crack of length  $2a = 2 \text{ mm}$  under  $\epsilon_0 = 5 \times 10^{-5}$  and  $E_0 = 0$ . The results for the permeable, open and impermeable crack models are shown

Elastic stiffnesses ( $\times 10^{10}$ N/m <sup>2</sup> )					Piezoelectric coefficients (C/m <sup>2</sup> )			Dielectric constants ( $\times 10^{-10}$ C/V m)	
$c_{11}$	$c_{12}$	$c_{13}$	$c_{33}$	$c_{44}$	$e_{31}$	$e_{33}$	$e_{15}$	$\epsilon_{11}$	$\epsilon_{33}$
12.0	7.7	7.7	11.4	2.4	-17.3	21.2	20.2	226	235

**Table 1.** Material properties of C-91.

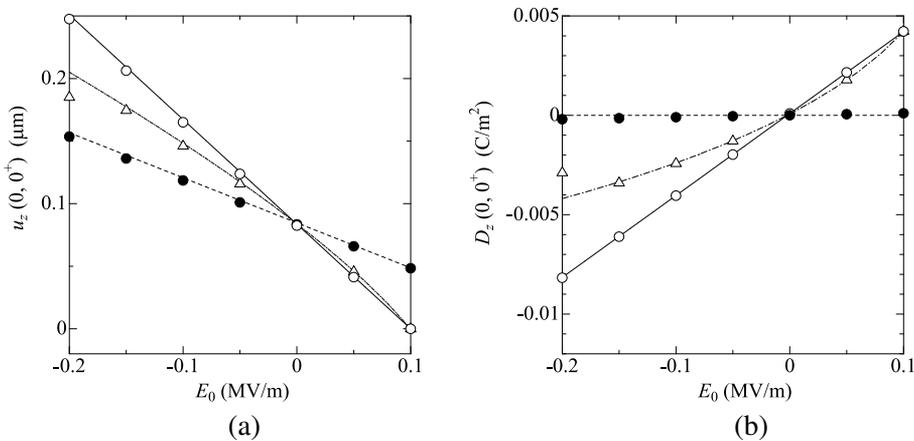


**Figure 2.** (a) Displacement  $u_z(x, 0^+)$ ; (b) electric potential  $\phi(x, 0^+)$ ; (c) normal component of electric displacement  $D_z(x, 0^+)$ ; and (d) tangential component of electric field  $E_x(x, 0^+)$  along the upper crack surface for an infinite piezoelectric material C-91 under uniform strain. Here  $l, h \rightarrow \infty$ ,  $a = 1$  mm,  $\epsilon_0 = 5 \times 10^{-5}$  and  $E_0 = 0$ . The permeable model is represented by the solid line, open by the dot-dashed line, and the impermeable by the dashed line.

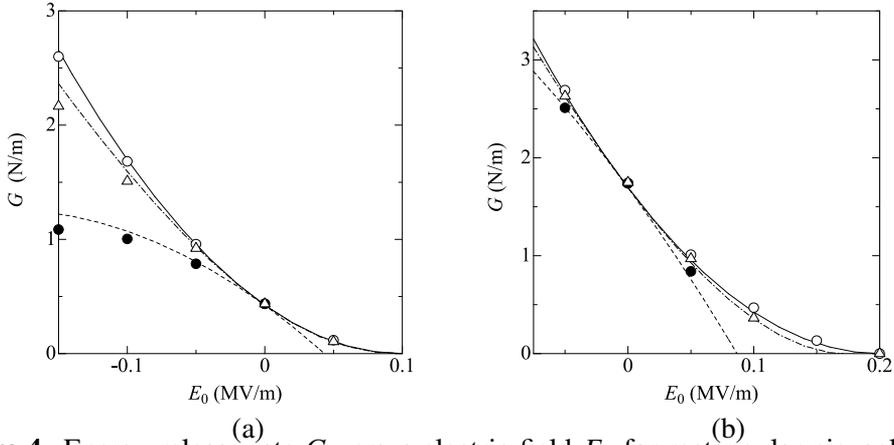
for comparison purposes. Little difference among three piezoelectric crack models is observed. The rest of Figure 2 shows the electric potential  $\phi(x, 0^+)$ , normal component of electric displacement  $D_z(x, 0^+)$  and tangential component of electric field  $E_x(x, 0^+)$  along the upper crack surface. There are differences among the crack models. It is noted that the open and impermeable crack models reduce the continuity of the tangential components of the electric field across the crack surface.

Figure 3a presents the crack opening displacement  $u_z(0, 0^+)$  at the center of the crack versus electric field  $E_0$  from the finite element analysis without the polarization switching effect. The rectangular piezoelectric material C-91 with a crack of length  $2a = 2$  mm has a length  $2l = 20$  mm and width  $2h = 20$  mm, and is under applied displacement  $u_0 = 0.5 \mu\text{m}$  corresponding to the uniform strain  $5 \times 10^{-5}$  for the uncracked material. For comparison, the results for the infinite piezoelectric material ( $l, h \rightarrow \infty, \epsilon_0 = 5 \times 10^{-5}$ ) obtained from the theoretical analysis are included. The results for the finite element analysis agree with the theoretical analysis data. Figure 3b shows similar results for the normal component of electric displacement  $D_z(0, 0^+)$ .

Figure 4a shows the dependence of the energy release rate  $G$  on  $E_0$ . The results for the infinite piezoelectric material obtained from the theoretical analysis are also shown. The energy release rates are lower for positive electric fields and higher for negative electric fields under applied displacement. In the impermeable case, a negative energy release rate is produced. The energy release rate for the permeable crack in the infinite piezoelectric material under applied stress is independent of the electric field (not shown). Figure 4b shows the similar results under  $u_0 = 1 \mu\text{m}$  with  $\epsilon_0 = 10^{-4}$ . A negative energy release rate is also produced for the open crack model. The parameters for the impermeable and open crack models have questionable physical significance.

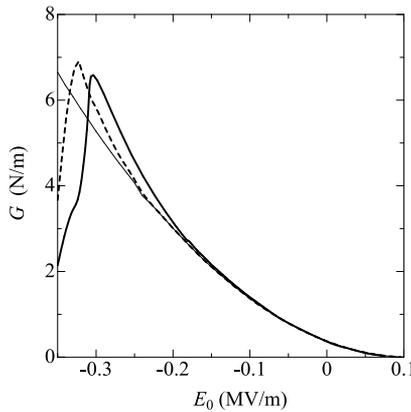


**Figure 3.** (a) Crack center displacement  $u_z(0, 0^+)$  and (b) normal component of electric displacement  $D_z(0, 0^+)$  versus electric field  $E_0$  for rectangular piezoelectric material C-91 under applied displacement for finite element analysis data  $l = h = 10$  mm, whereas the theoretical prediction is made with  $l, h \rightarrow \infty$ .  $a = 1$  mm,  $\epsilon_0 = 5 \times 10^{-5}$ , and  $u_0 = 0.5 \mu\text{m}$ . The permeable theory is represented by the solid line and data with open circles, open by the dot-dashed line and triangles, and the impermeable by the dashed line and solid circles.



**Figure 4.** Energy release rate  $G$  versus electric field  $E_0$  for rectangular piezoelectric material C-91 under applied displacement. For finite element analysis data  $l = h = 10$  mm, whereas the theoretical prediction is made with  $l, h \rightarrow \infty$ ; also  $a = 1$  mm. (a)  $\epsilon_0 = 5 \times 10^{-5}$  and  $u_0 = 0.5 \mu\text{m}$ , (b)  $\epsilon_0 = 10^{-4}$  and  $u_0 = 1 \mu\text{m}$ . For legend, see Figure 3.

Figure 5 displays the variation of  $G$  with electric field  $E_0$  for the permeable crack model from the finite element analysis with and without the polarization switching effect. For the polarization switching effect, the predictions by the criteria based on work (12) and energy density (13) are shown. The rectangular piezoelectric material C-91 ( $2l = 5$  mm,  $2h = 5$  mm) with a crack ( $2a = 2$  mm) is under applied displacement  $u_0 = 0.125 \mu\text{m}$  corresponding to the uniform strain  $5 \times 10^{-5}$  for the uncracked material. Positive electric fields decrease the values of  $G$ , while negative electric fields have an opposite effect. A monotonically increasing negative  $E_0$  causes polarization switching. The value of electric field associated with the switching is  $-0.25$  MV/m for the work-based criterion, while it is approximately  $-0.17$  MV/m for

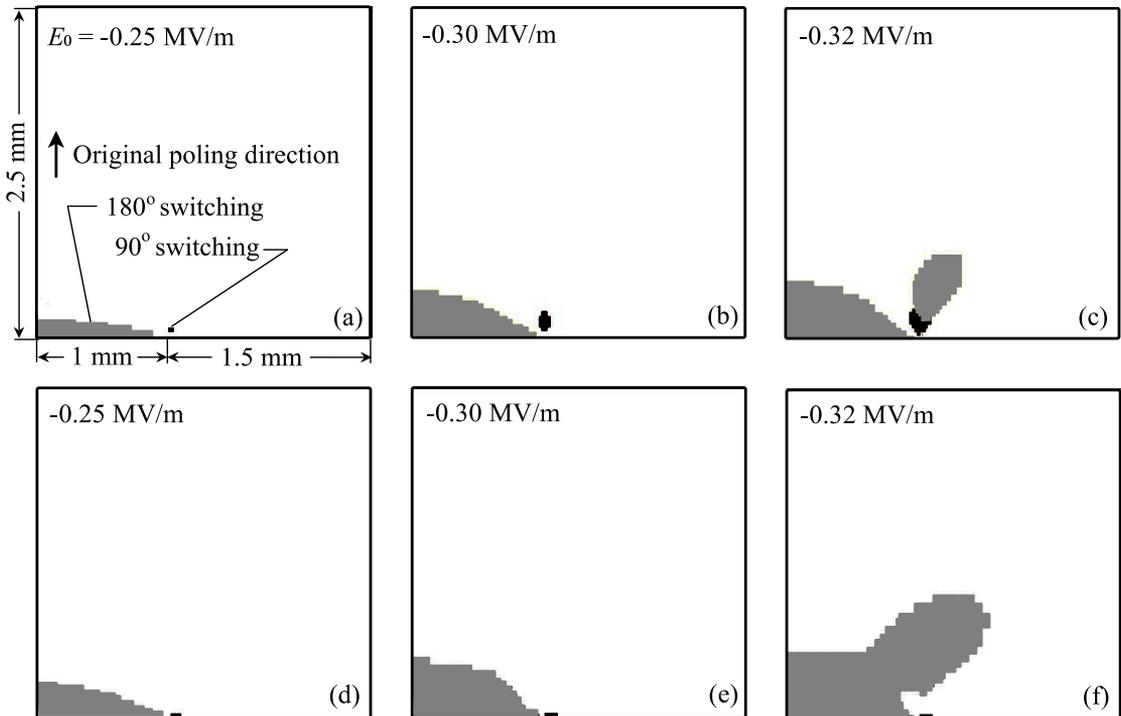


**Figure 5.** Energy release rate  $G$  versus electric field  $E_0$  for rectangular piezoelectric material C-91 under applied displacement in the permeable model.  $l = h = 2.5$  mm,  $a = 1$  mm, and  $u_0 = 0.125 \mu\text{m}$ . Thin line gives prediction without polarization switching; the dashed line gives work-based and thick line gives energy density-based switching effect.

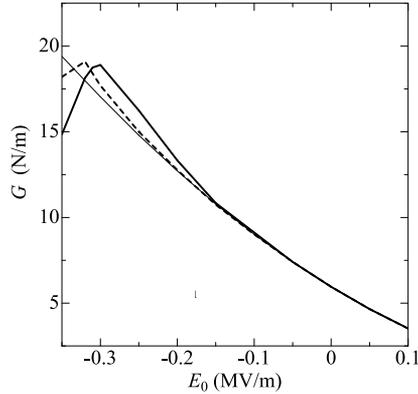
the criterion based on the energy density. When the negative  $E_0$  increases further,  $G$  with the polarization switching effect becomes larger than that without the switching effect. After  $E_0$  reaches about  $-0.325$  ( $-0.305$ ) MV/m, polarization switching in a local region, based on the work (energy density), leads to an unexpected decrease in  $G$  for the permeable crack. Our previous experimental study [Shindo et al. 2003] showed a significant nonlinearity in the fracture load due to polarization switching. The nonlinear effect caused by polarization switching may affect the piezoelectric crack behavior.

Figure 6 shows the  $180^\circ$  and  $90^\circ$  switching zones near the permeable crack tip in the rectangular piezoelectric material C-91 ( $2l = 5$  mm,  $2h = 5$  mm,  $2a = 2$  mm) under  $u_0 = 0.125$   $\mu\text{m}$  for various values of  $E_0$ . Predictions resulting from different criteria are presented. The size of the  $180^\circ$  ( $90^\circ$ ) switching zone behind (ahead of) the crack tip increases at first when the negative  $E_0$  is increased, and the difference between energy release rate results with and without switching effect becomes larger at a higher negative  $E_0$ . As the negative  $E_0$  continues increasing, the area of the  $180^\circ$  switching zone grows ahead of the crack tip. Unexpected decrease in  $G$  is attributed to  $180^\circ$  switching ahead of the crack tip. In the impermeable case, the region ahead of the crack tip is found to undergo  $180^\circ$  switching due to the large negative electric field, and the region behind the crack tip has  $90^\circ$  switching because of the large intensified electric field  $E_x$  [Kalyanam and Sun 2005].

The applied displacement may enhance the polarization switching depending on its magnitude. The critical value of the electric field associated with the polarization switching decreases (relative to  $u_0 =$

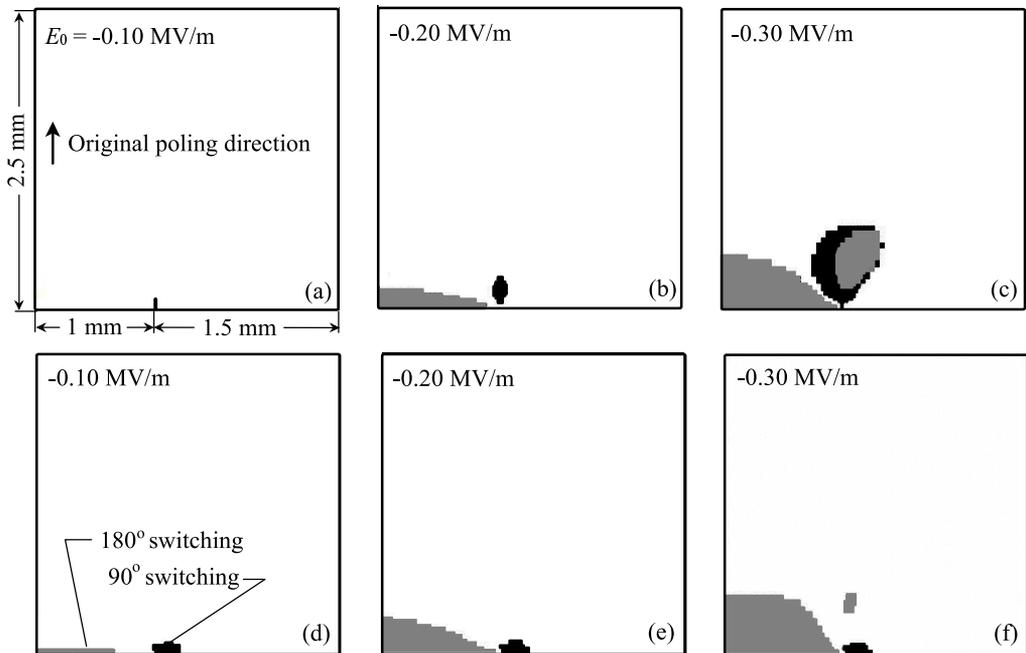


**Figure 6.** Polarization switching zone induced by displacement  $u_0 = 0.125$   $\mu\text{m}$  and electric field  $E_0$  of (a, d)  $-0.25$ , (b, e)  $-0.30$ , (c, f)  $-0.32$  MV/m based on different criteria: (a–c) work and (d–f) energy density.

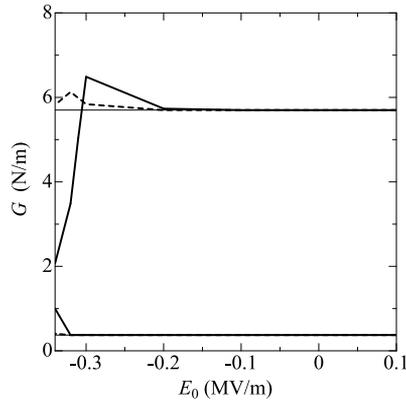


**Figure 7.** Energy release rate  $G$  versus electric field  $E_0$  for rectangular piezoelectric material C-91 under applied displacement in the permeable model.  $l = h = 2.5$  mm,  $a = 1$  mm, and  $u_0 = 0.5$   $\mu\text{m}$ . For legend, see Figure 5.

0.125  $\mu\text{m}$ ) when  $u_0 = 0.5$   $\mu\text{m}$  is applied, as shown in Figure 7. After  $E_0$  reaches about  $-0.21$  ( $-0.15$ ) MV/m, the  $G$  with the switching effect, based on the work (energy density), deviates from the curve without the switching effect. This is due to the  $180^\circ$  switching behind the crack tip; see Figure 8. As



**Figure 8.** Polarization switching zone induced by displacement  $u_0 = 0.5$   $\mu\text{m}$  and electric field  $E_0$  of (a, d)  $-0.10$ , (b, e)  $-0.20$ , (c, f)  $-0.30$  MV/m based on different criteria: (a–c) work; (d–f) energy density.



**Figure 9.** Energy release rate  $G$  versus electric field  $E_0$  for rectangular piezoelectric material C-91 under applied stress in the permeable model.  $l = h = 2.5$  mm and  $a = 1$  mm. We present results for  $\sigma_l = 22.8$  and  $5.70$  MPa. For legend, see [Figure 5](#).

$E_0$  reaches about  $-0.32$  ( $-0.30$ ) MV/m,  $G$  falls. In the experimental data [[Sun and Park 2000](#)], crack length deviated from the linear function of the electric field for the case of a larger load, especially for negative electric fields. By including the polarization switching effect of the energy release rate, the observed nonlinear dependence of piezoelectric crack behavior on the electric field is explained.

[Figure 9](#) shows the energy release rate  $G$  versus electric field  $E_0$  under applied stress. The rectangular piezoelectric material C-91 ( $2l = 5$  mm,  $2h = 5$  mm) with a permeable crack ( $2a = 2$  mm) is subjected to the stress  $\sigma_l = 22.8$  MPa, corresponding to the uniform strain  $2 \times 10^{-4}$  for the uncracked material without the electric field. We also present data for  $\sigma_l = 5.70$  MPa. The results for positive  $E_0$  under applied stress are different from those under applied displacement, and the energy release rate for the permeable crack in the rectangular piezoelectric material is independent of the positive  $E_0$ . The behavior of the energy release rate for negative  $E_0$  is complicated because of the polarization switching phenomena.

## 6. Conclusions

Theoretical and finite element analyses are presented for the cracked piezoelectric materials under tension. Based on the results of this study, the following conclusions may be inferred:

- (1) Piezoelectric crack face boundary conditions strongly affect the electric field effect characteristics of the electromechanical behavior and fracture mechanics parameters such as energy release rate.
- (2) The energy release rate criteria for the open and impermeable crack models led to negative values which are unphysical. The energy release rate for the permeable crack always remains positive.
- (3) For the permeable crack in the rectangular piezoelectric material, the positive electric field decreases the energy release rate under applied displacement. If the negative electric field is applied, localized polarization switching occurs due to electroelastic field concentrations near the crack tip, and the switching causes a sudden change in the energy release rate under applied displacement or stress.

- (4) The higher mechanical loading level decreases the critical value of the electric field associated with the polarization switching, and the localized 180° switching ahead of the crack tip can significantly influence the energy release rate.

### Appendix A

We consider an infinite piezoelectric material with a permeable crack under applied strain  $\epsilon_0$  and electric field  $E_0$ . The crack face boundary and loading conditions become

$$\sigma_{zx}(x, 0) = 0 \quad (0 \leq x < \infty), \quad \sigma_{zz}(x, 0) = 0 \quad (0 \leq x < a), \quad u_z(x, 0) = 0 \quad (a \leq x < \infty), \quad (\text{A.1})$$

$$E_x(x, 0) = E_x^c(x, 0) \quad (0 \leq x < a), \quad \phi(x, 0) = 0 \quad (a \leq x < \infty), \quad (\text{A.2})$$

$$D_z(x, 0) = D_z^c(x, 0) \quad (0 \leq x < a), \quad (\text{A.3})$$

$$\epsilon_{zz}(x, z) = \epsilon_0 \quad (0 \leq x < \infty, z \rightarrow \infty), \quad E_z(x, z) = E_0 \quad (0 \leq x < \infty, z \rightarrow \infty). \quad (\text{A.4})$$

Fourier transform is applied to Equations (3) and the results satisfying the loading conditions (A.4) are

$$\begin{aligned} u_x(x, z) &= \frac{2}{\pi} \sum_{j=1}^3 \int_0^\infty a_j A_j(\alpha) \exp(-\gamma_j \alpha z) \sin(\alpha x) d\alpha + \left( \frac{c_{13}}{c_{13}^2 - c_{33}c_{11}} (\sigma_l + e_1 E_0) + \frac{e_{31}}{c_{11}} E_0 \right) x, \\ u_z(x, z) &= \frac{2}{\pi} \sum_{j=1}^3 \int_0^\infty \frac{1}{\gamma_j} A_j(\alpha) \exp(-\gamma_j \alpha z) \cos(\alpha x) d\alpha + \frac{c_{11}}{c_{33}c_{11} - c_{13}^2} (\sigma_l + e_1 E_0) z, \\ \phi(x, z) &= -\frac{2}{\pi} \sum_{j=1}^3 \int_0^\infty \frac{b_j}{\gamma_j} A_j(\alpha) \exp(-\gamma_j \alpha z) \cos(\alpha x) d\alpha - E_0 z, \end{aligned} \quad (\text{A.5})$$

where  $A_j(\alpha)$  are the unknowns to be solved for,  $a_j$  and  $b_j$  stand for expressions

$$a_j = \frac{(e_{31} + e_{15})(c_{33}\gamma_j^2 - c_{44}) - (c_{13} + c_{44})(e_{33}\gamma_j^2 - e_{15})}{(c_{44}\gamma_j^2 - c_{11})(e_{33}\gamma_j^2 - e_{15}) + (c_{13} + c_{44})(e_{31} + e_{15})\gamma_j^2}, \quad b_j = \frac{(c_{44}\gamma_j^2 - c_{11})a_j + (c_{13} + c_{44})}{e_{31} + e_{15}},$$

and  $\gamma_j^2$  are the roots of the characteristic equation  $a_0\gamma^6 + b_0\gamma^4 + c_0\gamma^2 + d_0 = 0$  with

$$\begin{aligned} a_0 &= c_{44}(c_{33}\epsilon_{33} + e_{33}^2), & d_0 &= -c_{11}(c_{44}\epsilon_{11} + e_{15}^2), \\ b_0 &= -2c_{44}e_{15}\epsilon_{33} - c_{11}e_{33}^2 - c_{33}(c_{44}\epsilon_{11} + c_{11}\epsilon_{33}) + \epsilon_{33}(c_{13} + c_{44})^2 \\ &\quad + 2e_{33}(c_{13} + c_{44})(e_{31} + e_{15}) - c_{44}^2\epsilon_{33} - c_{33}(e_{31} + e_{15})^2, \\ c_0 &= 2c_{11}e_{15}\epsilon_{33} + c_{44}e_{15}^2 + c_{11}(c_{33}\epsilon_{11} + c_{44}\epsilon_{33}) - \epsilon_{11}(c_{13} + c_{44})^2 \\ &\quad - 2e_{15}(c_{13} + c_{44})(e_{31} + e_{15}) + c_{44}^2\epsilon_{11} + c_{44}(e_{31} + e_{15})^2. \end{aligned}$$

Application of the Fourier transform to Equation (4)<sub>3</sub> yields

$$\phi^c = \frac{2}{\pi} \int_0^\infty C(\alpha) \sinh(\alpha z) \cos(\alpha x) d\alpha, \quad (0 \leq x < a),$$

where  $C(\alpha)$  is also unknown.

By applying the crack face boundary conditions of Equations (A.1) and (A.2), the unknowns  $A_j(\alpha)$  are related to the stress  $\sigma_l$  via

$$A_j(\alpha) = -\frac{\pi d_j a J_1(a\alpha)}{2 F \alpha} \sigma_l,$$

where  $J_1(\cdot)$  is the order one Bessel function of the first kind and  $F = \sum_{j=1}^3 g_j d_j$  with

$$\begin{aligned} d_1 &= \gamma_1(b_2 f_3 - b_3 f_2), & d_2 &= \gamma_2(b_3 f_1 - b_1 f_3), & d_3 &= \gamma_3(b_1 f_2 - b_2 f_1), \\ f_j &= c_{44}(a_j \gamma_j^2 + 1) - e_{15} b_j, & g_j &= c_{13} a_j - c_{33} + e_{33} b_j. \end{aligned}$$

The displacement  $u_z$  and electric potential  $\phi$  on the crack surface are given by

$$u_z(x, 0) = -\frac{b}{F} \sigma_l (a^2 - x^2)^{1/2}, \quad \phi(x, 0) = 0, \quad b = b_1(f_2 - f_3) + b_2(f_3 - f_1) + b_3(f_1 - f_2).$$

The tangential component of electric field  $E_x$  and the normal component of electric displacement  $D_z$  on the crack surface are

$$\begin{aligned} E_x(x, 0) &= 0, & D_z(x, 0) &= D_l - \frac{\sigma_l}{F} \sum_{j=1}^3 h_j d_j, \\ D_l &= \frac{e_{31} c_{13} - e_{33} c_{11}}{c_{13}^2 - c_{33} c_{11}} \sigma_0 + \left( \frac{e_{31}^2}{c_{11}} + \epsilon_{33} \right) E_0, & h_j &= e_{31} a_j - e_{33} - \epsilon_{33} b_j. \end{aligned}$$

The displacement component  $u_z$  and electric potential  $\phi$  near the crack tip can be written as

$$\begin{aligned} u_z &= -\frac{K_I}{F} \left( \frac{r}{\pi} \right)^{1/2} \sum_{j=1}^3 \frac{d_j}{\gamma_j} [(\cos^2 \theta + \gamma_j^2 \sin^2 \theta)^{1/2} - \cos \theta]^{1/2}, \\ \phi &= \frac{K_I}{F} \left( \frac{r}{\pi} \right)^{1/2} \sum_{j=1}^3 \frac{b_j d_j}{\gamma_j} [(\cos^2 \theta + \gamma_j^2 \sin^2 \theta)^{1/2} - \cos \theta]^{1/2}, \end{aligned} \tag{A.6}$$

where the polar coordinates  $r$  and  $\theta$  are defined by  $r = [(x - a)^2 + z^2]^{1/2}$ ,  $\theta = \tan^{-1} z/(x - a)$ . The singular parts of the stress  $\sigma_{zz}$  and electric displacement  $D_z$  in the neighborhood of the crack tip are

$$\begin{aligned} \sigma_{zz} &= \frac{K_I}{2F(\pi r)^{1/2}} \sum_{j=1}^3 g_j d_j \left[ \frac{(\cos^2 \theta + \gamma_j^2 \sin^2 \theta)^{1/2} + \cos \theta}{\cos^2 \theta + \gamma_j^2 \sin^2 \theta} \right]^{1/2}, \\ D_z &= \frac{K_I}{2F(\pi r)^{1/2}} \sum_{j=1}^3 h_j d_j \left[ \frac{(\cos^2 \theta + \gamma_j^2 \sin^2 \theta)^{1/2} + \cos \theta}{\cos^2 \theta + \gamma_j^2 \sin^2 \theta} \right]^{1/2}. \end{aligned} \tag{A.7}$$

By using the concept of crack closure energy and the asymptotic behavior of electroelastic fields near the crack tip illustrated in Equations (A.6) and (A.7), the energy release rate  $G$  in Equation (10) for the permeable crack model can be obtained.

**Appendix B**

A solution procedure for the impermeable crack model in the infinite piezoelectric material is outlined here. The crack face electric boundary condition for the impermeable crack model is

$$D_z(x, 0) = 0 \quad (0 \leq x < a), \quad \phi(x, 0) = 0 \quad (a \leq x < \infty). \tag{B.8}$$

The unknowns  $A_j(\alpha)$  in Equations (A.5) can be found using the same approach as in the permeable case. By applying the crack face boundary conditions of Equations (A.1) and (B.8), the unknowns  $A_j(\alpha)$  are related to  $\sigma_l$  and  $D_l$  as follows:

$$\begin{aligned} \frac{f_1}{\gamma_1} A_1(\alpha) + \frac{f_2}{\gamma_2} A_2(\alpha) + \frac{f_3}{\gamma_3} A_3(\alpha) &= 0, \\ \frac{1}{\gamma_1} A_1(\alpha) + \frac{1}{\gamma_2} A_2(\alpha) + \frac{1}{\gamma_3} A_3(\alpha) &= -\frac{\pi}{2F'} \frac{a}{\alpha} J_1(a\alpha) (F_{22}\sigma_l - F_{12}D_l), \\ \frac{b_1}{\gamma_1} A_1(\alpha) + \frac{b_2}{\gamma_2} A_2(\alpha) + \frac{b_3}{\gamma_3} A_3(\alpha) &= \frac{\pi}{2F'} \frac{a}{\alpha} J_1(a\alpha) (F_{21}\sigma_l - F_{11}D_l). \end{aligned}$$

where

$$\begin{aligned} F_{11} &= \frac{1}{b} \sum_{j=1}^3 g_j d_j, & F_{12} &= \frac{1}{b} \sum_{j=1}^3 g_j l_j, & F_{21} &= \frac{1}{b} \sum_{j=1}^3 h_j d_j, & F_{22} &= \frac{1}{b} \sum_{j=1}^3 h_j l_j, \\ F' &= F_{11}F_{22} - F_{12}F_{21}, & l_1 &= \gamma_1(f_2 - f_3), & l_2 &= \gamma_2(f_3 - f_1), & l_3 &= \gamma_3(f_1 - f_2). \end{aligned}$$

The displacement  $u_z$ , electric potential  $\phi$ , tangential component of electric field  $E_x$  and normal component of electric displacement  $D_z$  on the crack surface are given by

$$\begin{aligned} u_z(x, 0) &= -\frac{F_{22}\sigma_l - F_{12}D_l}{F'} (a^2 - x^2)^{1/2}, & \phi(x, 0) &= -\frac{F_{21}\sigma_l - F_{11}D_l}{F'} (a^2 - x^2)^{1/2}, \\ E_x(x, 0) &= -\frac{F_{21}\sigma_l - F_{11}D_l}{F'} \frac{x}{(a^2 - x^2)^{1/2}}, & D_z(x, 0) &= 0. \end{aligned}$$

The energy release rate  $G^I$  for the impermeable crack model is

$$\begin{aligned} G^I &= -\frac{1}{2F'^2} \left[ \left( F' \sum_{j=1}^3 \frac{s_j}{\gamma_j} - \sum_{k=1}^3 h_k s_k \sum_{j=1}^3 \frac{b_j s_j}{\gamma_j} \right) K_I^2 \right. \\ &\quad \left. + \left( \sum_{k=1}^3 h_k t_k \sum_{j=1}^3 \frac{b_j s_j}{\gamma_j} + \sum_{k=1}^3 h_k s_k \sum_{j=1}^3 \frac{b_j t_j}{\gamma_j} - F' \sum_{j=1}^3 \frac{t_j}{\gamma_j} \right) K_I K_D - \left( \sum_{k=1}^3 h_k t_k \sum_{j=1}^3 \frac{b_j t_j}{\gamma_j} \right) K_D^2 \right], \tag{B.9} \end{aligned}$$

where  $s_j = d_j F_{22} - l_j F_{21}$  and  $t_j = d_j F_{12} - l_j F_{11}$ . In Equation (B.9) the stress and the electric displacement intensity factors are given by, respectively,

$$K_I = \lim_{x \rightarrow a^+} [2\pi(x - a)]^{1/2} \sigma_{zz}(x, 0) = \sigma_l(\pi a)^{1/2}, \quad K_D = \lim_{x \rightarrow a^+} [2\pi(x - a)]^{1/2} D_z(x, 0) = D_l(\pi a)^{1/2}.$$

### Appendix C

The solutions for the open crack model in the infinite piezoelectric material can be derived as follows. The crack face electric boundary condition for the open crack model becomes

$$D_z^+ = D_z^- \quad (0 \leq x < a), \quad D_z^+(u_z^+ - u_z^-) = \epsilon_0(\phi^- - \phi^+) \quad (0 \leq x < a), \quad \phi(x, 0) = 0 \quad (a \leq x < \infty), \quad (\text{C.10})$$

where the superscripts + and - denote the upper and lower crack surfaces, respectively. By applying the crack face boundary conditions of Equations (A.1) and (C.10), the unknowns  $A_j(\alpha)$  in Equations (A.5) are related to  $\sigma_l$  and  $D_l$  as follows:

$$\begin{aligned} \frac{f_1}{\gamma_1} A_1(\alpha) + \frac{f_2}{\gamma_2} A_2(\alpha) + \frac{f_3}{\gamma_3} A_3(\alpha) &= 0, \\ \frac{1}{\gamma_1} A_1(\alpha) + \frac{1}{\gamma_2} A_2(\alpha) + \frac{1}{\gamma_3} A_3(\alpha) &= -\frac{\pi}{2F'} \frac{a}{\alpha} J_1(a\alpha) (F_{22}\sigma_l + F_{12}(D_0 - D_l)), \\ \frac{b_1}{\gamma_1} A_1(\alpha) + \frac{b_2}{\gamma_2} A_2(\alpha) + \frac{b_3}{\gamma_3} A_3(\alpha) &= \frac{\pi}{2F'} \frac{a}{\alpha} J_1(a\alpha) (F_{21}\sigma_l + F_{11}(D_0 - D_l)), \end{aligned}$$

where

$$D_0 = -\epsilon_0 \frac{F_{21}\sigma_l + F_{11}(D_0 - D_l)}{F_{22}\sigma_l + F_{12}(D_0 - D_l)}.$$

If  $\epsilon_0 = 0$ ,  $D_0$  is equal to zero. When  $\epsilon_0$  becomes very large, the expression for  $D_0$  above shows that  $D_0 \rightarrow D_l - (F_{21}/F_{11})\sigma_l$ .

The displacement, electric potential, tangential component of electric field and normal component of electric displacement on the crack surface are

$$\begin{aligned} u_z(x, 0) &= -\frac{F_{22}\sigma_l + F_{12}(D_0 - D_l)}{F'} (a^2 - x^2)^{1/2}, \quad \phi(x, 0) = -\frac{F_{21}\sigma_l + F_{11}(D_0 - D_l)}{F'} (a^2 - x^2)^{1/2}, \\ E_x(x, 0) &= -\frac{F_{21}\sigma_l + F_{11}(D_0 - D_l)}{F'} \frac{x}{(a^2 - x^2)^{1/2}}, \quad D_z(x, 0) = D_0. \end{aligned}$$

Energy release rate  $G^0$  for the open crack model is given by (B.9) with  $K_D = (D_l - D_0)(\pi a)^{1/2}$ .

### References

- [Fu and Zhang 2000] R. Fu and T. Y. Zhang, "Effects of an electric field on the fracture toughness of poled lead zirconate titanate ceramics", *J. Am. Ceram. Soc.* **83**:5 (2000), 1215–1218.
- [Hao and Shen 1994] T.-H. Hao and Z.-Y. Shen, "A new electric boundary condition of electric fracture mechanics and its applications", *Eng. Fract. Mech.* **47**:6 (1994), 793–802.
- [Hwang et al. 1995] S. C. Hwang, C. S. Lynch, and R. M. McMeeking, "Ferroelectric/ferroelastic interactions and a polarization switching model", *Acta Metall. Mater.* **43**:5 (1995), 2073–2084.
- [Kalyanam and Sun 2005] S. Kalyanam and C. T. Sun, "Modeling of electrical boundary condition and domain switching in piezoelectric materials", *Mech. Mater.* **37**:7 (2005), 769–784.
- [Landis 2004] C. M. Landis, "Energetically consistent boundary conditions for electromechanical fracture", *Int. J. Solids Struct.* **41**:22-23 (2004), 6291–6315.
- [McMeeking 1999] R. M. McMeeking, "Crack tip energy release rate for a piezoelectric compact tension specimen", *Eng. Fract. Mech.* **64**:2 (1999), 217–244.

- [McMeeking 2004] R. M. McMeeking, “The energy release rate for a griffith crack in a piezoelectric material”, *Eng. Fract. Mech.* **71**:7-8 (2004), 1149–1163.
- [Narita et al. 2003] F. Narita, Y. Shindo, and K. Horiguchi, “Electroelastic fracture mechanics of piezoelectric ceramics”, pp. 89–101 in *Mechanics of electromagnetic material systems structures*, WIT Press, Southampton, 2003. Y. Shindo, ed.
- [Narita et al. 2005] F. Narita, Y. Shindo, and K. Hayashi, “Bending and polarization switching of piezoelectric laminated actuators under electromechanical loading”, *Comput. Struct.* **83**:15-16 (2005), 1164–1170.
- [Park and Sun 1995] S. Park and C.-T. Sun, “Fracture criteria for piezoelectric ceramics”, *J. Am. Ceram. Soc.* **78**:6 (1995), 1475–1480.
- [Shindo et al. 2002] Y. Shindo, H. Murakami, K. Horiguchi, and F. Narita, “Evaluation of electric fracture properties of piezoelectric ceramics using the finite elementsingle-edge precracked-beam methods”, *J. Am. Ceram. Soc.* **85**:5 (2002), 1243–1248.
- [Shindo et al. 2003] Y. Shindo, F. Narita, K. Horiguchi, Y. Magara, and M. Yoshida, “Electric fracture and polarization switching properties of piezoelectric ceramic pzt studied by the modified small punch test”, *Acta Mater.* **51**:16 (2003), 4773–4782.
- [Shindo et al. 2004] Y. Shindo, M. Yoshida, F. Narita, and K. Horiguchi, “Electroelastic field concentrations ahead of electrodes in multilayer piezoelectric actuators: experiment and finite element simulation”, *J. Mech. Phys. Solids* **52**:5 (2004), 1109–1124.
- [Shindo et al. 2005] Y. Shindo, F. Narita, and M. Mikami, “Double torsion testing and finite element analysis for determining the electric fracture properties of piezoelectric ceramics”, *J. Appl. Phys.* **97** (2005), 114109.
- [Sun and Achuthan 2004] C.-T. Sun and A. Achuthan, “Domain-switching criteria for ferroelectric materials subjected to electrical and mechanical loads”, *J. Am. Ceram. Soc.* **87**:3 (2004), 395–400.
- [Sun and Park 2000] C. T. Sun and S. B. Park, “Measuring fracture toughness of piezoceramics by vickers indentation under the influence of electric fields”, *Ferroelectrics* **248** (2000), 79–95.
- [Wang and Mai 2003] B. L. Wang and Y.-W. Mai, “On the electrical boundary conditions on the crack surfaces in piezoelectric ceramics”, *Int. J. Eng. Sci.* **41**:6 (2003), 633–652.
- [Xu and Rajapakse 2001] X.-L. Xu and R. K. N. D. Rajapakse, “On a plane crack in piezoelectric solids”, *Int. J. Solids Struct.* **38**:42-43 (2001), 7643–7658.
- [Yoshida et al. 2003] M. Yoshida, F. Narita, Y. Shindo, M. Karaiwa, and K. Horiguchi, “Electroelastic field concentration by circular electrodes in piezoelectric ceramics”, *Smart Mater. Struct.* **12** (2003), 972–978.

Received 16 Jun 2006. Revised 13 Apr 2007. Accepted 20 Apr 2007.

YASUhide SHINDO: [shindo@material.tohoku.ac.jp](mailto:shindo@material.tohoku.ac.jp)

Department of Materials Processing, Graduate School of Engineering, Tohoku University, Aoba-yama 6-6-02, Sendai 980-8579, Japan

FUMIO NARITA: [narita@material.tohoku.ac.jp](mailto:narita@material.tohoku.ac.jp)

Department of Materials Processing, Graduate School of Engineering, Tohoku University, Aoba-yama 6-6-02, Sendai 980-8579, Japan

FUMITOSHI SAITO: Department of Materials Processing, Graduate School of Engineering, Tohoku University, Aoba-yama 6-6-02, Sendai 980-8579, Japan

# ACTIVE CONTROL SCHEMES BASED ON THE LINEARIZED TSCHAUNER–HEMPEL EQUATIONS TO MAINTAIN THE SEPARATION DISTANCE CONSTRAINTS FOR THE NASA BENCHMARK TETRAHEDRON CONSTELLATION

PEDRO A. CAPÓ-LUGO AND PETER M. BAINUM

The NASA benchmark tetrahedron constellation is a proposed satellite formation that requires a nominal separation distance at every apogee point. To maintain these separation distance constraints between any pair of satellites within the constellation, an open-loop scheme was developed based on the orbital elements. For a particular size of the NASA benchmark tetrahedron problem, the constellation maintains the separation distance conditions without perturbations. On the other hand, with perturbations, the constellation maintains the separation distance criteria for a limited number of orbits.

This scheme does not maintain the constellation together for the complete mission period. For this reason, the Tschauner–Hempel (TH) equations are used to maintain the separation distance criteria. Two control schemes are used to maintain the separation distance conditions of the tetrahedron constellation and are compared with each other to determine which one provides for minimum time and consumption.

## 1. Introduction

The proposed NASA benchmark tetrahedron constellation [Carpenter et al. 2003] is a complex problem because of the different strategies used to maintain the separation distance constraints. This benchmark problem is divided in three different specific sizes that contain different orbital dimensions for every orbit. For three specific sizes, we have analyzed the proposed constellation without the use of an active control scheme in which the strategy was based only on the orbital elements; for details see [Capó-Lugo 2005; Capó-Lugo and Bainum 2005b]. With this strategy the constellation satisfied the separation distance constraints for a short period of time without perturbations. When perturbations are added, the constellation violates the separation distance constraints in a limited number of complete orbits. After the first pair of satellites violates the separation distance conditions, an active control scheme is needed to maintain the separation distance criteria for an additional short period of time.

The motion of a pair of satellites around Earth is explained by the linearized Tschauner–Hempel (TH) equations. These equations describe the rendezvous of a pair of satellites in an elliptical orbit in which these satellites have a relative separation distance. To maintain the separation distance between a pair of satellites within the constellation, the linear quadratic regulator (LQR) is used as the active control scheme, but two different approaches are used with this active control scheme. Tan et al. [1999; 2002] and Bainum et al. [2004] (BST) developed an active control scheme which adapted in a piecewise manner the varying term in the linearized TH equations to correct the separation distance of an along-track

---

*Keywords:* tetrahedron constellation, linear quadratic regulator, elliptical orbits.

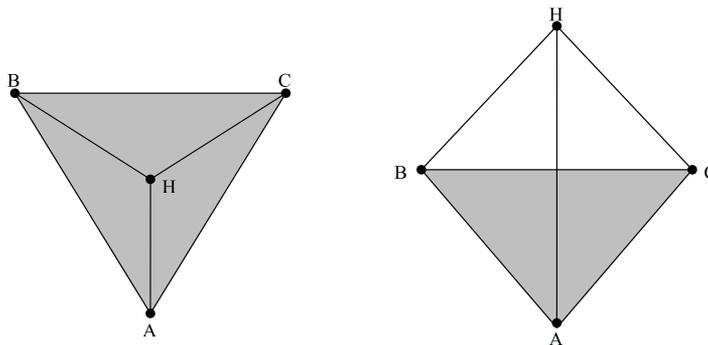
Research supported by the Alliances for Graduate Education and Professoriate (AGEP) Program.

constellation (or string of pearls). Carter and Humi [1987] and Carter [1990] (CH) developed a different control scheme based on the Pontryagin minimum principle, but a different formulation based on the CH approach is developed for the LQR control scheme. With these two techniques, one specific size of the NASA benchmark tetrahedron constellation is used to determine how the active control scheme is affected by these orbital dimensions and the different weighting matrices used in the LQR strategy. After the correction of the positions and velocities for a pair of satellites is performed, the Satellite Tool Kit [STK 2003] software is used to determine if the constellation is going to hold for another short period of time before a pair of satellites within the constellation violates the separation distance constraints. Hence, two active control schemes will be tested with one orbital size to understand their different responses to the correction of the separation distance between a pair of satellites.

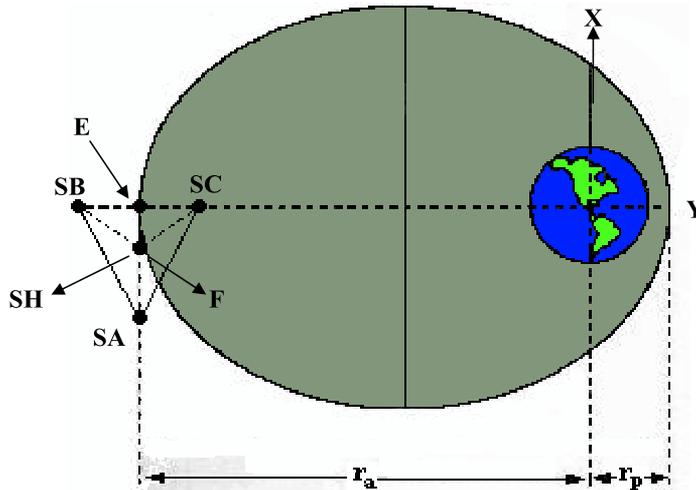
## 2. Tetrahedron definition

The NASA benchmark tetrahedron configuration is similar to a pyramid, but the base of this configuration is an equilateral triangle with an apex point above the centroid of the triangle in a different plane [Carpenter et al. 2003; Capó-Lugo 2005; Capó-Lugo and Bainum 2005b]. Figure 1 shows the top and front view of the configuration. Points  $A$ ,  $B$ ,  $C$ , and  $H$  are the nominal positions of the satellites in the constellation, but throughout the paper these points will be also referred to as  $SA$ ,  $SB$ ,  $SC$ , and  $SH$ . Points  $B$  and  $C$  are nominally situated along the line of apsides, and points  $H$  and  $A$  are the satellites nominally orbiting around the centroid in a different orbital plane and in the equilateral triangle, respectively. The nominal separation distance between any two subsatellites at apogee is 10 km, and the separation error at subsequent apogees should be within 10%, giving an acceptable range between 9 and 11 km [Carpenter et al. 2003]. At other points in the orbit, the minimum separation distance between any pairs of subsatellites should be 1 km [Carpenter et al. 2003]. The purpose of this constellation is to measure the electromagnetic field of the Earth.

The positions of the satellites are determined from the reference point with respect to the Earth Coordinate Inertial (ECI) frame, given by point  $E$  in Figure 2. In this problem, the configuration is assumed such that the satellites arrive at the initial apogee point by some predetermined launch sequence. Figure 2 shows the tetrahedron configuration in the  $x$ - $y$  plane. This configuration is situated at the apogee point where  $r_a$  and  $r_p$  are the radii of apogee and perigee, respectively. As mentioned above,  $SB$  and  $SC$  are



**Figure 1.** Top view (left side) and front view (right side) of the tetrahedron configuration.



**Figure 2.** Two dimensional view of the configuration at apogee.

situated along the line of apsides ( $Y$  direction),  $SA$  forms the equilateral triangle, and  $SH$  is over the centroid in a different orbital plane. [Table 1](#) shows the initial positions of the constellation at the apogee point with respect to the ECI frame [[Capó-Lugo 2005](#); [Capó-Lugo and Bainum 2005b](#)].

### 3. Definition of the specific sizes (or phases)

The benchmark problem has four phases with a mission period of two years, but this research is only concerned with the three phases that contain the restrictions to maintain the separation distance constraints. [Table 2](#) details these three phases in terms of the orbital elements [[Carpenter et al. 2003](#); [Capó-Lugo 2005](#); [Capó-Lugo and Bainum 2005b](#)]. The fourth phase for the NASA benchmark problem is a lunar swing-by which is not considered here. [Table 2](#) shows the dimensions considered for the three phases.

The inclination angle in the third phase is not specified in the benchmark problem because the constellation must be in a near polar orbit, so this orbital inclination angle is chosen to be 85 degrees. As [Table 2](#) shows, the last phase has the largest orbit, smallest eccentricity and largest orbital period. Through this paper only phase I is analyzed with the LQR active control scheme.

Axis (ECI frame)	Ref. Point (km)	SA (km)	SB (km)	SC (km)	Centroid (km)	SH (km)
x	0	$-\overline{AE}$	0	0	$-\overline{EF}$	$-\overline{EF}$
y	$-r_a$	$-r_a$	$-(r_a + 5)$	$-(r_a - 5)$	$-r_a$	$-\frac{r_a}{2}$
z	0	0	0	0	0	$\overline{HF}$

**Table 1.** Initial satellite position at apogee with respect to the ECI frame, where  $\overline{AE} = 5\sqrt{3}$ ,  $\overline{EF} = 5/\sqrt{3}$ , and  $\overline{HF} = 10/\sqrt{3}$ .

Dimensions	First phase	Second phase	Third phase
Radius of perigee ( $r_p$ )	1.2 ER	1.2 ER	10 ER
Radius of apogee ( $r_a$ )	12 ER	30 ER	40 ER
Semimajor axis ( $a$ )	42,095.7 km	99,498.92 km	159,453.4 km
Eccentricity ( $e$ )	0.818	0.923	0.6
Inclination angle ( $i$ )	18.5°	18.5°	85.0°
Period (days)	1	2	7

**Table 2.** Dimensions and properties for the three phases; ER means Earth radius.

### 4. Development of the equations of motion

The derivation of the equations of motion follows that of [Carter and Humi \[1987\]](#) who derived a set of equations to describe the rendezvous motion between a pair of satellites in an elliptical orbit for a general Keplerian orbit. For this application, the separation distance between a pair of satellites within the constellation is needed to be maintained at the apogee point to satisfy the separation distance constraints of the NASA benchmark problem [[Capó-Lugo and Bainum 2005b](#)].

The equations of motion are derived for a maneuvering satellite and a reference (target) satellite or point which is orbiting about the Earth in an elliptical orbit as shown in [Figure 3](#). The maneuvering satellite is assumed to have a scalar point mass  $m(t)$  and an applied thrust vector  $T(t)$  projected in the reference axis system. The target satellite is acted on by a Newtonian gravitational force directed toward the center of the Earth.  $R(t)$  is the vector measured from the center of the Earth to the reference or target satellite, and  $r(t)$  is the vector determined from the center of the Earth to the maneuvering satellite.  $\rho(t)$  (dashed line in [Figure 3](#)) is the vector measured from the reference satellite to the maneuvering satellite and describes their relative separation distance. The coordinate system is defined by two conditions. First,  $x_1$  is opposed to the motion of the maneuvering satellite and perpendicular to the  $x_2$ -axis, whose positive direction is along  $R(t)$ . Secondly,  $x_3$  is positive when the right handed system is completed.  $\omega$  is the relative angular velocity of the satellites about the Earth.

[Carter and Humi \[1987\]](#) developed a set of equations which explains the movement of a maneuvering spacecraft relative to the reference satellite in an elliptical orbit as shown in [Figure 3](#). This set of equations is dependent on the true anomaly angle  $\theta$  and is given by

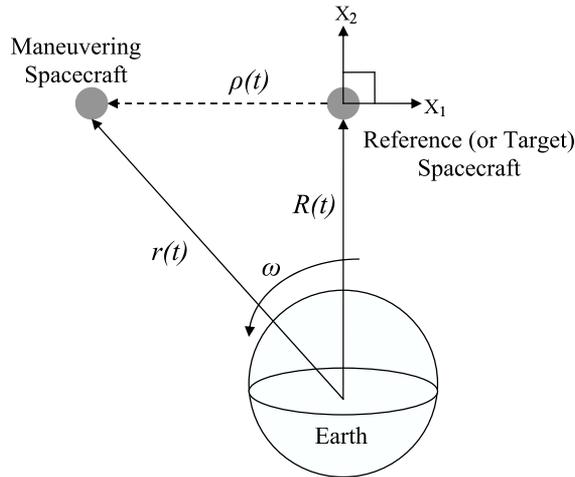
$$(1 + e \cos \theta)x_1'' - (2e \sin \theta)x_1' = (e \cos \theta)x_1 - (2e \sin \theta)x_2 + 2(1 + e \cos \theta)x_2' + a_1, \tag{1}$$

$$(1 + e \cos \theta)x_2'' - (2e \sin \theta)x_2' = (2e \sin \theta)x_1 - 2(1 + e \cos \theta)x_1' + (3 + e \cos \theta)x_2 + a_2, \tag{2}$$

$$(1 + e \cos \theta)x_3'' - (2e \sin \theta)x_3' = -x_3 + a_3, \tag{3}$$

$$a_j = \frac{T_j}{m} \left( \frac{h^6}{\mu^4} \right) (1 + e \cos \theta)^{-3}, \quad \varphi' = \frac{d\varphi}{d\theta}, \quad \varphi'' = \frac{d^2\varphi}{d\theta^2},$$

where  $j = 1, 2, 3$  and  $\varphi$  is any function of the true anomaly angle. Equations (1)–(3) describe the Keplerian motion of the maneuvering satellite relative to the reference or target satellite for the special case where  $a_i$  (or  $T_i$ ) are zero. These equations can be solved analytically using transformations from



**Figure 3.** Reference and maneuvering satellite motion about Earth.

[Carter and Humi 1987]:

$$\{(1 + e \cos \theta)x_j\}' = (-e \sin \theta)x_j + (1 + e \cos \theta)x_j',$$

$$\frac{1}{1 + e \cos \theta} \{(1 + e \cos \theta)x_j'\}' = (1 + e \cos \theta)x_j'' - 2(e \sin \theta)x_j',$$

which then give

$$\frac{1}{1 + e \cos \theta} \{(1 + e \cos \theta)^2 x_1'\}' = (e \cos \theta)x_1 + 2\{(1 + e \cos \theta)x_2'\}' + a_1, \quad (4)$$

$$\frac{1}{1 + e \cos \theta} \{(1 + e \cos \theta)^2 x_2'\}' = (3 + e \cos \theta)x_2 - 2\{(1 + e \cos \theta)x_1'\}' + a_2, \quad (5)$$

$$\frac{1}{1 + e \cos \theta} \{(1 + e \cos \theta)^2 x_3'\}' = -x_3 + a_3. \quad (6)$$

Since

$$y_j = (1 + e \cos \theta)x_j, \quad y_j' = (1 + e \cos \theta)x_j' - (e \sin \theta)x_j,$$

and

$$y_j'' = (1 + e \cos \theta)x_j'' - (2e \sin \theta)x_j' - (e \cos \theta)x_j,$$

Equations (4)–(6) reduce to

$$y_1'' = 2y_2' + a_1, \quad y_2'' = 3\kappa y_2 - 2y_1' + a_2, \quad y_3'' = -y_3 + a_3, \quad (7)$$

where  $\kappa = \mu r/h^2 = 1/(1 + e \cos \theta)$  is determined from the well known equation of a Keplerian orbit (or equation of a conic section). Equations (7) are called the rendezvous linearized Tschauner–Hempel (TH) equations for the motion of a pair of satellites in an elliptical orbit. A control function  $u(t)$  can be used to represent the change in thrust  $T(\theta)$  and mass  $m(\theta)$  with respect to the true anomaly angle

[Carter and Humi 1987] via

$$\frac{T_m}{m_0} u_j(\theta) = \frac{T_j(\theta)}{m(\theta)}, \quad (8)$$

where  $T_m$  and  $m_0$  are the maximum thrust and initial mass of the maneuvering satellite. One can introduce new state variables  $\xi$ ,  $\zeta$ , and  $\eta$  by defining [Athans and Falb 1966]

$$v = \frac{h^6 T_m}{\mu^4 m_0}, \quad \xi = \frac{y_1}{v}, \quad \zeta = \frac{y_2}{v}, \quad \eta = \frac{y_3}{v}. \quad (9)$$

Using Equations (8) and (9), the linearized equations for the motion of the maneuvering satellite (see Equation (7)), in state-based format, can be expressed in the following form:

$$\begin{bmatrix} \xi' \\ \zeta' \\ \eta' \\ \xi'' \\ \zeta'' \\ \eta'' \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 3\kappa & 0 & -2 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \xi \\ \zeta \\ \eta \\ \xi' \\ \zeta' \\ \eta' \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \kappa^3 & 0 & 0 \\ 0 & \kappa^3 & 0 \\ 0 & 0 & \kappa^3 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}. \quad (10)$$

## 5. BST control approach

Tan et al. [1999; 2002] used the linearized TH equations to determine an active control scheme to satisfy the separation distance constraints for a constellation in an along-track formation (or string of pearls).

The active control scheme used by these authors is the linear quadratic regulator (LQR) which is an optimal control. To determine this active control law, they used the following quadratic cost function:

$$J = \frac{1}{2} \int_{\theta_0}^{\theta_f} \left\{ [(x(\theta) - x_D)^T Q (x(\theta) - x_D)] + [(u(\theta))^T R (u(\theta))] \right\} d\theta, \quad (11)$$

where  $x(\theta)$ ,  $x_D$  describes the components of the actual a desired state vector, respectively,  $u(\theta)$  is the control signal that will be used to maintain the separation distance constraints,  $Q$  and  $R$  are  $n \times n$  positive semidefinite and  $m \times m$  positive definite weight functioning matrices, respectively. This cost function is used to minimize the difference in the errors between the state vector and the desired state vector, and a minimum time problem can be obtained to maintain the separation distance criteria.

For these authors,  $\kappa = (\mu r)/h^2$ , and they adapt the nonlinear term in a piecewise manner. The nonlinear term in Equation (10) can be adjusted in a number of ways [Strong 2000; Bainum et al. 2004]:

1. When it is assumed that  $r$  remains constant, that is,  $r(\theta) = h^2/\mu$ , true for a circle and relatively short displacements, then, the term becomes 3.
2. If the simulation is started at perigee or apogee, then, evaluate  $r$  at perigee and apogee, respectively, and treat as constant for a sufficiently short time after.
3. If several orbits are needed to correct the disturbance, then, use an average value of  $r$  with  $h = rv = \sqrt{(b^2\mu)/a}$ , where  $b$  is the semiminor axis.
4. A final consideration is to update 1 and 2 in a piecewise adaptive manner along the orbit.

## 6. CH control approach

Carter and Humi [1987] and Carter [1990] used the same linearized TH equations, but with

$$\kappa = 1/(1 + e \cos \theta)$$

and with the control matrix  $B$  defined differently. BST assumed the  $B$  matrix as a control signal that produces some force to maintain the satellites in their corresponding relative distance between the reference (or target) and maneuvering satellite. For Carter and Humi, the  $A$  and  $B$  matrices vary with the true anomaly angle because Marec [1979] established that to obtain a minimum time problem with a minimum consumption of control the true anomaly must be considered as part of the controls for an elliptic orbit.

Carter and Humi used the Pontryagin minimum principle to obtain an admissible control such that the control scheme is used to rendezvous in an optimal way between the target satellite and the reference satellite [Carter and Humi 1987; Carter 1990]. Their cost function for the optimal control is given by

$$J = \int_{\theta_0}^{\theta_f} \frac{|u_i(\theta)|}{(1 + e \cos \theta)^2} d\theta. \quad (12)$$

Using this LQR strategy, let us define the cost function via the varying terms in the  $A$  and  $B$  matrices as

$$J = \frac{1}{2} \int_{\theta_0}^{\theta_f} \left\{ \frac{(x(\theta) - x_D)^T Q (x(\theta) - x_D)}{1 + e \cos \theta} + \frac{(u(\theta))^T R (u(\theta))}{(1 + e \cos \theta)^2} \right\} d\theta. \quad (13)$$

This proposed cost function is in accordance with Marec's statement for a minimum-time problem in an elliptical orbit and is based on the cost function defined by Carter and Humi.

## 7. Development of the linear quadratic regulator

Pontryagin minimum (or maximum) principles have been used by different authors to obtain an admissible control which leads to an optimal control that maintains the relative distance between the maneuvering and target satellite (or reference point) [Carter and Humi 1987; Carter 1990; Carter and Brient 1992; Massari et al. 2004]. Carter and Brient [1992] show that these principles apply to any elliptical orbit if the cost function is defined by Equation (13). On the other hand, a digital optimal control has been implemented by Massari et al. [2004] to maintain the orbit of some satellites in elliptical orbits, using the form of Equations (4)–(6), but their cost function is not defined in terms of the true anomaly angle. In this section, the linear quadratic regulator (LQR) technique will be developed to satisfy a minimum-time problem defined by the cost functions in Equations (11) and (13).

The LQR optimal control approach will be implemented with the use of the cost function defined by Equation (11). The solution of the LQR problem leads to an optimal feedback system with the property that the components of the state vector  $x(\theta)$  are kept near the desired state vector  $x_D$  without excessive expenditure of control energy [Athans and Falb 1966]. The existence of the optimal control is obtained from the solution of the Hamilton–Jacobi equation which is defined everywhere to obtain a minimum-time problem.

Consider the angle varying system defined by Equation (10) in the form

$$x' = Ax + Bu + \psi(\theta) \quad (14)$$

and the cost functional defined by Equation (11).  $\psi(\theta)$  is a  $n \times 1$  disturbance column matrix that can contain different perturbations such as the J2 (Earth's oblateness) perturbation, the drag force, Moon's gravity, and the solar pressure. It is defined as a function of the true anomaly angle. The  $A$  and  $B$  matrices have dimensions of  $n \times n$  and  $n \times m$ , respectively. It is assumed that for any initial state, there exists an optimal control which can obtain a desirable minimum consumption problem. The Hamiltonian  $H$  for the system in Equation (14) and the cost function in Equation (11) can be defined as

$$H = \frac{1}{2}|x(\theta) - x_D|^2 Q + \frac{1}{2}|u(\theta)|^2 R + A(\theta)x(\theta) \cdot p(\theta) + B(\theta)u(\theta) \cdot p(\theta) + \psi(\theta) \cdot p(\theta).$$

The minimum principles are used to obtain the necessary conditions for the optimal control [Athans and Falb 1966]. The costate vector  $p(\theta)$  is the solution of the vector differential equations

$$p'(\theta) = -\frac{\partial H}{\partial x(\theta)} = -Q(x(\theta) - x_D) - A^T(\theta)p(\theta). \quad (15)$$

The optimal trajectory is given by

$$\frac{\partial H}{\partial u(\theta)} = 0 = u(\theta)R + B^T(\theta)p(\theta), \quad u(\theta) = -R^{-1}B^T(\theta)p(\theta). \quad (16)$$

Using Equation (16), the angle varying differential equations defined by the Hamiltonian above can be rewritten as

$$x' = A(\theta)x - S(\theta)p(\theta) + \psi(\theta), \quad S(\theta) = B(\theta)R^{-1}B^T(\theta). \quad (17)$$

$S(\theta)$  is a square  $n \times n$  matrix. Using the transversality conditions defined in [Athans and Falb 1966] the costate variables have the following relationship,

$$p(\theta) = k(\theta)x(\theta) + m(\theta). \quad (18)$$

The  $k(\theta)$  and  $m(\theta)$  matrices are  $n \times n$  and  $n \times 1$ , respectively. They depend on the final angle  $\theta_f$  and a weighting matrix in the final state [Athans and Falb 1966], but not on the initial state. The solutions of the state and costate vectors are related by Equation (18). Upon differentiating it with respect to the true anomaly angle and substituted into Equation (15), the costate variables can be written as

$$k'(\theta)x(\theta) + k(\theta)x'(\theta) + m'(\theta) = -Qx(\theta) - A^T(\theta)p(\theta) + Qx_D.$$

Substituting Equation (17) into the above equation, we get

$$k'(\theta)x(\theta) + k(\theta)\{A(\theta)x(\theta) + S(\theta)(k(\theta)x(\theta) + m(\theta)) + \psi(\theta)\} + m'(\theta) = -Qx(\theta) - A^T(\theta)p(\theta) + Qx_D,$$

which can be separated into two equations

$$k'(\theta) = -A^T(\theta)k(\theta) - k(\theta)A(\theta) + k(\theta)S(\theta)k(\theta) - Q, \quad (19)$$

$$m'(\theta) = (k(\theta)S(\theta) - A^T(\theta))m(\theta) + Qx_D - k(\theta)\psi(\theta). \quad (20)$$

Equations (19) and (20) are the Ricatti equation (RE) and the adjoint Ricatti equation (ARE), respectively. The RE and ARE must be solved backwards in time as explained in [Phillips and Nagle 1995]. This system can be solved using Runge–Kutta methods, but, this method always runs forward in time. For this reason, Euler's method is used to obtain an approximate solution since it can be applied backwards in time [Strang 1986; Borse 2000; Gerald and Wheatley 2004]. Euler's method must be applied until a

stable solution is obtained, but, instead of using this method to obtain a stable solution for the RE and ARE, the system can be defined continuously with respect to the true anomaly angle. In this way, a solution is obtained in the stable region. The following approximation can be applied for the system

$$\lim_{\Delta\theta \rightarrow \infty} k'(\theta) = \lim_{\Delta\theta \rightarrow \infty} m'(\theta) = \lim_{\Delta\theta \rightarrow \infty} \frac{k(\theta + \Delta\theta) - k(\theta)}{\Delta\theta} = \lim_{\Delta\theta \rightarrow \infty} \frac{m(\theta + \Delta\theta) - m(\theta)}{\Delta\theta} = 0.$$

Equations (19) and (20) become

$$0 = -A^T(\theta)k(\theta) - k(\theta)A(\theta) + k(\theta)S(\theta)k(\theta) - Q, \quad (21)$$

$$m(\theta) = (A^T(\theta) + k(\theta)S(\theta))^{-1}(Qx_D - k(\theta)\psi(\theta)). \quad (22)$$

The state vector can be solved using a numerical integration scheme such as the Runge–Kutta method since this integration process runs forward in time. Substituting Equation (18) into Equation (17), the state vector is defined as  $x'(\theta) = A(\theta)x(\theta) - S(\theta)k(\theta)x(\theta) - S(\theta)m(\theta) + \psi(\theta)$ . With Equations (16) and (18), the control vector is  $u(\theta) = -S(\theta)(k(\theta)x(\theta) + m(\theta))$ .

The following procedure is adapted to obtain a solution for the RE, the ARE, the state vector, and the control vector:

1. Use Equations (21) and (22) to obtain a solution for the RE and the ARE.
2. Substitute these values for  $k(\theta)$  and  $m(\theta)$  (continuous in time) into the state vector equation, and integrate forward in time using any numerical scheme.
3. Substitute these values for  $k(\theta)$ ,  $m(\theta)$ , and  $x(\theta)$  to determine  $u(\theta)$ .

## 8. BST approach to the LQR active control scheme

The LQR is determined from the cost function defined by Equation (11). The  $Q$  and  $R$  matrices do not vary with the true anomaly angle. Then, the RE and the ARE are

$$0 = -A^T(\theta)k_\infty - k_\infty A(\theta) + k_\infty S(\theta)k_\infty - Q, \quad m_\infty = (A^T(\theta) + k_\infty S(\theta))^{-1}(Qx_D - k_\infty \psi(\theta)).$$

$k_\infty$  and  $m_\infty$  are constants if the matrix is completely controllable and continuous with respect to the true anomaly angle. The state vector equation and the control vector are then given by

$$x'(\theta) = A(\theta)x(\theta) - Sk_\infty x(\theta) - Sm_\infty + \psi(\theta), \quad u(\theta) = -S(k_\infty x(\theta) + m_\infty).$$

These substitutions are performed because the authors adapt in a piecewise manner the angle varying term in the  $A$  matrix which is going to be constant for short periods of time.

## 9. Carter–Humi approach to the LQR active control scheme

For this control scheme, the LQR is complex because the cost function defined by Equation (13) varies with the true anomaly angle and therefore takes the form of an elliptical integral. Define it by

$$J = \frac{1}{2} \int_{\theta_0}^{\theta_f} \{ (x(\theta) - x_D)^T \tilde{Q} (x(\theta) - x_D) + (u(\theta))^T \tilde{R} u(\theta) \} d\theta, \quad \tilde{Q} = \frac{Q}{1 + e \cos \theta}, \quad \tilde{R} = \frac{R}{(1 + e \cos \theta)^2}.$$

Then, the RE and ARE, continuous with respect to the true anomaly angle, are defined as

$$0 = -A^T(\theta)k(\theta) - k(\theta)A(\theta) + k(\theta)\tilde{S}(\theta)k(\theta) - \tilde{Q}, \quad m(\theta) = (A^T(\theta) + k(\theta)\tilde{S}(\theta))^{-1}(\tilde{Q}x_D - k(\theta)\psi(\theta)),$$

where  $\tilde{S}(\theta) = B(\theta)\tilde{R}^{-1}B^T(\theta)$ . These equations vary with the true anomaly angle instead of being constant for short periods of time as for BST. For the Carter–Humi control schemes, the differential equations are defined by Equation (10), where  $\kappa = 1/(1 + e \cos \theta)$ . The state and control vectors are

$$x'(\theta) = A(\theta)x(\theta) - \tilde{S}(\theta)k(\theta)x(\theta) - \tilde{S}(\theta)m(\theta) + \psi(\theta), \quad u(\theta) = -\tilde{S}(\theta)(k(\theta)x(\theta) + m(\theta)).$$

The same procedure to calculate the values for the RE, ARE, state vectors, and control vectors will be applied for this scheme, but the true anomaly angle will vary with the position of the satellites.

### 10. Reference and maneuvering satellite

The reference and maneuvering satellite must be chosen such that the linearized TH equations can be used to maintain the separation distance constraints [Capó-Lugo 2005; Capó-Lugo and Bainum 2005b; 2005a]. The maneuvering satellite is assumed to have an applied thrust along its reference axes to correct the separation distance with respect to the reference satellite. With only the Earth’s oblateness (J2) perturbation, the satellites near the centroid (SA-SH) maintain the separation distance constraints for a long period of time, but SA-SB and SA-SC violate the separation distance constraints in 6 complete orbits for phase I and must be corrected first as shown in [Capó-Lugo 2005; Capó-Lugo and Bainum 2005b]. Since SH-SA does not violate the separation distance constraints for a simulation time of 30 days, these two satellites are nominally on their orbits and can be used as references to correct the drift of the other two satellites, namely, SB and SC. For the in-plane motion, SA will be used as the reference satellite to correct the positions and velocities of the other two satellites along the semimajor axis, making SB and SC the maneuvering satellites. SH is not considered to correct its separation distance conditions, but it can be used to correct the out of plane motion of the other satellites.

The J2 perturbation has major effects on low-earth orbit (LEO) satellites, but, for this constellation, phase I is greatly disturbed at the perigee and apogee points because the altitude of the perigee point is very close to the Earth. This perturbation causes the constellation to violate the separation distance constraints in a limited number of complete orbits. Following [Mishne 2004] the J2 perturbation can be defined in component form as

$$\psi(\theta) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ f_x \\ f_y \\ f_z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -\frac{3}{2}J_2\frac{\mu}{r^2}\left(\frac{R_e}{r}\right)^2(1 - 3\sin^2 i \sin^2 \theta) \\ -3J_2\frac{\mu}{r^2}\left(\frac{R_e}{r}\right)^2(\sin^2 i \sin \theta \cos \theta) \\ -3J_2\frac{\mu}{r^2}\left(\frac{R_e}{r}\right)^2(\sin i \cos i \sin \theta) \end{bmatrix}, \quad (23)$$

where  $i$  is the inclination angle, and  $\theta$  is the true anomaly angle. For the NASA benchmark tetrahedron constellation, the inclination angle for phase I is equal to 18.5°. Equation (23) is the disturbance matrix defined in Equation (14) in terms of the true anomaly angle.

Nominal coordinates	Separation distance
$\xi_N = -8.6602543 \text{ km}$	$\xi_S = -9.928 \text{ km}$
$\zeta_N = 4.7416 \text{ km}$	$\zeta_S = 4.7 \text{ km}$
$\eta_N = 1.5865 \text{ km}$	$\eta_S = 1.5 \text{ km}$
$\xi'_N = 3.49665 \times 10^{-4} \text{ km/s}$	$\xi'_S = 3.51 \times 10^{-4} \text{ km/s}$
$\zeta'_N = 0 \text{ km/s}$	$\zeta'_S = -1.4 \times 10^{-5} \text{ km/s}$
$\eta'_N = 0 \text{ km/s}$	$\eta'_S = -9 \times 10^{-6} \text{ km/s}$

**Table 3.** Nominal coordinates and initial separation distance for SA-SB system in phase I.

To initialize the procedure to calculate the Riccati and adjoint Riccati equations, state variables, and the optimal control explained in the previous section, [Table 3](#) shows the nominal coordinates and the separation distance [[Capó-Lugo 2005](#); [Capó-Lugo and Bainum 2005b](#); [2005a](#)]. The first column illustrates the nominal separation distance coordinates for which the constellation satisfies the separation distance conditions at the apogee point. The second column shows the separation distance coordinates when the pair SA-SB first violates the separation distance criteria. These separation distances are obtained from [[Capó-Lugo 2005](#)] when the J2 perturbation is added into the STK simulations.

The relative drift from the reference satellite to the maneuvering satellite, in state-based variables  $(\xi, \zeta, \eta)$ , can be calculated as:

$$x(\theta_0) = \begin{bmatrix} \xi \\ \zeta \\ \eta \\ \xi' \\ \zeta' \\ \eta' \end{bmatrix} = \begin{bmatrix} \xi_S - \xi_N \\ \zeta_S - \zeta_N \\ \eta_S - \eta_N \\ \xi'_S - \xi'_N \\ \zeta'_S - \zeta'_N \\ \eta'_S - \eta'_N \end{bmatrix}. \quad (24)$$

[Equation \(24\)](#) represents initial conditions for system of differential equations defined by [Equation \(10\)](#), with the desired state vector  $x_D$  used in [Equations \(11\)](#) and [\(13\)](#) set equal to zero. With this scheme the satellites will reduce the relative drift in the separation distance and the velocity for a pair of satellites within the constellation such that, at the next apogee point, the satellites will satisfy the separation distance constraints.

## 11. Results of the active control laws

The two active control laws are studied to determine if there exists a difference between the methods to correct the drift between a pair of satellites. The main question is: how much weight do the  $Q$  and  $R$  matrices need? The  $Q$  and  $R$  matrices can have any values within the maximum limits on the magnitude of the control and can change the response in the system. Through these simulations, the matrices will be weighted in different ways.

Equations of motion are written in terms of the true anomaly angle, but the time that it takes to correct the drift between a pair of satellites can be determined with the following equations [[Massari et al. 2004](#)]:

$$\tan(E/2) = \sqrt{(1-e)/(1+e)} \tan(\theta/2), \quad n(t - t_\pi) = E - e \sin E, \quad (25)$$

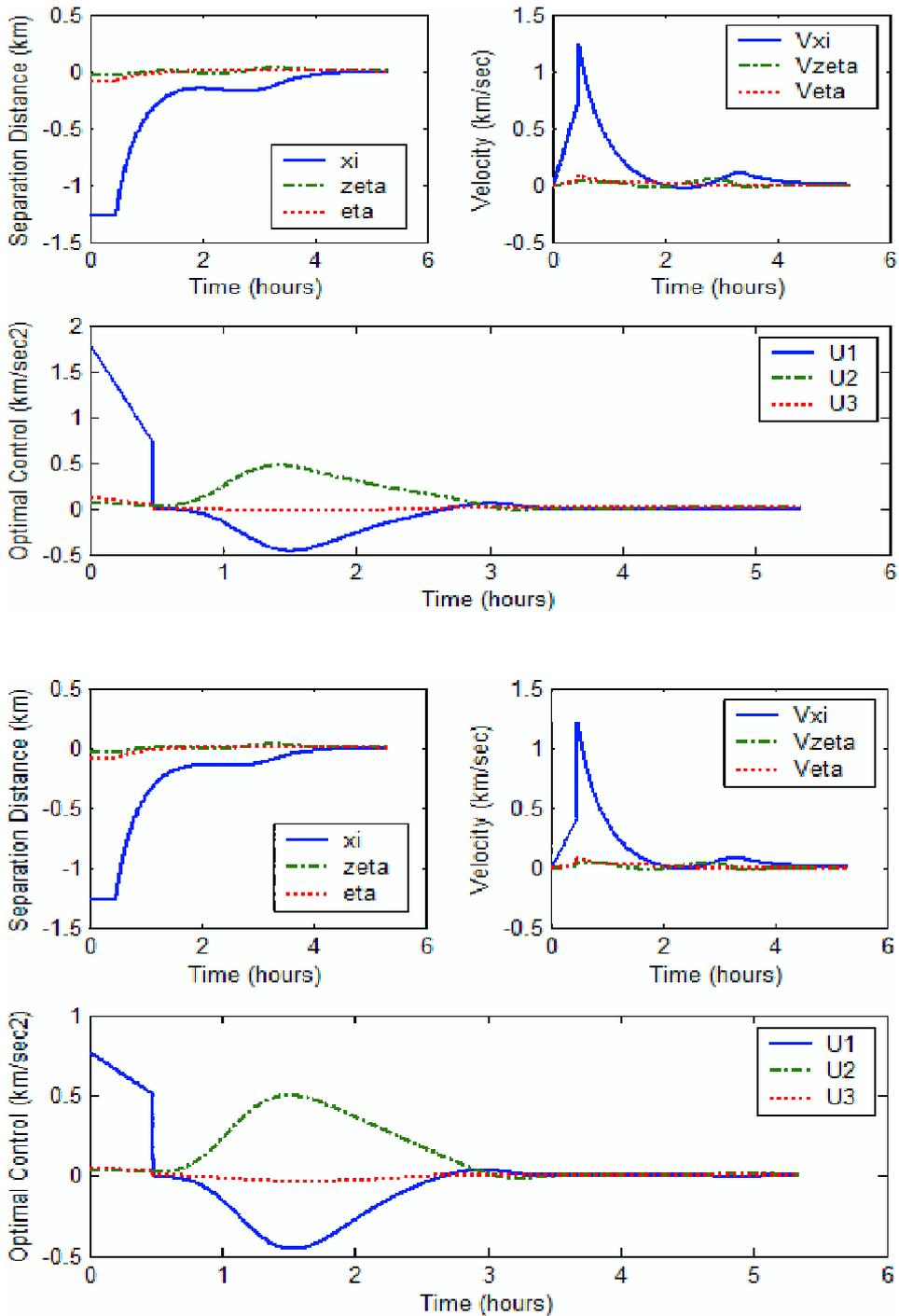
where  $E$  is the eccentric anomaly,  $n$  is the mean motion of a satellite in which  $n = \sqrt{\mu/a^3}$ ,  $t_\pi$  is the time at the perigee point, and  $t$  is the time of the satellite at some point in the orbit. This transformation is used to determine how much time is required for the satellites to make the corrections. The time is changed from seconds to hours because the period of the orbit is expressed in days as shown in Table 1, and a number of orbits can be obtained to understand when the maneuvering satellite finishes the corrections to the drift between a pair of satellites. The responses that will be obtained from the solution of the active control scheme will have particular units as follows: km for the correction in the separation distance, km/s for the velocity correction, and km/s<sup>2</sup> for the optimal control since it is expressed in terms of acceleration.

The optimal control  $u_j(\theta)$ , for  $j = 1, 2, 3$  also shows thrust levels because in Equation (10) the thrust  $T(\theta)$  is divided by the mass  $m(\theta)$  such that the terms can be expressed in terms of accelerations to develop the linearized TH equations. The directions of the axes in state variable format, defined by Equation (10), are as follows:  $\xi$  is positive against the motion of the spacecraft,  $\zeta$  is positive along the radial direction, and  $\eta$  is positive when the right hand system is completed. The thrust  $T(\theta)$  is specified in the same directions as  $\xi$ ,  $\zeta$ , and  $\eta$  in which the control function  $u(\theta)$  is defined over the same direction as the thrust  $T(\theta)$ .

The simulation begins at the apogee point where the constellation first violates the separation distance conditions and finishes at the following apogee point. Since the linear quadratic optimal controller is defined continuously in the true anomaly angle, the simulation can be expanded after one complete period, but, in some cases, the simulation time is shortened because, after the system of linear equations (10) comes into steady state, the same result is obtained through the complete simulation of the corresponding nonlinear equations. The Runge–Kutta method is used to integrate the linear equations forward in the true anomaly angle, where the step size is chosen to be 0.004 radians for this phase. This step size is small, but is used to diminish the error in the calculations. For the BST active control scheme, the  $\kappa$  term in Equation (10) is updated every 0.012 radians, but the active control scheme will depend mainly on the weighted matrices. In the simulations the following aliases are used: (1) xi  $\rightarrow$   $\xi$ , (2) zeta  $\rightarrow$   $\zeta$ , (3) eta  $\rightarrow$   $\eta$ , (4) Vxi  $\rightarrow$   $\xi'$ , (5) Vzeta  $\rightarrow$   $\zeta'$ , (6) Veta  $\rightarrow$   $\eta'$ , and (7)  $U1$ ,  $U2$ , and  $U3$  is the control vector in the directions of  $\xi$ ,  $\zeta$ , and  $\eta$  defined in Equation (10), respectively. This representation is used in the legend of all the simulations performed in this study.

The first set of initial conditions defined in Table 3 are used to determine the drift between the position and velocity of SA and SB using Equation (24). Table 2 details the values for the orbital elements defined in the state matrix  $A$ , the control matrix  $B$ , and Equation (25). The simulations are run with the BST active control scheme, and thereafter the Carter–Humi control scheme is implemented to compare the two different control techniques for the correction of the separation distance and velocity of a pair of satellites.

When  $Q = \text{diag}[20 \ 20 \ 20 \ 20 \ 20 \ 20]$  and  $R = \text{diag}[10 \ 10 \ 10]$  Figure 4 shows the results with the BST and CH active control schemes. The correction in the separation distance and the velocity of the maneuvering spacecraft shows that the system is stabilized in approximately 5 hr, but, for the CH active control scheme, the optimal control effort is less than in the BST active control scheme. The cost function



**Figure 4.** SA-SB separation distance correction using BST (top) and CH (bottom) active control scheme for  $Q = \text{diag}[20 \ 20 \ 20 \ 20 \ 20 \ 20]$  and  $R = \text{diag}[10 \ 10 \ 10]$ .

in the CH active control scheme varies with the true anomaly angle and uses the eccentricity term such that a minimum time and consumption problem can be obtained along the orbit.

If the  $Q$  matrix is split to weight the velocities  $\xi'$ ,  $\zeta'$ ,  $\eta'$  more than the positions  $\xi$ ,  $\zeta$ ,  $\eta$  while setting  $R = \text{diag}[20 \ 20 \ 20]$ , Figure 5 shows the results for the correction of the drift in the separation distance and velocities. In Figure 5 both control schemes show that the correction is going to take more time before a steady state response is obtained. In the BST active control scheme, the optimal control is going to be stabilized before 6 hr. In the CH active control scheme, the optimal control is stabilized after 8 hr, and the optimal control and the correction of the velocity shows lower levels than in the BST active control scheme. For the BST active control scheme, this weighting in the  $Q$  matrix will cause greater fuel expenditure, an undesired situation.

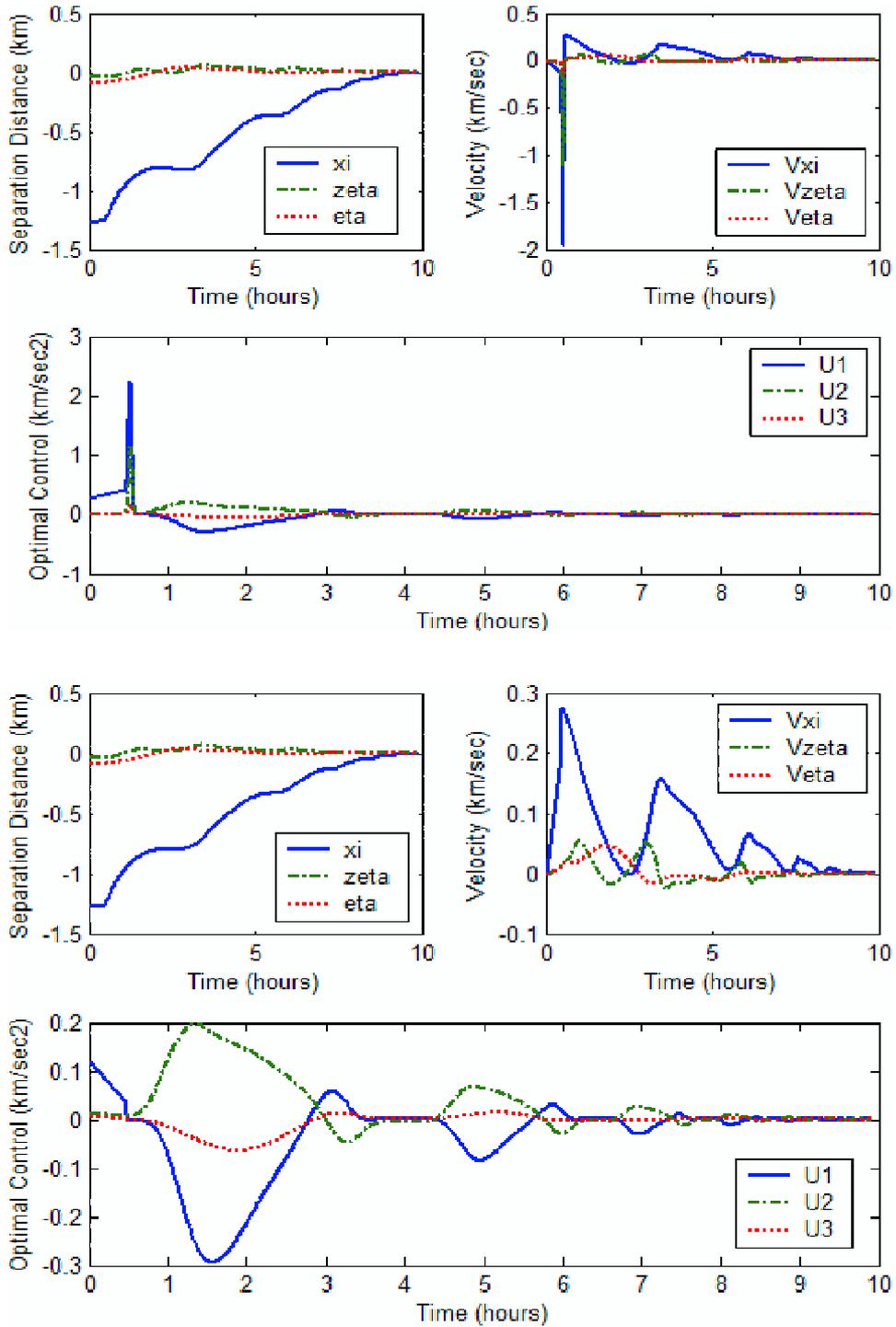
When the positions  $\xi$ ,  $\zeta$ ,  $\eta$  are weighted more than the velocities  $\xi'$ ,  $\zeta'$ ,  $\eta'$  Figure 6 shows the responses in the correction of the separation distance and velocity for the maneuvering spacecraft for both active control schemes. The weights for this case are  $Q = \text{diag}[20 \ 20 \ 20 \ 1 \ 1 \ 1]$  and  $R = \text{diag}[20 \ 20 \ 20]$ . For both active control schemes, the separation distance and the velocities are corrected faster compared to the previous case. The correction takes less than 2 hr for both active control schemes, but the CH active control schemes shows a lower level in the optimal control as well in the velocity.

Analyzing Equation (10), one sees that if the varying term in the  $A$  matrix is weighted more, then the active control scheme is going to take less time to stabilize the system; this is shown in Figure 6. On the other hand, when the velocities are weighted more, the system is going to take more time to stabilize the varying term; this is illustrated in Figure 5.

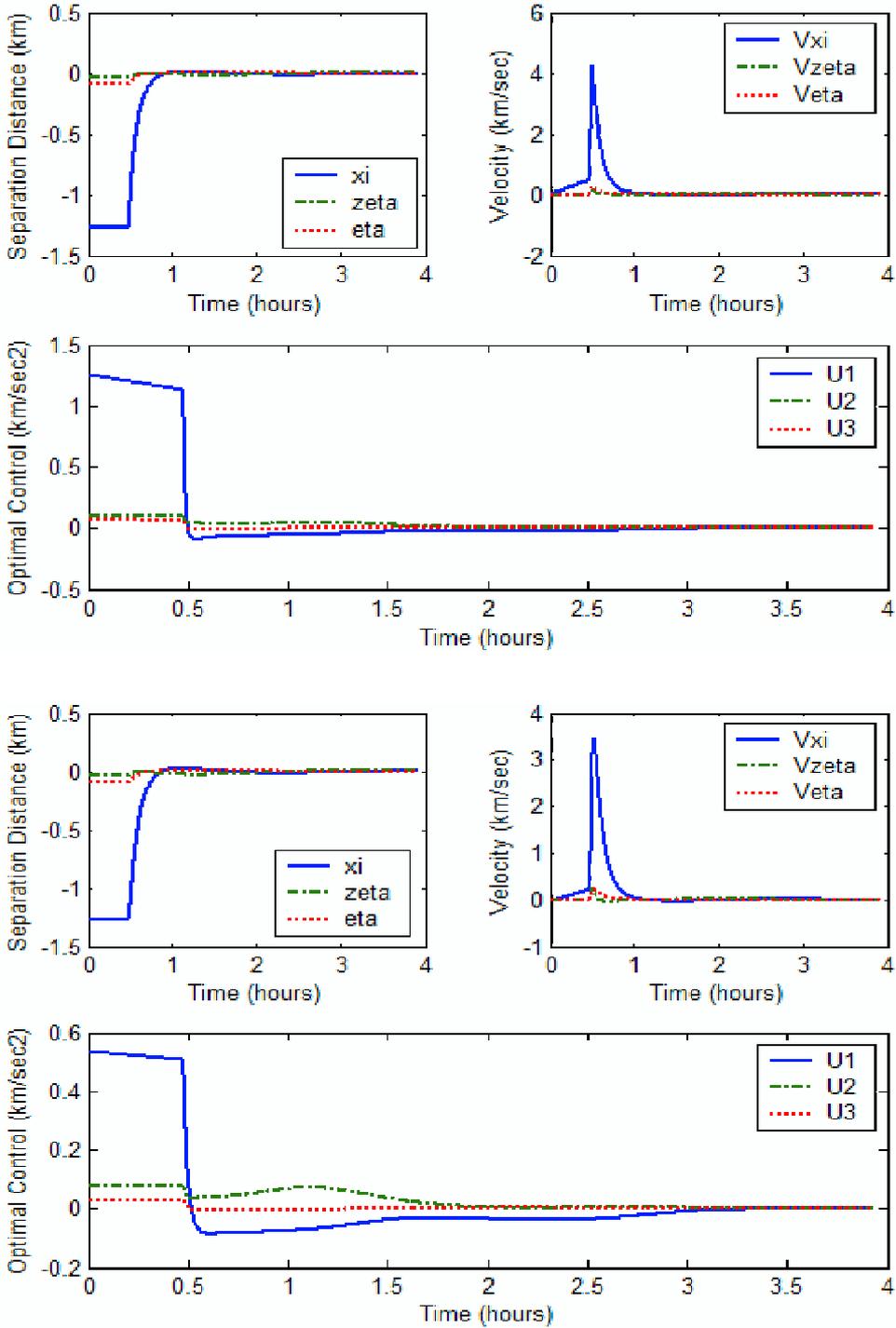
## 12. Examination of the drifts after the first correction

When the satellites move to the apogee point the constellation will satisfy the separation distance constraints, but the perturbations will still be present, and the constellation may hold for only a limited number of complete orbits. This situation may guarantee that the constellation will not violate the separation distance constraint at the following apogee point. The Satellite Tool Kit software [STK 2003] (STK) is used to propagate the constellation motion after the active control scheme has corrected the drift between a pair of satellites; in the last simulations, the correction in the drift of the separation distance is between SA and SC [Carpenter et al. 2003; Capó-Lugo 2005; Capó-Lugo and Bainum 2005b]. This pair of satellites violates the NASA benchmark tetrahedron conditions first. The STK motion is propagated from the apogee point after the correction is made, assuming that the satellites corrected the drift made by the J2 perturbation.

For phase I, Table 4 shows the initial conditions at the initial apogee point for the NASA benchmark tetrahedron constellation. The simulation for the correction in the positions and velocities between a pair of satellites for the BST and Carter–Humi active control schemes shows that the correction in the drift is less than  $1 \times 10^{-6}$  km for the separation distance and  $1 \times 10^{-6}$  km/sec for the velocities when the control schemes finish the correction. To start the simulation at the next apogee point, the drift is assumed to be  $1 \times 10^{-6}$  for the correction in the positions (km) and velocities (km/sec) between a pair of satellites after the correction is performed. This value is used because, if the correction for the J2 perturbation is made, the maneuvering satellites (SB and SC) satisfy the separation distance requirements at the following apogee point. This small drift is added into the initial conditions shown in Table 4 to determine if the



**Figure 5.** SA-SB separation distance correction using BST (top) and CH (bottom) active control scheme for  $Q = \text{diag}[1 \ 1 \ 1 \ 20 \ 20 \ 20]$  and  $R = \text{diag}[20 \ 20 \ 20]$ .

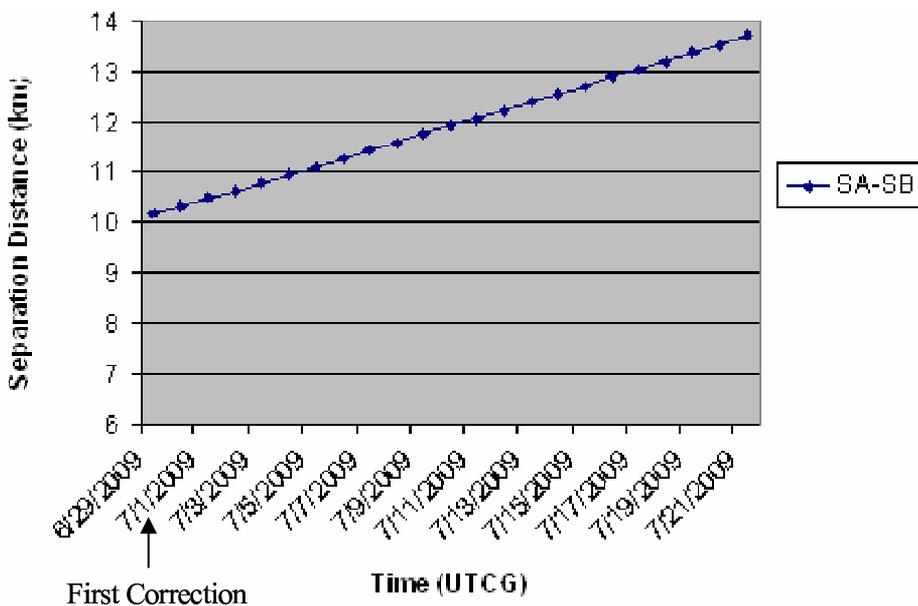


**Figure 6.** SA-SB separation distance correction using BST (top) and CH (bottom) active control scheme for  $Q = \text{diag}[20 \ 20 \ 20 \ 1 \ 1 \ 1]$  and  $R = \text{diag}[20 \ 20 \ 20]$ .

Satellite	$x$ (km)	$y$ (km)	$z$ (km)	$v_x$ (km/s)	$v_y$ (km/s)	$v_z$ (km/s)
SA	-8.66025403	-72,582.4525	-24,285.7489	0.973083288	0	0
SB	0	-72,587.1941	-24,287.3354	0.972733623	0	0
SC	0	-72,577.7109	-24,284.1624	0.973432881	0	0
SH	-2.88675134	-72,585.0433	-24,278.0058	0.973083324	0	0

**Table 4.** Initial coordinates and velocities for phase I.

constellation can be maintained for another six complete orbits. The constellation is propagated with  $J_2$  perturbation at the starting date of June 29, 2009 22:56 (UTCG). The orbit propagation will include only SA and SB because these satellites are considered in the correction of the drift. Figure 7 shows that the satellites in the in-plane motion will maintain the desired benchmark configuration for another 6 complete orbits before the constellation violates the separation distance constraints again. Also, the arrow indicates the time when the first correction is made. The constellation may violate the separation distance constraints earlier because the drift is assumed to be small at the apogee point, but, at least for six complete orbits, the constellation maintains the configuration. Moreover, the active control scheme ensures that the constellation will maintain the configuration at the following apogee point, even though the perturbations are still present when the correction is performed.



**Figure 7.** Satellite separation for in-plane motion with  $J_2$  perturbation only for phase I.

### 13. Conclusion

The NASA benchmark tetrahedron constellation will not satisfy the separation distance constraints with perturbations, and an active control scheme is needed to maintain the separation distance conditions. To define the motion of a pair of satellites about Earth in an elliptical orbit, the linearized Tschauner–Hempel (TH) equations are used. The linear quadratic regulator (LQR) technique is used as an active control strategy in which two different control schemes are used to maintain the separation distance constraints of the NASA Benchmark tetrahedron constellation depending on the varying term in the TH equations. The Bainum, Strong, and Tan (BST) active control scheme adapts in a piecewise manner the varying term in the TH equations; on the other hand, a different LQR control scheme is defined for the cost function using the Carter–Humi (CH) approach. The CH technique developed a different cost function based on the true anomaly angle and eccentricity of the orbit.

When the simulations to correct the drift for the first specific size (phase I) of the tetrahedron constellation are performed, the LQR control scheme shows a (relative) impulsive type response for both control schemes, but the CH active control scheme shows a lower thrust level than in the BST active control scheme because the CH active control scheme is varying with respect to the true anomaly angle and is explained in terms of the eccentricity. Furthermore, the eccentricity must be part of the equations of motion to obtain a better approximation of the ellipse. [Marec \[1979\]](#) established that for an eccentric orbit, the true anomaly angle must be defined in the cost function to obtain a minimum time and consumption problem. Given these results the LQR using the CH approach gives a lower consumption problem than in the BST control scheme because the BST control scheme keeps constant the varying term for short periods of time.

After the first correction is performed to maintain the separation distance conditions, the simulations show that the pair of satellites analyzed in this paper is going to satisfy the separation distance constraints for at least another 6 complete orbits before the constraints are violated again. For the simulations already presented in this paper, the pair of satellites within the constellation is going to satisfy the separation distance constraints at the next apogee point.

### References

- [Athans and Falb 1966] M. Athans and P. L. Falb, *Optimal control: an introduction to the theory and its applications*, McGraw-Hill, New York, 1966.
- [Bainum et al. 2004] P. M. Bainum, Z. Tan, and X. Duan, “[Review of station keeping strategies for elliptically orbiting constellations in along-track formation](#)”, pp. 350–353 in *Eighth pan american congress of applied mechanics* (Havana, Cuba), vol. 10, January 5–9 2004. also in *International Journal of Solids and Structures*, **42**: 21–22 (October 2005), pp. 5683–5691, PACAM VIII Special Issue.
- [Borse 2000] G. J. Borse, *Numerical methods with MATLAB, a resource for scientists and engineers*, International Thompson Publishing, Boston, 2000.
- [Capó-Lugo 2005] P. A. Capó-Lugo, “Strategies and control schemes to satisfy the separation distance constraints for the NASA benchmark tetrahedron constellation”, Master’s Thesis, Howard University, Washington D. C., December 2005.
- [Capó-Lugo and Bainum 2005a] P. A. Capó-Lugo and P. M. Bainum, “Implementation of the strategy for satisfying distance constraints for the NASA benchmark tetrahedron constellations”, pp. 1463–1482 in *Proceedings of the AAS/AIAA astrodynamics conference* (South Lake Tahoe, CA), edited by B. G. Williams et al., *Advances in the astronomical sciences* **123**, Univelt, San Diego, CA, August 7–11 2005. Paper AAS 05-344.

- [Capó-Lugo and Bainum 2005b] P. A. Capó-Lugo and P. M. Bainum, “Strategy for satisfying distance constraints for the NASA benchmark tetrahedron constellation”, pp. 775–793 in *Proceedings of the 15th annual AAS/AIAA spaceflight mechanics meeting* (Copper Mountain, CO), edited by D. A. Vallado et al., Advances in the astronautical sciences **120**, Univelt, San Diego, CA, January 23–27 2005. Paper AAS 05-153.
- [Carpenter et al. 2003] J. R. Carpenter, J. A. Leitner, and R. D. Folta, David C. and Burns, “Benchmark problems for spacecraft formation flying missions”, in *AIAA guidance, navigation, and control conference and exhibit* (Austin, TX), August 11–14 2003. Paper AIAA-2003-5364.
- [Carter 1990] T. Carter, “New form for the optimal rendezvous equations near a keplerian orbit”, *J. Guid. Control Dynam.* **13**:1 (1990), 183–186.
- [Carter and Brient 1992] T. Carter and J. Brient, “Fuel-optimal rendezvous for linearized equations of motion”, *J. Guid. Control Dynam.* **15**:6 (1992), 1411–1416.
- [Carter and Humi 1987] T. Carter and M. Humi, “Fuel-optimal rendezvous near a point in general keplerian orbit”, *J. Guid. Control Dynam.* **10**:6 (1987), 567–573.
- [Gerald and Wheatley 2004] C. F. Gerald and P. O. Wheatley, *Applied numerical analysis*, 7th ed., Addison-Wesley, New York, 2004.
- [Marec 1979] J. P. Marec, *Optimal space trajectories*, vol. 1, Studies in astronautics, Elsevier Scientific, Amsterdam, 1979.
- [Massari et al. 2004] M. Massari, R. Armellin, and A. E. Finzi, “Optimal trajectory generation and control for reconfiguration maneuvers of formation flying using low-thrust propulsion”, pp. 2461–2474 in *Proceedings of the 14th annual AAS/AIAA spaceflight mechanics meeting* (Maui, Hawaii), edited by S. L. Coffey et al., Advances in the astronautical sciences **119**, Univelt, San Diego, CA, February 8–12 2004. Paper AAS 04-258.
- [Mishne 2004] D. Mishne, “Controlling the out-of-plane motion of a follower satellite in a periodical relative trajectory, using angular rate information”, in *AIAA/AAS astrodynamics specialist conference and exhibit* (Providence, RI), August 16–19 2004. Paper AIAA 2004-5215.
- [Phillips and Nagle 1995] C. L. Phillips and H. T. Nagle, *Digital control system analysis and design*, 3rd ed., Prentice Hall, Englewood Cliffs, NJ, 1995.
- [STK 2003] *STK, Satellite Tool Kit Software*, Analytical Graphics, Inc., Malvern, PA, 2003. Version 5.0.
- [Strang 1986] G. Strang, *Introduction to applied mathematics*, Wellesley-Cambridge Press, Wellesley, Mass., 1986.
- [Strong 2000] A. Strong, *On the deployment and station keeping dynamics of n-body orbiting satellite constellations*, Ph.D. Dissertation, Howard University, Washington D. C., 2000.
- [Tan et al. 1999] Z. Tan, P. M. Bainum, and A. Strong, “A strategy for maintaining distance between satellites in an orbiting constellation”, pp. 343–354 in *Proceedings of the 9th Annual AAS/AIAA space flight mechanics meeting* (Breckenridge, CO), edited by R. H. Bishop et al., Advances in the astronautical sciences **102**, Univelt, San Diego, CA, February 7–10 1999. Paper AAS 99-125.
- [Tan et al. 2002] Z. Tan, P. M. Bainum, and A. Strong, “The implementation of maintaining constant distance between satellites in coplanar elliptic orbits”, *J. Astronaut. Sci.* **50**:1 (2002), 53–69.

Received 16 May 2006. Accepted 20 Apr 2007.

PEDRO A. CAPÓ-LUGO: [pcapo@howard.edu](mailto:pcapo@howard.edu)

Department of Mechanical Engineering, Howard University, 2300 Sixth Street NW, Washington, DC 20059, United States

PETER M. BAINUM: [pbainum@fac.howard.edu](mailto:pbainum@fac.howard.edu)

Department of Mechanical Engineering, Howard University, 2300 Sixth Street NW, Washington, DC 20059, United States

## EVALUATION OF EFFECTIVE MATERIAL PROPERTIES OF RANDOMLY DISTRIBUTED SHORT CYLINDRICAL FIBER COMPOSITES USING A NUMERICAL HOMOGENIZATION TECHNIQUE

HARALD BERGER, SREEDHAR KARI, ULRICH GABBERT, REINALDO RODRÍGUEZ RAMOS,  
JULIAN BRAVO CASTILLERO AND RAÚL GUINOVART DÍAZ

In this paper effective material properties of randomly distributed short fiber composites are calculated with a developed comprehensive tool for numerical homogenization. We focus on the influence of change in volume fraction and length/diameter aspect ratio of fibers. Two types of fiber alignments are considered: fiber orientations with arbitrary angles and parallel oriented fibers. The algorithm is based on a numerical homogenization technique using a unit cell model in connection with the finite element method. To generate the three-dimensional unit cell models with randomly distributed short cylindrical fibers, a modified random sequential adsorption algorithm is used, which we describe in detail. For verification of the algorithm and checking the influence of different parameters, unit cells with various fiber embeddings are created. Numerical results are also compared with those from analytical methods.

### 1. Introduction

Short fiber composites can be easily produced and have good mechanical properties. Since the mixture of short fibers and liquid resin can be manufactured by injection or compression molding, the production of parts with nearly arbitrary and very complicated shapes is possible. Composites consisting of spatially distributed short fibers have become popular in a wide variety of applications. Moreover, using spatial short fibers as reinforcing elements in a controlled manner can provide more balanced properties, which lead to an improved through-the-thickness stiffness/strength.

A classical problem in solid mechanics is the determination of effective elastic properties of a composite material made up of a statistically isotropic random distribution of isotropic and elastic short cylindrical fibers embedded in a continuous, isotropic and elastic matrix. Even though analytical and semianalytical models have been developed to homogenize fiber composites, they are often applicable only to specific cases. Numerical models seem to be a well-suited approach to describe the behavior of these materials, because there are no restrictions on the geometry, on material properties, on the number of phases (constituents) and on size. In order to obtain realistic predictions of a new material, micro-macro considerations are the appropriate approach. In this context the finite element method has been used to determine effective properties of the short fiber composites based on unit cell models.

---

*Keywords:* finite element method, unit cell, representative volume element, homogenization, short fibers, random sequential adsorption algorithm, effective material properties.

This work was supported by DFG Germany, Graduiertenkolleg 828 *Micro-Macro Interactions in Structured Media and Particle Systems*. This support is gratefully acknowledged.

A number of classical micromechanics theories have been developed. Using variational principles, Hashin [1962] and Hashin and Shtrikman [1963] established bounds on materials that could be considered as *mechanical mixtures of a number of different isotropic and homogeneous elastic phases* which are then treated as statistically isotropic and homogeneous. These two-point bounds were improved by three-point bounds [Milton 1982; Milton and Phan-Thien 1982], which incorporate information about the phase arrangement through the statistical correlation parameters. The dilute approximation can be used to model a dilute suspension of spherical elastic particles in continuous elastic phases. The interaction between particles is neglected. So the algorithm reduces to that of solving the problem of a spherical inclusion in an infinite matrix subjected to hydrostatic loading at infinity. Eshelby [1957; 1959] considered the problem of an ellipsoidal inclusion in an infinite isotropic matrix, assuming a well-defined matrix. That, however, is not always true in polycrystalline materials. A variety of properties can be exhibited, but there is no clearly defined matrix phase. In these cases the interactions between particles are more significant. The Mori–Tanaka method [Mori and Tanaka 1973] was designed to calculate the average internal stress in the matrix containing precipitates with eigenstrains. Benveniste [1987] reformulated it so that it could be applied to composite materials. He considered isotropic phases and ellipsoidal phases. Recently, Segurado and Llorca [2002] and Böhm et al. [2002] have assessed the effective coefficients of randomly distributed spherical particles using random sequential adsorption method and compared them with Hashin–Shtrikman bounds and other results from literature. Gusev et al. [2000] and Lusti et al. [2002] performed experiments of randomly distributed short cylindrical fiber composites and found good agreement with numerical results. However, due to the lack of literature which deals with randomly distributed short cylindrical fibers and the restriction to low volume fractions of fibers, we have been motivated to develop a numerical homogenization tool which extends the limits and provides the basis for investigation of composites with arbitrary inclusions. In our opinion micro-macro mechanical approaches offer new insights in the material behavior of such fiber composites, and may result in new procedures to develop realistic material models for design and optimization purposes.

## 2. Numerical homogenization

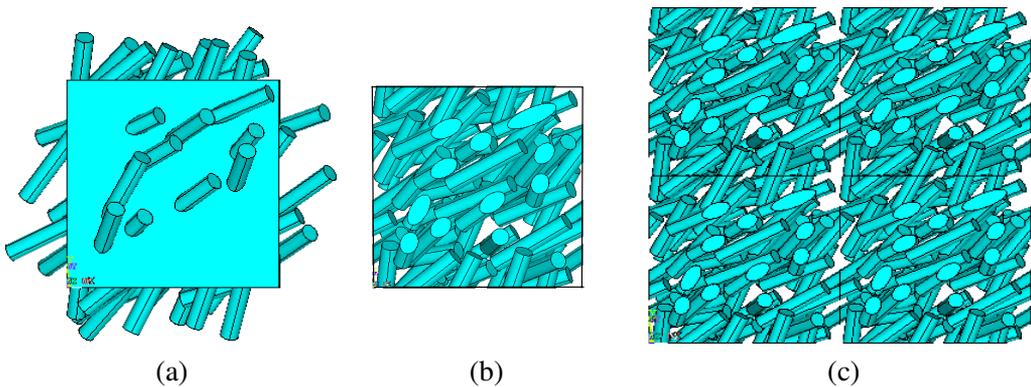
**2.1. Basic procedure.** The mechanical and physical properties of the constituent materials are always regarded as a small-scale/micro structure. To predict the overall behavior of the structure on a macro level, the knowledge of effective material properties is necessary. One of the most powerful tools to estimate such effective properties is the homogenization method. The main idea is to find a globally homogeneous medium equivalent to the original composite, such that the strain energy stored in both systems is approximately the same. The common approach to model the macroscopic properties of fiber composites is to create a unit cell or a representative volume element (RVE) that captures the major features of the underlying microstructure.

The RVE can generally be considered as a periodic part of the heterogeneous structure that is sufficiently large to be a statistically representative of the composite, that is, to effectively include a sampling of all microstructural heterogeneities that occur in the composite [Kanit et al. 2003]. To obtain the homogenized effective material properties, periodicity must be ensured for the mechanical behavior of the RVE by introducing periodic boundary conditions between opposite surfaces. By constructing several load cases with selected traction loads and selected shear loads in one direction and preventing strains in

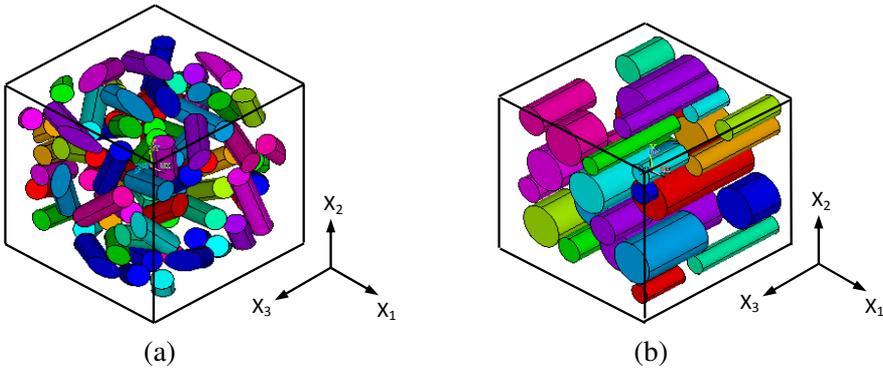
the other directions, all effective elasticity coefficients can be calculated from the constitutive relations by an averaging technique. This procedure is described in detail in [Berger et al. 2005a; 2005b] and shall not be explained here.

**2.2. Fiber generation by random sequential adsorption algorithm.** Creation of an RVE with randomly distributed cylindrical short fibers which fulfill certain restrictions, such as nonoverlapping, ensuring periodicity and the like, is a difficult task. Due to the statistical distribution of inclusions, the RVE can be modeled as a cube with unit size. For automatic generation of such RVEs, a modified random sequential adsorption algorithm [Hinrichsen et al. 1986] is used. Several input parameters can be given, including the size of RVE, diameter and length range of fibers, minimum distance between neighboring fibers, and desired volume fraction. The algorithm starts by creating the cylinder axis of the first fiber at a random position, with random length and with random angle. Subsequently new fibers are created with random distribution values. If the new fiber matches the restriction of nonoverlapping and sufficient distance to the earlier one, it is accepted; otherwise it is deleted. Furthermore, to ensure periodicity, if any surface of the cylinder cuts any of the cubic RVE surfaces it is copied to the opposite surface with the RVE size length. In this case one also checks all the restrictions; if it fails, the original and copied fibers are deleted. Concerning the later finite element generation, we would also like to ensure that some practical limitations are fulfilled. For instance, the cylinder surfaces should not be very close to the RVE surface as well as to corners of the RVE in order to avoid highly distorted finite elements during meshing. The generation of new fibers is repeated until the desired volume fraction is reached or no more fibers can be placed due to the aforementioned restrictions. Figure 1 shows a sample of generated fibers before cutting on the RVE surface, after cutting, and an ensemble of four RVEs which demonstrate that periodicity is maintained.

By modifying the input parameters it is possible to create RVEs with different fiber arrangements, such as, for example, fibers of same diameter, fibers of same length, or only parallel alignment of fibers. Combining these arrangements opens the possibility of generating RVEs which represent different types of fiber reinforced composites such as those presented in this paper. The possible maximum fiber volume fraction plays an important role. In general, for fibers of identical size the algorithm can generate up to



**Figure 1.** Generation of randomly distributed short fibers: (a) uncut fibers, (b) cut fibers, (c) periodicity demonstrated with four RVEs.



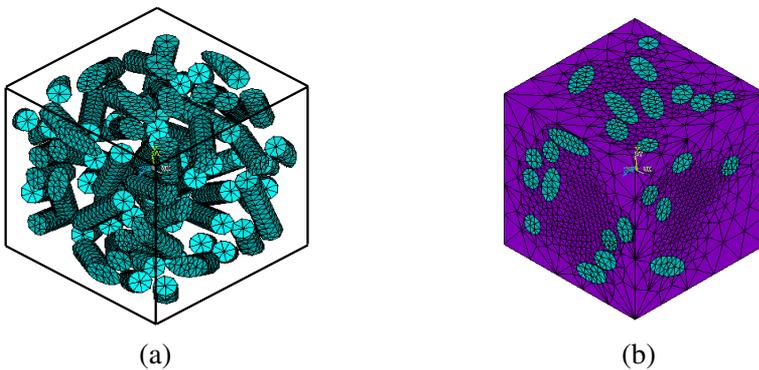
**Figure 2.** Types of composites: (a) ROF and (b) POF.

25% fiber volume fraction. For higher volume fractions, one must use fibers of different sizes, which can be generated by creating fibers with subsequently descending diameters. Using this approach, fiber volume fraction up to 40% can be achieved with minimum distortion of the finite elements.

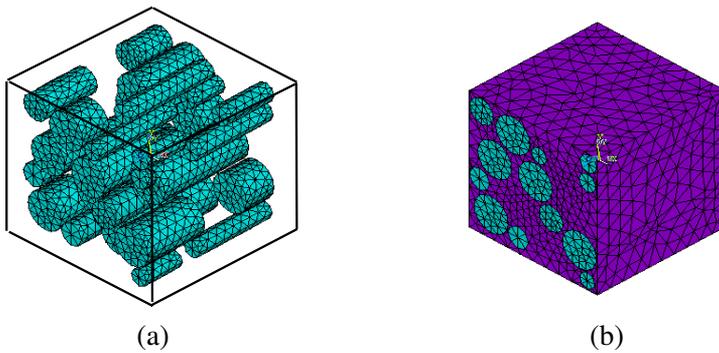
For calculating effective material properties of randomly distributed short fiber composites, we investigate two types of fiber arrangements: randomly oriented fibers (ROF) and parallel-oriented fibers (POF); see Figure 2. For POF composites, the fibers in the models are aligned along the  $x_3$ -axis. This is denoted as the longitudinal direction, while the perpendicular  $x_1x_2$  plane are the transverse directions.

**2.3. Finite element modeling.** All finite element calculations were performed with the commercial FE package ANSYS. The matrix and the fibers were meshed with 10 node tetrahedron elements with full integration. For the calculation of geometry of the fibers by random sequential adsorption algorithm, a special preprocessor was developed in FORTRAN programming language, which produces a partial input file for ANSYS. Cutting of the fibers on the RVE surfaces is carried out by geometrical modeling features of ANSYS. Figures 3 and 4 show samples of meshed RVEs for ROF and POF models.

To apply the periodic boundary conditions on the RVE, identical meshes are necessary on opposite surfaces. For this purpose a surface of the RVE is first meshed with blind plane elements; then this



**Figure 3.** Sample for meshed RVE for ROF: (a) only fibers, (b) fibers and matrix.



**Figure 4.** Sample for meshed RVE for POF: (a) only fibers, (b) fibers and matrix.

element configuration is copied to the opposite surface. Based on all meshed surfaces, three-dimensional meshing is carried out. In order to apply periodic boundary conditions, we must generate constraint equations between opposite nodal pairs. Here the ANSYS Parametric Design Language (APDL) is used to automate this process. This script language allows the common nodal pairs to be identified automatically by their coordinates. Furthermore, APDL is used to collect averaged stresses and strains from element solution as well as to calculate the effective elastic constants.

The combination of the FORTRAN preprocessor, APDL and ANSYS batch processing lets us automate the whole process. It also provides a powerful tool for the fast calculation of homogenized material properties for composites with a great variety of inclusion geometries.

### 3. Test models

We have investigated two types of short fiber composites: ROF and POF. In order to test the influence of various parameters, different RVEs were generated. Furthermore so that we can obtain statistically-averaged results for every configuration, five RVEs with different starting values for the random algorithm were generated. The material properties of the constituents used for the analysis to evaluate the effective material properties were taken from literature [Böhm et al. 2002] to verify the developed method with other solutions. Table 1 contains Young’s moduli and Poisson’s ratios for matrix and fibers.

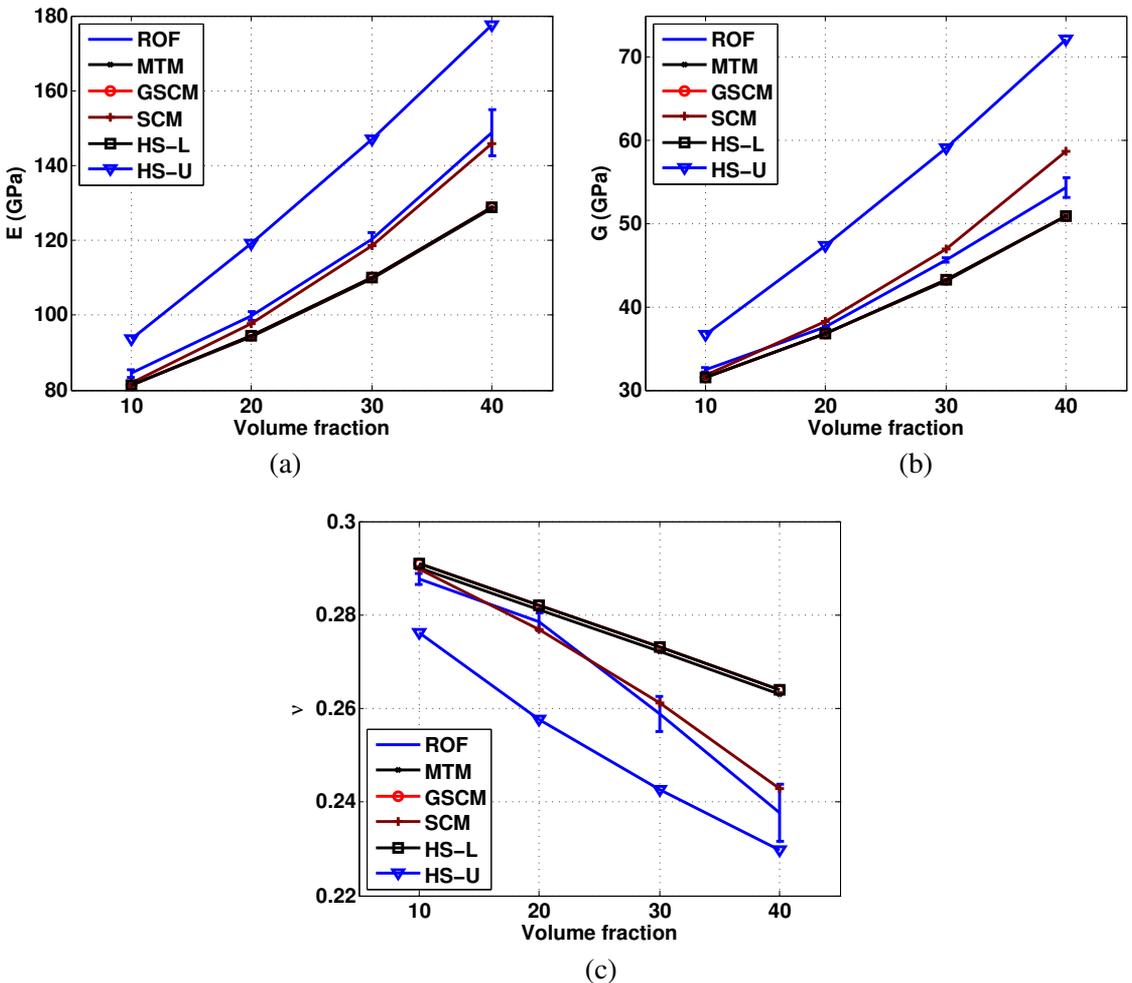
The calculated results were compared with different analytical methods such as Hashin–Shtrikman two-point bounds (HS) [Hashin and Shtrikman 1963], Mori–Tanaka estimates (MTM) [Mori and Tanaka 1973], the self-consistent method (SCM) [Li and Wang 2005], and the generalized self-consistent method (GSCM) [Christensen and Lo 1979]. We also performed studies to determine the influence of aspect ratio length/diameter of fibers on the effective material properties of these composites.

Constituent	Young’s modulus	Poisson’s ratio
Matrix Al2618-T4	70 GPa	0.3
Fiber SiC	450 GPa	0.17

**Table 1.** Material constants for constituents of the composite.

### 4. Results and discussion

**4.1. Variation of volume fraction.** Effective values of Young’s modulus  $E$ , shear modulus  $G$  and Poisson’s ratio  $\nu$  were evaluated for different volume fractions from 10% to 40% in steps of 10%; see Figure 5. Five samples of RVE models with randomly distributed short fibers (random angle of orientation, random diameter and length in a certain range) were generated for each volume fraction. In six particular load cases the RVEs were subjected to uniaxial tension as well as shear deformation along the three coordinate axes [Berger et al. 2005a; 2005b]. From these load cases nine material constants were calculated:  $E_{11}$ ,  $E_{22}$ ,  $E_{33}$ ,  $G_{12}$ ,  $G_{13}$ ,  $G_{31}$ ,  $\nu_{12}$ ,  $\nu_{23}$ ,  $\nu_{31}$ . Because of the statistically isotropy mean values  $E$ ,  $G$  and  $\nu$  from all directions were used for comparison with other methods. Furthermore, due to the random distribution



**Figure 5.** Variation of effective material properties for ROF composites with change in volume fraction and comparison with different analytical results: (a) Young’s modulus, (b) shear modulus, (c) Poisson’s ratio.

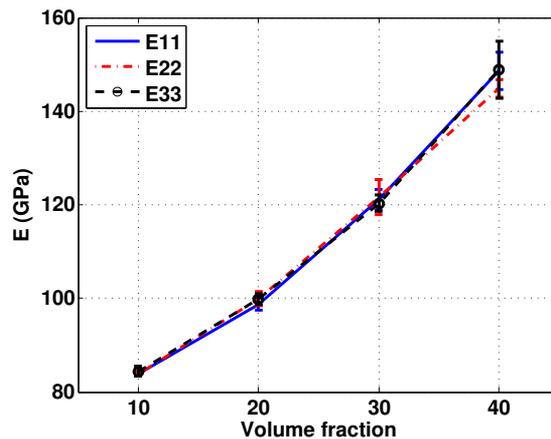
of fibers in each five samples, a certain variance can be observed for the effective values. The bounds of this variance are marked in the figures with vertical bars in the sense of a standard deviation.

The effective material properties, which were obtained for ROF composites using the presented numerical homogenization technique, lie within the lower (HS-L) and upper (HS-U) Hashin–Shtrikman bounds. The results of the analytical methods MTM and GSCM are always the same and are nearly identical with HS-L. The results of our solution (ROF) are nearer to the self-consistent method (SCM) for all volume fractions. The maximum difference between ROF and SCM is about 3%.

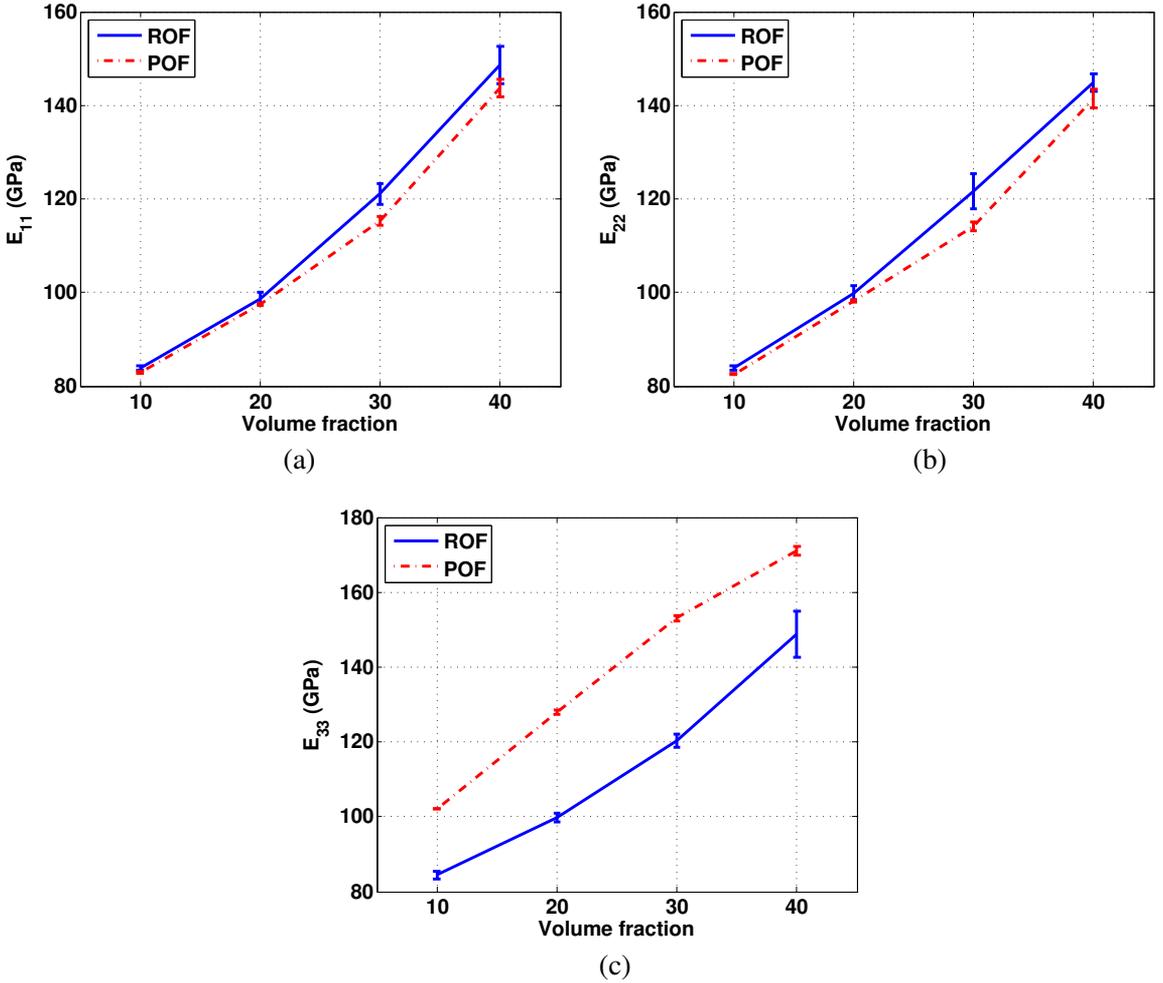
To show the nearly isotropic behavior of the ROF composite, in Figure 6 we plot effective Young's moduli in all coordinate directions as mean values from the five random samples for different volume fractions. Effective Young's moduli, which were obtained for the three coordinate directions, are nearly the same over the full investigated range of volume fraction; the maximum difference is less than 1.5%. This indicates a nearly isotropic macro behavior of the short fiber composite with randomly distributed fiber orientation.

Effective material properties obtained for POF composites were compared with ROF composites. Figure 7 shows the variation of effective Young's moduli for POF composites with change in volume fraction in three coordinate directions, and compares it with the results for ROF composites. The transverse Young's moduli of POF composites have slightly lower values compared to ROF composites. Nevertheless, along the longitudinal direction the POF effective material properties have higher values when compared with ROF composites. This is obvious because in case of POF composites, fibers are aligned along the longitudinal direction, which results in higher stiffness relative to the transverse directions. From Figure 7 it can also be seen that the effective Young's moduli  $E_{11}$  and  $E_{22}$  are nearly the same; this fact expresses transverse isotropy.

**4.2. Variation of fiber aspect ratio.** We have investigated the influence of aspect ratio length/diameter  $L/D$  of fibers on their effective material properties for ROF and POF composites. As  $L/D$  increases, the composite tends to a long fiber composite. The effective material properties were calculated at 10% volume fraction of fibers. Table 2 represents the variation of effective material constants  $E_{11}$ ,  $E_{22}$  and



**Figure 6.** Isotropy of effective material properties expressed by Young's moduli in three directions for ROF composites.



**Figure 7.** Variation of effective Young's moduli for all three directions with change in volume fraction for ROF and POF composites.

$E_{33}$  with change in aspect ratio  $L/D$  of the fibers for ROF and POF composites. From Table 2 it can be observed that with respect to change in aspect ratio of fibers, there are no significant variations in effective Young's moduli along the three coordinate directions for ROF composites. This is not true for POF composites, which show a significant variation in  $E_{33}$  with the increase in aspect ratio of fibers. Along the transverse direction,  $E_{11}$  and  $E_{22}$  of POF composites are slightly less than these parameters for ROF composites, but variations in the transverse Young's moduli with respect to the aspect ratio of fibers are not significant.

## 5. Conclusions

Numerical homogenization tools have been developed and presented for the evaluation of the effective material properties of short fiber reinforced composites. The effective material properties of randomly oriented fiber (ROF) and parallel-oriented fiber (POF) composites were obtained using these tools and

Aspect ratio $L/D$	$E_{11}$ (ROF)	$E_{11}$ (POF)	$E_{22}$ (ROF)	$E_{22}$ (POF)	$E_{33}$ (ROF)	$E_{33}$ (POF)
1	83.73	83.53	83.79	83.21	83.85	83.95
3	83.81	82.68	83.77	82.34	84.80	92.96
6	84.57	82.54	84.27	81.98	82.94	98.85
9	85.17	82.16	83.53	82.19	84.44	103.98
12	83.74	81.96	83.85	82.01	83.92	104.36

**Table 2.** Variation of effective Young's moduli (in GPa) with change in aspect ratio of fibers length/diameter ( $L/D$ ) for ROF and POF composites at 10% volume fraction.

compared with the results of different analytical methods. Our numerical predictions fit between the Hashin–Shtrikman bounds and are close to the results of the self-consistent approximation. We have also studied the influence of the aspect ratio of fibers on the effective material properties. These studies showed that there is no significant influence on effective material properties with increase of aspect ratio for ROF composites. However, POF composites show that along the longitudinal direction of the fibers the material behavior becomes stiffer as the aspect ratio increases.

Our investigation provides an insight into the more complex investigation of influencing factors for the macro behavior of fiber reinforced composites. We have shown that our method is reliable and offers the possibility for treatment of composites with arbitrary inclusions, for example, spheres and ellipsoids, with random distribution. Moreover, it allows the investigation of composites with more than two phases. The use of a modified random sequential adsorption algorithm allows the inclusions with different sizes to be generated so as to attain high volume fractions typical for many real composites.

The developed procedure, which combines a special geometrical preprocessor, ANSYS Parametric Design Language and ANSYS batch processing, provides a comprehensive tool for calculation of effective material properties of composites in a highly automated manner.

## References

- [Benveniste 1987] Y. Benveniste, "A new approach to the application of Mori–Tanaka's theory in composite materials", *Mech. Mater.* **6**:2 (1987), 147–157.
- [Berger et al. 2005a] H. Berger, S. Kari, U. Gabbert, R. Rodriguez-Ramos, J. Bravo-Castillero, and R. Guinovart-Diaz, "Calculation of effective coefficients for piezoelectric fiber composites based on a general numerical homogenization technique", *Compos. Struct.* **71**:3–4 (2005), 397–400.
- [Berger et al. 2005b] H. Berger, S. Kari, U. Gabbert, R. Rodriguez-Ramos, R. Guinovart, J. A. Otero, and J. Bravo-Castillero, "An analytical and numerical approach for calculating effective material coefficients of piezoelectric fiber composites", *Int. J. Solids Struct.* **42**:21–22 (2005), 5692–5714.
- [Böhm et al. 2002] H. J. Böhm, A. Eckschlagler, and W. Han, "Multi-inclusion unit cell models for metal matrix composites with randomly oriented discontinuous reinforcements", *Comput. Mater. Sci.* **25**:1–2 (2002), 42–53.
- [Christensen and Lo 1979] R. M. Christensen and K. H. Lo, "Solutions for effective shear properties in three phase sphere and cylinder models", *J. Mech. Phys. Solids* **27**:4 (1979), 315–330.
- [Eshelby 1957] J. D. Eshelby, "The determination of the elastic field of an ellipsoidal inclusion and related problems", **241**:1226 (1957), 376–396.
- [Eshelby 1959] J. D. Eshelby, "The elastic field outside an ellipsoidal inclusion", **252**:1271 (1959), 561–569.
- [Gusev et al. 2000] A. A. Gusev, P. J. Hine, and I. M. Ward, "Fiber packing and elastic properties of a transversely random unidirectional glass/epoxy composite", *Compos. Sci. Technol.* **60**:4 (2000), 535–541.

- [Hashin 1962] Z. Hashin, “The elastic moduli of heterogeneous materials”, *J. Appl. Mech. (Trans. ASME)* **29** (1962), 143–150.
- [Hashin and Shtrikman 1963] Z. Hashin and S. Shtrikman, “A variational approach to the theory of the elastic behavior of multiphase materials”, *J. Mech. Phys. Solids* **11**:2 (1963), 127–140.
- [Hinrichsen et al. 1986] L. Hinrichsen, J. Feder, and T. Ossang, “Geometry of random sequential adsorption”, *J. Stat. Phys.* **44**:5-6 (1986), 793–827.
- [Kanit et al. 2003] T. Kanit, S. Forest, I. Galliet, V. Mounoury, and D. Jeulin, “Determination of the size of the representative volume element for random composites: statistical and numerical approach”, *Int. J. Solids Struct.* **40**:13-14 (2003), 3647–3679.
- [Li and Wang 2005] L. X. Li and T. J. Wang, “A unified approach to predict overall properties of composite materials”, *Mater. Charact.* **54**:1 (2005), 49–62.
- [Lusti et al. 2002] H. R. Lusti, P. J. Hine, and A. A. Gusev, “Direct numerical predictions for the elastic and thermoelastic properties of short fibre composites”, *Compos. Sci. Technol.* **62**:15 (2002), 1927–1934.
- [Milton 1982] G. W. Milton, “Bounds on the elastic and transport properties of two-component composites”, *J. Mech. Phys. Solids* **30**:3 (1982), 177–191.
- [Milton and Phan-Thien 1982] G. W. Milton and N. Phan-Thien, “New bounds on effective elastic moduli of two-component materials”, **380**:1779 (1982), 305–331.
- [Mori and Tanaka 1973] T. Mori and K. Tanaka, “Average stress in matrix and average elastic energy of materials with misfitting inclusions”, *Acta Metall.* **21**:5 (1973), 571–574.
- [Segurado and Llorca 2002] J. Segurado and J. Llorca, “A numerical approximation to the elastic properties of sphere-reinforced composites”, *J. Mech. Phys. Solids* **50**:10 (2002), 2107–2121.

Received 30 Jul 2006. Revised 2 Apr 2007. Accepted 20 Apr 2007.

HARALD BERGER: [harald.berger@mb.uni-magdeburg.de](mailto:harald.berger@mb.uni-magdeburg.de)

*Institute of Mechanics, University of Magdebur, Universitaetsplatz 2, D-39106, Magdeburg, Germany*

SREEDHAR KARI: [sreedhar.kari@nottingham.ac.uk](mailto:sreedhar.kari@nottingham.ac.uk)

*Institute of Mechanics, University of Magdebur, Universitaetsplatz 2, D-39106, Magdeburg, Germany*

ULRICH GABBERT: [ulrich.gabbert@mb.uni-magdeburg.de](mailto:ulrich.gabbert@mb.uni-magdeburg.de)

*Institute of Mechanics, University of Magdebur, Universitaetsplatz 2, D-39106, Magdeburg, Germany*

REINALDO RODRÍGUEZ RAMOS: [reinaldo@matcom.uh.cu](mailto:reinaldo@matcom.uh.cu)

*Facultad de Matemática y Computación, Universidad de La Habana, San Lázaro y L, CP 10400, Vedado, Havana 4, Cuba*

JULIAN BRAVO CASTILLERO: [jbravo@matcom.uh.cu](mailto:jbravo@matcom.uh.cu)

*Facultad de Matemática y Computación, Universidad de La Habana, San Lázaro y L, CP 10400, Vedado, Havana 4, Cuba*

RAÚL GUINOVRT DÍAZ: [guino@matcom.uh.cu](mailto:guino@matcom.uh.cu)

*Facultad de Matemática y Computación, Universidad de La Habana, San Lázaro y L, CP 10400, Vedado, Havana 4, Cuba*

# NATURAL CONVECTION FLUID FLOW AND HEAT TRANSFER IN POROUS MEDIA

ELSA BÁEZ AND ALFREDO NICOLÁS

Natural convection and heat transfer of fluid flow are studied numerically inside a rectangular cavity with inclination filled with a porous medium. The mass and momentum equations are given by the Darcy equations coupled with the thermal energy equation through the unsteady Boussinesq approximation. The two-dimensional restriction in terms of the stream function and vorticity variables is considered. The study is analyzed in terms of several values of the parameters that determine the evolution of the flow: the Rayleigh Number, the aspect ratio of the cavity and the angle of inclination.

## 1. Introduction

The mass and momentum equations in natural convection fluid flow in a porous medium are given by the Darcy equations coupled with the thermal energy equation through the unsteady Boussinesq approximation to deal with an incompressible structure. In this work the dimensionless problem is formulated in terms of the stream function and vorticity variables; then, the computation of the pressure is avoided and the incompressibility condition is satisfied automatically. Regarding the numerical method, once a convenient second order time discretization is performed, a nonlinear elliptic system is obtained which is solved through a fixed point iterative process. The iterative process leads to the solution of uncoupled, well-conditioned, symmetric linear elliptic problems for which very efficient solvers exist regardless of the space discretization.

The study of natural convection and heat transfer of fluid flow in a porous medium has important technological applications: storage and preservation of grains and cereals; solar energy collectors; filter systems; transport of radioactive wastes through the soil; and postaccident heat removal in nuclear reactors. Our numerical study is carried out on tilted rectangular cavities. The study is realized through the parameters that influence directly the behavior and evolution of the flow: the Rayleigh Number  $Ra$ , the aspect ratio of the cavity  $A$ , and the angle of inclination  $\phi$ .

We mention below two categories of research in connection with natural convection problems that arise when opposing walls of a cavity are subjected to a temperature gradient and where the other set of walls is insulated—the subject of the present work.

- (i) The steady problem: [Vasseur et al. \[1987\]](#) studied analytically and numerically the flow in a tilted rectangular cavity and observed that the maximum heat transfer, for a given  $Ra$ , is obtained when the cavity is heated from below, with  $\phi$  in the range  $90^\circ < \phi < 180^\circ$ . They found that this maximum takes place for values of  $\phi$  approaching  $90^\circ$  whenever  $Ra$  increases. [Moya et al. \[1987\]](#) studied the problem in tilted horizontal rectangular cavities for Rayleigh number  $Ra = 100$  and found

---

*Keywords:* natural convection, heat transfer, tilted cavity, Boussinesq approximation.

multiple cellular flow. [Sen et al. \[1987\]](#) studied the multiplicity of solutions, considering vertical and horizontal inclined cavities, and showed analytical and numerical results for Rayleigh numbers  $Ra \leq 500$  with small angles.

- (ii) The unsteady problem : [Baytas \[2000\]](#) showed results in a tilted square cavity for Rayleigh numbers  $10^2$ ,  $10^3$ , and  $10^4$ . [Saeid and Pop \[2004\]](#) studied the transient evolution for Rayleigh numbers with values of  $10^2 - 10^4$  in a square cavity, and reported the final time when the steady state is reached.

Results for  $Ra = 10^2$  and  $10^3$  are reported to validate the numerical method; these flows, obtained from the unsteady problem, agree with the ones obtained by other authors solving either the steady problem or the unsteady one but using different methods. Results for  $Ra \geq 10^2$  with  $A = 4$  and  $8$  are also reported, which to the best of our knowledge, have not been presented before. To assure that these new flows are correct, a time step and mesh independence studies have been made. All the results are complemented with their local and global Nusselt numbers on the hot wall, and the extreme values of the stream function. Actually, results with the numerical method described here are reported in [\[Báez and Nicolás 2006\]](#), where the numerical method is extended to include natural convection flows in homogeneous fluids (the evolution of some oscillatory, time-dependent, flows is also described therein); however, the results for porous media reported here are different from those shown in that work.

## 2. Mathematical models

### *Nomenclature*

$W$	width of the cavity
$H$	height of the cavity
$A$	aspect ratio of the cavity ( $=H/W$ )
$\rho$	density
$\rho_0$	reference density
$T$	temperature
$T_r$	reference temperature
$\beta$	thermal expansion coefficient
$k$	permeability of porous medium
$\eta$	thermal diffusivity
$\mu$	dynamic viscosity
$\nu$	kinematic viscosity ( $=\frac{\mu}{\rho_0}$ )
$t$	dimensionless time
$\mathbf{u}$	dimensionless velocity vector ( $\mathbf{u} = (u_1, u_2)$ )
$p$	dimensionless pressure
$\theta$	dimensionless temperature
$\psi$	stream function
$Ra$	Rayleigh number
$\phi$	angle of inclination of the cavity
$g$	gravity constant
$\mathbf{e}$	unitary vector in the gravity direction

$Nu$  local Nusselt number  
 $\overline{Nu}$  global Nusselt number

Natural convection flow of a thermal viscous fluid assumed to be Newtonian is considered under the well known Boussinesq approximation in the presence of a gravitational field. Let  $\Omega \subset R^N$  ( $N = 2, 3$ ) be the region of the flow of an unsteady, viscous, and thermal fluid, and  $\Gamma$  the boundary of the region. This kind of flow may be governed by the following dimensionless system of equations with incompressible structure:

$$\begin{cases} t > 0 : \\ \mathbf{u} + \nabla p = Ra \theta \mathbf{e}, & \text{in } \Omega, \text{ (a)} \\ \nabla \cdot \mathbf{u} = 0, & \text{in } \Omega, \text{ (b)} \\ \theta_t - \nabla^2 \theta + \mathbf{u} \cdot \nabla \theta = 0, & \text{in } \Omega. \text{ (c)} \end{cases} \tag{1}$$

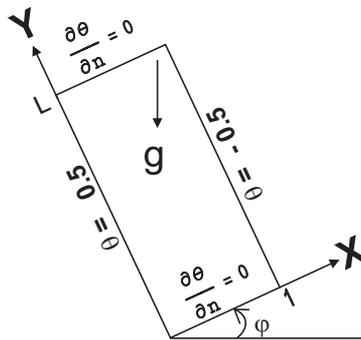
Equations (1)a–b are the Darcy equations in primitive variables  $\mathbf{u}$  and  $p$  coupled with the temperature Equation (1)c; Equation (1)b is known as the *incompressibility condition*. The Rayleigh number is given by

$$Ra = \frac{k\beta Lg(T_h - T_c)}{\eta\nu},$$

with  $T_h$  as the constant temperature on the hot wall,  $T_c$  that of the cold wall, and  $L$  as a reference length. The system must be supplemented with initial conditions  $\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x})$  and  $\theta(\mathbf{x}, 0) = \theta_0(\mathbf{x})$  in  $\Omega$ , and boundary conditions: for instance  $\mathbf{u} = \mathbf{f}$  and  $B\theta = 0$  on  $\Gamma$ ,  $t \geq 0$ , where  $B$  is a temperature boundary operator that can involve Dirichlet, Neumann or mixed boundary conditions. Figure 1 shows the geometry of the model considered.

Restricting the system in Equation (1) to a two-dimensional region  $\Omega$ , applying the rotational in both sides of the momentum equation, and considering that  $\nabla \cdot \mathbf{u} = 0$ , imply the existence of a function  $\psi$ , called the stream function, such that

$$u_1 = \frac{\partial \psi}{\partial y}, \quad u_2 = -\frac{\partial \psi}{\partial x}, \tag{2}$$



**Figure 1.** Geometry of the model.

the following scalar system is then obtained, where the unitary vector  $\mathbf{e}$  has been replaced by the contribution of the angle of inclination  $\phi$  of the region  $\Omega$  through  $\mathbf{e} = (\sin \phi, \cos \phi)$ :

$$\begin{cases} t > 0 : \\ -\nabla^2 \psi = \text{Ra} \left( \frac{\partial \theta}{\partial x} \cos \phi - \frac{\partial \theta}{\partial y} \sin \phi \right), & \text{in } \Omega, \\ \theta_t - \nabla^2 \theta + \mathbf{u} \cdot \nabla \theta = 0, & \text{in } \Omega. \end{cases} \tag{3}$$

This system represents the Boussinesq approximation in stream-function vorticity variables of system Equation (1) in primitive variables. As pointed out in Section 1, the pressure has been eliminated since the curl of the gradient is zero and the incompressibility condition is satisfied automatically by Equation (2). This work is concerned with natural convection in rectangular cavities, and so the equations are set in  $\Omega = (0, W) \times (0, H)$ . For viscous fluids  $\mathbf{u} = \mathbf{0}$  on solid and fixed walls, all the walls of the cavities are solid and fixed in natural convection, and so, by Equation (2),  $\psi$  is constant and this constant can be chosen to be 0.

The local Nusselt number  $\text{Nu}$  measures the heat transfer at each point on the hot wall where the temperature is specified, and the global Nusselt number  $\overline{\text{Nu}}$  measures the total rate of heat transfer on the same wall. These nondimensional parameters are defined by

$$\begin{aligned} \text{local Nusselt number: } \quad \text{Nu}(y) &= A \left| \frac{\partial \theta}{\partial x} \right|_{x=0}; \\ \text{global Nusselt number: } \quad \overline{\text{Nu}} &= \int_0^H \text{Nu}(y) \, dy. \end{aligned}$$

### 3. Numerical scheme

The time derivative  $\theta_t$  in the system of Equation (3) is approximated by

$$f_t(\mathbf{x}, (n + 1)\Delta t) \approx \frac{1.5 f^{n+1} - 2 f^n + 0.5 f^{n-1}}{\Delta t}, \quad n \geq 1, \quad \mathbf{x} \in \Omega, \tag{4}$$

where  $\Delta t > 0$  is the time discretization step,  $f^r$  is an approximation of  $f(x, r \Delta t)$ , and where it is known that Equation (4) is a second order approximation for sufficiently smooth function  $f$ .

Then, once Equation (4) is applied to  $\theta_t$  the following nonlinear elliptic system is obtained, incorporating the boundary condition for  $\psi$  and  $\theta$  as discussed before:

$$\begin{cases} -\nabla^2 \psi^{n+1} = \text{Ra} \left( \frac{\partial \theta^{n+1}}{\partial x} \cos \phi - \frac{\partial \theta^{n+1}}{\partial y} \sin \phi \right) & \text{in } \Omega, & \psi^{n+1} = 0 & \text{on } \Gamma, \\ \alpha \theta^{n+1} - \nabla^2 \theta^{n+1} + \mathbf{u}^{n+1} \cdot \nabla \theta^{n+1} = f_\theta & \text{in } \Omega, & B \theta^{n+1} = 0 & \text{on } \Gamma, \end{cases} \tag{5}$$

where  $\alpha = \frac{1.5}{\Delta t}$ ,  $f_\theta = \frac{2\theta^n - 0.5\theta^{n-1}}{\Delta t}$ ,  $\mathbf{u}$  in terms of  $\psi$  is given by Equation (2), and  $B$  is still the temperature boundary operator.

Renaming  $(\psi^{n+1}, \theta^{n+1})$  by  $(\psi, \theta)$  to simplify the notation, we must solve at each time level, for Equation (5), a nonlinear elliptic system of the form

$$\begin{cases} -\nabla^2 \psi = \text{Ra} \left( \frac{\partial \theta}{\partial x} \cos \phi - \frac{\partial \theta}{\partial y} \sin \phi \right) & \text{in } \Omega, & \psi = 0 & \text{on } \Gamma \\ \alpha \theta - \nabla^2 \theta + \mathbf{u} \cdot \nabla \theta = f_\theta & \text{in } \Omega, & B \theta = 0 & \text{on } \Gamma. \end{cases} \tag{6}$$

To obtain  $(\psi^1, \theta^1)$  in Equation (5), a first order approximation is applied to the temporal derivative with a smaller time step to maintain second order precision; an elliptic system like the one in Equation (6) is also obtained.

Given

$$\Theta(\psi, \theta) \equiv (\alpha I - \nabla^2\theta) + \mathbf{u} \cdot \nabla\theta - f_\theta \text{ in } \Omega,$$

then system Equation (6) is equivalent to

$$\begin{cases} -\nabla^2\psi = \text{Ra}(\frac{\partial\theta}{\partial x} \cos\phi - \frac{\partial\theta}{\partial y} \sin\phi) & \text{in } \Omega, & \psi|_\Gamma = 0, \\ \Theta(\psi, \theta) = 0 & \text{in } \Omega, & B\theta|_\Gamma = 0. \end{cases} \tag{7}$$

System Equation (7) is solved with the following fixed point iterative process:

$$\left\{ \begin{array}{l} \text{Knowing } \theta^0 = \theta^n \text{ solve until convergence on } \theta \\ -\nabla^2\psi^{m+1} = \text{Ra}(\frac{\partial\theta^m}{\partial x} \cos\phi - \frac{\partial\theta^m}{\partial y} \sin\phi) \quad \text{in } \Omega, \quad \psi^{m+1} = 0 \quad \text{on } \Gamma, \quad \text{(a)} \\ \theta^{m+1} = \theta^m - \lambda(\alpha I - \nabla)^{-1}\Theta(\psi^{m+1}, \theta^m) \quad \text{in } \Omega, \quad B\theta^{m+1} = 0 \quad \text{on } \Gamma, \lambda > 0, \quad \text{(b)} \\ \text{and take } (\psi^{n+1}, \theta^{n+1}) = (\psi^{m+1}, \theta^{m+1}). \end{array} \right. \tag{8}$$

The partial differential equation problem for  $\theta^{m+1}$  in Equation (8)b is equivalent to

$$(\alpha I - \nabla^2)\theta^{m+1} = (\alpha I - \nabla^2)\theta^m - \lambda\Theta(\psi^{m+1}, \theta^m) \quad \text{in } \Omega, \quad B\theta^{m+1}|_\Gamma = 0.$$

Therefore, at each iteration of each time level, uncoupled, well-conditioned, symmetric elliptic linear problems associated with the operators  $-\nabla^2$  and  $\alpha I - \nabla^2$  must be solved.

For linear elliptic problems, very efficient solvers exist regardless of the space discretization. The results in this work are obtained with the second-order approximation of the Fishpack solver [Adams et al. 1980], where the algebraic linear systems are solved with an efficient cyclic reduction iterative method [Sweet 1977]. As mentioned before, the first time derivative  $\theta_t$  is approximated by Equation (4), which is a second-order approximation, whereas the first space derivatives of  $\psi$  in Equation (2) to obtain  $\mathbf{u}$  in Equation (6), the normal derivative of the boundary condition for  $\theta$  (described later), and the first space derivative for the local Nusselt number are approximated by the centered second-order finite difference approximation in interior points and by Equation (4) in boundary points. To approximate the integral in the global Nusselt number the second-order trapezoid rule (in the entire interval) is used. All these kinds of approximations imply that the whole discrete problem relies on second-order approximations.

#### 4. Numerical results

The initial condition for the temperature is given by  $\theta(\mathbf{x}, 0) = 0$ . The discretization parameters, time step  $\Delta t$  and the size of the mesh  $h_x \times h_y$ , will be specified in each case under study. In the iterative process, the parameter  $\lambda$  is chosen as  $\lambda = 0.7$  and the stopping absolute criterion as  $10^{-5}$ .

The results are reported through the streamlines of the stream function and the isotherms of the temperature; most of the isocontours are specified for each case, otherwise they are the default ones. All the results are also complemented with their local Nu and global  $\overline{\text{Nu}}$  Nusselt numbers, in order to see the local and global heat transfer, as well as the extreme values of the stream function  $\psi$ .

The results shown correspond to steady state flows from the unsteady problem. They are the converged asymptotic steady state as time  $t$  approaches  $+\infty$  (large time, in practice). To reach convergence to an asymptotic steady state a stopping criterion must be given for the final time  $T_{ss}$  when it occurs. Since  $T_{ss}$  is the time when the solution does not change any more with respect to time at any spatial point occupied by the fluid [Nicolás and Bermúdez 2005],  $T_{ss}$  is determined with the point-wise discrete  $L_\infty$  absolute criterion in the closure  $\bar{\Omega}$  of the cavity

$$\theta : \|\theta_{hx,hy}^{n+1} - \theta_{hx,hy}^n\|_\infty,$$

with tolerance  $10^{-5}$ . For  $Ra = 10^2$  and  $Ra = 10^3$  results are shown to validate the numerical method mainly in the square cavity with results other authors have obtained solving the steady problem or the unsteady one but using a different method. New results, to the best of our knowledge reported here for the first time, are presented for  $Ra \geq 10^2$  with aspect ratios  $A = 4, 8$ . To support the validity of these results, a time step and mesh independence studies have been made with the point-wise discrete  $L_\infty$  relative error in  $\bar{\Omega}$

$$\begin{cases} \Delta t \text{ fixed :} & \frac{\|f_{hx1,hy1;\Delta t} - f_{hx2,hy2;\Delta t}\|_\infty}{\|f_{hx1,hy1;\Delta t}\|_\infty}, \\ \{h_x, h_y\} \text{ fixed :} & \frac{\|f_{hx,hy;\Delta t1} - f_{hx,hy;\Delta t2}\|_\infty}{\|f_{hx,hy;\Delta t1}\|_\infty}. \end{cases}$$

The specific temperature boundary condition to be considered, described so far in the boundary operator  $B$ , is given by

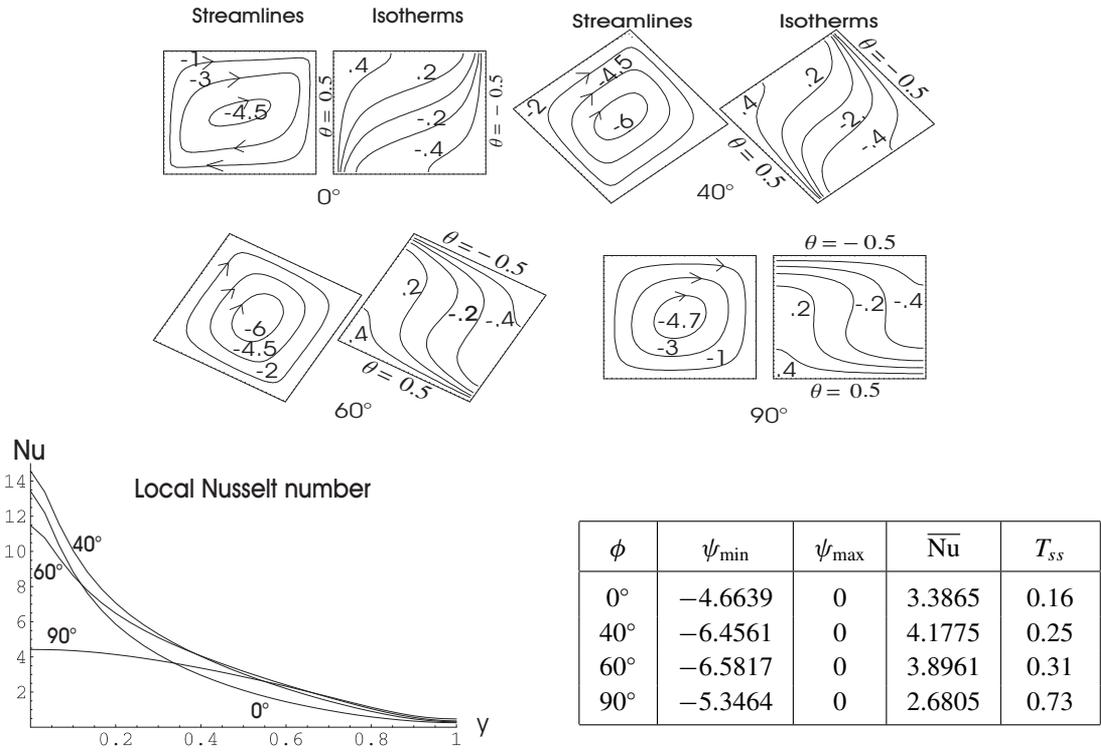
$$\frac{\partial \theta}{\partial y} = 0 \quad \text{on } \Gamma|_{y=0,W}, \quad \theta = 0.5 \quad \text{on } \Gamma|_{x=0}, \quad \theta = -0.5 \quad \text{on } \Gamma|_{x=1};$$

that is, horizontal walls are adiabatic and the hot and cold wall are the left and the right wall, respectively.

Figure 2 shows the streamlines and isotherms for  $Ra = 10^2$  in the unit square, that is  $A = 1$ , and  $0^\circ \leq \phi \leq 90^\circ$ . This range of angles means that heating from the lateral (left) wall to heating on the bottom wall is considered.

It is observed from the streamlines of Figure 2 that one main cell only is obtained for the range of angles considered and therefore, from the local Nusselt number graphic, one maximum is also obtained for heat transfer. In this case, from  $\psi_{\min}$ , it follows that the fluid motion is slower when the heating comes either from the lateral or bottom wall, that is when  $\phi = 0^\circ$  or  $\phi = 90^\circ$ , than for the other angles. It is also observed, from  $\bar{Nu}$  in Figure 1, that the heat transfer is smaller for these angles than for the others.

Figure 3 shows the results for  $Ra = 10^3$  and some angles  $0^\circ \leq \phi \leq 360^\circ$  in the unit square cavity also. For  $\phi = 0^\circ$ , it is observed that the fluid is heated along the left wall causing the less dense fluid to rise toward the top wall; this fluid is then cooled on the right wall, becomes denser, and then falls to the bottom of the cavity, originating a rotating clockwise cell in the streamlines. A similar situation occurs for  $40^\circ$  and  $330^\circ$ ; however, the opposite effect occurs for  $130^\circ$ . In this case, the main cell rotates counterclockwise since the hot wall is below the cold one and the isotherms show that the hot fluid is localized toward the top of the right part of the tilted cavity while the cold fluid resides toward the bottom of the left part. Secondary cells appear for some angles, for instance  $\phi = 330^\circ$ . The graphic of the local Nusselt number shows that the maximum value for  $0^\circ, 40^\circ, \text{ and } 330^\circ$  is reached on the bottom wall



**Figure 2.** Results for  $Ra = 10^2$ ,  $A = 1$ ,  $\Delta t = 10^{-2}$  and  $h_x \times h_y = \frac{1}{30} \times \frac{1}{30}$ .

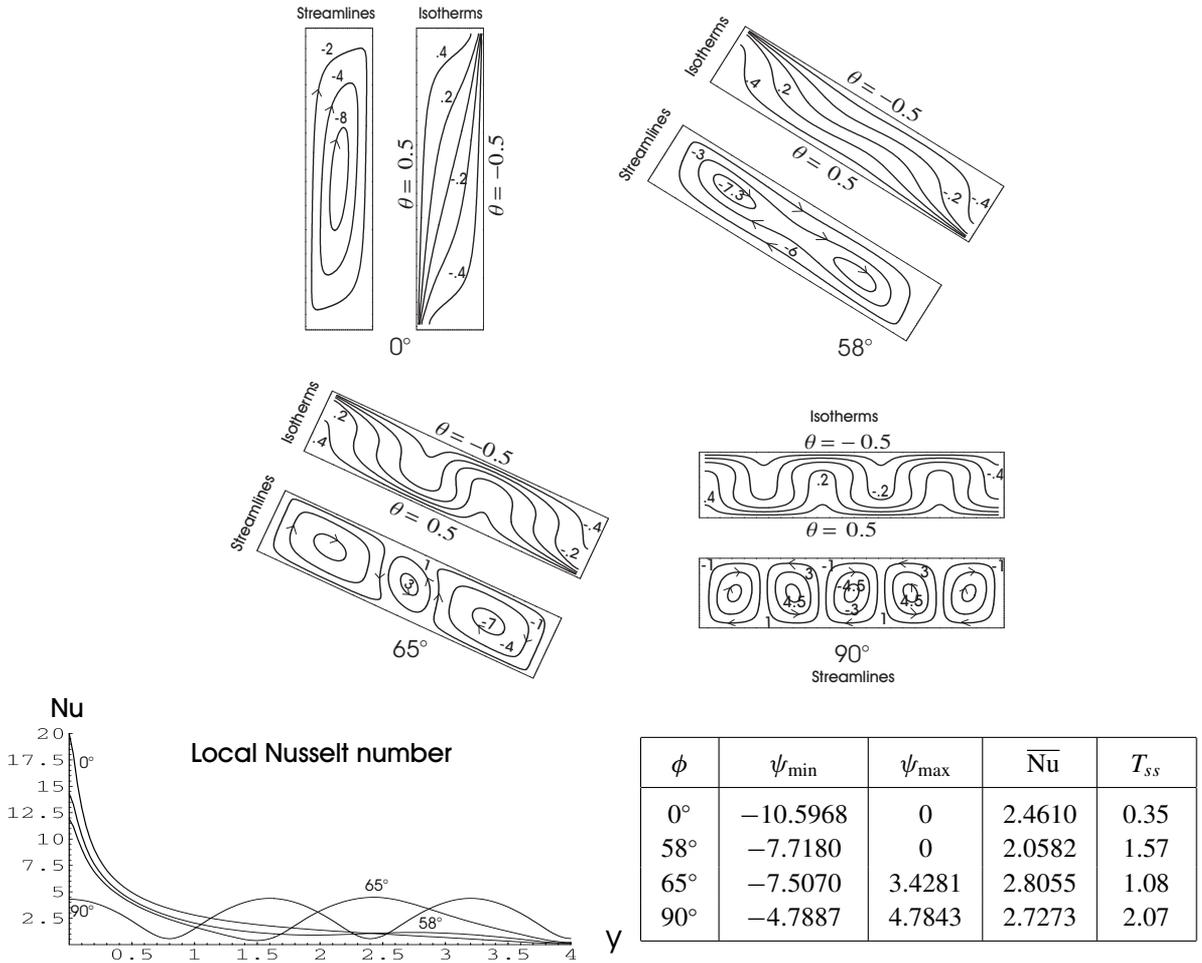
whereas for  $130^\circ$  the opposite occurs, that is, the maximum is reached on the top wall because of the buoyancy effect.

The corresponding results for the minimum and maximum values of the stream function,  $\psi_{\min}$  and  $\psi_{\max}$ , the global Nusselt number  $\overline{Nu}$ , and the final time  $T_{ss}$  when the flow reaches the steady state are displayed in Figure 3. The extreme values of the stream function indicate an increase in the fluid motion for the angles  $\phi = 40^\circ$  and  $\phi = 130^\circ$ , implying an increase of heat transfer, because of  $\overline{Nu}$ , although this is higher for the first angle than for the second, whereas for  $330^\circ$  less fluid motion is obtained implying a diminution of the heat transfer.

The above results for  $Ra = 10^2$  and  $Ra = 10^3$  as well as others obtained for  $Ra = 10^2$  and  $Ra = 10^4$  with  $0^\circ \leq \phi \leq 360^\circ$  are in agreement with those reported in [Baytas 2000] from the unsteady problem also but using a different method. For these three values of  $Ra$  one main cell is obtained for almost all values of  $\phi$  and secondary cells appear for some angles when  $Ra = 10^3$  and  $10^4$ , as pointed out before for  $Ra = 10^3$ .

As already mentioned, the new results are shown for values of Rayleigh number  $Ra \geq 10^2$  with aspect ratios  $A \geq 1$ . For  $Ra = 10^2$  and the aspect ratio is augmented to  $A = 4$ , it can be seen in Figure 4 that there appears from one cell for  $\phi = 0^\circ$  to five cells when the cavity is heated on the bottom, that is for  $\phi = 90^\circ$ . The local Nusselt number graphic shows one maximum for  $\phi = 0^\circ$  and  $\phi = 58^\circ$ , two for  $\phi = 65^\circ$  and three for  $\phi = 90^\circ$ .





**Figure 4.** Results for  $Ra = 10^2$ ,  $A = 4$ ,  $\Delta t = 10^{-2}$  and  $h_x \times h_y = \frac{1}{30} \times \frac{4}{120}$ .

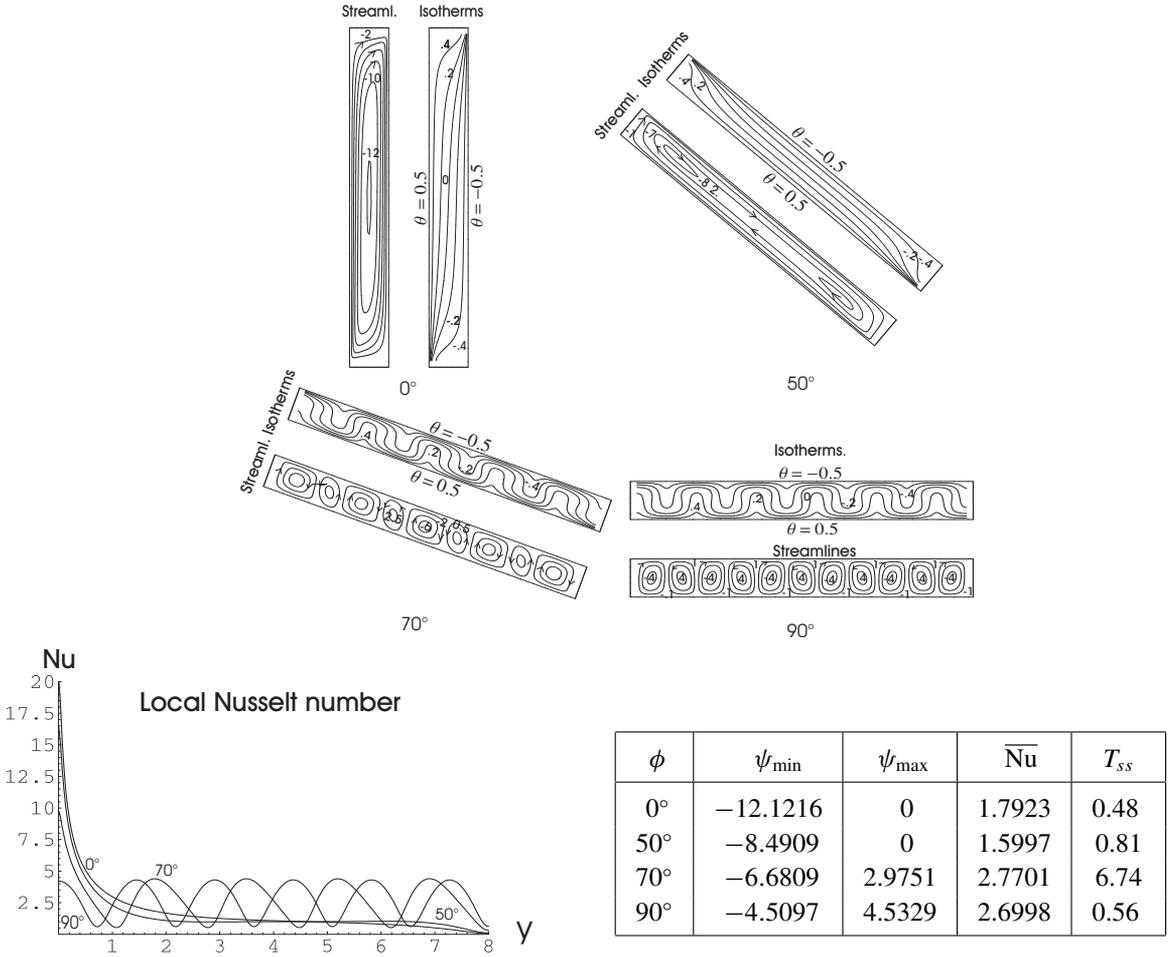
To demonstrate that the flows in Figures 4 and 5 are correct, a time step and mesh independence studies were performed in the vertical case,  $\phi = 0^\circ$ , with  $A = 8$  in Figure 5, for three mesh sizes and three times as follows:

- (1) time step fixed,  $\Delta t = 10^{-2}$  and  $(h_x, h_y) = (1/30, 8/240), (1/45, 8/360), (1/60, 8/480)$ ;
- (2) mesh size fixed  $(h_x, h_y) = (1/30, 1/240)$  and  $\Delta t = 10^{-2}, 5 \times 10^{-3}, 2.5 \times 10^{-3}$ .

The respective discrepancies are:

- (1) less than  $6 \times 10^{-2}\%$  (at most  $3.8 \times 10^{-1}\%$  for stream function and  $5.9 \times 10^{-1}\%$  for temperature);
- (2) at most  $4.96 \times 10^{-2}\%$  ( $2.44 \times 10^{-2}\%$  for stream function and  $4.96 \times 10^{-2}\%$  for temperature).

The corresponding minima of the stream function  $\psi$  in each case (the maximum value is always zero) are:



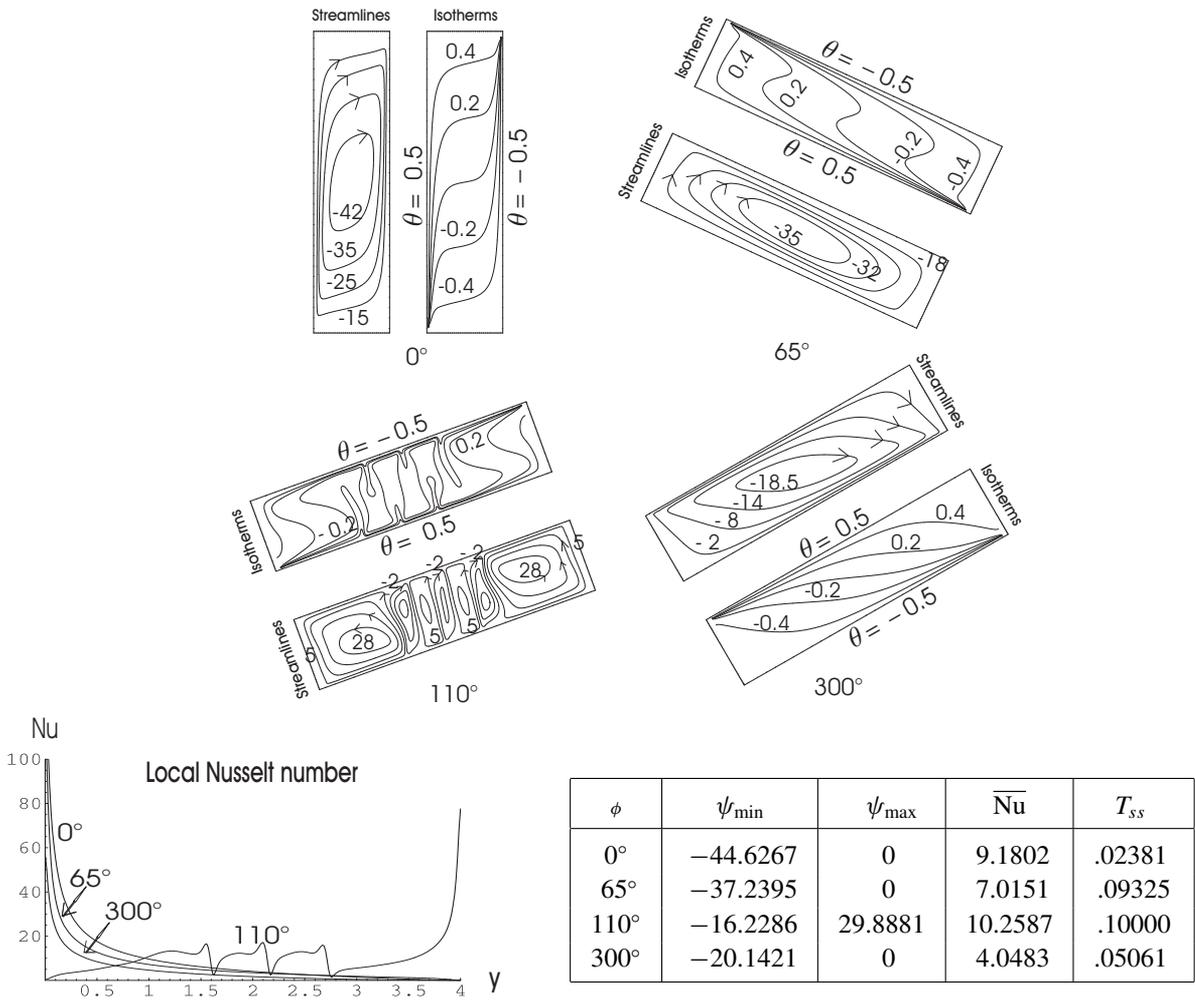
**Figure 5.** Results for  $Ra = 10^2$ ,  $A = 8$ ,  $\Delta t = 10^{-2}$  and  $h_x \times h_y = \frac{1}{30} \times \frac{8}{240}$ .

- (1)  $\min = -12.1216, -12.1206, -12.1280$ , respectively;
- (2)  $\min = -12.1216, -12.1223, -12.1238$ , respectively.

Therefore, the results shown in [Figure 5](#) are chosen as the correct ones.

For the corresponding easier case with  $A = 4$  in [Figure 4](#) something similar occurs.

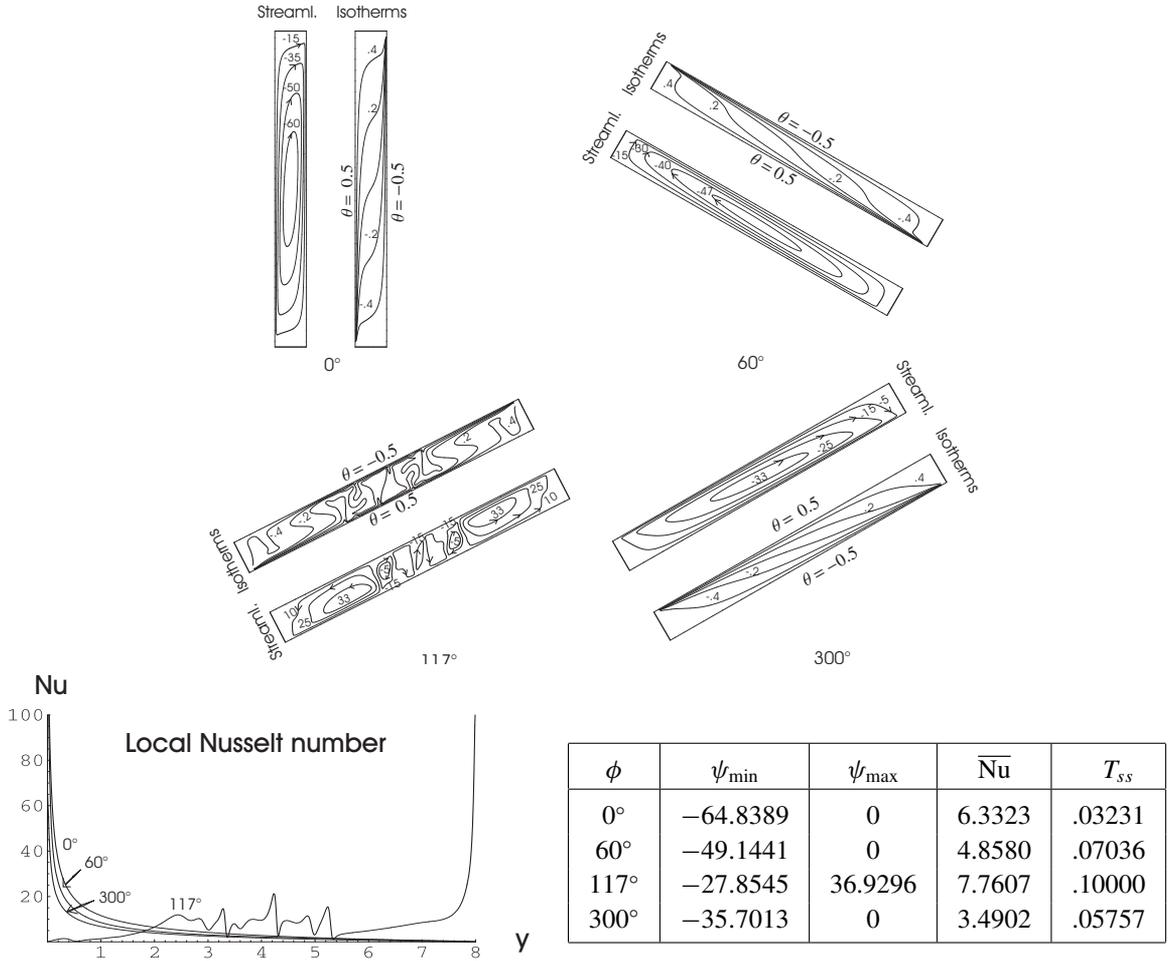
In [Figure 6](#), results for  $Ra = 10^3$  and some angles  $0^\circ \leq \phi \leq 360^\circ$  with  $A = 4$  are presented. A single large cell appears for angles  $\phi = 0^\circ$  and  $\phi = 65^\circ$  rotating clockwise. Experiments show that multiple cells appear for angles  $66^\circ \leq \phi \leq 114^\circ$ , the result for  $\phi = 110^\circ$  is an example of this situation: seven main cells appear, five small cells occupying the middle of the cavity enclosed by two big ones in the extremes of the cavity and rotating in opposite direction each other. The graphic of the local Nusselt number shows that the maximum heat transfer for this angle is localized on the top of the cavity due to the reverse effect of the buoyancy force, in comparison to the other angles; the other maximums are a consequence of the fact that multiple cells appear.



**Figure 6.** Results for  $Ra = 10^3$ ,  $A = 4$ ,  $\Delta t = 10^{-5}$  and  $h_x \times h_y = \frac{1}{70} \times \frac{4}{280}$ .

From Figure 6 it is also observed that a stronger fluid motion occurs for  $0^\circ$  than for  $110^\circ$ ; however, the corresponding value of the global Nusselt number for  $0^\circ$ , when there is a single cell, indicates a smaller heat transfer than for  $110^\circ$ , when multiple cells appear.

Figure 7 pictures numerical results for the same Rayleigh number with  $A = 8$  and several angles  $0^\circ \leq \phi \leq 360^\circ$ . Multiple cells may appear for some angles, as shown for  $\phi = 117^\circ$ , which indicate a more complex fluid motion. On the other hand, Figure 5 shows that with some angles the fluid motion is stronger than others, but the corresponding value of the heat transfer is smaller. Complemented with other experiments, not shown here, it may be concluded that this situation is characteristic of aspect ratios  $A > 1$ .



**Figure 7.** Results for  $Ra = 10^3$ ,  $A = 8$ ,  $\Delta t = 10^{-5}$  and  $h_x \times h_y = \frac{1}{70} \times \frac{8}{560}$ .

To show that the flows in Figures 6 and 7 are correct, a time step and mesh independence studies were performed in the vertical case,  $\phi = 0^\circ$ , with  $A = 8$  in Figure 7, for three mesh sizes and three times as follows:

- (1) time step fixed,  $\Delta t = 10^{-5}$  and  $(h_x, h_y) = (1/70, 8/560), (1/105, 8/840), (1/140, 8/1120)$ ;
- (2) mesh size fixed  $(h_x, h_y) = (1/70, 8/560)$  and  $\Delta t = 10^{-5}, 5 \times 10^{-6}, 2.5 \times 10^{-6}$ .

The respective discrepancies are:

- (1) less than 2% (at most 1.67% for stream function and 1.59% for temperature);
- (2) at most 5.16% (5.16% for stream function and 2.15% for temperature).

The corresponding minima of the stream function  $\psi$  in each case (the maximum value is always zero) are:

- (1)  $\min = -64.8389, -64.8197, -64.8379$ , respectively;

(2)  $\min = -64.8389, -64.8415, -64.8351$ , respectively.

Therefore, the results shown in [Figure 7](#) are chosen as the correct ones.

For the corresponding easier case with  $A = 4$  in [Figure 6](#) something similar occurs.

## 5. Conclusions

From the numerical experiments we observe that when  $Ra$  increases, the time step and the spatial mesh size must necessarily be significantly diminished—which can become a problem computationally speaking, at least in the current form of the numerical method. The results obtained for several values of the Rayleigh number, the aspect ratio, and the angle of inclination of the cavity indicate that the flow is affected whenever the value of each of these parameters changes. The fluid motion is strong not only when  $Ra$  increases but also when  $A$  increases and  $Ra$  is fixed. The global Nusselt number shows also that the heat transfer increases as  $Ra$  increases, but when this value is fixed and the aspect ratio is larger, the heat transfer is smaller. For angles where multiple cells appear the global heat transfer is higher than for those with a single cell. Moreover, there exist two maxima of the global heat transfer, as a function of the angle  $\phi$ , with  $0^\circ \leq \phi \leq 90^\circ$ : one for angles when a single cell appears and one more for angles with multiple cells. About the time  $T_{ss}$  necessary to reach a steady state, for  $\phi$  fixed, it can be observed that  $T_{ss}$  is smaller whenever  $Ra$  is higher; however, this time increases whenever  $A$  increases. It may be pointed out that with some modifications of the numerical method, the case with variable porosity [[Marcondes et al. 2001](#)] and variable anisotropy [[Nguyen et al. 1994](#)] can be also explored as well as viscous effects near walls through the Brinkman extension [[Rees 1999](#)].

## Acknowledgements

The authors would like to thank the anonymous reviewers for their remarks which have improved the presentation of the paper as well as J. W. Eischen and G. Monsivais, coeditors of this issue, for their kind invitation to submit this paper.

## References

- [Adams et al. 1980] J. Adams, P. Swarztrauber, and R. Sweet, “FISHPACK: a package of Fortran subprograms for the solution of separable elliptic PDE’s”, technical report, The National Center for Atmospheric Research, Boulder, CO, 1980.
- [Báez and Nicolás 2006] E. Báez and A. Nicolás, “2D natural convection flows in tilted cavities: porous media and homogeneous fluids”, *Int. J. Heat Mass Tran.* **49**:25-26 (2006), 4773–4785.
- [Baytas 2000] A. C. Baytas, “Entropy generation for natural convection in an inclined porous cavity”, *Int. J. Heat Mass Tran.* **43**:12 (2000), 2089–2099.
- [Marcondes et al. 2001] F. Marcondes, J. M. De Medeiros, and J. M. Gurgel, “Numerical analysis of natural convection in cavities with variable porosity”, *Numer. Heat Tr. A Appl.* **40**:4 (2001), 403–420.
- [Moya et al. 1987] S. L. Moya, E. Ramos, and M. Sen, “Numerical study of natural convection in a tilted rectangular porous material”, *Int. J. Heat Mass Tran.* **30**:4 (1987), 741–756.
- [Nguyen et al. 1994] H. D. Nguyen, S. Paik, and R. W. Douglass, “Study of double-diffusive convection in layered anisotropic porous media”, *Numer. Heat Tr. B Fund.* **26**:4 (1994), 489–505.
- [Nicolás and Bermúdez 2005] A. Nicolás and B. Bermúdez, “2D thermal/isothermal incompressible viscous flows”, *Int. J. Numer. Meth. Fl.* **48**:4 (2005), 349–366.

- [Rees 1999] D. A. S. Rees, “Darcy-Brinkman free convection from a heated horizontal surface”, *Numer. Heat Tr. A Appl.* **35**:2 (1999), 191–204.
- [Saeid and Pop 2004] N. H. Saeid and I. Pop, “Transient free convection in a square cavity filled with a porous medium”, *Int. J. Heat Mass Tran.* **47**:8-9 (2004), 1917–1924.
- [Sen et al. 1987] M. Sen, P. Vasseur, and L. Robillard, “Multiple steady states for unicellular natural convection in an inclined porous layer”, *Int. J. Heat Mass Tran.* **30**:10 (1987), 2097–2113.
- [Sweet 1977] R. Sweet, “A cyclic reduction algorithm for solving block tridiagonal systems of arbitrary dimension”, *SIAM J. Numer. Anal.* **14**:4 (1977), 706–720.
- [Vasseur et al. 1987] P. Vasseur, M. G. Satish, and L. Robillard, “Natural convection in a thin, inclined, porous layer exposed to a constant heat flux”, *Int. J. Heat Mass Tran.* **30**:3 (1987), 537–549.

Received 21 Apr 2006. Accepted 20 Apr 2007.

ELSA BÁEZ: [ebj@xanum.uam.mx](mailto:ebj@xanum.uam.mx)

Depto. Matemáticas, Ed. AT-Diego Bricio, UAM-I, 09340 Mexico D.F., Mexico

ALFREDO NICOLÁS: [anc@xanum.uam.mx](mailto:anc@xanum.uam.mx)

Depto. Matemáticas, Ed. AT-Diego Bricio, UAM-I, 09340 Mexico D.F., Mexico

## STARK LADDER RESONANCES IN ACOUSTIC WAVEGUIDES

GUILLERMO MONSIVAIS AND RAUL ESQUIVEL-SIRVENT

We present a theoretical study on how to obtain a Wannier–Stark ladder in the transmission spectra of an acoustic wave traveling through a waveguide of variable cross section. Starting from Webster’s equation for the acoustic pressure, we derive the necessary conditions to obtain the Wannier–Stark ladder. Furthermore, we present a numerical calculation for the transmission spectra when a Wannier–Stark ladder is present. This ladder is characterized by a family of well defined peaks, equidistant in frequency.

### 1. Introduction

Band structures of the energy spectrum of the electrons are the basis of electronic devices. When electrons travel through a periodic structure such as a crystal, the constructive and destructive interference gives rise to bands in the energy spectra of the electrons [Brillouin 1953; Guo 2006]. Similarly, an electromagnetic wave traveling through a structure with a dielectric function that varies periodically will exhibit a band structure in its frequency spectrum. This gives rise to photonic crystals and its applications in light flow control as described by Joannopoulos et al. [1995]. In an elastic structure with a specific impedance that varies periodically, the transmission spectra as a function of frequency, elastic waves will also show a band structure [Esquivel-Sirvent and Cocolletzi 1994].

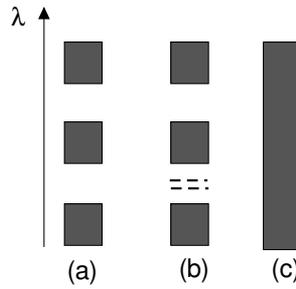
Physically, the band structure represents regions of allowed and forbidden propagation as shown in Figure 1.  $\lambda$  represents the frequency (or energy) of a wave (or electron) traveling through a system. Figure 1(a) corresponds to a periodic system in which  $\lambda$  can get only certain allowed values indicated by the dark zones. The regions where no values of  $\lambda$  are permissible are known as forbidden bands or gaps. When the periodicity is only slightly modified, for example at only one site of the structure, the band structure shows localized states. This is, there are certain allowed values of  $\lambda$  for which transmission is allowed in an otherwise forbidden region. This is indicated by the dotted lines in Figure 1(b). Finally, Figure 1(c) shows that  $\lambda$  can take any value when there is no periodicity.

An interesting case of broken periodicity give rise to Wannier–Stark ladders (WSL) that will be discussed in the next section. By imposing a particular condition on the configuration of the system, it is possible to destroy the band structure of an otherwise periodic system, and obtain sharply localized states that are equidistant in frequency or energy. This resonances were first predicted in quantum mechanics by Wannier [1962] in connection with the energy spectrum of an electron traveling through a crystal in a dc electric field. As in the case of periodic system, it has been shown that WSL exist in photonic crystals [Monsivais et al. 1990], elastic systems [Mateos and Monsivais 1994] and piezoelectric systems [Monsivais et al. 2003].

---

*Keywords:* Stark, acoustic, resonances.

Partial support from CONACyT Grant No. 44306 and DGAPA-UNAM Grant No. IN101605 is acknowledged.



**Figure 1.** Allowed values of  $\lambda$  in three different structures: (a) periodic; (b) periodic with a defect, where the localized states are shown by dotted lines; (c) disordered.

In this paper we study the formation of WSL in acoustic waveguides. This system is simpler than the elastic case since there is no mode conversion. In addition, for low frequencies, it is easier to realize experimentally.

## 2. Theory

Up to constants, we can write Schrodinger's equation for stationary solutions as

$$-\frac{d^2\psi}{dx^2} + U(x)\psi = \lambda\psi. \quad (1)$$

If the potential  $U(x)$  is periodic, that is,  $U(x) = U(x + np)$  where  $p$  is the period and  $n$  a positive integer, the system will show a band structure [Brillouin 1953]. We can break the periodicity by requiring that

$$U(x + np) = U(x) + npF, \quad (2)$$

where  $F$  is a constant. In the case of an electron of charge  $q$  traveling through a crystal,  $F = qE$ , where  $E$  is a constant electric field. Making the change of variable  $x = x' + npF$ , Equation (1) becomes

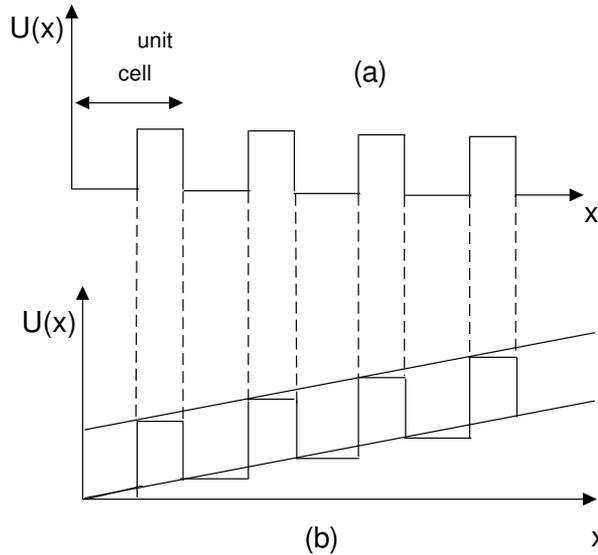
$$-\frac{d^2\varphi(x')}{dx'^2} + U(x' + ndF)\varphi(x') = \lambda\varphi(x').$$

Finally, using the property of  $U(x)$  given by Equation (2) we have

$$-\frac{d^2\varphi(x')}{dx'^2} + U(x')\varphi(x') = (\lambda - npF)\varphi(x'),$$

where  $\varphi(x') = \psi(x' + npF)$ . Comparing Equations (1) and (2), we see that if  $\lambda$  is an eigenvalue, then so is  $\lambda - npF$ . The difference between two neighboring eigenvalues is exactly  $pF$ ; it is this that gives rise to a Wannier–Stark ladder.

We should mention, however, that the mathematics described above presents many subtleties [Zak 1968] and a rigorous description is very difficult. For this reason the existence of WSL in quantum mechanics was a controversial matter for twenty years, from the prediction of Wannier [1962] in the 1960's until the experimental observation of WSL in superlattices [Mendez et al. 1988] and the results of numerical calculations in, both in the 1980's.



**Figure 2.** Schematic representation of the function  $U(x)$  given by Equation (2). (a) Periodic case obtained by setting  $F = 0$ . (b) If  $F \neq 0$  the periodicity of  $U(x)$  is broken and the solution of Equation (3) gives rise to a WSL.

The formulation discussed in this paper associated with the properties of the acoustic waveguide presents several differences compared with the quantum mechanical case. However, it suffers from similar mathematical subtleties, and, therefore, our discussion cannot be a rigorous demonstration of the existence of acoustic WSL. However, our numerical calculation will show that the naive formulation, in fact, predicts the correct result.

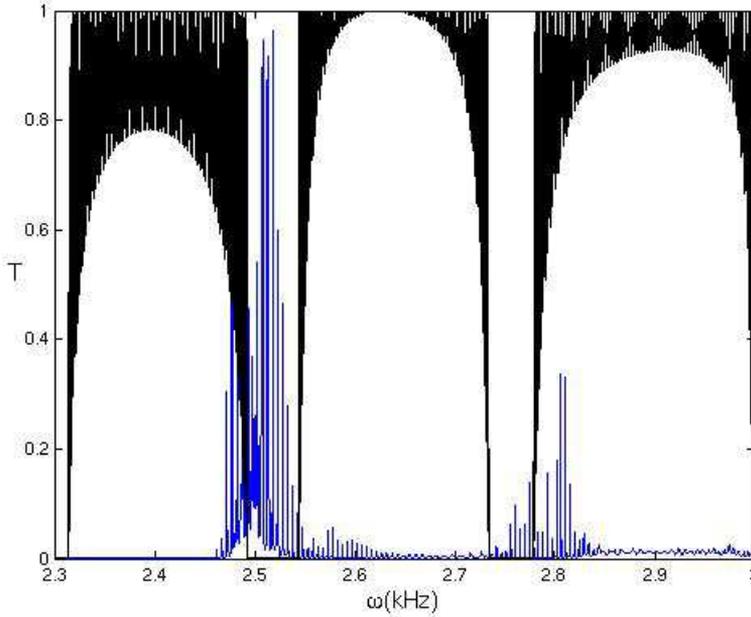
Now we will show how the above ideas can be adapted to an acoustic waveguide. Consider the equation for a waveguide with variable cross section  $S(x)$  and symmetry axis parallel to the  $x$ -axis. Webster’s equation for pressure  $p(x, t)$  is

$$\frac{\partial^2 p(x, t)}{\partial t^2} = c^2 \left( \frac{1}{S(x)} \frac{\partial}{\partial x} \left[ S(x) \frac{\partial p(x, t)}{\partial x} \right] \right),$$

where  $c$  is the speed of sound in the waveguide. This equation can be transformed to a Schrodinger-like one by introducing a function  $f(x, t)$  defined by  $p(x, t) = f(x, t)/S(x)^{1/2}$ . This particular choice comes from the fact that  $f(x, t)$  is proportional to the potential energy per unit area of the acoustic wave [Forbes et al. 2003]. Thus, Webster’s equation takes the form

$$\frac{\partial^2 f(x, t)}{\partial t^2} = c^2 \left( \frac{\partial^2 f(x, t)}{\partial x^2} - U(x) f(x, t) \right), \quad U(x) = \frac{1}{S(x)^{1/2}} \frac{d^2 S(x)^{1/2}}{dx^2}. \quad (3)$$

Equation (3) is separable in the variables  $x, t$ , and its solutions are of the form  $f(x, t) = X(x)T(t) = X(x) \exp(i\omega t)$ , where  $\omega$  is the wave frequency. Substituting this ansatz into Equation (3) yields the



**Figure 3.** Transmission spectra for an acoustic waveguide. Period of unit cell is 5 cm. Black curve shows result for a periodic system made of 60 unit cells. For  $F = 1$  the band structure disappears and WSL emerges, demonstrated by the blue curve.

following equation for  $X(x)$ :

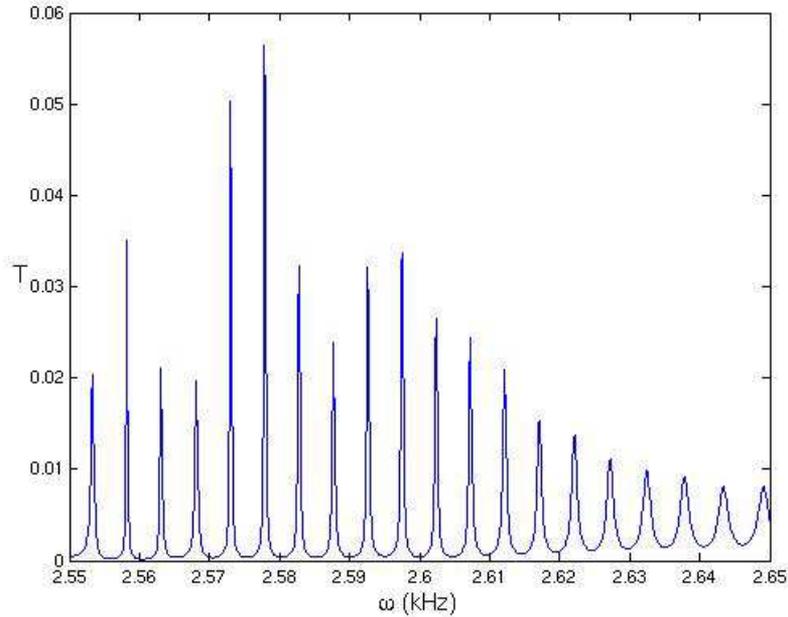
$$\frac{d^2 X(x)}{dx^2} + \left[ \frac{\omega^2}{c^2} - U(x) \right] X(x) = 0, \quad (4)$$

which has the form of Schrodinger's equation. By imposing condition (2) for the  $U(x)$ , the existence of Wannier–Stark ladders can be expected. We can now construct the function  $U(x)$  subject to the required condition. The actual variation of the cross section  $S(x)$  is obtained by solving Equation (3). However, this is not done in this paper, and will be reported elsewhere. First we notice that when  $F = 0$ , we have a periodic function constructed by repeating a unit cell as shown in Figure 2(a). When  $F \neq 0$  we choose a function profile as that shown in Figure 2(b). This profile will be used in the numerical calculations presented in the next section.

### 3. Numerical results

In this section we consider a finite system in order to have a model to analyze the existence of WSL numerically. Since for a finite system Equation (2) is only satisfied for a finite region of space, it can be expected that WSL formation will not be perfect.

The transmission spectra are calculated for a particular function  $U(x)$  of Figure 2(b). To perform the calculations, we consider that the changes in  $U(x)$  imply a change in the cross section of the waveguide. This in turn implies a change in the acoustic impedance that also depends on  $S(x)$ . At each change



**Figure 4.** Detail of WSL in [Figure 3](#): sharp, equidistant peaks are evident.

in the cross section, the boundary conditions are given by the continuity of  $X(x)$  and of its derivative  $dX(x)/dx$ . A transfer matrix approach is used to calculate the transmission spectra [[Esquivel-Sirvent and Cocolletzi 1994](#)].

In [Figure 3](#) we show two curves. The black line shows the transmission spectra for a periodic system made of 60 unit cells as described in [Figure 2\(a\)](#). The band structure is clearly seen. The number of oscillations in the regions of high transmission is equal to the number of unit cells. In the case of an infinite number of unit cells, the transmission will be equal to one in these regions. If we break the periodicity by setting  $F = 1$ , we obtain the blue curve. The periodicity is broken and the band structure is replaced by a series of sharp transmission peaks. This is the Wannier–Stark ladder. In [Figure 4](#) we show a detail of the transmission spectra. The spikes are equally spaced and are sharply peaked as expected. For these calculations we took a unit cell of period 5 cm. As mentioned before, the peaks in WSL are equidistant, but their separation is not  $pF$  as predicted by the theory since the system we analyzed was finite.

The parameters we choose to construct the function  $U(x)$  were adequate to obtain a WSL. However, we have observed that not any choice of parameters gives rise to WSL. It is not possible to know when a Stark ladder will emerge, except by trial and error.

#### 4. Conclusions

In this paper we have demonstrated the basic principles to obtain the Wannier–Stark ladder in the transmission spectra of an acoustic waveguide. Starting with Webster’s equation, we find an equivalent

Schrodinger-like equation that exhibits a Stark ladder for a suitable choice of a function  $U(x)$ . In our case, this function is related to the changes in cross section of the acoustic waveguide.

Our numerical studies show that the Stark ladder in acoustic waveguides in fact exist, even though this cannot be rigorously proven from the theoretical analysis, as was the case in quantum mechanics.

## References

- [Brillouin 1953] L. Brillouin, *Wave propagation in periodic structures: electric filters and crystal lattices*, Dover, New York, 1953.
- [Esquivel-Sirvent and Coccoletzi 1994] R. Esquivel-Sirvent and G. H. Coccoletzi, “Band structure for the propagation of elastic waves in superlattices”, *J. Acoust. Soc. Am.* **95**:1 (1994), 86–90.
- [Forbes et al. 2003] B. J. Forbes, E. R. Pike, and D. B. Sharp, “The acoustical Klein–Gordon equation: the wave-mechanical step and barrier potential functions”, *J. Acoust. Soc. Am.* **114**:3 (2003), 1291–1302.
- [Guo 2006] W. Guo, “Optical band gaps as a result of destructive superposition of scattered waves”, *Am. J. Phys.* **74**:7 (2006), 595–599.
- [Joannopoulos et al. 1995] J. D. Joannopoulos, R. D. Meade, and J. N. Winn, *Photonic crystals: molding the flow of light*, Princeton University Press, Princeton, NJ, 1995.
- [Mateos and Monsivais 1994] J. L. Mateos and G. Monsivais, “Stark-ladder resonances in elastic waves”, *Physica A* **207**:1-3 (1994), 445–451.
- [Mendez et al. 1988] E. E. Mendez, F. Agullo-Rueda, and J. M. Hong, “Stark localization in GaAs-GaAlAs superlattices under an electric field”, *Phys. Rev. Lett.* **60**:23 (1988), 2426–2429.
- [Monsivais et al. 1990] G. Monsivais, M. Castillo-Mussot, and F. Claro, “Stark-ladder resonances in the propagation of electromagnetic waves”, *Phys. Rev. Lett.* **64**:12 (1990), 1433–1436.
- [Monsivais et al. 2003] G. Monsivais, R. Rodríguez-Ramos, R. Esquivel-Sirvent, and L. Fernández-Alvarez, “Stark-ladder resonances in piezoelectric composites”, *Phys. Rev. B* **68**:17 (2003), 174109–174120.
- [Wannier 1962] G. H. Wannier, “Dynamics of band electrons in electric and magnetic fields”, *Rev. Mod. Phys.* **34**:4 (1962), 645–655.
- [Zak 1968] J. Zak, “Stark ladder in solids?”, *Phys. Rev. Lett.* **20**:26 (1968), 1477–1481.

Received 18 Jul 2006. Accepted 20 Apr 2007.

GUILLERMO MONSIVAIS: [monsi@fisica.unam.mx](mailto:monsi@fisica.unam.mx)

*Instituto de Física, Universidad Nacional Autónoma de México, Apdo. Postal 20-364, Distrito Federal, 01000, Mexico*

RAUL ESQUIVEL-SIRVENT: [raul@fisica.unam.mx](mailto:raul@fisica.unam.mx)

*Instituto de Física, Universidad Nacional Autónoma de México, Apdo. Postal 20-364, Distrito Federal, 01000, Mexico*

# TENSION BUCKLING IN MULTILAYER ELASTOMERIC ISOLATION BEARINGS

JAMES M. KELLY AND SHAKHZOD M. TAKHIROV

Seismic isolators are constructed from multiple layers of elastomer (usually natural rubber) reinforced with steel plates; they are, therefore, very stiff in the vertical direction, but soft in the horizontal direction. The buckling of these bearings under compression load is a well-understood phenomenon and has been widely studied. It is therefore unexpected that the buckling analysis for compression predicts that the isolator can buckle in tension at a load close to that for buckling in compression. The linear elastic model that leads to both compression and tension buckling is an extremely simple one, so it might be argued that the tensile buckling may be an artifact of the model itself rather than a property of the isolator. To test the simple theoretical model we have conducted a numerical simulation study using a finite element model of a multilayer elastomeric bearing. We find that the prediction of tensile buckling by the simple linear elastic theory is indeed accurate and not an artifact of the model.

## 1. Introduction

Seismic isolation using multilayer elastomeric isolators has been used in the United States and around the world for more than 20 years. The isolators are constructed from many layers of elastomer reinforced with steel shims, and are very stiff in the vertical direction, but soft in the horizontal direction. This enables them to carry the weight of a building, but cause the building to have a fundamental natural frequency that is both lower than that of the same building, if conventionally founded, and the dominant frequencies of strong ground motion.

They appear to be very stable, though the low shear stiffness causes a buckling phenomenon. However, it is straightforward to design them to have a large safety factor against buckling. The buckling of these bearings under compression load is a well-understood phenomenon, and has been widely studied. Buckling theory is based on a linearly elastic analysis. Although the elastomer is not really linearly elastic, the deformation is predominantly one of shear, and typical elastomers used in bearings are very close to linear over a large range of shear strain. While approximate, the linear theory is relatively accurate and adequate for most design purposes.

However, buckling analysis for compression has been used to make an unexpected prediction that the isolator can buckle in tension at a load close to that for buckling in compression. Of course, there are many examples of strange systems that buckle in tension, but these are entirely pathological in that the tension forces are always transferred to compression elements that produce the instability. This is not the case here. The buckling process is in fact tensile. The linear elastic model that leads to both compression and tension buckling is an extremely simple one, so it might be argued that tensile buckling may be an artifact of the model itself and not of the isolator.

---

*Keywords:* steel-reinforced elastomeric seismic isolators, tension and compression buckling, linear theory, nonlinear finite element analysis.

For this reason we had undertaken a numerical simulation study using a finite element model of a multilayer elastomeric bearing to check whether the prediction of tensile buckling by the simple linear elastic theory is, in fact, accurate. We found this to be the case. The essential point is that the mechanics of the isolator in tension are the mirror image of that for the isolator in compression. In particular, when the isolator is in compression below the buckling load but laterally displaced, the layers in the center experience rotations that give the vertical load a component along the layer causing a shear deformation. In tension, the layers in the center experience rotations in the opposite direction giving a shear deformation due to the tensile force that permits the top of the isolator to move upwards by a much larger displacement than that which could be sustained in pure tension with no lateral displacement.

While the bearings are normally in compression, base-isolated tall buildings in near-fault locations can lead to situations where peripheral bearings in the isolation system can be required to take some amount of tension. This tension is caused by global overturning of the building produced by the lateral inertial force at the center of the mass of the isolated building. The maximum inertial force and the resulting maximum overturning movement occur at the same time as the maximum lateral displacement of the isolators; at first sight this would seem to be a critical situation. The value of the buckling analysis is that it demonstrates that the condition of the isolator in tension and shear is not as dire as had been feared. In tension, the layers in the center experience a rotation which allows a shear deformation caused by the tensile force and permits the top of the isolator to move upwards by a much larger displacement than that which could be sustained in pure tension with no lateral displacement. Thus the simultaneous occurrence of tension and shear in the isolator prevents the development of damage due to cavitation.

## 2. Overview

Earliest theoretical approaches to study the stability of rubber bearings by [Haringx \[1948; 1949a; 1949b\]](#) were based on linearity of the rubber material and small displacements. Theoretical predictions of the decrease in horizontal stiffness with increasing axial load based on Haringx's theory were made by [Gent \[1964\]](#) and [Derham and Thomas \[1981\]](#). [Simo and Kelly \[1984\]](#) used finite element modeling to study the variation of lateral load-displacement behavior under increasing axial load.

An extensive experimental study of low shape factor elastomeric isolators used for base isolation was given by [Aiken et al. \[1989\]](#). Buckling tests were conducted on doweled bearings; they consisted of applying monotonically increasing axial load to a bearing with the top of the bearing free to displace in the horizontal direction. The tests showed that the analytical formula gives a higher value of the critical load than the experimentally measured one.

[Roeder et al. \[1987\]](#) and [Stanton et al. \[1990\]](#) studied the stability of laminated elastomeric bearings experimentally and theoretically with due consideration given to axial shortening. [Buckle and Kelly \[1986\]](#) studied the stability of elastomeric bearings using a model bridge deck tested using a shaking table. Since the bearings were doweled, bearing overturning or rollover was clearly evident in these tests. [Koh and Kelly \[1986; 1988; 1989\]](#) developed a viscoelastic stability model and a mechanical model based on bearing test results. A comprehensive study of the basic theory and its application to design issues, including the stability problem, was presented by [Kelly \[1997\]](#).

Experimental determination of critical buckling behavior of steel-reinforced bearings at high shear strain was conducted by [Buckle and Liu \[1994\]](#). Studies by [Nagarajaiah and Ferrell \[1999\]](#) introduced

a nonlinear analytical model based on the Koh–Kelly model and included large displacements, large rotations, and nonlinearity of the rubber. The model was verified through experimental results. Further experimental work on the stability of elastomeric bearings was presented by Buckle et al. [2002]. It was shown that the critical buckling load decreases with increasing horizontal displacement or shear strain. Finite element analysis results conducted on a plane model of the bearings with a coarse mesh were compared with the experimental results.

Some results on finite element analysis of the multilayered elastomeric bearings can be found in several papers and reports; see, for example, [Takayama et al. 1992].

### 3. Formulation of elementary stability theory of multilayer elastomeric isolators

The elementary theory for the buckling of isolation bearings treats the bearing as a continuous homogeneous beam in which plane sections normal to the undeformed axis remain plane but not necessarily normal to the deformed axis. The deformation is defined by three functions  $u(x)$ ,  $v(x)$ ,  $\psi(x)$ , which are the axial and lateral displacements of the centroidal axis and the rotation of a section normal to the undeformed axis, respectively. The overall shear deformation  $\gamma(x)$  of the section is the difference between the rotations of the centroidal axis and the section, namely,  $\gamma(x) = v'(x) - \psi(x)$ .

The internal forces on the deformed plane section are the axial load  $N(x)$  normal to the section, the shear force  $V(x)$  parallel to the section and the bending moment  $M(x)$ , as shown in Figure 1. These internal forces are related to the deformation quantities through

$$N(x) = E_c A_s u'(x), \quad V(x) = G A_s (v' - \psi), \quad M(x) = E I_s \psi'(x).$$

$A_s$  is the cross sectional area  $A$  increased by  $h/t_r$ , where  $h$  is the total height of the bearing (rubber plus steel),  $t_r$  is the total thickness of rubber. and  $E_c$  is the compression modulus of the bearing. The value of  $E_c$  for a single rubber layer is controlled by the shape factor  $S = (\text{loaded area})/(\text{force-free area})$ , which is a dimensionless aspect ratio of a single layer of the elastomer. For example,

$$S = \begin{cases} S = b/t, & \text{infinite strip of width } 2b \text{ with a single layer thickness } t, \\ S = R/2t, & \text{circular pad of diameter } R \text{ and thickness } t, \\ S = a/4t, & \text{square of side } a \text{ and thickness } t. \end{cases}$$

In a circular isolator  $E_c = 6GS^2$ , while in the long strip isolator that will be used for the numerical simulation  $E_c = 4GS^2$ . The increase in the area is needed to account for the fact that the steel does not deform in the composite beam.  $I_s$  is the effective moment of inertia of the cross section. This is modified the same way as  $A$ . There is an additional modification to account for the fact that in the isolator the pressure distribution that generates the internal bending moment is a cubic parabola in contrast to regular beam theory where the bending stress distribution is linear. The effective bending stiffness, accounting for these two effects is [Kelly 1997]

$$E I_s = \begin{cases} \frac{1}{3} E_c I \frac{h}{t_r}, & \text{circular bearing,} \\ \frac{1}{5} E_c I \frac{h}{t_r}, & \text{strip bearing.} \end{cases}$$

Under the kinematic assumptions of this beam theory, the downward deflection of the top of the composite column representing the isolator and the resulting external work done by the applied load  $P$  are, respectively,

$$v(x) = \frac{1}{2} \int_0^h (2v'\psi - \psi^2)dx, \quad W_{\text{ext}} = -Pv(x) = -\frac{P}{2} \int_0^h (2v'\psi - \psi^2)dx.$$

The internal stored energy of the column is given by

$$W_{\text{int}} = \frac{1}{2} \int_0^h N(x)u'dx + \frac{1}{2} \int M(x)\psi'(x)dx + \frac{1}{2} \int V(x)(v'(x) - \psi(x))dx.$$

Application of the method of virtual work to the total work  $W_{\text{int}} + W_{\text{ext}}$  with respect to the virtual displacements  $\delta u(x)$ ,  $\delta v(x)$ ,  $\delta \psi(x)$  leads to the following set of equilibrium equations

$$N' = 0, \quad V' - P\psi = 0, \quad M' + V + P(v' - \psi) = 0. \tag{1}$$

The boundary conditions for Equations (1) are shown in Figure 1; the external loads applied to the column at  $x = 0$  are the axial load  $P$  (or  $T$ ), a transverse load  $H_o$  and a moment  $M_o$ .

From Equation (1)<sub>1</sub> we have  $N = \text{const} = -P$ . Integrating the second equation using the consistent boundary condition gives

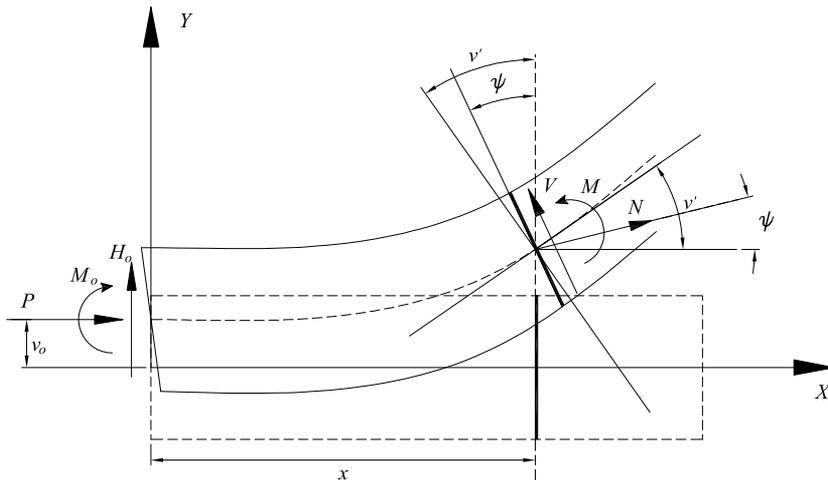
$$V - P\psi = -H_o, \tag{2}$$

and inserting  $V = P\psi - H_o$  into the third equation gives  $M' + Pv' = H_o$ , which can be integrated to

$$M + Pv = M_o + Pv_o + H_o x, \tag{3}$$

where  $v_o = v(0)$ . These two integrated versions of equilibrium equations (2) and (3) are the starting point for the stability analysis of the isolator. When the constitutive equations are included, these become

$$EI_s\psi' + Pv = Pv_o + M_o + H_o x, \quad GA_s(v' - \psi) - P\psi = -H_o.$$



**Figure 1.** Geometry and loads in theoretical study of buckling.

The second equation above can be used in two ways to give a pair of uncoupled equations for  $v$  and  $\psi$ . First we can write  $\psi$  in terms of  $v$  in the form  $\psi = (GA_s v' + H_o)/(GA_s + P)$ , and substitute  $\psi'$  into the first equation to get

$$EI_s \frac{GA_s}{GA_s + P} v'' + Pv = Pv_o + M_o + H_o x. \quad (4)$$

We can also express  $v$  in terms of  $\psi$  as  $v' = (P + GA_s)\psi/(GA_s) - (H_o)/(GA_s)$ , which, when substituted into the first, leads to

$$EI_s \frac{GA_s}{GA_s + P} \psi'' + P\psi = H_o. \quad (5)$$

Thus, Equations (4) and (5) for two kinematic variables have similar form, but different right hand sides. The most general form of solution, taking into account the connections between  $v$  and  $\psi$ , is

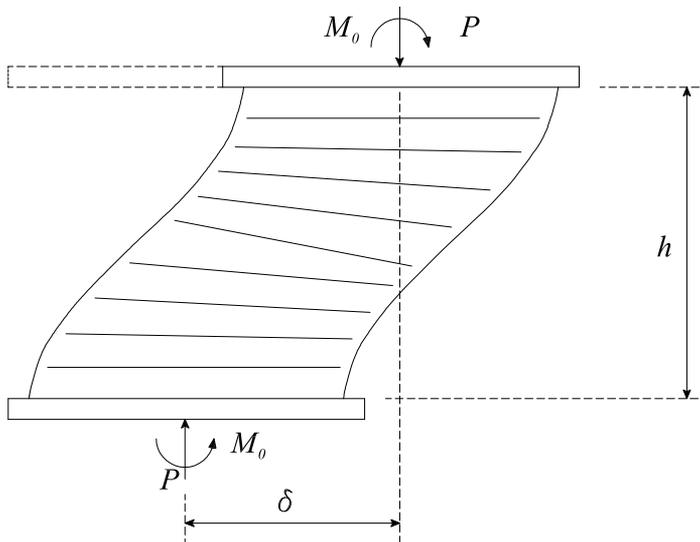
$$v = A \cos \alpha x + B \sin \alpha x + v_o + \frac{M_o}{P} + \frac{H_o}{P} x, \quad \psi = \alpha \beta B \cos \alpha x - \alpha \beta A \sin \alpha x + \frac{H_o}{P},$$

where  $A$  and  $B$  are constants of integration and  $\alpha$ ,  $\beta$  are given by

$$\alpha^2 = \frac{P(P + GA_s)}{EI_s GA_s}, \quad \beta = \frac{GA_s}{P + GA_s}.$$

These equations are now used to predict the buckling behavior of a bearing in an isolation system. As shown in Figure 2, the isolator is constrained against displacement and rotation at the bottom, against rotation at the top, but is free to move laterally at the top, which give boundary conditions  $v(0) = 0$ ,  $\psi(0) = 0$ ,  $\psi(h) = 0$ ,  $H_o = 0$  leading to  $\alpha h = \pi$ . The deformed configuration becomes

$$v(x) = \frac{1}{2} \delta_h \left( 1 - \cos \frac{\pi x}{h} \right), \quad \psi(x) = \frac{1}{2} \alpha \beta \delta_h \sin \frac{\pi x}{h}.$$



**Figure 2.** Boundary conditions for isolation bearing under a vertical load  $P$ . The bearing buckles with no lateral force constraint, but is prevented from rotating at each end.

The above result  $\alpha^2 = \pi^2/h^2$  means that

$$\frac{P(P + GA_s)}{GA_s} = \frac{\pi^2 EI_s}{h^2} = P_E,$$

where  $P_E$  is the Euler load for the column without shear deformation. Denoting  $GA_s$  by  $P_S$ , the above equation for the buckling load can be recast as a quadratic  $P^2 + PP_S - P_S P_E = 0$ , which gives two critical loads: a compression load  $P_C$  and a tension load  $P_T$

$$P_C = \frac{-P_S + \sqrt{P_S^2 + 4P_S P_E}}{2}, \quad P_T = \frac{-P_S - \sqrt{P_S^2 + 4P_S P_E}}{2}. \tag{6}$$

For all reasonable values of the shape factor  $S$ , the  $P_S$  is so much less than  $P_E$  that it can be neglected, and the two critical loads can be approximated by

$$P_{C,T} = \pm \sqrt{P_S P_E}. \tag{7}$$

The significance of the tensile critical load becomes clear when we replace the generic load  $P$  in the formulae for  $\alpha$  and  $\beta$  by  $-T$ , with the assumption  $T \geq GA_s$ , giving

$$\alpha^2 = \frac{T(T - GA_s)}{EI_s GA_s}, \quad \beta = -\frac{GA_s}{T - GA_s},$$

showing that the buckled shape in tension  $v(x)$  is the same as that in compression, but the rotation  $\psi(x)$  is reversed, and the central layers of the bearing are rotated in the direction that facilitates the upward movement of the top through a rotated shear deformation. Because of the natural symmetry of shear, there is an intrinsic symmetry here between compression and tension.

#### 4. Modification for change in length of column prior to buckling

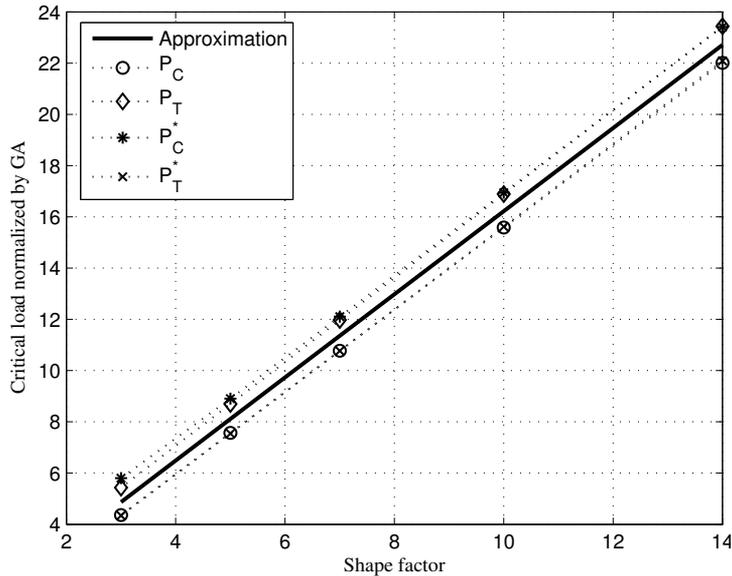
One interesting aspect of the comparison between the buckling loads predicted by the theory and those obtained by the numerical simulation is the fact that while the theory has the buckling load in tension always slightly higher than that in compression for the same shape factor (the difference is  $GA_s$ ), in the numerical results the buckling load in compression is always higher than that in tension. The reason for this is that theory neglects the change in length due to the axial load, whereas in the simulation when buckling is initiated, the bearing has shortened or lengthened. For smaller values of the shape factor, the changes in length can be significant.

Using the approximation  $GA_s \ll P$  and the values of  $GA_s, EI_s$  for a long strip bearing, the critical pressures  $p_{crit} = P/A$  are given in the theoretical analysis by  $p_{crit}/G = \pm 2\pi bS/(\sqrt{15}t_r)$ , where  $t_r$  is the total thickness of rubber in the bearing. To compare the theory with the simulation we replace  $t_r$  by

$$t_r = t_r^o \left( 1 \mp \frac{P_{crit}}{4GS^2} \right),$$

where the minus sign is for compression and the plus for tension. The buckling loads are then given by

$$\frac{p_{crit}}{G} = \pm \frac{2\pi bS}{\sqrt{15}t_r^o \left( 1 \mp \frac{P_{crit}}{4GS^2} \right)}.$$



**Figure 3.** Theoretical buckling loads with and without global bearing deformation: Equation (6) versus (8); solid line corresponds to load approximation from (7).

Solving the last expression for  $p_{\text{crit}}/G$  and taking into account that  $P = pA$  leads to

$$\frac{P_{C,T}^*}{GA} = \pm 2S^2 \left\{ 1 - \sqrt{1 \mp \frac{2\pi b}{St_r^0 \sqrt{15}}} \right\}. \quad (8)$$

Compression and tension critical loads computed by means of Equation (8) are presented in Figure 3.

It is easy to see that the value in compression is always larger than that in tension. The difference between  $|P_{\text{crit}}/GA|$  in compression and tension is approximately  $2(2\pi^2 b^2)/(15t_r^0{}^2)$ . For the case of the numerical models studied we have  $b = t_r^0$ . The difference is about  $8/3$  and, thus, becomes less important with increasing  $S$ .

## 5. Numerical modeling of buckling in tension

In order to verify the tension buckling predicted by the simple analytical theory we use the general purpose finite element ABAQUS application [ABAQUS 2001] to model a steel-reinforced bearing and to study the buckling behavior in both tension and compression [Kelly and Takhirov 2004].

**5.1. Modeling details.** Five finite element models of a bearing are created. These numerical models have the same width and total rubber thickness, and steel shims have the same thickness also. The only difference between them is the shape factor of the bearing. The total thickness of rubber layers is the same in each model, but the thickness of a single rubber layer varies from model to model. The correspondence between the model name and the shim thickness is given in Table 1.

The finite element analysis on the elastomeric bearings is restricted to plane strain. The  $Oy$ -axis of the coordinate system is a vertical axis that extends across the steel shims and rubber layers, while the

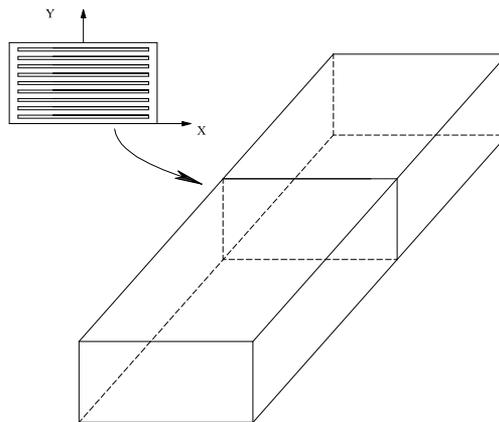
Model	Steel shim thickness	Total rubber thickness	Width	Rubber layer thickness	Shape factor
Model 1	2.60	80.01	160.02	5.72	14
Model 2	2.60	80.01	160.02	8.00	10
Model 3	2.60	80.01	160.02	11.43	7
Model 4	2.60	80.01	160.02	16.00	5
Model 5	2.60	80.01	160.02	26.67	3

**Table 1.** Numerical models with various shape factors. All lengths in mm.

horizontal axis  $Ox$  corresponds to the lateral direction of the bearing as shown in [Figure 4](#). Generally steel-reinforced rubber bearings have a hole in the middle of the steel plates and a rubber cover on the traction-free sides of the bearing. In order to create a model close to the theoretical one given earlier, the hole in the middle and the rubber cover are not included in the consideration.

In the analysis the end plates of the bearing are assumed to be undeformable. Therefore, in the numerical model, the top rubber layer of the bearing is connected to an absolutely rigid surface with the reference point in the middle at which the vertical load is applied. The bottom surface of the bearing is fixed. Two vertical sides of the bearing model are traction-free. The top surface is restrained against rotation around the  $Oz$ -axis (out of the plane), but is free to move horizontally.

Linearly elastic material properties are assumed for the steel plates with Young's modulus and Poisson's ratio equal to 200, 000 MPa and 0.3, respectively. Rubber materials have very little compressibility compared to their shear flexibility; these materials are usually modeled by a hyperelastic material model. ABAQUS [ABAQUS 2001] has a special family of hybrid elements to model the fully incompressible behavior seen in a rubber material. The following assumptions are made in modeling a rubber material: (1) elastic, (2) isotropic, (3) (almost) incompressible, and (4) includes nonlinear geometric effects.



**Figure 4.** Geometry and coordinate axes of numerical simulation models.

Rubber model	$C_{10}$	$C_{01}$	$C_{20}$	$C_{11}$	$C_{02}$	$D_1$
Polynomial	193.4	-0.1	-0.8	0.2	0	0
neo-Hookean	345.0	0	0	0	0	$9.7 \times 10^{-7}$

**Table 2.** Material parameters for two rubber models.  $C_{ij}$  in kPa,  $D_1$  in  $\text{kPa}^{-1}$ .

Hyperelastic materials are described in terms of a strain energy potential  $U$ , which defines the strain energy stored in the material per unit of reference volume in the initial configuration as a function of the strain at that point in the material. The rubber is selected as a polynomial hyperelastic material of the second order. In this case, the strain energy potential has the form

$$U = \sum_{i+j=1}^2 C_{ij}(I_1 - 3)^i(I_2 - 3)^j + \frac{(J^{\text{el}} - 1)^2}{D_1},$$

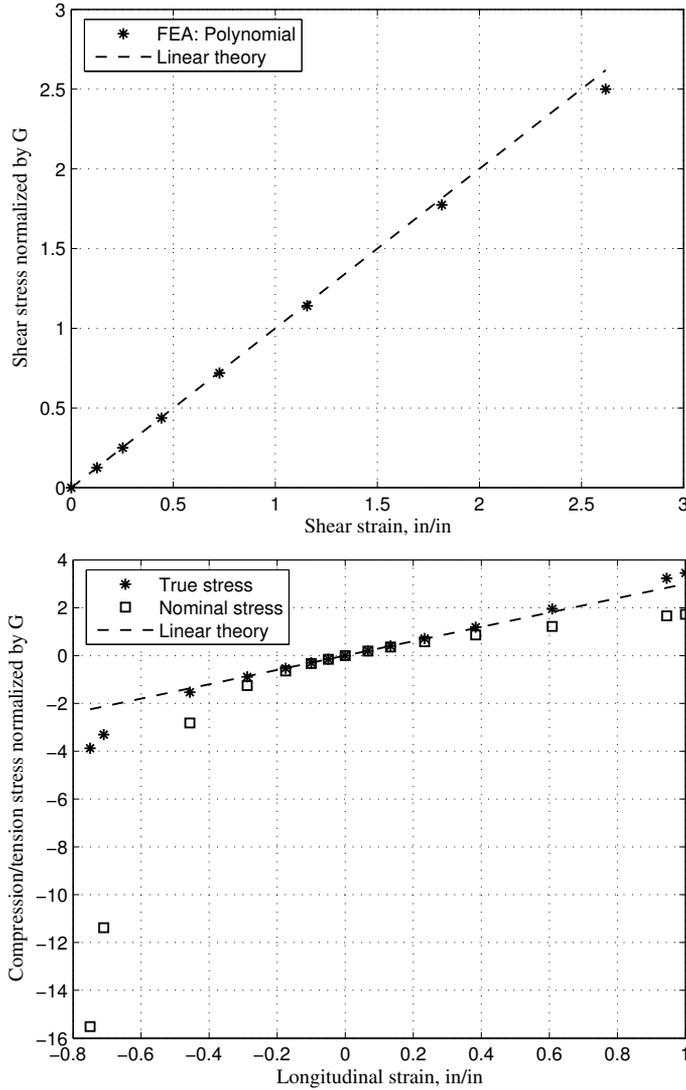
where  $C_{ij}$  and  $D_1$  are the material parameters,  $I_1$  and  $I_2$  are the first and the second invariants of the deviatoric strain, and  $J^{\text{el}}$  is the elastic volume ratio.

Two rubber models are included in the consideration: a fully incompressible (polynomial) model, and an almost incompressible (neo-Hookean) model. The material parameters of the rubber can be expressed in terms of initial shear modulus  $G$ , and initial bulk modulus  $K$  via  $G = 2(C_{10} + C_{01})$ ,  $K = 2/D_1$ . The values of the material parameters for both rubber models are presented in Table 2. Since  $D_1$  is not equal to zero for the neo-Hookean model, this model allows some compressibility in the rubber material.

A supplemental study on the properties of the rubber models is conducted on a rubber cylinder and a rubber layer. The cylinder is used for the rubber material study in tension and compression. The layer, representing one single layer of the rubber locked between two rigid horizontal surfaces, is used to study behavior of the rubber material in shear with no vertical load. While both rubber materials are linearly elastic up to about 250% strain in shear, they exhibit a significant nonlinearity in tension or compression as shown in Figure 5.

**5.2. Critical buckling load in compression and tension.** The model is studied by a classical buckling analysis scheme available in ABAQUS. First the buckling mode of each bearing model is determined. Very small imperfections of about 1% of the steel layer thickness are introduced in the model; they are based on the buckling mode obtained in the buckling mode analysis. The postbuckling behavior is followed up to about 30% shear deformation.

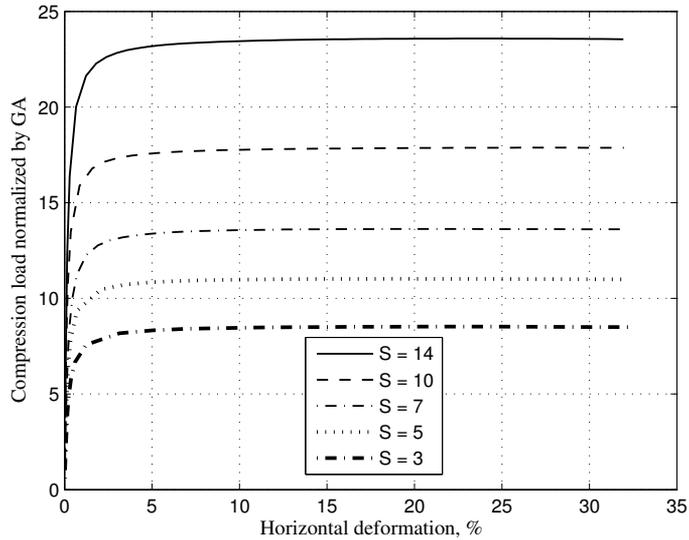
The buckling analysis of the numerical bearing models reveals the following results. All models have significant horizontal drift caused by a large vertical load. The curves of the compression vertical load versus horizontal drift for all numerical models are shown in Figure 6. The critical buckling load increases with the increase in the shape factor. All plots show similar behavior with the majority of the change happening up to a 3% deformation after which they flatten out when the bearing is buckling. Figure 7 presents the corresponding curves of buckling in tension. The critical load is again dependent on the shape factor increasing with it. The buckling in tension is more sudden, so the point at which buckling begins moves very close to the vertical axis, and the shear when the buckling starts in tension



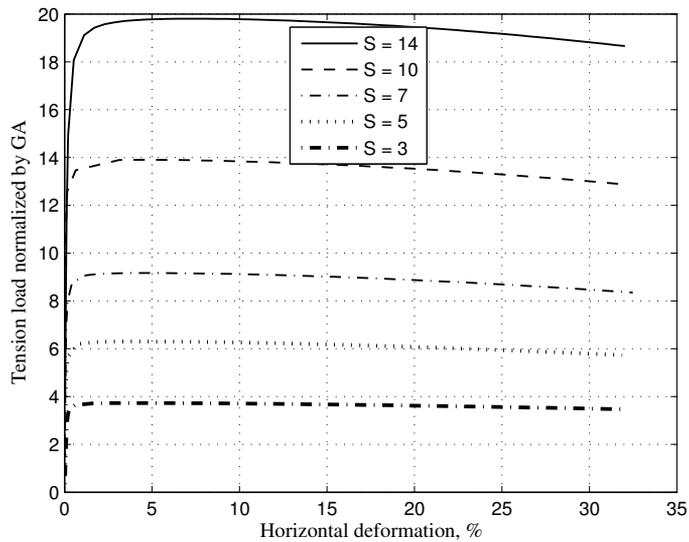
**Figure 5.** Properties of rubber material in (top) shear and (bottom) compression/tension deformation (polynomial rubber material).

can be as low as 0.2%; see Model 5 with the smallest shape factor. Increase in the shape factor moves this point closer to the 3% critical strain obtained for the compression buckling.

Theoretical critical force versus the shape factor for the corresponding numerical model was discussed earlier; see Figure 3. Theoretical buckling compression load is always less than the absolute value of the tension load for the simple theoretical solution presented in Equation (6), which is not consistent with the numerical analysis results shown in Figure 8. The critical load determined by Equation (8) takes into account shortening or elongation of the bearing in the vertical direction. Therefore, it correlates better with numerical results. As the latter show, the compression buckling load is always greater than the absolute value of the tension buckling load shown in Figure 8 by dashed and dot-dashed lines. Figure 8



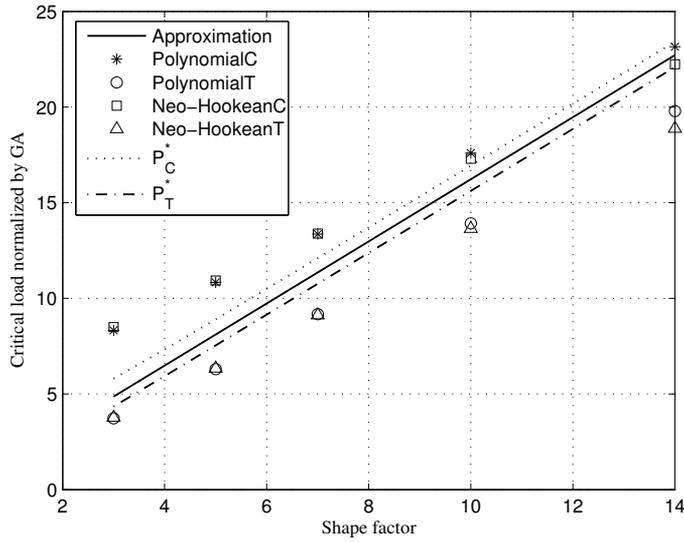
**Figure 6.** Buckling diagram for all models (compression).



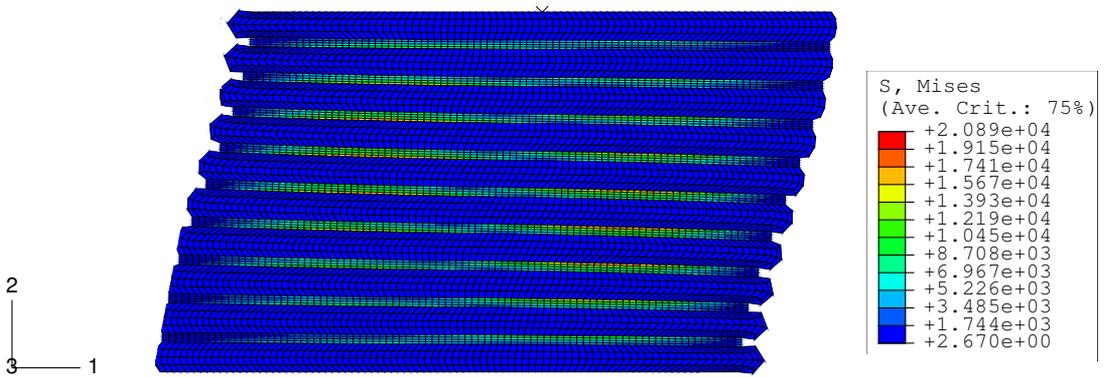
**Figure 7.** Buckling diagram for all models (tension).

also shows no significant differences between critical buckling load for incompressible (polynomial) and compressible (neo-Hookean) rubber materials. As a typical result, Figures 9–12 show deformed shapes of two numerical models in compression and tension.

The numerical results on buckling behavior of the bearing have satisfactory correlation with the theoretical solutions presented earlier. The theoretical study and the finite element analysis lead to the following conclusions. The critical buckling load increases with the shape factor; this behavior is almost

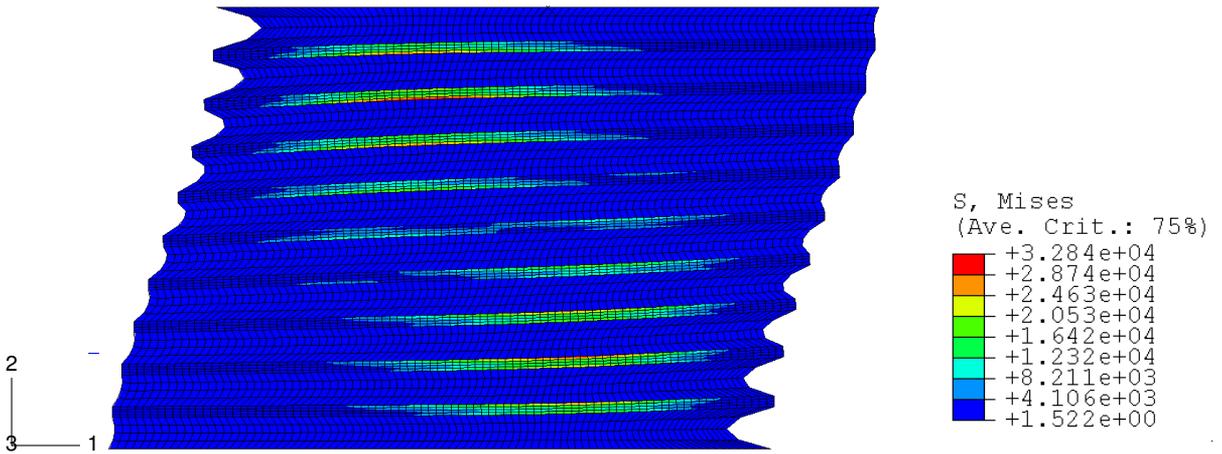


**Figure 8.** Critical buckling load (normalized by GA) versus shape factor.

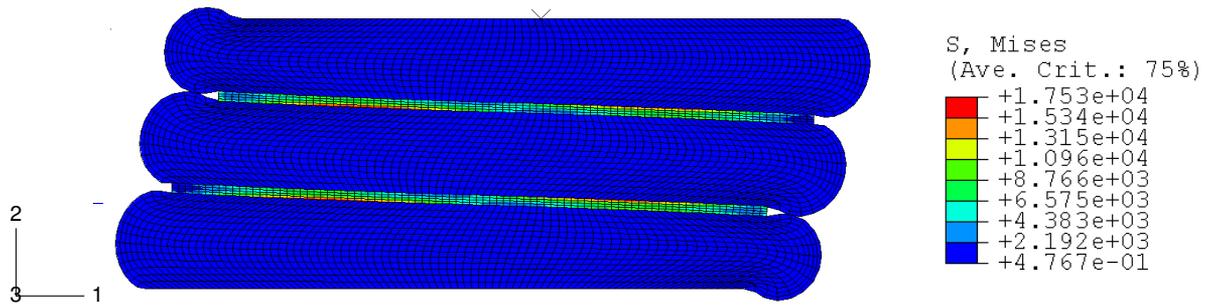


**Figure 9.** Buckled shape and Mises stresses for Model 2 (compression).

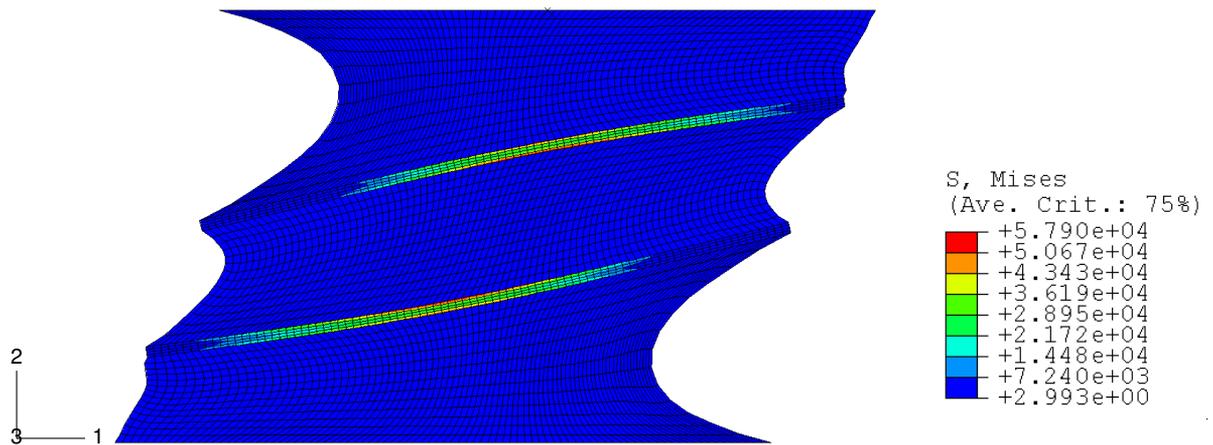
linear as predicted by the theory. The numerical compression buckling load is almost always higher than the theoretically estimated one for all bearings. The compression buckling load is always higher than the absolute value of the corresponding tension load due to nonlinear geometry effects noted for the simple rubber cylinder model. The numerical models have different postbuckling behaviors in compression and tension. In compression, the vertical load remains almost the same after the buckling occurs, and the bearing deflects horizontally. In contrast, the vertical load in tension slowly decreases with horizontal deflection.



**Figure 10.** Buckled shape and Mises stresses for Model 2 (tension).



**Figure 11.** Buckled shape and Mises stresses for Model 5 (compression).



**Figure 12.** Buckled shape and Mises stresses for Model 5 (tension).

## 6. Conclusions

We have shown that numerical analysis confirms that the buckling of a bearing in tension is not an artifact of the way the theory has been set up. Of course, it must be acknowledged that a bearing will experience cavitation before the buckling load can be achieved. However, the theory shows that under high seismic loading, the isolators that experience tension will avoid the possibility of cavitation since the tension loading due to the overturning of the structure is accompanied by large lateral shear. Moreover, due to the interaction between shear and the vertical stiffness, the tension stresses are much less than they would be if the tension force were applied in the absence of shear.

It has been the purpose of this paper to demonstrate that the condition of the isolator in tension and shear is not as dire as has been feared. The analysis has shown that the mechanics of the isolator in tension are the mirror image of those for the isolator in compression. In particular, when the isolator is in compression below the buckling load but laterally displaced, the layers in the center experience a rotation which gives the vertical load a component along the layer and turns the compression displacement into a shear deformation. In tension the rubber layers in the center of the bearing experience a rotation in the opposite direction which allows a shear deformation caused by the tensile force and permits the top of the isolator to move upwards by a much larger displacement than that which could be sustained in pure tension with no lateral displacement. The elastomer can sustain only small strains in the state of triaxial stress generated by pure tension on a multilayer isolator with a large shape factor, but can sustain shear strains on the order of 500–600%. Thus the simultaneous occurrence of tension and shear allows the isolator to avoid the damaging effects of cavitation.

## References

- [ABAQUS 2001] *ABAQUS 6.2 user manual*, Version 6.2, Hibbit Karlsson Sorensen Inc., Pawtucket, RI, 2001.
- [Aiken et al. 1989] I. D. Aiken, J. M. Kelly, and F. F. Tajirian, “Mechanics of low shape factor elastomeric seismic isolation bearings”, Technical report UCB/EERC-89/13, Earthquake Engineering Research Center, University of California, Berkeley, CA, 1989.
- [Buckle and Kelly 1986] I. G. Buckle and J. M. Kelly, “Properties of slender elastomeric isolation bearings during shake table studies of a large-scale model bridge deck”, pp. 247–269 in *Joint sealing and bearing systems for concrete structures*, vol. 1, American Concrete Institute, Detroit, MI, 1986.
- [Buckle and Liu 1994] I. G. Buckle and H. Liu, “Experimental determination of critical loads of elastomeric isolators at high shear strain”, *NCEER Bull.* **8**:3 (1994), 1–5.
- [Buckle et al. 2002] I. Buckle, S. Nagarajaiah, and K. Ferrell, “Stability of elastomeric isolation bearings: experimental study”, *J. Struct. Eng. ASCE* **128**:1 (2002), 3–11.
- [Derham and Thomas 1981] C. J. Derham and A. G. Thomas, “The design of seismic isolation bearings”, pp. 21–36 in *Control of seismic response of piping systems and other structures by base isolation*, Earthquake Engineering Research Center, University of California, Berkeley, CA, 1981.
- [Gent 1964] A. N. Gent, “Elastic stability of rubber compression springs”, *J. Mech. Eng. Sci.* **6**:4 (1964), 318–326.
- [Haringx 1948] J. A. Haringx, “On highly compressible helical springs and rubber rods and their application for vibration-free mountings, I”, *Philips Res. Rep.* **3** (1948), 401–449.
- [Haringx 1949a] J. A. Haringx, “On highly compressible helical springs and rubber rods and their application for vibration-free mountings, II”, *Philips Res. Rep.* **4** (1949), 49–80.
- [Haringx 1949b] J. A. Haringx, “On highly compressible helical springs and rubber rods and their application for vibration-free mountings. III”, *Philips Res. Rep.* **4** (1949), 206–220.

- [Kelly 1997] J. M. Kelly, *Earthquake-resistant design with rubber*, 2nd ed., Springer, London, 1997.
- [Kelly and Takhirov 2004] J. M. Kelly and S. M. Takhirov, "Analytical and numerical study on buckling of elastomeric bearings with various shape factors", Technical report UCB/EERC-2004/03, Earthquake Engineering Research Center, University of California, Berkeley, CA, 2004.
- [Koh and Kelly 1986] C. G. Koh and J. M. Kelly, "Effects of axial load on elastomeric bearings", Technical report UCB/EERC-86/12, Earthquake Engineering Research Center, University of California, Berkeley, CA, 1986.
- [Koh and Kelly 1988] C. G. Koh and J. M. Kelly, "[A simple mechanical model for elastomeric bearings used in base isolation](#)", *Int. J. Mech. Sci.* **30**:12 (1988), 933–943.
- [Koh and Kelly 1989] C. G. Koh and J. M. Kelly, "Viscoelastic stability model for elastomeric isolation bearings", *J. Struct. Eng. ASCE* **115**:2 (1989), 285–302.
- [Nagarajaiah and Ferrell 1999] S. Nagarajaiah and K. Ferrell, "[Stability of elastomeric seismic isolation bearings](#)", *J. Struct. Eng. ASCE* **125**:9 (1999), 946–954.
- [Roeder et al. 1987] C. W. Roeder, J. F. Stanton, and A. W. Taylor, "Performance of elastomeric bearings", NCHRP Rep. 298, Trans. Res. Board, National Research Council, Washington, D. C., 1987.
- [Simo and Kelly 1984] J. C. Simo and J. M. Kelly, "[Finite element analysis of the stability of multilayer elastomeric bearings](#)", *Eng. Struct.* **6**:3 (1984), 162–174.
- [Stanton et al. 1990] J. F. Stanton, G. Scroggins, A. W. Taylor, and C. W. Roeder, "Stability of laminated elastomeric bearings", *J. Eng. Mech. ASCE* **116**:6 (1990), 1351–1371.
- [Takayama et al. 1992] M. Takayama, H. Tada, and R. Tanaka, "[Finite-element analysis of laminated rubber bearing used in base-isolation system](#)", *Rubber Chem. Technol.* **65**:1 (1992), 46–62. Rubber Division, ACS.

Received 31 Jul 2006. Accepted 20 Apr 2007.

JAMES M. KELLY: [jmkelly@berkeley.edu](mailto:jmkelly@berkeley.edu)

*Earthquake Engineering Research Center, 1301 South 46th St., Bldg. 451, University of California, Berkeley, Richmond, CA 94804, United States*

SHAKHZOD M. TAKHIROV: [takhirov@berkeley.edu](mailto:takhirov@berkeley.edu)

*Earthquake Engineering Research Center, 1301 South 46th St., Bldg. 451, University of California, Berkeley, Richmond, CA 94804, United States*

# REPRESENTATIVE VOLUME ELEMENT AND EFFECTIVE ELASTIC PROPERTIES OF OPEN CELL FOAM MATERIALS WITH RANDOM MICROSTRUCTURES

SERGEY KANAUN AND OLEKSANDR TKACHENKO

This work is devoted to the problem of the numerical simulation of effective elastic properties of open-cell foam materials. The Laguerre tessellation procedure is used for the construction of skeletons of random foam microstructures with prescribed distributions of cell diameters. A four-parametric approximation of the ligament shapes in the open-cell foams is proposed. A version of the finite element method, based on the Timoshenko beam finite element, is developed for calculating stresses and strains in the foam ligaments and the solution of the homogenization problem. The size of the representative volume element for reliable calculations of the effective elastic properties of the foam materials is evaluated on the basis of a series of numerical experiments. The dependences of the effective elastic properties of the open-cell foams on cell size distributions and on ligament shapes are obtained and analyzed.

## 1. Introduction

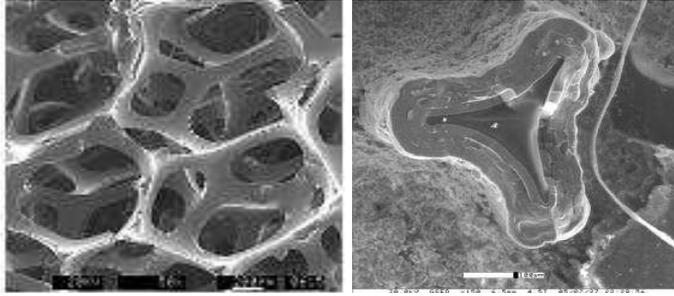
Physical and mechanical properties of carbon foam materials have been extensively studied in recent decades. Interest in these very light and highly thermoconductive materials has increased because of many important areas of their applications. A typical microstructure of the open-cell foams is shown on the left-hand side of [Figure 1](#). It consists of a set of ligaments (fiber like elements) connected to a number of nodes that are, in fact, irregular lumps. The cross section of a typical ligament is presented on the right-hand side of [Figure 1](#). For applications, it is important to predict the dependence of a foam material's mechanical and physical properties on the details of its microstructure. Effective elastic properties of foam-like systems have been studied by many authors [[Gibson and Ashby 1982](#); [Warren and Kraynik 1997](#); [Christensen 2000](#); [Roberts and Garboczi 2002](#); [Gong et al. 2005](#)]. Analytical equations for the Young moduli and the Poisson ratios of such materials were obtained for regular structures, and approximation of an actual random foam structure by a regular one allows us to describe experimental behavior of the effective elastic properties of such materials in some region of microstructural parameters [[Gong et al. 2005](#)]. Nevertheless, in the framework of the regular approach, many important details of mechanical and physical behavior of random foams cannot be described (see discussions of this problem in [Roberts and Garboczi \[2002\]](#) and [Zhu et al. \[2000\]](#)).

In the works of [Roberts and Garboczi \[1999; 2001; 2002\]](#), [Zhu et al. \[2000\]](#), and [Kadashevich and Stoyan \[2005\]](#), elastic properties of the open-cell foams by direct numerical simulations of the elastic

---

*Keywords:* open-cell foams, representative volume element, random space tessellation, homogenization problem, finite element method, elastic constants.

This work was supported by the Texas A& M University (USA) and CONACYT (Mexico) joint program (Proposal 3-050102-1), and AFSOR (Grant FA9550-06-1-0347).



**Figure 1.** Scanning electron microscopy image of an open-cell carbon foam microstructure (left), and the cross section of a typical ligament (right).

fields inside a representative volume element (RVE) of the foam material. The process of numerically simulating the effective elastic properties of open-cell foams consists of the following steps: First, the microstructure (skeleton) of the foam material is constructed using statistical models of the actual foam microstructure. As a rule, the Voronoi tessellation procedure is used to produce a set of polyhedron cells inside the chosen RVE. The geometry of the open-cell foam in the RVE is defined by using the edges of the Voronoi polyhedra as the axes of the ligaments, and by choosing a certain approximation of the ligament shapes. Next, the finite element technique is applied for calculation of stresses and strains in the ligaments for given boundary conditions on the surface of the RVE. Finally, the values of the effective elastic constants of the foam are obtained by averaging detailed strain and stress fields over the RVE.

The first problem that must be addressed in carrying out numerical simulations is the appropriate choice of RVE size. If the RVE is a cube, one has to point out the number of cells that should be taken inside this cube by the numerical simulations in order to obtain reliable values for the effective elastic constants of the foam material. Usually, the number of cells in the RVE is restricted by capacities of available computers and software. Standard finite element packages permit the consideration of RVEs that contain hundred of cells. Nevertheless, in some works (see [Kadashevich and Stoyan 2005]) the authors came to the conclusion that the number of cells in the RVE should be more than a thousand in order to obtain reliable values of the effective elastic constants.

Another problem that must be addressed in carrying out numerical simulations of the effective properties of foams is the construction of the foam microstructures with the law of the cell size distribution that corresponds to the one observed in the actual foams. In fact, the Voronoi tessellation procedure does not permit the simulation of microstructures with predetermined distributions of the cell sizes. To be exact, it is impossible to point out the positions of seed points in the RVE that produce the Voronoi polyhedra with the prescribed distribution of diameters.

The influence of ligament shape on the properties of open-cell foams is another important problem that has not been considered sufficiently in the literature. The ligaments in the actual foams have rather complex geometry. An adequate description of this geometry, and an analysis of its influence on the effective properties, are important tasks of numerical simulations.

The paper is focused on the above-mentioned problems, and its structure is as follows:

- (1) In [Section 2](#), the problem of computer simulation of skeletons of the open-cell foams is considered. We use the Voronoi and Laguerre tessellation procedures in order to construct a set of cells inside

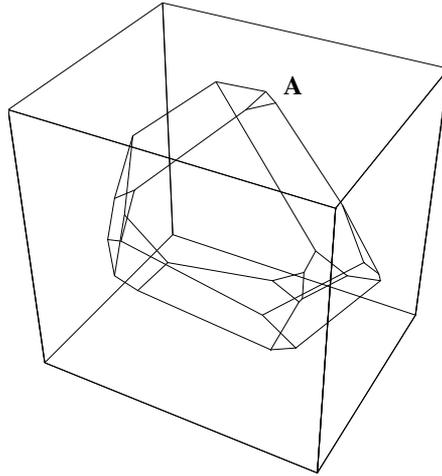
a cubic RVE. The Laguerre procedure allows us to simulate foam skeletons with a given known distribution of cell diameters.

- (2) In [Section 3](#), an analytical four-parametric approximation of the ligament form is proposed. This approximation reflects the most important features of the ligaments in actual foams.
- (3) In [Section 4](#), a version of the finite element method based on the Timoshenko beam theory is developed for calculation of stresses and strains in the ligaments of the open-cell foam structures. In this method, every ligament is considered as one finite element (super element), and it results that the RVE with several thousands of cells may be considered. The appropriate size of the cubic RVE is assessed in [Section 5](#). Performing series of numerical experiments we show that the number of cells in the RVE should be about 900–1000 in order to obtain reliable values of the effective elastic constants of the open-cell foams.
- (4) The influence of the ligament form and the cell size distribution on the effective elastic properties of foams is studied in [Section 6](#). Some details of the proposed method and its possible area of application are discussed in the conclusion.

## 2. Computer simulation of skeletons of open-cell foam materials

A conventional method for carrying out computer simulations of the microstructures of open-cell foams, shown in [Figure 1](#), is based on the Voronoi tessellation algorithm. In our study, this algorithm is used in the following specific form. Let us consider a cube  $V : \{|x_1| < 1, |x_2| < 1, |x_3| < 1\}$  centered at the origin of the Cartesian coordinate system,  $(x_1, x_2, x_3)$ . First, a random set  $S_{(0,0,0)}$  of the so-called seed points, homogeneously distributed inside this cube, is generated. After that, the sets  $S_{(i,j,k)}$  ( $i, j, k = 0, \pm 1$ ) that have mirror-like symmetry to the set  $S_{(0,0,0)}$  with respect to the planes  $x_i = \pm 1$  ( $i = 1, 2, 3$ ), are constructed. For instance, the set  $S_{(1,0,0)}$  is mirror-like to the set  $S_{(0,0,0)}$  with respect to the plane  $x_1 = 1$ , the set  $S_{(0,1,0)}$  is symmetric to the set  $S_{(0,0,0)}$  with respect to the plane  $x_2 = 1$ , etc. The union of these sets  $\bar{S} = S_{(0,0,0)} \cup S_{(1,0,0)} \cup \dots \cup S_{(0,0,-1)}$  is the set of seed points under consideration. The Voronoi polyhedron that corresponds to a seed point  $x^{(i)}$  with the coordinates  $(x_1^{(i)}, x_2^{(i)}, x_3^{(i)})$  consists of the points of three-dimensional space that are closer to the point  $x^{(i)}$  than to any other seed point  $x^{(j)}$ . The Voronoi polyhedra that correspond to the seed points of the set  $S_{(0,0,0)}$  perform a tessellation of the original cube  $V$ . Because of the mirror symmetry of the sets  $S_{(i,j,k)}$  ( $i, j, k = 0, \pm 1$ ) with respect to  $S_{(0,0,0)}$ , the borders of the cube  $V$  belong to the polyhedron surfaces. For the realization of the Voronoi tessellation procedure, the algorithm proposed by [Tenemura et al. \[1983\]](#) was adopted in this work.

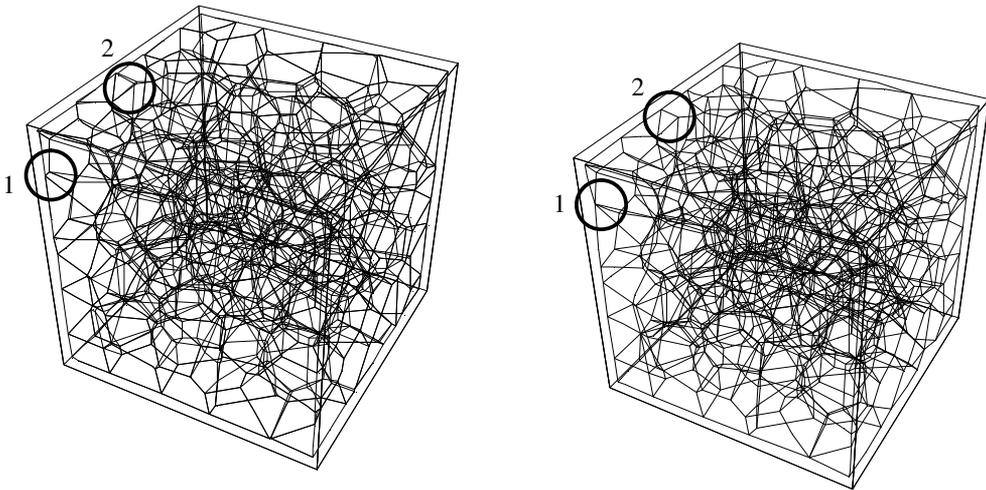
A typical polyhedron obtained by the Voronoi tessellation procedure is shown in [Figure 2](#). The edges of this polyhedron compose the ligament axes of the future foam microstructure, and the polyhedron vertices are the nodes where the ligaments are connected. Note that very often the Voronoi tessellation produces polyhedra whose vertices are situated too closely to each other (e.g. the set  $A$  of nodes in [Figure 2](#)), and the lengths of the ligaments that connect these close nodes are very short. Because every ligament has a finite volume, these short ligaments turn out to be inside other ligaments that connect more distant nodes. It means that the obtained tessellation model should be altered (cleaned): the clusters of the nodes that are closer to each other than a characteristic size of the ligament cross-sections should be joined into one node. An additional reason for such a procedure is that the elastic deformation of very short ligaments



**Figure 2.** A typical polyhedron obtained by the Voronoi 3D-space tessellation process.

cannot be described by the beam theory that is used in the finite element technique adopted in this study (Section 4). If the number of such short ligaments is large, the error of the calculation of elastic fields based on the beam finite elements may be essential.

In Figure 3, the Voronoi tessellation of a cube with two hundred cells is presented. The seed points corresponding to these structure were homogeneously distributed inside the cube under an additional condition that the distance between two different points  $x^{(i)}$  and  $x^{(j)}$  should be more than 0.3. The



**Figure 3.** The Voronoi tessellation of the cube (left) and the same structure after elimination of two small polyhedron surfaces (right). (In the regions 1 and 2 in the right picture, small polyhedron surfaces are eliminated).

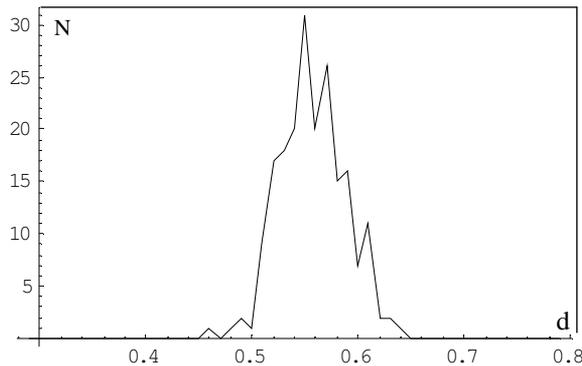
images of the tessellation before and after the "cleaning" procedure are presented in the left and right hand sides of Figure 3. In the right hand side of Figure 3, the nodes that are closer than 0.05 are joined into one node. If the original left hand side structure consists of 3855 ligaments with the maximal and minimal lengths  $0.5559$  and  $1.26 \cdot 10^{-6}$  and the mean length  $0.138$ , the final structure contains 1971 ligaments of the maximal and minimal lengths  $0.557$  and  $0.05$ , and the mean ligament length is equal to  $0.205$ . The difference between two microstructures may be observed in regions 1 and 2 in the left and right hand sides of Figure 3.

The distribution function  $f(d)$  of the diameters  $d$  of the cells shown in Figure 3 is presented in Figure 4. Note that this distribution is difficult to control in the framework of the Voronoi tessellation procedure. In actual carbon foam materials, however, the cell size distributions are rather specific. In many cases, the experimental histograms of the cell diameters are close to a linear distribution law:  $f(d) \approx ad$ , where  $a$  is an appropriate constant (K. Lafdi, 2004, private communication). The diameter  $d$  of a cell is calculated from the equation  $d = 2\sqrt[3]{3v/(4\pi)}$ , where  $v$  is the volume of the cell.

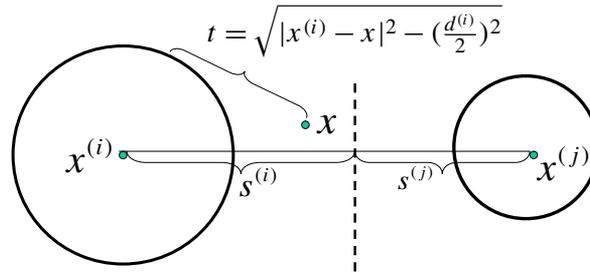
For simulation of the microstructure of the foam materials with a prescribed distribution of cell diameters, the so-called Laguerre tessellation procedure may be used. The application of this procedure consists of two steps. First, a set of balls of random diameters  $d^{(i)}$ , whose distribution coincides with the distribution of cell diameters of the simulated foam, is generated. After that, these balls are closely packed inside the RVE (cube  $V$ ). As a result, we obtain a set of ball centers  $x^{(i)}$  and a set of diameters  $d^{(i)}$  associated with these centers. Next, this information is used for the Laguerre tessellation of the cube  $V$ . Following Aurenhammer and Klein [2000], let us consider a ball  $B^{(i)}$  of the diameters  $d^{(i)}$  centered at the point  $x^{(i)}$ . A positive scalar  $t(x^{(i)}, x)$ ,

$$t(x^{(i)}, x) = \sqrt{|x^{(i)} - x|^2 - \left(\frac{d^{(i)}}{2}\right)^2}, \tag{1}$$

is the distance from a point  $x$  ( $x \notin B^{(i)}$ ) to the surface  $S^{(i)}$  of  $B^{(i)}$  along the tangent line to this surface (see Figure 5). The  $i$ th Laguerre polyhedron consists of the union of the points of the  $i$ th ball and the points  $x$  for which the parameter  $t(x^{(i)}, x)$  is less than such a parameter  $t(x^{(j)}, x)$  for any other  $j$ th ball,



**Figure 4.** A typical histogram of the cell diameters obtained after the Voronoi tessellation process.



**Figure 5.** The Laguerre tessellation algorithm; the point  $x$  belongs to the  $i$ th Laguerre polyhedron if the parameter  $t$  is smaller for the point  $x^{(i)}$  than for any other seed point  $x^{(j)}$ .

$i \neq j$ . It is possible to show that the common border of the Laguerre polyhedron that correspond to two neighbor balls  $i$  and  $j$  centered at the points  $x^{(i)}$  and  $x^{(j)}$ ,  $|x^{(i)} - x^{(j)}| \geq (d^{(i)} + d^{(j)})/2$ , is orthogonal to the interval connecting the centers of these balls, and that the border plane intersects this interval in the proportion  $s^{(i)}/s^{(j)}$ ,

$$\frac{s^{(i)}}{s^{(j)}} = \frac{4|x^{(i)} - x^{(j)}|^2 + (d^{(i)})^2 - (d^{(j)})^2}{4|x^{(i)} - x^{(j)}|^2 + (d^{(j)})^2 - (d^{(i)})^2}. \quad (2)$$

The intervals  $s^{(i)}$  and  $s^{(j)}$  are indicated in [Figure 5](#), and the dashed line in this figure is the border of the Laguerre polyhedra. It is shown in [\[Aurenhammer and Klein 2000\]](#) that all Laguerre polyhedra are convex and span three-dimensional space.

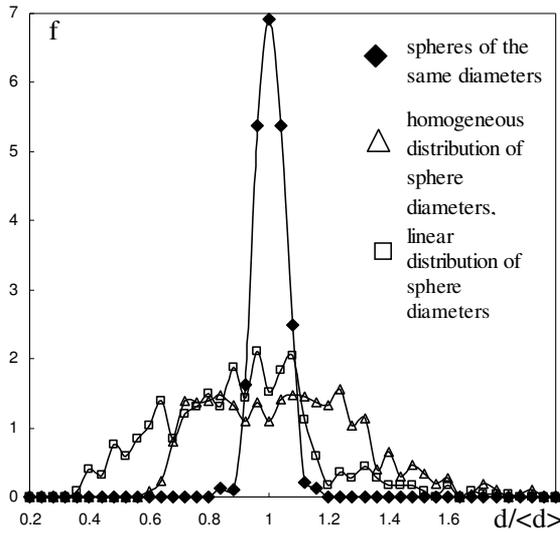
The algorithm of packing used in this work is based on the following procedure. We start with the first ball centered at the origin, and the center  $x^{(i)}$  of the  $i$ th ball is defined such that  $x^{(i)}$  has minimal distance from the center of the cube, and  $|x^{(i)} - x^{(j)}| \geq (d^{(i)} + d^{(j)})/2$  for  $j = 1, 2, 3, \dots, i - 1$ . The number of the balls is increased until it is impossible to find a center for the next ball inside the cube.

The algorithm of [Tenemura et al. \[1983\]](#) is also adopted for carrying out the Laguerre tessellation procedure.

[Figure 6](#) presents the histograms of three distribution functions  $f(d)$  of the diameters of the Laguerre polyhedra for different distributions of the diameters of the initial balls. The line with black rhombs corresponds to a set of balls of approximately the same diameters, the line with triangles corresponds to a homogeneous distribution of the ball diameters in the region  $(0.5 \langle d \rangle, 1.5 \langle d \rangle)$ , and the line with squares corresponds to a linear distribution of the ball diameters in the interval  $(0.3 \langle d \rangle, 1.2 \langle d \rangle)$ . Here  $\langle d \rangle$  is the mean value of the ball diameters.

### 3. Approximation of the form a typical ligament in the open-cell foams

As seen from [Figure 1](#), the cross-section of a typical ligament has a quasitriangular form (see also the discussion of the forms of ligaments in carbon foams in [\[Gong et al. 2005\]](#)). In this study, we treat a typical ligament as a direct beam with a quasitriangular cross-section that changes along the ligament



**Figure 6.** The distribution function of cell diameters after the Laguerre tessellation for random packing of spheres of approximately the same diameter (line with black rhombs), for packing of spheres with the homogeneous distribution of the diameters in the interval  $\{0.5 \langle d \rangle, 1.5 \langle d \rangle\}$  (line with triangles), and for packing of spheres with linear distribution of diameters on the interval  $\{0.3 \langle d \rangle, 1.2 \langle d \rangle\}$  (line with squares).

axis. Taking into account that function  $f(\zeta)$  of a complex variable  $\zeta$ ,

$$f(\zeta) = R \left( \frac{1}{\zeta} + \frac{\zeta^2}{a_1} \right), \quad a_1 \geq 2,$$

maps a unit circle in the  $\zeta$ -plane into a quasitriangular region in the  $w = f(\zeta)$ -plane, we define the border of the ligament cross-section by the equations

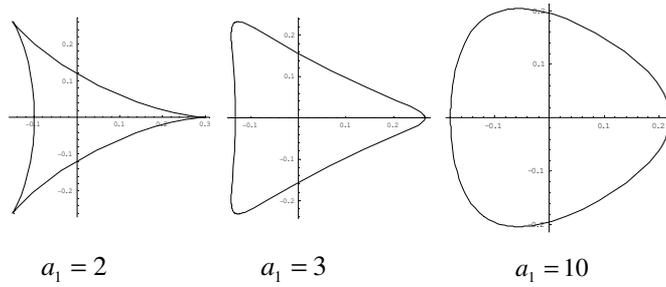
$$\begin{aligned} y(\varphi, x) &= R(x) \left( \cos(\varphi) + \frac{\cos(2\varphi)}{a_1} \right), \\ z(\varphi, x) &= R(x) \left( -\sin(\varphi) + \frac{\sin(2\varphi)}{a_1} \right). \end{aligned} \tag{3}$$

Here  $\varphi$  is the angle parameter,  $0 \leq \varphi < 2\pi$ ,  $(y, z)$  are the Cartesian coordinates in the plane of the ligament cross-section, and the coordinate  $x$  is directed along the ligament axis. The function  $R(x)$  that defines the change of the ligament along its axis is taken in the form

$$R(x) = a_2 [1 - a_3 \xi (1 - \xi)], \quad \xi = \frac{x}{l}, \quad 0 < x < l, \tag{4}$$

which reflects the experimental fact that the ligaments are thinner in the middle region than in the regions near their ends:  $x = 0, l$ .

Parameters  $a_1, a_2, a_3$  and  $l$  define the global shape of the ligament. The cross-sections of the ligament by the plane  $x_3 = 0$  is presented in [Figure 7](#) for the parameters  $a_2 = 0.3$  and  $a_1 = 2, 3, 10$ . The shape



**Figure 7.** Ligament cross-sectional area different values of the parameter  $a_1$  in Equation (3),  $a_2 = 0.3$ ,  $\xi = 0$ .

of a typical ligament that corresponds to Equation (3) and Equation (4) is presented in Figure 8 for the parameters  $a_1 = 2.5$ ,  $a_2 = 0.3$ ,  $a_3 = 0.4$ .

The proposed approximation allows us to calculate the basic geometrical characteristics of the ligament in closed analytical form. For instance, the area  $S$  of the cross-section of the ligament is

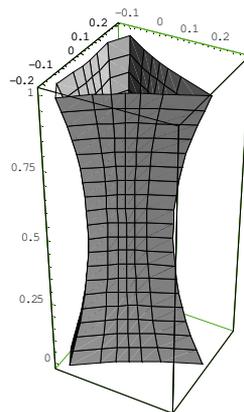
$$S(a_1, a_2, a_3, \xi) = \pi a_2^2 \frac{(a_1^2 - 2)}{a_1^2} [1 - 4a_3^2 \xi (1 - \xi)]^2. \tag{5}$$

The main moment of inertia  $J$  of the cross-section takes the form

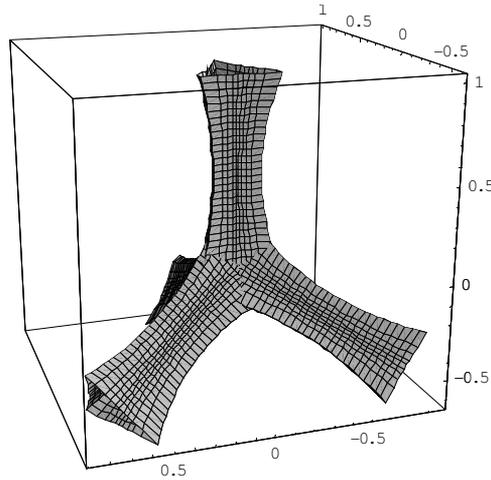
$$J(a_1, a_2, a_3, z) = \pi a_2^4 \frac{(a_1^4 - 2a_1^2 - 2)}{4a_1^4} [1 - 4a_3^2 \xi (1 - \xi)]^4. \tag{6}$$

The volume  $V_l$  of the ligament is

$$V_l(a_1, a_2, a_3) = \pi a_2^2 l \frac{(a_1^2 - 2)}{15a_1^2} (15 - 20a_3 + 8a_3^2). \tag{7}$$



**Figure 8.** The shape of a ligament defined by Equations (3) and (4) for  $a_1 = 2.5$ ,  $a_2 = 0.3$ ,  $a_3 = 0.4$ .



**Figure 9.** A typical node of the foam structure where four ligaments intersect.

A typical element of the microstructure of the foam material consists of four ligaments connected in a node. Such an element is shown in Figure 9. For calculating the volume of the hard phase in the RVE, one has to take into account the volume  $V_{int}$  of the intersection of the ligaments in the nodes. The volume  $V_{int}$  may be evaluated approximately as the sum of four pyramids with bases of area  $S_{int}$ , and we can write

$$S_{int} = S(a_1, a_2, a_3, \frac{a_2}{l}), \quad V_{int} = \frac{4}{3}a_2S_{int}. \tag{8}$$

Here the function  $S(a_1, a_2, a_3, \xi)$  is defined in Equation (5). Finally, the volume  $V_h$  of the hard phase in the RVE may be evaluated from the equation

$$V_h = N_l V_{l-2a_2} + N_n V_{int}.$$

Here  $N_l$  is the number of the ligaments,  $N_n$  is the number of the nodes in the RVE,  $V_{l-2a_2}$  is the volume of the ligament without two zones of length  $a_2$  near its ends

$$V_{l-2a_2} = \pi a_2^2 l \frac{(a_1^2 - 2)}{15a_1^2} (1 - 2\lambda)[15 + 4a_3(2a_3 - 5) - 8a_3(1 - \lambda)\lambda (5 - 2a_1 - 6a_1\lambda(1 - \lambda))], \quad \lambda = \frac{a_2}{l}.$$

Thus, the geometrical structure of the open-cell foam inside a cubic region  $V$  is defined if the coordinates of all the nodes and connections between them are indicated as well as the parameters  $a_1, a_2, a_3$  of all the ligaments that perform such connections.

#### 4. The finite element method

The version of the finite element method used in this work for calculating elastic fields in the ligaments of open-cell foams is based on a special beam element. The geometry of a typical beam was described in the previous section, and its material is assumed to be elastic with a Young modulus  $E$  and Poisson ratio  $\nu$ . The Timoshenko beam model is used to describe the deformation of the beam element.

Let us consider a beam element of length  $l$ , and introduce a local Cartesian coordinate system  $(x, y, z)$  with the origin at the left end of the beam and the  $x$ -coordinate directed along the beam axis. The ends of the beam (nodes) are labeled with numbers 1 and 2. The vectors of the nodal displacements  $\{u_x^{(i)}, u_y^{(i)}, u_z^{(i)}\}$  and rotations  $\{\theta_x^{(i)}, \theta_y^{(i)}, \theta_z^{(i)}\}$  compose the generalized vector of displacements

$$\mathbf{d}^{(i)} = \{u_x^{(i)}, u_y^{(i)}, u_z^{(i)}, \theta_x^{(i)}, \theta_y^{(i)}, \theta_z^{(i)}\}^T$$

of the  $i$ th node ( $i = 1, 2$ ). The two vectors  $\{\mathbf{d}^{(1)}, \mathbf{d}^{(2)}\}^T$  completely define the deformation of the beam. Hence, the degrees of freedom per node are equal to 6 (3 displacements and 3 rotations). The vectors of the nodal forces  $\{f_x^{(i)}, f_y^{(i)}, f_z^{(i)}\}$  and moments  $\{m_x^{(i)}, m_y^{(i)}, m_z^{(i)}\}$  compose the vector of generalized force  $\mathbf{p}^{(i)} = \{f_x^{(i)}, f_y^{(i)}, f_z^{(i)}, m_x^{(i)}, m_y^{(i)}, m_z^{(i)}\}^T$ . Here  $\{\}^T$  is the transposed vector.

We derive the beam stiffness matrix by using the direct method of [Cook et al. \[1989\]](#). The stiffness matrix  $\mathbf{K}$  of the beam element is defined by the equation

$$\mathbf{K} \begin{Bmatrix} \mathbf{d}^{(1)} \\ \mathbf{d}^{(2)} \end{Bmatrix} = \begin{Bmatrix} \mathbf{p}^{(1)} \\ \mathbf{p}^{(2)} \end{Bmatrix}, \quad \mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}. \tag{9}$$

Here  $\mathbf{K}_{ij}$  are the block elements of the stiffness matrix that are constructed below. There are two vector equations of equilibrium connecting the nodal forces and moments:

$$\begin{aligned} \mathbf{p}^{(1)} + \mathbf{R}_{12}\mathbf{p}^{(2)} &= \{\mathbf{0}\}, \\ \mathbf{R}_{21}\mathbf{p}^{(1)} + \mathbf{p}^{(2)} &= \{\mathbf{0}\}, \end{aligned} \tag{10}$$

where matrices  $\mathbf{R}_{12}$  and  $\mathbf{R}_{21}$  have forms

$$\mathbf{R}_{12} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & L & 0 & 1 & 0 \\ 0 & -L & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{R}_{21} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -L & 0 & 1 & 0 \\ 0 & L & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{R}_{12} = \mathbf{R}_{21}^{-1}.$$

The strain energy  $U$  of the Timoshenko beam can be expressed as follows [\[Pilkey et al. 2003\]](#)

$$\begin{aligned} U = \frac{1}{2} & \left( \int_0^L \frac{N^2(x)}{ES(x)} dx + \int_0^L \frac{M_y^2(x)}{EJ(x)} dx + \int_0^L \frac{M_z^2(x)}{EJ(x)} dx \right) \\ & + \frac{1}{2} \left( \int_0^L \frac{M_x^2(x)}{\mu \widehat{J}_p(x)} dx + \int_0^L \frac{\alpha_s V_y^2(x)}{\mu S(x)} dx + \int_0^L \frac{\alpha_s V_z^2(x)}{\mu S(x)} dx \right), \end{aligned}$$

where  $S, J$  are the area and the main moment of inertia of the beam cross-section, respectively (see [Section 3](#)),  $\widehat{J}_p$  is the corrected polar moment of inertia that results from the solution of the torsion problem for a beam with a noncircular cross section

$$\widehat{J}_p = \pi a_2^4 (1 - 4a_1^{-2} + 2a_1^{-4}) (1 - 4a_3^2 \xi (1 - \xi))^4,$$

and  $\alpha_s$  is the shear deformation coefficient [Pilkey et al. 2003, p. 750].  $F_x$ ,  $F_y$ ,  $F_z$  are internal axial and shear forces,  $M_x$ ,  $M_y$ ,  $M_z$  are internal torque and bending moments.

Let us fix displacements at node 2 ( $\mathbf{d}^{(2)} = \mathbf{0}$ ) and apply the force  $\mathbf{p}^{(1)}$  to node 1. In this case, the internal forces in the beam are

$$\begin{aligned} F_x(x) &= -f_x^{(1)}, & F_y(x) &= -f_y^{(1)}, & F_z(x) &= -f_z^{(1)}, \\ M_x(x) &= -m_x^{(1)}, & M_y(x) &= -m_y^{(1)} - f_z^{(1)}x, & M_z(x) &= -m_z^{(1)} + f_y^{(1)}x. \end{aligned} \quad (11)$$

The Castigliano theorem ( $\frac{\partial U}{\partial \mathbf{p}} = \mathbf{d}$ ) together with equilibrium Equation (10) give us the equations

$$\mathbf{S}_{11}\mathbf{p}^{(1)} = \mathbf{d}^{(1)}, \quad \mathbf{R}_{21}\mathbf{p}^{(1)} + \mathbf{p}^{(2)} = \{\mathbf{0}\}, \quad (12)$$

where  $\mathbf{S}_{11}$  is the flexibility matrix

$$\mathbf{S}_{11} = \begin{bmatrix} s_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & s_{22} & 0 & 0 & 0 & -s_{26} \\ 0 & 0 & s_{33} & 0 & s_{35} & 0 \\ 0 & 0 & 0 & s_{44} & 0 & 0 \\ 0 & 0 & s_{35} & 0 & s_{55} & 0 \\ 0 & -s_{26} & 0 & 0 & 0 & s_{66} \end{bmatrix}. \quad (13)$$

The components  $s_{ij}$  of this matrix are defined by the equations

$$\begin{aligned} s_{11} &= \frac{L}{E} \int_0^1 \frac{1}{S(\xi)} d\xi, & s_{22} &= s_{33} = \frac{L^3}{E} \int_0^1 \frac{\xi^2}{J(\xi)} d\xi + \frac{\alpha_s L}{\mu} \int_0^1 \frac{1}{S(\xi)} d\xi, \\ s_{44} &= \frac{L}{\mu} \int_0^1 \frac{1}{\widehat{J}_p(\xi)} d\xi, & s_{55} &= s_{66} = \frac{L}{E} \int_0^1 \frac{1}{J(\xi)} d\xi, \\ s_{35} &= \frac{L^2}{E} \int_0^1 \frac{\xi}{J(\xi)} d\xi, & s_{26} &= \frac{L^2}{E} \int_0^1 \frac{\xi}{J(\xi)} d\xi, \quad \xi = \frac{x}{L}. \end{aligned} \quad (14)$$

Next, we fix displacements at node 1 ( $\mathbf{d}^{(1)} = \mathbf{0}$ ) and apply the force  $\mathbf{p}^{(2)}$  to node 2. In this case, the internal forces in the beam are

$$\begin{aligned} N(x) &= f_x^{(2)}, & V_y(x) &= f_y^{(2)}, & V_z(x) &= f_z^{(2)}, \\ M_x(x) &= m_x^{(2)}, & M_y(x) &= m_y^{(2)} - f_z^{(2)}(L-x), & M_z(x) &= m_z^{(2)} + f_y^{(2)}(L-x). \end{aligned} \quad (15)$$

The Castigliano theorem with the equilibrium Equation (10) give

$$\mathbf{S}_{22}\mathbf{p}^{(2)} = \mathbf{d}^{(2)}, \quad \mathbf{p}^{(1)} + \mathbf{R}_{12}\mathbf{p}^{(2)} = \{\mathbf{0}\}, \quad (16)$$

where  $\mathbf{S}_{22}$  is the flexibility matrix

$$\mathbf{S}_{22} = \begin{bmatrix} s_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & s_{22} & 0 & 0 & 0 & \tilde{s}_{26} \\ 0 & 0 & s_{33} & 0 & -\tilde{s}_{35} & 0 \\ 0 & 0 & 0 & s_{44} & 0 & 0 \\ 0 & 0 & -\tilde{s}_{35} & 0 & s_{55} & 0 \\ 0 & \tilde{s}_{26} & 0 & 0 & 0 & s_{66} \end{bmatrix}, \tag{17}$$

$$\tilde{s}_{35} = \tilde{s}_{26} = \int_0^L \frac{L-x}{EJ(x)} dx = Ls_{55} - s_{35}.$$

Finally, the blocks of stiffness matrix  $\mathbf{K}$  in Equation (9) are defined by the equations

$$\begin{aligned} \mathbf{K}_{11} &= \mathbf{S}_{11}^{-1}, & \mathbf{K}_{21} &= -\mathbf{R}_{21}\mathbf{K}_{11}, \\ \mathbf{K}_{22} &= \mathbf{S}_{22}^{-1}, & \mathbf{K}_{12} &= -\mathbf{R}_{12}\mathbf{K}_{22}, \end{aligned} \tag{18}$$

$$\mathbf{K}_{11} = \begin{bmatrix} k_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & k_{22} & 0 & 0 & 0 & k_{26} \\ 0 & 0 & k_{33} & 0 & -k_{35} & 0 \\ 0 & 0 & 0 & k_{44} & 0 & 0 \\ 0 & 0 & -k_{35} & 0 & k_{55} & 0 \\ 0 & k_{26} & 0 & 0 & 0 & k_{66} \end{bmatrix}, \quad \mathbf{K}_{12} = \begin{bmatrix} -k_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & -k_{22} & 0 & 0 & 0 & k_{26} \\ 0 & 0 & -k_{33} & 0 & -k_{35} & 0 \\ 0 & 0 & 0 & -k_{44} & 0 & 0 \\ 0 & 0 & k_{35} & 0 & d_{55} & 0 \\ 0 & -k_{26} & 0 & 0 & 0 & d_{66} \end{bmatrix},$$

$$\mathbf{K}_{21} = \mathbf{K}_{12}^T, \quad \mathbf{K}_{22} = \begin{bmatrix} k_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & k_{22} & 0 & 0 & 0 & -\tilde{k}_{26} \\ 0 & 0 & k_{33} & 0 & \tilde{k}_{35} & 0 \\ 0 & 0 & 0 & k_{44} & 0 & 0 \\ 0 & 0 & \tilde{k}_{35} & 0 & k_{55} & 0 \\ 0 & -\tilde{k}_{26} & 0 & 0 & 0 & k_{66} \end{bmatrix},$$

$$\begin{aligned} k_{11} &= s_{11}^{-1}, & k_{22} &= k_{33} = d_z s_{55}, \\ k_{44} &= s_{44}^{-1}, & k_{55} &= k_{66} = d_y s_{22}, \\ k_{26} &= k_{35} = d s_{26}, & \tilde{k}_{26} &= \tilde{k}_{35} = d \tilde{s}_{26}, \\ k_{55} &= d(Ls_{35} - s_{33}), & k_{66} &= d(Ls_{26} - s_{22}), \end{aligned} \tag{19}$$

$$d = (s_{22}s_{66} - s_{26}^2)^{-1}.$$

By the calculation of the elastic energy of the open-cell structures, there appears a problem of accounting for the elastic energy of the regions of the beam intersections (nodes). The elastic energy of the node regions increases together with the volume concentration of the hard phase of the foams. In this study, we suppose that the nodal force vector  $\mathbf{p}$  corresponding to the considered beam does not vary in the node region, and coincides with its value at the point of the beam connections. Let the node have the

coordinate  $x = l$  in the local coordinate system, and the characteristic size of the node region be  $dl$ . The components of the strain tensor in this region are calculated as follows

$$\varepsilon_{xx} = \frac{\partial u_x}{\partial x} = \frac{N(l)}{ES(l)}, \quad \gamma_{xy} = \alpha_s \frac{F_y(l)}{\mu S(l)}, \quad \gamma_{xz} = \alpha_s \frac{F_z(l)}{\mu S(l)}, \quad (20)$$

and the parameters of rotation angle altering are

$$\frac{d\theta_x}{dx} = \frac{M_x(l)}{\mu \widehat{J}_p(l)}, \quad \frac{d\theta_z}{dx} = \frac{M_y(l)}{EJ(l)}, \quad \frac{d\theta_y}{dx} = \frac{M_z(l)}{EJ(l)}. \quad (21)$$

Using these equations one can determine the corresponding part of the energy by integrating [Equation \(20\)](#) and [Equation \(21\)](#) along the  $dl$  element. If the ligament is very short ( $L < 2dl$ ), it is considered to be a beam of a constant cross-section. Note that the accepted geometrical model of the ligament ([Section 2](#)) allows us to calculate all the integrals in the above equations in closed analytical forms.

If the stiffness matrices of all the beam elements are constructed, the final system of the EFM may be obtained by the displacement method [[Cook et al. 1989](#)]. This system follows from the conditions that the displacement vectors of the beam ends connected at one node are the same for all these beams.

## 5. Representative volume element of the open-cell foam materials

Let us go to the calculation of the effective elastic properties of the foam material. The procedure of simulation of the foam microstructure inside a cubic region,  $V$  described in [Section 2](#), give us the skeleton of the foam microstructure inside the RVE: the coordinates of the nodes and the rule of their connections by the ligaments. The parameters of the ligaments mentioned in [Section 3](#) are the other part of information necessary for performance of the EFM calculations. The version of the EFM considered in the previous section is used to calculate displacements, angles of rotations, forces and moments at all the nodes of the beam structure by the prescribed boundary conditions on the surface of the RVE (cube  $V$ ). Note that the ligaments that are placed on the surface of cube  $V$  should be deleted from the skeleton in order to obtain reliable values of the effective elastic constants of the foam material. These ligaments appear as a result of the mirror reflections of the seed point inside  $V$  with respect to the sides of the cube (see [Section 2](#)). They don't correspond to the actual foam structure and provoke excess of rigidity of the considered RVE.

In the nodes that are on the surface  $\Omega$  of the cube  $V$  one has to define static and kinematic conditions that are necessary for uniqueness of the solution of the elasticity problem. Let us define the components of the displacement vector  $u_i^{(k)}$  at the surface nodes  $x^{(k)}$  ( $x^{(k)} \subset \Omega$ ) by the following equation (affine deformation)

$$u_i^{(k)} = \varepsilon_{ij} x_j^{(k)}, \quad (22)$$

where  $\varepsilon_{ij}$  is a fixed symmetric tensor. All angles of rotations of the surface nodes are assumed to be equal to zero. These boundary conditions are sufficient in order to obtain displacements, forces and moments at every node of the beam structure including the border nodes.

Let us go to the homogenization problem that is the determination of a homogeneous elastic material equivalent to the given foam material. It means that the elastic module tensor  $C^*$  of the equivalent material should coincide with the tensor that connects the mean values of the stress  $\langle \sigma_{ij} \rangle$  and strain  $\langle \varepsilon_{ij} \rangle$

tensors over the RVE of the foam material

$$\langle \sigma_{ij} \rangle = C_{ijkl}^* \langle \varepsilon_{kl} \rangle, \tag{23}$$

$$\langle \sigma_{ij} \rangle = \frac{1}{V} \int_V \sigma_{ij}(x) dv, \quad \langle \varepsilon_{ij} \rangle = \frac{1}{V} \int_V \varepsilon_{ij}(x) dv. \tag{24}$$

Let us consider a volume of the equivalent homogeneous material that coincides with the RVE and is loaded with the forces  $f_j(x) = n_k(x)\sigma_{kj}(x)$  on its surface  $\Omega$ . Here  $\sigma_{kj}(x)$  is the stress tensor,  $n_i$  is the external normal to  $\Omega$ . The surface integral

$$\int_{\Omega} f_j(x)x_i d\Omega$$

may be transformed in a volume integral using the Gauss theorem as follows

$$\int_{\Omega} f_j(x)x_i d\Omega = \int_{\Omega} n_k(x)\sigma_{kj}(x)x_i d\Omega = \int_V \partial_k [\sigma_{kj}(x)x_i] dv = \int_V [\partial_k \sigma_{kj}(x)] x_i dv + \int_V \sigma_{ij}(x) dv.$$

Because of the equilibrium equation for the stress tensor  $\sigma_{kj}(x)$  ( $\partial_k \sigma_{kj}(x) = 0$ ), the first integral in the right hand side of this equation disappears, and for the mean stress field  $\langle \sigma_{ij} \rangle$  over the cube  $V$  we obtain

$$\langle \sigma_{ij} \rangle = \frac{1}{V} \int_V \sigma_{ij}(x) dv = \frac{1}{V} \int_{\Omega} f_j(x)x_i d\Omega. \tag{25}$$

It follows from this equation that in the case of the beam structure, the mean stress tensor inside the cubic RVE may be calculated as follows

$$\langle \sigma_{ij} \rangle = \frac{1}{8} \sum_{x^{(k)} \subset \Omega} F_j^{(k)} x_i^{(k)}, \tag{26}$$

where  $F_j^{(k)}$  is the vector of the concentrated force acting in the surface node  $x^{(k)}$ . It is taken into account that the volume of the cube  $V$  is equal to 8.

For the affine surface deformation Equation (22), the mean strain tensor  $\langle \varepsilon_{ij} \rangle$  defined in Equation (24) coincides with the tensor  $\varepsilon_{ij}$  presented in boundary conditions Equation (22). Thus, using (23) and (26) one can calculate the components of the tensor  $C_{ijkl}^*$  (tensor of the effective elastic modules) if the forces  $F_j^{(k)}$  in the surface nodes are obtained from the solution of the elasticity problem for the considered beam structure.

Application of other boundary conditions on the surface of the cube  $V$  faces additional computational difficulties. For instance, one can define the forces  $F_j^{(k)}$  and the moments at all the surface nodes and calculate the mean strain tensor from Equation (26). For a homogeneous material, the mean strain field defined in Equation (24) may be transformed in a surface integral

$$\langle \varepsilon_{ij} \rangle = \frac{1}{V} \int_V \varepsilon_{ij}(x) dv = \frac{1}{V} \int_{\Omega} n_i(x)u_j(x) d\Omega. \tag{27}$$

Here  $\varepsilon_{ij}(x) = \partial_{(i} u_{j)}(x)$ ,  $u_i(x)$  is the displacement vector, parentheses in indices mean symmetrization. Thus, for calculation of the mean strain field one has to calculate the integral in the right hand side of Equation (27) from the solution of the elasticity problem for the beam structure. Note that the FEM provides the values of displacements  $u_j$  only in a finite number of the surface nodes. Thus, in order to find the mean strain field  $\langle \varepsilon_{ij} \rangle$  from Equation (27), one has to interpolate function  $u(x)$  on all the points of surface  $\Omega$ , and after that, to calculate the surface integral in the right hand side of Equation (27) numerically. Such interpolation and integration are sources of additional numerical errors that cannot be avoided if the force boundary conditions on  $\Omega$  are used.

Very often in the literature, for the numerical solution of the homogenization problems, periodic boundary conditions on the surface of the RVE are applied. Note that for the cubic volume element discussed in Section 2, such conditions cannot be imposed. In the framework of the beam FEM, the periodic conditions are to be formulated in a finite number of surface nodes. But the position of the nodes on the opposite sides of cube  $V$  are not symmetric, and strictly speaking, the periodic boundary conditions cannot be formulated. On the other hand, one can consider a tessellation process using not mirror but periodic continuation of the seed points  $S_{(0,0,0)}$  inside the cube  $V$  on all three-dimensional space. In this case, the Voronoi or Laguerre polyhedra corresponded to the original seed point set  $S_{(0,0,0)}$  inside  $V$  will compose not a cubic region, but a region with a rather complex, nonplane surface. The calculation of the mean strain field over such a region using Equation (27) faces an additional difficulty of definition of the external normal  $n(x)$  at all the points of such a surface. The latter contains many angle points where normals are not defined. Note that the displacement vectors are calculated at the nodes that are the vertices of the polyhedra. Thus, for the calculation of the mean strain field over the region  $V$ , one has to interpolate the displacement field onto all points of a nonplane surface  $\Omega$ , and then to calculate the integral in the right hand side of Equation (27) numerically. Both operations are connected with some numerical errors. That is why in this study, only kinematic boundary conditions (22) are considered.

Let us go to the problem of definition of the number  $N$  of cells inside the RVE that are sufficient to obtain reliable values of the effective elastic constants of the foam material. In the series of numerical experiments, the Voronoi tessellation procedure was used for generation of the foam microstructures. The seed points were independently and homogeneously distributed inside the cube by the condition that these points cannot be closer than a distance  $h$  from each other. The distance  $h$  was chosen in order to generate the set of seed points in a reasonable time. Circular, cylindrical, ligaments with parameters:  $a_1 = 100$ ,  $a_3 = 0$  were considered; parameter  $a_2$  depends on the volume concentration of the hard phase. The material of the ligaments was taken to be isotropic with the Young modulus  $E$  and the Poisson ratio  $\nu = 0.3$ .

We have considered the increasing number of the cells from one hundred to fifteen hundreds. For the calculation of the effective elastic moduli, the following six different kinematic boundary conditions were used. Extension of the cube in the direction of  $x_1$ ,  $x_2$  or  $x_3$ -axes ( $\varepsilon_{ij}^{(1)} = \delta_{1i}\delta_{1j}$ ,  $\varepsilon_{ij}^{(2)} = \delta_{2i}\delta_{2j}$ ,  $\varepsilon_{ij}^{(3)} = \delta_{3i}\delta_{3j}$ ), and three independent shear deformations on the cube surface:

$$\varepsilon_{ij}^{(4)} = \delta_{1(i}\delta_{2j)}, \quad \varepsilon_{ij}^{(5)} = \delta_{1(i}\delta_{3j)}, \quad \varepsilon_{ij}^{(6)} = \delta_{2(i}\delta_{3j)}.$$

We calculate the effective elastic Young and shear moduli in the directions of the coordinate axes for the increasing number of the cells inside  $V$ . We also evaluate anisotropy of the effective moduli with respect to tension and shear deformations. Because for an isotropic material, the Young  $E_*$  and shear  $\mu_*$  moduli

are connected by the equation

$$\mu_* = \frac{E_*}{2(1 + \nu_*)}, \quad (28)$$

one can introduce anisotropy parameter  $\alpha$  defined by the equation

$$\alpha = \frac{2(1 + \nu_*)\mu_*}{E_*}, \quad (29)$$

and the closer to 1 the value of  $\alpha$ , the closer to isotropy the symmetry of the tensor of the effective elastic moduli  $C^*$  of the cubic RVE.

Another parameter  $\beta$

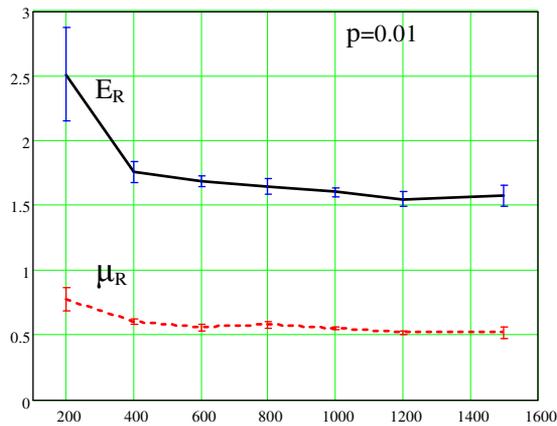
$$\beta = \frac{1}{\langle E_* \rangle} \sqrt{\langle (E_* - \langle E_* \rangle)^2 \rangle} \quad (30)$$

characterizes the dispersion of the effective Young modulus over the realizations of the foam structures. Here the average is taken over three directions and over the realizations of the random microstructures for a fixed number of cells inside the RVE.

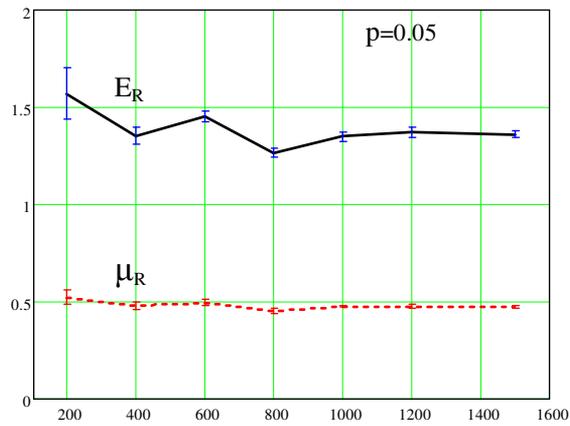
In Figures 10, 11, and 12, the dependences of the reduced effective Young  $E_R$  and shear  $\mu_R$  moduli of the RVE on the number  $N$  of cells inside the RVE are presented. The reduced moduli are defined by the equations

$$E_R = \frac{E_*}{p^2 E}, \quad \mu_R = \frac{\mu_*}{p^2 \mu}, \quad p = \frac{\rho_*}{\rho}. \quad (31)$$

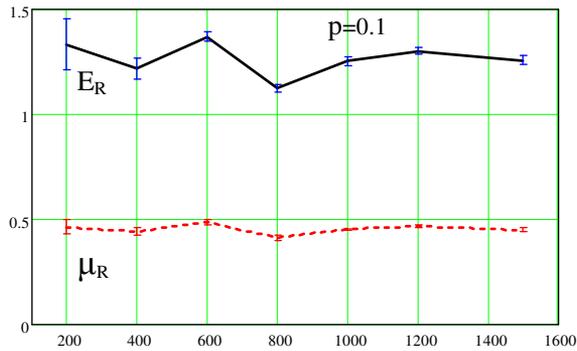
Here  $E$ ,  $\mu$ ,  $\rho$  are the Young modulus, shear modulus, and density of the foam hard phase,  $\rho_*$  is the density of the foam,  $p$  is the volume concentration of the hard phase. The graphs in Figure 10 correspond to the volume concentration of the hard phase  $p = 0.01$ , in Figure 11 to  $p = 0.05$ , and in Figure 12 to  $p = 0.1$ . The same dependences for the Poisson ratio  $\nu_*$  are presented in Figure 13. Dependences of the anisotropy coefficient  $\alpha$  and dispersion coefficient  $\beta$  on the number  $N$  of cells in the RVE are in Figures 14 and 15 for  $p = 0.01$ . The vertical bars in these figures show the dispersions of the numerical results



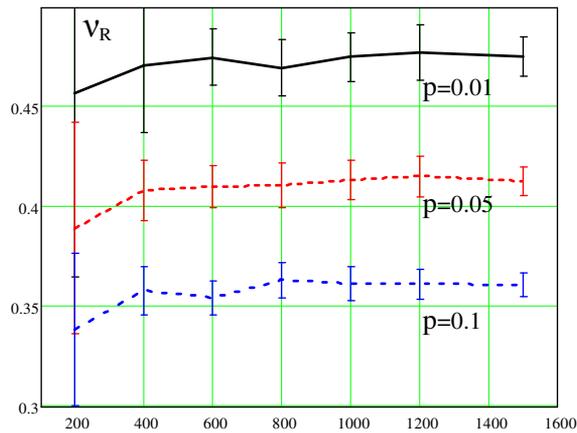
**Figure 10.** The dependences of the reduced Young  $E_R$  and shear  $\mu_R$  moduli (Equation (31)) on the number  $N$  of cells in the RVE for the volume concentration of the hard phase  $p = 0.01$ .



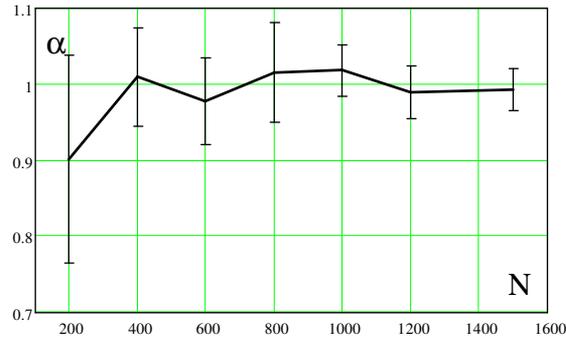
**Figure 11.** The same graphs as in Figure 10 for  $p = 0.05$ .



**Figure 12.** The same graphs as in Figure 10 for  $p = 0.1$ .



**Figure 13.** The dependence of the Poisson ratio  $\nu_*$  of the foam on the number  $N$  of cells in the RVE.

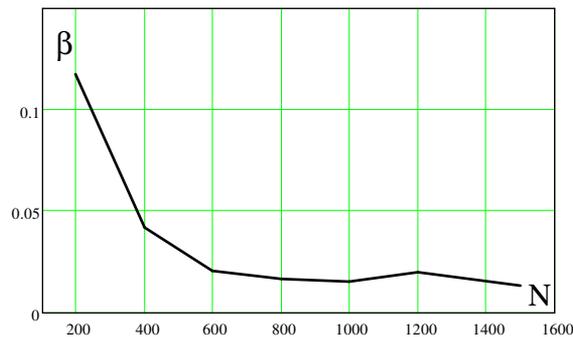


**Figure 14.** The dependence of the isotropy parameter  $\alpha$  (Equation (29)) on the number  $N$  of cells in the RVE,  $p = 0.01$ .

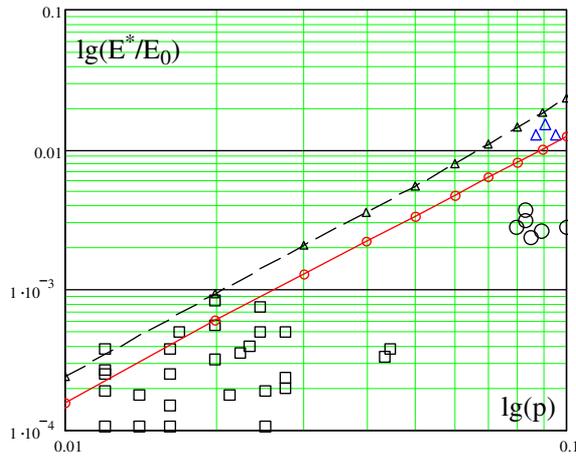
among the realizations of the microstructures with fixed values of the cell number  $N$ . For every value of  $N$ , 5–7 realizations of the random structures were taken.

The main conclusion that can be made up from these graphs is that the RVE should contain about 900–1000 cells in order to obtain reliable values of the effective elastic properties of the open-cell foam materials. For such a number of cells in the RVE ( $N = 1000$ ), the dependences of the relative effective Young modulus  $E_*/E$  of the foams on the volume concentration  $p$  of the hard phase are presented in Figure 16. Square points in this figure are experimental data of Gibson and Ashby [1982], triangle points are the data of Liderman [1971], and circles are the data of Hagiwara and Green [1987]. The line with black dots is the result of our simulations for the foam with circular cylindrical ligaments ( $a_1^{-1} = 0$ ), and the line with black triangles corresponds to the foams with triangle cylindrical ligaments ( $a_1^{-1} = 0.5$ ). Note that the detailed information about the ligament shapes is absent in the mentioned experimental works.

If the number of cells inside the RVE is taken smaller than indicated above, dispersion of the numerical values of the effective elastic constants for different realizations of the foam microstructures increases (Figure 15). It is necessary also to emphasize that the mean value of these constants over the realizations

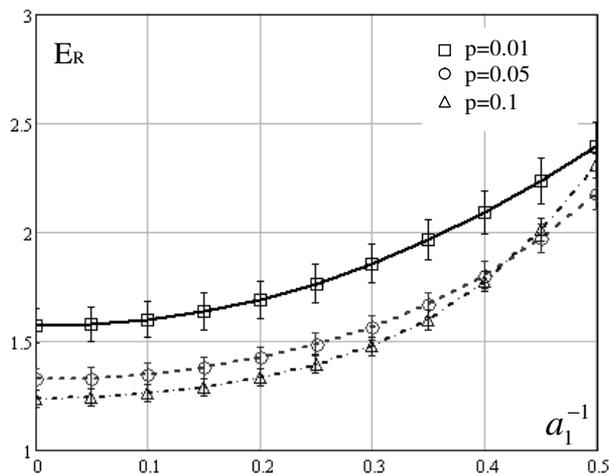


**Figure 15.** The dependence of the dispersion coefficient  $\beta$  of the effective Young modulus (Equation (30)) on the number  $N$  of cells in the RVE,  $p = 0.01$ .

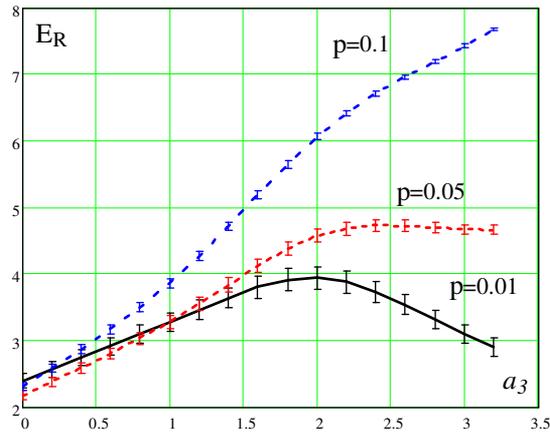


**Figure 16.** The dependences of the relative effective Young modulus ( $E_*/E$ ) on the volume concentration  $p$  of the hard phase of the foam. Triangle points correspond to the data from [Liderman 1971]; square points from [Gibson and Ashby 1982]; circle points from [Hagiwara and Green 1987]; the line with dots are theoretical predictions for the foams with circular cylindrical ligaments ( $a_1^{-1} = 0, a_3 = 0$ ), the line with triangles is the prediction for triangular cylindrical ligaments ( $a_1^{-1} = 0.5, a_3 = 0$ ).

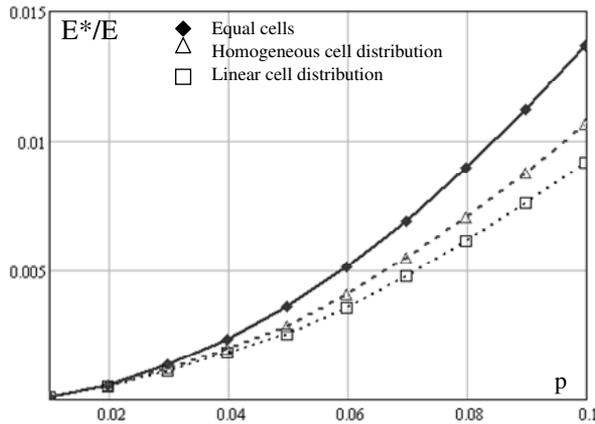
does not coincide with the mean values of the constants for the RVE with a sufficiently large number of cells (Figure 16). The same fact was noted in [Kanit et al. 2003], where the problem of the appropriate size of the RVE for random polycrystalline materials was considered.



**Figure 17.** The dependence of the reduced effective Young modulus  $E_R$  on the shape of the cross-section (parameter  $a_1$ ) for cylindrical ligaments  $a_3 = 0$ .



**Figure 18.** The dependence of the reduced effective Young modulus  $E_R$  on the ligament axial altering (parameter  $a_3$ ) for triangle ligaments ( $a_1^{-1} = 0.5$ ).



**Figure 19.** The dependences of the relative effective Young modulus ( $E_*/E$ ) on the law of distribution of cell diameters and the volume concentration  $p$  of the hard phase for the foams with circular cylindrical ligaments ( $a_1^{-1} = 0, a_3 = 0$ ). The line with black rhombs correspond to approximately equivalent cell diameters, the line with triangles to a homogeneous distribution of cell diameters, and the line with squares to a linear distribution of the diameters.

### 6. Dependences of elastic properties of open-cell foams on the ligament shapes and the law of the cell size distribution

The effective Young moduli of the foams with ligaments of various shapes were calculated according to the proposed algorithm. The results of such calculations are presented in Figures 17 and 18. As it was indicated in Section 2, the parameter  $a_1$  ( $0 < a_1^{-1} < 0.5$ ) describes the shape of the cross-sections of the ligament, and parameter  $a_3$  defines the altering of the ligament cross-section along its axis ( $0 < a_3 < 4$ ,

$a_3 = 0$  corresponds to a cylindrical ligament). As it is seen from these graphs, the reduced Young modulus  $E_R$  is more sensitive to the parameter  $a_3$  altering than to altering the parameter  $a_1$ .

The dependences of relative Young modulus  $E_*/E$  of the foams on the distribution of cell diameters inside the RVE and on the volume concentration  $p$  of the hard phase are presented in Figure 19. In this figure, the line with black rhombs corresponds to the foams with almost equal diameters of the cells (the distribution function of the cell diameters is presented in Figure 6 also by the line with black rhombs), the line with triangles corresponds to the homogeneous distribution of cell diameters in the interval  $(0.5 \langle d \rangle, 1.5 \langle d \rangle)$ , and the line with squares correspond to the linear distribution of the cell diameters in the interval  $(0.3 \langle d \rangle, 1.2 \langle d \rangle)$  (the corresponding distribution functions are in Figure 6). As it is seen from these graphs, for a fixed volume concentration of the hard phase, the foams with a wide distribution of cell diameters have lower elastic modules than the foams with approximately the same diameters of cells.

## 7. Conclusion

This work proposes a method of calculating the effective elastic properties of open-cell foam materials in two stages: simulation of the microstructure of the foam inside a RVE of such a material, and application of a specific version of the FEM for the calculation of stresses and strains in the ligaments of the foam structure. The Laguerre tessellation algorithm adapted in this work allows us to simulate the foam microstructures with any prescribed cell size distribution law. But this algorithm is more complex than the conventional Voronoi algorithm. It requires carrying out the procedure of close packing of spheres with the given distribution of the diameters inside the RVE. The coordinates of the centers of the spheres obtained after the packing, and the diameters of the spheres associated with every center, are initial data for the Laguerre tessellation process.

The size of the RVE is a crucial problem for the numerical simulation of the properties of the open-cell foams. As is shown in this work, the size of the RVE, or the minimal number  $N$  of the cells inside the RVE that is necessary to obtain reliable values of the effective elastic constants, depends on the volume concentration  $p$  of the hard phase. For small volume concentrations ( $p = 0.01$ ) this number turns out to be about 900–1000, and the value of  $N$  decreases when  $p$  increases:  $N$  is about 800 for  $p = 0.05$ , and about 400 for  $p = 0.1$ .

The influence of the form of the ligaments on the effective elastic properties is essential, but altering the ligament cross sections along the ligament axes affects the values of the effective elastic moduli more strongly than altering the shape of the cross sections.

The influence of the cell diameter distributions on the effective elastic properties of the foams is also notable. For the foams with a wide distribution of cell diameters, the effective Young moduli turn out to be less than the modules of the foams with approximately equal cells, and the difference grows with the volume concentration  $p$  of the hard phase. It was also indicated that for the same volume concentration of the hard phase, the foams with a linear distribution of cell diameters have lower Young moduli than the foams with homogeneous distribution of cell diameters.

## Acknowledgement

The authors thank Dr. Ozden Ochoa, Dr. Khalid Lafdi, and Dr. Dominique Jeulin for fruitful discussions.

## References

- [Aurenhammer and Klein 2000] F. Aurenhammer and R. Klein, “Voronoi diagrams”, pp. 201–290 in *Handbook of computational geometry*, edited by J.-R. Sack and J. Urrutia, Elsevier, Amsterdam, 2000.
- [Christensen 2000] R. M. Christensen, “Mechanics of cellular and other low density materials”, *Int. J. Solids Struct.* **37**:1-2 (2000), 93–104.
- [Cook et al. 1989] R. D. Cook, D. S. Malkus, and M. E. Plesha, *Concepts and applications of finite element analysis*, 3rd ed., Wiley, New York, 1989.
- [Gibson and Ashby 1982] L. J. Gibson and M. F. Ashby, “The mechanics of three dimensional cellular materials”, *P. Roy. Soc. Lon. A Mat.* **382**:1782 (1982), 43–59.
- [Gong et al. 2005] L. Gong, K. S., and J. W.-Y., “Compressive response of open-cell foams, I: Morphology and elastic properties”, *Int. J. Solids Struct.* **42**:5-6 (2005), 1355–1378.
- [Hagiwara and Green 1987] H. Hagiwara and D. J. Green, “Elastic behavior of open-cell alumina”, *J. Am. Ceram. Soc.* **70**:11 (1987), 811–815.
- [Kadashevich and Stoyan 2005] I. Kadashevich and D. Stoyan, “Micro-mechanical analysis of ACC”, pp. 219–228 in *Autoclaved aerated concrete*, Taylor and Francis, London, 2005.
- [Kanit et al. 2003] T. Kanit, S. Forest, I. Galliet, V. Mounoury, and D. Jeulin, “Determination of the size of the representative volume element for random composites: statistical and numerical approach”, *Int. J. Solids Struct.* **40**:13-14 (2003), 3647–3679.
- [Liderman 1971] J. M. Liderman, “The prediction of the tensile properties of flexible foams”, *J. Appl. Polym. Sci.* **15**:3 (1971), 693–703.
- [Pilkey et al. 2003] W. D. Pilkey, , and W. Wunderlich, *Mechanics of structures, variational and computational methods*, 2nd ed., CRC Press, Boca Raton, FL, 2003.
- [Roberts and Garboczi 1999] A. P. Roberts and E. J. Garboczi, “Elastic properties of a tungsten-silver composite by reconstruction and compilation”, *J. Mech. Phys. Solids* **47**:10 (1999), 2029–2055.
- [Roberts and Garboczi 2001] A. P. Roberts and E. J. Garboczi, “Elastic properties of model random three-dimensional closed-cell cellular solids”, *Acta Mater.* **49**:2 (2001), 189–197.
- [Roberts and Garboczi 2002] A. P. Roberts and E. J. Garboczi, “Elastic properties of model random three-dimensional open-cell solids”, *J. Mech. Phys. Solids* **50**:1 (2002), 33–55.
- [Tenemura et al. 1983] M. Tenemura, T. Ogawa, and N. Ogita, “A new algorithm for three-dimensional voronoi tessellation”, *J. Comput. Phys.* **51**:2 (1983), 191–207.
- [Warren and Kraynik 1997] W. E. Warren and A. M. Kraynik, “Linear behavior of a low-density kelvin foam with open cells”, *J. Appl. Mech. (Trans. ASME)* **64** (1997), 787–793.
- [Zhu et al. 2000] H. X. Zhu, J. R. Hobdell, and A. H. Windle, “Effect of cell irregularity on the elastic properties of open-cell foams”, *Acta Mater.* **48**:20 (2000), 4893–4900.

Received 19 Jun 2006. Accepted 8 May 2007.

SERGEY KANAUN: [kanaoun@itesm.mx](mailto:kanaoun@itesm.mx)

*Departamento de Ingeniería Mecánica, Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Estado de México, Carretera Lago de Guadalupe 4 km Atizapan, Edo de México, 52926, Mexico*

OLEKSANDR TKACHENKO: [oleksandr.tkachenko@itesm.mx](mailto:oleksandr.tkachenko@itesm.mx)

*Departamento de Ingeniería Mecánica, Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Estado de México, Carretera Lago de Guadalupe 4 km Atizapan, Edo de México, 52926, Mexico*

## ELASTIC WANNIER–STARK LADDERS IN TORSIONAL WAVES

GUILLERMO MONSIVAIS, RAFAEL A. MÉNDEZ-SÁNCHEZ, ALFREDO DÍAZ DE ANDA,  
JORGE FLORES, LUIS GUTIÉRREZ AND ALEJANDRO MORALES

We study the normal modes of torsional waves in an elastic beam consisting of a set of  $N$  cuboids of varying heights. We present experimental, theoretical, and numerical results. We show that some analogies to the Wannier–Stark ladders resonances, originally introduced by Wannier in 1962, are exhibited by this classical system. The original ladders studied by Wannier consist of a series of equidistant energy levels for the electrons in a crystal in the presence of a static external electric field with the nearest-neighbor level spacing proportional to the intensity of the external field. For the case of torsional waves in the beam we have observed a similar behavior, namely, the vibrations of the beam show resonances of equidistant frequencies with the nearest-neighbor spacing proportional to parameter  $\gamma$  associated with the geometry of the beam analogously to the electric field. However, this analogy is not perfect; we address the origin of the differences.

### 1. Introduction

Since the discovery of the wavelike behavior of particles whose size is on the order of atomic dimensions or smaller, several analogies between quantum systems, that is, systems whose dynamics is governed by quantum mechanics, and classical systems have been observed. This is particularly true when the undulatory properties of the particles are important and interference phenomena play the relevant role. Thus, one frequently finds analogies in optics, electromagnetism, acoustics, elasticity and the like. It is in the case of optics where more analogies have been studied; see [Monsivais et al. 1990; Sheng 1995; Joannopoulos et al. 1995; Soukoulis 1996; de Sterke et al. 1998; Sapienza et al. 2003; Agarwal et al. 2004] and references therein. In some cases the analogies are not exact, and new and interesting characteristic effects appear for each field. In other cases the classical systems that potentially can present analogies, do not exist in a natural way, but can be built from an appropriate combination of other systems.

In this paper we study the analogy of the quantum mechanical phenomenon known as Wannier–Stark Ladder Resonances (WSLR) in a special type of classical elastic system. The study includes experimental, theoretical and numerical results. The existence of WSLR and their associated Bloch Oscillations (BO) in quantum mechanics have been controversial for many years, but by now some of their properties seem to be theoretically on firm ground. The BO were predicted by Bloch [1928]; see also [Zener 1934; James 1949; Wannier 1955]. They consist of a counterintuitive behavior of electrons in a crystal, which is under the action of a static external electric field. According to Bloch, this static field produces an oscillatory movement of the electrons inside the crystal. This strange prediction was a controversial matter for more than 60 years [Hart and Emin 1988]. The controversy waxed due to Wannier’s 1962 discovery [Wannier

---

*Keywords:* Wannier–Stark ladders, elastic waves, EMAT.

This work has been supported by UNAM DGAPA-PAPIIT Projects IN111307 and IN104102.

1962; 1969; Zak 1968; 1969]. This discovery establishes that the electronic energy spectrum consists of a series of energies  $E_1, E_2, \dots$  such that the nearest-neighbor level spacing is constant and proportional to the intensity of the external field. This set of energies forms the so called Wannier–Stark ladder. These amazing characteristics contrast with what occurs in a nonelectrified crystal, where the electrons travel through the whole periodic structure (Bloch waves) and where the constructive and destructive interferences give rise to an energy band spectrum, which consists of bands of allowed energy and regions where no values of the energy are permissible, forbidden bands or gaps [Brillouin 1946]. Thus, according to Bloch and Wannier, when the electric field modifies the periodic potential, the band structure is destroyed and states are localized.

These predictions were very important since the band structure is the basis of electronic devices. However, when Bloch and Wannier made their predictions it was impossible to test them experimentally. On the one hand, the BO are difficult to observe because the electrons lose their coherence in times shorter than the expected period of the oscillations. On the other hand, the Wannier–Stark ladders are difficult to observe because the width of the levels is larger than the separation between levels. Obviously both effects are related since a short lifetime of a state implies a wider associated energy level. The first confirmation of the Bloch–Wannier model came from the observations of the WSLR that appeared first in numerical experiments [Rabinovitch 1977; Banavar and Coon 1978], and thereafter from the laboratory [Méndez et al. 1988]. This was around 20 years after the prediction of Wannier. Later on, in 1992, the BO were also observed [Feldmann et al. 1992; Leo et al. 1992; Dekorsy et al. 1994; Lyssenko et al. 1998]. This occurred when the semiconductor superlattices were built [Esaki and Tsu 1970], since in these systems the period of BO is shorter. Actually, there exists considerable literature on the WSLR and the BO, and now it is recognized that the original ideas of Bloch and Wannier are essentially correct. The BO are due to the fact that when an electron in the crystal is accelerated by the electric field, its velocity is increased until it reaches the end of the Brillouin zone, where it is dispersed and its velocity decreases. This effect continues until the velocity is equal to zero and changes sign returning to the original position inside the crystal. Then, the electron is again accelerated by the field and the cycle is repeated. Under these circumstances the wave functions are localized in the zones where the oscillatory movement occurs. We should mention, however, that there always exists a probability that the electron tunnels to other regions of the crystal (Zener tunneling). Whenever this probability is small, the BO can be present.

The origin of the WSLR can be understood as follows. Consider the time independent Schrödinger's equation for an electron of charge  $e$  in a one-dimensional crystal in the presence of a static external electric field  $F$ ,

$$-\frac{1}{2} \frac{d^2 \psi}{dx^2} + V(x) \psi = E \psi, \quad (1)$$

where we are using a system of units in which the electron mass and Plank's constant are set to 1. The function  $V(x)$  is the potential acting on the electron,  $E$  is the energy of the electron and  $e|\psi|^2$  is the charge density. In our case  $V(x)$  has the form  $V(x) = V_p(x) + eFx$ , where  $V_p(x)$  is the periodic potential due to the atoms of the crystal, that is  $V_p(x) = V_p(x + np)$ ,  $p$  being the period and  $n$  an arbitrary integer. When  $F = 0$ , the potential is periodic and the system will show a band structure [Brillouin 1946]. However, when  $F \neq 0$ , the periodicity is broken and the potential acquires the crucial property  $V(x + np) = V(x) + neFp$ . Making the change of variable  $x = x' + np$  and using the above

property of  $V(x)$ , Equation (1) becomes

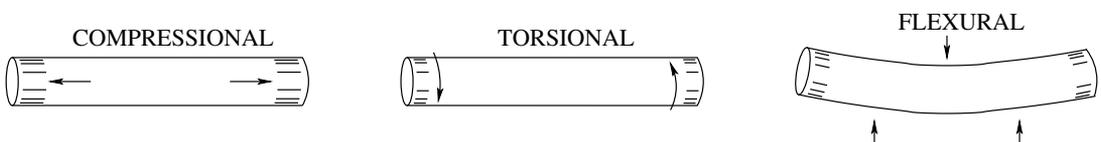
$$-\frac{1}{2} \frac{d^2\varphi(x')}{dx'^2} + V(x' + np)\varphi(x') = E\varphi(x') \quad \Longrightarrow \quad -\frac{1}{2} \frac{d^2\varphi(x')}{dx'^2} + V(x')\varphi(x') = (E - neFp)\varphi(x'), \quad (2)$$

where  $\varphi(x') = \psi(x' + np)$ . Comparing Equations (1) and (2), we see that if  $E$  is an eigenvalue, then  $E - neFp$  is also an eigenvalue. The difference between two consecutive eigenvalues is  $eFp$  and the Wannier–Stark ladder is formed. We should mention, however, that the simple mathematical derivation just described above is rather subtle [Zak 1968; 1969; Wannier 1969; Rabinovitch and Zak 1971] and a rigorous description is very difficult. For this reason, as mentioned before, the existence of the Wannier–Stark ladder in quantum mechanics was a controversial matter for many years. Actually, one finds that the energies forming a ladder are indeed a set of resonances embedded in a continuous energy spectrum. This is why the Wannier–Stark ladders are called WSLR.

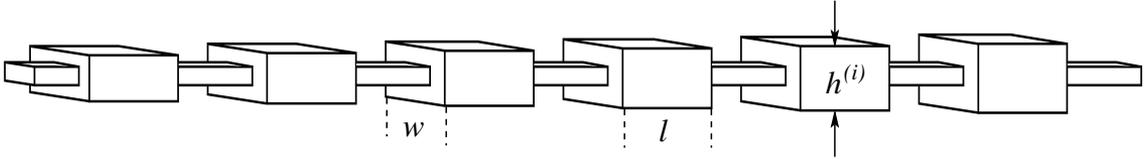
Several systems whose behavior is governed by classical physics and which present analogous phenomena to the band structure, BO, and WSLR, have been studied up to now. For example, an electromagnetic wave traveling through a structure with a dielectric function that varies periodically will exhibit a band structure, which in turn gives rise to photonic crystals. Applications of photonic crystals in light flow control have been described by Joannopoulos et al. [1995]. In an elastic structure with a specific impedance that varies periodically, the transmission spectra of elastic waves will also show a band structure [Esquivel-Sirvent and Coccoletzi 1994]. For theoretical studies of the BO and WSLR analogies see [Monsivais et al. 1990; Mateos and Monsivais 1994; de Sterke et al. 1998; Monsivais et al. 2003] and references therein. However, there are relatively few experimental studies [Sapienza et al. 2003; Agarwal et al. 2004], and it is just in this context that this paper is placed. We show experimentally that it is possible to find a WSLR-like structure in the spectrum of frequencies associated with torsional waves of special beams. We also use a numerical model whose predictions are in excellent agreement with our measurements. We have used our numerical model to show that the separation between the frequencies depends linearly on the parameter that plays the role of the static external electric field. However, we will see that the analogies are not exact.

## 2. The physical system

The system analyzed in this paper is a special elastic beam described below. It is well known that in any elastic system there exist several types of waves. However, for the case of beams, and in the range of frequencies and wave lengths used in this work, it can be assumed that only three types of vibrations exist: compressional, torsional and flexural; see Figure 1. For the present study we have considered only torsional vibrations.



**Figure 1.** Three types of vibrations in a beam: compressional, torsional and flexural.



**Figure 2.** Beam used to obtain Wannier–Stark ladder resonances. Cuboids have same length  $l = 5$  cm and width  $w = 1.905$  cm, but different heights.  $v^{(i)} = (1 + i\gamma)v$ ,  $i = 1, 2, \dots, 15$  with  $v = 2027.3$  m/s and  $\gamma = 0.02786$ . The width, height, and length of the small cuboids are  $w' = 5$  mm,  $h' = 5$  mm, and  $l' = 6$  mm, respectively.

The elastic system is depicted in Figure 2. It was constructed by machining a solid aluminum piece whose original shape was a beam with rectangular cross section. The result of the machining is a set on  $N$  cuboids or subbeams of constant width  $w$  and constant length  $l$ . They have different heights  $h^{(i)}$ , with  $w, h^{(i)} \ll l$  for  $i = 1, 2, \dots, N$ . These cuboids are separated by other small cuboids of dimensions  $w', l', h' \ll l$ , where  $w' = h'$ . We have observed that the behavior of machined systems is different from the behavior of similar systems constructed by welding their different parts. This property can be used to carry out nondestructive tests on the systems.

We now discuss the rule used to assign the values of the heights  $h^{(i)}$ . The procedure is different from the formulation used by Wannier just discussed above because the torsional waves are described by an equation different from the one describing the electrons in an electrified crystal. To design the beam we are guided by a qualitative analysis of what could be called an independent beam model in which each body oscillates independently from the rest.

It is well known that torsional waves are described by the equation [Graff 1975]

$$\frac{\partial^2 \theta}{\partial x^2} - \left(\frac{1}{v}\right)^2 \frac{\partial^2 \theta}{\partial t^2} = 0,$$

where  $v$  is the velocity of the waves and  $\theta = \theta(x, t)$  the angle of rotation of the cross section at point  $x$  and time  $t$ . The  $x$ -axis lies on the axis of the beam. We now apply this equation to the torsional normal modes of the  $i$ -th cuboid, for  $i = 1, 2, \dots, N$ , with free ends. If we denote by  $n_n$  the number of nodes of this mode and by  $\omega_{n_n}^{(i)}$  its angular frequency, the above equation becomes

$$\frac{\partial^2 \theta}{\partial x^2} + (k_{n_n}^{(i)})^2 \theta = 0,$$

where  $k_{n_n}^{(i)}$  is the wave number of the mode given by  $k_{n_n}^{(i)} = \omega_{n_n}^{(i)}/v^{(i)} = 2\pi/\lambda_{n_n}^{(i)}$ , where  $\lambda_{n_n}^{(i)}$  is the wave length and  $v^{(i)}$  the velocity of the waves in the  $i$ -th cuboid. It is clear that the length  $l$  of a cuboid with free ends is related to  $\lambda_{n_n}^{(i)}$  via  $l = \lambda_{n_n}^{(i)} n_n/2$ , which implies that the angular frequency  $\omega_{n_n}^{(i)}$  is given by

$$\omega_{n_n}^{(i)} = \pi v^{(i)} n_n/l. \tag{3}$$

To obtain a set of equidistant frequencies (for a given value of  $n_n$ ) we look for a set of velocities  $\{v^{(i)}\}$  such that  $v^{(i)} = (1 + i\gamma)v$ , where  $v$  is an arbitrary constant velocity. The parameter  $\gamma$  is dimensionless.

We then obtain

$$\omega_{n_n}^{(i)} = \frac{\pi v(1+i\gamma)}{l} n_n, \quad \Delta_{n_n} \equiv \Delta\omega_{n_n}^{(i)} = \frac{\pi v\gamma}{l} n_n, \quad (4)$$

for the frequencies and their differences (for a given value of  $n_n$ )  $\Delta\omega_{n_n}^{(i)} = \omega_{n_n}^{(i)} - \omega_{n_n}^{(i-1)}$ . Since the latter are independent of  $i$ , we have dropped the index  $i$  and defined  $\Delta_{n_n}$ . In this model of independent beams, we therefore obtain sets of equidistant frequencies for each value of  $n_n$ . The required set of velocities  $\{v^{(i)}\}$  can be obtained by taking appropriate values for the heights  $h^{(i)}$  as described below. We should mention that this procedure is not possible for cylindrical bars, since the torsional wave velocity in cylindrical bars is independent of the radius.

There is, however, another possibility to have equidistant frequencies. It consists in taking different lengths  $l^{(i)}$  for the different subbeams in Equation (3), with  $l^{(i)} = l/(1+i\gamma)$ , but this possibility will not be considered here. The corresponding analysis has been published recently [Gutiérrez et al. 2006].

To obtain the velocities  $\{v^{(i)}\}$  we have used the expression derived by Navier [1827] for the torsional velocity in the  $i$ -th cuboid

$$v^{(i)} = \sqrt{\frac{G}{\rho}} \sqrt{\frac{\alpha^{(i)}}{I^{(i)}}},$$

where  $I^{(i)} = (h^{(i)}w^3 + [h^{(i)}]^3w)/12$  is the moment of inertia with respect to the axis of the system,  $\rho$  is the density,  $G$  is the shear modulus, and  $\alpha^{(i)}$  is given by

$$\alpha^{(i)} = \frac{256}{\pi^6} \sum_{m=0}^{\infty} \sum_{p=0}^{\infty} \frac{1}{(2m+1)^2(2p+1)^2} \frac{h^{(i)}w}{[(2m+1)/h^{(i)}]^2 + [(2p+1)/w]^2}.$$

We can solve these equations to obtain the values of  $h^{(i)}$  such that  $v^{(i)}$  take the required value  $(1+i\gamma)v$ . Figure 3 shows a plot of  $v$  as a function of  $h$  for particular values of the parameters  $w$  and  $\sqrt{G/\rho}$ . This figure also shows a comparison with experimental results.

We now return to discuss the properties of the whole beam constructed by machining a solid piece as shown in Figure 2. We have used the above procedure to calculate the  $N$  heights  $\{h^{(i)}\}$  of the cuboids or subbeams forming the beam. When the parameter  $\gamma$  is equal to zero, a locally periodic beam is formed. This kind of locally periodic beams shows a discrete band spectrum [Morales et al. 2002]. If we break the periodicity by setting  $\gamma \neq 0$ , a completely different spectrum occurs. The discrete band structure disappears and the new spectrum resembles the WSLR. We see from Equation (4) that  $\gamma$  here plays the role of the electric field  $F$  for the quantum mechanical ladders.

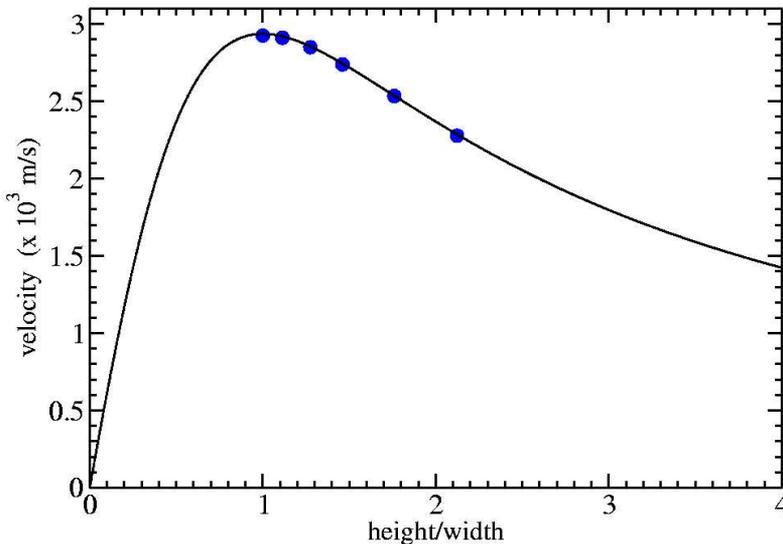
Before presenting the calculations of the normal modes for this system and showing numerical and experimental results, let us make a qualitative analysis to see what type of spectrum could be expected from the independent beam model. At the lowest frequencies, the wavelength  $\lambda$  is of the same order of magnitude as  $L \approx lN$  and the whole beam is excited. When  $\omega$  increases and  $\lambda$  becomes of the order of  $l$ , the length of the subbeams, the state equivalent to its lowest normal mode is excited. This occurs at the first subbeam since it has the smallest velocity  $v^{(1)} = (1+\gamma)v$ ; furthermore, this corresponds to the subbeam with the smallest  $h^{(i)}$ ; see Figure 3. The rest of the subbeams are out of resonance, so the amplitude decreases as one moves away from subbeam 1. Therefore, the state is localized around the latter. In some sense this was to be expected since we are disturbing a periodic structure to obtain a disordered one-dimensional system, which always shows localized wave amplitudes. Increasing the

exciting frequency by  $\Delta_1$ , the subbeam with velocity  $v^{(2)} = (1 + 2\gamma)v$ , that is, subbeam 2, will now be excited and the rest will be out of resonance. The amplitude of the vibrations therefore decreases as the distance from subbeam 2 increases. The wave amplitude is again localized but now around subbeam 2; it has a similar shape as the wave amplitude that subbeam 1 had previously. The same arguments apply when subbeam  $i$  of velocity  $v^{(i)} = (1 + i\gamma)v$  is excited.

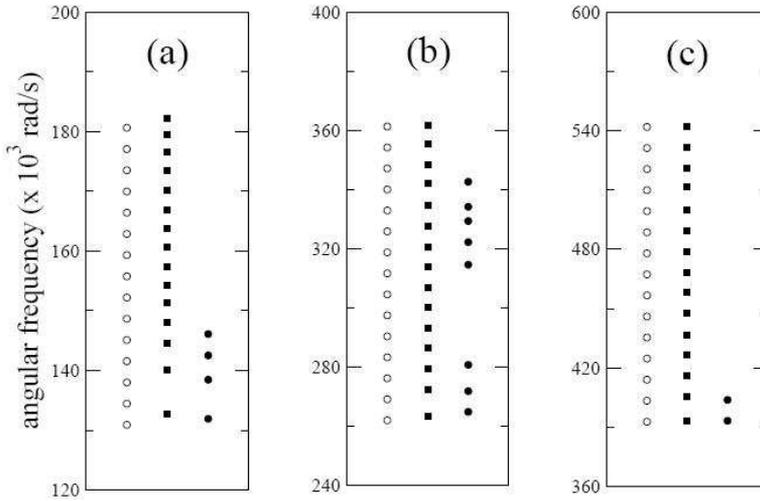
What we have done is to produce a finite Wannier–Stark ladder, that is,  $N$  localized states with constant difference in frequency given by Equation (4). However, more ladders exist since normal modes with two or more nodes can also be excited in each subbeam. For instance, taking  $n_n = 2$  in Equation (4), a second ladder is obtained. This ladder is different from the first one because the frequency difference is now twice the one of the lower ladder, as can be seen from Equation (4). The states are again localized and all have similar shape. A third ladder exists with  $\Delta_3 = 3\Delta_1$  and so on for the other values of  $n_n$ . The difference between the quantum-mechanical WSLR [Thommen et al. 2004] and the ladders discussed here is that in the latter the spacing between resonances is not the same for different ladders.

To measure normal modes frequencies and amplitudes we have used an Electromagnetic Acoustic Transducer (EMAT) which has been recently developed [Morales et al. 2001; 2002]. This EMAT is versatile and operates at low frequencies. It can selectively excite or detect compressional, flexural or torsional vibrations. We have also calculated these quantities using the transfer matrix method for torsional waves with the following boundary conditions between the different sections of the beam

$$\theta^{(i)}(x)|_{x=x_i} = \theta^{(i+1)}(x)|_{x=x_i}, \quad \beta^{(i)} \frac{d\theta^{(i)}(x)}{dx} \Big|_{x=x_i} = \beta^{(i+1)} \frac{d\theta^{(i+1)}(x)}{dx} \Big|_{x=x_i},$$

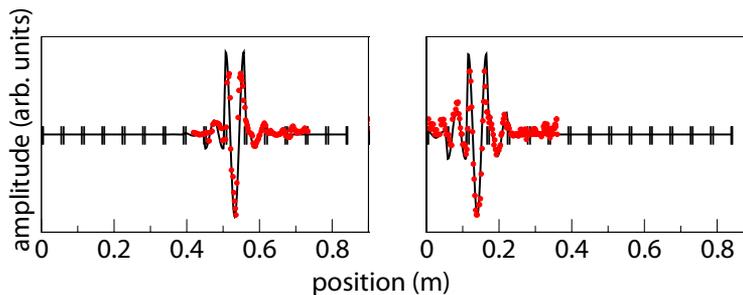


**Figure 3.** Navier prediction for velocity  $v$  as a function of height  $h$  for cuboids (continuous line). Experimental values (points) fit the prediction. Here  $\sqrt{G/\rho} = 3190$  m/s and  $w = 1.905$  cm.

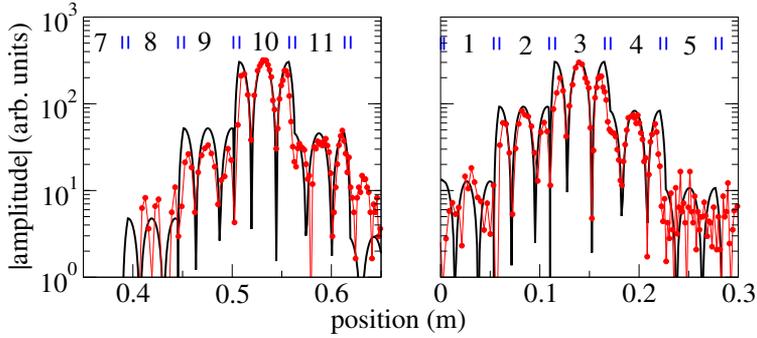


**Figure 4.** Normal mode angular frequencies of beam in Figure 2 yielding the elastic Wannier–Stark effect: (a)  $n_n = 1$ , (b)  $n_n = 2$ , (c)  $n_n = 3$ . In each figure, left column shows frequencies from the independent beam model, middle column — from the transfer matrix model, right column — measured in the laboratory.

where  $\beta^{(i)}$  is the square of the cross-section area of the  $i$ -th beam. Free end boundary conditions were used as discussed by Morales et al. [2002]. Our calculation shows explicitly that the frequency difference  $\Delta_{n_n}$  is proportional to the parameter  $\gamma$ . Furthermore, as mentioned above, for  $\gamma = 0$  a discrete band spectrum appears, and as  $\gamma$  grows the levels of each band separate to form the WSLR. In Figure 4 we show the theoretical normal mode frequencies and the values obtained from the independent beam model of the system shown in Figure 2 for  $\gamma = 0.02786$ . We see that this model provides a rather good first approximation. As also shown in this figure, the experimental values are very well reproduced by the theoretical values.



**Figure 5.** Two wave amplitudes for torsional waves of beam in Figure 2 associated with the second ladder  $n_n = 2$ , the left one localized on the tenth subbeam, the right one on the third. Double small vertical lines along beam axis indicate the position of small cuboids. Points are experimental values, while solid line gives calculated values.



**Figure 6.** Logarithmic plot of the absolute value of wave amplitudes shown in [Figure 5](#). Points are experimental values, while solid line gives calculated values.

One can clearly see from [Figure 4](#) that the states form a set of WSLR. The first band composed by extended modes is not displayed in this graph. Notice that the frequencies in the extremes of each ladder do not have the same frequency difference as those in the middle of the ladder. This is due to a border effect in the wave amplitudes localized near the free ends [[Gutiérrez et al. 2006](#)].

In [Figure 5](#) we show as an example the comparison between theoretical and experimental wave amplitudes for two states of the second WSLR. These are localized around particular subbeams. For example, in [Figure 5](#) (left) the tenth subbeam resonates while in [Figure 5](#) (right) the third subbeam resonates. Both wave amplitudes have the same form and, as expected, the amplitudes show two nodes at the resonating subbeams. Note that we again have excellent agreement between experiment and theory, where we have used the one-dimensional transfer method in spite of the fact that  $w$  and  $h^{(i)}$  are not much smaller than  $l$  as required by the method. We have also calculated the wave amplitudes for states of the first ladder. Localization is again observed, and the amplitudes show one node at the resonating subbeams.

In [Figure 6](#) we show the theoretical and experimental values of the logarithm of the wave amplitude corresponding to the states of [Figure 5](#). The plots show that the envelopes of the amplitudes of the normal modes are exponentially localized.

### 3. Conclusions

In this paper we have constructed an elastic analogue of the WSLR in the torsional frequency spectra of some special beams. Starting with the independent beam model, we find the appropriate geometry of the bars in order to have the WSLR. The geometry is related to the cross section of the subbeams forming the whole beam, as indicated by the Navier formula, which we tested experimentally for the first time. In contrast. In contrast with the optical analogue of [[Sapienza et al. 2003](#); [Agarwal et al. 2004](#)], we have observed the WSLR directly. Furthermore, we have measured the wave amplitudes, including phases, which show exponential localization. We also observed higher WSLR. Our numerical studies are in close agreement with experimental results. The elastic WSLR have potential applications in the design of systems with localized vibrations.

## References

- [Agarwal et al. 2004] V. Agarwal, J. A. del Río, G. Malpuech, M. Zamfirescu, A. Kavokin, D. Coquillat, D. Scalbert, M. Vladimirova, and B. Gil, “Photon Bloch oscillations in porous silicon optical superlattices”, *Phys. Rev. Lett.* **92** (2004), #097401.
- [Banavar and Coon 1978] J. R. Banavar and D. D. Coon, “Widths and spacing of Stark ladder levels”, *Phys. Rev. B* **17**:10 (1978), 3744–3749.
- [Bloch 1928] F. Bloch, “Über die Quantenmechanik der Electronen in Kristallgittern”, *Z. Phys.* **52** (1928), 555–600.
- [Brillouin 1946] L. Brillouin, *Waveguide propagation in periodic structures; electric filters and crystal lattices*, McGraw Hill, New York, 1946. Reprinted Dover, New York, 1953.
- [Dekorsy et al. 1994] T. Dekorsy, P. Leisching, K. Köhler, and H. Kurz, “Electro-optic detection of Bloch oscillations”, *Phys. Rev. B* **50**:11 (1994), 8106–8109.
- [Esaki and Tsu 1970] L. Esaki and R. Tsu, “Superlattice and negative differential conductivity in semiconductors”, *IBM J. Res. Dev.* **14**:1 (1970), 61–65.
- [Esquivel-Sirvent and Coccoletzi 1994] R. Esquivel-Sirvent and G. H. Coccoletzi, “Band structure for the propagation of elastic waves in superlattices”, *J. Acoust. Soc. Am.* **95**:1 (1994), 86–90.
- [Feldmann et al. 1992] J. Feldmann, K. Leo, J. Shah, D. A. B. Miller, J. E. Cunningham, T. Meier, G. von Plessen, A. Schulze, P. Thomas, and S. Schmitt-Rink, “Optical investigation of Bloch oscillations in a semiconductor superlattice”, *Phys. Rev. B* **46**:11 (1992), 7252–7255.
- [Graff 1975] K. F. Graff, *Wave motion in elastic solids*, Clarendon, Oxford, 1975. Reprinted Dover, New York, 1991.
- [Gutiérrez et al. 2006] L. Gutiérrez, A. D. de Anda, J. Flores, R. A. Méndez-Sánchez, G. Monsivais, and A. Morales, “Wannier–Stark ladders in one-dimensional elastic systems”, *Phys. Rev. Lett.* **97** (2006), #114301.
- [Hart and Emin 1988] C. F. Hart and D. Emin, “Time evolution of a Bloch electron in a constant electric field”, *Phys. Rev. B* **37**:11 (1988), 6100–6104.
- [James 1949] H. M. James, “Electronic states in perturbed periodic systems”, *Phys. Rev.* **76**:11 (1949), 1611–1624.
- [Joannopoulos et al. 1995] J. D. Joannopoulos, R. D. Meade, and J. N. Winn, *Photonic crystals: molding the flow of light*, Princeton University Press, Princeton, NJ, 1995.
- [Leo et al. 1992] K. Leo, P. H. Bolivar, F. Brüggemann, R. Schwedler, and K. Köhler, “Observation of Bloch oscillations in a semiconductor superlattice”, *Solid State Commun.* **84**:10 (1992), 943–946.
- [Lyssenko et al. 1998] V. G. Lyssenko, M. Sudzius, F. Loser, and G. Valusis, “Bloch oscillations in semiconductor superlattices”, *Adv. Solid State Phys.* **38** (1998), 257–268.
- [Mateos and Monsivais 1994] J. L. Mateos and G. Monsivais, “Stark-ladder resonances in elastic waves”, *Physica A* **207**:1-3 (1994), 445–451.
- [Méndez et al. 1988] E. E. Méndez, F. Agulló-Rueda, and J. M. Hong, “Stark localization in GaAs-GaAlAs superlattices under an electric field”, *Phys. Rev. Lett.* **60**:23 (1988), 2426–2429.
- [Monsivais et al. 1990] G. Monsivais, M. del Castillo-Mussot, and F. Claro, “Stark-ladder resonances in the propagation of electromagnetic waves”, *Phys. Rev. Lett.* **64**:12 (1990), 1433–1436.
- [Monsivais et al. 2003] G. Monsivais, R. Rodríguez-Ramos, R. Esquivel-Sirvent, and L. Fernández-Álvarez, “Stark-ladder resonances in piezoelectric composites”, *Phys. Rev. B* **68**:17 (2003), #174109.
- [Morales et al. 2001] A. Morales, L. Gutiérrez, and J. Flores, “Improved eddy current driver-detector for elastic vibrations”, *Am. J. Phys.* **69**:4 (2001), 517–522.
- [Morales et al. 2002] A. Morales, J. Flores, L. Gutiérrez, and R. A. Méndez-Sánchez, “Compressional and torsional wave amplitudes in rods with periodic structures”, *J. Acoust. Soc. Am.* **112**:5 (2002), 1961–1967.
- [Navier 1827] M. Navier, “Mémoire sur lois de l’équilibre et du mouvement des corps solides élastiques”, *Mém. Acad. Roy. Sci. Inst. France* **7** (1827), 375–393.
- [Rabinovitch 1977] A. Rabinovitch, “Stark ladders in finite crystals!”, *Phys. Lett. A* **59**:6 (1977), 475–477.

- [Rabinovitch and Zak 1971] A. Rabinovitch and J. Zak, “Electrons in crystals in a finite-range electric field”, *Phys. Rev. B* **4**:8 (1971), 2358–2370.
- [Sapienza et al. 2003] R. Sapienza, P. Costantino, D. Wiersma, M. Ghulinyan, C. J. Oton, and L. Pavesi, “Optical analogue of electronic Bloch oscillations”, *Phys. Rev. Lett.* **91**:26 (2003), #263902.
- [Sheng 1995] P. Sheng, *Introduction to wave scattering, localization, and mesoscopic phenomena*, Academic Press, New York, 1995.
- [Soukoulis 1996] C. M. Soukoulis (editor), *Photonic band gap materials* (Elounda, Greece, 1995), NATO ASI series E **315**, Kluwer, Dordrecht, 1996.
- [de Sterke et al. 1998] C. M. de Sterke, J. N. Bright, P. A. Krug, and T. E. Hammon, “Observation of an optical Wannier–Stark ladder”, *Phys. Rev. E* **57**:2 (1998), 2365–2370.
- [Thommen et al. 2004] Q. Thommen, J. C. Garreau, and V. Zehnlé, “Quantum interference in a driven washboard potential”, *Am. J. Phys.* **72**:8 (2004), 1017–1025.
- [Wannier 1955] G. H. Wannier, “Possibility of a Zener effect”, *Phys. Rev.* **100**:4 (1955), 1227–1227.
- [Wannier 1962] G. H. Wannier, “Dynamics of band electrons in electric and magnetic fields”, *Rev. Mod. Phys.* **34**:4 (1962), 645–655.
- [Wannier 1969] G. H. Wannier, “Stark ladder in solids? A reply”, *Phys. Rev.* **181**:3 (1969), 1364–1365.
- [Zak 1968] J. Zak, “Stark ladder in Solids?”, *Phys. Rev. Lett.* **20**:26 (1968), 1477–1481.
- [Zak 1969] J. Zak, “Stark ladder in solids? A reply to a reply”, *Phys. Rev.* **181**:3 (1969), 1366–1367.
- [Zener 1934] C. Zener, “A theory of the electrical breakdown of solid dielectrics”, *P. Roy. Soc. Lond. A Mat.* **145**:855 (1934), 523–529.

Received 31 Jul 2006. Accepted 8 May 2007.

GUILLERMO MONSIVAIS: [monsi@fisica.unam.mx](mailto:monsi@fisica.unam.mx)

*Instituto de Física, Universidad Nacional Autónoma de México, Apdo. Postal 20-364, México 01000 DF, México*

RAFAEL A. MÉNDEZ-SÁNCHEZ: [mendez@ce.fis.unam.mx](mailto:mendez@ce.fis.unam.mx)

*Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México, P.O. Box 48-3, 62251 Cuernavaca Mor., México*

ALFREDO DÍAZ DE ANDA: [ada@ce.fis.unam.mx](mailto:ada@ce.fis.unam.mx)

*Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México, P.O. Box 48-3, 62251 Cuernavaca Mor., México*

JORGE FLORES: [flores@fisica.unam.mx](mailto:flores@fisica.unam.mx)

*Instituto de Física, Universidad Nacional Autónoma de México, Apdo. Postal 20-364, México 01000, DF, México*

LUIS GUTIÉRREZ: [luisg@ce.fis.unam.mx](mailto:luisg@ce.fis.unam.mx)

*Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México, P.O. Box 48-3, 62251 Cuernavaca Mor., México*

ALEJANDRO MORALES: [mori@fis.unam.mx](mailto:mori@fis.unam.mx)

*Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México, P.O. Box 48-3, 62251 Cuernavaca Mor., México*

## A NEW VARIABLE DAMPING SEMIACTIVE DEVICE FOR SEISMIC RESPONSE REDUCTION OF CIVIL STRUCTURES

ORLANDO CUNDUMI AND LUIS E SUÁREZ

A semiactive mechanism, called a VDSA (variable damping semiactive device), is proposed to reduce the seismic response of structures. It is composed of two fixed-orifice viscous fluid dampers installed in the form of a V whose top ends are attached to a floor and their lower ends to a collar that moves along a vertical rod. By varying the VDSA position one obtains an optimal instantaneous damping added to the structure. The position of the moving end is calculated with an algorithm based on a variation of the instantaneous optimal control theory which includes a generalized LQR (linear quadratic regulator) scheme. This modified algorithm, referred to as  $Qv$ , is based on the minimization of a performance index  $J$  quadratic in the state vector, the control force vector, and an absolute velocity vector. Two variants of the algorithm are used to present numerical simulations of the controlled seismic response of a single and a MDOF (multi-degree-of-freedom) structure.

### 1. Introduction

Civil engineering structures are typically designed to rely on their strength and ductility to withstand the large forces imposed on them by strong earthquakes. A number of modern mechanical devices have been proposed in the last two decades to reduce the structural response. They are known collectively as protective devices and they include added viscoelastic dampers, viscous fluid dampers, frictional dampers, tuned-mass dampers, and base isolation systems. The devices themselves and their design methodology are referred to as passive control systems. At the highest level of sophistication for seismic protection are the so called active control systems. Although these devices provide in theory the uppermost response reduction, they also required a large amount of energy to operate and their robustness and reliability are questionable.

In between these passive and active systems are the semiactive devices which, as the name indicates, combine the features of the former two protective systems. The force (and thus the energy) required to operate a semiactive device is much less than for an active system. To calculate the control forces that operate the passive devices, it is necessary to know the response of the structure by measuring it with sensors. A proper numerical algorithm processes this information and calculates how the properties of the (formerly) passive device should be modified.

Semiactive control systems have only recently been considered for applications to large civil structures. We believe that the application of these systems to civil engineering structures was first reported by Hrovat et al. [1983]. Several devices that can deliver changeable variable damping such as variable orifice dampers [Symans and Constantinou 1997; Kurata et al. 1999; Kurata et al. 2000] and hydraulics dampers [Kawashima et al. 1992; Patten et al. 1993; Sack et al. 1994; Patten et al. 1996] have been

---

*Keywords:* semiactive systems, control algorithms, earthquake engineering, seismic response, added damping.

proposed. Variable stiffness devices have also been proposed by Kobori et al. [1993], Nagarajaiah and Mate [1998], and Gluck et al. [2000]. Furthermore, numerous algorithms have been developed for selecting the appropriate damping coefficient [Yang et al. 1987; Soong 1990; Sadeck and Mohraz 1998; Cundumi 2005; Cundumi and Suárez 2006b]. The list of references is meant only to provide a few relevant examples; a comprehensive review of these systems is beyond the scope of this discussion.

The present paper describes the implementation of a VDSA device. In contrast to semiactive dampers described in the technical literature, the damper coefficient  $c$  is not controlled by modifying the size of an orifice in the piston, but by changing the position of the damper. The required damping coefficient is calculated by means of two instantaneous optimal control algorithms: the (*closed-loop control* and the *closed-open-loop control*). It is shown that both algorithms are effective in reducing the response. The damping coefficient  $c(t)$  during the response can be adjusted between an upper limit  $c_{\max}$  and a lower value  $c_{\min}$ .

This paper contains in detail the formulation and the results of a paper presented at the 9th Pan American Congress of Applied Mechanics, in Mérida, Mexico [Cundumi and Suárez 2006a].

## 2. The modified algorithm $Qv$

It is well known that the equations of motion of a structure modeled as a MDOF system and subjected to a base acceleration  $\ddot{x}_g(t)$  at all its supports are given by

$$[M]_{n \times n} \{\ddot{x}(t)\} + [C]_{n \times n} \{\dot{x}(t)\} + [K]_{n \times n} \{x(t)\} = -[M]_{n \times n} \{E\} \ddot{x}_g(t), \quad (1)$$

where  $[M]$ ,  $[C]$  and  $[K]$  are the mass, damping and stiffness matrix, respectively, the vectors  $\{\ddot{x}(t)\}$ ,  $\{\dot{x}(t)\}$  and  $\{x(t)\}$  contain the relative (with respect to the foundation) acceleration, velocity and displacement of each dynamic degree of freedom of the structure,  $\{E\}$  is the vector of influence coefficients, and  $n$  is the number of degrees of freedom. If all the degrees of freedom of the structural model coincide with the direction of the applied ground motion, then the vector  $\{E\}$  is simply a vector with ones  $\{I\}$ .

If the structure is outfitted with  $r$  semiactive dampers, the previous equations of motion must be changed as follows:

$$[M]_{n \times n} \{\ddot{x}(t)\} + [C]_{n \times n} \{\dot{x}(t)\} + [K]_{n \times n} \{x(t)\} = -[M]_{n \times n} \{E\} \ddot{x}_g(t) + [D]_{n \times r} \{u(t)\}. \quad (2)$$

The matrix  $[D]$  defines the locations of the controllers,  $r$  is the number of controllers and  $\{u(t)\}$  is the  $r$ -dimensional control force vector. The location of the controllers (or the VDSA devices in our case) will be determined via a trial and error process by trying to maximize the effect of the devices. No attempt is made to determine the optimal position of the devices in an analytical way.

To solve the system of equations of motion, Equation (2), by transforming them into a set of uncoupled equations, it is convenient to change into a system of  $2n$  first order differential equations. In Linear System Theory this method is referred to as the state-space representation. Introducing the following

response vector and matrices,

$$\{z(t)\} = \left\{ \begin{array}{l} \{x(t)\} \\ \{\dot{x}(t)\} \end{array} \right\}, \quad [A] = \left[ \begin{array}{c|c} 0 & I \\ \hline -M^{-1}K & -M^{-1}C \end{array} \right],$$

$$[B] = \left[ \begin{array}{c} 0 \\ M^{-1}D \end{array} \right], \quad [H] = \left[ \begin{array}{c} 0 \\ -E \end{array} \right].$$

Equation (2) can be written in the form:

$$\{\dot{z}(t)\}_{2n \times 1} = [A]_{2n \times 2n} \{z(t)\} + [B]_{2n \times 2r} \{u(t)\} + [H]_{2n \times 1} \ddot{x}_g(t).$$

To define the variation of the control forces in  $\{u(t)\}$  one needs to select a control algorithm. In this study, two algorithms (closed-loop control and closed-open-loop control) have been developed based on the instantaneous optimal control theory. They are referred to here as the modified algorithms  $Qv$ . As usual, this type of algorithm is based on the minimization of a performance index  $J$  quadratic in the state vector  $\{z(t)\}$  and in the control force  $\{u(t)\}$ . However, in the modified algorithm a quadratic form of the absolute velocity  $\{\dot{x}_a(t)\}$  is added to  $J$ . A penalty on the state vector is imposed through a matrix  $Q$ , on the control vector through a matrix  $R$  and on the absolute velocity vector through a matrix  $Qv$ .  $Q$  and  $Qv$  are two symmetric positive semidefinite weighting matrices of size  $2n \times 2n$  and  $n \times n$ , respectively, and  $R$  is an  $r \times r$  positive definite weighting matrix. The performance index takes the form:

$$J = \int_0^{t_f} \left[ \{z(t)\}^T [Q] \{z(t)\} + \{\dot{x}_a(t)\}^T [Qv] \{\dot{x}_a(t)\} + \{u(t)\}^T [R] \{u(t)\} \right] dt,$$

where  $t_f$  is the duration of excitation. Usually the excitation is not included in the definition of performance indices. However, it was found that for a semiactive device with variable damping such as the one presented in this work, including the excitation in the definition of  $J$  through the absolute velocity has a beneficial effect on the effectiveness of the device.

The absolute velocity vector is computed as

$$\{\dot{x}_a(t)\} = [A_v]_{n \times 2n} \{z(t)\} + \{S_v\}_{n \times 1} \dot{x}_g(t),$$

where  $[A_v] = [0 \mid I]$ ,  $\{S_v\} = \{1\}$ ,  $[I]$  is an  $n \times n$  identity matrix,  $\{1\}$  is a vector of 1's of length  $n$ , and  $\dot{x}_g(t)$  is the ground velocity.

The procedure to define the control and response vectors in the modified algorithm  $Qv$  can be found in [Cundumi 2005]. Here only the final results are reported.

For the closed-loop control case, the variables  $\{u(t)\}$  and  $\{z(t)\}$  can be obtained as follows:

$$\{u(t)\} = -\frac{\Delta t}{2} [R]^{-1} [B]^T \left[ [A_2] \{z(t)\} + [A_3] \dot{x}_g(t) \right],$$

$$\{z(t)\} = \left[ [I] + \frac{\Delta t^2}{4} [B][R]^{-1}[B]^T[A_2] \right]^{-1} \left[ [T] \{d(t - \Delta t)\} - \frac{\Delta t^2}{4} [B][R]^{-1}[B]^T[A_3] \dot{x}_g(t) + \frac{\Delta t}{2} [H] \ddot{x}_g(t) \right],$$

where  $\Delta t$  is the constant time step,  $[A_2] = [Q] + [A_v]^T [Qv] [A_v]$ ,  $[A_3] = [A_v]^T [Qv] [S_v]$  and  $\{d(t - \Delta t)\}$  contains the displacement  $x$  and the velocity  $\dot{x}$  at time  $t - \Delta t$ .

For the closed-open-loop control case,  $\{u(t)\}$  and  $\{z(t)\}$  are calculated with the following equations:

$$\{u(t)\} = \frac{\Delta t}{4} [R]^{-1} [B]^T [P] \{z(t)\} + \{p(t)\}, \tag{3}$$

$$\{z(t)\} = \left[ [I] - \frac{\Delta t^2}{8} [B][R]^{-1}[B]^T [P] \right]^{-1} \left[ [T] \{d(t - \Delta t)\} + \frac{\Delta t^2}{8} [B][R]^{-1}[B]^T \{p(t)\} + \frac{\Delta t}{2} [H] \ddot{x}_g(t) \right]. \tag{4}$$

In Equations (3) and (4),  $[P]$  is the Riccati matrix and  $\{p(t)\}$  represents the open-loop control.

$$[P] = - [Q] + 2[A_v]^T [Q_v][A_v] \left[ [I] + \frac{\Delta t^2}{8} [Q][B][R]^{-1}[B]^T \right]^{-1}$$

$$\{p(t)\} = - \left[ \frac{\Delta t^2}{8} [Q][B][R]^{-1}[B]^T + [I] \right]^{-1} \left[ [Q] \left[ [T] \{d(t - \Delta t)\} + \frac{\Delta t}{2} [H] \ddot{x}_g(t) \right] + 2[A_v]^T [Q_v] \{S_v\} \dot{x}_g(t) \right]. \tag{5}$$

### 3. Equations of motion of SDOF structures controlled with the VDSA device

The system considered is shown schematically in Figure 1. It consists of a single degree of freedom structure (SDOF) with the proposed variable damping system installed. The dampers have fixed-constant damping coefficient  $C_{oA}$  and  $C_{oB}$ . The structure consists of a mass  $m$  distributed at the roof level, a massless frame that provides stiffness  $k$  to the system, and the natural (inherent) damping of the structure is represented by a damper with constant  $C_s$ . This coefficient can be defined as  $2\xi m\omega_n$  where  $\xi$  is the inherent (original) damping ratio and  $\omega_n$  is the natural frequency of the SDOF system. This model may be considered as an idealization of a one-story structure. In reality, each structural member (column, beam) of the structure contributes to the inertial (mass), elastic (stiffness), and energy dissipation (damping)

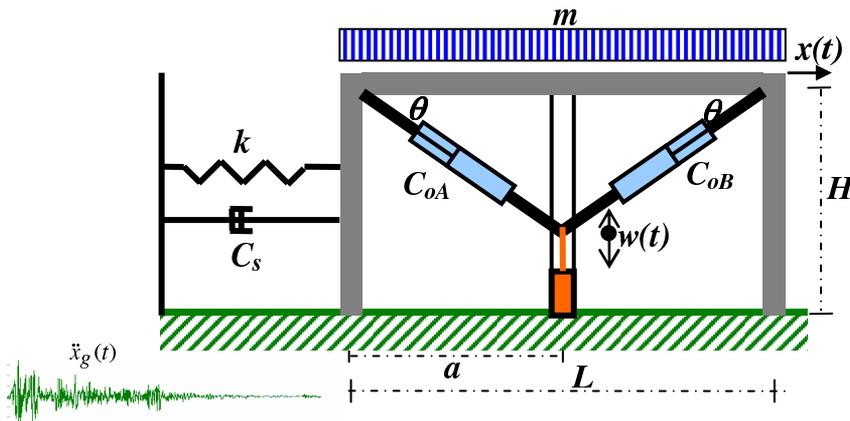


Figure 1. Single degree of freedom model of a structure with a VDSA device.

properties of the structure. In the idealized system, however, these properties are concentrated in three separate, pure components: a lumped mass, a linear spring, and a linear viscous damper.

To move the common lower end of the two dampers one needs to use an actuator, for instance a hydraulic actuator with fast reaction. The information required to determine the movement of the VDSA device includes the relative displacement and velocity of the mass  $m$  and the ground velocity  $\dot{x}_g(t)$ .

Figure 2 displays the velocities that govern the forces produced by the dampers:  $\dot{x}(t)$  is the relative velocity of the floor and  $\dot{w}(t)$  is the velocity of the bottom end of the dampers. As shown in the Figure 2, the damping force is proportional to the difference between the components of the velocities  $\dot{x}(t)$  and  $\dot{w}(t)$  along the axis of the dampers  $A$  and  $B$  of the VDSA device.

Using Figures 1 and 2, it can be shown that the equation of motion for the SDOF structure subjected to the horizontal component of an earthquake-induced ground acceleration is

$$m\ddot{x}(t) + (C_s + (C_{oA} + C_{oB}) \cos^2 \theta(t))\dot{x}(t) + kx(t) = -m\ddot{x}_g(t) + \frac{1}{2}(C_{oA} - C_{oB}) \sin 2\theta(t)\dot{w}(t), \quad (6)$$

where

$$\cos^2 \theta(t) = \frac{a^2}{a^2 + [H - w(t)]^2}, \quad \sin 2\theta(t) = \frac{2a[H - w(t)]}{a^2 + [H - w(t)]^2}, \quad a = \frac{L}{2},$$

and  $a$ ,  $H$ , and  $L$  are the dimensions shown in Figure 1.

For a structure with two dampers in a fixed position, the second term in the right hand side of the equation of motion, Equation (6), vanishes. This term arises due to the component of the velocity of the lower end of the dampers in the direction of the axis of the device. Rewriting Equation (6) in a space-state representation leads to

$$\begin{aligned} \begin{Bmatrix} \dot{z}_1(t) \\ \dot{z}_2(t) \end{Bmatrix} &= \begin{bmatrix} 0 & 1 \\ -m^{-1}k & -m^{-1}(C_s + (C_{oA} + C_{oB}) \cos^2 \theta(t)) \end{bmatrix} \begin{Bmatrix} z_1(t) \\ z_2(t) \end{Bmatrix} \\ &+ \begin{bmatrix} 0 \\ \frac{1}{2}m^{-1}(C_{oA} - C_{oB}) \sin 2\theta(t) \end{bmatrix} \dot{w}(t) + \begin{bmatrix} 0 \\ -1 \end{bmatrix} \ddot{x}_g(t), \end{aligned}$$

where  $z_1(t) = x(t)$  and  $z_2(t) = \dot{x}(t)$ .

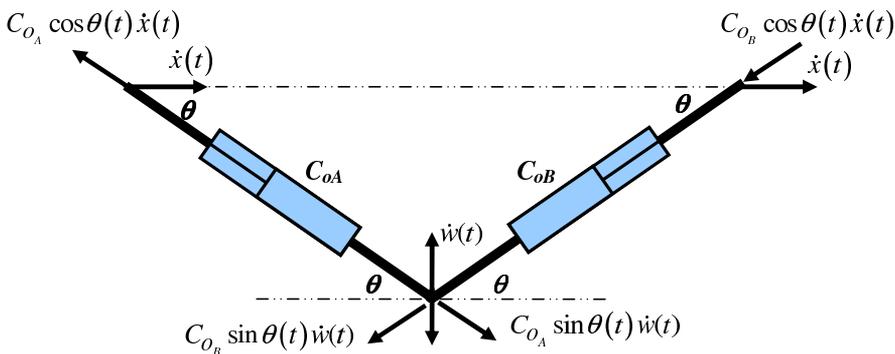


Figure 2. End velocities of the VDSA device installed in an SDOF structure.

These state-space equations can be solved by decoupling them with the complex eigenvectors of the matrix in the right hand side, provided that the displacement  $w(t)$  of the bottom support of the dampers is known. The term  $w(t)$  is determined by using one of the two modified algorithms  $Qv$  described in the previous section.

For practical reasons, the position  $w(t)$  of the common joint of the VDSA device (which governs the damping provided to the structure), must be bounded between two limiting values  $w_{\min}$  and  $w_{\max}$ . Thus the effective instantaneous damping in the structure can be represented by a dashpot with a variable coefficient given by

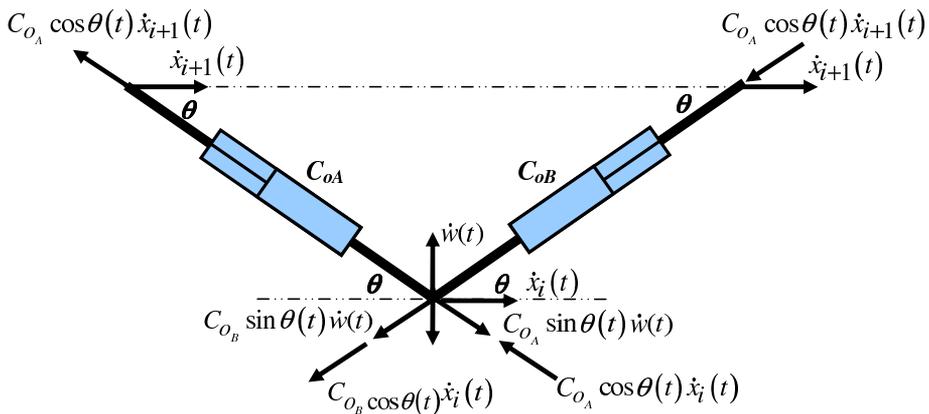
$$C(t) = \begin{cases} C_s + (C_{oA} + C_{oB}) \frac{a^2}{a^2 + [H - w_{\min}]^2}, & \text{for } w(t) < w_{\min}, \\ C_s + (C_{oA} + C_{oB}) \frac{a^2}{a^2 + [H - w(t)]^2}, & \text{for } w_{\min} < w(t) < w_{\max}, \\ C_s + (C_{oA} + C_{oB}) \frac{a^2}{a^2 + [H - w_{\max}]^2}, & \text{for } w(t) > w_{\max}. \end{cases} \quad (7)$$

#### 4. Equations of motion of MDOF structures controlled with the VDSA device

The application of the VDSA device to MDOF systems is similar to the SDOF case. When the VDSA device is installed between the  $i$ th and  $i + 1$ th building floors (and above the first level), the damping force generated by the VDSA device is related to the velocities  $\dot{x}_i(t)$ ,  $\dot{x}_{i+1}(t)$  and  $\dot{w}(t)$  as shown in Figure 3.

The equation of motion for a MDOF system with the device installed between the  $i$ th and  $(i + 1)$ th floor is

$$[M]\{\ddot{x}(t)\} + ([C_s] + [C_1] + [C_2])\{\dot{x}(t)\} + [K]\{x(t)\} = -[M]\{r\}\ddot{x}_g(t) - \{D\}\dot{w}(t). \quad (8)$$



**Figure 3.** End velocities of the VDSA device installed between two floors of an MDOF building.

The matrices  $[C_1]$  and  $[C_2]$  and the vector  $\{D\}$  are defined in terms of three vectors with only one or two nonzero elements. The vectors  $\{e_1\}$ ,  $\{e_2\}$  and  $\{e_3\}$ , with length  $n$ , are:

$$\begin{aligned} \{e_1\}^T &= [0, 0, \dots, 0, 1, 0, \dots, 0] && \text{with } 1 \text{ at column } i + 1 \\ \{e_2\}^T &= [0, 0, \dots, 0, -1, 0, \dots, 0] && \text{with } -1 \text{ at column } i \\ \{e_3\}^T &= [0, 0, \dots, 0, -1, 1, \dots, 0] && \text{with } -1 \text{ at column } i, 1 \text{ at column } i + 1. \end{aligned} \quad (9)$$

Using the three vectors in Equation (9), the matrices  $[C_1]$  and  $[C_2]$  and the vector  $\{D\}$  can be written as:

$$\begin{aligned} [C_1] &= (C_{oA} + C_{oB}) \cos^2 \theta(t) \{e_1\} \{e_3\}^T, \\ [C_2] &= (C_{oA} + C_{oB}) \cos^2 \theta(t) \{e_2\} \{e_1\}^T, \\ \{D\} &= \frac{1}{2} (C_{oA} - C_{oB}) \sin 2\theta(t) \{e_1\}. \end{aligned} \quad (10)$$

Substituting Equation (10) into Equation (8) and solving for  $\{\ddot{x}(t)\}$  leads to

$$\{\ddot{x}(t)\} = -[M]^{-1}([C_s] + [C_1] + [C_2])\{\dot{x}(t)\} - [M]^{-1}[K]\{x(t)\} - [M]^{-1}\{D\}\dot{w}(t) - \{r\}\ddot{x}_g(t). \quad (11)$$

Defining four matrices  $[A_c]$ ,  $[B_c]$ ,  $[D_c]$ , and  $[E_c]$  as follows

$$\begin{aligned} [A_c] &= -[[M]^{-1}[K]], \\ [B_c] &= -[[M]^{-1}([C_s] + [C_1] + [C_2])], \\ [D_c] &= -[[M]^{-1}\{D\}], \\ [E_c] &= [\{r\}], \end{aligned}$$

and introducing the components of a state vector  $\{z_1(t)\} = \{x(t)\}$ ,  $\{z_2(t)\} = \{\dot{x}(t)\}$ , Equation (11) can be written as:

$$\{\dot{z}_2(t)\} = [A_c]\{z_1(t)\} + [B_c]\{z_2(t)\} + [D_c]\dot{w}(t) - [E_c]\dot{x}_g(t). \quad (12)$$

To obtain the equations of motion in the state-space form, Equation (12) is supplemented with the following identity:

$$\{\dot{z}_1(t)\} = \{z_2(t)\}. \quad (13)$$

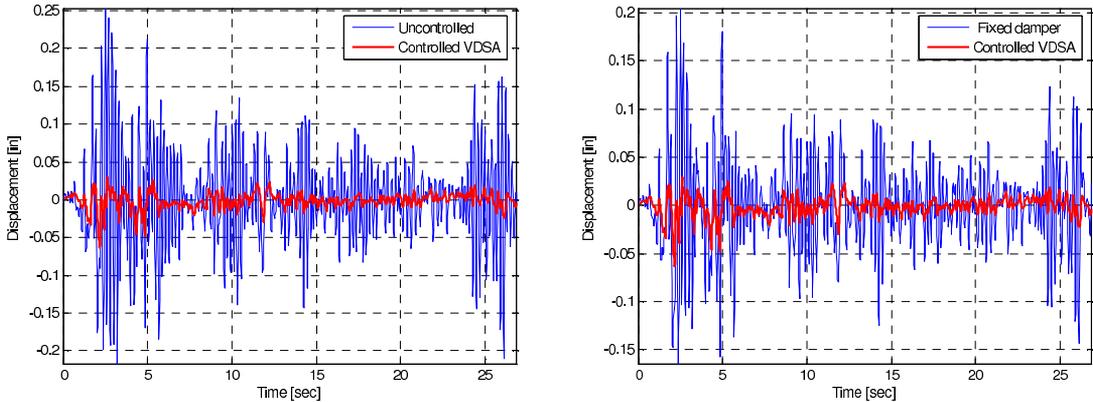
Equation (12) and Equation (13) can now be combined into the following equation of motion in the state-space:

$$\begin{Bmatrix} \dot{z}_1(t) \\ \dot{z}_2(t) \end{Bmatrix} = \begin{bmatrix} [O_c] & [I_c] \\ [A_c] & [B_c] \end{bmatrix} \begin{Bmatrix} z_1(t) \\ z_2(t) \end{Bmatrix} + \begin{bmatrix} [O_c] \\ [D_c] \end{bmatrix} \dot{w}(t) - \begin{bmatrix} [O_c] \\ [E_c] \end{bmatrix} \dot{x}_g(t).$$

The vertical position of the lower end of the VDSA device  $w(t)$  required in each instant of time, along with the damping coefficient for the coupled system, can be computed with Equation (7) using one of the two modified algorithms  $Qv$ .

### 5. Numerical examples

Two examples are presented in this paper to illustrate the effectiveness of the VDSA device in reducing the seismic response: an SDOF system and an MDOF structure. The response obtained by applying the



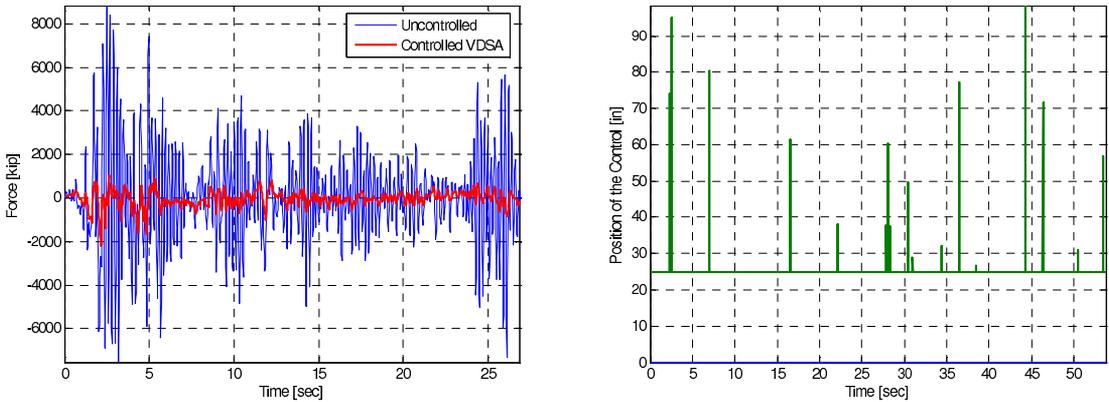
**Figure 4.** Left: relative displacement of the SDOF system due to the El Centro record for uncontrolled versus VDSA; right: fixed damper versus VDSA (closed-loop control).

closed-loop control modified algorithm  $Qv$  and the closed-open-loop control modified algorithm  $Qv$  are compared against the response of the uncontrolled structures. In addition, the response of the structures fitted with passive dampers is included in the comparisons. The structures are subjected to the horizontal component of three different earthquakes. First, the record of the well-known El Centro earthquake in the Imperial Valley, California on May 18, 1940, is considered. This record had a peak ground acceleration (PGA) of 0.348 g. Next, the record of the San Fernando, California earthquake of February 9, 1971, with a PGA of 1.007 g is used. Finally, the record of the Friuli, Italy earthquake of May 6, 1976, with a PGA of 0.4788 g is applied to the structures. The accelerations are sampled at equal time intervals of 0.02 sec.

**Example 1.** The first example is a SDOF frame with a weight of 13,630 kip and natural period of 0.20 sec. The damping coefficient of the dampers  $A$  and  $B$  of the VDSA device were 20 kip-sec/in and 10 kip-sec/in, respectively. The results obtained are compared with those of the uncontrolled structure with a damping ratio of 5%. The weighting matrix  $Q$  is selected as  $[I] \times 10^2$ , where  $[I]$  is the identity matrix. The matrices  $Qv$  and  $R$  become scalars with values equal to  $10^1$  and  $10^{-4}$ , respectively. The original damping ratio of the structure in which the VDSA system was installed was taken to be 2%. For this case of passive damping, the damping ratio was set equal to 30%, and the damping coefficient of the dampers is 40 kip-sec/in.

The response of the SDOF system to the base acceleration of the El Centro earthquake is presented first. [Figure 4](#) shows a comparison of the relative displacement time histories for the uncontrolled structure (left), the structure with fixed dampers, and the structure controlled with the VDSA device (right). Only the first twenty-seven seconds of the response are shown. [Figure 5](#) (left) presents the total shear force at the base of the structure as a function of time for the uncontrolled structure and the structure controlled with the VDSA device. The time variation of the position of the lower end of the VDSA device,  $w(t)$ , is presented in [Figure 5](#) (right). In this case the graph shows the variation of the position of the device for the full duration of the earthquake excitation.

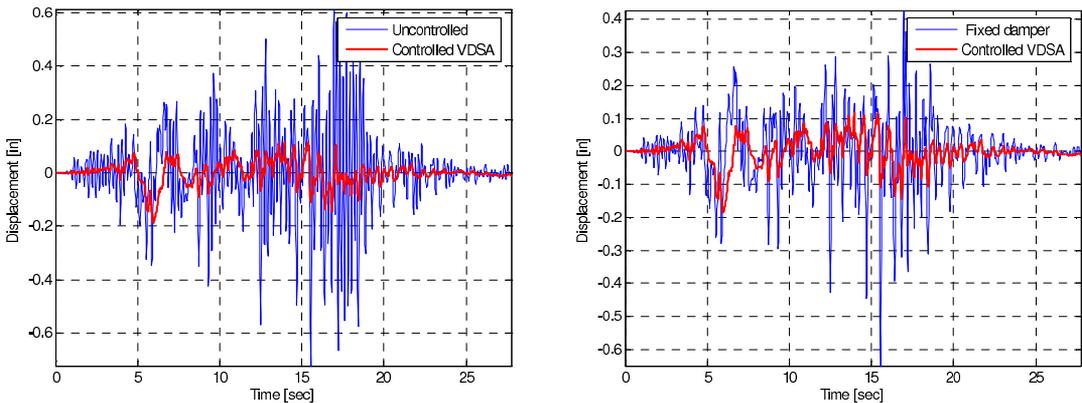
It can be noticed from [Figure 5](#) (right) that the displacement  $w(t)$  of the lower end of the VDSA device needs to change quite rapidly, actually in fractions of a second. This may pose a problem if a hydraulic actuator is used to push or pull the VDSA device because it may require an actuator with very



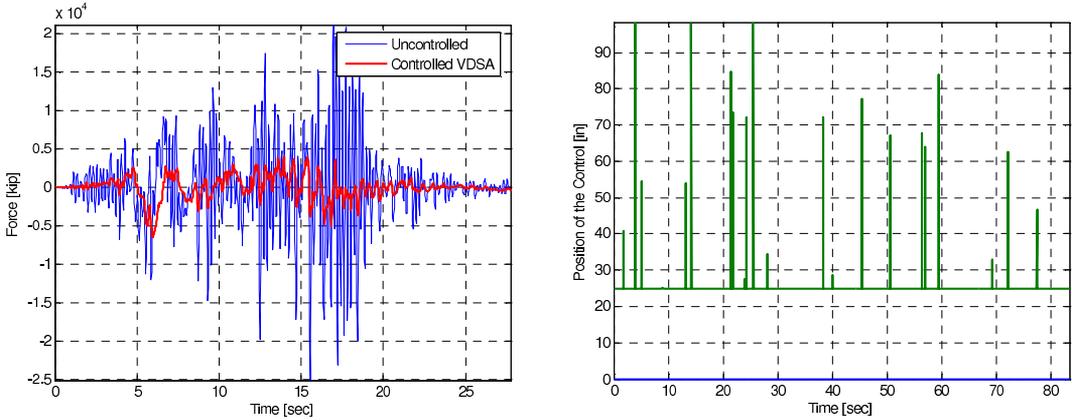
**Figure 5.** Left: uncontrolled versus VDSA-controlled base shear of the SDOF system. Right: variation of the position of the VDSA device for the El Centro record (closed-loop control).

high performance characteristics. In any case, for future work, and before an experimental verification of the proposed protective system is undertaken, one would ideally include a model of a nonideal actuator to study its effect on the response reduction.

The fact that the VDSA device continues to move even when the excitation diminishes (Figure 5 (right)) may be intriguing at first sight. The reason for this behavior is that both control algorithms try to minimize the response even if its magnitude is not large. In other words, during the strong motion part of the ground acceleration the semiactive control system reduces the response of the structure by about the same degree than during the final phase of the excitation. In theory, to avoid this behavior one could use in the definition of the performance index weighting matrices that vary with time. However, this will considerably increase the required computational time which may, in turn, create other problems due to the time lag between the response measurement and the actuator engagement. The effect on the



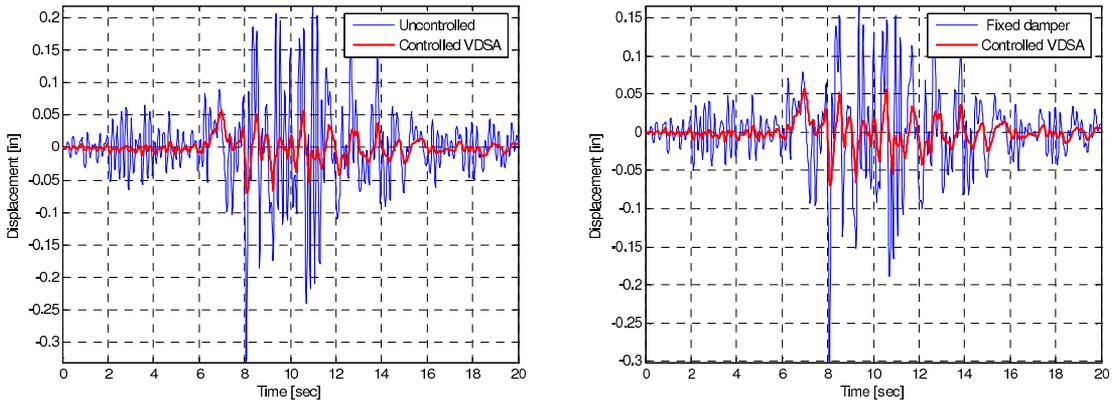
**Figure 6.** Relative displacement of the SDOF system for the San Fernando record for (left) uncontrolled versus VDSA, and (right) fixed damper versus VDSA (closed-loop control).



**Figure 7.** Left: uncontrolled versus VDSA-controlled base shear of the SDOF system. Right: variation of the position of the VDSA device for the San Fernando record (closed-loop control).

structure of the feature portrayed in Figure 5 (right) during low intensity seismic motions should be studied experimentally.

The previous response calculations were repeated with the San Fernando record. The responses compared are those obtained with the original (uncontrolled) structure, with fixed dampers and with the VDSA device. Figure 6 shows the time variation of the displacements for the three conditions whereas Figure 7 (left) displays the base shear time histories. The first forty-two seconds of the response is shown. Figure 7 (right) shows the variation of the height of the lower end of the device  $w(t)$  for the San Fernando earthquake. Similar observations to those made for the El Centro earthquake can also be repeated here regarding the nature of the time variation of  $w(t)$ .



**Figure 8.** Relative displacement of the SDOF system for the Friuli record for (left) uncontrolled versus VDSA, and (right) fixed damper versus VDSA (closed-loop control).

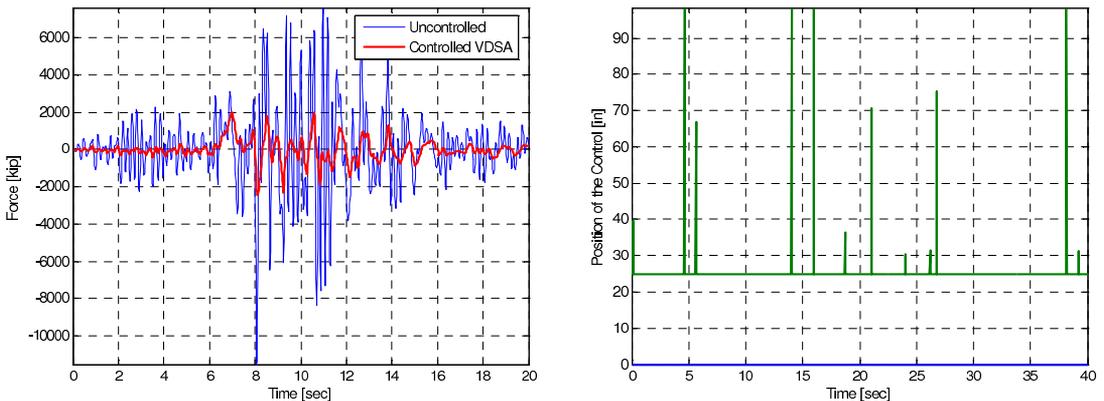
Earthquake	Displacement			Total base shear		
	Uncont.	Fixed damper	VDSA	Uncont.	Damper Fixed	VDSA
	[in]	[in]	[in]	[kip]	[kip]	[kip]
El Centro	0.2536	0.2049	0.0630	8828.0	7133.1	2192.1
San Fernando	0.7219	0.7072	0.1848	25132.0	24620.0	6435.3
Friuli	0.3315	0.3019	0.0696	11540.0	10509.0	2424.3

**Table 1.** Maximum response of the SDOF structure without control and with a passive and semiactive system (closed-loop control).

The next set of results corresponds to the 1976 Friuli accelerogram. Again, the responses compared are the relative displacement of the mass, and the sum of the shear forces in the columns, in both uncontrolled and controlled mode with fixed dampers and with the VDSA device. The results are presented in Figure 8 and 9 (left) for the first twenty seconds of the response. Figure 9 (right), shows the variation of the position of the VDSA device.

Table 1 shows a summary of the maximum responses obtained for the SDOF structure of Section 5 when the closed-loop modified algorithm  $Qv$  was used. Table 2 is similar to Table 1 but displays the controlled response using the closed-open-loop control algorithm. Both tables demonstrate the advantages of using the VDSA device. In addition, the tables show that both the closed-loop and closed-open-loop control modified algorithms  $Qv$  provide almost the same results. The fact that both algorithms yielded similar results coincides with the results observed in the area of active structural control whenever the closed-loop and closed-open loop control formulations are used.

**Example 2.** A six-story building was selected to show, via a numerical simulation, the implementation of the VDSA device and to illustrate its effectiveness in reducing the seismic response of a MDOF structure. For simplicity, the structure is modeled as a shear building with one DOF per floor (the lateral



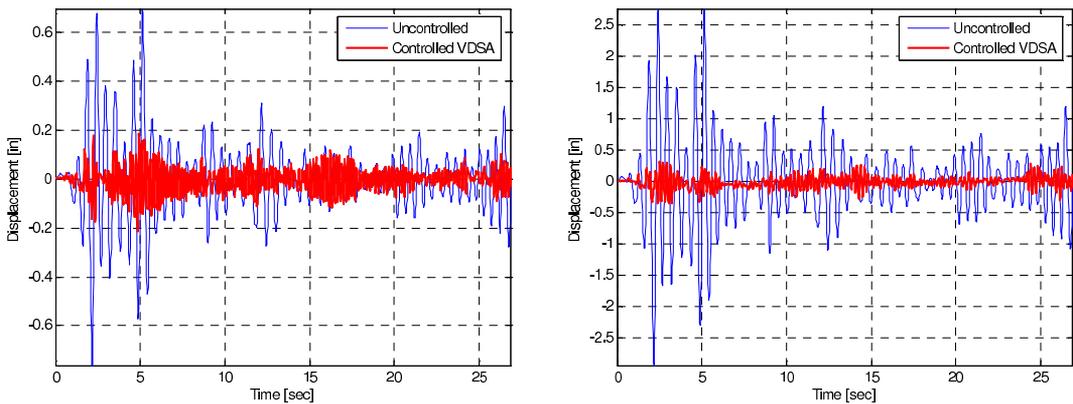
**Figure 9.** Left: uncontrolled versus VDSA-controlled base shear of the SDOF system. Right: variation of the position of the VDSA device for the Friuli record (closed-loop control).

Earthquake	Displacement			Total base shear		
	Uncont.	Fixed damper	VDSA	Uncont.	Fixed damper	VDSA
	[in]	[in]	[in]	[kip]	[kip]	[kip]
El Centro	0.2536	0.2049	0.0616	8828.0	7133.1	2144.5
San Fernando	0.7219	0.7072	0.1816	25132.0	24620.0	6322.2
Friuli	0.3315	0.3019	0.0678	11540.0	10509.0	2361.6

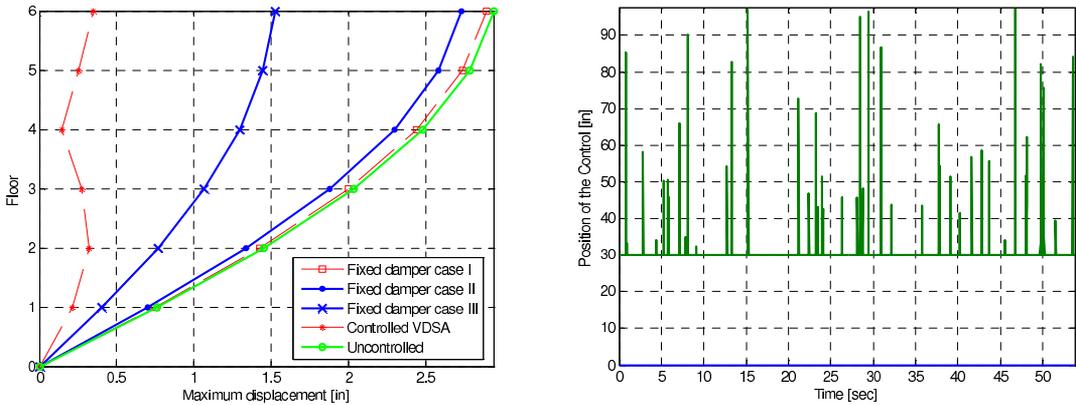
**Table 2.** Maximum response of the SDOF structure without control and with a passive and semiactive system (closed-open-loop control).

displacement). The total lateral stiffness coefficients of the columns are  $k_i = 5,315$  kip/in and the floor weights are  $W_i = 2,205$  kip. The damping ratio of the uncontrolled structure is assumed to be 5% for all the modes. For the case where the dampers are installed in a fixed position, the damping ratio provided by them is selected to be 30% for the first mode. The damping coefficients of the dampers of the VDSA device are 25 kip-sec/in and 10 kip-sec/in for the dampers  $A$  and  $B$ , respectively. The results obtained are compared with those obtained for the uncontrolled structure and also with the response calculated with fixed dampers. In the latter case three configurations, identified as I, II and III, were considered. Each corresponds to increasing number of fixed dampers: in case I, a single damper was installed at the first floor; in case II three dampers were placed on the three lower floors and in case III a damper was installed at each of the six floors. The VDSA device was assumed to be installed in the fourth floor. This position was found to be the best one by a simple trial and error process. For closed-loop and closed-open-loop control algorithms  $Qv$  the weighting matrices  $Q$  and  $Qv$  were selected as  $[I] \times 10^4$  and  $[I] \times 10^2$ , respectively, where  $[I]$  is an identity matrix. Matrix  $R$  is, in this case, a scalar with a value equal to  $10^{-1}$ .

The first result for the MDOF structure is the response to the ground acceleration due to the 1940 El Centro earthquake. The relative displacements computed for the uncontrolled structure are compared with a similar response quantity but for the structure controlled with the VDSA device. The time trace of



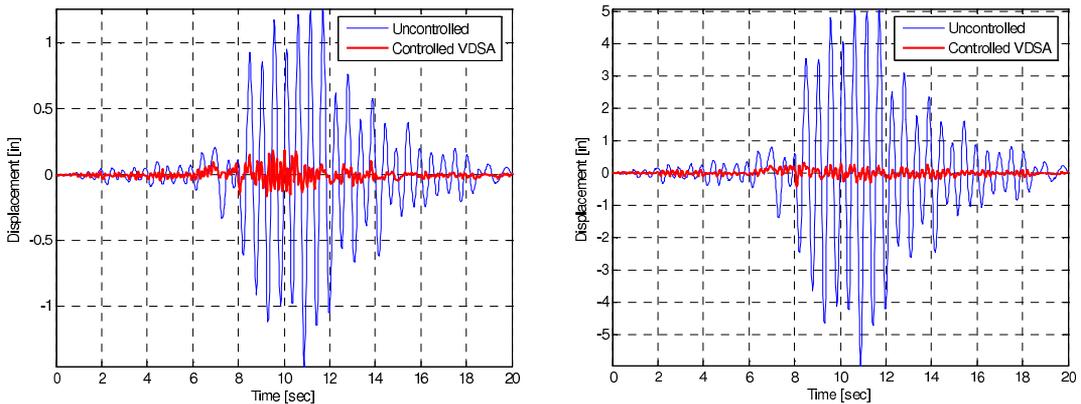
**Figure 10.** Relative displacements of the 6-story building for the El Centro record, (left) first floor, and (right) top floor — uncontrolled versus VDSA (closed-loop control).



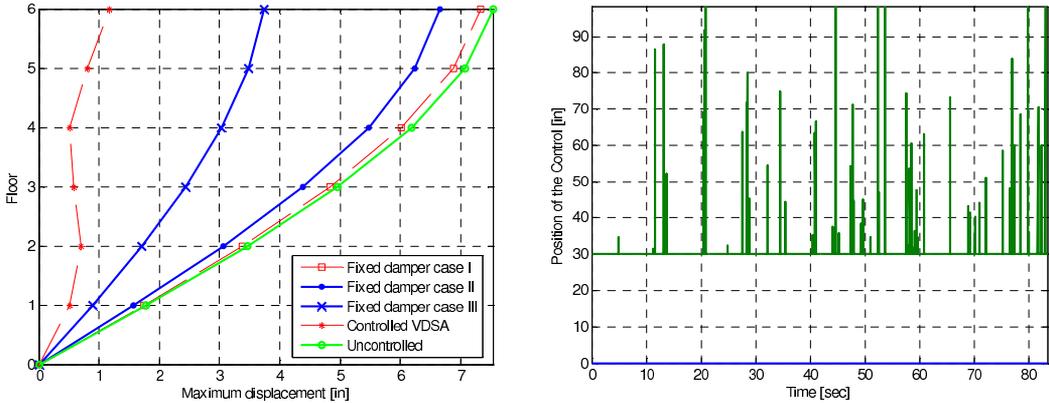
**Figure 11.** Left: maximum floor displacements for all cases. Right: variation of the position of the VDSA device — El Centro record (closed-loop control).

the relative displacements of the first and six floors are presented in Figure 10 (left and right, respectively). The maximum relative displacements in the six floors for all the cases considered are shown in Figure 11 (left). There are five cases considered: the original structure, the structure with a single VDSA device in the fourth floor, and the three fixed damper configurations I, II and III, previously described. Clearly, the response reduction achieved by the VDSA system is remarkable, even when compared to the case in which all floors are provided with viscous dampers at the maximum practical range. The variation of the control device position for the El Centro record is presented in Figure 11 (right).

The previous analyses were repeated for the San Fernando ground motion. Figure 12 displays the time variation of the displacement response for the first and top floor of the original structure and controlled with the VDSA device. The next set of results displayed in Figure 13 is the maximum relative displacements of the six floors for all cases studied (left) and the vertical position of the lower end of the device (right).



**Figure 12.** Relative displacements of the 6-story building for the San Fernando record, (left) first floor, and (right) top floor — uncontrolled versus VDSA (closed-loop control).

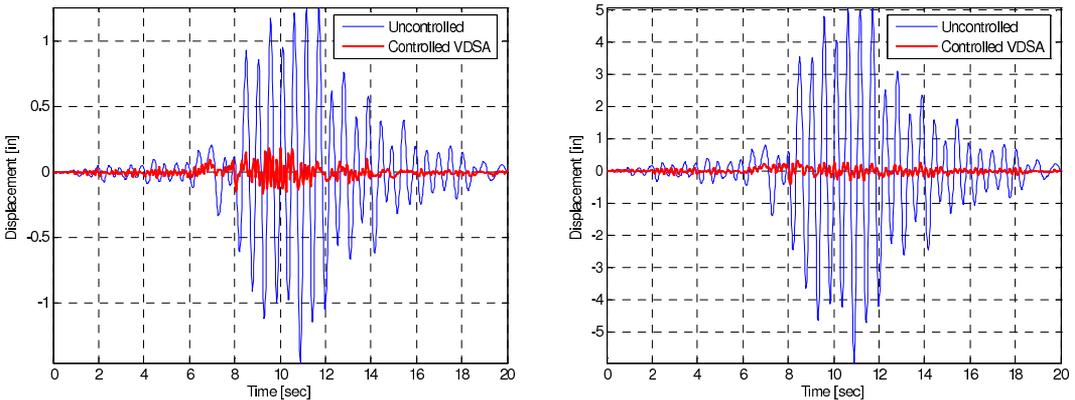


**Figure 13.** Left: maximum floor displacements for all cases. Right: variation of the position of the VDSA — San Fernando record (closed-loop control).

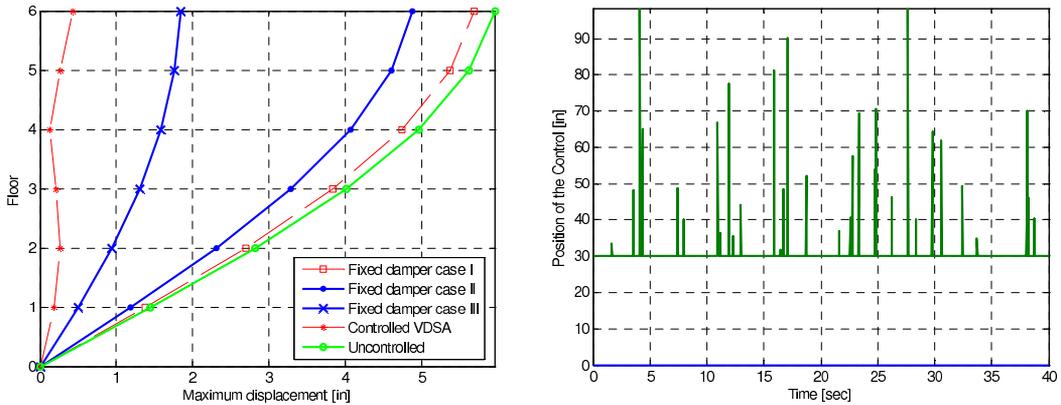
The last set of results corresponds to the response of the 6-story building subjected to the acceleration of the Friuli earthquake. Only the first twenty seconds of the response is shown. **Figure 14** displays the relative displacement time histories of the structure in uncontrolled mode (left) and controlled mode (right) with the variable dampers for the first and top floor. **Figure 15** shows the maximum relative displacements of the six floors for the five cases analyzed (left) and the variation of the position of the VDSA device (right).

Here also the maximum responses obtained using the closed-loop and closed-open-loop control algorithms were practically the same. Only small differences in the form that varies the position of the VDSA device were found [Cundumi 2005].

A summary of the response of the structure in three conditions: a) uncontrolled, b) fitted with fixed (passive) dampers using the configurations I, II and III, and c) controlled with the proposed semiactive device, are compared in Tables 3–5. These tables present the maximum relative displacement for all floors when the El Centro, San Fernando, and Friuli ground motions were applied at the base. It can be



**Figure 14.** Relative displacements of the 6-story building for the Friuli record, (left) first floor, and (right) top floor — uncontrolled versus VDSA (closed-loop control).



**Figure 15.** Left: maximum floor displacements for all cases. Right: variation of the position of the VDSA — Friuli record (closed-loop control).

noticed that the VDSA device is effective in reducing the relative displacements and shear forces. The results presented in this paper correspond to the VDSA device installed in the fourth floor where the best response reduction was obtained. However, although it is not presented here, comparable reductions were obtained with the VDSA device positioned in other floors. As expected, the best results obtained with passive control were for case III in which the dampers were installed in all six floors of the building. In this passive case the fixed dampers were assigned a damping coefficient equal to the maximum recommended practical limit. The reduction in the maximum displacements of the building with the VDSA device compared to the passive control (case III) ranges from 50–80%. When the maximum displacement reduction achieved with the proposed system is compared with the original structure, the decrease in the top floor response varies from 88–93%.

Finally, the effectiveness in the response reduction in MDOF structures with the VDSA device controlled by the modified closed-loop and closed-open-loop control algorithm  $Qv$  was observed to be the same. In other words, in terms of performance, there was no advantage in using one methodology over the other.

Floor	Displacement				
	Uncont. [in]	Fixed damper Case I [in]	Fixed damper Case II [in]	Fixed damper Case III [in]	VDSA [in]
6th	2.939	2.894	2.732	1.529	0.352
5th	2.786	2.742	2.586	1.448	0.252
4th	2.483	2.443	2.301	1.296	0.141
3rd	2.036	2.000	1.881	1.069	0.273
2nd	1.454	1.427	1.340	0.769	0.320
1st	0.762	0.747	0.700	0.405	0.215

**Table 3.** Maximum displacements of the 6-story building without control and with a passive and semiactive system for the El Centro record (closed-loop control).

Floor	Displacement				
	Uncont.	Fixed damper Case I	Fixed damper Case II	Fixed damper Case III	VDSA
	[in]	[in]	[in]	[in]	[in]
6th	7.528	7.321	6.650	3.731	1.178
5th	7.066	6.873	6.244	3.485	0.800
4th	6.185	6.018	5.467	3.035	0.519
3rd	4.957	4.823	4.380	2.432	0.585
2nd	3.463	3.371	3.057	1.715	0.701
1st	1.785	1.738	1.575	0.897	0.519

**Table 4.** Maximum displacements of the 6-story building without control and with a passive and semiactive system for the San Fernando record (closed-loop control).

Floor	Displacement				
	Uncont.	Fixed damper Case I	Fixed damper Case II	Fixed damper Case III	VDSA
	[in]	[in]	[in]	[in]	[in]
6th	5.951	5.689	4.877	1.850	0.428
5th	5.617	5.369	4.602	1.759	0.271
4th	4.960	4.741	4.063	1.579	0.132
3rd	4.010	3.832	3.283	1.307	0.210
2nd	2.813	2.688	2.304	0.946	0.269
1st	1.452	1.383	1.184	0.505	0.186

**Table 5.** Maximum displacements of the 6-story building without control and with a passive and semiactive system for the Friuli record (closed-loop control).

## 6. Conclusions

The results presented in Tables 1 and 2 for the SDOF structure, selected as the first example, indicate that the maximum relative displacements due to the El Centro, San Fernando, and Friuli accelerograms were reduced by 75.2%, 74.4%, and 79.0%, respectively, compared to the case when the structure had its original 5% damping ratio. The results were obtained by using the modified closed-loop control algorithm  $Qv$ . When the modified closed-open-loop control algorithm  $Qv$  was used to define the position of the VDSA device, the reductions were 75.7%, 74.8% and 79.6%, that is, practically the same. When compared to the case in which the SDOF system was fitted with fixed dampers, the peak relative displacements were reduced by 19.2%, 2.0%, and 8.9% for the El Centro, San Fernando, and Friuli earthquakes.

In another example a 6-story shear building was used to numerically examine the performance of the proposed semiactive dampers. The maximum relative displacements of all floors for the three seismic records were presented in Table 3, 4 and 5, respectively. The reductions obtained with the VDSA device in the top floor displacements were 88.0% (for El Centro), 84.4% (for San Fernando) and 92.8% (for Friuli). Both algorithms led to the same results. To compare the effectiveness of the VDSA device with viscous dampers in a fixed position, three configurations were selected for the latter case. The

best results were observed when the structure had passive dampers installed in all floors (a configuration identified as Case III). In this case the reduction in the peak displacement at the top floor was 48%, 50.4%, and 68.9% for the El Centro, San Fernando and Friuli accelerograms. The reduction in the displacements of the lower floors is not as dramatic as in the top floor. However, the proposed device was capable of achieving a notable decrease even for the lowest floor. For example, the reductions in the peak displacements of the first floor obtained with the VDSA device were 71.3%, 70.9%, and 87.2% for El Centro, San Fernando and Friuli record, respectively. These percentages should be compared with the 46.8%, 49.8%, and 65.2% reduction obtained by installing fixed dampers in all the floors.

The objective of this paper was to introduce the concept of a novel variable damping device in which the damping provided to the structure can be changed by varying the orientation of two dampers with constant coefficients. A preliminary verification of the performance of the proposed device was done via numerical simulations. However, it is recognized that there are still many issues that need to be studied analytically and even more importantly, experimentally. For instance, the actuator was assumed to be ideal, that is, no actuator dynamics were included in the simulations. The final corroboration of the concept must be done through a thorough experimental program.

## 7. Acknowledgment

The authors would like to thank the reviewers of the paper for their valuable comments.

## References

- [Cundumi 2005] O. Cundumi, *A variable damping semiactive device for control of the seismic response of buildings*, Ph.D. Dissertation, the University of Puerto Rico at Mayagüez, Department of Civil Engineering, Mayagüez, Puerto Rico, 2005.
- [Cundumi and Suárez 2006a] O. Cundumi and L. E. Suárez, "A new variable damping semi-active (VDSA) device for seismic response reduction of civil structures", in *The IX pan american congress of applied mechanics (PACAM IX)*, Mérida, Mexico, January 2–6 2006a.
- [Cundumi and Suárez 2006b] O. Cundumi and L. Suárez, "Seismic response reduction using semi-active control with a new variable damping device and modified algorithm  $Qv$ ", in *Proceedings of the eight U.S. national conference on earthquake engineering*, San Francisco, California, April 18–22 2006b.
- [Gluck et al. 2000] J. Gluck, Y. Ribakov, and A. N. Dancygier, "Selective control of base-isolated structures with CS dampers", *Earthq. Spectra* **16**:3 (2000), 593–606.
- [Hrovat et al. 1983] D. Hrovat, P. Barak, and M. Rabins, "Semi-active versus passive or active tuned mass dampers for structural control", *J. Eng. Mech., ASCE* **109**:3 (1983), 691–705.
- [Kawashima et al. 1992] K. Kawashima, S. Unjoh, H. Iida, and N. Niwa, "Effectiveness of the variable damper for reducing seismic response of highway bridges", pp. 479–493 in *Proceedings of the 2nd U.S.-Japan workshop on earthquake protective systems for bridges*, PWRI, Tsukuba Science City, Japan, 1992.
- [Kobori et al. 1993] T. Kobori, M. Takahashi, T. Nasu, N. Niwa, and K. Ogasawara, "Seismic response controlled structure with active variable stiffness system", *Earthquake & Eng. Struc. Dyn.* **22**:9 (1993), 925–941.
- [Kurata et al. 1999] N. Kurata, T. Kobori, M. Takahashi, N. Niwa, and H. Midorikawa, "Actual seismic response controlled building with semi-active damper system", *Earthquake. & Eng. Struc. Dyn.* **28**:11 (1999), 1427–1447.
- [Kurata et al. 2000] N. Kurata, T. Kobori, M. Takahashi, T. Ishibashi, N. Niwa, J. Tagami, and H. Midorikawa, "Forced vibration test of a building with semi-active damper system", *Earthquake. & Eng. Struc. Dyn.* **29**:5 (2000), 629–645.
- [Nagarajaiah and Mate 1998] S. Nagarajaiah and D. Mate, "Semi-active control of continuously variable stiffness system", pp. 397–405 in *Proceedings of the 2nd world conference on structural control*, Kyoto, Japan, 1998.

- [Patten et al. 1993] W. N. Patten, R. L. Sack, W. Yen, C. Mo, and H. C. Wu, “Seismic motion control using semi-active hydraulic force actuators”, pp. 727–736 in *Proceedings of the ATC-17-1 seminar on seismic isolation, passive energy dissipation, and active control*, Redwood City, California, 1993.
- [Patten et al. 1996] W. N. Patten, R. L. Sack, and Q. He, “Controlled semi-active hydraulic vibration absorber for bridges”, *J. Struct. Eng., ASCE* **122**:2 (1996), 187–192.
- [Sack et al. 1994] R. L. Sack, C. C. Kuo, H. C. Wu, L. Liu, and W. N. Patten, “Seismic motion control via semi-active hydraulic actuators”, pp. 311–320 in *Proceedings of the 5th national conference of earthquake engineering*, El Cerrito, California, 1994.
- [Sadeck and Mohraz 1998] F. Sadeck and B. Mohraz, “Semiaactive control algorithms for structures with variable dampers”, *J. Eng. Mech.* **124**:9 (1998), 981–990.
- [Soong 1990] T. T. Soong, *Active structural control: theory and practice*, John Wiley & Sons, New York, 1990.
- [Symans and Constantinou 1997] M. Symans and M. C. Constantinou, “Seismic testing of a building structure with a semi-active fluid damper control system”, *Earthquake. & Eng. Struc. Dyn.* **26**:7 (1997), 759–777.
- [Yang et al. 1987] J. N. Yang, A. Akbarpour, and P. Ghaemmaghami, “New optimal control algorithms for structural control”, *J. Eng. Mech., ASCE* **113**:9 (1987), 1369–1386.

Received 22 May 2007. Accepted 23 May 2007.

ORLANDO CUNDUMI: [ocundumi@caribbean.edu](mailto:ocundumi@caribbean.edu)

Department of Engineering, Caribbean University, Bayamon Campus, P.O. Box 493, Bayamon, PR 00960, United States

LUIS E SUÁREZ: [lsuarez@uprm.edu](mailto:lsuarez@uprm.edu)

Civil Engineering Department, University of Puerto Rico, P.O. Box 9041, Mayaguez, PR 00681-9041, United States

(continued from back cover)

Natural convection fluid flow and heat transfer in porous media	<b>E. BÁEZ AND A. NICOLÁS</b>	<b>1571</b>
Stark ladder resonances in acoustic waveguides	<b>G. MONSIVAIS AND R. ESQUIVEL-SIRVENT</b>	<b>1585</b>
Tension buckling in multilayer elastomeric isolation bearings	<b>J. M. KELLY AND S. M. TAKHIROV</b>	<b>1591</b>
Representative volume element and effective elastic properties of open cell foam materials with random microstructures	<b>S. KANAUN AND O. TKACHENKO</b>	<b>1607</b>
Elastic Wannier–Stark ladders in torsional waves	<b>G. MONSIVAIS, R. A. MÉNDEZ-SÁNCHEZ, A. DÍAZ DE ANDA, J. FLORES, L. GUTIÉRREZ AND A. MORALES</b>	<b>1629</b>
A new variable damping semiactive device for seismic response reduction of civil structures	<b>O. CUNDUMI AND L. E SUÁREZ</b>	<b>1639</b>

# Journal of Mechanics of Materials and Structures

Volume 2, Nº 8      October 2007

---

Special issue

## Ninth Pan American Congress of Applied Mechanics

- Preface**      J. W. EISCHEN AND G. MONSIVAIS 1395
- Optimization of a satellite with composite materials**  
J. A. C. AMBRÓSIO, M. AUGUSTA NETO AND R.ÉRIO PEREIRA LEAL 1397
- Geometric analysis of the dynamics of a double pendulum**  
J. AWREJCEWICZ AND D. SENDKOWSKI 1421
- Numerical simulation of granular materials in a rotating tumbler**      H. TAPIA-MCCLUNG 1431
- Density measurements in a supersonic jet**  
C. ELIZABETH STERN, J. MANUEL ALVARADO AND C. AGUILAR 1437
- Frequency and spatial response of basilar membrane vibration in a three-dimensional gerbil cochlear model**  
Y. YOON, S. PURIA AND C. R. STEELE 1449
- Flushing of the port of Ensenada using a SIBEO wave-driven seawater pump**  
S. PETER CZITROM, C. CORONADO AND I. NUÑEZ 1459
- Symmetry analysis of extreme areal Poisson's ratio in anisotropic crystals**  
L. WHEELER AND C. Y. GUO 1471
- A three dimensional contact model for soil-pipe interaction**  
N. P. RUBIO, D. ROEHL AND C. ROMANEL 1501
- Calculation of inertial properties of the malleus-incus complex from micro-CT imaging**  
J. H. SIM, S. PURIA AND C. R. STEELE 1515
- Electroelastic intensification and domain switching near a plane strain crack in a rectangular piezoelectric material**  
Y. SHINDO, F. NARITA AND F. SAITO 1525
- Active control schemes based on the linearized Tschauner–Hempel equations to maintain the separation distance constraints for the NASA benchmark Tetrahedron Constellation**  
P. A. CAPÓ-LUGO AND P. M. BAINUM 1541
- Evaluation of effective material properties of randomly distributed short cylindrical fiber composites using a numerical homogenization technique**  
H. BERGER, S. KARI, U. GABBERT, R. RODRIGUEZ-RAMOS,  
J. BRAVO-CASTILLERO AND R. GUINOVART-DIAZ 1561

(Continued inside back cover)