Mathematical Problems in Engineering

Theory, Methods, and Applications

Models, Methods, and Applications of Dynamics and Control in Engineering Sciences: State of the Art

Models, Methods, and Applications of Dynamics and Control in Engineering Sciences: State of the Art

Guest Editors: José Manoel Balthazar, Paulo Batista Gonçalves, Stefano Lenci, and Yuri Vladimirovich Mikhlin

Editor-in-Chief

Contents

Editorial

Models, Methods, and Applications of Dynamics and Control in Engineering Sciences: State of the Art

Jose Manoel Balthazar,[1] Paulo Batista Gonçalves,[2] Stefano Lenci,[3] and Yuri Vladimirovich Mikhlin[4]

[1] *Universidade Estadual Paulista (UNESP), Rio Claro, SP, Brazil*

[2] *Catholic University, Rio de Janeiro, Brazil*

[3] *Polytechnic University of Marche, Ancona, Italy*

[4] *National Technical University, Kharkov, Ukraine*

Correspondence should be addressed to Jose Manoel Balthazar, jmbaltha@rc.unesp.br

The goal of this special issue of Mathematical Problems in Engineering (MPE) is to present papers, containing complete reviews on modern dynamics models, available in the current literature, classify them, and discuss their applications and limitations. It was assumed to consider papers containing a discussion on appropriate models and control criteria and on various applications on Engineering and Sciences and suggest directions for further works. The critical reviews, a description of open problems, and future developments were hailed. 93 papers were presented for this issue, and 21 of them were accepted. There are 9 review papers and 12 research articles. In particular, there are reviews on the feedback designs for linear descriptors, modeling of ship dynamics, continuous models for discrete media, applications of fractional calculus in engineering, nonlinear fault tolerant control, simulation of a target tracking system with backlash compensation, dynamics of furuta's pendulum, and identification approaches for discrete event systems. One review is devoted to advances in structural control of civil engineering in China. There are 8 articles which are devoted to investigations in nonlinear dynamics; 8 articles are devoted to the control problems; 4 articles treat numerical approaches and problems of simulation. The majority of accepted articles contains modeling, as well as analytical and/or numerical investigations. The contents of the 9 reviews papers are briefly summarized in the following.

(1) *"Modeling of ship roll dynamics and its coupling with heave and pitch,"* by R. A. Ibrahim and I. M. Grace. An overview of the roll dynamic stability under random sea waves has been presented in terms of the sample stability condition and response statistical moments.

(2) "*Some applications of fractional calculus in engineering,*" by J. A. T. Machado, et al. This article presents several case studies on the implementation of FC-based models and control systems, with the advantages of using the FC theory in different areas of science and engineering being demonstrated.

(3) "*Improved continuous models for discrete media,*" by I. V. Andrianov et al. This paper focuses on continuous models derived from a discrete microstructure. Various continualization procedures that take into account the nonlocal interaction between variables of the discrete media are analysed.

(4) "*Furuta's pendulum: a conservative nonlinear model for theory and practise,*" by J. Á. Acosta. A detailed dynamical model of the available laboratory Furuta's pendulum is provided. A survey of all the nonlinear controllers designed with that quasiconservative model and successfully tested through experiments is also reported.

(5) "*A comparative analysis of recent identification approaches for discrete event systems,*" by A. P. Estrada-Vargas et al. This paper overviews identification approaches of DES found in the literature and then it provides a comparative analysis of three recent and innovative contributions.

(6) "*A survey of some recent results on nonlinear fault tolerant control,*" by M. Benosman. This paper presents a survey of some of the results obtained last years on nonlinear Fault tolerant control (NFTC).

(7) "*Advances in structural control in civil engineering in china,*" by Hongnan Li and Linsheng Huo. The aim of this paper is to review a state of the art of researches and applications of structural control in civil engineering in China. It includes the passive control, active control, hybrid control and semiactive control.

(8) "*Time domain simulation of a target tracking system with backlash compensation,*" by M. Gruzman et al. This paper presents the modeling and control of a target tracking system assembled into a moving body. Special attention is given to the problems caused by backlash in gear transmission.

(9) "*Robust state-derivative feedback LMI-based designs for linear descriptor systems,*" by F. A. Faria et al. Necessary and sufficient stability conditions based on LMI for state-derivative feedback of linear descriptor systems are presented. The proposed methods are LMI-based designs that, when feasible, can be efficiently solved by convex programming techniques.

The contents on the research articles are somehow more spread and are summarized in the following.

(10) "*Applied model-based analysis and synthesis for the dynamics, guidance, and control of an autonomous undersea vehicle,*" by Kangsoo Kim and Tamaki Ura. In this article, a model-based analysis and synthesis to the following three research fields in AUV(Autonomous Undersea Vehicle) design and development have been presented: dynamic system modelling of an AUV, motion control design and tracking control application, and an optimal guidance of an AUV in environmental disturbances.

(11) "*Control of limit cycle oscillations of a two-dimensional aeroelastic system,*" by M. Ghommem et al. Linear and nonlinear controls are implemented on a rigid airfoil undergoing pitch and plunge motions. The method of multiple scales is applied to the governing system of equations to derive the normal form of the Hopf bifurcation near the flutter onset. On the other hand, the nonlinear control can be efficiently implemented to convert subcritical to supercritical Hopf bifurcations and to significantly reduce LCO amplitudes.

(12) "*Investigation of aeroelastic mode distribution for aircraft wing model in subsonic air flow,*" by Marianna A. Shubov et al. The presented paper is concerned with numerical investigation of two problems arising in the area of theoretical aeroelasticity. Namely, it has been shown that analytical formulas representing the asymptotical distribution of aeroelastic modes for a specific aircraft wing model can be used by practitioners.

(13) "*Nonlinear and dynamic aerodynamic models for commercial transport aircraft with adverse weather effects,*" by Ray C. Chang et al. This paper presents a numerical modeling method based on a fuzzy-logic algorithm which is presented to estimate the aerodynamic models for a twin-jet transport by using the flight data from the flight data recorder (FDR).

(14) "*Limiting phase trajectories and resonance energy transfer in a system of two coupled oscillators,*" by L. I. Manevitch et al. In this paper, the authors have extended the concept of the limiting phase trajectories (LPTs) to dissipative oscillatory systems. Using this concept, they have constructed an approximate solution describing the maximum energy exchange between coupled oscillators.

(15) "*A wavelet interpolation galerkin method for the simulation of MEMS devices under the effect of squeeze film damping,*" by Pu Li and Yuming Fang. This paper presents a new wavelet interpolation Galerkin method for the numerical simulation of MEMS devices under the effect of squeeze film damping. Simulation results show that the results of the wavelet interpolation Galerkin method match the experimental data better than that of the finite difference method by about 10%.

(16) "*Modeling and fuzzy PDC control and its application to an oscillatory TLP structure,*" by Cheng-Wu Chen. An analytical solution is derived to describe the wave-induced flow field and surge motion of a deformable platform structure controlled with fuzzy controllers in an oceanic environment. In the controller design procedure, a parallel distributed compensation (PDC) scheme is utilized to construct a global fuzzy logic controller by blending all local state feedback controllers.

(17) "*Adaptive fuzzy tracking control for a permanent magnet synchronous motor via backstepping approach,*" by Jinpeng Yu et al. Based on adaptive fuzzy control approach and backstepping technique, an adaptive fuzzy control scheme is proposed to control a permanent magnet synchronous motor. The proposed controllers guarantee that the tracking error converges to a small neighborhood of the origin and all the closed-loop signals are bounded. Simulation results illustrate an effectiveness of the presented method.

(18) "*Design for motor controller in hybrid electric vehicle based on vector frequency conversion technology,*" by Jing Lian et al. This paper carries on the research of motor control technology of application HEV, designs HEV motor controller based on vector frequency conversion technology, and completes the controller's software and

hardware design. In accordance with the characteristics of the HEV electromagnetic interference source, propagation path, and CAN bus communication, this paper proposes the corresponding antijamming measures, which greatly improves the antijamming ability of the CAN bus system and controller working stability.

(19) "*Biologically inspired robotic arm control using an artificial neural oscillator,*" by Woosung Yang et al. The authors have presented a control scheme for technically achieving a biologically inspired self-adapting robotic motion. In contrast to existing works that were only capable of rhythmic pattern generation for simple tasks, the presented approach allows the robot arm to precisely trace a trajectory correctly through entrainment. With this, the proposed method is verified through more complex behaviors of the real robot arm under unknown environmental changes.

(20) "*Shape sensitivity analysis in flow models using a finite-difference approach,*" by I. Akhtar et al. The authors consider flows past an elliptic cylinder with varying thickness ratios. Shape sensitivities (derivatives of flow variables with respect to thickness ratio) computed by finite-difference approximations are used to compute the POD sensitivity vectors. Numerical studies test shows the accuracy of the new bases to represent flow solutions over a range of parameter values.

(21) "*Investigations on actuator dynamics through theoretical and finite element approach,*" by Somashekhar S. Hiremath and M. Singaperumal. This paper gives a new approach for modeling the fluid-structure interaction of servovalve component-actuator. The analyzed valve is a precision flow control valve-jet pipe electrohydraulic servovalve. The positioning of an actuator depends upon the flow rate from control ports, in turn depends on the spool position. Theoretical investigation is made for No-Load condition and Load condition for an actuator. These are used in finite element modeling of an actuator. The fluid-structure-interaction (FSI) is established between the piston and the fluid cavities at the piston end. The fluid cavities were modeled with special purpose hydrostatic fluid elements while the piston is modeled with brick elements.

As a final remark, we want to thank all the authors for their contributions and we hope that the present special issue can be useful in the development of the research in the area on nonlinear dynamics and control.

Jose Manoel Balthazar
Paulo Batista Gonçalves
Stefano Lenci
Yuri Vladimirovich Mikhlin

Review Article

Modeling of Ship Roll Dynamics and Its Coupling with Heave and Pitch

R. A. Ibrahim and I. M. Grace

Department of Mechanical Engineering, Wayne State University, Detroit, MI 48202, USA

Correspondence should be addressed to R. A. Ibrahim, ibrahim@eng.wayne.edu

In order to study the dynamic behavior of ships navigating in severe environmental conditions it is imperative to develop their governing equations of motion taking into account the inherent nonlinearity of large-amplitude ship motion. The purpose of this paper is to present the coupled nonlinear equations of motion in heave, roll, and pitch based on physical grounds. The ingredients of the formulation are comprised of three main components. These are the inertia forces and moments, restoring forces and moments, and damping forces and moments with an emphasis to the roll damping moment. In the formulation of the restoring forces and moments, the influence of large-amplitude ship motions will be considered together with ocean wave loads. The special cases of coupled roll-pitch and purely roll equations of motion are obtained from the general formulation. The paper includes an assessment of roll stochastic stability and probabilistic approaches used to estimate the probability of capsizing and parameter identification.

1. Introduction

Generally, ships can experience three types of displacement motions (heave, sway or drift, and surge) and three angular motions (yaw, pitch, and roll) as shown in Figure 1. The general equations of motion have been developed either by using Lagrange's equation (see, e.g., [1–4]) or by using Newton's second law (see, e.g., [5–7]). In order to derive the hydrostatic and hydrodynamic forces and moments acting on the ship, two approaches have been used in the literature. The first approach utilizes a mathematical development based on a Taylor expansion of the force function (see, e.g., [8–12]). The second group employs the integration of hydrodynamic pressure acting on the ship's wetted surface to derive the external forces and moments (see, e.g., [13–18]). Stability against capsizing in heavy seas is one of the fundamental requirements in ship design. Capsizing is related to the extreme motion of the ship and waves. Of the six motions of the ship, the roll oscillation is the most critical motion that can lead to the ship capsizing. For small angles of roll motions, the response of ships can be described by a linear equation. However, as the amplitude of oscillation increases,
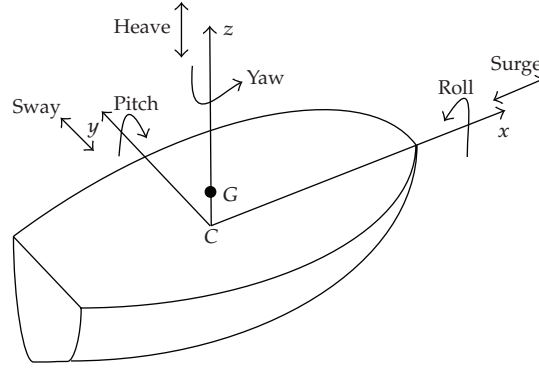
Figure 1: Ship schematic diagram showing the six degrees of freedom.

nonlinear effects come into play. Nonlinearity can magnify small variations in excitation to the point where the restoring force contributes to capsizing. The nonlinearity is due to the nature of restoring moment and damping. The environmental loadings are nonlinear and beyond the control of the designer. The nonlinearity of the restoring moment depends on the shape of the righting arm diagram.

Abkowitz [19] presented a significant development of the forces acting on a ship in surge, sway, and yaw motions. He used Taylor series expansions of the hydrodynamic forces about a forward cruising speed. The formulation resulted in an unlimited number of parameters and can model forces to an arbitrary degree of accuracy. Thus, it can be reduced to linear and extended to nonlinear equations of motion. Later, Abkowitz [20, 21], Hwang [22], and Källström and Åström [23] provided different approaches to estimate the coefficients of these models. Son and Nomoto [24] extended the work of Abkowitz [19] to include ship roll motion in deriving the forces and moments acting on the ship. Ross [25] developed the nonlinear equations of motion of a ship maneuvering through waves using Kirchhoff's [26] convolution integral formulation of the added mass. Kirchhoff's [26] equations are a set of relations used to obtain the equations of motion from the derivatives of the system kinetic energy. They are special cases of the Euler-Lagrange equations. The derived equations also give the Coriolis and centripetal forces [27, 28].

Rong [29] considered some problems of weak and strong nonlinear sea loads on floating marine structures. The weak nonlinear problem considers hydrodynamic loads on marine structures due to wave-current-body interaction. The strong nonlinear problem considers slamming loads acting on conventional and high-speed vessels. Theoretical and numerical methods to analyze wave-current interaction effects on large-volume structure were developed. The theory is based on matching a local solution to a far-field solution. It is known that large-amplitude ship motions result in strongly nonlinear, even chaotic behavior [30]. The current trends toward high-speed and unique hull-form vessels in commercial and military applications have broadened the need for robust mathematical approaches to study the dynamics of these innovative ships.

Various models of roll motion containing nonlinear terms in damping and restoring moments have been studied by many researchers [31–33]. Bass and Haddara [34, 35] considered various forms for the roll damping moment and introduced two techniques to identify the parameters of the various models together with a methodology for their

evaluation. Taylan [36] demonstrated that different nonlinear damping and restoring moment formulations reported in the literature have resulted in completely different roll amplitudes, and further yielded different ship stability characteristics. Since ship capsizing is strongly dependent on the magnitude of roll motions, an accurate estimation of roll damping is crucial to the prediction of the ship motion responses. Moreover, the designer should consider the influence of waves on roll damping, especially nonlinear roll damping of large-amplitude roll motion, and subsequently on ship stability.

Different models for the damping moment introduced in the equation of roll motion were proposed by Dalzell [37], Cardo et al. [38], and Mathisen and Price [39]. They contain linear-quadratic or linear-cubic terms in the angular roll velocity. El-Bassiouny [40] studied the dynamic behavior of ships roll motion by considering different forms of damping moments consisting of the linear term associated with radiation, viscous damping, and a cubic term due to frictional resistance and eddies behind bilge keels and hard bilge corners.

This paper presents the derivation of the equations of motion based on physical grounds. The equations of motion will then be simplified to consider the roll-pitch coupling, which is very critical in studying the problem of ship capsizing. It begins with a basic background and terminology commonly used in Marine Engineering. This is followed by considering the hydrostatics of ships in calm water and the corresponding contribution due to sea waves. An account of nonlinear damping in ship roll oscillation will be made based on the main results reported in the literature. The paper includes an overview of ship roll dynamic stability and its stochastic modes, probability methods used in estimating ship capsizing and parameter identification.

2. Background and Terminology

One needs to be familiar with naval architecture terminology. This includes key stability terms that are used in the design and analysis of navigation vessels and their structure components. A list of the main terms is provided in the appendix. Those terms described in this section are written in italics. The purpose of this section is to introduce the fundamental concept of ship roll hydrostatic stability.

A floating ship displaces a volume of water whose weight is equal to the weight of the ship. The ship will be buoyed up by a force equal to the weight of the displaced water. The *metacenter M* shown in Figure 2(a) is the point through which the *buoyant* forces act at small angles of *list*. At these small angles the center of buoyancy tends to follow an arc subtended by the *metacentric* radius $\overline{BM}$, which is the distance between the *metacenter* and the *center of buoyancy B*. As the vessels' *draft* changes so does the *metacenter* moving up with the center of buoyancy when the draft increases and vice versa when the draft decreases. For small angle stability it is assumed that the *metacenter* does not move.

The *center of Buoyancy B* is the point through which the buoyant forces act on the wetted surface of the *hull*. The position of the *center of buoyancy* changes depending on the attitude of the vessel in the water. As the vessel increases or reduces its *draft* (drawing or pulling), its *center of buoyancy* moves up or down, respectively, caused by a change in the water displaced. As the vessel *lists* the *center of buoyancy* moves in a direction governed by the changing shape of the submerged part of the *hull* as demonstrated in Figure 2(b). For small angles, the *center of buoyancy* moves towards the side of the ship that is becoming more submerged. This is true for small angle stability and for vessels with sufficient *freeboard*. When the water line reaches and moves above the main deck level a relatively smaller volume of
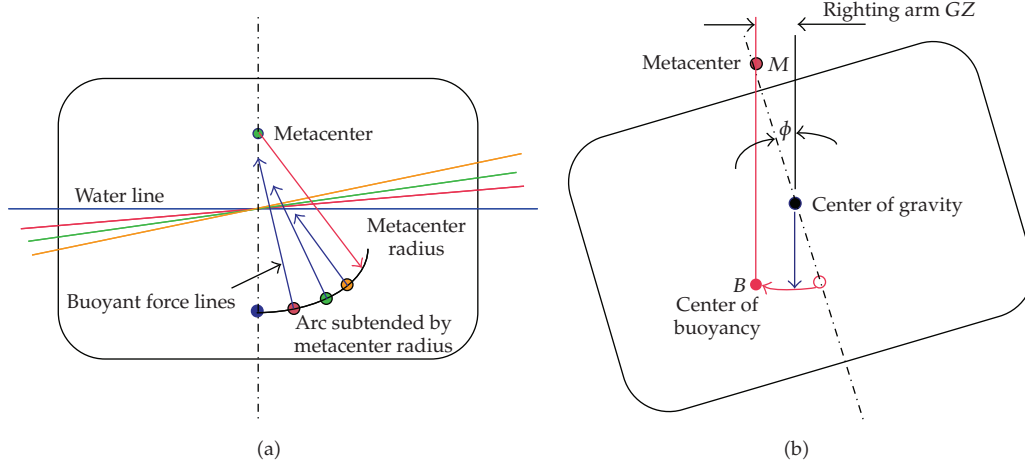
Figure 2: (a) Possible locations of the metacenter and (b) the righting arm.

the *hull* is submerged on the lower side for every centimeter movement as the water moves up the deck. The *center buoyancy* will now begin to move back towards the centerline.

As a vessel rolls its *center of buoyancy* moves off the centerline. The center of gravity, however, remains on the centerline. For small roll angles up to 10°, depending on *hull* geometry, the *righting arm* $\overline{GZ}$ is

$$\overline{GZ} = \overline{GM} \sin \phi. \tag{2.1}$$

It can be seen that the greater the *metacenter* height the greater the *righting arm* and therefore the greater the force restoring the vessel (*righting moment*) to the upright position one. When the *metacenter* is at or very near the centre of gravity then it is possible for the vessel to have a permanent *list* due to the lack of an adequate righting arm. Note that this may occur during loading operations. A worst case occurs when the *metacenter* is located substantially below the center of gravity as shown in Figure 3. This situation will lead to the ship *capsizing*. As long as the *metacenter* is located above the center of gravity, the righting arm has a stabilizing effect to bring the ship back to its normal position. If, on the other hand, the righting arm is displaced below the center of gravity, the ship will lose its roll stability and *capsize*.

Hydrostatic and hydrodynamic characteristics of ships undergo changes because of the varying underwater volume, *centers of buoyancy* and gravity and pressure distribution. Another factor is the effect of forward speed on ship stability and motions, particularly on rolling motion in synchronous *beam* waves. Taylan [41] examined the influence of forward speed by incrementing its value and determining the roll responses at each speed interval. Various characteristics of the $\overline{GZ}$ curve for a selected test vessel were found to change systematically.

The roll stability of a ship is usually measured by the stability diagram shown in Figure 4. The diagram shows the dependence of the *righting arm* on the roll angle (*list*) and is an important design guide for roll stability.
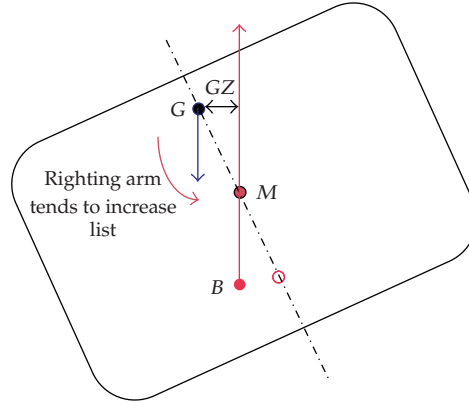
Figure 3: Negative ship stability.
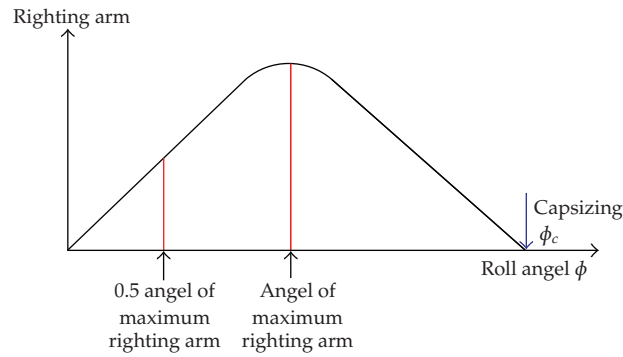
Figure 4: Dependence of the righting arm on the roll angle.

The roll oscillation of a ship is associated with a restoring moment to stabilize the ship about the x-axis given by the expression

$$M_x = W\overline{GM}\sin\phi. \tag{2.2}$$

where is the weight of water of displaced volume of the ship which is equal to the weight of the ship. If the ship experiences pitching motion of angle θ the righting arm will be raised by an increment $\overline{GM}\sin\phi\sin\theta$. In the case the net roll moment becomes

$$M_x = W\overline{GM}\sin\phi(1 + \sin\theta) \approx W\overline{GM}\phi(1 + \theta). \tag{2.3}$$

Note that the static stability is governed by the minimum value that the metacenter height, $\overline{GM}$, should have and the shape of the static stability curve with respect to the roll angle. This approach is still being applied in the assessment of stability criterion. The dynamic stability approach, on the other hand, is based on the equation of rolling motion. This involves constructing a model for a ship rolling in a realistic sea. The linear restoring parameters can be easily obtained from ship hydrostatics.
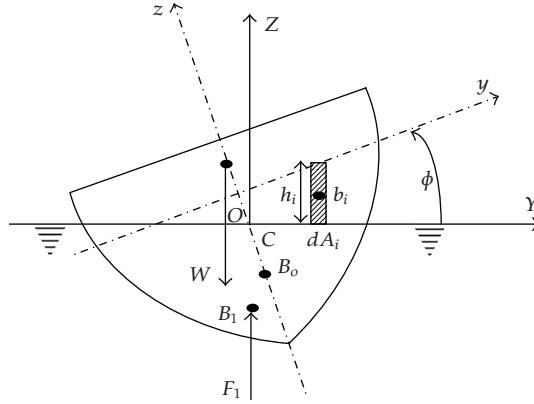
Figure 5: Ship schematic diagrams showing hydrostatic forces in a displaced position.

The curve for righting arm, known also as the restoring lever, has been represented by an odd-order polynomial up to different degrees [42–45]. Different representations of the restoring moment have been proposed in the literature. For example, Roberts [46, 47], Falzarano and Zhang [48], Huang et al. [49], and Senjanović et al. [50] represented $M_x(\phi)$ by the polynomial

$$M_x(\phi) = k_1\phi + k_3\phi^3 + k_5\phi^5 + k_7\phi^7 + \cdots, \tag{2.4}$$

where $k_1 > 0$, $k_3 < 0$, $k_5 > 0$, and $k_7 < 0$ for a damaged vessel, but $k_7 = 0$ for an intact vessel. Moshchuk et al. [51] proposed the following representation:

$$M_x(\phi) = k_0 \sin\left(\frac{\pi\phi}{\phi_s}\right) + \tilde{\gamma}\left(\frac{\pi\phi}{\phi_s}\right), \tag{2.5}$$

where ϕ_s is the capsizing angle, and the function $\tilde{\gamma}(\pi\phi/\phi_s)$ accounts for the difference between the exact function $M_x(\phi)$ and $k_0 \sin(\pi\phi/\phi_s)$.

3. Heave-Pitch-Roll Equations of Motion

Consider a ship sitting in its static equilibrium position with a submerged volume v_0. During its motion, its instantaneous submerged volume is v_1, and the difference in the submerged volume is $\delta v = v_1 - v_0$. The inertial frame of axes is XYZ with unit vectors $\mathbf{I}$, $\mathbf{J}$, and $\mathbf{K}$ along X-, Y-, and Z-axes, respectively. On the other hand, the body frame that moves with the ship is xyz with unit vectors $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ along x-, y-, and z-axes, respectively. Figure 5 shows the instantaneous buoyant center located at point B_1 and the corresponding instantaneous force is $\mathbf{F}_1 = \rho g v_1 \mathbf{K} = \rho g(v_0 + \delta v)\mathbf{K}$. The weight of the ship is $\mathbf{W} = -\rho g v_0 \mathbf{K}$. In this case the instantaneous restoring hydrostatic force is

$$\mathbf{F}_H = \rho g \delta v \mathbf{K}. \tag{3.1}$$

The restoring moment is the resultant between the moments of weight and instantaneous buoyancy

$$\mathbf{M}_H = \rho g \left[v_0 \mathbf{K} \times \left(\overrightarrow{OG} - \overrightarrow{OB}_o \right) + \int \overrightarrow{Ob}_i \times dv_i \mathbf{K} \right], \tag{3.2}$$

where dv_i is the volume of the infinitesimal prism of height h_i, $\overrightarrow{OG} = \bar{z}_G \mathbf{k}$, $\bar{z}_G$ is the center of mass location from O, $\overrightarrow{OB}_o = \bar{z}_{BO}\mathbf{k}$, $\overrightarrow{Ob}_i \approx x_{Ai}\mathbf{I} + y_{Ai}\mathbf{J} + (h_i/2)\mathbf{K}$, $\mathbf{k} = \sin\theta\mathbf{I} - \sin\phi\mathbf{J} + \mathbf{K}$, $\bar{z}_G$ is the vertical coordinate of the center of gravity, $\delta v = \int h_i dA_i = \int h dA$, and $\bar{z}_{BO}$ is the vertical coordinate of the center of gravity of the submerged volume, and (x_A, y_A) are the coordinates of the elemental prism in the instantaneous plane with respect to the inertial frame $CXYZ$. Substituting these parameters in (3.2) gives

$$\mathbf{M}_H = -\mathbf{I}\rho g \left[v_0(\bar{z}_{BO} - \bar{z}_G)\sin\phi - \int y_A h dA \right] - \mathbf{J}\rho g \left[v_0(\bar{z}_{BO} - \bar{z}_G)\sin\theta + \int x_A h dA \right]. \tag{3.3}$$

The elemental prism height $h_i = h$ can be written in terms of the heave displacement z of the origin O above the water level, the pitch, θ, and roll, ϕ, angles as

$$h_i = -z - y_{Ai}\sin\phi + x_{Ai}\sin\theta. \tag{3.4}$$

The volume variation δv is

$$\begin{aligned} \delta v &= \int dv \\ &= \int (-z - y_A\sin\phi + x_A\sin\theta)dA \\ &= -z\int dA - \sin\phi\int y_A dA + \sin\theta\int x_A dA. \end{aligned} \tag{3.5}$$

The above summations are dependent on z, ϕ, and θ. They represent the following geometric properties:

$$\int dA = A(z, \phi, \theta) \quad \text{area of instantaneous plane of floatation,}$$

$$\int y_A dA = A_x(z, \phi, \theta) \quad \text{first static moment of the area about } x\text{-axis,} \tag{3.6}$$

$$\int x_A dA = A_y(z, \phi, \theta) \quad \text{first static moment of area about } y\text{-axis.}$$

In this case, one may write the volume variation in the form

$$\delta v = -zA(z,\phi,\theta) - \sin\phi A_x(z,\phi,\theta) + \sin\theta A_y(z,\phi,\theta). \tag{3.7}$$

Note the above summations could have been replaced by integrals. The instantaneous restoring hydrostatic force given by (3.1) takes the form

$$\mathbf{F}_H = -\rho g\left[zA(z,\phi,\theta) + \sin\phi A_x(z,\phi,\theta) - \sin\theta A_y(z,\phi,\theta)\right]\mathbf{K}. \tag{3.8}$$

In scalar form, the absolute value of the restoring force is

$$F_H = \rho g\left[zA(z,\phi,\theta) + \sin\phi A_x(z,\phi,\theta) - \sin\theta A_y(z,\phi,\theta)\right]. \tag{3.9}$$

The summations in (3.3) can also be written in terms of (3.4) as

$$\begin{aligned}
\int y_A h\,dA &= \int y_A(-z - y_A\sin\phi + x_A\sin\theta)\,dA \\
&= -zA_x(z,\phi,\theta) - \sin\phi I_{xx}(z,\phi,\theta) + \sin\theta I_{xy}(z,\phi,\theta), \\
\int x_A h\,dA &= \int x_A(-z - y_A\sin\phi + x_A\sin\theta)\,dA \\
&= -zA_y(z,\phi,\theta) - \sin\phi I_{xy}(z,\phi,\theta) + \sin\theta I_{yy}(z,\phi,\theta),
\end{aligned} \tag{3.10}$$

where

$$I_{xx}(z,\phi,\theta) = \int y_A^2\,dA, \qquad I_{yy}(z,\phi,\theta) = \int x_A^2\,dA, \qquad I_{xy}(z,\phi,\theta) = \int x_A y_A\,dA. \tag{3.11}$$

Introducing (3.10) into (3.3) and writing the result in the absolute and scalar form give

$$\begin{aligned}
M_{xH} &= \rho g\left[A_x z + v_0(\overline{z}_{BO} - \overline{z}_G)\sin\phi + I_{xx}\sin\phi - I_{xy}\sin\theta\right], \\
M_{yH} &= \rho g\left[-A_y z - I_{xy}\sin\phi + v_0(\overline{z}_{BO} - \overline{z}_G)\sin\theta + I_{yy}\sin\theta\right].
\end{aligned} \tag{3.12}$$

Note that the geometrical parameters (3.6) and (3.11) depend on the instantaneous displacements of the ship (z, ϕ, θ). These properties may be expanded in multivariable Taylor series around the average position, that is,

$$A(z, \phi, \theta) = A_0 + \left.\frac{\partial A}{\partial z}\right|_0 z + \left.\frac{\partial A}{\partial \theta}\right|_0 \theta + \frac{1}{2}\left.\frac{\partial^2 A}{\partial z^2}\right|_0 z^2 + \left.\frac{\partial^2 A}{\partial z \partial \theta}\right|_0 z\theta + \frac{1}{2}\left.\frac{\partial^2 A}{\partial \phi^2}\right|_0 \phi^2 + \frac{1}{2}\left.\frac{\partial^2 A}{\partial \theta^2}\right|_0 \theta^2,$$

$$A_x(z, \phi, \theta) = \left.\frac{\partial A_x}{\partial \phi}\right|_0 \phi + \left.\frac{\partial^2 A_x}{\partial z \partial \phi}\right|_0 z\phi + \left.\frac{\partial^2 A_x}{\partial \phi \partial \theta}\right|_0 \phi\theta,$$

$$A_y(z, \phi, \theta) = \left.A_y\right|_0 + \left.\frac{\partial A_y}{\partial z}\right|_0 z + \left.\frac{\partial A_y}{\partial \theta}\right|_0 \theta + \frac{1}{2}\left.\frac{\partial^2 A_y}{\partial z^2}\right|_0 z^2 + \left.\frac{\partial^2 A_y}{\partial z \partial \theta}\right|_0 z\theta + \frac{1}{2}\left.\frac{\partial^2 A_y}{\partial \phi^2}\right|_0 \phi^2 + \frac{1}{2}\left.\frac{\partial^2 A_y}{\partial \theta^2}\right|_0 \theta^2,$$

$$I_{xx}(z, \phi, \theta) = \left.I_{xx}\right|_0 + \left.\frac{\partial I_{xx}}{\partial z}\right|_0 z + \left.\frac{\partial I_{xx}}{\partial \theta}\right|_0 \theta + \frac{1}{2}\left.\frac{\partial^2 I_{xx}}{\partial z^2}\right|_0 z^2 + \left.\frac{\partial^2 I_{xx}}{\partial z \partial \theta}\right|_0 z\theta + \frac{1}{2}\left.\frac{\partial^2 I_{xx}}{\partial \phi^2}\right|_0 \phi^2 + \frac{1}{2}\left.\frac{\partial^2 I_{xx}}{\partial \theta^2}\right|_0 \theta^2,$$

$$I_{yy}(z, \phi, \theta) = \left.I_{yy}\right|_0 + \left.\frac{\partial I_{yy}}{\partial z}\right|_0 z + \left.\frac{\partial I_{yy}}{\partial \theta}\right|_0 \theta + \frac{1}{2}\left.\frac{\partial^2 I_{yy}}{\partial z^2}\right|_0 z^2 + \left.\frac{\partial^2 I_{yy}}{\partial z \partial \theta}\right|_0 z\theta + \frac{1}{2}\left.\frac{\partial^2 I_{yy}}{\partial \phi^2}\right|_0 \phi^2 + \frac{1}{2}\left.\frac{\partial^2 I_{yy}}{\partial \theta^2}\right|_0 \theta^2,$$

$$I_{xy}(z, \phi, \theta) = \left.\frac{\partial I_{xy}}{\partial \phi}\right|_0 \phi + \left.\frac{\partial^2 I_{xy}}{\partial z \partial \phi}\right|_0 z\phi + \left.\frac{\partial^2 I_{xy}}{\partial \phi \partial \theta}\right|_0 \phi\theta.$$

$$(3.13)$$

A_0, $\left.A_y\right|_0$, $\left.I_{xx}\right|_0$, and $\left.I_{yy}\right|_0$ are the geometric properties evaluated at the average plane of floatation. Note that the variation of first moment of area about the x-axis is dependent on an odd order of roll angle. That dependence does not exist in variations of other geometrical parameters. Paulling and Rosenberg [8] showed that the dependencies of the heave and pitch coefficients on roll are of even order, while the coefficients in roll due to heave and pitch are odd.

The restoring hydrodynamic force and moments given by (3.8) and (3.3) take the form

$$F_H = \rho g \left\{ z A_0 - \left.A_y\right|_0 \sin\theta + \left.\frac{\partial A}{\partial z}\right|_0 z^2 + \left.\frac{\partial A}{\partial \theta}\right|_0 z(\theta + \sin\theta) + \left.\frac{\partial A_x}{\partial \phi}\right|_0 \phi \sin\phi \right.$$

$$- \left.\frac{\partial A_y}{\partial \theta}\right|_0 \theta \sin\theta + \frac{1}{6}\left.\frac{\partial^2 A}{\partial z^2}\right|_0 z^3 + \left.\frac{\partial^2 A}{\partial z \partial \theta}\right|_0 z^2(\theta + \sin\theta) + \frac{1}{2}\left.\frac{\partial^2 A}{\partial \phi^2}\right|_0 z\phi\left(\frac{\phi}{2} + \sin\phi\right)$$

$$\left. + \left.\frac{\partial^2 A_x}{\partial \phi \partial \theta}\right|_0 \phi\left(\theta \sin\phi + \frac{\phi}{2}\sin\theta\right) + \left.\frac{\partial^2 A}{\partial \theta^2}\right|_0 z\theta\left(\frac{\theta}{2} + \sin\theta\right) - \frac{1}{2}\left.\frac{\partial^2 A_y}{\partial \theta^2}\right|_0 \theta^2 \sin\theta \right\},$$

$$M_{xH} = \rho g \left\{ v_0 (\overline{z}_{BO} - \overline{z}_G) \sin\phi + I_{xx}|_0 z \sin\phi + \left.\frac{\partial I_{xx}}{\partial z}\right|_0 z(\phi + \sin\phi) + \left.\frac{\partial I_{xx}}{\partial \theta}\right|_0 (\theta \sin\phi + \phi \sin\theta) \right.$$

$$+ \left.\frac{\partial^2 I_{xx}}{\partial z^2}\right|_0 z^2 \left(\phi + \frac{1}{2}\sin\phi\right) + \frac{1}{2}\left.\frac{\partial^2 I_{xx}}{\partial \theta^2}\right|_0 \theta(\theta \sin\phi + \phi \sin\theta)$$

$$\left. + \left.\frac{\partial^2 I_{xx}}{\partial z \partial \theta}\right|_0 z(\phi\theta + \theta \sin\phi + \phi \sin\phi) + \frac{1}{2}\left.\frac{\partial^2 I_{xx}}{\partial \phi^2}\right|_0 \phi^2 \sin\phi \right\},$$

$$M_{yH} = \rho g \left\{ v_0 (\overline{z}_{BO} - \overline{z}_G) \sin\theta - A_y|_0 z + I_{yy}|_0 \sin\theta - \left.\frac{\partial A_y}{\partial z}\right|_0 z^2 + \left.\frac{\partial I_{yy}}{\partial z}\right|_0 z(\theta + \sin\theta) \right.$$

$$- \left.\frac{\partial I_{xy}}{\partial \phi}\right|_0 \phi \sin\phi + \left.\frac{\partial I_{yy}}{\partial \theta}\right|_0 \theta \sin\theta - \frac{1}{6}\left.\frac{\partial^2 A_y}{\partial z^2}\right|_0 z^3 + \left.\frac{\partial^2 I_{yy}}{\partial z^2}\right|_0 z^2\left(\theta + \frac{1}{2}\sin\theta\right)$$

$$- \left.\frac{\partial^2 I_{xy}}{\partial z \partial \phi}\right|_0 z\phi\left(\frac{\phi}{2} + \sin\phi\right) + \left.\frac{\partial^2 I_{yy}}{\partial z \partial \theta}\right|_0 z\theta\left(\frac{\theta}{2} + \sin\theta\right) - \left.\frac{\partial^2 I_{xy}}{\partial \phi \partial \theta}\right|_0 \phi\theta \sin\phi$$

$$\left. + \frac{1}{2}\left.\frac{\partial^2 I_{yy}}{\partial \phi^2}\right|_0 \phi^2 \sin\theta + \frac{1}{2}\left.\frac{\partial^2 I_{yy}}{\partial \theta^2}\right|_0 \theta^2 \sin\theta \right\}.$$

$$(3.14)$$

In achieving the above equations use has been made of the following equalities verified by Neves and Rodríguez [11]:

$$\left.\frac{\partial A}{\partial \theta}\right|_0 = -\left.\frac{\partial A_y}{\partial z}\right|_0, \qquad \left.\frac{\partial A_x}{\partial \phi}\right|_0 = \left.\frac{\partial I_{xx}}{\partial z}\right|_0, \qquad \left.\frac{\partial I_{xx}}{\partial \theta}\right|_0 = -\left.\frac{\partial I_{xy}}{\partial \phi}\right|_0,$$

$$\left.\frac{\partial I_{yy}}{\partial z}\right|_0 = -\left.\frac{\partial A_y}{\partial \theta}\right|_0, \qquad \left.\frac{\partial^2 A}{\partial \phi^2}\right|_0 = \left.\frac{\partial^2 A_x}{\partial z \partial \phi}\right|_0 = \left.\frac{\partial^2 I_{xx}}{\partial z^2}\right|_0,$$

$$\left.\frac{\partial^2 A}{\partial \theta^2}\right|_0 = -\left.\frac{\partial^2 A_y}{\partial z \partial \theta}\right|_0 = \left.\frac{\partial^2 I_{yy}}{\partial z^2}\right|_0, \qquad \left.\frac{\partial^2 A}{\partial z \partial \theta}\right|_0 = -\left.\frac{\partial^2 A_y}{\partial z^2}\right|_0,$$

$$-\left.\frac{\partial^2 I_{xx}}{\partial z \partial \theta}\right|_0 = -\left.\frac{\partial^2 I_{xy}}{\partial z \partial \phi}\right|_0 = \left.\frac{\partial^2 A_x}{\partial \phi \partial \theta}\right|_0 = -\left.\frac{\partial^2 A_y}{\partial \phi^2}\right|_0, \qquad \left.\frac{\partial^2 I_{xx}}{\partial \theta^2}\right|_0 = -\left.\frac{\partial^2 I_{xy}}{\partial \phi \partial \theta}\right|_0 = \left.\frac{\partial^2 I_{yy}}{\partial \phi^2}\right|_0,$$

$$\left.\frac{\partial^2 I_{yy}}{\partial z \partial \theta}\right|_0 = -\left.\frac{\partial^2 A_y}{\partial \theta^2}\right|_0.$$

$$(3.15)$$

3.1. Wave Motion Effects

The influence of incident sea waves of arbitrary direction along the hull is to change the average submerged shape defined by the instantaneous position of the wave. These waves exert external forces and moments in heave, roll, and pitch in addition they introduce an additional restoring forces and moments. For the case of head sea, Neves and Rodríguez [11, 12] considered the Airy linear theory in representing longitudinal waves (along x-axis) defined by the expression [14]

$$\eta(x, y, t; \chi) = \eta_0 \cos(kx + \omega_e t), \tag{3.16}$$

where η_0 is the wave amplitude, $k = \omega_w^2/g = 2\pi/\lambda$ is the wave number, ω_w is the wave frequency, λ is the wave length, g is the gravitational acceleration, χ is the wave incidence, and $\omega_e = (\omega_w - kU \cos \chi)$ is the encounter frequency of the wave by the ship when the ship advances with speed U.

Note that h_i expressed by (3.4) should read

$$h_i = -[z - \eta(x_{Ai}, y_{Ai}, t)] - y_{Ai} \sin \phi + x_{Ai} \sin \theta. \tag{3.17}$$

The contributions of longitudinal waves to the restoring force, F_η, and the restoring moments, $M_{x\eta}$ and $M_{y\eta}$, obtained using Taylor series expansion about the average position up to third-order terms are given by the expressions [11, 12]

$$F_\eta = \left.\frac{\partial^2 F_\eta}{\partial \eta \partial z}\right|_0 z + \left.\frac{\partial^2 F_\eta}{\partial \eta \partial \theta}\right|_0 \theta + \left.\frac{\partial^3 F_\eta}{\partial \eta \partial z^2}\right|_0 z^2 + \left.\frac{\partial^3 F_\eta}{\partial \eta \partial z \partial \theta}\right|_0 z\theta + \left.\frac{\partial^3 F_\eta}{\partial \phi^2 \partial \eta}\right|_0 \phi^2 + \left.\frac{\partial^3 F_\eta}{\partial \eta^2 \partial \theta}\right|_0 \theta$$

$$+ \left.\frac{\partial^3 F_\eta}{\partial \eta \partial \theta^2}\right|_0 \theta^2,$$

$$M_{x\eta} = \left.\frac{\partial^2 M_{x\eta}}{\partial \eta \partial \phi}\right|_0 \phi + \left.\frac{\partial^3 M_{x\eta}}{\partial \eta \partial z \partial \phi}\right|_0 z\phi + \left.\frac{\partial^3 M_{x\eta}}{\partial \eta^2 \partial \phi}\right|_0 \phi + \left.\frac{\partial^3 M_{x\eta}}{\partial \eta \partial \phi \partial \theta}\right|_0 \phi\theta,$$

$$M_{y\eta} = \left.\frac{\partial^2 M_{y\eta}}{\partial \eta \partial z}\right|_0 z + \left.\frac{\partial^2 M_{y\eta}}{\partial \eta \partial \theta}\right|_0 \theta + \left.\frac{\partial^3 M_{y\eta}}{\partial \eta^2 \partial z}\right|_0 z + \left.\frac{\partial^3 M_{y\eta}}{\partial \eta \partial z^2}\right|_0 z\theta + \left.\frac{\partial^3 M_{y\eta}}{\partial \phi^2 \partial \eta}\right|_0 \phi^3 + \left.\frac{\partial^3 M_{y\eta}}{\partial \eta^2 \partial \theta}\right|_0 \theta$$

$$+ \left.\frac{\partial^3 M_{y\eta}}{\partial \theta^2 \partial \eta}\right|_0 \theta^2, \tag{3.18}$$

where the derivatives of the above equations are given by the following expressions:

$$
\left.\frac{\partial^2 F_\eta}{\partial\eta\partial z}\right|_0 = 2\rho g \int_L \frac{\partial y}{\partial z}\eta\,dx,
$$

$$
\left.\frac{\partial^2 F_\eta}{\partial\eta\partial\theta}\right|_0 = -2\rho g \int_L x\frac{\partial y}{\partial z}\eta\,dx,
$$

$$
\left.\frac{\partial^3 F_\eta}{\partial\phi^2\partial\eta}\right|_0 = -\rho g \int_L \left[2y\left(\frac{\partial y}{\partial z}\right)^2 + y\right]\eta\,dx,
$$

$$
\left.\frac{\partial^3 F_\eta}{\partial\eta\partial z^2}\right|_0 = \left.\frac{\partial^3 F_\eta}{\partial\eta\partial z\partial\theta}\right|_0 = \left.\frac{\partial^3 F_\eta}{\partial\eta^2\partial\theta}\right|_0 = \left.\frac{\partial^3 F_\eta}{\partial\eta\partial\theta^2}\right|_0 = 0,
$$

$$
\left.\frac{\partial^2 M_{x\eta}}{\partial\eta\partial\phi}\right|_0 = 2\rho g \int_L y^2\frac{\partial y}{\partial z}\eta\,dx,
$$

$$
\left.\frac{\partial^3 M_{x\eta}}{\partial\eta\partial z\partial\phi}\right|_0 = -2\rho g \int_L \left[2y\left(\frac{\partial y}{\partial z}\right)^2 + y\right]\eta\,dx,
$$

$$
\left.\frac{\partial^3 M_{x\eta}}{\partial\eta^2\partial\phi}\right|_0 = \rho g \int_L \left[2y\left(\frac{\partial y}{\partial z}\right)^2 + y\right]\eta^2\,dx,
$$

$$
\left.\frac{\partial^3 M_{x\eta}}{\partial\eta\partial\phi\partial\theta}\right|_0 = 2\rho g \int_L \left[2xy\left(\frac{\partial y}{\partial z}\right)^2 + xy\right]\eta\,dx,
$$

$$
\left.\frac{\partial^2 M_{y\eta}}{\partial\eta\partial z}\right|_0 = -2\rho g \int_L x\frac{\partial y}{\partial z}\eta\,dx,
$$

$$
\left.\frac{\partial^2 M_{y\eta}}{\partial\eta\partial\theta}\right|_0 = 2\rho g \int_L x^2\frac{\partial y}{\partial z}\eta\,dx,
$$

$$
\left.\frac{\partial^3 M_{y\eta}}{\partial\phi^2\partial\eta}\right|_0 = \rho g \int_L \left[2xy\left(\frac{\partial y}{\partial z}\right)^2 + xy\right]\eta\,dx,
$$

$$
\left.\frac{\partial^3 M_{y\eta}}{\partial\eta^2\partial z}\right|_0 = \left.\frac{\partial^3 M_{y\eta}}{\partial\eta\partial z^2}\right|_0 = \left.\frac{\partial^3 M_{y\eta}}{\partial\theta^2\partial\eta}\right|_0 = \left.\frac{\partial^3 M_{y\eta}}{\partial\eta^2\partial\theta}\right|_0 = 0.
$$

(3.19)

3.2. Ships Roll Damping

The surface waves introduce inertia and drag hydrodynamic forces. The inertia force is the sum of two components. The first is a buoyancy force acting on the structure in the fluid due to a pressure gradient generated from the flow acceleration. The buoyancy force is equal to the mass of the fluid displaced by the structure multiplied by the acceleration of the flow. The second inertia component is due to the added mass, which is proportional to

the relative acceleration between the structure and the fluid. This component accounts for the flow entrained by the structure. The drag force is the sum of the viscous and pressure drags produced by the relative velocity between the structure and the flow. This type of hydrodynamic drag is proportional to the square of the relative velocity.

Viscosity plays an important role in ship responses especially at large-amplitude roll motions in which the wave radiation damping is relatively low. The effect of the *bilge keel* on the roll damping was first discussed by Bryan [52]. Hishida [53–55] proposed an analytical approach to roll damping for ship hulls in simple oscillatory waves. The regressive curve of the roll damping obtained from the experiments by Kato [56] has been widely used in the prediction of ship roll motions. Since amplitudes and frequencies are varying in random waves, the hydrodynamic coefficients are time-dependent and irregular. Several experimental investigations have been conducted to measure the effect of *bilge keels* on the roll damping (see, e.g., [57–66]).

It was indicated by Bishop and Price [67] that existing information on the structural damping of ships is far from satisfactory. It cannot be calculated and it can only be measured in the presence of hydrodynamic damping, whose nature and magnitude are also somewhat obscure. Yet it is very important. Much less is known about antisymmetric responses to waves, either as regards the means of estimating them or the appropriate levels of hull damping. Vibration at higher frequencies, due to excitation by machinery (notably propellers), is limited by structural damping to a much greater extent than it is by the fluid actions of the sea. Damping measurements at these frequencies therefore give more accurate estimates of hull damping. The damping moment of ships is related to multiplicity of factors such as hull shape, loading condition, *bilge keel*, rolling frequency, and range of rolling angle. For small roll angles, the damping moment is directly proportional to the angular roll velocity. But with increasing roll angle, nonlinear damping will become significant. Due to the occurrence of strong viscous effects, the roll damping moment cannot be computed by means of potential theory. Himeno [68] provided a detailed description of the equivalent damping coefficient and expressed it in terms of various contributions due to hull skin friction damping, hull eddy shedding damping, free surface wave damping, lift force damping, and bilge keel damping. The viscous damping is due to the following sources.

(i) Wave-making moment, B_W.

(ii) Skin-friction damping moment, B_F.

(iii) The moment resulting from the bare hull arising from separation and eddies mostly near the bilge keels, B_E.

(iv) Lift damping moment due to an apparent angle of attack as the ship rolls, B_L.

(v) Bilge-keel damping moment, B_{BK}.

Damping due to bilge keels can be decomposed into the following components.

(i) Bilge keels moment due to normal force, B_{BKN}.

(ii) Moment due to interaction between hull and bilge keel, B_{BKH}.

(iii) Modification to wave making due to the presence of bilge keels, B_{BKW}.

The damping components B_F, B_L, B_W, and B_{BKW} are linear, while B_E, B_{BKN}, and B_{BKH} are nonlinear. The linear and nonlinear damping moments can be expressed as follows:

$$B_{\text{lin}} = B_F + B_L + B_W + B_{BKW},$$

$$B_{\text{nonlin}} = B_E + B_{BKN} + B_{BKH}. \tag{3.20}$$

A pseudospectral model for nonlinear ship-surface wave interactions was developed by Lin et al. [69]. The algorithm is a combination of spectral and boundary element methods. All possible ship-wave interactions were included in the model. The nonlinear bow waves at high Froude numbers from the pseudospectral model are much closer to the experimental results than those from linear ship wave models. One of the main problems in modeling ship-wave hydrodynamics is solving for the forcing (pressure) at the ship boundary. With an arbitrary ship, singularities occur in evaluating the velocity potential and the velocities on the hull. Inaccuracies in the evaluation of the singular terms in the velocity potential result in discretization errors, numerical errors, and excessive computational costs. Lin and Kuang [70, 71] presented a new approach to evaluating the pressure on a ship. They used the digital, self-consistent, ship experimental laboratory (DiSSEL) ship motion model to test its effectiveness in predicting ship roll motion. It was shown that the implementation of this roll damping component improves significantly the accuracy of numerical model results. Salvesen [72] reported some results pertaining numerical methods such as large amplitude motion program (LAMP) used to evaluate hydrodynamic performance characteristics. These methods were developed for solving fully three-dimensional ship-motions, ship-wave-resistance and local-flow problems using linearized free-surface boundary conditions. Lin et al. [73] examined the capabilities of the 3D nonlinear time-domain Large Amplitude Motion Program (LAMP) for the evaluation of fishing vessels operating in extreme waves. They extended their previous work to the modeling of maritime casualties, including a time-domain simulation of a ship capsizing in stern quartering seas.

The damping characteristics of a variety of ship shapes and offshore structures undergoing roll oscillation in the presence of ocean waves have been assessed by Chakrabarti [74]. Chakrabarti [74] relied on empirical formulas derived from a series of model experiments reported by Ikeda [75] and Ikeda et al. [76–78]. These experiments were performed on two-dimensional shapes. The damping roll moment $B(\dot{\phi})$ is nonlinear and may be expressed by the expression [40, 74]

$$B(\dot{\phi}) = c_1 \dot{\phi} + c_2 \dot{\phi}|\dot{\phi}| + c_3 \dot{\phi}^3 + \cdots = \sum_{k=1}^{K} c_k \dot{\phi}|\dot{\phi}|^{k-1}. \tag{3.21}$$

The first term is the usual linear viscous damping, the second is the quadratic damping term originally developed by Morison et al. [79]. It is in phase with the velocity but it is quadratic because the flow is separated and the drag is primarily due to pressure rather than the skin friction. Sarpkaya and Isaacson [80] provided a critical assessment of Morison's equation, which describes the forces acting on a pile due to the action of progressive waves. The third term is cubic damping. The total damping may be replaced by an equivalent viscous term in the form

$$B(\dot{\phi}) = c_{\text{eq}} \dot{\phi}, \tag{3.22}$$

where c_{eq} is the equivalent damping coefficient. This coefficient can be expressed in terms of the nonlinear coefficients as

$$c_{\text{eq}} = c_1 + \frac{8}{3\pi} c_2 (\omega \phi_0) + \frac{3}{4} c_3 (\omega \phi_0)^2, \tag{3.23}$$

where ω is the wave frequency and ϕ_0 is the amplitude of the ship roll angle.

Dalzell [37] replaced the nonlinear damping term $\dot{\phi}|\dot{\phi}|$ by an equivalent smooth nonlinear polynomial given by

$$\dot{\phi}|\dot{\phi}| = \sum_{k=1,3,\dots} \alpha_k \frac{\dot{\phi}^k}{\left(\dot{\phi}_c\right)^{k-2}} \cong \frac{5}{16} \dot{\phi}_c \dot{\phi} + \frac{35}{48} \frac{\dot{\phi}^3}{\dot{\phi}_c}, \tag{3.24}$$

where $\dot{\phi}_c$ is the maximum amplitude of roll velocity. The numerical coefficients α_k were estimated by using least-square fitting.

Haddara [81] employed the concept of the random decrement in the damping identification of linear systems. He extended the concept of the random decrement for a ship performing rolling motion in random beam waves. Wave excitation was assumed to be a Gaussian white noise process. The equations were used to identify the parameters of the nonlinear roll damping moment. Wu et al. [82] conducted an experimental investigation to measure the nonlinear roll damping of a ship in regular and irregular waves.

3.3. Ship Inertia Forces and Moments

The inertia forces and moments in heave, roll, and pitch motions are mainly due to the ship mass and mass moment of inertia and the corresponding added mass terms. These are well documented in Neves and Rodríguez [11] and are given in the form

$$F_{ZI} = (m + Z_{\ddot{z}}) \ddot{z} + Z_{\ddot{\theta}} \ddot{\theta},$$

$$M_{xI} = \left(J_{xx} + K_{\ddot{\phi}} \right) \ddot{\phi}, \tag{3.25}$$

$$M_{yI} = \left(J_{yy} + K_{\ddot{\theta}} \right) \ddot{\theta} + M_{\ddot{z}} \ddot{z},$$

where m is the ship mass, J_{xx} and J_{yy} are the ship mass moment of inertia about roll and pitch axes, $Z_{\ddot{z}}$ is the hydrodynamic added mass in heave, $Z_{\ddot{\theta}}$ is the hydrodynamic added inertia in heave due to pitch motion, inertia $K_{\ddot{\phi}}$ and $K_{\ddot{\theta}}$ are the hydrodynamic added polar mass moment of inertia about the ship roll and pitch axes, respectively, and $M_{\ddot{z}}$ is the added inertia in pitch due to heave motion. The added inertia parameters may be evaluated using the potential theory as described by Salvesen et al. [83] and Meyers et al. [84].

3.4. Governing Equations of Motion

Applying Newton's second law, the equations governing heave-roll-pitch motion may be written in a form.

The heave equation of motion is

$$
(m + Z_{\ddot{z}})\ddot{z} + Z_{\ddot{\theta}}\ddot{\theta} + C_{\dot{z}}\dot{z} + \rho g \left\{ zA_0 - A_y\big|_0 \sin\theta + \left.\frac{\partial A}{\partial z}\right|_0 z^2 + \left.\frac{\partial A}{\partial \theta}\right|_0 z(\theta + \sin\theta) + \left.\frac{\partial A_x}{\partial \phi}\right|_0 \phi\sin\phi \right.
$$

$$
- \left.\frac{\partial A_y}{\partial \theta}\right|_0 \theta\sin\theta + \frac{1}{6}\left.\frac{\partial^2 A}{\partial z^2}\right|_0 z^3 + \left.\frac{\partial^2 A}{\partial z \partial \theta}\right|_0 z^2(\theta + \sin\theta)
$$

$$
+ \frac{1}{2}\left.\frac{\partial^2 A}{\partial \phi^2}\right|_0 z\phi\left(\frac{\phi}{2} + \sin\phi\right) + \left.\frac{\partial^2 A_x}{\partial \phi \partial \theta}\right|_0 \phi\left(\theta\sin\phi + \frac{\phi}{2}\sin\theta\right)
$$

$$
+ \left.\frac{\partial^2 A}{\partial \theta^2}\right|_0 z\theta\left(\frac{\theta}{2} + \sin\theta\right) - \frac{1}{2}\left.\frac{\partial^2 A_y}{\partial \theta^2}\right|_0 \theta^2\sin\theta \bigg\}
$$

$$
+ 2\rho g z \int_L \frac{\partial y}{\partial z}\eta(t)dx - 2\rho g \theta \int_L x\frac{\partial y}{\partial z}\eta(t)dx - \rho g \phi^2 \int_L \left[2y\left(\frac{\partial y}{\partial z}\right)^2 + y\right]\eta(t)dx = Z(t).
$$

$$(3.26)$$

The roll moment equation of motion taking into account the beam sea hydrodynamic wave excitation moment, $\Phi(t)$, is

$$
\left(J_{xx} + K_{\ddot{\phi}}\right)\ddot{\phi} + \sum_{k=1}^{K} c_k \dot{\phi}|\dot{\phi}|^{k-1} + \rho g \left\{ v_0(\bar{z}_{BO} - \bar{z}_G)\sin\phi + I_{xx}\big|_0 z\sin\phi + \left.\frac{\partial I_{xx}}{\partial z}\right|_0 z(\phi + \sin\phi) \right.
$$

$$
+ \left.\frac{\partial I_{xx}}{\partial \theta}\right|_0 (\theta\sin\phi + \phi\sin\theta) + \left.\frac{\partial^2 I_{xx}}{\partial z^2}\right|_0 z^2\left(\phi + \frac{1}{2}\sin\phi\right)
$$

$$
+ \frac{1}{2}\left.\frac{\partial^2 I_{xx}}{\partial \theta^2}\right|_0 \theta(\theta\sin\phi + \phi\sin\theta)
$$

$$
+ \left.\frac{\partial^2 I_{xx}}{\partial z \partial \theta}\right|_0 z(\phi\theta + \theta\sin\phi + \phi\sin\phi)
$$

$$
+ \frac{1}{2}\left.\frac{\partial^2 I_{xx}}{\partial \phi^2}\right|_0 \phi^2\sin\phi \bigg\}
$$

$$
+ 2\rho g \phi \int_L y^2 \frac{\partial y}{\partial z}\eta(t)dx - 2\rho g z \phi \int_L \left[2y\left(\frac{\partial y}{\partial z}\right)^2 + y\right]\eta(t)dx + \rho g \phi \int_L \left[2y\left(\frac{\partial y}{\partial z}\right)^2 + y\right]\eta^2(t)dx
$$

$$
+ 2\rho g \phi \theta \int_L \left[2xy\left(\frac{\partial y}{\partial z}\right)^2 + xy\right]\eta(t)dx = \Phi(t).
$$

$$(3.27)$$

The pitch moment equation of motion taking into account the beam sea hydrodynamic wave excitation moment, $\Theta(t)$, is

$$(J_{yy} + K_{\dot\theta})\ddot\theta + M_{\dot z}\ddot z + C_{\dot\theta}\dot\theta + \rho g \left\{ v_0(\overline z_{BO} - \overline z_G)\sin\theta - A_y\big|_0 z + I_{yy}\big|_0 \sin\theta - \frac{\partial A_y}{\partial z}\bigg|_0 z^2 \right.$$

$$+ \frac{\partial I_{yy}}{\partial z}\bigg|_0 z(\theta + \sin\theta) - \frac{\partial I_{xy}}{\partial \phi}\bigg|_0 \phi\sin\phi + \frac{\partial I_{yy}}{\partial \theta}\bigg|_0 \theta\sin\theta$$

$$- \frac{1}{6}\frac{\partial^2 A_y}{\partial z^2}\bigg|_0 z^3 + \frac{\partial^2 I_{yy}}{\partial z^2}\bigg|_0 z^2\left(\theta + \frac{1}{2}\sin\theta\right)$$

$$- \frac{\partial^2 I_{xy}}{\partial z\partial \phi}\bigg|_0 z\phi\left(\frac{\phi}{2} + \sin\phi\right) + \frac{\partial^2 I_{yy}}{\partial z\partial \theta}\bigg|_0 z\theta\left(\frac{\theta}{2} + \sin\theta\right)$$

$$\left. - \frac{\partial^2 I_{xy}}{\partial\phi\partial\theta}\bigg|_0 \phi\theta\sin\phi + \frac{1}{2}\frac{\partial^2 I_{yy}}{\partial\phi^2}\bigg|_0 \phi^2\sin\theta + \frac{1}{2}\frac{\partial^2 I_{yy}}{\partial\theta^2}\bigg|_0 \theta^2\sin\theta \right\}$$

$$- 2\rho g z\int_L x\frac{\partial y}{\partial z}\eta(t)dx + 2\rho g\theta\int_L x^2\frac{\partial y}{\partial z}\eta(t)dx + \rho g\phi^3\int_L \left[2xy\left(\frac{\partial y}{\partial z}\right)^2 + xy\right]\eta(t)dx = \Theta(t),$$

$$(3.28)$$

where $C_{\dot z}$ and $C_{\dot\theta}$ are linear damping coefficients associated with heave and pitch motions, respectively. $Z(t)$, $\Phi(t)$, and $\Theta(t)$ are the external excitations due to sea waves. One can extract from the above three equations the coupled roll-pitch equations of motion or the purely roll equation of motion. Nayfeh et al. [85] described two different mechanisms that cause roll instabilities in ships. An approximate solution based on the method of multiple scales was presented together with different simulations using the Large Amplitude Motions Program (LAMP) code to determine linear parameters of the heave, pitch, and roll response. A methodology for nonlinear system identification that combines the method of multiple scales and higher-order statistics was also proposed.

3.4.1. Coupled Roll-Pitch Equations of Motion

Considering the coupled roll-pitch equations of motion, (3.27) and (3.28) take the form

$$\left(J_{xx} + K_{\ddot\phi}\right)\ddot\phi + \sum_{k=1}^{K} c_k\dot\phi\big|\dot\phi\big|^{k-1} + \rho g\left\{ v_0(\overline z_{BO} - \overline z_G)\sin\phi + \frac{\partial I_{xx}}{\partial\theta}\bigg|_0 (\theta\sin\phi + \phi\sin\theta) \right.$$

$$\left. + \frac{1}{2}\frac{\partial^2 I_{xx}}{\partial\theta^2}\bigg|_0 \theta(\theta\sin\phi + \phi\sin\theta) + \frac{1}{2}\frac{\partial^2 I_{xx}}{\partial\phi^2}\bigg|_0 \phi^2\sin\phi \right\}$$

$$+ 2\rho g \phi \int_L y^2 \frac{\partial y}{\partial z} \eta(t) dx + \rho g \phi \int_L \left[2y \left(\frac{\partial y}{\partial z} \right)^2 + y \right] \eta^2(t) dx$$

$$+ 2\rho g \phi \theta \int_L \left[2xy \left(\frac{\partial y}{\partial z} \right)^2 + xy \right] \eta(t) dx = \Phi(t),$$

$$(J_{yy} + K_{\ddot{\theta}}) \ddot{\theta} + C_{\theta} \dot{\theta} + \rho g \left\{ v_0 (\overline{z}_{BO} - \overline{z}_G) \sin \theta + I_{yy}|_0 \sin \theta - \left. \frac{\partial I_{xy}}{\partial \phi} \right|_0 \phi \sin \phi + \left. \frac{\partial I_{yy}}{\partial \theta} \right|_0 \theta \sin \theta \right.$$

$$\left. - \left. \frac{\partial^2 I_{xy}}{\partial \phi \partial \theta} \right|_0 \phi \theta \sin \phi + \frac{1}{2} \left. \frac{\partial^2 I_{yy}}{\partial \phi^2} \right|_0 \phi^2 \sin \theta + \frac{1}{2} \left. \frac{\partial^2 I_{yy}}{\partial \theta^2} \right|_0 \theta^2 \sin \theta \right\}$$

$$+ 2\rho g \theta \int_L x^2 \frac{\partial y}{\partial z} \eta(t) dx + \rho g \phi^3 \int_L \left[2xy \left(\frac{\partial y}{\partial z} \right)^2 + xy \right] \eta(t) dx = \Theta(t).$$

$$(3.29)$$

Note that the nonlinear coupling terms may result in nonlinear internal resonances among pitch and roll motions (see, e.g., [2, 3]).

3.4.2. Roll Equation of Motion

The prediction of ship stability during the early stages of design is very important from the point of a vessel's safety. Of the six motions of a ship, the critical motion leading to capsize is the rolling motion. Thus for studying roll stability in beam seas one should consider the nonlinear roll equation

$$\left(J_{xx} + K_{\ddot{\phi}} \right) \ddot{\phi} + \sum_{k=1}^{K} c_k \dot{\phi} |\dot{\phi}|^{k-1} + \rho g \left[v_0 (\overline{z}_{BO} - \overline{z}_G) + \frac{1}{2} \left. \frac{\partial^2 I_{xx}}{\partial \phi^2} \right|_0 \phi^2 \right] \sin \phi$$

$$+ 2\rho g \phi \int_L y^2 \frac{\partial y}{\partial z} \eta(t) dx + \rho g \phi \int_L \left[2y \left(\frac{\partial y}{\partial z} \right)^2 + y \right] \eta^2(t) dx = \Phi(t).$$

$$(3.30)$$

In formulating the roll equation in beam seas one should realize that the hydrodynamic roll moments on the ship are dependent on the relative motion of ship and wave, rather than upon the absolute roll motion. In a beam sea the relative roll is defined as $(\phi - \alpha)$, where α is the local wave slope in a long-crested regular beam sea. In this case, the nonlinear equation of roll motion may be written in the form [86]

$$J_{xx} \ddot{\phi} = -\delta J_{xx} (\ddot{\phi} - \ddot{\alpha}) - H(\dot{\phi} - \dot{\alpha}) - E(\phi - \alpha) + B, \qquad (3.31)$$

where δJ_{xx} is the roll added inertia and B is the bias moment created by several sources such as a steady beam wind, a shift of cargo, water, or ice on deck. Setting $\phi_r = \phi - \alpha$, (3.31) takes the form (see, e.g., [87])

$$(J_{xx} + \delta J_{xx}) \ddot{\phi}_r + H \dot{\phi}_r + E \phi_r = B - J_{xx} \ddot{\alpha}. \qquad (3.32)$$

Wright and Marshfield [86] solved (3.32) for small nonlinear restoring moment and small linear and cubic damping near the resonance frequency using three different approximate techniques: perturbation method, averaging method, and harmonic balance. Lin and Salvesen [88] presented an assessment of the Large Amplitude Motion Program (LAMP) for evaluating ship performance in extreme seas. The study included a time domain simulation of a ship capsizing in beam seas. It was shown that capsizing can happen due to dynamic effects even for ships that satisfy the minimum righting arm requirement. Surendran and Reddy [89, 90] evaluated the performance of a ship in beam seas using strip theory. The critical condition in the rolling motion of a ship is when it is subjected to synchronous beam waves (i.e., the encounter frequency coincides with the wave frequency). They considered various representations of damping and restoring terms to identify the effect of wave amplitude, wave frequency, and metacentric height (was represented by a quintic polynomial).

Contento et al. [91] reported some results of experimental tests on nonlinear rolling in a regular beam sea of a Ro-Ro ship model by varying both the wave steepness and the wave frequency. They adopted a parameter estimation technique based on the least squares fitting of the stationary numerical solution of the nonlinear rolling motion differential equation. It was possible to extract information on the damping model and on the linear and nonlinear damping coefficients. These exhibit a quite strong dependence on frequency that reduces the efficiency of constant coefficients rolling equation to simulate large amplitude nonlinear rolling. The results indicate that a good quality prediction model of nonlinear rolling cannot be based on constant coefficients time-domain simulations. The analysis indicates also a marked dependence of the effective wave slope coefficient on wave amplitude. The effect of the excitation modeling on the fitting capability of the nonlinear roll motion equation to experimental data was studied by Francescutto et al. [92]. Several frequency dependent and constant effective wave slope coefficients were derived for five different scale models corresponding to different ship typologies by a parameter identification technique. Later, Francescutto and Contento [93] studied the steady rolling response in a regular beam sea of a 1 : 50 scale model of a destroyer in the bare hull condition. In view of the softening characteristics of the restoring moment, bifurcations with jump in amplitude and phase at two different wave frequencies were observed experimentally. Exact numerical solutions were used to obtain reliable values of the coefficients of the mathematical model to be used for the roll motion simulation.

Mahfouz [94] presented a robust method for the identification of linear and nonlinear damping and restoring parameters in the equation describing the rolling motion of a ship using only its measured response at sea. The parameters were identified using a combination of the random decrement technique, auto- and cross-correlation functions, a linear regression algorithm, and a neural-network technique. The proposed method would be particularly useful in identifying the nonlinear damping and restoring parameters for a ship rolling under the action of unknown excitations effected by a realistic sea.

3.5. Memory Effect

Note that the previous formulation did not account for the hydrodynamic memory effect. The hydrodynamic load due to the ship motion is a function of its frequency of oscillation. When the ship oscillates, waves will be generated on the free surface. As time increases, these waves will propagate outward from the body, but they continue to affect the fluid pressure

and hence the body force for all subsequent times [14]. In the time domain, this force or moment can be represented by a convolution integral of the impulse response function as outlined by Cummins [95], that is,

$$F_{ij} = -\alpha_{ij}(\infty)\dot{V}_j - \int_{-\infty}^{t} K_{ij}(t - \tau)V_j(\tau)d\tau, \quad i, j = 1, 2, 3, \tag{3.33}$$

where $i, j = 1, 2, 3$ indicate surge, sway, and yaw, respectively. $V_j(\tau)$ is the ship velocity along the axis j, α_{ij} is the ship added mass, and $K_{ij}(t - \tau)$ is the retardation function and can be expressed in terms of the velocity potential function φ as

$$K_{ij}(t - \tau) = \rho \iint_S \frac{\partial \varphi_j(t - \tau)}{\partial \tau} s_i \, d\sigma, \tag{3.34}$$

where s_i is the ith component of the normal vector of the surface element $d\sigma$. Chung and Bernitsas [96] evaluated these forces in details.

A component of this force initiated at a certain moment continues to attribute its influence on the system for a period of time. This is referred to as the hydrodynamic memory effect [96]. It was indicated that calculating this effect in the time domain is very time consuming. Tick [97] represented the convolution integral by a set of recursive differential equations with constant coefficients. These coefficients are determined by curve fitting the added mass and damping in the frequency domain. This method was used for estimating the memory effect on ship maneuvering by McCreight [98], a single-point mooring tanker by Jiang et al. [99] and Sharma et al. [100], and other motions by Schmiechen [101, 102].

4. Ship Roll Dynamic Stability

Traditional ship stability analysis compares the vessel righting arm curve to a standard or to a steady wind heeling moment (see, e.g., [103, 104]). Modern analysis methods of ship stability are based on analyzing the vessel's roll motion response either by simulation or using modern methods of dynamical systems. One method of analyzing nonlinear dynamical systems used by numerous researchers is the analysis of the so-called safe basin [105]. This method consists of numerically integrating a grid of initial conditions in order to determine which initial conditions will lead to bounded motions (safe basin) and which will yield unbounded motions (i.e., capsizing). There are two cases that are generally analyzed and these are the unbiased and biased cases. The unbiased case involves the symmetric ship while the biased case is for an asymmetric ship. The asymmetry may be due to a steady wind moment, cargo shifting, or an asymmetric ice accretion.

The upper bound of the wave excitation amplitude, beyond which the ship becomes dynamically unstable, is governed by both the damping factor and the excitation frequency. This is reflected by the reduction of the safe basin area with the excitation amplitude. The stability fraction known in the literature as the normalized safe basin area or Safe Basin Integrity Factor [106–109]. It is usually obtained by estimating the ratio of the area of the stable region in the phase plane (area of the safe basin) to the total area encompassed by the homoclinic orbit, which is defined to be the safe basin in the absence of excitation. The stability fraction is strongly dependent on the excitation amplitude. For excitation amplitudes

less than a critical value, governed by the excitation frequency, there is no erosion at all for the safe basin. Above this critical value of the excitation amplitude, the area of the safe basin shrinks and the stability fraction drops. As the excitation frequency changes so does the critical excitation amplitude.

Froude [110] observed that ships have undesirable roll characteristics when the frequency of a small free oscillation in pitch is twice the frequency of a small free oscillation in roll. It was Paulling and Rosenberg [8] who formulated the analytical modeling of the coupled roll-pitch motion. This coupled motion is described by a set of nonlinear equations. If the nonlinear effect of roll is neglected, the pitch equation of motion is reduced to a linear differential equation, which is free from roll motion terms. When pitch equation is solved, its response appears as a coefficient to the restoring moment in the roll equation of motion in the well-known Mathieu equation. Nayfeh et al. [1] analyzed the nonlinear coupling of the pitch and roll modes in regular seas when their frequencies are in the ratio of two to one. When the frequency of encounter (excitation frequency) is near the pitch frequency, the pitch mode is excited if the encountered wave amplitude (excitation amplitude) is small. As the excitation amplitude increases, the amplitude of the pitch mode increases until it reaches a critical small value. As the excitation amplitude increases further, the pitch amplitude reaches a saturated value, and all of the extra energy is transferred to the roll mode.

4.1. Stochastic Roll Stability

Sea waves are not sinusoidal and may cause severe or dangerous ship motions. The probabilistic theory of ship dynamics was documented by Price and Bishop [111] and Lloyd [112]. If the nonlinear effect of roll is neglected, the pitch equation of motion is reduced to a linear differential equation, which is free from roll motion terms. When the pitch equation is solved, its response appears as a coefficient to the restoring moment of the roll motion, and the roll equation of motion is reduced to the Mathieu equation

$$\ddot{\phi} + 2\bar{\zeta}\omega_n\dot{\phi} + \omega_n^2[1 + \varepsilon\Theta(t)]\phi = \varepsilon M(t), \tag{4.1}$$

where ϕ is the roll angle, $\Theta(t)$ represents the pitch angle which is assumed to be a random stationary process, $M(t)$ represents the wave random excitation, $\bar{\zeta}$ is a linear damping factor, ω_n is the natural frequency of the ship roll oscillation, and ε is a small parameter.

The analysis of the ship roll stability in an irregular sea was addressed by some investigators (see, e.g., [113–115]). The stochastic stability and response of the ship roll motion in random seas have been predicted analytically using the stochastic averaging method [46, 47, 49, 116]. Roberts [47] analyzed (4.1) for the stochastic stability and statistical response moments using the stochastic averaging method. The sample stability condition of the roll angle amplitude was obtained in the form

$$\bar{\zeta} > \frac{\varepsilon^2\omega_n S_\Theta(2\omega_n)}{8}, \tag{4.2}$$

where $S_\Theta(2\omega_n)$ is the power spectral density of the random pitch process $\Theta(t)$ at frequency $2\omega_n$. Condition (4.2) reveals that the onset of instability is not affected by the forcing excitation $M(t)$. If this excitation is removed, the probability density function of the response

degenerates into a delta function. The stability conditions of the first moment of the response amplitude is

$$\bar{\zeta} > \frac{3}{16}\varepsilon^2 \omega_n S_\Theta(2\omega_n).$$

(4.3)

The stability condition of the second moment is

$$\bar{\zeta} > \frac{1}{4}\varepsilon^2 \omega_n S_\Theta(2\omega_n).$$

(4.4)

These stability conditions apply only for the case of random sea waves in the absence of ice effects. Haddara [117] obtained the autocorrelation function of the roll motion in irregular seas.

4.2. Probabilistic Roll Dynamics

In addition to the modes of stochastic stability outlined in the previous subsection, it is important to examine the ship probabilistic description in random seas. It is of great importance to estimate the probability of capsizing. Equally important is to identify ship's parameters in roll motion. Different probability approaches have been found very effective in studying these issues. For example, the path integral technique was applied to the roll nonlinear motion of a ship in irregular waves by Kwon et al. [118]. The exciting moment due to irregular waves was modeled as a nonwhite noise. Both damping and nonlinear restoring functions were included with the equivalent white-noise intensity. Lin and Yim [119] developed a stochastic analysis to examine the properties of chaotic roll motion and capsize of ships subjected to a periodic excitation with a random noise disturbance. They used a generalized Melnikov method to provide an upper bound on the domain of the potential chaotic roll motion. The associated Fokker-Planck equation governing the evolution of the probability density function of the roll motion was numerically solved by the path integral solution procedure to obtain joint probability density functions in state space. A chaotic response was found to take place near the homoclinic and heteroclinic orbits. The heteroclinic model emulates symmetric vessel capsize and the homoclinic model represents a vessel with an initial bias caused by water on deck. It was found that the presence of noise enlarges the boundary of the chaotic domains and bridges coexisting attracting basins in the local regimes. The probability of capsize was considered as an extreme excursion problem with the time-averaged probability density function as an invariant measure. In the presence of noise, the numerical results revealed that all roll motion trajectories that visit the regime near the heteroclinic orbit will eventually lead to capsize.

Another version of the path integration approach based on the Gauss-Legendre quadrature integration rule was proposed by Gu [120]. It was applied for estimating the probability density of the nonlinear roll motion of ships in stochastic beam seas. The ship roll motion was described by a nonlinear random differential equation that includes a nonlinear damping moment and restoring moment. The results include the time evolution of the ship response probability density as well as the tail region, which is very important for the system reliability analysis. Gu [121] derived an approximate stationary probability density function

and stationary mean out-crossing rate of the response of nonlinear roll-motion subjected to additive stochastic white noise excitations.

Yim et al. [122, 123] developed an analytical approach for the identification of ship parameters and calibration of their prediction capability using experimental results. They examined a three-degree-of-freedom fully coupled roll-heave-sway model, which features realistic and practical high-degree polynomial approximations of rigid body motion relations, hydrostatic and hydrodynamic forces and moments. System parameters of the model were identified using physical model test results from several regular wave cases. The predictive capability of the model is then calibrated using results from a random wave test case. Yim et al. [123] presented a computationally quasi-two-degree-of-freedom stochastic model describing the coupled roll-heave motions and a stability analysis of barges in random seas. Stochastic differential equations governing the evolution of probability densities of roll-heave and roll responses were derived using the Fokker-Planck formulation. Numerical results of roll responses using direct simulation in the time domain and the path integral solution technique in the probability domain were compared to determine the effects of neglecting the influence of heave on roll motion.

The case of small ships with water on deck subjected to random beam waves described by a periodic force and white noise perturbation was considered by Liu and Yougang [124] using the path integral solution. The random Melnikov mean square criterion was used to determine the parameter domain for the ship's stochastic chaotic motion. The evolution of the probability density function of the roll response was calculated by solving the stochastic differential equations using the path integral method. It was found that in the probability density function of the system has two peaks for which the response of the system was found to jump from one peak to another for large amplitudes of periodic excitation. Mamontov and Naess [125] developed a combined analytical-numerical approach referred to as the successive-transition method, which is essentially a version of the path-integration solution and is based on an analytical approximation for the transition probability density. The method was applied to a one-dimensional nonlinear Ito's equation describing the velocity of a ship maneuvering along a straight line under the action of the stochastic drag due to wind or sea waves. It was also used for the problem of ship roll motion up to its possible capsizing. It was indicated that the advantage of the proposed successive transition is that it provides an account for the damping matrix in the approximation.

Haddara and Zhang [126] developed an expression for the joint conditional probability density function for the ship roll angle and roll velocity in beam seas. The joint probability density function was expressed as a double series in the nondimensional roll angle and roll velocity. Jiang et al. [127] examined ships capsizing in random beam seas using the Melnikov function and the concept of phase-flux rates. Damping and wave excitation moments were treated as perturbations since they are relatively small compared with inertial effects and hydrostatic righting moments. Safe and unsafe areas were defined in the phase plane of the unperturbed system model to distinguish the qualitatively different ship motions of capsize and noncapsize. They derived expressions for the phase space flux rate. The correlation of phase space flux and capsize was investigated through extensive simulations. It was shown that these analytical tools provide reliable predictive information regarding the likelihood of a vessel capsize in a given sea state. Gu [128] and Tang et al. [129] employed the Melnikov function and phase space flux to examine the nonlinear roll motion of a fishing ship in random beam seas. They showed that the phase space flux is monotonically increasing as the significant wave height increases, while the safe basin is decreasing rapidly.

(a)

(b)
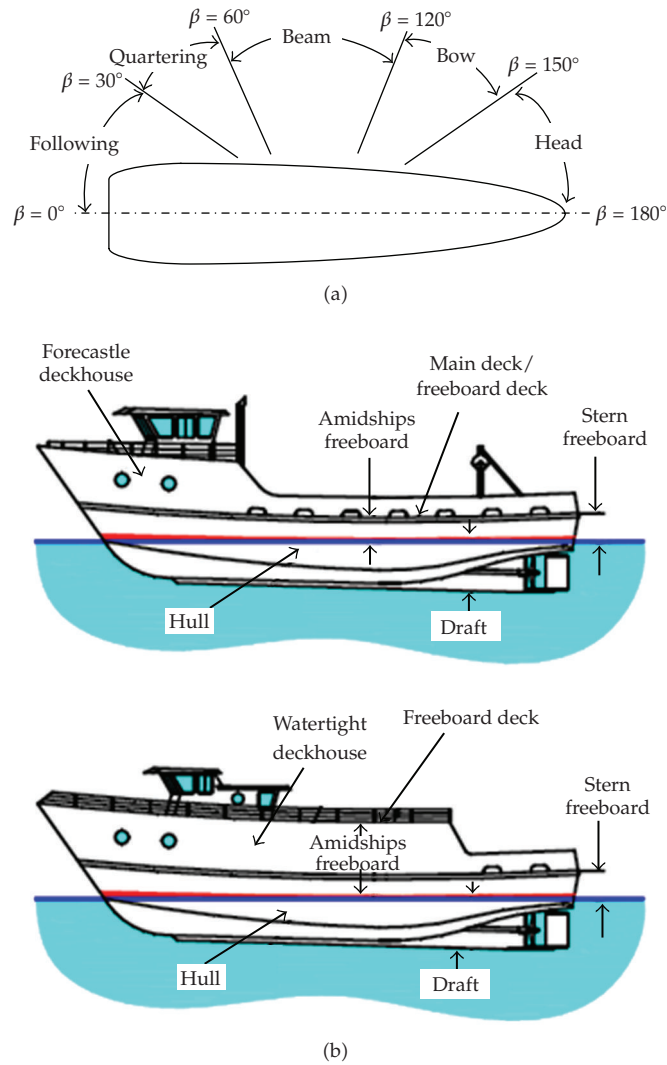
Figure 6: (a) Definition of incident wave directions. (b) Key parts of a ship structure.

Liu et al. [130] considered some methods for constructing safe basins and predicting the survival probability of ships in random beam waves. The nonlinear random roll differential equation was numerically solved in the time domain by considering the instantaneous state of ships and the narrowband wave energy spectrum. The safe basins were constructed for safe navigation, and the survival probability of ships was also estimated. In another work, Liu et al. [131] considered the random differential equation of roll motion in beam seas and the random Melnikov mean-square criterion was used to determine the threshold intensity for the onset of chaos. It was found that ships undergo stochastic chaotic motion when the real intensity of the white noise exceeds the threshold intensity. The stable probability density function of the roll response was found to possess two peaks and the random jump happened in the response of the system for high intensity of the white noise excitation.

Reliability of ship operations under Gaussian or nonGaussian random sea waves deals with the probability that the ship will not capsize. One may estimate the ship reliability in terms of the probabilistic characteristics of the time at which the roll motion first exits from the safe domain. When capsizing is defined by the first exit of response from a safe domain of operation, the reliability is referred to as a first-passage problem. For ships whose response is described by a Markov process, the mean value of the exit time is usually governed by a partial differential equation known as the Pontryagin-Vitt (PV) equation [132].

The first-passage problem of nonlinear roll oscillations in random seas has been considered by Roberts [46, 47], Cai and Lin [133], Cai et al. [134], and Moshchuk et al. [51, 135]. Roberts [46, 47] developed an approximate theory based on a combination of averaging techniques and the theory of Markov processes. His analysis resulted in a simple expression for the distribution of the ship roll angle. Cai et al. [134] adopted the same modeling and introduced a parametric excitation term. They used the modified version of quasiconservative averaging. Moshchuk et al. [51] determined the mean exit time of the perturbed ship motion by solving Pontryagin's partial differential equation using the method of asymptotic expansion. It was found that the mean exit time is extremely large for any excitation intensity less than a critical value above which it experiences exponential decay.

5. Closing Remarks

The nonlinear dynamic modeling of ship motions in roll, pitch, and heave has been formulated based on physical ground. The formulation has been adopted from the work of Neves and Rodríguez [11, 12]. One can use the coupled nonlinear equations motion to examine only the ship motion in roll oscillations under regular and random sea waves. Other issues related to this modeling deal with the effect of roll damping and hydrodynamic memory effect arising from the ship motion. An overview of the roll dynamic stability under random sea waves has been presented in terms of the sample stability condition and response statistical moments. Equally important are the probability of capsizing and the identification of parameters in roll motion. Different probability approaches have been found very effective in studying these issues such as the path integral technique and the generalized Melnikov method, which provides an upper bound on the domain of the potential chaotic roll motion. This paper has not addressed the interaction of roll dynamics with floating ice and the reader can consult the review article by Ibrahim et al. [136].

Appendix

Figure 6 clarifies some of the terminology defined in this Appendix.
Aft: toward the stern of the boat.
Beam: the width of a vessel also a structural component. Both Uses come from the Anglo-Saxon word beam, meaning, "tree".
Beam Sea: sea coming on the side of the ship.
Bilge: the lower point of inner hull of a ship.
Bow: the forward part of a boat. The word may come from the Old Icelandic bogr, meaning "shoulder."
Broach: the action of turning a vessel broadside to the waves.
Broadside: presenting the side of the ship.

Buoyancy: the upward push of water pressure, equal to the weight of the volume of water the ship displaces (W).

Capsize: to turn over.

Center of Buoyancy (B): the geometric center of the submerged hull, acting vertically upward.

Center of Flotation (F): the geometric center of the waterline plane, about which the ship trims fore and aft.

Center of Gravity (G): the center of all mass of the ship, acting vertically downward.

Displacement Volume (v): the volume of the underwater hull at any given waterline.

Displacement (W): the weight of water of the displaced volume of the ship, which equals the weight of the ship and cargo.

Draft: the depth of water a boat draws.

Fathom: six feet.

Following Sea: sea coming on the stern.

Forecastle: pronounced "fo'c's'l" and usually now spelled that way. Now the foredeck of a vessel, the term originally referred to a raised and fortified platform at the ship's bow. Used by archers in combat at sea as early as the 13th century.

Freeboard: that part of a ships sides above water, from the Anglo-Saxon framebord, meaning "the frame's side."

Head: (1) the uppermost or forward-most part of a ship (or of some specific part of a ship, such as the masthead, beakhead, stemhead, or whatever. (2) The bathroom. In the age of sail, the crew was quartered forward in the forecastle, and their latrine was located on the beakhead, over hanging the water (for obvious reasons).

Heading: the direction in which a vessel's bow points at any given time.

Headway: the forward motion of a boat. Opposite of sternway.

Heel: constant roll angle—such as caused by a side wind or turning of the vessel.

Hull: the main body of a vessel.

Keel: the centerline of a boat running fore and aft; the backbone of a vessel.

Knot: a measure of speed equal to one nautical mile (6076 feet) per hour.

Lee: the side sheltered from the wind.

Leeward: the direction away from the wind. Opposite of windward.

Leeway: the sideways movement of the boat caused by either wind or current.

List, Heel, and Roll: it is both a noun and a verb referring to ships upping to one side or the other due to poor trim, shifting cargo, or sinking. The word comes from the Anglo-Saxon lystan, meaning "to lean." Angular transverse inclinations. List describes a static inclination such as list due to side damage. Heel describes a temporary inclination generally involving motion, such as wind or turning, while roll indicates periodic inclination from side to side such as wave action.

Metacenter (M): when the ship is inclined at small angles, the metacenter is the intersection of the buoyant force with the ship centerline. If the metacenter is above the center of gravity then the ship is stable.

Midship: approximately in the location equally distant from the bow and stern.

Nautical Mile: one minute of latitude; approximately 6076 feet: about 1/8 longer than the statute mile of 5280 feet.

Naval Architecture: ship design: especially hull design, overall layout with attention to stability, sea keeping, and strength.

Port: the left side of a boat looking forward.

Quarter: the sides of a boat aft of amidships.

Quarter Sea: sea coming on a boat's quarter.

Reserve Buoyancy: the watertight volume between the waterline and the uppermost continuous watertight deck.

Righting Arm (also *Restoring Lever*): it is the horizontal distance between the vertical line passing through the buoyant center and the vertical line passing through the ship center of gravity.

Starboard: the right side of a boat when looking forward.

Stern: the rear of any vessel. The word came from the Norse Stjorn (pronounced "Styorn"), meaning "steering." It is the after part of the boat.

Thwartships: it means across the ship.

Trim: longitudinal tilt. Stern draft, bow draft.

Wake: moving waves, track, or path that a boat leaves behind it, when moving across the waters.

Waterline: a line painted on a hull which shows the point to which a boat sinks when it is properly trimmed.

Way: movement of a vessel through the water such as headway, sternway, or leeway.

Windward: toward the direction from which the wind is coming.

Yaw: to swing or steer off course, as when running with a quartering sea.

Acknowledgment

References

[1] A. H. Nayfeh, D. T. Mook, and L. R. Marshall, "Nonlinear coupling of pitch and roll modes in ship motions," *Journal of Hydronautics*, vol. 7, no. 4, pp. 145–152, 1973.

[2] A. H. Nayfeh, D. T. Mook, and L. R. Marshall, "Perturbation-energy approach for the development of the nonlinear equations of ship motion," *Journal of Hydronautics*, vol. 8, no. 4, pp. 130–136, 1974.

[3] A. H. Nayfeh and D. T. Mook, *Nonlinear Oscillations*, Pure and Applied Mathematics, John Wiley & Sons, New York, NY, USA, 1979.

[4] B. M. Suleiman, *Identification of finite-degree-of-freedom models for ship motions*, Ph.D. thesis, Virginia Polytechnic Institute and State University, Blacksburg, Va, USA, 2000.

[5] T. I. Fossen, *Guidance and Control of Ocean Vehicles*, John Wiley & Sons, New York, NY, USA, 1994.

[6] E. M. Lewandowski, *The Dynamics of Marine Craft Maneuvering and Seakeeping*, vol. 2 of *Advanced Series on Ocean Engineering*, World Scientific, Singapore, 2004.

[7] T. Perez, *Ship Motion Control; Course Keeping and Roll Stabilization Using Rudder and Fins*, Advances in Industrial Control, Springer, Berlin, Germany, 2005.

[8] J. R. Paulling and R. M. Rosenberg, "On unstable ship motions resulting from nonlinear coupling," *Journal of Ship Research*, vol. 3, no. 1, pp. 36–46, 1959.

[9] W. D. Kinney, "On the unstable rolling motions of ships resulting from nonlinear coupling with pitch including the effect of damping in roll," Institute of Engineering Research, University of California, Berkeley, Calif, USA, 1961.

[10] M. R. Haddara, "On the stability of ship motion in regular oblique waves," *International Shipbuilding Progress*, vol. 18, no. 207, pp. 416–434, 1971.

[11] M. A. S. Neves and C. A. Rodríguez, "On unstable ship motions resulting from strong non-linear coupling," *Ocean Engineering*, vol. 33, no. 14-15, pp. 1853–1883, 2006.

[12] M. A. S. Neves and C. A. Rodríguez, "Influence of non-linearities on the limits of stability of ships rolling in head seas," *Ocean Engineering*, vol. 34, no. 11-12, pp. 1618–1630, 2007.

[13] J. H. Vugts, "The hydrodynamic forces and ship motions on oblique waves," Tech. Rep. 150 S, Netherlands Ship Research Center, 1971.

[14] J. N. Newman, *Marine Hydrodynamics*, The MIT Press, Cambridge, Mass, USA, 1977.

[15] L. V. Edward, *Principles of Naval Architecture*, The Society of Naval Architects and Marine Engineers, Jersey City, NJ, USA, 1989.

[16] H. Chu, *Numerical modeling of ship motions in shallow water waves*, Ph.D. thesis, Florida Institute of Technology, Melbourne, Fla, USA, 1998.

[17] Y. Lee, *Non linear ship motion models to predict capsize in regular beam seas*, Ph.D. thesis, University of Michigan, Ann Arbor, Mich, USS, 2001.

[18] M. S. Khalid, *Simulation of Euler equations of motion and blended—nonlinear hydrodynamics for multi-hulled vessels*, Ph.D. thesis, University of Michigan, Ann Arbor, Mich, USA, 2007.

[19] M. A. Abkowitz, "Lectures on ship hydrodynamics—steering and maneuverability," Tech. Rep. Hy-5, Hydro- and Aerodynamics Laboratory, Lyngby, Denmark, 1964.

[20] M. A. Abkowitz, "System identification techniques for ship maneuvering trials," in *Proceedings of the Symposium on Control Theory and Navy Applications*, pp. 337–393, Monterey, Calif, USA, 1975.

[21] M. A. Abkowitz, "Measurements of hydrodynamic characteristics from ship maneuvering trials by system identification," *Transactions of the Society of Architects and Marine Engineering*, vol. 88, pp. 283–318, 1981.

[22] W. Y. Hwang, *Application of system identification to ship maneuvering*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Mass, USA, 1980.

[23] C. G. Källström and K. J. Åström, "Experiences of system identification applied to ship steering," *Automatica*, vol. 17, no. 1, pp. 187–198, 1981.

[24] K.-H. Son and K. Nomoto, "On the coupled motion of steering and rolling of a high-speed container ship," *Naval Architecture and Ocean Engineering*, vol. 20, pp. 73–83, 1982.

[25] A. Ross, *Nonlinear Manoeuvring Models for Ships: a Lagrangian Approach*, Ph.D. thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2008.

[26] G. Kirchhoff, *Advanced Engineering Mathematics*, John Wiley & Sons, New York, NY, USA, 1869.

[27] L. M. Milne-Thomson, *Theoretical Hydrodynamics*, MacMillan, London, UK, 1968.

[28] T. I. Fossen, *Marine Control Systems: Guidance, Navigation and Control of Ships*, Rigs and Underwater Vehicles, Marine Cyberentics AS, Trondheim, Norway, 2002.

[29] Z. Rong, *Studies of weak and strong nonlinear sea loads on floating marine structures*, M.S. thesis, Norges Tekniske Hoegskol, Trodheim, Norway, April 1994.

[30] L. Alford, A. Banik, V. Belenky, et al., "Discussion on analytical approaches to the study of vessel dynamics-outcomes of the two-part mini symposium at the 2007 SIAM conference on applications of dynamical systems," *Marine Technology*, vol. 45, no. 4, pp. 211–220, 2008.

[31] J. O. de Kat and J. R. Paulling, "The simulation of ship motions and capsizing in severe seas," *Transactions of the Society of Architects and Marine Engineering*, vol. 97, pp. 139–168, 1989.

[32] J. A. Witz, C. B. Ablett, and J. H. Harrison, "Roll response of semisubmersibles with nonlinear restoring moment characteristics," *Applied Ocean Research*, vol. 11, pp. 153–166, 1989.

[33] J.-P. F. Denise, "On the roll motion of barges," *Transactions of the Royal Institution of Naval Architects*, vol. 125, pp. 255–268, 1983.

[34] D. W. Bass and M. R. Haddara, "On the modeling of the nonlinear damping moment for the rolling motion of a ship," in *Proceedings of the International Association of Scientific Technology (IASTED) Symposium: Identification, Modeling and Simulation*, pp. 346–349, Paris, France, 1987.

[35] D. W. Bass and M. R. Haddara, "Nonlinear models of ship roll damping," *International Shipbuilding Progress*, vol. 35, no. 401, pp. 5–24, 1988.

[36] M. Taylan, "The effect of nonlinear damping and restoring in ship rolling," *Ocean Engineering*, vol. 27, no. 9, pp. 921–932, 2000.

[37] J. F. Dalzell, "A note on the form of ship roll damping," *Journal of Ship Research*, vol. 22, no. 3, pp. 178–185, 1978.

[38] A. Cardo, A. Francescutto, and R. Nabergoj, "On damping models in free and forced rolling motion," *Ocean Engineering*, vol. 9, no. 2, pp. 171–179, 1982.

[39] J. B. Mathisen and W. G. Price, "Estimation of ship roll damping coefficients," *Transactions Royal Institution of Naval Architects*, vol. 127, pp. 295–307, 1984.

[40] A. F. El-Bassiouny, "Nonlinear analysis for a ship with a general roll-damping model," *Physica Scripta*, vol. 75, no. 5, pp. 691–701, 2007.

[41] M. Taylan, "Effect of forward speed on ship rolling and stability," *Mathematical and Computational Applications*, vol. 9, no. 2, pp. 133–145, 2004.

[42] A. H. Nayfeh and B. Balachandran, *Applied Nonlinear Dynamics*, Wiley Series in Nonlinear Science, John Wiley & Sons, New York, NY, USA, 1995.

[43] L. Arnold, I. Chueshov, and G. Ochs, "Stability and capsizing of ships in random sea-a survey," *Nonlinear Dynamics*, vol. 36, no. 2–4, pp. 135–179, 2004.

[44] S. Surendran, S. K. Lee, J. V. R. Reddy, and G. Lee, "Non-linear roll dynamics of a Ro-Ro ship in waves," *Ocean Engineering*, vol. 32, no. 14-15, pp. 1818–1828, 2005.

[45] G. Bulian, "Nonlinear parametric rolling in regular waves–a general procedure for the analytical approximation of the GZ curve and its use in time domain simulations," *Ocean Engineering*, vol. 32, no. 3-4, pp. 309–330, 2005.

[46] J. B. Roberts, "A stochastic theory for nonlinear ship rolling in irregular seas," *Journal of Ship Research*, vol. 26, no. 4, pp. 229–245, 1982.

[47] J. B. Roberts, "Effect of parametric excitation on ship rolling motion in random waves," *Journal of Ship Research*, vol. 26, no. 4, pp. 246–253, 1982.

[48] J. Falzarano and F. Zhang, "Multiple degree of freedom global analysis of transient ship rolling motion," in *Proceedings of the ASME Winter Annual Meeting on Nonlinear Dynamics of Marine Vehicles*, J. M. Falzarano and F. Papoulias, Eds., vol. 51, pp. 57–72, New York, NY, USA, 1993.

[49] X. Huang, X. Gu, and W. Bao, "The probability distribution of rolling amplitude of a ship in high waves," in *Proceedings of the 5th International Conference on Stability of Ships and Ocean Vehicles*, Melbourne, Fla, USA, 1994.

[50] I. Senjanović, G. Ciprić, and J. Parunov, "Survival analysis of fishing vessels rolling in rough seas," *Philosophical Transactions of the Royal Society A*, vol. 358, no. 1771, pp. 1943–1965, 2000.

[51] N. K. Moshchuk, R. A. Ibrahim, R. Z. Khasminskii, and P. L. Chow, "Asymptotic expansion of ship capsizing in random sea waves-I. First-order approximation," *International Journal of Non-Linear Mechanics*, vol. 30, no. 5, pp. 727–740, 1995.

[52] G. H. Bryan, "The action of bilge keels," *Transactions of the Institute of Naval Architects*, vol. 42, pp. 198–239, 1900.

[53] T. Hishida, "Study on the wave making resistance for rolling ships—parts 1 and 2," *Journal of Society of Naval Architecture of Japan*, vol. 85, pp. 45–61, 1952.

[54] T. Hishida, "Study on the wave making resistance for rolling ships—parts 3 and 4," *Journal of Society of Naval Architecture of Japan*, vol. 86, pp. 277–283, 1954.

[55] T. Hishida, "Study on the wave making resistance for rolling ships—parts 5 and 6," *Journal of Society of Naval Architecture of Japan*, vol. 87, pp. 67–68, 1955.

[56] H. Kato, "On the frictional resistance of the roll of ships," *The Science of Naval Architects of Japan*, vol. 102, pp. 115–123, 1958.

[57] M. Martin, "Roll damping due to bilge keels," Tech. Rep., Iowa University, Institute of Hydraulics Research, Iowa City, Iowa, USA, 1958.

[58] M. Tanaka, "A study on the bilge keel—part 1," *Journal of Society of Naval Architecture of Japan*, vol. 101, pp. 99–105, 1957.

[59] M. Tanaka, "A study on the bilge keel—part 2," *Journal of Society of Naval Architecture of Japan*, vol. 103, p. 343, 1958.

[60] M. Tanaka, "A study on the bilge keel—part 3," *Journal of Society of Naval Architecture of Japan*, vol. 105, pp. 27–32, 1959.

[61] M. Tanaka, "A study on the bilge keel—part 4," *Journal of Society of Naval Architecture of Japan*, vol. 104, pp. 205–212, 1961.

[62] H. Kato, "Effect of bilge keels on the rolling of ships," *Journal of Society of Naval Architecture of Japan*, vol. 117, pp. 93–114, 1965.

[63] C. G. Moody, "The effect of bilge keels on the roll damping characteristics of the large bulk-oil carrier SS World Glory," Tech. Rep., David Taylor Model Basin, West Bethesda, Md, USA, 1961.

[64] L. E. Motter, "Ship motion and bilge keel studies for the CVA," Tech. Rep. OPNAVINST S5513.3A-17, David Taylor Model Basin, West Bethesda, Md, USA, 1967.

[65] H. Jones, "Bilge keel size investigation for an attack aircraft carrier (CVA-41)," Tech. Rep. DTNSRDC/SPD-0861-01, David Taylor Model Basin, West Bethesda, Md, USA, 1978.

[66] H. Jones, "Bilge keel size investigation for an attack aircraft carrier and vstol (vertical short takeoff and landing," Tech. Rep. DTNSRDC/SPD-0861-02, David Taylor Model Basin, West Bethesda, Md, USA, 1979.

[67] R. E. D. Bishop and W. G. Price, "A note on structural damping of ship hulls," *Journal of Sound and Vibration*, vol. 56, no. 4, pp. 495–499, 1978.

[68] Y. Himeno, "Prediction of ship roll damping: state-of-the-art," Tech. Rep. 239, Department of Naval Architecture and Marine Engineering, The University of Michigan, Ann Arbor, Mich, USA, September 1981.

[69] R.-Q. Lin, W. Kuang, and A. M. Reed, "Numerical modeling of nonlinear interactions between ships and surface gravity waves—part I: ship waves in calm water," *Journal of Ship Research*, vol. 49, no. 1, pp. 1–11, 2005.

[70] R.-Q. Lin and W. Kuang, "Nonlinear ship-wave interaction model—part 2: ship boundary condition," *Journal of Ship Research*, vol. 50, no. 2, pp. 181–186, 2006.

[71] R.-Q. Lin and W. Kuang, "Modeling nonlinear roll damping with a self-consistent, strongly nonlinear ship motion model," *Journal of Marine Science and Technology*, vol. 13, no. 2, pp. 127–137, 2008.

[72] N. Salvesen, "Five years of numerical naval ship hydrodynamics at DTNSRDC," Tech. Rep. DTNSRDC-79/082, David W. Taylor Naval Ship Research and Development Center, Bethesda, Md, USA, 1979.

[73] W. M. Lin, M. J. Meinhold, and N. Salvesen, "Nonlinear time-domain simulation of fishing vessel capsizing in quartering seas," Tech. Rep. SAIC-CG-D-17-98, Science Applications International, Annapolis, Md, USA, April 1998.

[74] S. Chakrabarti, "Empirical calculation of roll damping for ships and barges," *Ocean Engineering*, vol. 28, no. 7, pp. 915–932, 2001.

[75] Y. Ikeda, "Roll damping of ships," in *Proceedings of the Ship Motions, Wave Loads and Propulsive Performance in a Seaway, 1st Marine Dynamics Symposium*, pp. 241–250, The Society of Naval Architecture in Japan, 1984.

[76] Y. Ikeda, Y. Himeno, and N. Tanaka, "Components of roll damping of ships at forward speed," *Journal of Society of Naval Architecture of Japan*, vol. 143, pp. 121–133, 1978.

[77] Y. Ikeda, Y. Himeno, and N. Tanaka, "A prediction method for ships roll damping," Tech. Rep. 00405, Department of Naval Architecture, University of Osaka Prefecture, Osaka, Japan, 1978.

[78] Y. Ikeda, T. Fujiwara, and T. Katayama, "Roll damping of a sharp-cornered barge and roll control by a new-type stabilizer," in *Proceedings of the 3rd International Offshore and Polar Engineering Conference*, pp. 634–639, Singapore, June 1993.

[79] J. P. Morison, M. P. O'Brien, J. W. Johnson, and S. A. Schaaf, "The force exerted by surface waves on a pile," *Petroleum Transactions*, vol. 189, pp. 149–154, 1950.

[80] T. Sarpkaya and M. Isaacson, *Mechanics of Wave Forces on Offshore Structures*, Van Nostrand Reinhold, New York, NY, USA, 1981.

[81] M. R. Haddara, "On the random decrement for nonlinear rolling motion," in *Proceedings of the 11th International Offshore Mechanics and Arctic Engineering Symposium*, vol. 2 of *Safety and Reliability*, pp. 321–324, Calgary, Canada, 1992.

[82] X. Wu, L. Tao, and Y. Li, "Nonlinear roll damping of ship motions in waves," *Journal of Offshore Mechanics and Arctic Engineering*, vol. 127, no. 3, pp. 205–211, 2005.

[83] N. Salvesen, O. E. Tuck, and O. Faltinsen, "Ship motions and sea loads," *Transactions Society of Architects and Marine Engineering*, vol. 78, pp. 250–287, 1970.

[84] W. G. Meyers, D. J. Sheridan, and N. Salvesen, "Manual: NSRDC ship-motion and sea-load computer program," Tech. Rep. 3376, Naval Ship Research and Development Center, Carderock, Md, USA, 1975.

[85] A. H. Nayfeh, S. A. Ragab, and A. A. Al-Maaitah, "Effect of bulges on the stability of boundary layers," *Physics of Fluids*, vol. 31, no. 4, pp. 796–806, 1988.

[86] J. H. G. Wright and W. B. Marshfield, "Ship roll response and capsize beviour in beam seas," *Transactions of the Institute of Naval Architects*, no. 3, pp. 129–148, 1979.

[87] C. Kuo and Y. Odabasi, "Theoretical studies on intact stability of ships (Phase 2)," Department of Shipbuilding and Naval Architecture, University of Strathclyde, October 1974.

[88] W. M. Lin and N. Salvesen, "Nonlinear time-domain simulation of ship capsizing in beam seas," Final Report, Science Applications International, Annapolis, Md, USA, November 1997.

[89] S. Surendran and J. V. R. Reddy, "Roll dynamics of a Ro-Ro ship," *International Shipbuilding Progress*, vol. 49, no. 4, pp. 301–320, 2002.

[90] S. Surendran and J. V. R. Reddy, "Numerical simulation of ship stability for dynamic environment," *Ocean Engineering*, vol. 30, no. 10, pp. 1305–1317, 2003.

[91] G. Contento, A. Francescutto, and M. Piciullo, "On the effectiveness of constant coefficients roll motion equation," *Ocean Engineering*, vol. 23, no. 7, pp. 597–618, 1996.

[92] A. Francescutto, G. Contento, M. Biot, L. Schiffrer, and G. Caprino, "Effect of excitation modeling in the parameter estimation of nonlinear rolling," in *Proceedings of the 8th International Offshore and Polar Engineering Conference*, vol. 3, pp. 490–498, Monteral, Canada, May 1998.

[93] A. Francescutto and G. Contento, "Bifurcations in ship rolling: experimental results and parameter identification technique," *Ocean Engineering*, vol. 26, no. 11, pp. 1095–1123, 1999.

[94] A. B. Mahfouz, "Identification of the nonlinear ship rolling motion equation using the measured response at sea," *Ocean Engineering*, vol. 31, no. 17-18, pp. 2139–2156, 2004.

[95] W. E. Cummins, "The impulse response function and ship motions," *Schiffstechnik*, vol. 9, pp. 101–109, 1962.

[96] J.-S. Chung and M. M. Bernitsas, "Hydrodynamic memory effect on stability, bifurcation, and chaos of two-point mooring systems," *Journal of Ship Research*, vol. 41, no. 1, pp. 26–44, 1997.

[97] L. J. Tick, "Differential equations with frequency-dependent coefficients," *Journal of Ship Research*, vol. 3, no. 2, pp. 45–46, 1959.

[98] W. R. McCreight, "Ship maneuvering in waves," in *Proceedings of the 16th Symposium on Naval Hydrodynamics: Ship Wakes; Large Amplitude Waves; Real Fluid Effects in Ship Hydrodynamics; Fluid-Structure Interaction; Frontier Problems in Hydrodynamics*, pp. 456–469, National Research Council, US Naval Studies Board, University of California, 1986, US Office of Naval Research, USA.

[99] T. Jiang, T. E. Schellin, and S. D. Sharma, "Maneuvering simulation of a tanker moored in a steady current including hydrodynamic memory effect and stability analysis," in *Proceedings of the Royal Institution of Naval Architects (RINA) International Conference on Ship Maneuvering*, London, UK, 1987.

[100] S. D. Sharma, T. Jiang, and T. E. Schellin, "Dynamic instability and chaotic motions of a single-point-moored tanker," in *Proceedings of the 17th Office of Naval Research (ONR) Symposium on Naval Hydrodynamics*, The Hague, The Netherlands, 1988.

[101] M. Schmiechen, "Zur kollisiondynamik von Schiffen," *Jahrbuch der Schiffbautechnischen Gesellschaft*, vol. 68, 1974.

[102] M. Schmiechen, "Equation for non-quasi-steady ship motion," in *Proceedings of the 14th International Towing Tank Conference*, Ottawa, Canada, 1975.

[103] A. Francescutto, A. Serra, and S. Scarpa, "A critical analysis of weather criterion for intact stability of large passenger vessels," in *Proceedings of the International Conference on Offshore Mechanics and Arctic Engineering (OMAE '01)*, vol. 1, pp. 829–836, 2001.

[104] K. Spyrou, "A basis for developing a rational alternative to the weather criterion: problems and capabilities," in *Proceedings of the 6th International Ship Stability Workshop*, October 2002.

[105] J. M. T. Thompson, "Designing against capsize in beam seas: recent advances and new insights," *Applied Mechanics Reviews*, vol. 50, no. 5, pp. 307–324, 1997.

[106] A. N. Lansbury, J. M. T. Thompson, and H. B. Stewart, "Basin erosion in the twin-well Duffing oscillator: two distinct bifurcations scenarios," *International Journal of Bifurcation and Chaos*, vol. 2, pp. 505–532, 1992.

[107] J. M. T. Thompson, "Loss of engineering integrity due to the erosion of absolute and transient basin boundaries," in *Proceedings of IUTAM Symposium 'Nonlinear Dynamics in Engineering Systems'*, W. Schiehlen, Ed., Springer, Berlin, Germany, August 1989.

[108] J. M. T. Thompson, "Chaotic phenomena triggering the escape from a potential well," *Proceedings Royal Society London A*, vol. 421, pp. 195–225, 1989.

[109] J. M. T. Thompson and F. A. McRobie, "Intermediate bifurcations and global dynamics of driven oscillators," in *Proceedings of the 1st European Nonlinear Oscillations Conference*, E. Kreuzer and G Schmidt, Eds., vol. 72 of *Mathematical Research*, pp. 107–128, Academie, Hamburg, Germany, 1993.

[110] W. Froude, "Remarks on Mr. Scott Russell's paper on rolling," *Transactions of the Institute of Naval Research*, vol. 4, pp. 232–275, 1863.

[111] W. G. Price and R. E. D. Bishop, *Probabilistic Theory of Ship Dynamics*, Chapman & Hall, London, UK, 1974.

[112] A. R. J. M. Lloyd, *Seakeeping—Ship Behavior in Rough Weather*, Ellis Howood, Chichester, UK, 1989.

[113] W. G. Price, "A stability analysis of the roll motion of a ship in an irregular seaway," *International Shipbuilding Progress*, vol. 22, no. 247, pp. 103–112, 1975.

[114] M. R. Haddara, "A study of stability of the mean and variance of rolling motion in random waves," in *Proceedings of the International Conference on Stability of Ships and Ocean Vehicles*, University of Strathclyde, Glasgow, UK, 1975.

[115] P. K. Muhuri, "A study of the stability of the rolling motion of a ship in an irregular seaway," *International Shipbuilding Progress*, vol. 27, no. 310, pp. 139–142, 1980.

[116] J. B. Roberts and N. M. C. Dacunha, "Roll motion of a ship in random beam waves: comparison between theory and experiment," *Journal of Ship Research*, vol. 29, no. 2, pp. 112–126, 1985.

[117] M. R. Haddara, "Note on the power spectrum of nonlinear rolling motion," *International Shipbuilding Progress*, vol. 30, no. 342, pp. 41–44, 1983.

[118] S. Kwon, D. Kim, and J. Chung, "Application of path integral solution to ship rolling motion," in *Proceedings of the 3rd International Offshore and Polar Engineering Conference (ISOPE '93)*, pp. 657–660, Singapore, 1993.

[119] H. Lin and S. C. S. Yim, "Chaotic roll motion and capsize of ships under periodic excitation with random noise," *Applied Ocean Research*, vol. 17, no. 3, pp. 185–204, 1995.

[120] J.-Y. Gu, "Calculation of ship rolling probability using a new path integration method," *Journal of Ship Mechanics*, vol. 10, no. 6, pp. 43–52, 2006.

[121] J.-Y. Gu, "Probability analysis of ship nonlinear roll-motion excited by white noise," *Journal of Ship Mechanics*, vol. 10, no. 3, pp. 17–25, 2006.

[122] S. C. Yim, T. Nakhata, W. A. Bartel, and E. T. Huang, "Coupled nonlinear barge motions—part I: deterministic models development, identification and calibration," in *Proceedings of the 23rd ASME International Conference on Offshore Mechanics and Arctic Engineering (OMAE '04)*, vol. 1, pp. 281–291, Vancouver, Canada, 2004.

[123] S. C. Yim, T. Nakhata, and E. T. Huang, "Coupled nonlinear barge motions—part II: deterministic models stochastic models and stability analysis," in *Proceedings of the 23rd ASME International Conference on Offshore Mechanics and Arctic Engineering (OMAE '04)*, vol. 1A, pp. 293–302, Vancouver, Canada, 2004.

[124] L. Liu and T. Yougang, "Stability of ships with water on deck in random beam waves," *Journal of Vibration and Control*, vol. 13, no. 3, pp. 269–280, 2007.

[125] E. Mamontov and A. Naess, "An analytical-numerical method for fast evaluation of probability densities for transient solutions of nonlinear Ito's stochastic differential equations," *International Journal of Engineering Science*, vol. 47, no. 1, pp. 116–130, 2009.

[126] M.R. Haddara and Y. Zhang, "On the joint probability density function of non-linear rolling motion," *Journal of Sound and Vibration*, vol. 169, no. 4, pp. 562–569, 1994.

[127] C. Jiang, A. W. Troesch, and S. W. Shaw, "Highly nonlinear rolling motion of biased ships in random beam seas," *Journal of Ship Research*, vol. 40, no. 2, pp. 125–135, 1996.

[128] J.-Y. Gu, "Nonlinear rolling motion of ship in random beam seas," *Journal of Marine Science and Technology*, vol. 12, no. 4, pp. 273–279, 2004.

[129] Y.-G. Tang, J.-Y. Gu, H.-Y. Zheng, and H.-X. Li, "Study on the ship capsize in random beam seas using Melnikov method," *Journal of Ship Mechanics*, vol. 8, no. 5, pp. 27–34, 2004 (Chinese).

[130] L.-Q. Liu, Y.-G. Tang, H.-Y. Zheng, and J.-Y. Gu, "Analysis method of capsizing probability in time domain for ships in random beam waves," *Journal of Tianjin University Science and Technology*, vol. 39, no. 2, pp. 165–169, 2006 (Chinese).

[131] L.-Q. Liu, Y.-G. Tang, and H.-X. Li, "Stochastic chaotic motion of ships in beam seas," *Journal of Marine Science and Technology*, vol. 15, no. 2, pp. 123–128, 2007.

[132] L. S. Pontryagin, A. A. Andonov, and A. A. Vitt, "On statistical consideration of dynamical systems," *Journal of Experimental and Theoretical Physics*, vol. 3, no. 3, pp. 165–172, 1933.

[133] G. Q. Cai and Y. K. Lin, "On statistics of first-passage failure," *Journal of Applied Mechanics*, vol. 61, pp. 93–99, 1994.

[134] G. Q. Cai, J. S. Yu, and Y. K. Lin, "Ship rolling in random sea," in *Stochastic Dynamics and Reliability of Nonlinear Ocean Systems*, R. A. Ibrahim and Y. K. Lin, Eds., vol. 77, pp. 81–88, ASME, New York, NY, USA, 1994.

[135] N. K. Moshchuk, R. Z. Khasminskii, R. A. Ibrahim, and P. L. Chow, "Asymptotic expansion of ship capsizing in random sea waves-II. Second-order approximation," *International Journal of Non-Linear Mechanics*, vol. 30, no. 5, pp. 741–757, 1995.

[136] R. A. Ibrahim, N. G. Chalhoub, and J. Falzarano, "Interaction of ships and ocean structures with ice loads and stochastic ocean waves," *Applied Mechanics Reviews*, vol. 60, no. 5, pp. 246–290, 2007.

Review Article

Some Applications of Fractional Calculus in Engineering

J. A. Tenreiro Machado, Manuel F. Silva, Ramiro S. Barbosa, Isabel S. Jesus, Cecília M. Reis, Maria G. Marcos, and Alexandra F. Galhano

Institute of Engineering of Porto, Porto, Portugal

Correspondence should be addressed to J. A. Tenreiro Machado, jtm@isep.ipp.pt

Fractional Calculus (FC) goes back to the beginning of the theory of differential calculus. Nevertheless, the application of FC just emerged in the last two decades, due to the progress in the area of chaos that revealed subtle relationships with the FC concepts. In the field of dynamical systems theory some work has been carried out but the proposed models and algorithms are still in a preliminary stage of establishment. Having these ideas in mind, the paper discusses FC in the study of system dynamics and control. In this perspective, this paper investigates the use of FC in the fields of controller tuning, legged robots, redundant robots, heat diffusion, and digital circuit synthesis.

1. Introduction

The generalization of the concept of derivative $D^\alpha[f(x)]$ to noninteger values of α goes back to the beginning of the theory of differential calculus. In fact, Leibniz, in his correspondence with Bernoulli, L'Hôpital and Wallis (1695), had several notes about the calculation of $D^{1/2}[f(x)]$. Nevertheless, the development of the theory of Fractional Calculus (FC) is due to the contributions of many mathematicians such as Euler, Liouville, Riemann, and Letnikov [1–3].

The FC deals with derivatives and integrals to an arbitrary order (real or, even, complex order). The mathematical definition of a derivative/integral of fractional order has been the subject of several different approaches [1–3]. For example, the Laplace definition of a fractional derivative of a signal $x(t)$ is

$$D^\alpha x(t) = L^{-1}\left\{ s^\alpha X(s) - \sum_{k=0}^{n-1} s^k D^{\alpha-k-1} x(t)\big|_{t=0} \right\}, \tag{1.1}$$

where $n - 1 < \alpha \leq n, \alpha > 0$. The Grünwald-Letnikov definition is given by ($\alpha \in \mathfrak{R}$):

$$D^\alpha x(t) = \lim_{h \to 0} \left[\frac{1}{h^\alpha} \sum_{k=0}^{\infty} (-1)^k \binom{\alpha}{k} x(t - kh) \right],$$

$$\binom{\alpha}{k} = \frac{\Gamma(\alpha + 1)}{\Gamma(k + 1)\Gamma(\alpha - k + 1)},$$

(1.2)

where Γ is the Gamma function and h is the time increment. However, (1.2) shows that fractional-order operators are "global" operators having a memory of all past events, making them adequate for modeling memory effects in most materials and systems.

The Riemann-Liouville definition of the fractional-order derivative is ($\alpha > 0$):

$$_aD_t^\alpha f(t) = \frac{1}{\Gamma(n - \alpha)} \frac{d^n}{dt^n} \int_a^t \frac{f(\tau)}{(t - \tau)^{\alpha - n + 1}} d\tau, \quad n - 1 < \alpha < n,$$

(1.3)

where $\Gamma(x)$ is the Gamma function of x.

Based on the proposed definitions it is possible to calculate the fractional-order integrals/derivatives of several functions (Table 1). Nevertheless, the problem of devising and implementing fractional-order algorithms is not trivial and will be the matter of the following sections.

In recent years FC has been a fruitful field of research in science and engineering [1–6]. In fact, many scientific areas are currently paying attention to the FC concepts and we can refer its adoption in viscoelasticity and damping, diffusion and wave propagation, electromagnetism, chaos and fractals, heat transfer, biology, electronics, signal processing, robotics, system identification, traffic systems, genetic algorithms, percolation, modeling and identification, telecommunications, chemistry, irreversibility, physics, control systems as well as economy, and finance [7–18].

Bearing these ideas in mind, Sections 2–6 present several applications of FC in science and engineering. In Section 2, it is presented the application of FC concepts to the tuning of PID controllers and, in Section 3, the application of a fractional-order PD controller in the control of the leg joints of a hexapod robot. Then in Section 4, it is presented the fractional dynamics in the trajectory control of redundant manipulators. Next, in Section 5, it is introduced the fractional characteristics of heat diffusion along a media and, in Section 6 it is shown the application of FC to circuit synthesis using evolutionary algorithms. Finally, the main conclusions are presented in Section 7.

2. Tuning of PID Controllers Using Fractional Calculus Concepts

The PID controllers are the most commonly used control algorithms in industry. Among the various existent schemes for tuning PID controllers, the Ziegler-Nichols (Z-N) method is the most popular and is still extensively used for the determination of the PID parameters. It is well known that the compensated systems, with controllers tuned by this method, have generally a step response with a high percent overshoot. Moreover, the Z-N heuristics are only suitable for plants with monotonic step response.

Table 1: Fractional-order integrals of several functions.

$\varphi(x), \ x \in \mathfrak{R}$	$(I_+^\alpha \varphi)(x), \ x \in \mathfrak{R}, \ \alpha \in C$
$(x-a)^{\beta-1}$	$\dfrac{\Gamma(\beta)}{\Gamma(\alpha+\beta)}(x-a)^{\alpha+\beta-1}, \ \mathrm{Re}(\beta) > 0$
$e^{\lambda x}$	$\lambda^{-\alpha}e^{\lambda x}, \ \mathrm{Re}(\lambda) > 0$
$\begin{cases} \sin(\lambda x) \\ \cos(\lambda x) \end{cases}$	$\lambda^{-\alpha}\begin{cases} \sin(\lambda x - \alpha\pi/2), \\ \cos(\lambda x - \alpha\pi/2), \end{cases} \ \lambda > 0, \ \mathrm{Re}(\alpha) > 1$
$e^{\lambda x}\begin{cases} \sin(\gamma x) \\ \cos(\gamma x) \end{cases}$	$\dfrac{e^{\lambda x}}{(\lambda^2+\gamma^2)^{\alpha/2}}\begin{cases} \sin(\gamma x - \alpha\phi), \ \phi = \arctan(\gamma/\lambda) \\ \cos(\gamma x - \alpha\phi), \ \gamma > 0, \ \mathrm{Re}(\lambda) > 1 \end{cases}$

In this section, we study a methodology for tuning PID controllers such that the response of the compensated system has an almost constant overshoot defined by a prescribed value. The proposed method is based on the minimization of the integral of square error (ISE) between the step responses of a unit feedback control system, whose open-loop transfer function $L(s)$ is given by a fractional-order integrator and that of the PID compensated system [7].

Figure 1 illustrates the fractional-order control system that will be used as reference model for the tuning of PID controllers. The open-loop transfer function $L(s)$ is defined as ($\alpha \in \mathfrak{R}^+$):

$$L(s) = \left(\frac{\omega_c}{s}\right)^\alpha, \tag{2.1}$$

where ω_c is the gain crossover frequency, that is, $|L(j\omega_c)| = 1$. The parameter α is the slope of the magnitude curve, on a log-log scale, and may assume integer as well as noninteger values. In this study we consider $1 < \alpha < 2$, such that the output response may have a fractional oscillation (similar to an underdamped second-order system). This transfer function is also known as the Bode's ideal loop transfer function since Bode studies on the design of feedback amplifiers in the 1940s [19].

The Bode diagrams of amplitude and phase of $L(s)$ are illustrated in Figure 2. The amplitude curve is a straight line of constant slope -20α dB/dec, and the phase curve is a horizontal line positioned at $-\alpha\pi/2$ rad. The Nyquist curve is simply the straight line through the origin, $\arg L(j\omega) = -\alpha\pi/2$ rad.

This choice of $L(s)$ gives a closed-loop system with the desirable property of being insensitive to gain changes. If the gain changes, the crossover frequency ω_c will change, but the phase margin of the system remains PM $= \pi(1-\alpha/2)$ rad, independent of the value of the gain. This can be seen from the curves of amplitude and phase of Figure 2.

The closed-loop transfer function of fractional-order control system of Figure 1 is given by

$$G(s) = \frac{L(s)}{1+L(s)} = \frac{1}{(s/\omega_c)^\alpha + 1}, \quad 1 < \alpha < 2. \tag{2.2}$$
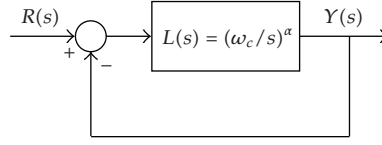
Figure 1: Fractional-order control system with open-loop transfer function $L(s)$.
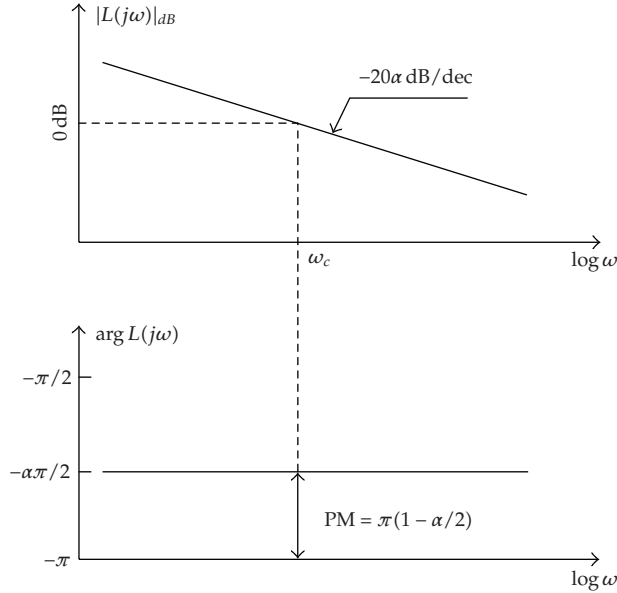
Figure 2: Bode diagrams of amplitude and phase of $L(j\omega)$ for $1 < \alpha < 2$.

The unit step response of $G(s)$ is given by the expression:

$$y_d(t) = L^{-1}\left\{\frac{1}{s}G(s)\right\} = L^{-1}\left\{\frac{\omega_c^\alpha}{s(s^\alpha + \omega_c^\alpha)}\right\} = 1 - \sum_{n=0}^{\infty}\frac{\left[-(\omega_c t)^\alpha\right]^n}{\Gamma(1 + \alpha n)} = 1 - E_\alpha\left[-(\omega_c t)^\alpha\right]. \qquad (2.3)$$

For the tuning of PID controllers, we address the fractional-order transfer function (2.2) as the reference system [8]. With the order α and the crossover frequency ω_c we can establish the overshoot and the speed of the output response, respectively. For that purpose we consider the closed-loop system shown in Figure 3, where $G_c(s)$ and $G_p(s)$ are the PID controller and the plant transfer functions, respectively.

The transfer function of the PID controller is

$$G_c(s) = \frac{U(s)}{E(s)} = K\left(1 + \frac{1}{T_i s} + T_d s\right), \qquad (2.4)$$

where $E(s)$ is the error signal and $U(s)$ is the controller's output. The parameters K, T_i, and T_d are the proportional gain, the integral time constant, and the derivative time constant of the controller, respectively.
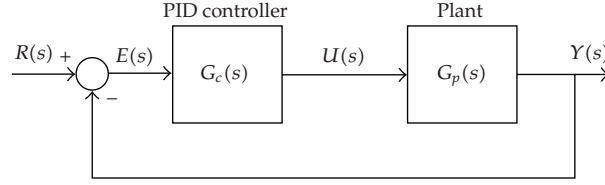
Figure 3: Closed-loop control system with PID controller $G_c(s)$.

The design of the PID controller will consist on the determination of the optimum PID set gains (K, T_i, T_d) that minimize J, the integral of the square error (ISE), defined as

$$J = \int_0^\infty \left[y(t) - y_d(t) \right]^2 dt, \tag{2.5}$$

where $y(t)$ is the step response of the closed-loop system with the PID controller (Figure 3) and $y_d(t)$ is the desired step response of the fractional-order transfer function (2.2) given by (2.3).

To illustrate the effectiveness of proposed methodology we consider the third-order plant transfer function:

$$G_p(s) = \frac{K_p}{(s+1)^3} \tag{2.6}$$

with nominal gain $K_p = 1$.

Figure 4 shows the step responses and the Bode diagrams of phase of the closed-loop system with the PID for the transfer function $G_p(s)$ for gain variations around the nominal gain $(K_p = 1)$ corresponding to $K_p = \{0.6, 0.8, 1.0, 1.2, 1.4\}$, that is, for a variation up to $\pm 40\%$ of its nominal value. The system was tuned for $\alpha = 3/2$ (PM = 45°), $\omega_c = 0.8\,\text{rad/s}$. We verify that we get the same desired iso-damping property corresponding to the prescribed (α, ω_c) values.

In fact, we observe that the step responses have an almost constant overshoot independent of the variation of the plant gain around the gain crossover frequency ω_c. Therefore, the proposed methodology is capable of producing closed-loop systems robust to gain variations and step responses exhibiting an iso-damping property. The proposed method was tested on several systems revealing good results. It was also compared with other tuning methods showing comparable or superior results [8].

3. Fractional PD^α Control of a Hexapod Robot

Walking machines allow locomotion in terrain inaccessible to other type of vehicles, since they do not need a continuous support surface, but at the cost of higher requirements for leg coordination and control. For these robots, joint level control is usually implemented through a PID-like scheme with position feedback. Recently, the application of the theory of FC to robotics revealed promising aspects for future developments [9]. With these facts in mind, this section compares different Fractional PD^α robot controller tuning, applied to the joint

(a)

(b)
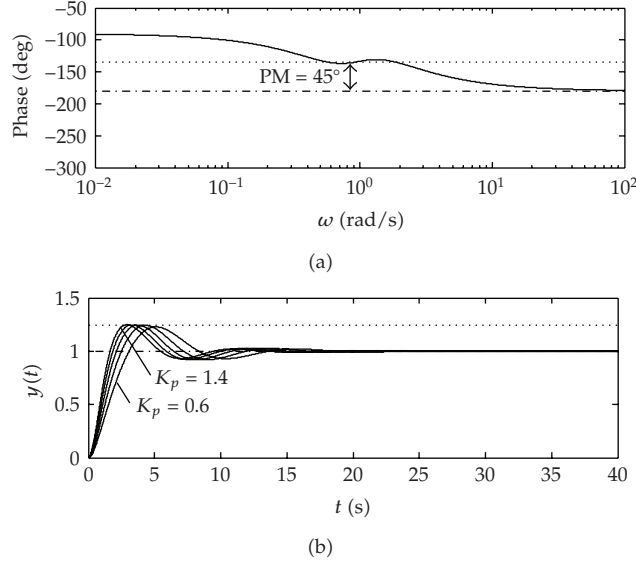
Figure 4: Bode phase diagrams and step responses for the closed-loop system with a PID controller for $G_p(s)$. The PID parameters are $K = 1.9158$, $T_i = 1.1407$, and $T_d = 0.9040$.

control of a walking system (Figure 5) with $n = 6$ legs, equally distributed along both sides of the robot body, having each three rotational joints (i.e., $j = \{1, 2, 3\} \equiv \{\text{hip}, \text{knee}, \text{ankle}\}$) [10].

During this study leg joint $j = 3$ can be either mechanical actuated or motor actuated (Figure 5). For the mechanical actuated case, we suppose that there is a rotational pre-tensioned spring-dashpot system connecting leg links L_{i2} and L_{i3}. This mechanical impedance maintains the angle between the two links while imposing a joint torque [10].

Figure 5 presents the dynamic model for the hexapod body and foot-ground interaction. It is considered robot body compliance because walking animals have a spine that allows supporting the locomotion with improved stability. The robot body is divided in n identical segments (each with mass $M_b n^{-1}$) and a linear spring-damper system (with parameters defined so that the body behaviour is similar to the one expected to occur on an animal) is adopted to implement the intrabody compliance [10]. The contact of the ith robot feet with the ground is modelled through a nonlinear system [11], being the values for the parameters based on the studies of soil mechanics [11].

The general control architecture of the hexapod robot is presented in Figure 6 [12]. In this study we evaluate the effect of different PD^α, $\alpha \in \mathfrak{R}$, controller implementations for $G_{c1}(s)$, while G_{c2} is a proportional controller with gain $Kp_j = 0.9$ ($j = 1, 2, 3$). For the PD^α algorithm, implemented through a discrete-time 4th-order Padé approximation ($a_{ij}, b_{ij} \in \mathfrak{R}$, $j = 1, 2, 3$), we have

$$G_{c1j}(z) \approx Kp_j + K\alpha_j \frac{\sum_{i=0}^{i=u} a_{ij}\ z^{-i}}{\sum_{i=0}^{i=u} b_{ij}\ z^{-i}}, \tag{3.1}$$

where Kp_j and $K\alpha_j$ are the proportional and derivative gains, respectively, and α_j is the fractional order, for joint j. Therefore, the classical PD^1 algorithm occurs when the fractional order $\alpha_j = 1.0$.
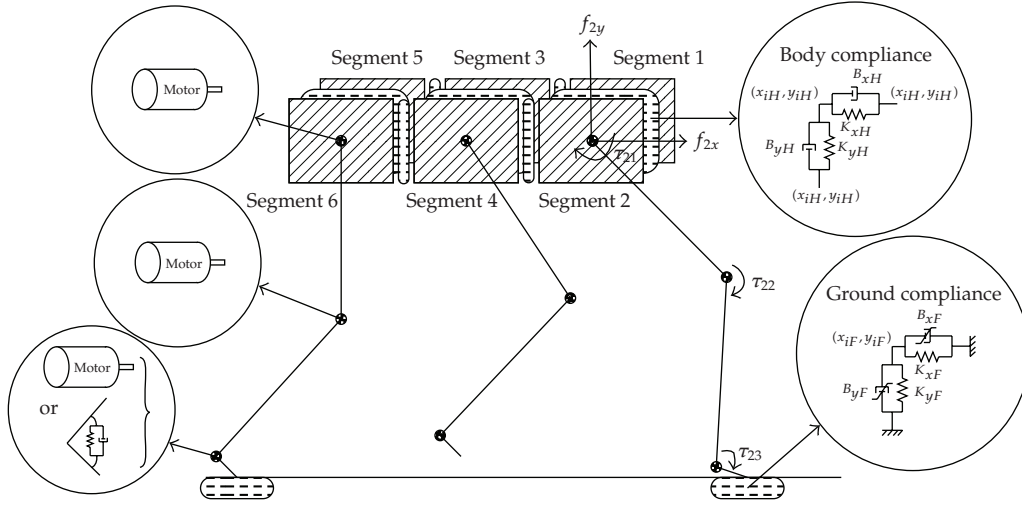
Figure 5: Model of the robot body and foot-ground interaction.

It is analysed the system performance of the different PD$^\alpha$ tuning, during a periodic wave gait at a constant forward velocity V_F, for two cases: two leg joints are motor actuated and the ankle joint is mechanical actuated and the three leg joints are fully motor actuated [10].

The analysis is based on the formulation of two indices measuring the mean absolute density of energy per traveled distance (E_{av}) and the hip trajectory errors (ε_{xyH}) during walking, according to

$$E_{av} = \frac{1}{d} \sum_{i=1}^{n} \sum_{j=1}^{m} \int_{0}^{T} |\boldsymbol{\tau}_{ij}(t)\dot{\boldsymbol{\theta}}_{ij}(t)| dt \quad \left[\mathbf{J}\mathbf{m}^{-1}\right],$$

$$\varepsilon_{xyH} = \sum_{i=1}^{n} \sqrt{\frac{1}{N_s} \sum_{k=1}^{N_s} \left(\Delta_{ixH}{}^2 + \Delta_{iyH}{}^2\right)} \quad [\mathbf{m}], \tag{3.2}$$

$$\Delta_{ixH} = x_{iHd}(k) - x_{iH}(k), \qquad \Delta_{iyH} = y_{iHd}(k) - y_{iH}(k).$$

To tune the different controller implementations we adopt a systematic method, testing and evaluating several possible combinations of parameters, for all controller implementations. Therefore, we adopt the $G_{c1}(s)$ parameters that establish a compromise in what concerns the simultaneous minimisation of E_{av} and ε_{xyH}. Moreover, it is assumed high-performance joint actuators, with a maximum actuator torque of $\tau_{ij\text{Max}} = 400$ Nm, and the desired angle between the foot and the ground (assumed horizontal) is established as $\theta_{i3hd} = -15°$. We tune the PD$^\alpha$ joint controllers for different values of the fractional order α_j while making $\alpha_1 = \alpha_2 = \alpha_3$.

We start by considering that leg joints 1 and 2 are motor actuated and joint 3 is mechanical actuated. For this case we tune the PD$^\alpha$ joint controllers for different values of the fractional order α_j, with step $\Delta\alpha_j = 0.1$, namely, $\alpha_j = \{-0.9, -0.8, \ldots, +0.9\}$. Afterwards,
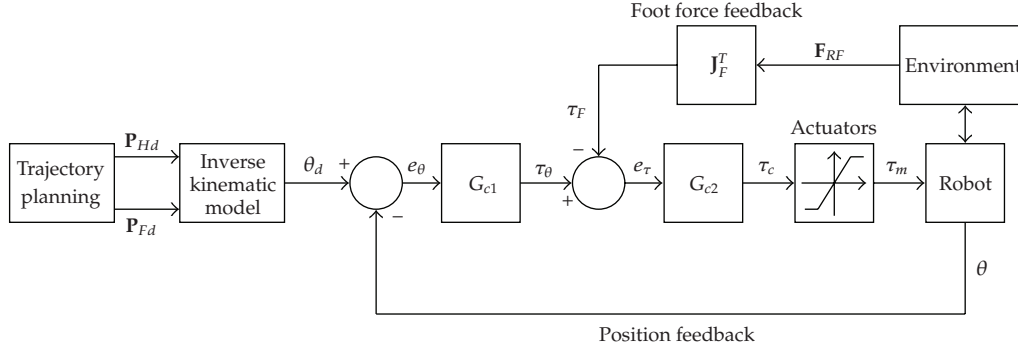
Figure 6: Hexapod robot control architecture.
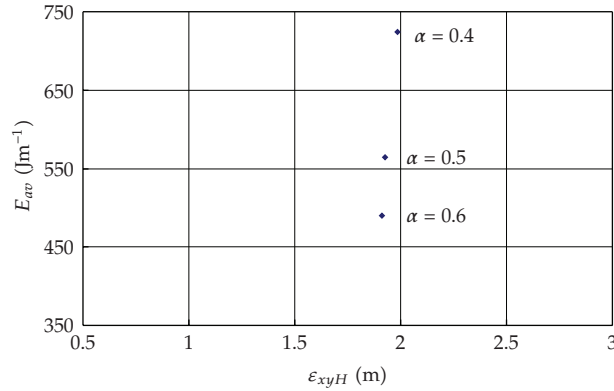
Figure 7: Locus of E_{av} versus ε_{xyH} for the different values of α in the $G_{c1}(s)$ tuning, when establishing a compromise between the minimisation of E_{av} and ε_{xyH}, with $G_{c2} = 0.9$, joints 1 and 2 motor actuated and joint 3 mechanical actuated.

we consider that joint 3 is also motor actuated, and we repeat the controller tuning procedure versus α_j.

For the first situation under study, we verify that the value of $\alpha_j = 0.6$ (Figure 7), being the gains of the PD^{α} controller $K_{p1} = 2500$, $K_{\alpha1} = 800$, $K_{p2} = 300$, $K_{\alpha2} = 100$ and the parameters of the mechanical spring-dashpot system for the ankle actuation $K_3 = 1$, $B_3 = 2$, presents the best compromise situation between the simultaneous minimisation of ε_{xyH} and E_{av}.

Regarding the case when all joints are motor actuated, Figure 8 presents the best controller tuning for different values of α_j. The experiments reveal the superior performance of the PD^{α} controller for $\alpha_j \approx 0.5$, with $K_{p1} = 15000$, $K_{\alpha1} = 7200$, $K_{p2} = 1000$, $K_{\alpha2} = 800$, and $K_{p3} = 150$, $K_{\alpha3} = 240$.

For $\alpha_j = \{0.1, 0.2, 0.3, 0.4\}$ the results are very poor and for $\alpha_j = \{-0.9, \ldots, -0.1\} \cup \{+0.9\}$, the hexapod locomotion is unstable. Furthermore, we conclude that the best case corresponds to all leg joints being motor actuated.

In conclusion, the experiments reveal the superior performance of the FO controller for $\alpha_j \approx 0.5$ and a robot with all motor actuated joints, as can be concluded analysing the curves for the joint actuation torques τ_{1jm} (Figure 9) and for the hip trajectory tracking errors Δ_{1xH} and Δ_{1yH} (Figure 10).
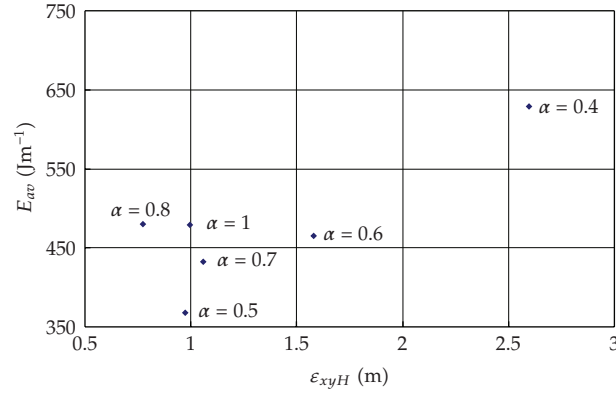
Figure 8: Locus of E_{av} versus ε_{xyH} for the different values of α in the $G_{c1}(s)$ tuning, when establishing a compromise between the minimisation of E_{av} and ε_{xyH}, with $G_{c2} = 0.9$ and all joints motor actuated.

(a)

(b)

--- Joint $j = 3$ motor actuated
— Joint $j = 3$ mechanical actuated

(c)
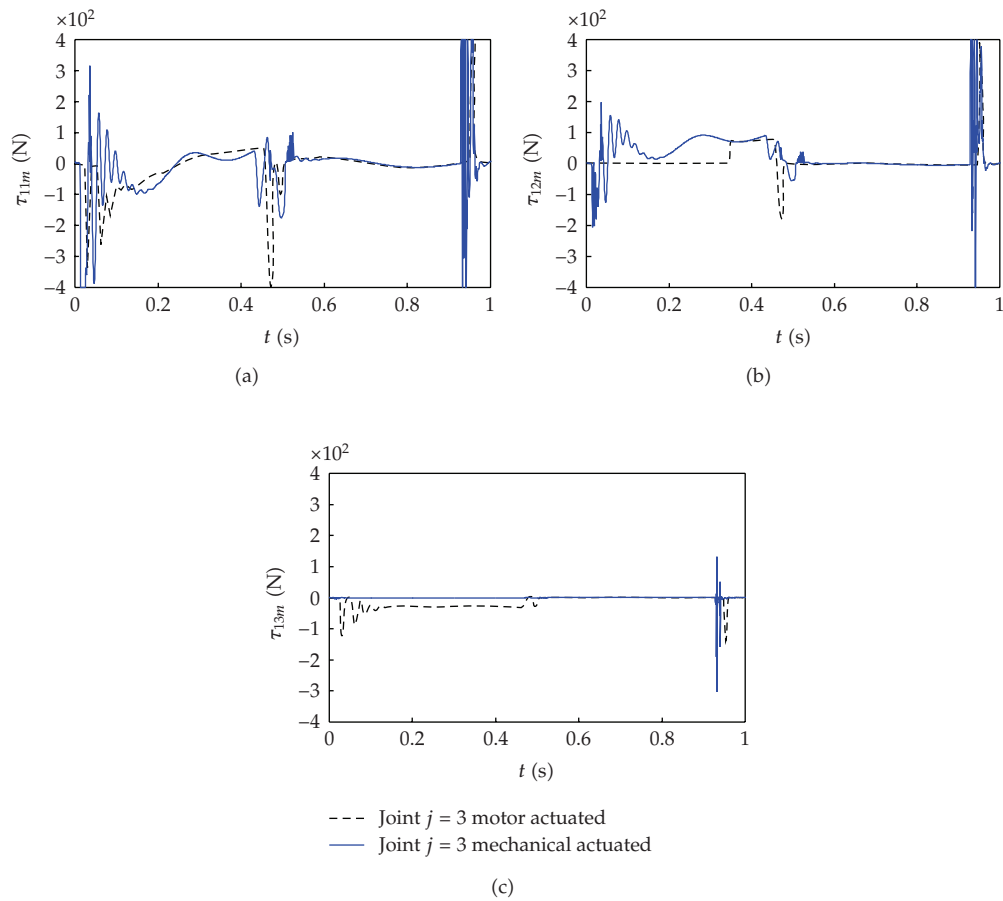
Figure 9: Plots of τ_{1jm} versus t, with joints 1 and 2 motor actuated and joint 3 mechanical actuated and all joints motor actuated, for $\alpha_j = 0.5$.

(a)   (b)
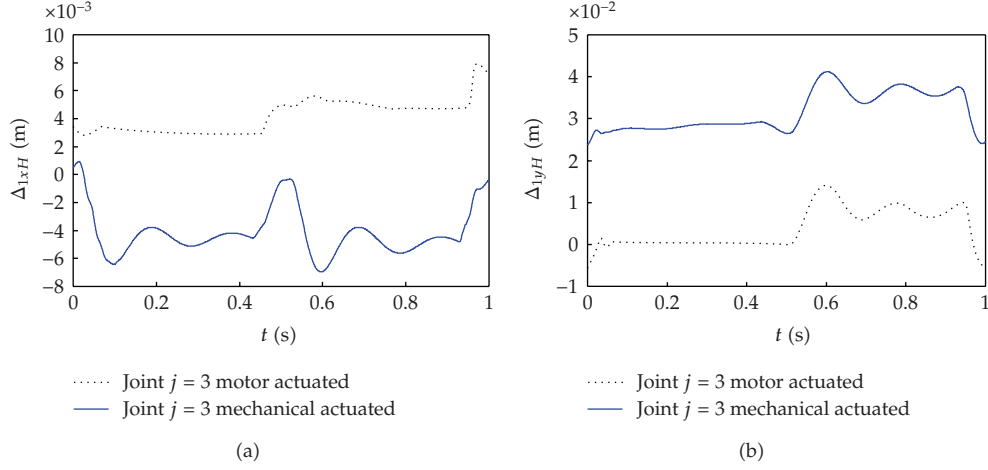
Figure 10: Plots of Δ_{1xH} and Δ_{1yH} versus t, with joints 1 and 2 motor actuated and joint 3 mechanical actuated and all joints motor actuated, for $\alpha_j = 0.5$.
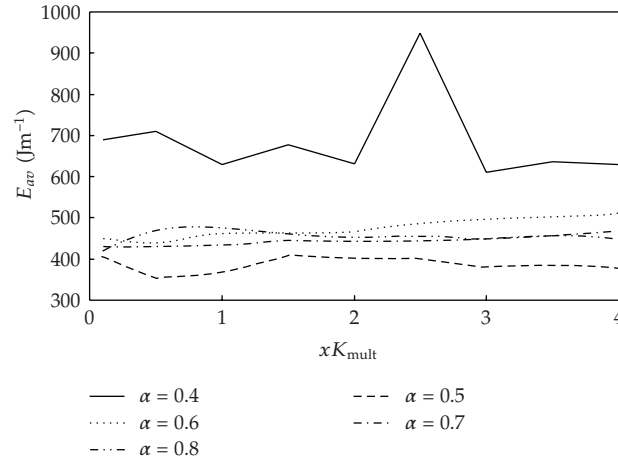
Figure 11: Performance index E_{av} versus K_{mult} for the $G_{c1}(s)$ PD$^\alpha$ controller tuning having all joints motor actuated.

Since the objective of the walking robots is to walk in natural terrains, in the sequel it is examined how the different controller tunings behave under different ground properties, considering that all joints are motor actuated. For this case, and considering the previously tuning controller parameters, the values of $\{K_{xF}, B_{xF}, K_{yF}, B_{yF}\}$ are varied simultaneously through a multiplying factor K_{mult} that is varied in the range $[0.1, 4.0]$. This variation for the ground model parameters allows the simulation of the ground behaviour for growing stiffness, from peat to gravel [11].

The performance measure E_{av} versus the multiplying factor of the ground parameters K_{mult} is presented on Figure 11. Analysing the system performance from the viewpoint of the index E_{av}, it is possible to conclude that the best PD$^\alpha$ implementation occurs for the fractional order $\alpha_j = 0.5$. Moreover, it is clear that the performances of the different controller implementations are almost constant on all range of the ground parameters, with

the exception of the fractional order $\alpha_j = 0.4$. For this case, E_{av} presents a significant variation with K_{mult}. Therefore, we conclude that the controller responses are quite similar, meaning that these algorithms are robust to variations of the ground characteristics [12].

4. Fractional Dynamics in the Trajectory Control of Redundant Manipulators

A redundant manipulator is a robotic arm possessing more degrees of freedom (*dof*) than those required to establish an arbitrary position and orientation of the end effector. Redundant manipulators offer several potential advantages over non-redundant arms. In a workspace with obstacles, the extra degrees of freedom can be used to move around or between obstacles and thereby to manipulate in situations that otherwise would be inaccessible [20–23].

When a manipulator is redundant, it is anticipated that the inverse kinematics admits an infinite number of solutions. This implies that, for a given location of the manipulator's gripper, it is possible to induce a self-motion of the structure without changing the location of the end effecter. Therefore, the arm can be reconfigured to find better postures for an assigned set of task requirements.

Several kinematic techniques for redundant manipulators control the gripper through the rates at which the joints are driven, using the pseudoinverse of the Jacobian [22, 24]. Nevertheless, these algorithms lead to a kind of chaotic motion with unpredictable arm configurations.

Having these ideas in mind, Section 4.1 introduces the fundamental issues for the kinematics of redundant manipulators. Based on these concepts, Section 4.2 presents the trajectory control of a three *dof* robot. The results reveal a chaotic behavior that is further analyzed in Section 4.3.

4.1. Kinematics of Redundant Manipulators

A kinematically redundant manipulator has more *dof* than those required to define an arbitrary position and orientation of the gripper. In Figure 12 is depicted a planar manipulator with $k \in \aleph$ rotational (R) joints that is redundant for $k > 2$. When a manipulator is redundant it is anticipated that the inverse kinematics admits an infinite number of solutions. This implies that, for a given location of the manipulator's gripper, it is possible to induce a self-motion of the structure without changing the location of the gripper. Therefore, redundant manipulators can be reconfigured to find better postures for an assigned set of task requirements but, on the other hand, have a more complex structure requiring adequate control algorithms.

We consider a manipulator with n degrees of freedom whose joint variables are denoted by $\mathbf{q} = [q_1, q_2, \ldots, q_n]^T$. We assume that a class of tasks, we are interested in can be described by m variables, $\mathbf{x} = [x_1, x_2, \ldots, x_m]^T$ ($m < n$) and that the relation between $\mathbf{q}$ and $\mathbf{x}$ is given by

$$\mathbf{x} = f(\mathbf{q}), \tag{4.1}$$

where f is a function representing the direct kinematics.
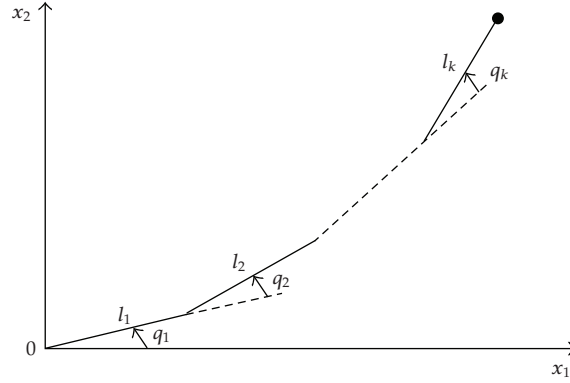
Figure 12: A planar redundant planar manipulator with k rotational joints.

Differentiating (4.1) with respect to time yields

$$\dot{\mathbf{x}} = \mathbf{J}(\mathbf{q})\dot{\mathbf{q}}, \tag{4.2}$$

where $\dot{\mathbf{x}} \in \mathfrak{R}^m$, $\dot{\mathbf{q}} \in \mathfrak{R}^n$, and $\mathbf{J}(\mathbf{q}) = \partial f(\mathbf{q})/\partial \mathbf{q} \in \mathfrak{R}^{m \times n}$. Hence, it is possible to calculate a path $\mathbf{q}(t)$ in terms of a prescribed trajectory $\mathbf{x}(t)$ in the operational space. We assume that the following condition is satisfied:

$$\max \text{rank}\{\mathbf{J}(\mathbf{q})\} = m. \tag{4.3}$$

Failing to satisfy this condition usually means that the selection of manipulation variables is redundant and the number of these variables m can be reduced. When condition (4.3) is verified, we say that the degree of redundancy of the manipulator is $n - m$. If, for some $\mathbf{q}$ we have

$$\text{rank}\{\mathbf{J}(\mathbf{q})\} < m \tag{4.4}$$

then the manipulator is in a singular state. This state is not desirable because, in this region of the trajectory, the manipulating ability is very limited.

Many approaches for solving redundancy [25, 26] are based on the inversion of (4.2). A solution in terms of the joint velocities is sought as

$$\dot{\mathbf{q}} = \mathbf{J}^{\#}(\mathbf{q})\dot{\mathbf{x}}, \tag{4.5}$$

where $\mathbf{J}^{\#}$ is one of the generalized inverses of the $\mathbf{J}$ [26–28]. It can be easily shown that a more general solution to (4.2) is given by

$$\dot{\mathbf{q}} = \mathbf{J}^{+}(\mathbf{q})\dot{\mathbf{x}} + [\mathbf{I} - \mathbf{J}^{+}(\mathbf{q})\mathbf{J}(\mathbf{q})]\dot{\mathbf{q}}_0, \tag{4.6}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix and $\dot{\mathbf{q}}_0 \in \mathfrak{R}^n$ is a $n \times 1$ arbitrary joint velocity vector and $\mathbf{J}^{+}$ is the pseudoinverse of the $\mathbf{J}$. The solution (4.6) is composed of two terms. The
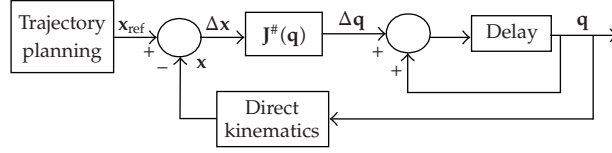
Figure 13: Block diagram of the closed-loop inverse kinematics algorithm with the pseudoinverse.

first term is relative to minimum norm joint velocities. The second term, the *homogeneous solution*, attempts to satisfy the additional constraints specified by $\dot{\mathbf{q}}_0$. Moreover, the matrix $\mathbf{I} - \mathbf{J}^+(\mathbf{q})\mathbf{J}(\mathbf{q})$ allows the projection of $\dot{\mathbf{q}}_0$ in the null space of $\mathbf{J}$. A direct consequence is that it is possible to generate internal motions that reconfigure the manipulator structure without changing the gripper position and orientation [27–30]. Another aspect revealed by the solution of (4.6) is that repetitive trajectories in the operational space do not lead to periodic trajectories in the joint space. This is an obstacle for the solution of many tasks because the resultant robot configurations have similarities with those of a chaotic system.

4.2. Robot Trajectory Control

The direct kinematics and the Jacobian of a 3-link planar manipulator with rotational joints (3*R* robot) has a simple recursive nature according with the expressions:

$$
\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} l_1 C_1 + l_2 C_{12} + l_3 C_{123} \\ l_1 S_1 + l_2 S_{12} + l_3 S_{123} \end{bmatrix},
$$

$$
\mathbf{J} = \begin{bmatrix} -l_1 S_1 - \cdots - l_3 S_{123} & \cdots & -l_3 S_{123} \\ l_1 C_1 + \cdots + l_3 C_{123} & \cdots & l_3 C_{123} \end{bmatrix},
$$

(4.7)

where l_i is the length of link i, $q_{i\ldots k} = q_i + \cdots + q_k$, $S_{i\ldots k} = \mathrm{Sin}(q_{i\ldots k})$, and $C_{i\ldots k} = \mathrm{Cos}(q_{i\ldots k})$.

During all the experiments it is considered $\Delta t = 10^{-3}$ seconds, $L_{\mathrm{TOT}} = l_1 + l_2 + l_3 = 3$, and $l_1 = l_2 = l_3$.

In the closed-loop pseudoinverse's method the joint positions can be computed through the time integration of the velocities according with the block diagram of the inverse kinematics algorithm depicted in Figure 13, where $\mathbf{x}_{\mathrm{ref}}$ represents the vector of reference coordinates of the robot gripper in the operational space.

Based on (4.7) we analyze the kinematic performances of the 3*R*-robot when repeating a circular motion in the operational space with frequency $\omega_0 = 7.0\,\mathrm{rad\,s}^{-1}$, centre at distance $r = [x_1^2 + x_2^2]^{1/2}$ and radius ρ.

Figure 14 shows the joint positions for the inverse kinematic algorithm (4.5) for $r = \{0.6, 2.0\}$ and $\rho = \{0.3, 0.5\}$. We observe that the following hold.

   (i) For $r = 0.6$ occur unpredictable motions with severe variations that lead to high joint transients [13]. Moreover, we verify a low-frequency signal modulation that depends on the circle being executed.

   (ii) For $r = 2.0$ the motion is periodic with frequency identical to $\omega_0 = 7.0\,\mathrm{rad\,s}^{-1}$.

(a)

(b)

(c)

(d)

$r = 0.6$

$r = 2$
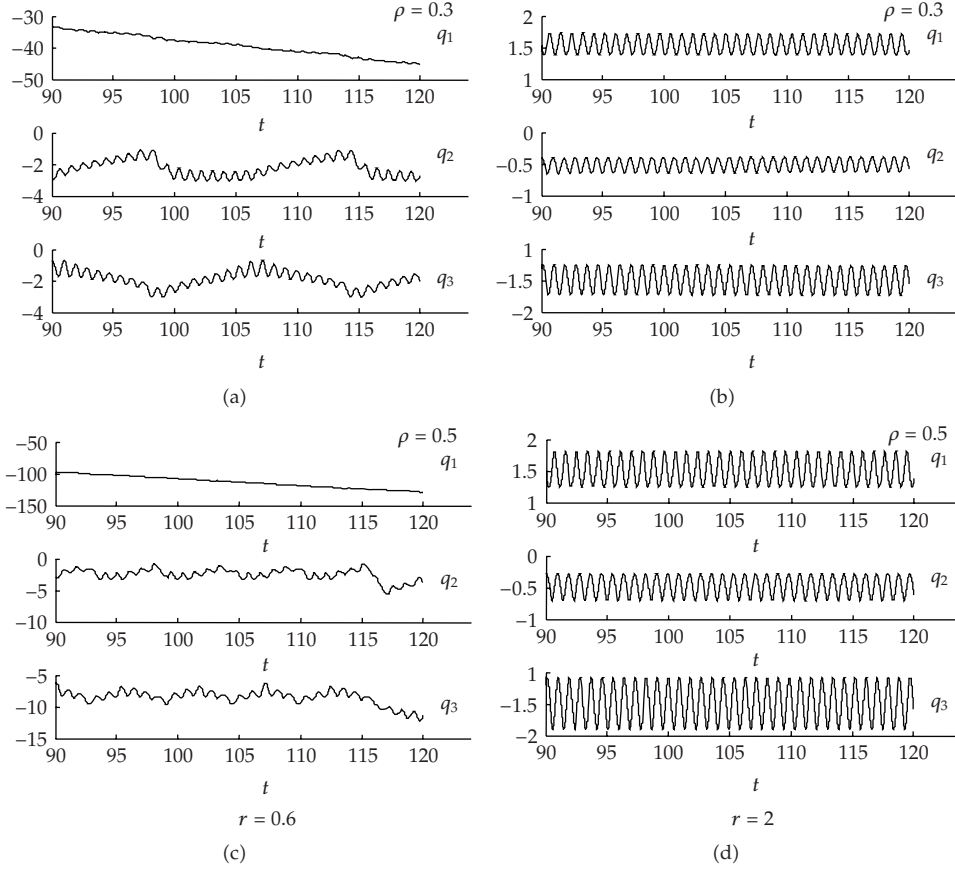
Figure 14: The $3R$-robot joint positions versus time using the pseudoinverse method for $r = \{0.6, 2.0\}$ and $\rho = \{0.3, 0.5\}$.

4.3. Analysis of the Robot Trajectories

In the previous subsection we verified that the pseudoinverse-based algorithm leads to unpredictable arm configurations. In order to gain further insight into the pseudoinverse nature several distinct experiments are devised in the sequel during a time window of 300 cycles. Therefore, in a first set of experiments we calculate the Fourier transform of the $3R$-robot joints velocities for a circular repetitive motion with frequency $\omega_0 = 7.0\,\mathrm{rad\,s^{-1}}$, radius $\rho = \{0.1, 0.3, 0.5, 0.7\}$, and radial distances $r \in]0, L_{\mathrm{TOT}} - \rho[$.

Figure 15 shows $|F\{\dot{q}_2(t)\}|$ versus the frequency ratio ω_0/ω and the distance r where $F\{\}$ represents the Fourier operator. Is verified an interesting phenomenon induced by the gripper repetitive motion ω_0 because a large part of the energy is distributed along several subharmonics. These fractional-order harmonics (*foh*) depend on r and ρ making a complex pattern with similarities with those revealed by chaotic systems. Furthermore, we observe the existence of several distinct regions depending on r.

For example, selecting in Figure 15 several distinct cases, namely for $r = \{0.08, 0.30, 0.53, 1.10, 1.30, 2.00\}$, we have the different signal Fourier spectra clearly visible in Figure 16.
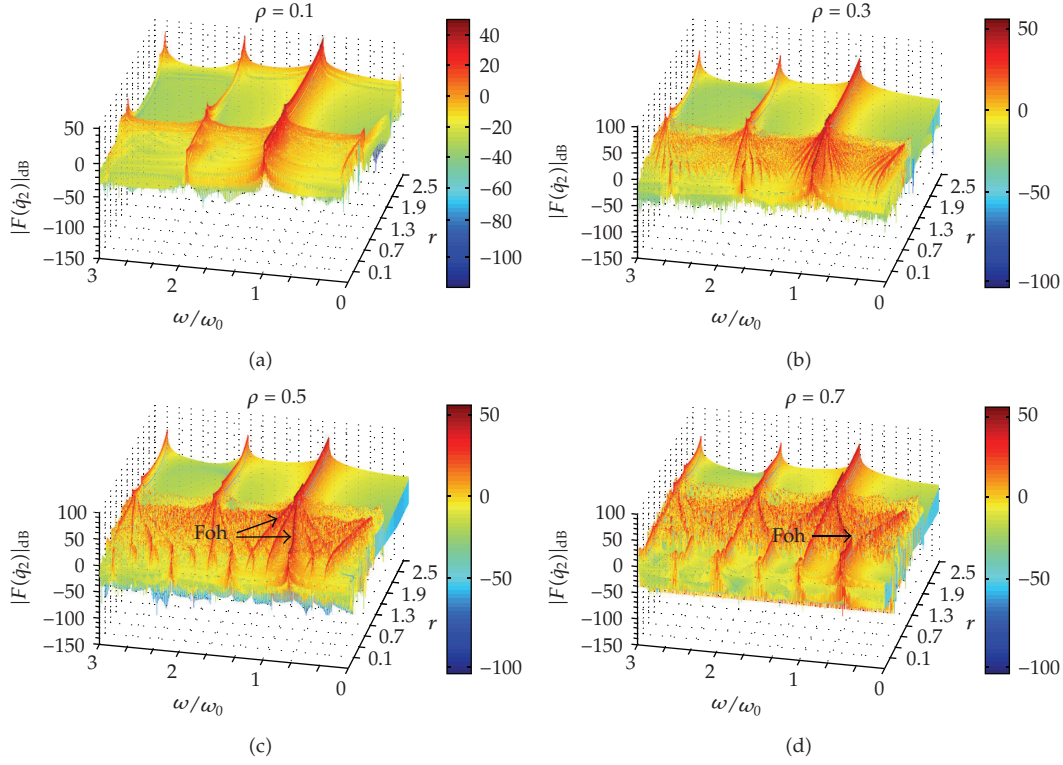
Figure 15: $|F\{\dot{q}_2(t)\}|$ of the 3R-robot during 300 cycles versus r and ω/ω_0, for $\rho = \{0.1, 0.3, 0.5, 0.7\}$, $\omega_0 = 7.0 \, \text{rad s}^{-1}$.

In the authors' best knowledge the *foh* are aspects of fractional dynamics [14, 15, 31], but a final and assertive conclusion about a physical interpretation is a matter still to be explored.

For joints velocities 1 and 3 the results are similar to the verified ones for joint velocity 2.

5. Heat Diffusion

The heat diffusion is governed by a linear one-dimensional partial differential equation (PDE) of the form:

$$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2}, \tag{5.1}$$

where k is the diffusivity, t is the time, u is the temperature, and x is the space coordinate. However, (5.1) involves the solution of a PDE of parabolic type for which the standard theory guarantees the existence of a unique solution [16].

For the case of a planar perfectly isolated surface we usually apply a constant temperature U_0 at $x = 0$ and analyzes the heat diffusion along the horizontal coordinate x.
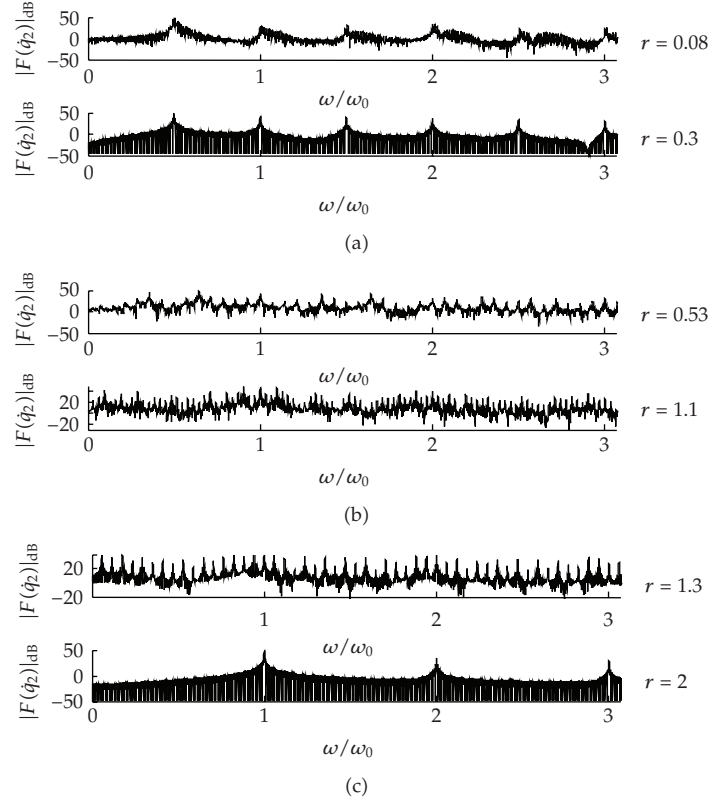
Figure 16: $|F\{\dot{q}_2(t)\}|$ of the 3*R*-robot during 300 cycles versus the frequency ratio ω/ω_0, for $r = \{0.08, 0.30, 0.53, 1.10, 1.30, 2.00\}$, $\rho = 0.7$, $\omega_0 = 7.0\,\mathrm{rad\,s^{-1}}$.

Under these conditions, the heat diffusion phenomenon is described by a noninteger-order model:

$$U(x,s) = \frac{U_0}{s}G(s) \qquad G(s) = e^{-x\sqrt{s/k}}, \tag{5.2}$$

where x is the space coordinate, U_0 is the boundary condition, and $G(s)$ is the system transfer function.

In our study, the simulation of the heat diffusion is performed by adopting the Crank-Nicholson implicit numerical integration based on the discrete approximation to differentiation as [16, 17]

$$-ru[j+1, i+1] + (2+r)u[j+1, i] - ru[j+1, i-1] = ru[j, i+1] + (2-r)u[j, i] + u[j, i-1], \tag{5.3}$$

where $r = \Delta t(\Delta x^2)^{-1}$, $\{\Delta x, \Delta t\}$, and $\{i, j\}$ are the increments and the integration indices for space and time, respectively.
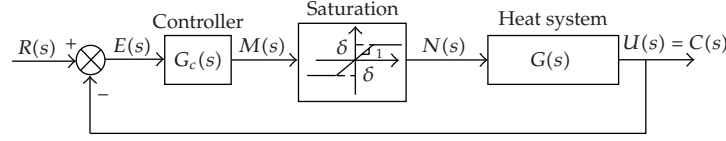
Figure 17: Closed-loop system with PID controller $G_c(s)$.

5.1. Control Strategies

The generalized PID controller $G_c(s)$ has a transfer function of the form

$$G_c(s) = K\left[1 + \frac{1}{T_i s^\alpha} + T_d s^\beta\right],\tag{5.4}$$

where α and β are the orders of the fractional integrator and differentiator, respectively. The constants K, T_i, and T_d are correspondingly the proportional gain, the integral time constant, and the derivative time constant.

Clearly, taking $(\alpha, \beta) = \{(1, 1), (1, 0), (0, 1), (0, 0)\}$ we get the classical $\{PID, PI, PD, P\}$ controllers, respectively.

The $PI^\alpha D^\beta$ controller is more flexible and gives the possibility of adjusting more carefully the closed-loop system characteristics.

In the following two subsections, we analyze the system of Figure 17 by adopting the classical integer-order PID and a fractional PID^β, respectively.

5.2. PID Tuning Using the Ziegler-Nichols Rule

In this subsection, we analyze the closed-loop system with a conventional PID controller given by the transfer function (5.4) with $\alpha = \beta = 1$. Usually, the PID parameters (K, T_i, T_d) are tuned by using the so-called Ziegler-Nichols open loop (ZNOL) method [17]. The ZNOL heuristics are based on the approximate first-order plus dead-time model:

$$\widehat{G}(s) = \frac{K_p}{\tau s + 1} e^{-sT}.\tag{5.5}$$

For the heat system, the resulting parameters are $\{K_p, \tau, T\} = \{0.52, 162, 28\}$ leading to the PID constants $\{K, T_i, T_d\} = \{18.07, 34.0, 8.5\}$.

A step input is applied at $x = 0.0\,\text{m}$ and the closed-loop response $c(t)$ is analyzed for $x = 3.0\,\text{m}$, without actuator saturation (Figure 18). We verify that the system with a PID controller, tuned through the ZNOL heuristics, does not produce satisfactory results giving a significant overshoot ov and a large settling time t_s, namely, $\{t_s, t_p, t_r, ov(\%)\} \equiv \{44.8, 27.5, 12.0, 68.56\}$, where t_p represents the peak time and t_r the rise time. We consider
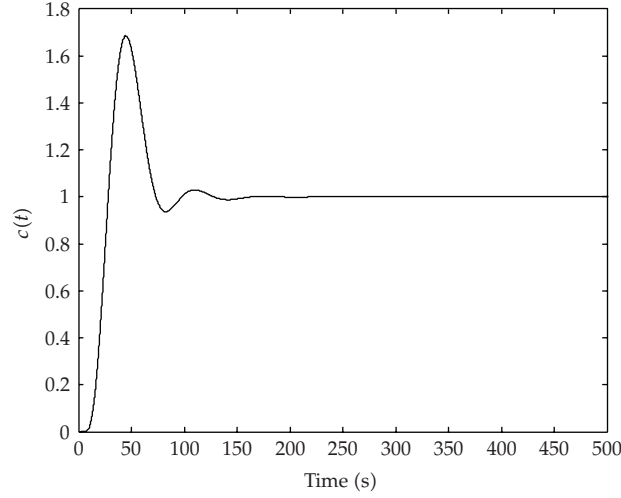
Figure 18: Step responses of the closed-loop system for the PID controller and $x = 3.0$ m.

two indices that measure the response error, namely, the integral square error (ISE) and the integral time square error (ITSE) criteria defined as

$$
\begin{aligned}
\text{ISE} &= \int_0^\infty [r(t) - c(t)]^2 dt, \\
\text{ITSE} &= \int_0^\infty t[r(t) - c(t)]^2 dt.
\end{aligned}
\tag{5.6}
$$

We can use other performance criteria such as the integral absolute error (IAE) or the integral time absolute error (ITAE); however, in the present case, the ISE and the ITSE criteria have produced the best results and are adopted in the study.

In this case, the ZNOL PID tuning leads to the values (ISE, ITSE) = $(27.53, 613.97)$. The poor results indicate again that the method of tuning may not be the most adequate for the control of the heat system.

In fact, the inherent fractional dynamics of the system lead us to consider other configurations. In this perspective, we propose the use of fractional controllers tuned by the minimization of the indices ISE and ITSE.

5.3. PID$^\beta$ Tuning Using Optimization Indices

In this subsection, we analyze the closed-loop system under the action of the PID$^\beta$ controller given by the transfer function (5.4) with $\alpha = 1$ and $0 \leq \beta \leq 1$. The fractional derivative term $T_d s^\beta$ in (5.4) is implemented through a fourth-order Padé discrete rational transfer function. It used a sampling period of $T = 0.1$ second.

The PID$^\beta$ controller is tuned by the minimization of an integral performance index. For that purpose, we adopt the ISE and ITSE criteria.

A step reference input $R(s) = 1/s$ is applied at $x = 0.0$ m and the output $c(t)$ is analyzed for $x = 3.0$ m, without actuator saturation. The heat system is simulated for 3000

K-ISE
K-ITSE
•   K-ZNOL

(a)

T_i-ISE
T_i-ITSE
• T_i-ZNOL

(b)

T_d-ISE
T_d-ITSE
• T_d-ZNOL

(c)
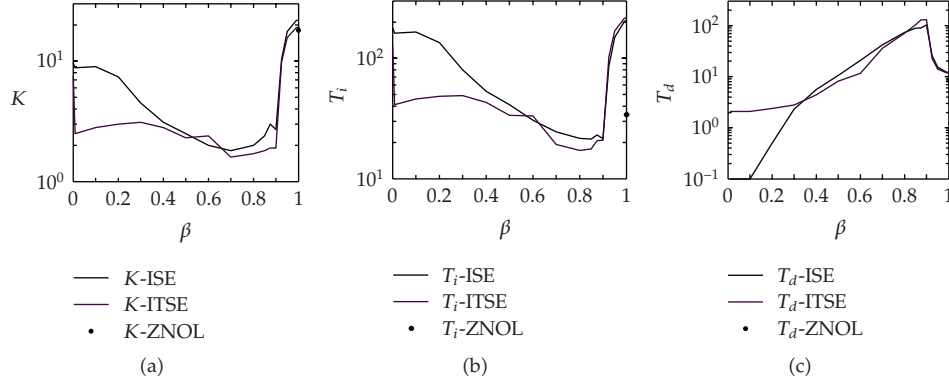
Figure 19: The PID$^\beta$ parameters (K, T_i, T_d) versus β for the ISE and ITSE optimization criteria. The dot represents the PID-ZNOL.

seconds. Figure 19 illustrates the variation of the fractional PID parameters (K, T_i, T_d) as function of the order's derivative β, for the ISE and the ITSE criteria. The dots represent the values corresponding to the classical PID (ZNOL-tuning) addressed in the previous section.

The curves reveal that for $\beta < 0.4$ the parameters (K, T_i, T_d) are slightly different, for the two ISE and ITSE criteria, while for $\beta \geq 0.4$ they lead to almost similar values. This fact indicates a large influence of a weak-order derivative on system's dynamics.

To further illustrate the performance of the fractional-order controllers a saturation nonlinearity is included in the closed-loop system of Figure 17 and inserted in series with the output of the controller $G_c(s)$. The saturation element is defined as

$$n(m) = \begin{cases} m, & |m| < \delta, \\ \delta\, \text{sign}\,(m), & |m| \geq \delta. \end{cases} \tag{5.7}$$

The controller performance is evaluated for $\delta = \{20, \ldots, 100\}$ and $\delta = \infty$ which corresponds to a system without saturation. We use the same fractional-PID parameters obtained without considering the saturation nonlinearity.

Figures 20 and 21 show the step responses of the closed-loop system and the corresponding controller output, for the PID$^\beta$ tuned in the ISE and ITSE perspectives for $\delta = 10$ and $\delta = \infty$, respectively. The controller parameters $\{K, T_i, T_d, \beta\}$ correspond to the minimization of those indices leading to the values ISE: $\{K, T_i, T_d, \beta\} \equiv \{3, 23, 90.6, 0.875\}$ and ITSE: $\{K, T_i, T_d, \beta\} \equiv \{1.8, 17.6, 103.6, 0.85\}$.

The step responses reveal a large diminishing of the overshoot and the rise time when compared with the integer PID, showing a good transient response and a zero steady-state error. The PID$^\beta$ leads to better results than the classical PID controller tuned through the ZNOL rule. These results demonstrate the effectiveness of the fractional algorithms when used for the control of fractional-order systems. The step response and the controller output are also improved when the saturation level δ is diminished.

Figure 22 depicts the ISE and ITSE indices for $0 \leq \beta \leq 1$, when $\delta = \{20, \ldots, 100\}$ and $\delta = \infty$. We verify the existence of a minimum for $\beta = 0.875$ and $\beta = 0.85$ for the ISE and ITSE cases, respectively. Furthermore, the higher the δ the lower the value of the index.
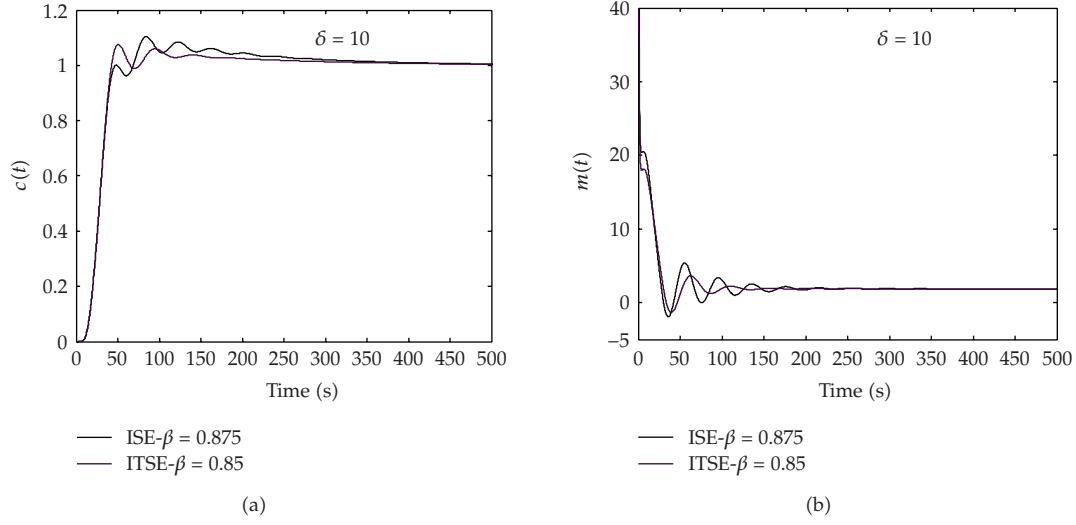
Figure 20: Step responses of the closed-loop system and the controller output for the ISE and the ITSE indices, with a PID$^\beta$ controller, $\delta = 10$ and $x = 3.0$ m.
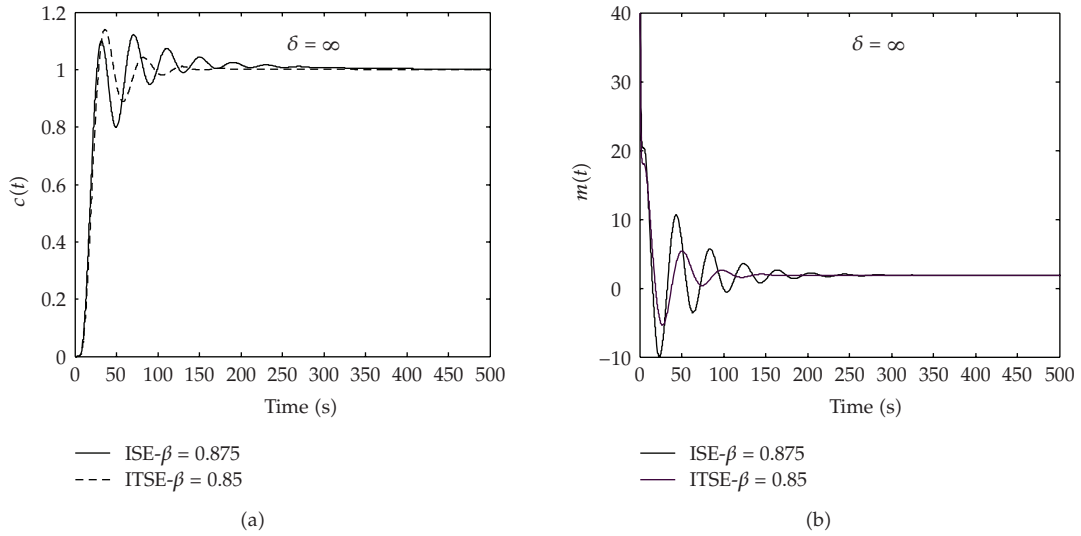
Figure 21: Step responses of the closed-loop system and the controller output for the ISE and the ITSE indices, with a PID$^\beta$ controller, $\delta = \infty$ and $x = 3.0$ m.

Figures 23 and 24 show the variation of the settling time t_s, the peak time t_p, the rise time t_r, and the percent overshoot $ov(\%)$, for the closed-loop response tuned through the minimization of the ISE and the ITSE indices, respectively.

In the ISE case t_s, t_p, and t_r diminish rapidly for $0 \leq \beta \leq 0.875$, while for $\beta > 0.875$ the parameters increase smoothly. For the ITSE, we verify the same behavior for $\beta = 0.85$. On the other hand, $ov(\%)$ increases smoothly for $0 \leq \beta \leq 0.7$, while for $\beta > 0.7$ it decreases very quickly, both for the ISE and the ITSE indices.
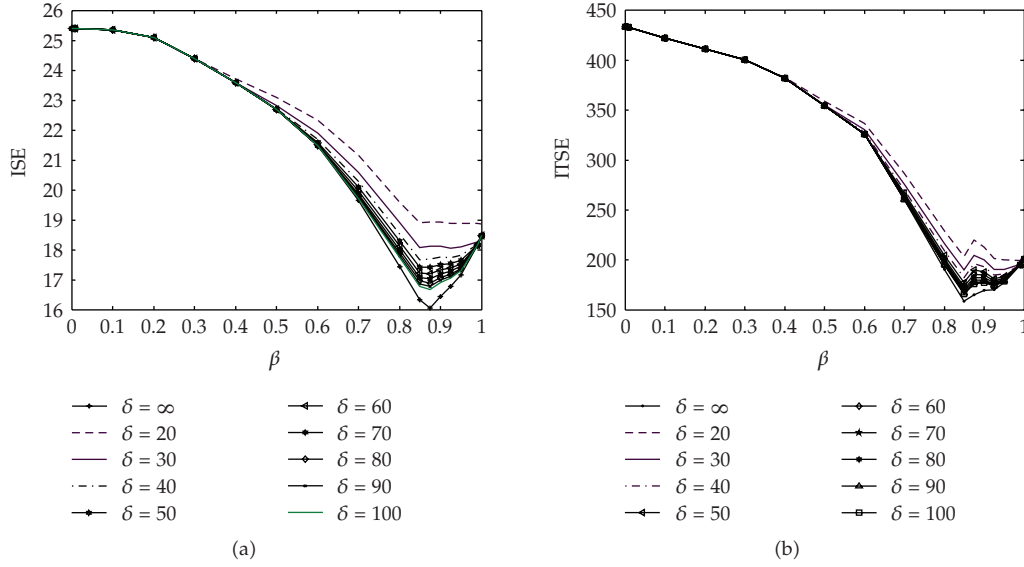
Figure 22: ISE and ITSE versus $0 \leq \beta \leq 1$ for $\delta = \{20, \dots, 100\}$ and $\delta = \infty$.

In conclusion, for $0.85 \leq \beta \leq 0.875$ we get the best controller tuning, superior to the performance revealed by the classical integer-order scheme.

6. Circuit Synthesis Using Evolutionary Algorithms

In recent decades evolutionary computation (EC) techniques have been applied to the design of electronic circuits and systems, leading to a novel area of research called Evolutionary Electronics (EE) or Evolvable Hardware (EH). EE considers the concept for automatic design of electronic systems. Instead of using human conceived models, abstractions, and techniques, EE employs search algorithms to develop implementations not achievable with the traditional design schemes, such as the Karnaugh or the Quine-McCluskey Boolean methods.

Several papers proposed designing combinational logic circuits using evolutionary algorithms and, in particular, genetic algorithms (GAs) [32, 33] and hybrid schemes such as the memetic algorithms (MAs) [34].

Particle swarm optimization (PSO) constitutes an alternative evolutionary computation technique, and this paper studies its application to combinational logic circuit synthesis. Bearing these ideas in mind, the organization of this section is as follows. Section 6.1 presents a brief overview of the PSO. Section 6.2 describes the PSO-based circuit design, while Section 6.3 exhibits the simulation results.

6.1. Particle Swarm Optimization

In literature about PSO the term 'swarm intelligence' appears rather often and, therefore, we begin by explaining why this is so.
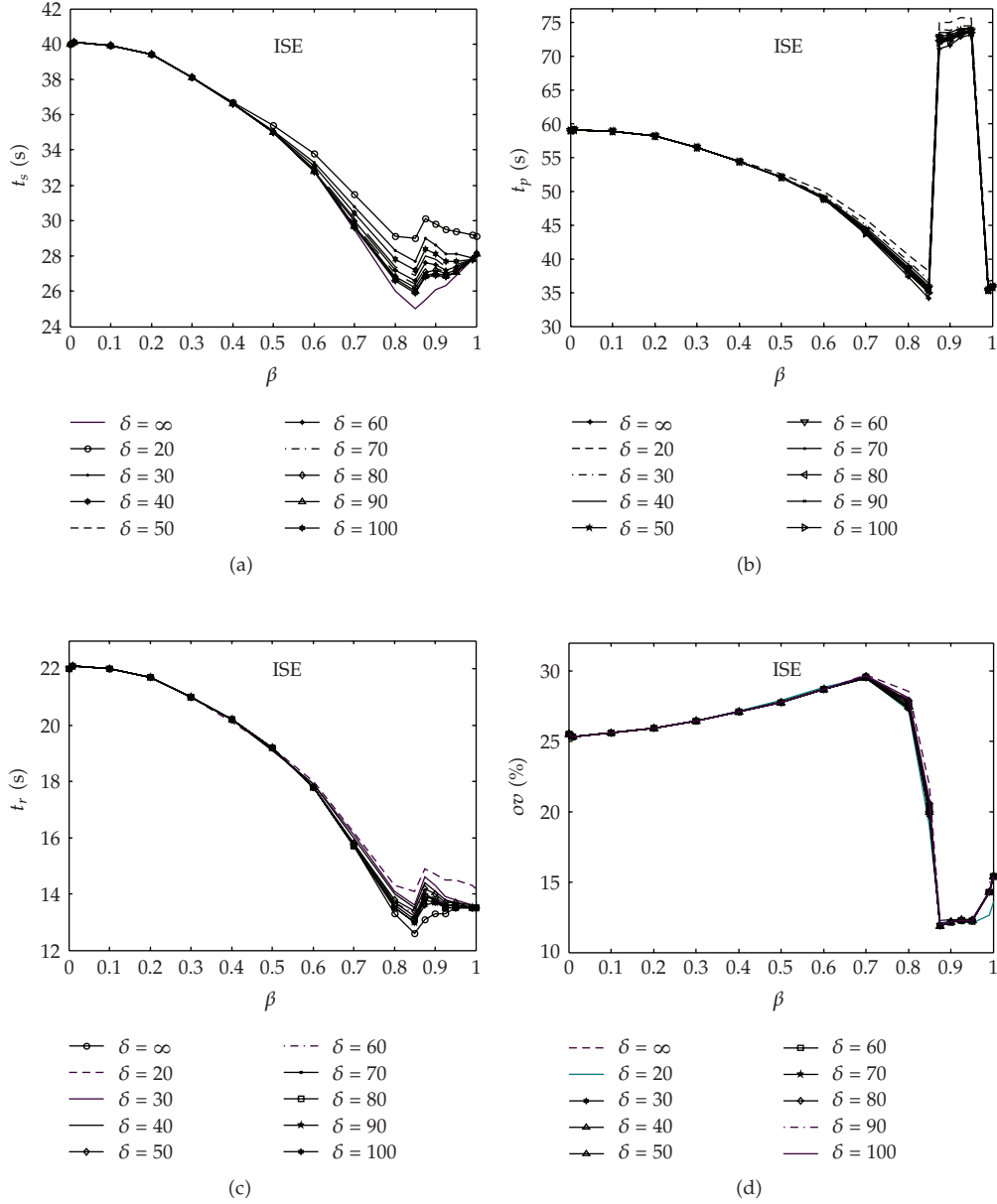
Figure 23: Parameters t_s, t_p, t_r, $ov(\%)$ for the step responses of the closed-loop system for the ISE indice, with a PID^β controller, when $\delta = \{20, \ldots, 100\}$ and $\delta = \infty$, $x = 3.0\,\mathrm{m}$.

Noncomputer scientists (ornithologists, biologists, and psychologists) did early research, which led into the theory of particle swarms. In these areas, the term "swarm intelligence" is well known and characterizes the case when a large number of individuals are able of accomplish complex tasks. Motivated by these facts, some basic simulations of swarms were abstracted into the mathematical field. The usage of swarms for solving simple tasks in nature became an intriguing idea in algorithmic and function optimization.
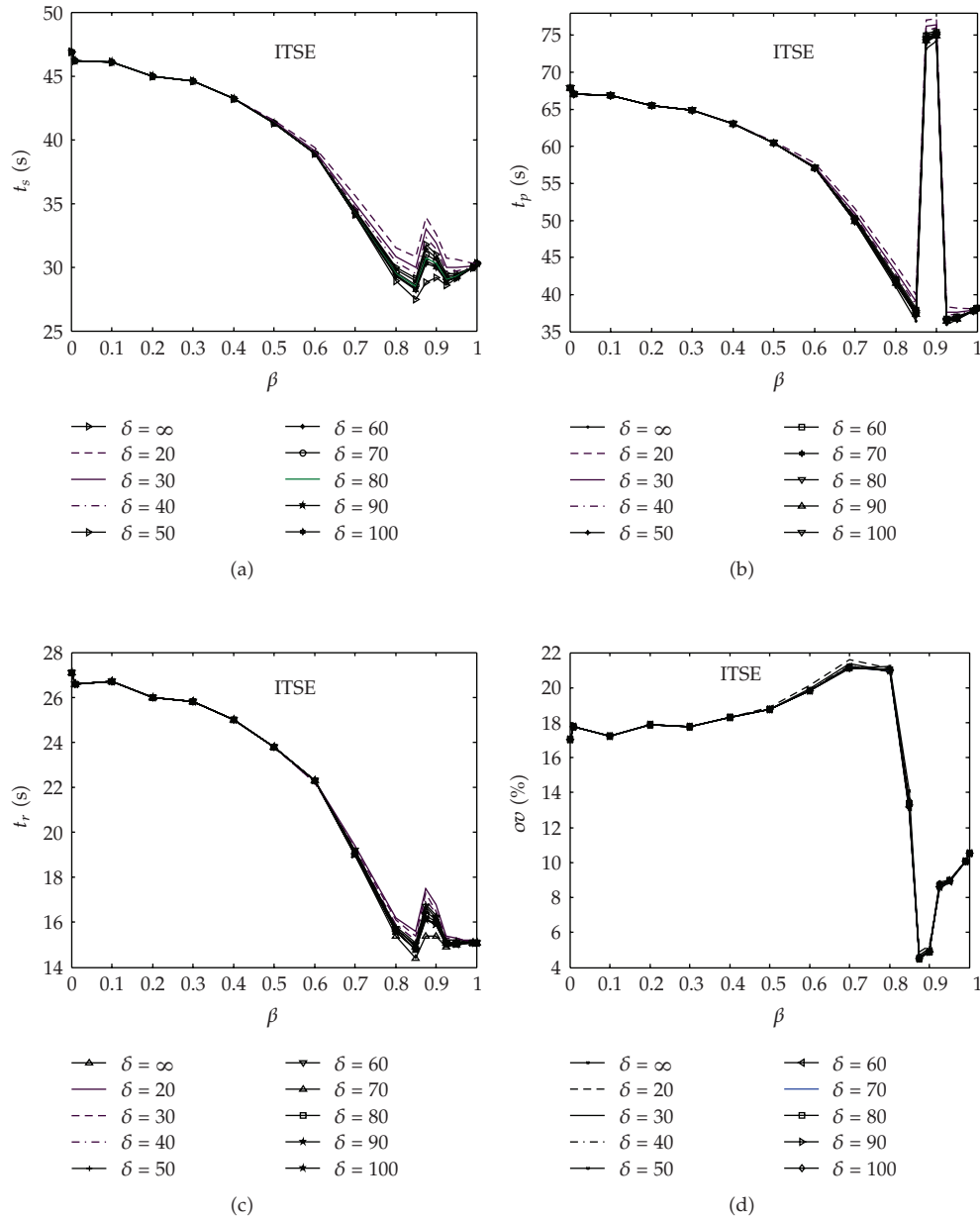
Figure 24: Parameters t_s, t_p, t_r, $ov(\%)$ for the step responses of the closed-loop system for the ITSE indice, with a PID$^\beta$ controller, when $\delta = \{20, \ldots, 100\}$ and $\delta = \infty$, $x = 3.0$ m.

Eberhart and Kennedy were the first to introduce the PSO algorithm [35], which is an optimization method inspired in the collective intelligence of swarms of biological populations, and was discovered through simplified social model simulation of bird flocking, fishing schooling, and swarm theory.

In the PSO, instead of using genetic operators, as in the case of GAs, each particle (individual) adjusts its flying according with its own and its companions experiences. Each

particle is treated as a point in a D-dimensional space and is manipulated as described in what follows in the original PSO algorithm:

$$v_{id} = v_{id} + c_1 \text{rand}(\ )(p_{id} - x_{id}) + c_2 \text{Rand}(\ )(p_{gd} - x_{id}),$$ (6.1a)

$$x_{id} = x_{id} + v_{id},$$ (6.1b)

where c_1 and c_2 are positive constants, rand() and Rand() are two random functions in the range $[0, 1]$, $X_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ represents the ith particle, $P_i = (p_{i1}, p_{i2}, \ldots, p_{iD})$ is the best previous position (the position giving the best fitness value) of the particle, the symbol g represents the index of the best particle among all particles in the population, and $V_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$ is the rate of the position change (velocity) for particle i.

However, (6.1a) and (6.1b) represent the flying trajectory of a population of particles. Also, (6.1a) describes how the velocity is dynamically updated and (6.1b) the position update of the "flying" particles. Moreover, (6.1b) is divided in three parts, namely the momentum, the cognitive and the social parts. In the first part the velocity cannot be changed abruptly: it is adjusted based on the current velocity. The second part represents the learning from its own flying experience. The third part consists on the learning group flying experience [36].

The first new parameter added into the original PSO algorithm is the inertia weigh. The dynamic equation of PSO with inertia weigh is modified to be

$$v_{id} = wv_{id} + c_1 \text{rand}(\ )(p_{id} - x_{id}) + c_2 \text{Rand}(\ )(p_{gd} - x_{id}),$$ (6.2a)

$$x_{id} = x_{id} + v_{id},$$ (6.2b)

where w constitutes the inertia weigh that introduces a balance between the global and the local search abilities. A large inertia weigh facilitates a global search while a small inertia weigh facilitates a local search.

Another parameter, called constriction coefficient k, is introduced with the hope that it can insure a PSO to converge. A simplified method of incorporating it appears in (6.3), where k is function of c_1 and c_2 as it is presented as follows:

$$v_{id} = k\left[v_{id} + c_1 \text{rand}(\ )(p_{id} - x_{id}) + c_2 \text{Rand}(\ )(p_{gd} - x_{id})\right],$$

$$x_{id} = x_{id} + v_{id},$$ (6.3)

$$k = 2\left(2 - \phi - \sqrt{\phi^2 - 4\phi}\right)^{-1},$$ (6.4)

where $\phi = c_1 + c_2$, $\phi > 4$.

There are two different PSO topologies, namely, the global version and the local version. In the global version of PSO, each particle flies through the search space with a velocity that is dynamically adjusted according to the particle's personal best performance achieved so far and the best performance achieved so far by all particles. On the other hand, in the local version of PSO, each particle's velocity is adjusted according to its personal best and the best performance achieved so far within its neighborhood. The neighborhood of each particle is generally defined as topologically nearest particles to the particle at each side.

1. Initialize the population
2. Calculate the fitness of each individual in the population
3. Reproduce selected individuals to form a new population
4. Perform evolutionary operations such as crossover and mutation on the population
5. Loop to step 2 until some condition is met

Figure 25: Evolutionary computation algorithm.

1. Initialize the population
2. Calculate the fitness of each individual in the population
3. Reproduce selected individuals to form a new population
4. Perform evolutionary operations such as crossover and mutation on the population
5. Apply a local search algorithm
5. Loop to step 2 until some condition is met

Figure 26: Memetic algorithm.

PSO is an evolutionary algorithm simple in concept, easy to implement and computationally efficient. Figures 25, 26, and 27 present a generic EC algorithm, a hybrid algorithm, more precisely a MA and the original procedure for implementing the PSO algorithm, respectively.

The different versions of the PSO algorithms are the real-value PSO, which is the original version of PSO and is well suited for solving real-value problems; the binary version of PSO, which is designed to solve binary problems; and the discrete version of PSO, which is good for solving the event-based problems. To extend the real-value version of PSO to binary/discrete space, the most critical part is to understand the meaning of concepts such as trajectory and velocity in the binary/discrete space.

Kennedy and Eberhart [35] use velocity as a probability to determine whether x_{id} (a bit) will be in one state or another (zero or one). The particle swarm formula of (6.1a) remains unchanged, except that now p_{id} and x_{id} are integers in $[0.0, 1.0]$ and a logistic transformation $S(v_{id})$ is used to accomplish this modification. The resulting change in position is defined by the following rule:

$$\text{if } [\text{rand}(\,) < S(v_{id})] \text{ then } x_{id} = 1; \text{ else } x_{id} = 0, \tag{6.5}$$

where the function $S(v)$ is a sigmoid limiting transformation and rand( ) is a random number selected from a uniform distribution in the range $[0.0, 1.0]$.
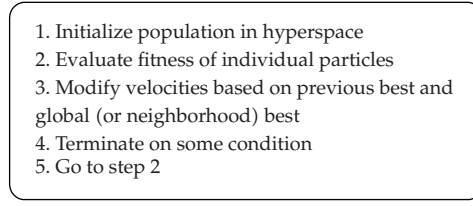
Figure 27: Particle swarm optimization process.
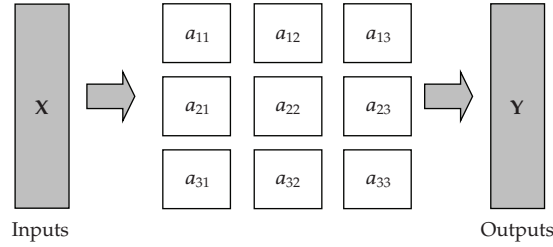
Figure 28: A 3×3 matrix representing a circuit with input X and output Y.

6.2. PSO Based Circuit Design

We adopt a PSO algorithm to design combinational logic circuits. A truth table specifies the circuits and the goal is to implement a functional circuit with the least possible complexity. Four sets of logic gates have been defined, as shown in Table 2, being *Gset 2* the simplest one (i.e., a RISC-like set) and *Gset 6* the most complex gate set (i.e., a CISC-like set). Logic gate named WIRE means a logical no-operation.

In the PSO scheme the circuits are encoded as a rectangular matrix **A** (*row × column* = $r \times c$) of logic cells as represented in Figure 28.

Three genes represent each cell: *<input1><input2><gate type>*, where *input1* and *input2* are one of the circuit inputs, if they are in the first column, or one of the previous outputs, if they are in other columns. The gate type is one of the elements adopted in the gate set. The chromosome is formed with as many triplets as the matrix size demands (e.g., triplets = $3 \times r \times c$). For example, the chromosome that represents a 3×3 matrix is depicted in Figure 29.

The initial population of circuits (particles) has a random generation. The initial velocity of each particle is initialized with zero. The following velocities are calculated applying (6.2a) and the new positions result from using (6.2b). This way, each potential solution, called particle, flies through the problem space. For each gene is calculated the corresponding velocity. Therefore, the new positions are as many as the number of genes in the chromosome. If the new values of the input genes result out of range, then a re-insertion function is used. If the calculated gate gene is not allowed a new valid one is generated at random. These particles then have memory and each keeps information of its previous best position (*pbest*) and its corresponding fitness. The swarm has the *pbest* of all the particles and the particle with the greatest fitness is called the global best (*gbest*).

The basic concept of the PSO technique lies in accelerating each particle towards its *pbest* and *gbest* locations with a random weighted acceleration. However, in our case we also use a kind of mutation operator that introduces a new cell in 10% of the population. This mutation operator changes the characteristics of a given cell in the matrix. Therefore, the

0	1	2		24	25	26	Genes
Input	Input	Gate	...	Input	Input	Gate	
a_{11}			...	a_{33}			Matrix element

Figure 29: Chromosome for the 3×3 matrix of Figure 28.

mutation modifies the gate type and the two inputs, meaning that a completely new cell can appear in the chromosome.

To run the PSO we have also to define the number P of individuals to create the initial population of particles. This population is always the same size across the generations, until reaching the solution.

The calculation of the fitness function F_s in (6.6) has two parts, f_1 and f_2, where f_1 measures the functionality and f_2 measures the simplicity. In a first phase, we compare the output Y produced by the PSO-generated circuit with the required values Y_R, according with the truth table, on a bit-per-bit basis. By other words, f_1 is incremented by one for each correct bit of the output until f_1 reaches the maximum value f_{10} that occurs when we have a functional circuit. Once the circuit is functional, in a second phase, the algorithm tries to generate circuits with the least number of gates. This means that the resulting circuit must have as much genes *<gate type> ≡ <wire>* as possible. Therefore, the index f_2, that measures the simplicity (the number of null operations), is increased by *one* (*zero*) for each *wire* (*gate*) of the generated circuit, yielding

$$f_{10} = 2^{ni} \times no,$$

$$f_1 = f_1 + 1 \text{ if } \{\text{bit } i \text{ of } \mathbf{Y}\} = \{\text{bit } i \text{ of } \mathbf{Y_R}\}, \quad i = 1, \dots, f_{10},$$

$$f_2 = f_2 + 1 \text{ if } gate\ type = wire, \tag{6.6}$$

$$F_s = \begin{cases} f_1, & F_s < f_{10}, \\ f_1 + f_2, & F_s \geq f_{10}, \end{cases}$$

where *ni* and *no* represent the number of inputs and outputs of the circuit.

The concept of dynamic fitness function F_d results from an analogy between control systems and the GA case, where we master the population through the fitness function. The simplest control system is the proportional algorithm; nevertheless, there can be other control algorithms, such as the proportional and the differential scheme.

In this line of thought, (6.6) is a static fitness function F_s and corresponds to using a simple proportional algorithm. Therefore, to implement a proportional-derivative evolution the fitness function needs a scheme of the type [18]

$$F_d = F_s + KD^{\mu}[F_s], \tag{6.7}$$

where $0 \leq \mu \leq 1$ is the differential fractional-order and $K \in \Re$ is the "gain" of the dynamical term.

(a)

(b)

(c)
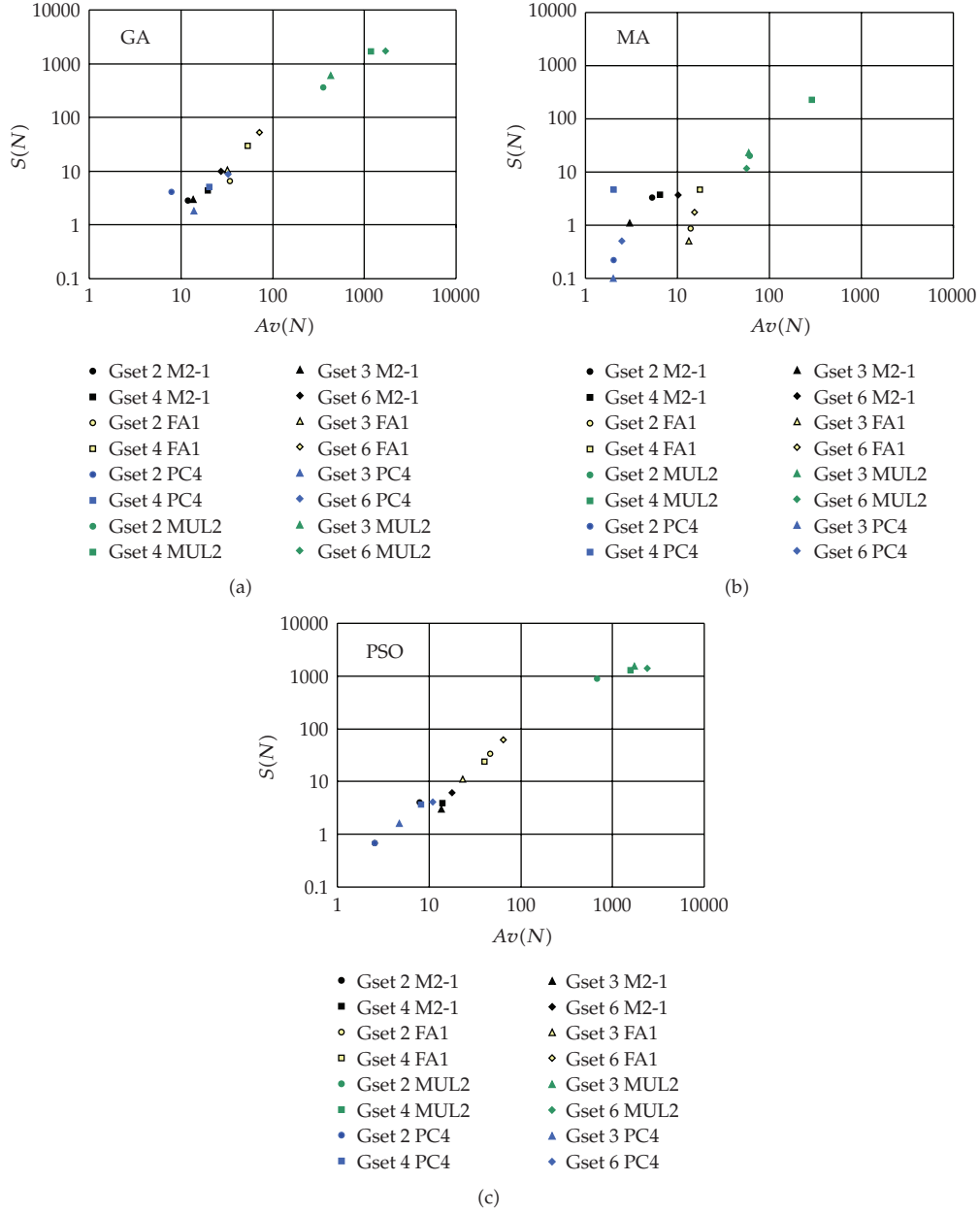
Figure 30: $S(N)$ versus $Av(N)$ with $P = 3000$ and F_s for the GA, the MA, and the PSO algorithms.

6.3. Experiments and Results

A reliable execution and analysis of an EC algorithm usually requires a large number of simulations to provide a reasonable assurance that the stochastic effects are properly considered. Therefore, in this study are developed $n = 20$ simulations for each case under analysis.

(a)

(b)

(c)
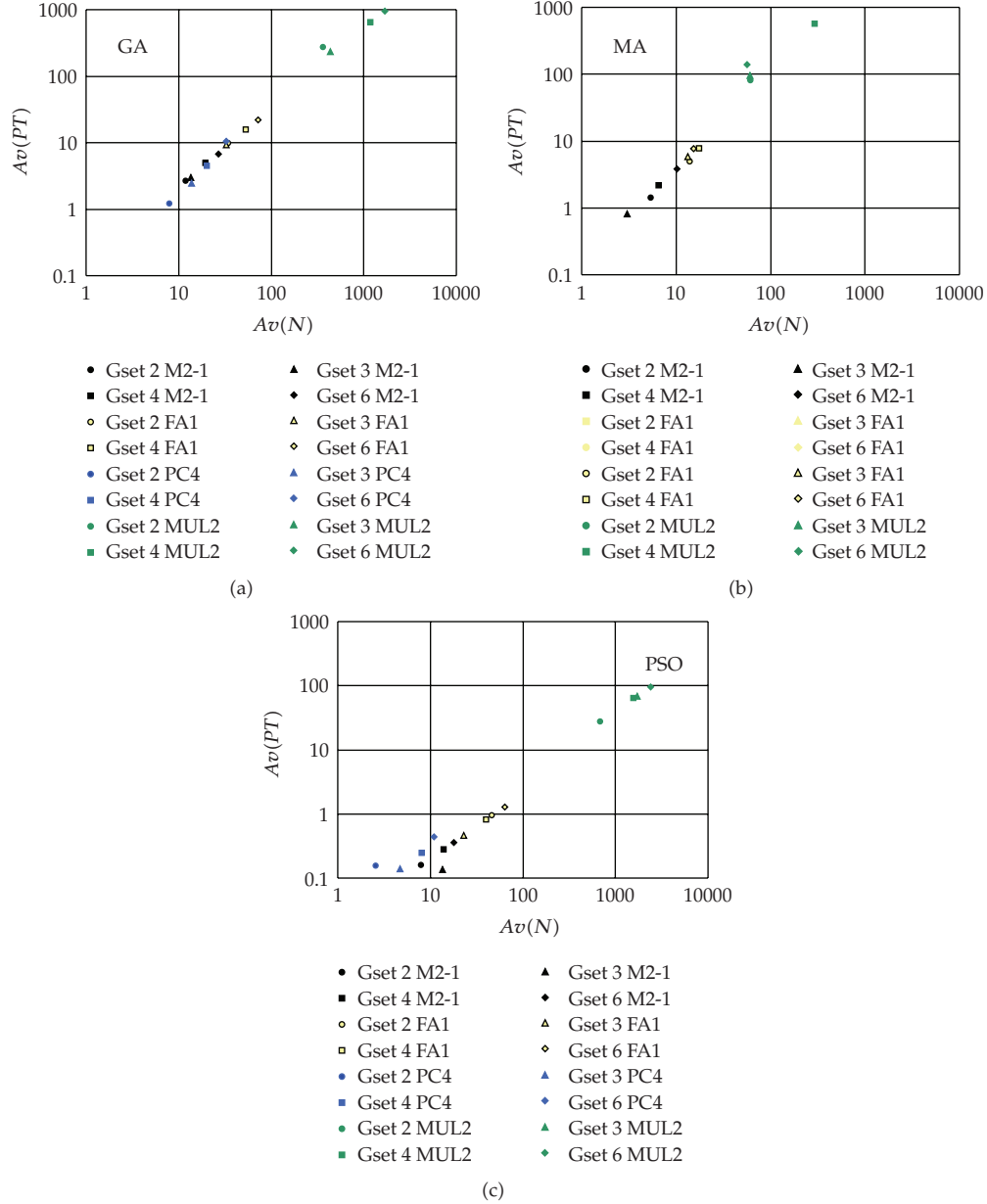
Figure 31: $Av(PT)$ versus $Av(N)$ with $P = 3000$ and F_s for the GA, the MA, and the PSO algorithms.

The experiments consist on running the three algorithms {GA, MA, PSO} to generate a typical combinational logic circuit, namely, a 2-to-1 multiplexer (M2-1), a 1-bit full adder ($FA1$), a 4-bit parity checker ($PC4$) and a 2-bit multiplier ($MUL2$), using the fitness scheme described in (6.6) and (6.7). The circuits are generated with the gate sets presented in Table 2 and $P = 3000$, $w = 0.5$, $c_1 = 1.5$, and $c_2 = 2$.

Figure 30 depicts the standard deviation of the number of generations to achieve the solution $S(N)$ versus the average number of generations to achieve the solution $Av(N)$

Table 2: Gate sets.

Gate set	Logic gates
Gset 2	{AND, XOR, WIRE}
Gset 3	{AND, OR, XOR, WIRE}
Gset 4	{AND, OR, XOR, NOT, WIRE}
Gset 6	{AND, OR, XOR, NOT, NAND, NOR, WIRE}

Table 3: The parameters (a, b) and (c, d).

Algorithm	a	b	c	d
GA	0.0365	1.602	0.1526	1.1734
MA	0.0728	1.2602	0.2089	1.3587
PSO	0.2677	1.1528	0.0141	1.1233

for the algorithms {GA, MA, PSO}, the circuits {M2-1, FA1, PC4, MUL2}, and the gate sets {2, 3, 4, 6}. In these figure, we can see that the $MUL2$ circuit is the most complex one, while the $PC4$ and the M2-1 are the simplest circuits. It is also possible to conclude that *Gset 6* is the less efficient gate set for all algorithms and circuits.

Figure 30 reveals that the plots follow a power law:

$$S(N) = a[Av(N)]^b \quad a, b \in \Re. \tag{6.8}$$

Table 3 presents the numerical values of the parameters (a, b) for the three algorithms.

In terms of $S(N)$ versus $Av(N)$, the MA algorithm presents the best results for all circuits and gate sets. In what concerns the other two algorithms, the PSO is superior (inferior) to the GA for complex (simple) circuits.

Figure 31 depicts the average processing time to obtain the solution $Av(PT)$ versus the average number of generations to achieve the solution $Av(N)$ for the algorithms {GA, MA, PSO}, the circuits {M2-1, FA1, PC4, MUL2} and the gate sets {2, 3, 4, 6}. When analysing these charts it is clear that the PSO algorithm demonstrates to be around ten times faster than the MA and the GA algorithms.

These plots follow also a power law:

$$Av(PT) = c[Av(N)]^d \quad c, d \in \Re. \tag{6.9}$$

Table 3 shows parameters (c, d) and we can see that the PSO algorithm has the best values.

Figures 32 and 33 depict the standard deviation of the number of generations to achieve the solution $S(N)$ and the average processing time to obtain the solution $Av(PT)$, respectively, versus the average number of generations to achieve the solution $Av(N)$ for the PSO algorithm using F_d, the circuits {M2-1, FA1, PC4, MUL2}, and the gate sets {2, 3, 4, 6}. We conclude that F_d leads to better results in particular for the $MUL2$ circuit and for the $Av(PT)$.

Figures 34 and 35 present a comparison between F_s and F_d.

In terms of $S(N)$ versus $Av(N)$ it is possible to say that the MA algorithm presents the best results. Nevertheless, when analysing Figure 31, that shows $Av(PT)$ versus $Av(N)$ for reaching the solutions, we verify that the PSO algorithm is very efficient, in particular, for the more complex circuits.
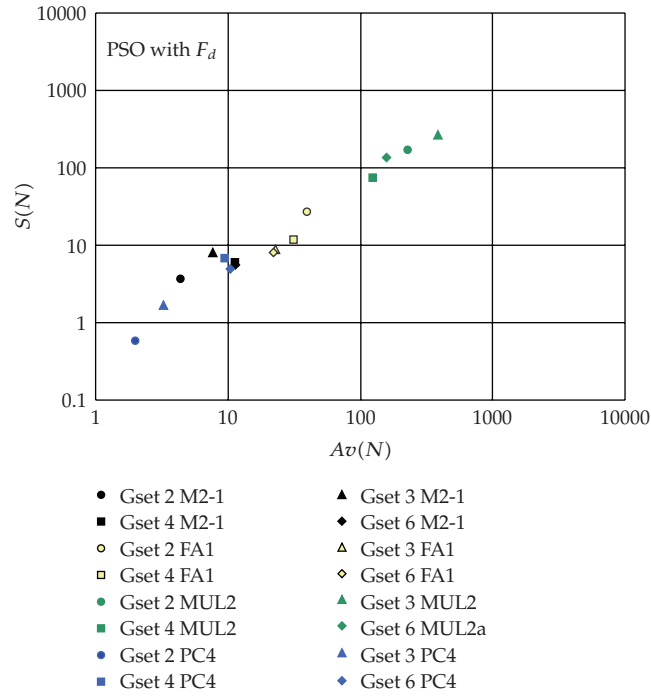
Figure 32: $S(N)$ versus $Av(N)$ for the PSO algorithm, $P = 3000$ and F_d.
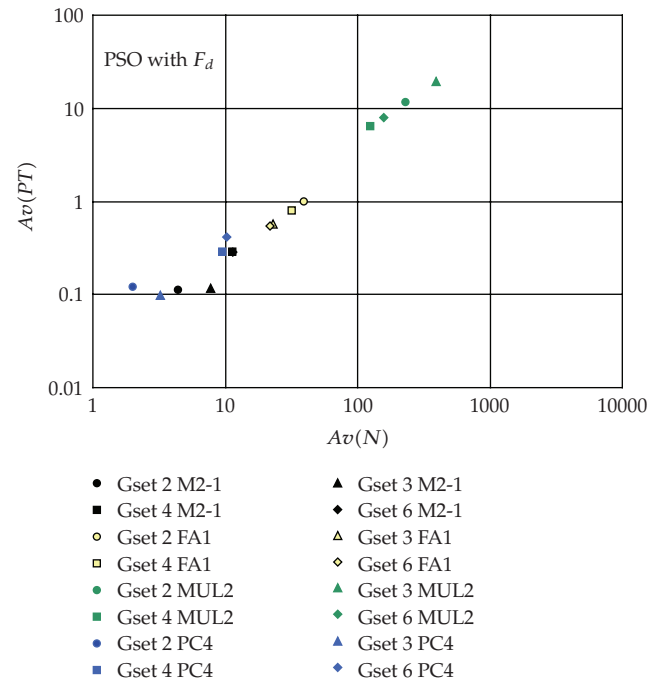
Figure 33: $Av(PT)$ versus $Av(N)$ for the GA, $P = 3000$ and F_d.
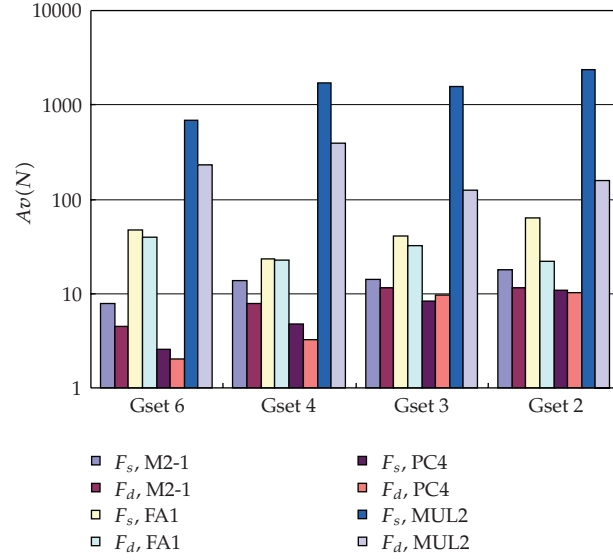
Figure 34: $Av(N)$ for the PSO algorithm, $P = 3000$ using F_s and F_d.
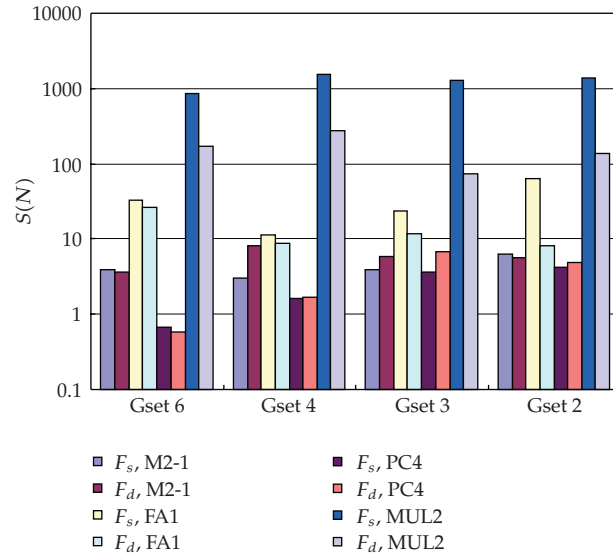
Figure 35: $S(N)$ for the PSO algorithm, $P = 3000$ using F_s and F_d.

The PSO-based algorithm for the design of combinational circuits follows the same profile as the other two evolutionary techniques presented in this paper.

Adopting the study of the $S(N)$ versus $Av(N)$ for the three evolutionary algorithms, the MA algorithm presents better results over the GA and the PSO algorithms. However, in what concerns the processing time to achieve the solutions, the PSO outcomes clearly the GA and the MA algorithms. Moreover, applying the F_d the results obtained are improved further in all gate sets and in particular for the more complex circuits.

7. Conclusions

Fractional Calculus (FC) goes back to the beginning of the theory of differential calculus. Nevertheless, the application of FC just emerged in the last two decades, due to the progress in the area of chaos that revealed subtle relationships with the FC concepts.

Recently FC has been a fruitful field of research in science and engineering and many scientific areas are currently paying wider attention to the FC concepts. In the field of dynamical systems theory, some work has been carried out but the proposed models and algorithms are still in a preliminary stage of establishment. This article presented several case studies on the implementation of FC-based models and control systems, being demonstrated the advantages of using the FC theory in different areas of science and engineering. In fact, this paper studied a variety of different physical systems, namely

  (i) tuning of PID controllers using fractional calculus concepts;

  (ii) fractional PD^{α} control of a hexapod robot;

  (iii) fractional dynamics in the trajectory control of redundant manipulators;

  (iv) heat diffusion;

  (v) circuit synthesis using evolutionary algorithms.

It has been recognized the advantageous use of this mathematical tool in the modeling and control of these dynamical systems, and the results demonstrate the importance of Fractional Calculus and motivate for the development of new applications.

References

[1] K. B. Oldham and J. Spanier, *The Fractional Calculus: Theory and Applications of Differentiation and Integration to Arbitrary Order*, vol. 11 of *Mathematics in Science and Engineering*, Academic Press, New York, NY, USA, 1974.

[2] K. S. Miller and B. Ross, *An Introduction to the Fractional Calculus and Fractional Differential Equations*, A Wiley-Interscience Publication, John Wiley & Sons, New York, NY, USA, 1993.

[3] I. Podlubny, *Fractional Differential Equations*, vol. 198 of *Mathematics in Science and Engineering*, Academic Press, San Diego, Calif, USA, 1999.

[4] R. Hilfer, Ed., *Applications of Fractional Calculus in Physics*, World Scientific Publishing, Singapore, 2000.

[5] A. Oustaloup, *La Commande CRONE: Commande Robuste d'Ordre Non Entier*, Editions Hermès, Paris, France, 1991.

[6] A. Oustaloup, *La Dérivation Non Entière: Théorie, Synthèse et Applications*, Editions Hermès, Paris, France, 1995.

[7] R. S. Barbosa, J. A. T. Machado, and I. M. Ferreira, "PID controller tuning using fractional calculus concepts," *Fractional Calculus & Applied Analysis*, vol. 7, no. 2, pp. 119–134, 2004.

[8] R. S. Barbosa, J. A. T. Machado, and I. M. Ferreira, "Tuning of PID controllers based on bode's ideal transfer function," *Nonlinear Dynamics*, vol. 38, no. 1–4, pp. 305–321, 2004.

[9] M. F. Silva, J. A. T. Machado, and A. M. Lopes, "Comparison of fractional and integer order control of an hexapod robot," in *Proceedings of International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 5 of *19th Biennial Conference on Mechanical Vibration and Noise*, pp. 667–676, ASME, Chicago, Ill, USA, September 2003.

[10] M. F. Silva, J. A. T. Machado, and I. S. Jesus, "Modelling and simulation of walking robots with 3 dof legs," in *Proceedings of the 25th IASTED International Conference on Modelling, Identification and Control (MIC '06)*, pp. 271–276, Lanzarote, Spain, 2006.

[11] M. F. Silva, J. A. T. Machado, and A. M. Lopes, "Position/force control of a walking robot," *Machine Intelligence and Robot Control*, vol. 5, pp. 33–44, 2003.

[12] M. F. Silva and J. A. T. Machado, "Fractional order PD^{α} joint control of legged robots," *Journal of Vibration and Control*, vol. 12, no. 12, pp. 1483–1501, 2006.

[13] F. Duarte and J. A. T. Machado, "Chaotic phenomena and fractional-order dynamics in the trajectory control of redundant manipulators," *Nonlinear Dynamics*, vol. 29, no. 1–4, pp. 315–342, 2002.

[14] J. A. T. Machado, "Analysis and design of fractional-order digital control systems," *Systems Analysis Modelling Simulation*, vol. 27, no. 2-3, pp. 107–122, 1997.

[15] J. A. T. Machado, "Discrete-time fractional-order controllers," *Fractional Calculus & Applied Analysis*, vol. 4, no. 1, pp. 47–66, 2001.

[16] J. A. T. Machado, I. S. Jesus, J. B. Cunha, and J. K. Tar, "Fractional dynamics and control of distributed parameter systems," *Intelligent Systems at the Service of Mankind*, vol. 2, pp. 295–305, 2006.

[17] I. S. Jesus, R. S. Barbosa, J. A. T. Machado, and J. B. Cunha, "Strategies for the control of heat diffusion systems based on fractional calculus," in *Proceedings of the IEEE International Conference on Computational Cybernetics (ICCC '06)*, Budapest, Hungary, 2006.

[18] C. Reis, J. A. T. Machado, and J. B. Cunha, "Evolutionary design of combinational circuits using fractional-order fitness," in *Proceedings of the 5th Nonlinear Dynamics Conference (EUROMECH '05)*, pp. 1312–1321, 2005.

[19] H. W. Bode, *Network Analysis and Feedback Amplifier Design*, Van Nostrand, New York, NY, USA, 1945.

[20] E. S. Conkur and R. Buckingham, "Clarifying the definition of redundancy as used in robotics," *Robotica*, vol. 15, no. 5, pp. 583–586, 1997.

[21] S. Chiaverini, "Singularity-robust task-priority redundancy resolution for real-time kinematic control of robot manipulators," *IEEE Transactions on Robotics and Automation*, vol. 13, no. 3, pp. 398–410, 1997.

[22] C. A. Klein and C. C. Huang, "Review of pseudoinverse control for use with kinematically redundant manipulators," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 13, no. 2, pp. 245–250, 1983.

[23] T. Yoshikawa, *Foundations of Robotics: Analysis and Control*, MIT Press, Cambridge, Mass, USA, 1988.

[24] J. S. Bay, "Geometry and prediction of drift-free trajectories for redundant machines under pseudoinverse control," *International Journal of Robotics Research*, vol. 11, no. 1, pp. 41–52, 1992.

[25] R. G. Roberts and A. A. Maciejewski, "Singularities, stable surfaces, and the repeatable behavior of kinematically redundant manipulators," *International Journal of Robotics Research*, vol. 13, no. 1, pp. 70–81, 1994.

[26] K. L. Doty, C. Melchiorri, and C. Bonivento, "A theory of generalized inverses applied to robotics," *International Journal of Robotics Research*, vol. 12, no. 1, pp. 1–19, 1993.

[27] Y. Nakamura, *Advanced Robotics: Redundancy and Optimization*, Addinson-Wesley, New York, NY, USA, 1991.

[28] B. Siciliano, "Kinematic control of redundant robot manipulators: a tutorial," *Journal of Intelligent and Robotic Systems*, vol. 3, no. 3, pp. 201–212, 1990.

[29] W. J. Chung, Y. Yorm, and W. K. Chung, "Inverse kinematics of planar redundant manipulators via virtual links with configuration index," *Journal of Robotic Systems*, vol. 11, no. 2, pp. 117–128, 1994.

[30] S. Seereeram and J. T. Wen, "A global approach to path planning for redundant manipulators," *IEEE Transactions on Robotics and Automation*, vol. 11, no. 1, pp. 152–159, 1995.

[31] M. da Graça Marcos, F. B. M. Duarte, and J. A. T. Machado, "Complex dynamics in the trajectory control of redundant manipulators," *Nonlinear Science and Complexity*, pp. 134–143, 2007.

[32] S. J. Louis, G. J. E. Rawlins, and G. J. Designer, "Genetic algorithms: genetic algorithms in structure design," in *Proceedings of the 4th International Conference on Genetic Algorithms*, 1991.

[33] D. E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Addison-Wesley, New York, NY, USA, 1989.

[34] C. Reis, J. A. T. Machado, and J. B. Cunha, "An evolutionary hybrid approach in the design of combinational digital circuits," *WSEAS Transactions on Systems*, vol. 4, no. 12, pp. 2338–2345, 2005.

[35] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1942–1948, November 1995.

[36] Y. Shi and R. C. Eberhart, "A modified particle swarm optimizer," in *Proceedings of the International Conference on Evolutionary Computation*, pp. 69–73, May 1998.

Review Article
Improved Continuous Models for Discrete Media

I. V. Andrianov,[1] J. Awrejcewicz,[2] and D. Weichert[1]

[1] *Institute of General Mechanic, RWTH Aachen University, Templergraben 64, 52056 Aachen, Germany*
[2] *Department of Automatics and Biomechanics, Technical University of Łódź,*
  1/15 Stefanowski St., 90-924 Łódź, Poland

Correspondence should be addressed to I. V. Andrianov, andrianov@iam.rwth-aachen.de

The paper focuses on continuous models derived from a discrete microstructure. Various continualization procedures that take into account the nonlocal interaction between variables of the discrete media are analysed.

1. Introduction

In the recent years new classes of ultra-dispersive and nanocrystalline materials [1–3], which require a different modern approach than that of a classical continuous media, have been proposed. For example, the nanocrystalline material is represented by a regular or quasiregular lattice with small size bodies (domains, granules, fullerenes, nanotubes, or clusters nanoparticles) possessing internal degrees of freedom occupying the lattice sites [4]. The situation mentioned is also a characteristic for various problems of nanomechanics [5, 6], because the transition of a material to the nanostructural state is accompanied by dimensional effects in its mechanical properties. Models established on a classical continuous media cannot govern high frequency vibrations, material behavior in the vicinity of cracks and on the fronts of destruction waves, and during phase transitions [7]. It is not possible to reach and overcome bifurcation points, that is, thresholds of lattice stability under catastrophic deformations [8]. Wave dispersion in granular materials [9–11] represents an important example of microstructure effects too. Microstructural effects are essential for correct description of softening phenomena [12] in damage mechanics [13–16] and in the theory of plasticity [17, 18]. As it is mentioned in [19], "From the behaviour of calcium waves in living cells to the discontinuous propagation of action potentials in the heart or chains of neurons and from chains of chemical reactors or arrays of Josephson junctions to optical waveguides, dislocations and the DNA double strand, the relevant models of physical reality are inherently discrete."

The effects mentioned may be analyzed within the frame of discrete models, using molecular dynamics [20], quasicontinuum analysis [21, 22], or other numerical approaches. However, a sought result is difficult to obtain using high tech computers in an economical way. For example, modern practical problems are still intractable for molecular dynamics-based analysis, even if the highest computing facility is at disposal.

Situation can be described by Dirac's words [23]: "The physical laws necessary for the mathematical theory of a large part of physics and whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be solvable. Therefore, it becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computations."

Hence, refinement of the existing theory of continuous media for the purpose of more realistic predictions seems to be the only viable alternative. In connection with this, one of the most challenging problems in multiscale analysis is that of finding continuous models for discrete, atomistic models. In statistical physics, these questions were already addressed 100 years ago, but many problems remain open even today. Most prominent is the question how to obtain irreversible thermodynamics as a macroscopic limit from microscopic models that are reversible. In this paper we consider another part of this field that is far from thermodynamic fluctuation. We are interested in reversible, macroscopic limits of atomic models. Debye approach is the simplest model of this type [24], but it does not take into account spatial dispersion.

Therefore, continuous modeling of micro- and nanoeffects plays a crucial role in mechanics. It seems that the simplest approach to realize this idea relies on a modification of the classical modeling keeping both hypothesis of continuity and main characteristic properties of a discrete structure. Here four strategies exist.

Phenomenological approach: additional terms are added to the energy functional or to the constitutive relation. The structure and character of these terms are postulated [25–31]. Phenomenological approach is very useful and accurate enough in the applied sciences when it is necessary to solve problems of practical significance [32–34]. But progress of natural sciences makes us look for ways of substantiated derivation of constitutive relations from "first principles." In this connection one can recall the 6th Hilbert's problem (mathematical treatment of the axioms of physics) [35]: "Boltzmann's work on the principles of mechanics suggests the problem of developing mathematically the limiting processes, there merely indicated, which lead from the atomistic view to the laws of motion of continua. Conversely, one might try to derive the laws of the rigid bodies' motion by a limiting process from a system of axioms depending upon the idea of continuously varying conditions of a material filling all space continuously, these conditions being defined by parameters. The question on the equivalence of different systems of axioms is always of great theoretical interest."

Statistical approach: starting from an inhomogeneous classical continuum, average values of the state variables are computed to produce enhanced field equations [36]. Unfortunately, great mathematical problems can not give possibility of wide usage of this approach.

Homogenization approach is based on Γ- and G-limit technique [37–43] and gave many important pure mathematical results, but not new insights for physics.

Continualization procedures are based on various approximations of local (discrete) operator by the nonlocal one. For this aim Taylor series [44, 45], one- and two-point Padé approximants [46–62], composite equations [63, 64], and other approaches are used.
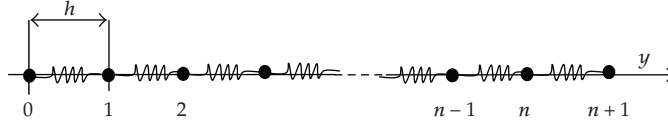
Figure 1: A chain of elastically coupled masses.

Independently of the strategy employed, the constitutive relation of the homogenized material becomes nonlocal. Mathematically this implies integral relations between stresses and strains, in which the stress in a point is not only related to the strain in the same point but also instantly to strains in the neighbouring points. We will analyse these approaches further.

The paper is organized as follows. Section 2 is devoted to some general remarks concerning a chain of elastically linear coupled masses. In Sections 3 and 4 the classical continuous approximation and so-called "splash effect" for 1D case are described. Section 5 is devoted to anticontinuum limit. Improved continuous approximations are studied is Sections 6 and 7 for natural and forced oscillations. Waves in 1D discrete and continuous media are compared in Section 8. 2D problems are analysed in Section 9. Nonlinear phenomenons are studied in Section 10. In Section 11 we analyse some possible generalization of described approaches. Section 12 is devoted to Navier-Stokes equations. Section 13 presents brief concluding remarks. In appendixes to the paper we describe one- and two-point Padé approximations, continuum limit of Toda lattice, and correspondence between functions of discrete arguments and approximating analytical functions.

2. A Chain of Elastically Coupled Masses

In this section we follow paper [65]. We study a chain of $n+1$ material points with the same masses m, located in equilibrium states in the points of the axis x with coordinates jh ($j = 0, 1, \ldots, n, n+1$) and suspended by elastic couplings of stiffness c (Figure 1).

Owing to the Hooke's law the elastic force acting on the jth mass is as follows:

$$\sigma_j(t) = c\big[y_{j+1}(t) - y_j(t)\big] - c\big[y_j(t) - y_{j-1}(t)\big] = c\big[y_{j-1}(t) - 2y_j(t) + y_{j+1}(t)\big],$$
$$j = 1, 2, \ldots, n, \tag{2.1}$$

where $y_j(t)$ is the displacement of the jth material point from its static equilibrium position.

Applying the 2nd Newton's law one gets the following system of ODEs governing chain dynamics:

$$m y_{jtt}(t) = c\big[y_{j-1}(t) - 2y_j(t) + y_{j+1}(t)\big], \quad j = 1, 2, \ldots, n. \tag{2.2}$$

System (2.2) can be cast into the following form:

$$m\sigma_{jtt}(t) = c\big(\sigma_{j+1} - 2\sigma_j + \sigma_{j-1}\big), \quad j = 1, \ldots, n. \tag{2.3}$$

Let the chain ends be fixed

$$y_o(t) = y_{n+1}(t) = 0. \tag{2.4}$$

In general, the initial conditions have the following form:

$$y_j(t) = \varphi_j^{(0)}; \quad y_{jt}(t) = \varphi_j^{(1)} \quad \text{for } t = 0. \tag{2.5}$$

As it has been shown in [65], for any solution of the BVP (2.2), (2.4), (2.5) the total energy is constant. Besides, the solutions mentioned so far are nonasymptotic and stable due to the Lyapunov stability definition.

A solution to the BVP (2.2), (2.4), (2.5) can be expressed by elementary functions applying the discrete variant of the method of variables separation. For this purpose normal vibrations are constructed of the form

$$y_j(t) = C_j T(t), \quad j = 1,\ldots,n, \tag{2.6}$$

where constants C_j are defined via solution of the following eigenvalue problem:

$$-\lambda C_j = C_{j+1} - 2C_j + C_{j-1}, \quad j = 1,\ldots,n, \ C_0 = C_{n+1} = 0. \tag{2.7}$$

Function $T(t)$ satisfies the following equation:

$$mT_{tt} + c\lambda T = 0. \tag{2.8}$$

Eigenvalues of the problem (2.7) follow [65]

$$\lambda_k = 4\sin^2 \frac{k\pi}{2(n+1)}, \quad k = 1,2,\ldots,n. \tag{2.9}$$

A solution to (2.8) has the form $T = A\exp(i\omega t)$. Hence, (2.8) and (2.9) yield the following Lagrange formula for frequencies ω_k of discrete system:

$$\omega_k = 2\sqrt{\frac{c}{m}} \sin \frac{k\pi}{2(n+1)}, \quad k = 1,2,\ldots,n. \tag{2.10}$$

Since all values λ_k are distinct, then all eigenvalues are different. Therefore, each of the eigenvalues is associated with one eigenvector $\mathbf{C}_k(C_1^{(k)}, C_2^{(k)}, \ldots, C_n^{(k)})$ of the form

$$\mathbf{C}_k = \cos ec \frac{k\pi}{n+1} \left(\sin \frac{k\pi}{n+1}, \sin \frac{2k\pi}{n+1}, \ldots, \sin \frac{nk\pi}{n+1} \right), \quad k = 1,2,\ldots,n. \tag{2.11}$$

Eigenvectors are mutually orthogonal; whereas

$$|\mathbf{C}_k|^2 = \frac{n+1}{2}\cos ec^2 \frac{k\pi}{n+1}, \quad k = 1,2,\ldots,n. \tag{2.12}$$

Each of the eigenfrequencies (2.10) is associated with a normal vibration

$$y_j^{(k)}(t) = C_j^{(k)}[A_k\cos(\omega_k t) + B_k\sin(\omega_k t)], \quad k = 1,2,\ldots,n. \tag{2.13}$$

A general solution of the BVP (2.3)–(2.5) is obtained as a result of superposition of normal vibrations

$$y_j(t) = \sum_{k=1}^{n} C_j^{(k)}[A_k\cos(\omega_k t) + B_k\sin(\omega_k t)], \quad j = 1,\ldots,n. \tag{2.14}$$

Let us study now the problem of chain masses movement under action of a unit constant force on the point number zero. Motion of such system is governed by (2.3) with the following boundary and initial conditions:

$$\sigma_o(t) = 1, \qquad \sigma_{n+1}(t) = 0,$$
$$\sigma_j(t) = \sigma_{jt}(t) = 0 \quad \text{for } t = 0. \tag{2.15}$$

In what follows the initial BVP with nonhomogeneous boundary conditions (2.3), (2.15) will be reduced to that of homogeneous boundary condition for (2.3) with nonhomogeneous initial conditions, and then the method of superposition of normal vibrations can be applied. The formulas of normal forms obtained so far can be applied in a similar way with exchange of $y_j(t)$ for $\sigma_j(t)$. In effect, the following exact solution to the BVP (2.3), (2.15) is obtained [65]:

$$\sigma_j = \frac{1}{n+1}\sum_{k=1}^{n}\sin\frac{\pi k j}{n+1}ctg\frac{\pi k}{2(n+1)}[1 - \cos(\omega_k t)], \quad j = 1,2,\ldots,n. \tag{2.16}$$

3. Classical Continuous Approximations

For large values of n usually continuous approximation of discrete problem is applied. In our case, described by (2.3), (2.15), it takes the form of

$$m\sigma_{tt}(x,t) = ch^2\sigma_{xx}(x,t), \tag{3.1}$$

$$\sigma(0,t) = 1, \qquad \sigma(l,t) = 0, \tag{3.2}$$

$$\sigma(x,0) = \sigma_t(x,0) = 0, \tag{3.3}$$

where $l = (n+1)h$.

BVP (3.1)–(3.3) can be used, for example, for modeling of stresses in van couplings of the rolling stocks [66].

Having in hand a solution to continuous BVP (3.1)–(3.3), one obtains a solution of a discrete problem due to the following formulas:

$$\sigma_j(t) = \sigma(jh,t), \quad j = 0, 1, \ldots, n, \ n+1. \tag{3.4}$$

Formally, the approximation described so far can be obtained in the following way. Let us denote the difference operator occurring in (2.3) as D, that is,

$$m\sigma_{jtt}(t) = cD\sigma(t). \tag{3.5}$$

Applying the translation operator $\exp(h\partial/\partial x)$, one gets [67]

$$D = \exp\left(h\frac{\partial}{\partial x}\right) + \exp\left(-h\frac{\partial}{\partial x}\right) - 2 = -4\sin^2\left(-\frac{ih}{2}\frac{\partial}{\partial x}\right). \tag{3.6}$$

Let us explain (3.6) in more details. The McLaurin formula for infinitely many times differentiable function $F(x)$ has the following form:

$$F(x+1) = \left[1 + \frac{\partial}{\partial x} + \frac{1}{2!}\frac{\partial^2}{\partial x^2} + \cdots\right]F(x) = \exp\left(\frac{\partial}{\partial x}\right)F(x). \tag{3.7}$$

Observe that $\exp(h\partial/\partial x)$ belongs to the so-called pseudodifferential operators. Using relations (3.5)–(3.7), we cast (2.3) into pseudo-differential equation of the following form:

$$m\frac{\partial^2\sigma}{\partial t^2} + 4c\sin^2\left(-\frac{ih}{2}\frac{\partial}{\partial x}\right)\sigma = 0. \tag{3.8}$$

On the other hand, splitting the pseudo-differential operator into the McLaurin series is as follows:

$$\sin^2\left(-\frac{ih}{2}\frac{\partial}{\partial x}\right) = -\frac{1}{2}\sum_{k=1}\frac{h^{2k}}{(2k)!}\frac{\partial^{2k}}{\partial x^{2k}} = -\frac{h^2}{4}\frac{\partial^2}{\partial x^2}\left(1 + \frac{h^2}{12}\frac{\partial^2}{\partial x^2} + \frac{h^4}{360}\frac{\partial^4}{\partial x^4} + \frac{h^6}{10080}\frac{\partial^6}{\partial x^6} + \cdots\right). \tag{3.9}$$

Keeping in right hand of (3.9) only the first term, one obtains a continuous approximation (3.1). Note that an application of the McLaurin series implies that displacements of the neighborhood masses differ slightly from each other. From a physical point of view, it means that we study vibrations of the chain with a few masses located on the spatial period (Figure 2); that is, the long wave approximation takes place. Note that the vertical axis in Figure 2 represents the displacement in the x direction, since the problem is 1D.
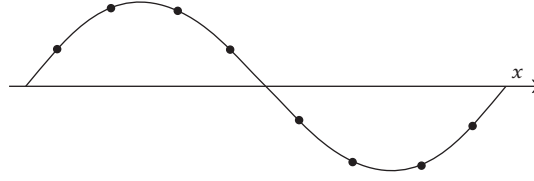
Figure 2: Solution form $\sigma = \sigma(x,t)$ in a fixed time instant $t = $ const (points—discrete system, curve—continuous system).

Table 1: Splashes.

n	8	16	32	64	128	256	$n \to \infty$
P_n	1.7561	2.0645	2.3468	2.6271	2.9078	3.1887	$P_n \to \infty$

Continuous system (3.1) possesses the following discrete infinite spectrum:

$$\alpha_k = \pi \sqrt{\frac{c}{m}} \frac{k}{n+1}, \quad k = 1, 2, \dots.\tag{3.10}$$

Relations (3.10) relatively good approximate low frequencies of discrete system (2.10), whereas the nth frequency α_k of a continuous system differs from the corresponding nth frequency ω_k of a discrete system of order of 50%. Accuracy of continuous approximations can be improved, what will be discussed further.

4. Splashes

One can obtain an exact solution to the BVP (3.1)–(3.3), using the d'Alembert method matched with operational calculus [65]:

$$\sigma(x,t) = H\left(nh \, \arcsin\left|\sin\left(\frac{\pi}{2n}\sqrt{\frac{c}{m}}t\right)\right| - x \right),\tag{4.1}$$

where $H(\cdots)$ is the Heaviside function.

From (4.1) one obtains the following estimation:

$$|\sigma(x,t)| \le 1.\tag{4.2}$$

It was believed that estimation (4.2) with a help of relation (3.4) can be applied also to a discrete system [68]. However, analytical as well as numerical investigations [66, 69–73] indicated a need to distinguish between global and local characteristics of a discrete system. In other words during investigation of lower frequency part a transition into continuous model is allowed. However, in the case of forced oscillations solutions to a discrete system may not be continuously transited into solution of a wave equation for $h \to 0$ [67]. Numerical investigations show that for given masses in a discrete chain quantity the $P_j = |\sigma_j(t)|$ may exceed the values of 1 in certain time instants (reported in [69, Table 5.1]).

Observe that splash amplitude does not depend on the parameter m/c. In addition, the amplitude of the chain vibrations increases with an increase of n; whereas its total energy does not depend on n. However, this is not a paradox. Namely, amplitude of vibrations has an order of sum of quantities $\sigma_j(t)$; whereas its potential energy order is represented by a sum of squares of the quantities mentioned [65].

On the other hand, a vibration amplitude of a mass with a fixed number is bounded for $n \to \infty$, but amplitude of vibrations of a mass with a certain number increasing with increase of n tends to infinity for $n \to \infty$ following $\ln n$ [65]. "In the language of mechanics what we just said means that when analyzing the so-called "local properties" of a one-dimensional continuous medium, one cannot treat the medium as the limiting case of a linear chain of point masses, obtained when the number of points increases without limit" [66].

It should be emphasized that a rigorous proof of the mentioned properties has been obtained for a case, when $n + 1$ is a simple digit or it is a power of 2. However, this assumption is not necessary, as it is mentioned in [65].

Earlier the same effect of continualization was predicted by Ulam, who wrote [74, pages 89, 90]: "The simplest problems involving an actual infinity of particles in distributions of matter appear already in classical mechanics." A discussion on these will permit us to introduce more general schemes which may possibly be useful in future physical theories.

Strictly speaking, one has to consider a true infinity in the distribution of matter in all problems of the physics of continua. In the classical treatment, as usually given in textbooks of hydrodynamics and field theory, this is, however, not really essential, and in most theories serves merely as a convenient limiting model of *finite* systems enabling one to use the algorithms of the calculus. The usual introduction of the continuum leaves much to be discussed and examined critically. The derivation of the equations of motion for fluids, for example, runs somewhat as follows. One images a very large number N of particles, say with equal masses constituting a net approximating the continuum, which is to be studied. The forces between these particles are assumed to be given, and one writes Lagrange equations for the motion of N particles. The finite system of ordinary differential equations becomes in the limit $N = \infty$ one or several *partial* differential equations. The Newtonian laws of conservation of energy and momentum are seemingly correctly formulated for the limiting case of the continuum. There appears at once, however, at least possible objection to the unrestricted validity of this formulation. For the very fact that the limiting equations imply tacitly the continuity and differentiability of the functions describing the motion of the continuum seems to impose various *constraints* on the possible motions of the approximating finite systems. Indeed, at any stage of the limiting process, it is quite conceivable for two neighbouring particles to be moving in opposite directions with a relative velocity which does not need to tend to zero as N becomes infinite; whereas the continuity imposed on the solution of the limiting continuum excludes such a situation. There are, therefore, constraints on the class of possible motions which are not explicitly recognized. This means that a viscosity or other type of constraints must be introduced initially, singling out "smooth" motions from the totality of all possible ones. In some cases, therefore, the usual differential equations of hydrodynamics may constitute a misleading description of the physical process.

Splash effect was observed numerically in 2D linear and 1D nonlinear case (A. M. Filimonov, private communication). By the way, due to this effect many papers "justifying" usual continualization can be treated as naive. For example, in [75] the continuous limit was derived based on the hypothesis that the microscopic displacements are equal to the macroscopic ones. In [75] authors supposed that the displacements of the particles which are
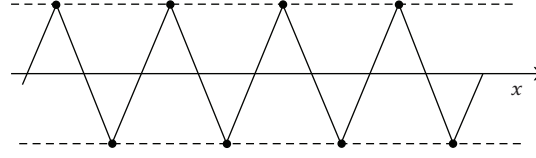
Figure 3: Saw-tooth vibrations of a mass chain.

connected by elastic springs are small in the following sense: $|y_i - y_j| \leq c\varepsilon$. This inequality can be justified only for lower part of spectrum for natural oscillations.

5. Anticontinuum Limit

A classical continuous approximation allows for relatively good description of a low part of the vibration spectrum of a finite chain of masses. In what follows we study now another limiting case, so-called anticontinuum limit, that is, completely uncoupled limit for lattice (Figure 3).

In this case one has $\sigma_k = (-1)^k \Omega$, and equation for Ω reads

$$m\Omega_{tt} + 4c\Omega = 0. \tag{5.1}$$

This is so-called anticontinuum limit.

In the case of vibrations close to the saw-tooth one, the short-wave approximation is applied "envelope continualization" [76–79] (Figure 4). Namely, first we use staggered transformation

$$\sigma_k = (-1)^k \Omega_k, \tag{5.2}$$

and then (2.3) and (2.15) are reduced to the following BVP:

$$m\Omega_{ktt} + c(4\Omega_k + \Omega_{k-1} - 2\Omega_k + \Omega_{k+1}) = 0, \tag{5.3}$$

$$\Omega_0 = 1, \qquad \Omega_{n+1} = 0, \tag{5.4}$$

$$\Omega_k = \Omega_{kt} = 0 \quad \text{for } t = 0, \ k = 1, 2, \ldots, n. \tag{5.5}$$

Then, the following relations are applied:

$$\Omega_{k-1} - 2\Omega_k + \Omega_{k+1} = -4\sin^2\left(-\frac{ih}{2}\frac{\partial}{\partial x}\right)\Omega = \left(h^2\frac{\partial^2}{\partial x^2} + \frac{h^4}{12}\frac{\partial^4}{\partial x^4} + \frac{h^6}{360}\frac{\partial^6}{\partial x^6} + \cdots\right)\Omega, \tag{5.6}$$

$$k = 0, 1, 2, \ldots, n, \quad n+1.$$
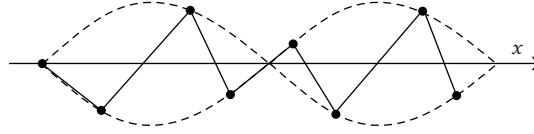
Figure 4: Envelope continualization.

Using (5.6), (5.3), and considering h^2 as a small parameter one gets (we take zeroth and first-order approximations only)

$$m\Omega_{tt} + 4c\Omega + ch^2\Omega_{xx} = 0. \tag{5.7}$$

Appropriate boundary and initial conditions for (5.7) follow

$$\begin{aligned} \Omega &= 1 \quad \text{for } x = 0, \ \Omega = 0 \quad \text{for } x = l, \\ \Omega &= \Omega_t = 0 \qquad\qquad\quad \text{for } t = 0. \end{aligned} \tag{5.8}$$

Observe that practically the whole frequency interval of discrete model is well approximated for two limiting cases, that is, for the case of the chain and for the case of the envelope.

6. Improved Continuous Approximations

In what follows we are going to construct improved continuous approximations. Modeling of such systems (nonlocal theories of elasticity) requires integral or gradient formulation. The integral formulation may be reduced to a gradient form by truncating the series expansion of the nonlocality kernel in the reciprocal space [80]. In what follows we apply the gradient formulation approach.

If, in the series (3.9), we keep three first terms, the following model is obtained:

$$m\frac{\partial^2 \sigma}{\partial t^2} = ch^2\left(\frac{\partial^2}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4}{\partial x^4} + \frac{h^4}{360}\frac{\partial^6}{\partial x^6}\right)\sigma. \tag{6.1}$$

However, a nontrivial problem regarding boundary conditions for (6.1) appears [81, 82]. The conditions mentioned can be defined only assuming the chain dynamics behavior for $k = -1, -2, -3$; $k = N + 2, N + 3, N + 4$. In other words, a boundary point is replaced by a boundary domain [44, 45]. In particular, in the case of periodic spatial extension "simple support" one gets

$$\sigma = \sigma_{xx} = \sigma_{xxxx} = 0 \quad \text{for } x = 0, l. \tag{6.2}$$

Assuming $\sigma_k(t) = 0$ for $k = -1, -2, -3$; $k = N+2, N+3, N+4$, instead of the boundary conditions (6.2) we have "clamping"

$$\sigma = \sigma_x = \sigma_{xxx} = 0 \quad \text{for } x = 0, l. \tag{6.3}$$

Comparison of nth frequency of a continuous system (6.1), (6.2) with that of a discrete system exhibits essential accuracy improvement (applying coefficient 2.1 instead of 2 in an exact solution yields an error of ~5%). Observe that an estimation of the accuracy of continuous approximation on the basis of comparison of discrete and continuous systems frequencies rather simple but yielding reliable results.

In a general case, keeping in series (3.9) N terms, one gets equations of the so-called intermediate continuous models [70]

$$m\frac{\partial^2 \sigma}{\partial t^2} = 2c \sum_{k=1}^{N} \frac{h^{2k}}{(2k)!} \frac{\partial^{2k} \sigma}{\partial x^{2k}}. \tag{6.4}$$

Boundary conditions for (6.4) have the following form:

$$\frac{\partial^{2k} \sigma}{\partial x^{2k}} = 0 \quad \text{for } x = 0, l, \ k = 0, 1, \ldots, N-1 \tag{6.5}$$

or

$$\sigma = 0, \qquad \frac{\partial^{2k-1} \sigma}{\partial x^{2k-1}} = 0 \quad \text{for } x = 0, l, \ k = 1, \ldots, N-1. \tag{6.6}$$

The corresponding BVPs are correct (and also stable during numerical solution) for odd N. In this case (6.4) is of hyperbolic type [70]. Application of intermediate continuous models allows catching the mentioned splashes effect.

For $N = 2k$, $k = 0, 1, \ldots$ intermediate continuous models (6.4) are unstable.

One also can mention the momentous elasticity theories [14, 25–27, 83–87]. For example, taking into account dependence of energy of deformation from the higher gradients of displacements leads to the concept of momentous stresses [88]. Le Roux was the first who showed the importance of the higher gradients of displacements [28]. In momentous theories of elasticity the method of macrocells is also widely used [89–91]. The dynamics of the continuous media is described by the equations of displacement of the centre of mass of a macro cell and by the equations of moments of various orders. The spectrum of this media tends to the complete spectrum of a linear chain with increasing number of moments. Critical reexamination of this approach can be found in [3].

The construction of intermediate continuous models is mainly based on the development of a difference operator into Taylor series. However, very often application of Padé approximants (PAs) is more effective for approximation [52, 92] (for description of PA see Appendix A). Collins [53] proposed to construct continuous models using PA. Then this approximation was widely used by Rosenau [55–60]. More exactly, Collins and Rosenau did not use PA straightforward; this was done later [61, 62, 93]. Sometimes these continuous models are called quasicontinuum approximations.

If only two terms are left in the series (3.9), then the PA can be cast into the following form:

$$\frac{\partial^2}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4}{\partial x^4} \approx \frac{\partial^2/\partial x^2}{1 - ((h^2/12)(\partial^2/\partial x^2))}. \tag{6.7}$$

For justification of this procedure Fourier or Laplace transforms can be used.
The corresponding quasicontinuum model reads

$$m\left(1 - \frac{h^2}{12}\frac{\partial^2}{\partial x^2}\right)\sigma_{tt} - ch^2\sigma_{xx} = 0. \tag{6.8}$$

The boundary conditions for (6.8) have the form

$$\sigma = 0 \quad \text{for } x = 0, l. \tag{6.9}$$

The error regarding estimation of n-th frequency in comparison to that of a discrete chain is of ~16.5%. However, (6.8) is of lower order in comparison to (6.1).

Equation (6.8) is usually called Love equation [94] (but as Love mentioned [95] this equation was obtained earlier by Rayleigh [96, 97]). Term σ_{xxtt} can be treated as the lateral inertia. Equation (6.8) has hyperbolic type.

Kaplunov et al. [98] refer to these type theories as theories with modified inertia.

Passage to (6.8) can be treating as regularization procedure. The model governed by (6.8) is unconditionally stable and propagating waves cannot transfer energy quicker than the velocity c. However, the model governed by (6.8) predicts that short waves transfer elastic energy with almost zero speed [99].

It is worth noting that (5.7) also can be regularized using PA as follows [100–102]:

$$m\left(1 - \frac{h^2}{4}\frac{\partial^2}{\partial x^2}\right)\Omega_{tt} + c\Omega = 0. \tag{6.10}$$

Finally, having in hand both long and short waves asymptotics, one may also apply two-point PA (description of two-point PA see Appendix A). In what follows we construct two-point PA using the first term of the series (3.9) as a one of limiting cases. We suppose

$$\sin^2\left(-\frac{ih}{2}\frac{\partial}{\partial x}\right) = -\frac{1}{2}\sum_{k=1}\frac{h^{2k}}{(2k)!}\frac{\partial^{2k}}{\partial x^{2k}} = -\frac{h^2}{4}\frac{\partial^2}{\partial x^2}\left(1 + \frac{h^2}{12}\frac{\partial^2}{\partial x^2} + \frac{h^4}{360}\frac{\partial^4}{\partial x^4} + \cdots\right)$$

$$\approx \frac{-(h^2/4)(\partial^2/\partial x^2)}{1 - a^2(\partial^2/\partial x^2)}. \tag{6.11}$$

The improved continuous approximation is governed by the following equation:

$$m\left(1 - a^2 h^2 \frac{\partial^2}{\partial x^2}\right)\sigma_{tt} - ch^2 \sigma_{xx} = 0. \tag{6.12}$$

Now we require the n-th frequency of a continuous and discrete system to coincide

$$\omega_n = 2\sqrt{\frac{c}{m}}\, \sin\frac{n\pi}{2(n+1)}. \tag{6.13}$$

For large value of n one may approximately assume that

$$\alpha_n \approx 2\sqrt{c/m} \tag{6.14}$$

with the boundary conditions (6.9).

Oscillations frequencies defined by the BVP (6.12), (6.9) are

$$\alpha_k = \pi\sqrt{\frac{c}{m}}\,\frac{k}{\sqrt{(n+1)^2 + a^2 k^2}}, \quad k = 1, 2, \ldots. \tag{6.15}$$

Now, using (6.14) one gets $a^2 = 0.25 - \pi^{-2}$.

The highest error in estimation of the eigenfrequencies appears for $k = [0.5(n+1)]$ and consists of 3%.

Observe that equations similar to the (6.14) are already known. Eringen [103–106] using a correspondence between the dispersion curves of the continuous approximation and Born-von Kármán model [82, 107] obtained $a^2 = 0.1521$. This value is very close to that proposed in [108, 109] on the basis of a certain physical interpretation, where also model (6.14) is referred to as the "dynamically consistent model."

Interesting, that Mindlin and Herrmann used very similar to two-point PA idea for construction their well-known equation for longitudinal waves in rod [110, 111].

Dispersion relation (6.15) does not satisfy condition

$$\frac{d\alpha_k}{dk} = 0 \tag{6.16}$$

at the end of first Brillouin zone [112]. That is why in [105, 113] so-called bi-Helmholtz type equation was proposed. In [113] this equation is

$$m\left(1 - a^2 h^2 \frac{\partial^2}{\partial x^2} + a_1^2 h^4 \frac{\partial^4}{\partial x^4}\right)\sigma_{tt} - ch^2 \sigma_{xx} = 0, \tag{6.17}$$

where $a_1 = 1/\pi$.

Boundary conditions associated with this equation are

$$\sigma = \sigma_{xx} = 0 \quad \text{for } x = 0, l \tag{6.18}$$

or

$$\sigma = \sigma_x = 0 \quad \text{for } x = 0, l. \tag{6.19}$$

7. Forced Oscillations

In order to study forced vibrations we begin with a classical continuous approximation. A solution to (3.1) is sought in the following form:

$$\sigma = 1 - \frac{x}{l} + u(x, t), \tag{7.1}$$

and the function $u(x, t)$ is defined by the following BVP:

$$m\frac{\partial^2 u}{\partial t^2} = ch^2 \frac{\partial^2 u}{\partial x^2},$$

$$u(0, t) = u(l, t) = 0, \tag{7.2}$$

$$u(x, 0) = -1 + \frac{x}{l}, \qquad u_t(x, 0) = 0.$$

Solution to the BVP (7.2) is obtained via Fourier method

$$\sigma = 1 - \frac{x}{l} - \frac{2}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}\sin\left(\frac{k\pi x}{l}\right)\cos(\alpha_k t), \tag{7.3}$$

where α_k defined by (3.10).

For the approximation the motion of the chain, one must keep in an infinite sum only n first harmonics

$$\sigma = 1 - \frac{x}{l} - \frac{2}{\pi}\sum_{k=1}^{n}\frac{1}{k}\sin\left(\frac{k\pi x}{l}\right)\cos(\alpha_k t). \tag{7.4}$$

Observe that solutions (2.16) and (7.4) differ regarding frequencies α_k and ω_k (relations (2.8) and (3.10), resp.). In addition, the coefficients in series (2.16) and (7.4) differ from each other, that is, projections onto normal vibrations of discrete and continuous systems, $(1/(n+1))ctg(\pi k/2(n+1))$ and $2/\pi k$, respectively, are strongly different for $k \gg 1$. This is because during projection into normal vibrations for the discrete system one must use

summation due k from 0 to n, whereas for the continuous system—integration due x from 0 to l. The problem occurring so far can be overcome using the Euler-McLaurin formulas [114]

$$\sum_{k=0}^{n+1} f(k) = \int_0^{n+1} f(x)dx + \frac{1}{2}[f(0) + f(n+1)] + \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j+1} B_j \left[\frac{d^j f(n+1)}{dx^j} - \frac{d^j f(0)}{dx^j} \right], \quad (7.5)$$

where B_i are the Bernoulli numbers, having the following values: $B_0 = 1, B_1 = -1/2, B_2 = 1/6, B_3 = 0, \ldots$.

In addition, the following formulas hold [115]:

$$B_n = -\frac{1}{n+1} \sum_{k=1}^{n} C_{n+1}^{k+1} B_{n-k}, \quad (7.6)$$

$$\sum_{k=0}^{n+1} \sin^2 \frac{k\pi j}{n+1} = \frac{n+1}{2}, \quad (7.7)$$

$$\sum_{k=0}^{n+1} \left(1 - \frac{j}{n+1}\right) \sin \frac{k\pi j}{n+1} = \frac{1}{n+1} ctg \frac{j\pi}{2(n+1)}. \quad (7.8)$$

The corresponding integrals read

$$\int_0^{n+1} \sin^2 \frac{\pi jx}{n+1} dx = \frac{n+1}{2}, \quad (7.9)$$

$$\int_0^{n+1} \left(1 - \frac{x}{l}\right) \sin \frac{\pi jx}{n+1} dx = \frac{2}{\pi j}. \quad (7.10)$$

Observe that the values of sum (7.7) and integral (7.9) coincide. Using the Euler-McLaurin formula one obtains

$$\sum_{k=0}^{n+1} \left(1 - \frac{j}{n+1}\right) \sin \frac{k\pi j}{n+1} = \int_0^{n+1} \left(1 - \frac{x}{l}\right) \sin \frac{\pi jx}{n+1} dx + \frac{1}{2}[\sin 0 + \sin(j\pi)] - \frac{j\pi}{6(n+1)} \cos 0 + \cdots$$

$$= \frac{2}{\pi j} \left[1 - \frac{\pi^2 j^2}{12(n+1)^2}\right].$$

$$\quad (7.11)$$

Owing to (7.11), one can construct a simple expression relatively good approximating sum (7.8) for arbitrary j values from $j = 1$ to $j = n$. For this purpose we change second term of the right-hand side of (7.11) in the following way:

$$\sum_{k=0}^{n+1} \left(1 - \frac{j}{n+1}\right) \sin \frac{k\pi j}{n+1} \approx \frac{2}{\pi j} \left[1 - \frac{j^2}{(n+1)^2}\right]. \quad (7.12)$$
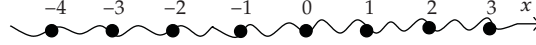
Figure 5: Infinite chain of masses.

8. Wave Processes in Discrete and Continuous 1D Media

Let us study wave propagation in the infinite 1D medium (Figure 5). Assuming that when time instant $t = 0$, a force P acts on a mass with number 0 in direction of the axis x. Governing equations can be written in the following way [116]:

$$m\frac{d^2y_j}{dt^2} = c(y_{j+1} - 2y_j + y_{j-1}), \quad j \neq 0,$$

$$m\frac{d^2y_0}{dt^2} = c(y_1 - 2y_0 + y_{-1}) + P. \tag{8.1}$$

We reduce (8.1) to the dimensionless form

$$\frac{\partial^2 Y_j}{\partial \tau^2} = Y_{j+1} - 2Y_j + Y_{j-1},$$

$$\frac{\partial^2 Y_0}{\partial \tau^2} = Y_1 - 2Y_0 + Y_{-1} + P_1. \tag{8.2}$$

Here $\tau = t\sqrt{c/m}$; $Y_j = y_j/h$; $P_1 = P/(ch)$.

Due to the linearity of the problem one can suppose $P_1 = 1$.

Let us apply Fourier transform to (8.2) [116]. Namely, multiplying left-and right-hand sides of (8.2) by $\exp(iqj), j = 0, \pm 1, \pm 2, \ldots$ and adding them one obtains

$$\frac{d^2U}{d\tau^2} + 4\sin^2\left(\frac{q}{2}\right)U = 1, \tag{8.3}$$

where $U = \sum_{j=-\infty}^{\infty} Y_j \exp(iqj)$.

Observe that the term $4\sin^2(q/2)$ can be developed in the vicinity of either $q = 0$ (classical continuous approximation)

$$4\sin^2\left(\frac{q}{2}\right) = q^2 + \cdots \tag{8.4}$$

or $q = \pi$ (envelope continualization)

$$4\sin^2\left(\frac{q}{2}\right) = 4 + \cdots . \tag{8.5}$$

Solving (8.3) and applying inverse Fourier transform one obtains wave velocity propagation in discrete media:

$$\frac{dY_j(\tau)}{d\tau} = \frac{A_1(j,\tau)}{\pi} = \frac{1}{\pi}\int_0^{\pi/2}\frac{\sin(2\tau\sin s)}{\sin s}\cos(2js)\,ds. \tag{8.6}$$

The obtained solution is exact one and will be further used for the error estimation of improved continuous models.

Applying the classical continuous approximation instead of system (8.2), the following wave equation with the Dirac delta-function in right hand-side is obtained:

$$\frac{\partial^2 Y}{\partial\tau^2} - \frac{\partial^2 Y}{\partial z^2} = \delta(z). \tag{8.7}$$

Here $z = x/h$.

The following relation between discrete and continuous systems holds:

$$Y_j(\tau) = Y(j,\tau), \quad j = 0,\pm 1,\pm 2,\ldots. \tag{8.8}$$

Now, applying the Fourier transform in x and the Laplace transform in τ, one obtains

$$\frac{\partial Y}{\partial\tau} = 0.5H(\tau - |z|). \tag{8.9}$$

So, a wave propagation in a discrete media strongly differs from this phenomenon in the classical continuous media. In what follows we give solutions on a basis of the theories of elasticity described so far. The intermediate continuous model is as follows:

$$\frac{\partial^2 Y}{\partial\tau^2} - \left(\frac{\partial^2}{\partial z^2} + \frac{1}{12}\frac{\partial^4}{\partial z^4} + \frac{1}{360}\frac{\partial^6}{\partial z^6}\right)Y = \frac{\sin(\pi z)}{\pi z}. \tag{8.10}$$

Let us explain the occurrence of term $\sin(\pi z)/(\pi z)$ in the right hand of (8.10) instead of Dirac delta-function (for detail description see Appendix C).

Applying both Fourier transform in z and Laplace transform in τ, the following equation governing wave velocity propagation is obtained:

$$\frac{\partial Y(z,\tau)}{\partial\tau} = \frac{A_2(z,\tau)}{\pi} = \frac{1}{\pi}\int_0^{\pi/2}\frac{\sin\left(2\tau s\sqrt{1-(1/3)s^2+(2/45)s^4}\right)}{s\sqrt{1-(1/3)s^2+(2/45)s^4}}\cos(2sz)\,ds. \tag{8.11}$$

The quasicontinuum model is as follows:

$$\left(1 - \frac{1}{12}\frac{\partial^2}{\partial z^2}\right)\frac{\partial^2 Y}{\partial\tau^2} - \frac{\partial^2 Y}{\partial z^2} = \left(1 - \frac{1}{12}\frac{\partial^2}{\partial z^2}\right)\frac{\sin(\pi z)}{\pi z}. \tag{8.12}$$

The associated wave velocity propagation follows:

$$\frac{\partial Y(z,\tau)}{\partial \tau} = \frac{A_3(z,\tau)}{\pi} = \frac{1}{\pi} \int_0^{\pi/2} \frac{\sqrt{1 + (1/3)s^2} \sin\left(2\tau s/\sqrt{1 + (1/3)s^2}\right)}{s} \cos(2sz)ds. \qquad (8.13)$$

Model (6.14) for our case has the following form:

$$\left(1 - a^2 \frac{\partial^2}{\partial z^2}\right)\frac{\partial^2 Y}{\partial \tau^2} - \frac{\partial^2 Y}{\partial z^2} = \left(1 - a^2 \frac{\partial^2}{\partial z^2}\right)\frac{\sin(\pi z)}{\pi z}. \qquad (8.14)$$

In this case the associated wave velocity propagation reads

$$\frac{\partial Y(z,\tau)}{\partial \tau} = \frac{A_4(z,\tau)}{\pi} = \frac{1}{\pi} \int_0^{\pi/2} \frac{\sqrt{1 + 4a^2 s^2} \sin\left(2\tau s/\sqrt{1 + 4a^2 s^2}\right)}{s} \cos(2sz)dz. \qquad (8.15)$$

Since analytical comparison of the obtained wave velocity propagation is difficult, we apply numerical procedures.

As it has been already mentioned, exact solution of the joined discrete problem governed by relation $A_1(j,\tau)$ will serve us as the standard solution used for estimation of other solutions. Results of computations are shown on Figures 6(a)–6(c) and Figures 7(a)–7(c). One can conclude that above described improved continuous models give possibility qualitatively taking into account effect of media discreteness: (i) occurrence of oscillations, being gradually damped at $x = $ const; (ii) infinite velocity of perturbations propagation (owing to assumption of instantaneous masses interaction during their draw near); (iii) occurrence of the quasi-front domain, where the stresses increase relatively fast, but without jump, as on the wave front set; velocities and deformations exponentially decay while the quasi-front $\tau - |z|$ increases finally becoming negligibly small.

Equation (8.10) has sixth order in spatial variable, (8.12) and (8.14) have second order in spatial variable, so, they are simpler for practice applications. Equations (8.12), and (8.14) give for case of wave motion qualitatively the similar results.

9. 2D Lattice

In order to analyse the 2D case we use 9-cell square lattice (Figure 8) [87, 117]. The central particle is supposed to interact with eight neighbours in the lattice. The mass centres of four of them are on horizontal and vertical lines, while the mass centres of the other four neighbouring particles lie along diagonals. Interactions between the neighbouring particles are modelled by elastic springs of three types. Horizontal and vertical springs with rigidity c_1

(a) Wave propagation for $z = 2.0$

(b) Wave propagation for $z = 10.0$

(c) Wave propagation for $z = 100.0$
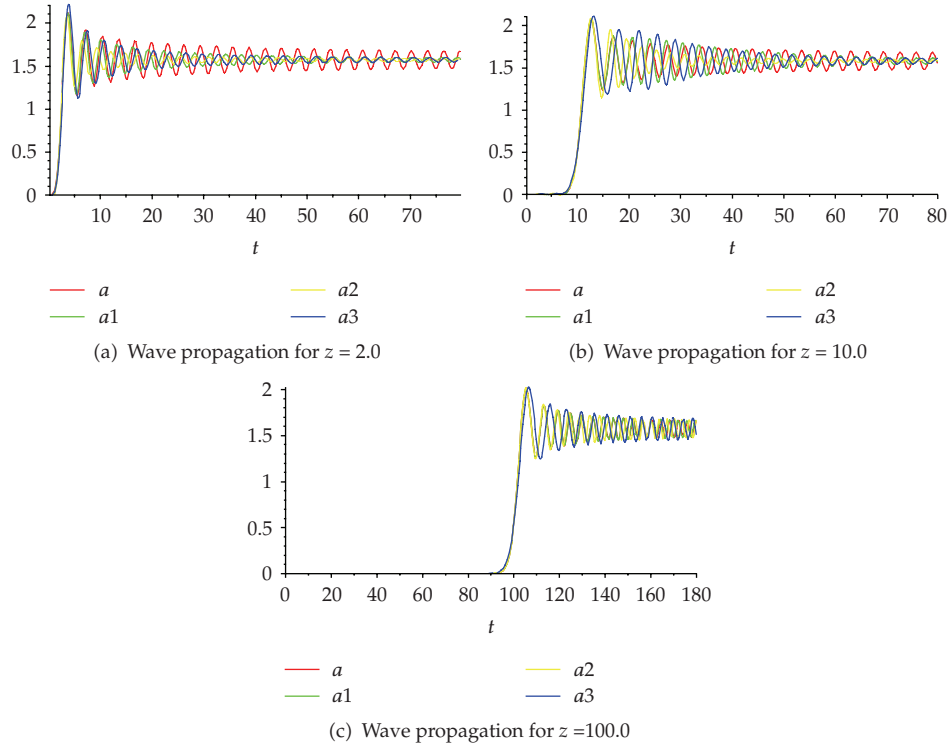
Figure 6

define interaction forces of extension/compression of the material. The diagonal longitudinal springs have rigidity c_0, and shear axial spring - c_2. Governing equations of motion are [87]

$$m\ddot{u}_{m,n} = c_1(u_{m,n-1} - 2u_{m,n} + u_{m,n+1}) + c_2(u_{m-1,n} - 2u_{m,n} + u_{m+1,n}) + 0.5c_0$$

$$\times (u_{m-1,n-1} + u_{m+1,n-1} + u_{m-1,n+1} + u_{m+1,n+1} + v_{m-1,n-1} - v_{m+1,n-1} - v_{m-1,n+1}$$

$$+ v_{m+1,n+1} - 4u_{m,n}),$$

$$m\ddot{v}_{m,n} = c_1(v_{m,n-1} - 2v_{m,n} + v_{m,n+1}) + c_2(v_{m-1,n} - 2v_{m,n} + v_{m+1,n}) + 0.5c_0$$

$$\times (v_{m-1,n-1} + v_{m+1,n-1} + v_{m-1,n+1} + v_{m+1,n+1} + u_{m-1,n-1} - u_{m+1,n-1} - u_{m-1,n+1}$$

$$+ u_{m+1,n+1} - 4v_{m,n}).$$

$$(9.1)$$

Here $u_{i\cdot j}, v_{i\cdot j}$ is the displacement vector for a particle situated at point (x_i, y_j), $x_i = ih, y_j = jh$.

The standard continualization procedure for (9.1) involves introducing a continuous displacement field $u(x, y), v(x, y)$ such that $u(x_m, y_n) = u_{m,n}; v(x_m, y_n) = v_{m,n}$, and

(a) Wave propagation for $\tau = 2.0$

(b) Wave propagation for $\tau = 10.0$

(c) Wave propagation for $\tau = 100.0$
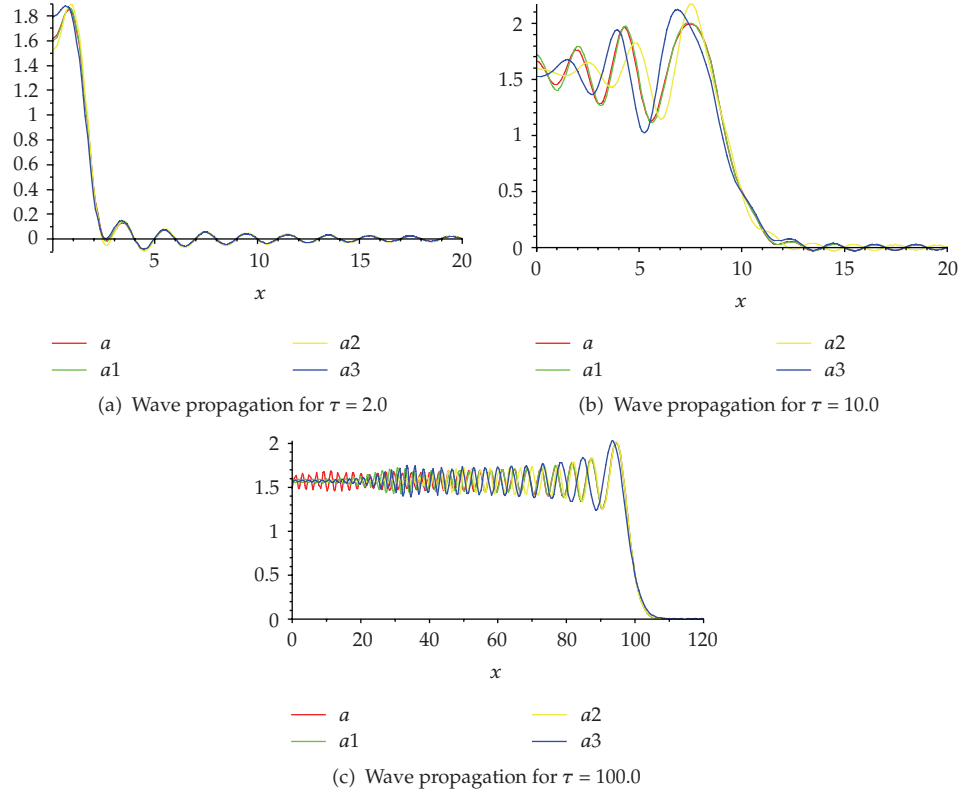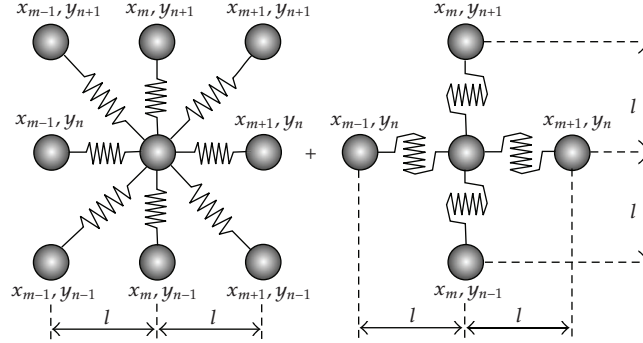
Figure 7

Figure 8: 2D lattice.

expanding $u_{m\pm1,n\pm1}, v_{m\pm1,n\pm1}$ into Taylor series around $u_{m,n}, v_{m,n}$. Second-order continuous theory in respect to the small parameter h is [87]

$$
m\frac{\partial^2 u}{\partial t^2} = c_1 h^2 \frac{\partial^2 u}{\partial x^2} + c_2 h^2 \frac{\partial^2 u}{\partial y^2} + 2c_0 h^2 \frac{\partial^2 v}{\partial x \partial y},
$$
$$
m\frac{\partial^2 v}{\partial t^2} = c_1 h^2 \frac{\partial^2 v}{\partial x^2} + c_2 h^2 \frac{\partial^2 v}{\partial y^2} + 2c_0 h^2 \frac{\partial^2 u}{\partial x \partial y}.
$$

(9.2)

Naturally, one can construct equations of higher order, but, as it is shown in [87], generally it is not possible to create asymptotic theories that do not possess extraneous solutions in the nonscalar context.

Using staggered transformations

$$u_k = (-1)^k u, \quad v_k = (-1)^k v, \tag{9.3}$$

one obtains in anticontinuum limit.

$$m\frac{\partial^2 u}{\partial t^2} = 4\left(c_1\frac{\partial^2 u}{\partial x^2} + c_2\frac{\partial^2 u}{\partial y^2}\right),$$

$$m\frac{\partial^2 v}{\partial t^2} = 4\left(c_1\frac{\partial^2 v}{\partial x^2} + c_2\frac{\partial^2 v}{\partial y^2}\right). \tag{9.4}$$

As in 1D case (see (5.1)) equations (9.4) do not contain parameter h^2. For this type of 2D lattice equations in anticontinuum limit are decoupled.

The existence of the continuous approximations (9.2) and (9.4) gives a possibility to construct the composite equations, which is uniformly suitable in the whole interval of the frequencies and the oscillation forms of the 2D lattice of masses. Let us emphasize, that the composite equations, due to Van Dyke [118], can be obtained as a result of synthesis of the limiting cases. The principal idea of the method of the composite equations can be formulated in the following way [118, page 195].

(i) Identify the terms in the differential equations, the neglection of which in the straightforward approximation is responsible for the nonuniformity.

(ii) Approximate those terms insofar as possible while retaining their essential character in the region of nonuniformity.

In our case the composite equations will be constructed in order to overlap (approximately) with (9.2) for long wave solution and with (9.4) for short wave solution. As a result of the described procedure one gets

$$m\left(1 - a^2h^2\frac{\partial^2}{\partial x^2} - a^2h^2\frac{\partial^2}{\partial y^2}\right)\frac{\partial^2 u}{\partial t^2}$$

$$= c_1h^2\frac{\partial^2 u}{\partial x^2} + c_2h^2\frac{\partial^2 u}{\partial y^2} - (c_1+c_2)h^4\gamma^2\frac{\partial^4 u}{\partial x^2\partial y^2} + 2c_0h^2\left[1 + \frac{1}{4(c_1+c_2)}\frac{\partial^2}{\partial t^2}\right]\frac{\partial^2 v}{\partial x\partial y},$$

$$m\left(1 - a^2h^2\frac{\partial^2}{\partial x^2} - a^2h^2\frac{\partial^2}{\partial y^2}\right)\frac{\partial^2 v}{\partial t^2}$$

$$= c_1h^2\frac{\partial^2 v}{\partial x^2} + c_2h^2\frac{\partial^2 v}{\partial y^2} - (c_1+c_2)h^4\gamma^2\frac{\partial^4 v}{\partial x^2\partial y^2} + 2c_0h^2\left[1 + \frac{1}{4(c_1+c_2)}\frac{\partial^2}{\partial t^2}\right]\frac{\partial^2 u}{\partial x\partial y}. \tag{9.5}$$

Here $a^2 = 0.25 - \pi^{-2}$; $\gamma^2 = (4 - \pi^2 + 8\pi^2\alpha^2)/(4\pi^2)$.

For 1D case from (9.5) one obtains (6.14) for u and the same equation for v. For small variability in spatial and time variables equations (9.5) can be approximated by (9.2), for large variability in spatial and time variables equations (9.5) can be approximated by (9.4).

10. Nonlinear Case

Fundamental methods of analysis of linear lattices are those based on either discrete or continuous Fourier transformations [44, 45, 119] and the integral Fourier operators method proposed by Maslov [67]. In a nonlinear case the similar methods, in general, are not applicable. However, there exists class of piece-wise acting forces-springs extensions relations, opening the doors for analytical tools application [120].

Besides, one may achieve even a solution to a nonlinear problem using the following observation [61, 62, 119, 121]. Let nonlinear springs in 1D problem have potential energy $U(y)$, where y is a difference between the neighborhood masses regarding their equilibrium positions. In this case the stretching force is $-\partial U(y)/\partial y$. The chain dynamics is governed by the following infinite system of ODEs:

$$\frac{\partial^2 r_n}{\partial t^2} = \frac{\partial U(r_{n+1})}{\partial r_{n+1}} - 2\frac{\partial U(r_n)}{\partial r_n} + \frac{\partial U(r_{n-1})}{\partial r_{n-1}}, \tag{10.1}$$

where $r_n = y_{n+1} - y_n$.

Assume that solution to the ODEs is a traveling wave, $r_n = R(z), z = n - vt$. Then (10.1) is cast into the following form:

$$v^2 R''(z) = U'(R(z-1)) - 2U'(R(z)) + U'(R(z+1)), \tag{10.2}$$

where $(\cdots)' = d(\cdots)/dr$.

Introducing notation $\Phi(z) = U'(R(z))$ and applying the Fourier transformation to (10.2) one gets

$$\overline{R}(q) = \int_{-\infty}^{\infty} R(z)\exp(-iqz)dz, \qquad \overline{\Phi}(q) = \int_{-\infty}^{\infty} \Phi(z)\exp(-iqz)dz. \tag{10.3}$$

In the conjugated space one obtains

$$v^2\overline{R}(q) = f(q)\overline{\Phi}(q), \tag{10.4}$$

where $f(q) = 4(\sin^2(q/2)/q^2)$.

Term $f(q)$ can be approximated by PA or two-point PA in the following way:

$$f(q) = \frac{1}{1 + q^2/12} \quad \text{or} \quad f(q) = \frac{1}{1 + a^2q^2}, \tag{10.5}$$

where $a^2 = 0.25 - \pi^{-2}$.

In result one obtains equation of continuous approximation

$$v^2\left(1 - \frac{1}{12}\frac{d^2}{dz^2}\right)R(z) = U'(R(z)) \tag{10.6}$$

or

$$v^2\left(1 - a^2\frac{d^2}{dz^2}\right)R(z) = U'(R(z)). \tag{10.7}$$

We show also a simultaneous application of PA matched with a perturbation procedure using an example of the Toda lattice [122]

$$m\frac{d^2y_n}{dt^2} = a\left[\exp(b(y_{n-1} - y_n)) - \exp(b(y_n - y_{n+1}))\right]. \tag{10.8}$$

In continuum limit one obtains (for details see **Appendix B**)

$$\frac{\partial y_1}{\partial t_1} + \frac{\partial}{\partial x}\left(1 + \frac{1}{24}\frac{\partial^2}{\partial x^2}\right)y_1 - y_1\frac{\partial y_1}{\partial x} = 0. \tag{10.9}$$

where: $t_1 = \sqrt{(ab/m)}t$; $y_1 = (b/2)y$.

One can obtain from (10.9) the Korteweg-de Vries (KdV) equation [122]. Here, we illustrate how one may get regularized long-wave equation [123–125]. Using PA one obtains

$$1 + \frac{1}{24}\frac{\partial^2}{\partial x^2} \sim \left(1 - \frac{1}{24}\frac{\partial^2}{\partial x^2}\right)^{-1}. \tag{10.10}$$

Then, (10.9) yields (up to the highest terms)

$$\left(1 - \frac{1}{24}\frac{\partial^2}{\partial x^2}\right)\frac{\partial y_1}{\partial t_1} + \frac{\partial y_1}{\partial x} - y_1\frac{\partial y_1}{\partial x} = 0. \tag{10.11}$$

On the other hand, there are nonlinear lattices with a special type nonlinearities allowing to achieve exact solutions as soliton and soliton-like solutions (Toda, Ablowitz and Ladik, *Langmuir,* Calogero, and so forth, [119, 126–130]) in the case of infinite lattices or in the case of occurrence of physically unrealistic boundary conditions. In many cases nonintegrable systems like a discrete variant of sine-Gordon equation possess soliton-like solution.

Lattice discreteness allows the existence of new types of localized structures that would not exist in the continuum limit. Nonlinearity transfers energy to higher wavenumbers but it can be suppressed and balanced by the bound of the spectrum of discrete systems. Such balance can generate highly localized structures, that is, the discrete breathers [100–102, 131]. Improved continuous models give possibility to construct such type of localized solutions [100–102].

On the other hand, it is more appropriate changing a system of Fermi-Pasta-Ulam not by KdV equation, but by Toda lattice. In the latter case one gets an asymptotic regarding nonlinear parameter instead of applying direct substitution of a discrete system by a continuous one. Occurrence of exact solutions of a discrete system allows applying the following approach; fast changeable solution part is constructed using a discrete model, whereas a continuous approximation is applied for slow components. The latter observation is very well exhibited by Maslov and Omel'yanov [130], who analysed Toda lattice with variable coefficients

$$m_k \frac{d^2 y_k}{dt^2} = -a_k \left[\exp b_k (y_k - y_{k-1}) - 1 \right] + a_{k+1} \left[\exp b_{k+1} (y_{k+1} - y_k) - 1 \right],$$

(10.12)

$$k = 0, \pm 1, \pm 2, \ldots.$$

They construct soliton solutions in the following way: rapidly changing part of soliton is constructed using Toda lattice with constant coefficients, and for slowly part of solution continuous approximation is used. Then these solutions are matched.

11. Some Generalization

Let us show that PA for constructing of improved continuous models can be used iteratively. For three terms in expansion (3.9) one has PA

$$\frac{(\partial^2 / \partial x^2)(1 + (h^2/20)\partial^2/\partial x^2)}{1 - ((h^2/30)(\partial^2/\partial x^2))},$$

(11.1)

and continuous model

$$m \left(1 - \frac{h^2}{30} \frac{\partial^2}{\partial x^2} \right) \sigma_{tt} - ch^2 \left(1 + \frac{h^2}{20} \frac{\partial^2}{\partial x^2} \right) \sigma_{xx} = 0$$

(11.2)

or

$$m \frac{\partial^2 \sigma}{\partial t^2} - ch^2 \left(1 + \frac{m}{30c} \frac{\partial^2}{\partial t^2} + \frac{h^2}{20} \frac{\partial^2}{\partial x^2} \right) \sigma_{xx} = 0.$$

(11.3)

Now we use PA as follows:

$$1 + \frac{m}{30c} \frac{\partial^2}{\partial t^2} + \frac{h^2}{20} \frac{\partial^2}{\partial x^2} = \frac{\left(1 + (m/30c)(\partial^2/\partial t^2) \right)^2}{\left(1 + (m/30c)(\partial^2/\partial t^2) - (h^2/20)(\partial^2/\partial x^2) \right)}$$

(11.4)

and obtain new improved continuous model

$$m \left(1 - \frac{7h^2}{60} \frac{\partial^2}{\partial x^2} \right) \frac{\partial^2 \sigma}{\partial t^2} + \frac{m^2}{30c} \left[1 - \frac{h^2}{30} \frac{\partial^2}{\partial x^2} \right] \frac{\partial^4 \sigma}{\partial t^4} - ch^2 \frac{\partial^2 \sigma}{\partial x^2} = 0.$$

(11.5)

Now let us consider the continuous models of the chain with two different particles in primitive cell, that is, the chain with alternating masses

$$my_{jtt}(t) - c\left[y_{j-1}(t) - 2y_j(t) + y_{j+1}(t)\right] = (m - M)y_{2jtt}(t); \quad j = 1, 2, 3, \ldots. \tag{11.6}$$

Continuous approximation for (11.6) is (rod with localized masses)

$$my_{tt}(x,t) - ch^2 y_{xx}(x,t) = (m - M)h \sum_{k=-\infty}^{\infty} \delta(x - 2hk)y_{tt}(x,t). \tag{11.7}$$

Improved continuous approximation for (11.6) can be written as follows:

$$
\begin{aligned}
&m\left(1 - \frac{h^2}{12}\frac{\partial^2}{\partial x^2}\right)y_{tt}(x,t) - ch^2 y_{xx}(x,t) \\
&\quad = (m - M)h\left(1 - \frac{h^2}{12}\frac{\partial^2}{\partial x^2}\right)\sum_{k=-\infty}^{\infty} \frac{\sin \pi(x - 2hk)}{\pi(x - 2hk)}y_{tt}(x,t)
\end{aligned}
\tag{11.8}
$$

or

$$
\begin{aligned}
&m\left(1 - a^2\frac{\partial^2}{\partial x^2}\right)y_{tt}(x,t) - ch^2 y_{xx}(x,t) \\
&\quad = (m - M)h\left(1 - a^2\frac{\partial^2}{\partial x^2}\right)\sum_{k=-\infty}^{\infty} \frac{\sin \pi(x - 2hk)}{\pi(x - 2hk)}y_{tt}(x,t).
\end{aligned}
\tag{11.9}
$$

Homogenization and other asymptotic approaches can be use for this continuous system [132, Section 3]).

12. Modified Navier-Stokes Equations

The Navier-Stokes equations have a long and glorious history but remain extremely challenging, for example, the issue of existence of physically reasonable solutions of these equations in 3D case was chosen as one of the seven millennium "million dollar" prize problems of the Clay Mathematical Institute [133]. Many famous mathematicians (Smale et al. [134], Yudovich [135]) also wrote about this problem as one of the most important problems of the 21st century. The 3D problem remains open until today [136], although in the 1950s Ladyzhenskaya obtained the key result of global unique solvability of the initial boundary value problem for the 2D Navier-Stokes Equations Ladyzhenskaya believed that in the 3D case the Navier-Stokes Equations, even with very smooth initial data, do not provide

uniqueness of their solutions on an arbitrary time interval. She introduced the modified Navier-Stokes equations [133, 137–139] (see also [140]):

$$u_t - \frac{\partial}{\partial x_k}\left[\left(\nu_0 + \nu_1 u_x^2\right)u_{x_k}\right] + u_k u_{x_k} = -\mathrm{grad}\, p + f(x,t),$$

$$\mathrm{div}\, u = 0, \quad k = 1,2,3$$

(12.1)

or

$$u_t + \mathrm{rot}\left[\left(\nu_0 + \nu_1 \mathrm{rot}^2 u\right)\mathrm{rot}\, u\right] + u_k u_{x_k} = -\mathrm{grad}\, p + f(x,t),$$

$$\mathrm{div}\, u = 0, \quad k = 1,2,3.$$

(12.2)

Equations (12.1), (12.2) differ from the classical ones only in large-velocity gradients and coincide with them for $\nu_1 = 0$. For (12.1) and (12.2) Ladyzhenskaya proved global unique solvability under reasonably wide assumptions.

As it is mentioned in [59], "From the kinetic point of view these equations represent the long-wavelength limit. However, it often happens that the dynamics predicted by the hydrodynamics equations involves relatively short-wavelength scales. This, formally at least, contradicts the conditions under which the macroscopic system was derived." Note that the modified Navier-Stokes equations (12.1), (12.2) allow for taking into account higher order harmonics influence via artificially separated terms.

More investigations in direction of improvement of the Navier-Stokes equations are carried out in references [133, 137–139].

It should be emphasized a role of mathematical approaches in proving various physical theories [141]. Although nowadays a physical experiment plays a crucial role in verification of many novel theories, but it is difficult sometimes to realize a sure and proper experimental verification and validation of the developed theories. In those cases it seems that the proper support of mathematical rigorous approaches may serve as a tool of "experimental" validation of the introduced theories. In other words the rigorous mathematical approaches validating the physical theories can be treated as experimental ones for the so called "theoretical mathematics."

Therefore, the modified Navier-Stokes equations are more suitable for fluid dynamics description for large Re from a point of view of theoretical mathematics. For small Re they are more close to the Navier-Stokes equations. In the latter case one gets a unique global solutions to the initial-boundary value problems as well as the solvability of the stationary boundary value problems for arbitrary Re numbers.

Note that the modified Navier-Stokes equations (12.1), (12.2) allow for taking into account higher order harmonics influence via artificially introduced terms. It is clear that it will be very challenging in getting the modified Navier-Stokes equations starting in modeling from a molecular structure, as it has been demonstrated so far using simple examples of discrete lattice. Here we can mention paper by Rosenau, who applied regularization procedure to the Chapman-Enskog expansion [59]. This approach, however, requires further investigations.

13. Conclusion

Classical molecular dynamics simulations have become prominent as a tool for elucidating complex physical phenomena, such as solid fracture and plasticity. However, the length and time scales probed using molecular dynamics are still fairly limited. To overcome this problem it is possible to use molecular dynamics only in localized regions in which the atomic-scale dynamics is important, and a continuous simulation method (e.g., FEM) everywhere else [142]. Then the problem of coupling molecular dynamics and continuum mechanics simulations appears. In this case, one can apply a concept of bridging domain [142–144]. We think that in the latter case an application for describing bridging domain of improved continuous models can result in efficient computational time saving.

Let us compare an impact of the methods illustrated and discussed so far on the improved continuous models. Continuous models based on the PA and two-point PA can be applied in the case of 1D problems (in spite of some artificially constructed examples [51]). Intermediate continuous models and BVP obtained on the basis of the composite equations can be used for problems of any dimensions.

It will be very interesting to use for study investigation of 2D lattices 2D PA [145].

Let us finally discuss problems closely connected with brittle fracture of elastic solids [120, 146–150]. Observe that a continuous model, which does not include material structure, is not suitable since any material crack occurs on the material structure level. This is a reason why some of the essential properties of damages of the classical continuous model theory are not exhibited. For instance, in a discrete medium the propagated waves transport part of the energy from the elastic body into the crack edges. This is why application of nonlocal theories during explosion process modeling (e.g., for the composite materials) looks very promising. As it is mentioned in [151], "The continuous development of advanced materials ensures that a "one size fits all" approach will no longer serve the engineering community in terms of predicting and preventing fatigue failures and reducing their associated costs."

Appendices

A. Padé Approximants

Let us consider Padé approximants, which allow us to perform, to some extent, the most natural continuation of the power series. Let us formulate the definition [152]. Let

$$F(\varepsilon) = \sum_{i=0}^{\infty} c_i \varepsilon^i,$$

$$F_{mn}(\varepsilon) = \frac{\sum_{i=0}^{m} a_i \varepsilon^i}{\sum_{i=0}^{m} b_i \varepsilon^i},$$

(A.1)

where the coefficients a_i, b_i are determined from the following condition: the first $(m + n)$ components of the expansion of the rational function $F_{mn}(\varepsilon)$ in a McLaurin series coincide with the first $(m + n + 1)$ components of the series $F(\varepsilon)$. Then F_{mn} is called the $[m/n]$ PA. The set of F_{mn} functions for different m and n forms the Padé table.

Now we give the notion of two-point Padé approximants [152]. Let

$$
F(\varepsilon) =
\begin{cases}
\displaystyle\sum_{i=0}^{\infty} a_i \varepsilon^i & \text{when } \varepsilon \longrightarrow 0, \\[2ex]
\displaystyle\sum_{i=0}^{\infty} b_i \varepsilon^{-i} & \text{when } \varepsilon \longrightarrow \infty.
\end{cases}
\tag{A.2}
$$

The TPPA is represented by the rational function

$$
F(\varepsilon) = \frac{\sum_{k=0}^{m} a_k \varepsilon^k}{\sum_{k=0}^{n} b_k \varepsilon^k},
\tag{A.3}
$$

where $k + 1$ ($k = 0, 1, \ldots, n + m + 1$) coefficients of a McLaurin expansion, if $\varepsilon \to 0$, and $m + n + 1 - k$ coefficients of a Laurent series, if $\varepsilon \to \infty$, coincide with the corresponding coefficients of the series (A.2).

B. Continuum Limit of Toda Lattice

Here we describe construct of the continuum limit of the Toda lattice

$$
m\frac{d^2 y_n}{dt^2} = a\big[\exp\big(b(y_{n-1} - y_n)\big) - \exp\big(b(y_n - y_{n+1})\big)\big].
\tag{B.1}
$$

We follow [122]. Equation (B.1) can be rewritten as pseudodifferential one:

$$
m\frac{\partial^2 y}{\partial t^2} - 4a\left[\sinh\left(\frac{1}{2}\frac{\partial}{\partial x}\right)\right]^2 \exp(-by) = 0.
\tag{B.2}
$$

Equation (B.1) can be approximately factorized

$$
\left[\sqrt{\frac{m}{ab}}\frac{\partial}{\partial t} - 2\sinh\left(\frac{1}{2}\frac{\partial}{\partial x}\right) + \frac{b}{2}y\frac{\partial}{\partial x}\right]\left[\sqrt{\frac{m}{ab}}\frac{\partial}{\partial t} + 2\sinh\left(\frac{1}{2}\frac{\partial}{\partial x}\right) - \frac{b}{2}y\frac{\partial}{\partial x}\right]y = 0.
\tag{B.3}
$$

For wave spreading in right direction one obtains

$$
\left[\sqrt{\frac{m}{ah}}\frac{\partial}{\partial t} + 2\sinh\left(\frac{1}{2}\frac{\partial}{\partial x}\right) - \frac{h}{2}y\frac{\partial}{\partial x}\right]y = 0.
\tag{B.4}
$$

Formally expanding function $\sinh((1/2)(\partial/\partial x))$ in a Maclaurin series up to the terms of the third-order one obtains from (B.4)

$$
\sqrt{\frac{m}{ab}}\frac{\partial y}{\partial t} + \frac{\partial y}{\partial x} + \frac{1}{24}\frac{\partial^3 y}{\partial x^3} - \frac{b}{2}y\frac{\partial y}{\partial x} = 0.
\tag{B.5}
$$

Using variables transform $t_1 = \sqrt{(ab/m)}t; y_1 = (b/2)y$ one obtains

$$\frac{\partial y_1}{\partial t_1} + \frac{\partial y_1}{\partial x} + \frac{1}{24}\frac{\partial^3 y_1}{\partial x^3} - y_1\frac{\partial y_1}{\partial x} = 0. \tag{B.6}$$

C. Correspondence between Functions of Discrete Arguments and Approximating Analytical Functions

Let us suppose system

$$m\sigma_{jtt}(t) = c(\sigma_{j+1} - 2\sigma_j + \sigma_{j-1}) + s_j(t) - s_{j-1}(t), \quad j = 0,1,\ldots,n-1,n \tag{C.1}$$

$s_j(t)$ is the external force acting on jth point.

Further we follow [44].

In continuum limit we must construct a function $s(x,t)$ representimg a continuous approximation of the function of discrete argument $s_j(t)$. Observe that it is defined with an accuracy to any arbitrary function, which equals zero in nodal points $x = j$, $j = 0,1,2,\ldots,n$. For this reason, from a set of interpolating functions one may choose the smoothest function owing to filtration of fast oscillating terms. As it is shown in [44], the interpolating function is determined uniquely, if one requires its Fourier image $\bar{s}(q)$

$$\bar{s}(q) = \int_{-\infty}^{\infty} s(x)e^{-iqx}dx; \quad s(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\bar{s}(q)e^{iqx}dx \tag{C.2}$$

to differ from zero only on the segment $-\pi/h \leq \bar{s}(q) \leq \pi/h$. From this condition one obtains

$$s(x,t) = \sum_{k=1}^{n} s_k(t)\frac{\sin(\pi(x-kh)/h)}{\pi(x-kh)}. \tag{C.3}$$

For function

$$\frac{\sin(\pi x/h)}{\pi x}, \tag{C.4}$$

one has the following relation (from [115, formula 2.5.3.12])

$$\int_{-\infty}^{\infty}\frac{\sin(\pi x/h)}{\pi x}dx = 1. \tag{C.5}$$

For $h \to \infty$ function (C.4) turns into Dirac's delta function. So for momentous theories of elasticity one must replace Dirac's delta function to the function (C.4).

Acknowledgments

References

[1] A. I. Gusev, A. A. Rempel, and A. J. Magerl, *Disorder and Order in Strongly Nonstoichiometric Compounds: Transition Metal Carbides, Nitrides and Oxides*, Springer, Berlin, Germany, 2001.

[2] I. S. Pavlov and A. I. Potapov, "Structural models in mechanics of nanocrystalline media," *Doklady Physics*, vol. 53, no. 7, pp. 408–412, 2008.

[3] A. I. Potapov, "Waves of deformation in the media with inner structure," in *Proceedings of the International Workshop on Nonlinear Waves*, A. V. Gaponov-Grekhov and V. I. Nekorkin, Eds., pp. 125–140, Institute of Applied Physics RAS, Nizhnyi Novgorod, Russia, 2004.

[4] A. I. Potapov, I. S. Pavlov, and S. A. Lisina, "Acoustic identification of nanocrystalline media," *Journal of Sound and Vibration*, vol. 322, no. 3, pp. 564–580, 2009.

[5] E. H. Dowell and D. Tang, "Multiscale, multiphenomena modeling and simulation at the nanoscale: on constructing reduced-order models for nonlinear dynamical systems with many degrees-of-freedom," *Journal of Applied Mechanics*, vol. 70, pp. 328–338, 2003.

[6] G. Friesecke and R. D. James, "A scheme for the passage from atomic to continuum theory for thin films, nanotubes and nanorods," *Journal of the Mechanics and Physics of Solids*, vol. 48, no. 6-7, pp. 1519–1540, 2000.

[7] S. Pagano and R. Paroni, "A simple model for phase transitions: from the discrete to the continuum problem," *Quarterly of Applied Mathematics*, vol. 61, no. 1, pp. 89–109, 2003.

[8] E. Aero, A. Fradkov, B. Andrievsky, and S. Vakulenko, "Dynamics and control of oscillations in a complex crystalline lattice," *Physics Letters A*, vol. 353, no. 1, pp. 24–29, 2006.

[9] E. Grekova and P. Zhilin, "Basic equations of Kelvin's medium and analogy with ferromagnets," *Journal of Elasticity*, vol. 64, no. 1, pp. 29–70, 2001.

[10] E. Pasternak and H.-B. Mühlhaus, "Generalized homogenization procedure for granular material," *Journal of Engineering Mathematics*, vol. 52, pp. 199–229, 2005.

[11] I. S. Pavlov, A. I. Potapov, and G. A. Maugin, "A 2D granular medium with rotating particles," *International Journal of Solids and Structures*, vol. 43, no. 20, pp. 6194–6207, 2006.

[12] R. H. J. Peerlings, M. G. D. Geers, R. de Borst, and W. A. M. Brekelmans, "A critical comparison of nonlocal and gradient-enhanced softening continua," *International Journal of Solids and Structures*, vol. 38, no. 44-45, pp. 7723–7746, 2001.

[13] H. Askes and L. J. Sluys, "Explicit and implicit gradient series in damage mechanics," *European Journal of Mechanics A*, vol. 21, no. 3, pp. 379–390, 2002.

[14] C. S. Chang, H. Askes, and L. J. Sluys, "Higher-order strain/higher-order stress gradient models derived from a discrete microstructure, with application to fracture," *Engineering Fracture Mechanics*, vol. 69, no. 17, pp. 1907–1924, 2002.

[15] G. del Piero and L. Truskinovsky, "A one-dimensional model for localized and distributed failure," *Journal de Physique*, vol. 8, no. 8, pp. 95–102, 1998.

[16] A. Yavari, M. Ortiz, and K. Bhattacharya, "A theory of anharmonic lattice statics for analysis of defective crystals," *Journal of Elasticity*, vol. 86, no. 1, pp. 41–83, 2007.

[17] N. A. Fleck and J. W. Hutchinson, "A phenomenological theory for strain gradient effects in plasticity," *Journal of the Mechanics and Physics of Solids*, vol. 41, no. 12, pp. 1825–1857, 1993.

[18] R. C. Rogers and L. Truskinovsky, "Discretization and hysteresis," *Physica B*, vol. 233, no. 4, pp. 370–375, 1997.

[19] P. G. Kevrekidis and I. G. Kevrekidis, "Heterogeneous versus discrete mapping problem," *Physical Review E*, vol. 64, no. 5, Article ID 056624, 8 pages, 2001.

[20] A. M. Kosevich, *Deformation and Destruction of Microstructured Solids*, Fizmatlit, Moscow, Russia, 2007.

[21] W. E and B. Engquist, "The heterogeneous multiscale methods," *Communications in Mathematical Sciences*, vol. 1, no. 1, pp. 87–132, 2003.

[22] E. B. Tadmor, M. Ortiz, and R. Phillips, "Quasicontinuum analysis of defects in solids," *Philosophical Magazine A*, vol. 73, no. 6, pp. 1529–1563, 1996.

[23] P. A. M. Dirac, "Quantum mechanics of many-electron systems," *Proceedings of the Royal Society of London A*, vol. 123, no. 729, pp. 714–733, 1929.

[24] P. Debye, "Zur Theorie der spezifischen Wärmen," *Annals of Physics*, vol. 39, pp. 789–839, 1912.

[25] E. C. Aifantis, "Gradient deformation models at nano, micro, and macro scales," *Journal of Engineering Materials and Technology*, vol. 121, no. 2, pp. 189–202, 1999.

[26] E. Cosserat and F. Cosserat, *Théorie des Corps Déformables*, Libraire Scientifique A. Hermann et Fils, Paris, France, 1909.

[27] E. Hellinger, "Die allgemeinen Ansätze der Mechanik der Kontinua," in *Enzyklopadie der Mathematischen Wissenschaften*, F. Klein and G. Müller, Eds., vol. 4, pp. 602–694, Teubner, Leipzig, Germany, 1914.

[28] M. J. Le Roux, "Étude géométrique de la torsion et de la flexion dans la déformation infinitésimale d'un milieu continu," *Annales Scientifiques de l'École Normale Supérieure*, vol. 28, pp. 523–579, 1911.

[29] G. A. Maugin, "From Piola's manifold to Cosserats' structure," in *Geometry, Continua and Microstructure*, G. A. Maugin, Ed., pp. 113–120, Hermann, Paris, France, 1999.

[30] R. D. Mindlin, "Micro-structure in linear elasticity," *Archive for Rational Mechanics and Analysis*, vol. 16, pp. 51–78, 1964.

[31] N. Triantafyllidis and S. Bardenhagen, "On higher order gradient continuum theories in 1-D nonlinear elasticity. Derivation from and comparison to the corresponding discrete models," *Journal of Elasticity*, vol. 33, no. 3, pp. 259–293, 1993.

[32] A. S. J. Suiker, A. V. Metrikine, and R. De Borst, "Comparison of wave propagation characteristics of the Cosserat continuum model and corresponding discrete lattice models," *International Journal of Solids and Structures*, vol. 38, no. 9, pp. 1563–1583, 2001.

[33] A. S. J. Suiker, A. V. Metrikine, and R. Borst, "Dynamical behavior of a layer of discrete particles— part 1: analysis of body waves and eigenmodes," *Journal of Sound and Vibration*, vol. 240, no. 1, pp. 1–18, 2001.

[34] A. S. J. Suiker, A. V. Metrikine, and R. Borst, "Dynamical behavior of a layer of discrete particles— part 2: response to a uniform moving, harmonically vibrating loading," *Journal of Sound and Vibration*, vol. 240, no. 1, pp. 19–39, 2001.

[35] D. Hilbert, *Gesammelte Abhandlungen III*, Springer, Berlin, Germany, 1935.

[36] M. J. Beran, *Statistical Continuum Theories*, Wiley-Interscience, New York, NY, USA, 1968.

[37] X. Blanc, C. Le Bris, and P.-L. Lions, "From molecular models to continuum mechanics," *Archive for Rational Mechanics and Analysis*, vol. 164, no. 4, pp. 341–381, 2002.

[38] X. Blanc, C. Le Bris, and P.-L. Lions, "Atomistic to continuum limits for computational materials science," *Mathematical Modelling and Numerical Analysis*, vol. 41, no. 2, pp. 391–426, 2007.

[39] A. Braides, "Non-local variational limits of discrete systems," *Communications in Contemporary Mathematics*, vol. 2, no. 2, pp. 285–297, 2000.

[40] M. Charlotte and L. Truskinovsky, "Linear elastic chain with a hyper-pre-stress," *Journal of the Mechanics and Physics of Solids*, vol. 50, no. 2, pp. 217–251, 2002.

[41] G. Friesecke and F. Theil, "Validity and failure of the Cauchy-Born hypothesis in a two-dimensional mass-spring lattice," *Journal of Nonlinear Science*, vol. 12, no. 5, pp. 445–478, 2002.

[42] R. Paroni, "From discrete to continuum: a Young measure approach," *Zeitschrift für Angewandte Mathematik und Physik*, vol. 54, no. 2, pp. 328–348, 2003.

[43] B. Schmidt, "A derivation of continuum nonlinear plate theory from atomistic models," *Multiscale Modeling & Simulation*, vol. 5, no. 2, pp. 664–694, 2006.

[44] I. A. Kunin, *Elastic Media with Microstructure. I. One-Dimensional Models*, vol. 26 of *Springer Series in Solid-State Sciences*, Springer, Berlin, Germany, 1982.

[45] I. A. Kunin, *Elastic Media with Microstructure. II. Three-Dimensional Models*, vol. 44 of *Springer Series in Solid-State Sciences*, Springer, Berlin, Germany, 1983.

[46] I. V. Andrianov, "Continuous approximation of higher-frequency oscillation of a chain," *Doklady Akademii Nauk Ukrainskoj SSR, Seriya A*, vol. 2, pp. 13–15, 1991 (Russian).

[47] I. V. Andrianov, "Padé approximants and continuum models for a one-dimensional lattice of masses," *Matematicheskoe Modelirovanie*, vol. 18, no. 1, pp. 43–58, 2006 (Russian).

[48] I. V. Andrianov and J. Awrejcewicz, "Continuous models for chain of inertially linked masses," *European Journal of Mechanics A*, vol. 24, no. 3, pp. 532–536, 2005.

[49] I. V. Andrianov and J. Awrejcewicz, "Continuous models for 1D discrete media valid for higher-frequency domain," *Physics Letters A*, vol. 345, no. 1–3, pp. 55–62, 2005.

[50] I. V. Andrianov and J. Awrejcewicz, "Love and Rayleigh correction terms and Padé approximants," *Mathematical Problems in Engineering*, vol. 2007, Article ID 94035, 8 pages, 2007.

[51] I. V. Andrianov and J. Awrejcewicz, "Continuous models for 2D discrete media valid for higher-frequency domain," *Computers and Structures*, vol. 86, no. 1-2, pp. 140–144, 2008.

[52] I. V. Andrianov and J. Awrejcewicz, "New trends in asymptotic approaches: summation and interpolation methods," *Applied Mechanics Reviews*, vol. 54, no. 1, pp. 69–92, 2001.

[53] M. A. Collins, "A quasicontinuum approximation for solitons in an atomic chain," *Chemical Physics Letters*, vol. 77, no. 2, pp. 342–347, 1981.

[54] I. F. Obraztsov, I. V. Andrianov, and B. V. Nerubailo, "Continuum approximation for high-frequency oscillations of a chain and composite equations," *Soviet Physics Doklady*, vol. 36, no. 7, p. 522, 1991.

[55] Ph. Rosenau, "Dynamics of nonlinear mass-spring chains near the continuum limit," *Physics Letters A*, vol. 118, no. 5, pp. 222–227, 1986.

[56] Ph. Rosenau, "Dynamics of dense lattices," *Physical Review B*, vol. 36, no. 11, pp. 5868–5876, 1987.

[57] Ph. Rosenau, "Dynamics of dense discrete systems," *Progress of Theoretical Physics*, vol. 89, no. 5, pp. 1028–1042, 1988.

[58] Ph. Rosenau, "Hamiltonian dynamics of dense chains and lattices: or how to correct the continuum," *Physics Letters A*, vol. 311, no. 1, pp. 39–52, 2003.

[59] Ph. Rosenau, "Extending hydrodynamics via the regularization of the Chapman-Enskog expansion," *Physical Review A*, vol. 40, no. 12, pp. 7193–7196, 1989.

[60] M. B. Rubin, Ph. Rosenau, and O. Gottlieb, "Continuum model of dispersion caused by an inherent materil characteristics length," *Journal of Applied Physics*, vol. 77, no. 5, pp. 4054–4063, 1995.

[61] J. A. D. Wattis, "Approximations to solitary waves on lattices. II. Quasi-continuum methods for fast and slow waves," *Journal of Physics A*, vol. 26, no. 5, pp. 1193–1209, 1993.

[62] J. A. D. Wattis, "Quasi-continuum approximations to lattice equations arising from the discrete nonlinear telegraph equation," *Journal of Physics A*, vol. 33, no. 33, pp. 5925–5944, 2000.

[63] I. V. Andrianov, "The special feature of limiting transition from a discrete elastic media to a continuous one," *Journal of Applied Mathematics and Mechanics*, vol. 66, no. 2, pp. 261–265, 2002.

[64] I. V. Andrianov and J. Awrejcewicz, "On the average continuous representation of an elastic discrete medium," *Journal of Sound and Vibration*, vol. 264, no. 5, pp. 1187–1194, 2003.

[65] A. D. Myshkis, "Mixed functional-differential equations," *Journal of Mathematical Sciences*, vol. 129, no. 5, pp. 4111–4226, 2005.

[66] P. F. Kurchanov, A. D. Myshkis, and A. M. Filimonov, "Vibrations of rolling stock and a theorem of Kronecker," *Journal of Applied Mathematics and Mechanics*, vol. 55, no. 6, pp. 870–876, 1991.

[67] V. P. Maslov, *Operational Methods*, Mir, Moscow, Russia, 1976.

[68] N. E. Joukowsky, "The work of continuous and non-continuous traction devices in pulling a train from its position and at the beginning of its motion," in *Complete Collected Works, 8: Theory of Elasticity. Railways. Automobiles*, N. E. Joukowsky and A. P. Kotel'nikov, Eds., pp. 221–255, ONTI, Moscow, Russia, 1937.

[69] A. M. Filimonov, "Some unexpected results on the classical problem of vibrations of the string with *N* beads. The case of multiple frequencies," *Comptes Rendus de l'Académie des Sciences. Série I*, vol. 315, no. 8, pp. 957–961, 1992.

[70] A. M. Filimonov, "Continuous approximations of difference operators," *Journal of Difference Equations and Applications*, vol. 2, no. 4, pp. 411–422, 1996.

[71] A. M. Filimonov and A. D. Myshkis, "Asymptotic estimate of solution of one mixed difference-differential equation of oscillations theory," *Journal of Difference Equations and Applications*, vol. 4, no. 1, pp. 13–16, 1998.

[72] A. Filimonov, X. Mao, and S. Maslov, "Splash effect and ergodic properties of solution of the classic difference-differential equation," *Journal of Difference Equations and Applications*, vol. 6, no. 3, pp. 319–328, 2000.

[73] A. M. Filimonov, P. F. Kurchanov, and A. D. Myshkis, "Some unexpected results in the classical problem of vibrations of the string with *n* beads when *n* is large," *Comptes Rendus de l'Académie des Sciences. Série I*, vol. 313, no. 13, pp. 961–965, 1991.

[74] S. M. Ulam, *A Collection of Mathematical Problems*, Interscience Tracts in Pure and Applied Mathematics, no. 8, Interscience, New York, NY, USA, 1960.

[75] M. Berezhnyy and L. Berlyand, "Continuum limit for three-dimensional mass-spring networks and discrete Korn's inequality," *Journal of the Mechanics and Physics of Solids*, vol. 54, no. 3, pp. 635–669, 2006.

[76] A. M. Kosevich, *The Crystal Lattice: Phonons, Solitons, Dislocations, Superlattices*, Wiley-VCH, Berlin, Germany, 2005.

[77] Yu. A. Kosevich, "Nonlinear envelope-function equation and strongly localized vibrational modes in anharmonic lattices," *Physical Review B*, vol. 47, no. 6, pp. 3138–3152, 1993.

[78] A. S. Kovalev and A. M. Kosevich, "Self-localization of vibrations in a one-dimensional anharmonic chain," *Soviet Physics JETP*, vol. 40, pp. 891–896, 1975.

[79] A. F. Vakakis, L. I. Manevitch, Yu. V. Mikhlin, V. N. Pilipchuk, and A. A. Zevin, *Normal Modes and Localization in Nonlinear Systems*, Wiley Series in Nonlinear Science, John Wiley & Sons, New York, NY, USA, 1996.

[80] R. C. Picu, "On the functional form of non-local elasticity kernels," *Journal of the Mechanics and Physics of Solids*, vol. 50, no. 9, pp. 1923–1939, 2002.

[81] H. Askes and A. V. Metrikine, "Higher-order continua derived from discrete media: continualisation aspects and boundary conditions," *International Journal of Solids and Structures*, vol. 42, no. 1, pp. 187–202, 2005.

[82] M. Born and K. Huang, *Dynamical Theory of Crystal Lattices*, Clarendon Press, Oxford, UK; Oxford University Press, New York, NY, USA, 1988.

[83] A. Askar, *Lattice Dynamical Foundations of Continuum Theories: Elasticity, Piezoelectricity, Viscoelasticity, Plasticity*, vol. 2 of *Series in Theoretical and Applied Mechanics*, World Scientific, Philadelphia, Pa, USA, 1985.

[84] H. Askes and M. A. Gutiérrez, "Implicit gradient elasticity," *International Journal for Numerical Methods in Engineering*, vol. 67, no. 3, pp. 400–416, 2006.

[85] H. Askes, A. V. Metrikine, A. V. Pichugin, and T. Bennett, "Four simplified gradient elasticity models for the simulation of dispersive wave propagation," *Philosophical Magazine*, vol. 88, no. 28-29, pp. 3415–3443, 2008.

[86] G. A. Maugin, *Nonlinear Waves in Elastic Crystals*, Oxford Mathematical Monographs, Oxford University Press, Oxford, UK, 1999.

[87] A. V. Pichugin, H. Askes, and A. Tyas, "Asymptotic equivalence of homogenisation procedures and fine-tuning of continuum theories," *Journal of Sound and Vibration*, vol. 313, no. 3–5, pp. 858–874, 2008.

[88] V. I. Erofeyev, *Wave Processes in Solids with Microstrucutre*, World Scientific, Singapore, 2003.

[89] S. V. Dmitriev, A. A. Vasiliev, N. Yoshikawa, T. Shigenari, and Y. Ishibashi, "Multi-cell continuum approximation for discrete medium with microscopic rotations," *Physica Status Solidi B*, vol. 242, no. 3, pp. 528–537, 2005.

[90] E. A. Il'iushina, "On a model of continuous medium, taking into account the microstructure," *Journal of Applied Mathematics and Mechanics*, vol. 33, no. 5, pp. 917–923, 1969.

[91] E. A. Il'iushina, "A version of the couple stress theory of elasticity for a one-dimensional continuous medium with inhomogeneous periodic structure," *Journal of Applied Mathematics and Mechanics*, vol. 36, no. 6, pp. 1019–1026, 1972.

[92] I. V. Andrianov, J. Awrejcewicz, and R. G. Barantsev, "Asymptotic approaches in mechanics: new parameters and procedures," *Applied Mechanics Reviews*, vol. 56, no. 1, pp. 87–110, 2003.

[93] N. Tzirakis and P. G. Kevrekidis, "On the collapse arresting effects of discreteness," *Mathematics and Computers in Simulation*, vol. 69, no. 5-6, pp. 553–566, 2005.

[94] E. I. Grigolyuk and I. I. Selezov, *Nonclassical Theories of Vibrations of Beams, Plates and Shells*, VINITI, Moscow, Russia, 1973.

[95] A. E. H. Love, *A Treatise on the Mathematical Theory of Elasticity*, Cambridge University Press, Cambridge, UK, 4th edition, 1927.

[96] J. W. S. Rayleigh, *The Theory of Sound. Vol. 1*, Macmillan, London, UK, 1877.

[97] J. W. S. Rayleigh, *The Theory of Sound. Vol. 2*, Macmillan, London, UK, 1888.

[98] J. D. Kaplunov, L. Yu. Kossovich, and E. V. Nolde, *Dynamics of Thin Walled Elastic Bodies*, Academic Press, San Diego, Calif, USA, 1998.

[99] A. V. Metrikine, "On causality of the gradient elasticity models," *Journal of Sound and Vibration*, vol. 297, no. 3–5, pp. 727–742, 2006.

[100] B.-F. Feng, Y. Doi, and T. Kawahara, "Quasi-continuum approximation for discrete breathers in Fermi-Pasta-Ulam atomic chains," *Journal of the Physical Society of Japan*, vol. 73, no. 8, pp. 2100–2111, 2004.

[101] B.-F. Feng, Y. Doi, and T. Kawahara, "A regularized model equation for discrete breathers in anharmonic lattices with symmetric nearest-neighbor potentials," *Physica D*, vol. 214, no. 1, pp. 33–41, 2006.

[102] B.-F. Feng and T. Kawahara, "Discrete breathers in two-dimensional nonlinear lattices," *Wave Motion*, vol. 45, no. 1-2, pp. 68–82, 2007.

[103] A. C. Eringen, "On differential equations of nonlocal elasticity and solutions of screw dislocation and surface waves," *Journal of Applied Physics*, vol. 54, no. 9, pp. 4703–4710, 1983.

[104] A. C. Eringen, "Theory of nonlocal elasticity and some applications," *Res Mechanica*, vol. 21, no. 4, pp. 313–342, 1987.

[105] A. C. Eringen, "Vistas of nonlocal continuum physics," *International Journal of Engineering Science*, vol. 30, no. 10, pp. 1551–1565, 1992.

[106] A. C. Eringen, *Microcontinuum Field Theories. I. Foundations and Solids*, Springer, New York, NY, USA, 1999.

[107] M. Born and Th. von Kármán, "On fluctuations in spatial grids," *Physikalische Zeitschrift*, vol. 13, pp. 297–309, 1912.

[108] H. Askes and A. V. Metrikine, "One-dimensional dynamically consistent gradient elasticity models derived from a discrete microstructure—part 2: static and dynamic response," *European Journal of Mechanics A*, vol. 21, no. 4, pp. 573–588, 2002.

[109] A. V. Metrikine and H. Askes, "One-dimensional dynamically consistent gradient elasticity models derived from a discrete microstructure—part 1: generic formulation," *European Journal of Mechanics A*, vol. 21, no. 4, pp. 555–572, 2002.

[110] R. D. Mindlin and G. Herrmann, "A one-dimensional theory of compressional waves in an elastic rod," in *Proceedings of the 1st U.S. National Congress of Applied Mechanics*, pp. 187–191, ASME, New York, NY, USA, 1952.

[111] R. D. Mindlin and G. Herrmann, "Corrections," in *Proceedings of the 2nd U.S. National Congress of Applied Mechanics*, p. 233, ASME, New York, NY, USA, 1954.

[112] L. Brillouin, *Wave Propagation in Periodic Structures: Electric Filters and Crystal Lattices*, Dover, Mineola, NY, USA, 2nd edition, 2003.

[113] M. Lazar, G. A. Maugin, and E. C. Aifantis, "On a theory of nonlocal elasticity of bi-Helmholtz type and some applications," *International Journal of Solids and Structures*, vol. 43, no. 6, pp. 1404–1421, 2006.

[114] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, NY, USA, 1965.

[115] A. P. Prudnikov, Yu. A. Brychkov, and O. I. Marichev, *Integrals and Series. Vol. 1. Elementary Functions*, Gordon and Breach, New York, NY, USA, 1986.

[116] L. I. Slepyan, *Nonstationary Elastic Waves*, Sudostroyenie, Leningrad, Russia, 1972.

[117] A. V. Metrikine and H. Askes, "An isotropic dynamically consistent gradient elasticity model derived from a 2D lattice," *Philosophical Magazine*, vol. 86, no. 21-22, pp. 3259–3286, 2006.

[118] M. Van Dyke, *Perturbation Methods*, The Parabolic Press, Stanford, Calif, USA, 1975.

[119] A. Scott, *Nonlinear Science: Emergence and Dynamics of Coherent Structure*, vol. 8 of *Oxford Texts in Applied and Engineering Mathematics*, Oxford University Press, Oxford, UK, 2nd edition, 2003.

[120] L. I. Slepyan and L. V. Troyankina, "Fracture wave in a chain structure," *Journal of Applied Mechanics and Technical Physics*, vol. 25, no. 6, pp. 921–927, 1984.

[121] D. B. Duncan, J. C. Eilbeck, H. Feddersen, and J. A. D. Wattis, "Solitons on lattices," *Physica D*, vol. 68, pp. 1–11, 1993.

[122] M. Toda, *Theory of Nonlinear Lattices*, vol. 20 of *Springer Series in Solid-State Sciences*, Springer, Berlin, Germany, 2nd edition, 1989.

[123] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform*, vol. 4 of *SIAM Studies in Applied Mathematics*, SIAM, Philadelphia, Pa, USA, 1981.

[124] T. B. Benjamin, J. L. Bona, and J. J. Mahony, "Model equations for long waves in nonlinear dispersive systems," *Philosophical Transactions of the Royal Society A*, vol. 272, no. 1220, pp. 47–78, 1972.

[125] D. H. Peregrine, "Calculations of the development of an undular bore," *Journal of Fluid Mechanics*, vol. 25, no. 2, pp. 321–330, 1966.

[126] P. Deift and K. T.-R. McLaughlin, *A Continuum Limit of the Toda Lattice*, AMS, Providence, RI, USA, 1998.

[127] R. K. Dodd, J. C. Eilbeck, J. D. Gibbon, and H. C. Morris, *Solitons and Nonlinear Wave Equations*, Academic Press, London, UK, 1984.

[128] J. Giannoulis and A. Mielke, "The nonlinear Schrödinger equation as a macroscopic limit for an oscillator chain with cubic nonlinearities," *Nonlinearity*, vol. 17, no. 2, pp. 551–565, 2004.

[129] M. Kac and P. van Moerbeke, "On an explicitly soluble system of nonlinear differential equations related to certain Toda lattices," *Advances in Mathematics*, vol. 16, pp. 160–169, 1975.

[130] V. P. Maslov and G. A. Omel'yanov, *Geometric Asymptotics for Nonlinear PDE. I*, vol. 202 of *Translations of Mathematical Monographs*, American Mathematical Society, Providence, RI, USA, 2001.

[131] G. M. Chechin and G. S. Dzhelauhova, "Discrete breathers and nonlinear normal modes in monoatomic chains," *Journal of Sound and Vibration*, vol. 322, no. 3, pp. 490–512, 2009.

[132] L. I. Manevitch, I. V. Andrianov, and V. G. Oshmyan, *Mechanics of Periodically Heterogeneous Structures*, Foundations of Engineering Mechanics, Springer, Berlin, Germany, 2002.

[133] O. A. Ladyzhenskaya, "The sixth millennium problem: Navier-Stokes equations, existence and smoothness," *Russian Mathematical Surveys*, vol. 58, no. 2, pp. 251–286, 2003.

[134] S. Smale, "Mathematical problems for the next century," in *Mathematics: Frontiers and Perspective*, V. Arnold, M. Atiyah, P. Lax, and B. Mazur, Eds., AMS, Providence, RI, USA, 2000.

[135] V. I. Yudovich, "Problems and promisings of contemporary hydrodynamics," *Uspehi Mekhaniki*, vol. 1, pp. 61–101, 2002 (Russian).

[136] V. P. Maslov, "Coherent structures, resonances and asymptotic nonuniqueness for the Navier-Stokes equations for large Reynolds numbers," *Russian Mathematical Surveys*, vol. 41, no. 6, pp. 19–35, 1986.

[137] O. A. Ladyzhenskaya, "Some nonlinear problems of the continuum mechanics," in *Proceedings of the International Congress of Mathematicians*, pp. 561–573, Nauka, Moscow, Russia, 1968.

[138] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, NY, USA, 2nd edition, 1969.

[139] O. A. Ladyzhenskaya, "Modification of the Navier-Stokes equations for large velocity gradients," in *Boundary Value Problems of Mathematical Physics and Related Aspects of Function Theory II*, pp. 57–69, Consultants Bureau, New York, NY, USA, 1970.

[140] K. K. Golovkin, "New model equations of motion of a viscous fluid and their unique solvability," *Proceedings of the Steklov Institute of Mathematics*, vol. 102, pp. 29–50, 1967.

[141] A. Jaffe and F. Quinn, ""Theoretical mathematics": toward a cultural synthesis of mathematics and theoretical physics," *Bulletin of the American Mathematical Society*, vol. 29, no. 1, pp. 1–13, 1993.

[142] G. J. Wagner and W. K. Liu, "Coupling of atomistic and continuum simulations using a bridging scale decomposition," *Journal of Computational Physics*, vol. 190, no. 1, pp. 249–274, 2003.

[143] M. Xu and T. Belytschko, "Conservation properties of the bridging domain method for coupled molecular/continuum dynamics," *International Journal for Numerical Methods in Engineering*, vol. 76, no. 3, pp. 278–294, 2008.

[144] S. Zhang, R. Khare, Q. Lu, and T. Belytschko, "A bridging domain and strain computation method for coupled atomistic-continuum modelling of solids," *International Journal for Numerical Methods in Engineering*, vol. 70, no. 8, pp. 913–933, 2007.

[145] V. V. Vavilov, M. K. Tchobanou, and P. M. Tchobanou, "Design of multidimensional recursive systems through Padé type rational approximation," *Nonlinear Analysis: Modelling and Control*, vol. 7, no. 1, pp. 105–125, 2002.

[146] L. I. Slepyan, "Dynamics of a crack in a lattice," *Soviet Physics Doklady*, vol. 26, pp. 538–540, 1981.

[147] L. I. Slepyan, "Crack propagation in high-frequency lattice vibration," *Soviet Physics Doklady*, vol. 26, pp. 900–902, 1981.

[148] L. I. Slepyan, "The relation between the solutions of mixed dynamic problems for a continuous elastic medium and a lattice," *Soviet Physics Doklady*, vol. 27, pp. 771–772, 1982.

[149] L. I. Slepyan, "Dynamics of brittle fracture in media with a structure," *Mechanics of Solids*, vol. 19, no. 6, pp. 114–122, 1984.

[150] L. I. Slepyan and L. V. Troyankina, "Impact waves in a nonlinear chain," in *Plasticity and Fracture of Solids*, R. V. Gol'dstein, Ed., pp. 175–186, Nauka, Moscow, Russia, 1988.

[151] J. J. Kruzik, "Predicting fatigue failures," *Science*, vol. 325, pp. 156–157, 2009.

[152] G. A. Baker Jr. and P. Graves-Morris, *Padé Approximants*, vol. 59 of *Encyclopedia of Mathematics and Its Applications*, Cambridge University Press, Cambridge, UK, 2nd edition, 1996.

Review Article

Furuta's Pendulum: A *Conservative* Nonlinear Model for Theory and Practise

J. Á. Acosta

Departamento de Ingeniería de Sistemas y Automática, Escuela Técnica Superior de Ingenieros, Camino de los Descubrimientos s/n, 41092 Sevilla, Spain

Correspondence should be addressed to J. Á. Acosta, jaar@esi.us.es

Furuta's pendulum has been an excellent benchmark for the automatic control community in the last years, providing, among others, a better understanding of model-based Nonlinear Control Techniques. Since most of these techniques are based on invariants and/or integrals of motion then, the dynamic model plays an important role. This paper describes, in detail, the successful dynamical model developed for the available laboratory pendulum. The success relies on a basic dynamical model derived from Classical Mechanics which has been augmented to compensate the *non-conservative* torques. Thus, the *quasi-conservative* "practical" model developed allows to design all the controllers as if the system was strictly *conservative*. A survey of all the nonlinear controllers designed and experimentally tested on the available laboratory pendulum is also reported.

1. Introduction

The control of underactuated mechanical systems has attracted the interest of the control engineering community for many years. Underactuated mechanical systems are control systems with fewer control inputs than degrees of freedom. This lack of inputs is, precisely, the main difficulty to control them. Moreover, the control of underactuated mechanical systems is currently an active field of research due to their broad applications in Robotics, Aerospace, and Marine Vehicles. Some examples of underactuated mechanical systems include flexible-link robots, mobile robots, walking robots, robots on mobile platforms, cars, locomotive systems, snake-type and swimming robots, acrobatic robots, aircraft, spacecraft, helicopters, satellites, surface vessels, and underwater vehicles.

One of the most well-known underactuated mechanical systems is the pendulum which is an excellent benchmark for many control problems, attracting the attention of many

Figure 1: Available Furuta's pendulum.

researchers during years. The available version of the pendulum is the rotary pendulum shown in Figure 1, and this survey is focused on that laboratory pendulum where many control strategies have been designed and successfully tested. The rotary pendulum is one of the most complex two-degrees-of-freedom pendulum since its configuration provides centrifugal torques, which are quadratic in velocities, giving rise to many interesting and difficult control problems.

The rotary pendulum is also known as Furuta's pendulum. Furuta was the inventor, and has been studied by many authors such as Furuta himself [1] (see also [2–4] among others). It is a non-linear underactuated mechanical system that is unstable at the desired upright position. In fact, it shows, mainly, two different and very interesting control problems. The first one is to swing the pendulum up from the rest (down) to the upright (up) position, which is commonly solved with energy control strategies. Once the pendulum is close to the desired upright position, at low speed, a stabilization or balancing strategy "catches" it there. Other control problems are also quite interesting as the stabilisation of autonomous oscillations and control through bifurcation analysis. All of them are briefly commented in the survey focusing on all the designed and experimentally tested in the available laboratory pendulum.

The paper is organized as follows. In Section 2 the non-linear model for the available pendulum is derived in detail, including a dynamical-friction model. In Section 3, a survey of the non-linear strategies designed and tested on the available pendulum is given. A literature review showing the relations with other approaches are also briefly commented.

Finally, a conclusion section and two appendices, one describing the whole experimental control system and another with the MATLAB codes used for the non-trivial identification of parameters of the dynamical-friction model.

Notation and Definitions

Throughout the paper it is assumed that all functions and vector functions are sufficiently smooth. Moreover, concerning functions, vector functions, and matrices which are functions of some variables, we will write explicitly this dependence the first time they are defined. The norm of a vector $x \in \mathbb{R}^n$ will be denoted by $\|x\|$ and the square of the Euclidean norm by $\|x\|_2^2 = x^\top x$. Acronyms used: w.r.t. means "with respect to"; i.e. means "that is."

Consider a dynamical system affine in control as $\dot{x}(t) = f(x(t)) + g(x(t))u$, where $x \in \mathbb{R}^n$ is the state space vector, $u \in \mathbb{R}^m$ is the control input vector and where we assume that the origin is at $x = 0$ with $f(0) = 0$ and $g(0) \neq 0$. Thus,

(i) the *open-loop* system means the unforced system, that is, with $u = 0$;

(ii) a (smooth) static state-feedback means a control input of the form $u = \phi(x)$, also called controller, control law and stabiliser when a stabilisation problem is considered;

(iii) the *closed-loop* system means the system with a specified static state-feedback, i.e. $\dot{x} = f(x) + g(x)\phi(x)$;

(iv) *linearisation* means a *non-linear* static state-feedback that linearises fully or partially the dynamics;

(v) the *Jacobian linearisation* means the first approximation of the system dynamics through the *Jacobian* at the origin.

2. The *Conservative* Mechanical Model

The schematic of Furuta's pendulum is described in Figure 2, where the two degrees of freedom are shown: the angular positions of the arm (φ) and the pendulum (θ). The arm shaft (corresponding to the angle φ referred to the x-axis) is subject to a torque, while no torque is applied directly to the pendulum shaft (angle θ of pendulum referred to the z-axis). Thus, it is an underactuated mechanical system because it has only one control input and two degrees of freedom. Rigorously, the configuration space is a 2-torus, that is, $(\theta, \varphi) \in \mathcal{S}^1 \times \mathcal{S}^1$ and the corresponding tangent space is $(\theta, \varphi, \dot{\theta}, \dot{\varphi}) \in \mathcal{S}^1 \times \mathcal{S}^1 \times \mathbb{R}^2$.

We define the physical parameters of the system as

m: mass of the pendulum,

$2l$: total length of the pendulum,

r: length of the arm (radius),

g: acceleration of gravity,

$\mathcal{J}$: moment of inertia of the pendulum w.r.t. the pivot,

$\mathcal{J}_a$: moment of inertia of the arm and the motor,

K: constant torque,

u: control input,

τ: control torque ($\tau := Ku$).
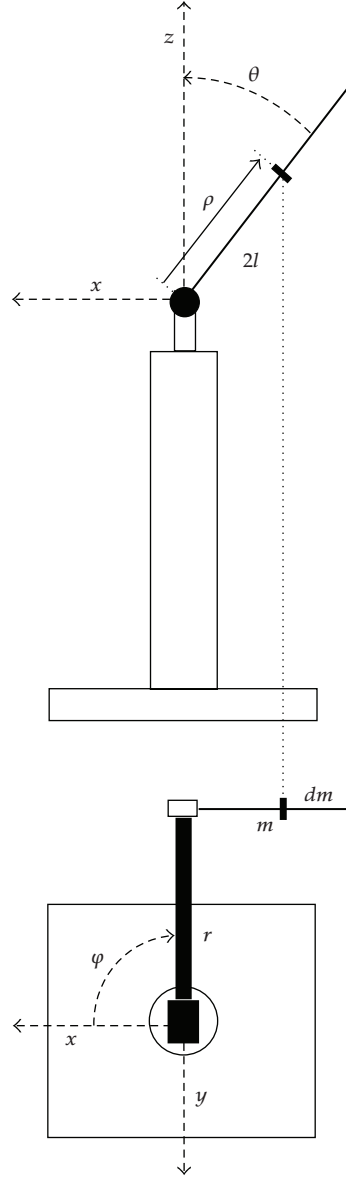
Figure 2: Reference frame of coordinates.

In what follows, energy methods of Classical Mechanics are used to derive the dynamical equations of motion. So, defining the reference frame of coordinates as in Figure 2 and assuming a linear distribution of the mass of the pendulum then, the coordinates of a differential mass element, namely ρ-mass element (see Figure 2), of the pendulum referred to that reference frame read

$$
\begin{aligned}
x_\rho &= r\cos\varphi - \rho\sin\theta\sin\varphi, \\
y_\rho &= r\sin\varphi + \rho\sin\theta\cos\varphi, \\
z_\rho &= \rho\cos\theta.
\end{aligned}
\tag{2.1}
$$

Since it is a simple mechanical system the total energy is composed by the sum of the kinetic and the potential energies. Thus, on the one hand the instantaneous kinetic energy of the system is composed by the kinetic energies of the pendulum and the arm (including the motor) and becomes

$$\mathcal{T}(\theta, \dot\theta, \dot\varphi) = \frac{1}{2}\int_0^{2l} v_\rho^2(\rho)\,dm + \frac{1}{2}\mathcal{J}_a\dot\varphi^2, \tag{2.2}$$

where v_ρ is the module of the velocity of the ρ-mass element of the pendulum, which is obtained differentiating (2.1) w.r.t. the time and adding the squares. Thus, it becomes

$$v_\rho^2(\rho) = r^2\dot\varphi^2 + \rho^2\dot\varphi^2\sin^2\theta + 2r\rho\dot\theta\dot\varphi\cos\theta + \rho^2\dot\theta^2. \tag{2.3}$$

Since by assumption $dm = m/2l\ d\rho$ then, the kinetic energy from (2.2) and (2.3) becomes

$$\mathcal{T}(\theta, \dot\theta, \dot\varphi) = \frac{1}{2}\begin{bmatrix}\dot\theta \\ \dot\varphi\end{bmatrix}^\top \begin{bmatrix}\mathcal{J} & mrl\cos\theta \\ mrl\cos\theta & \mathcal{J}_a + mr^2 + \mathcal{J}\sin^2\theta\end{bmatrix}\begin{bmatrix}\dot\theta \\ \dot\varphi\end{bmatrix}. \tag{2.4}$$

On the other hand, the potential energy of the system w.r.t. the given reference frame is only the one of the pendulum and, integrating for the ρ-mass element, becomes

$$\mathcal{V}(\theta) = mgl\ \ (\cos\theta - 1), \tag{2.5}$$

which is zero at the upright position of the pendulum. Now, with the energy of the whole system completely defined we use both standard and well-known energetic approaches, Euler-Lagrange and Hamilton, to derive the dynamical equations.

2.1. Euler-Lagrange Equations

Let define the vectors $q := (\theta, \varphi)^\top \in \mathcal{S}^1 \times \mathcal{S}^1$ and $\dot q := (\dot\theta, \dot\varphi)^\top \in \mathbb{R}^2$ and, just for compactness, another different set of physical parameters as

$$\alpha := \frac{mrl}{\mathcal{J}}, \qquad \beta := \frac{\mathcal{J}_a + mr^2}{\mathcal{J}}, \qquad \gamma := \frac{K}{mgl}, \qquad \omega_0^2 := \frac{mgl}{\mathcal{J}}. \tag{2.6}$$

Thus, to derive the Euler-Lagrange equations we calculate the Lagrangian of the system as the difference between the kinetic and potential energies of the whole system from equations (2.4) and (2.5), respectively. Then, the Lagrangian can be written as

$$\mathcal{L}(q, \dot{q}) = \frac{1}{2}\dot{q}^\top M(q)\dot{q} - \mathcal{U}(q), \tag{2.7}$$

where we have introduced the inertia matrix $M(q)$ that from (2.4) and with the new set of parameters (2.6) becomes

$$M(q) = 2\begin{bmatrix} 1 & \alpha\cos q_1 \\ \alpha\cos q_1 & \beta + \sin^2 q_1 \end{bmatrix}. \tag{2.8}$$

The well-known Euler-Lagrange equations [5] for this system take the form

$$\frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{q}_1}\right) - \frac{\partial\mathcal{L}}{\partial q_1} = 0,$$
$$\frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{q}_2}\right) - \frac{\partial\mathcal{L}}{\partial q_2} = \tau, \tag{2.9}$$

which can be rewritten as

$$M(q)\ddot{q} + C(q, \dot{q})\dot{q} + \frac{\partial\mathcal{U}}{\partial q}(q) = \begin{bmatrix} 0 \\ \tau \end{bmatrix}, \tag{2.10}$$

where $C(q, \dot{q})$ is the matrix of Coriolis and gyroscopic torques. Thus, the matrices defined in (2.10) are given by

$$C(q, \dot{q}) = 2\begin{bmatrix} 0 & -\dot{q}_2\sin q_1\cos q_1 \\ -\alpha\dot{q}_1\sin q_1 + \dot{q}_2\sin q_1\cos q_1 & \dot{q}_1\sin q_1\cos q_1 \end{bmatrix},$$
$$\frac{\partial\mathcal{U}}{\partial q}(q) = 2\begin{bmatrix} -\omega_0^2\sin q_1 \\ 0 \end{bmatrix}, \tag{2.11}$$

and so (2.10) reads

$$\ddot{q}_1 + \alpha\cos q_1\ddot{q}_2 - \frac{1}{2}\sin 2q_1\dot{q}_2^2 - \omega_0^2\sin q_1 = 0, \tag{2.12}$$

$$\alpha\cos\theta\ddot{q}_1 + \left(\beta + \sin^2 q_1\right)\ddot{q}_2 - \alpha\sin q_1\dot{q}_1^2 + \sin 2q_1\dot{q}_1\dot{q}_2 = \gamma\omega_0^2 u. \tag{2.13}$$

Notice that, if the matrix $C(q, \dot{q})$ is derived through Christoffel symbols [5, 6] (which is the case), then the matrix $\dot{M}(q) - 2C(q, \dot{q})$ is skew-symmetric.

Since, the inertia matrix $M(q)$ is positive definite and so invertible, then these equations can be rewritten as an affine-in-control system, just isolating the accelerations from (2.10) as

$$\ddot{q} = -M(q)^{-1}\left(C(q,\dot{q})\dot{q} + \frac{\partial \mathcal{U}}{\partial q}(q)\right) + M(q)^{-1}\begin{bmatrix} 0 \\ \tau \end{bmatrix}. \tag{2.14}$$

Denoting the determinant of the inertia matrix as $\Delta(q)$, then (2.14) becomes

$$\Delta\ddot{q}_1 = \left(-\alpha^2\dot{q}_1^2 + \left(\beta + \sin^2 q_1\right)\dot{q}_2^2 + 2\alpha\cos q_1\dot{q}_1\dot{q}_2\right)\sin q_1 \cos q_1 \\ + \beta\omega_0^2\sin q_1 + \omega_0^2\sin^3 q_1 - \alpha\gamma\omega_0^2\cos q_1 u, \tag{2.15}$$

$$\Delta\ddot{q}_2 = \alpha\left(\dot{q}_1^2 - \cos^2 q_1\dot{q}_2^2\right)\sin q_1 - \frac{1}{2}\left(2\dot{q}_1\dot{q}_2 + \alpha\omega_0^2\right)\sin 2q_1 + \omega_0^2\gamma u, \tag{2.16}$$

with $\Delta(q_1) = \mathcal{Q}(\beta + \sin^2 q_1 - \alpha^2\cos^2 q_1)$. Notice that (2.15)-(2.16) is a fourth-order non-linear system strongly coupled and the coordinate q_2 is cyclic. Remind that, a coordinate is said to be cyclic when it does not appear in the Lagrangian [7], and therefore the corresponding momentum is a constant of motion for the unforced system ($u = 0$). The latter means that the system is completely integrable. Roughly speaking, a completely integrable system is a system with the same number of integrals of motion and degrees of freedom.

2.1.1. Partial Linearisation

A common practise in control design, in order to reduce the complexity of the equations of motion, is to introduce a (non-linear) change in the control input to linearise the actuated equation (2.13), or (2.16), in the acceleration of the q_2-coordinate. Thus, from (2.16) and defining

$$\omega_0^2\gamma u := \Delta\bar{u} - \alpha\left(\dot{q}_1^2 - \cos^2 q_1\dot{q}_2^2\right)\sin q_1 + \frac{1}{2}\left(2\dot{q}_1\dot{q}_2 + \alpha\omega_0^2\right)\sin 2q_1, \tag{2.17}$$

system (2.12)-(2.13) becomes

$$\ddot{q}_1 - \frac{1}{2}\sin 2q_1\dot{q}_2^2 - a\sin q_1 = -b\cos q_1\bar{u}, \tag{2.18}$$

$$\ddot{q}_2 = \bar{u},$$

where $a := \omega_0^2$, $b := \alpha$, and $\bar{u}$ is the new control input. Notice that $(q,\dot{q},u) \rightsquigarrow (q,\dot{q},\bar{u})$ is a globally defined change of coordinates and control. The system in $(q,\dot{q},\bar{u})$-coordinates can be

seen as an Euler-Lagrange system where, through the Routh transformation [7], the energy is an integral of motion of the unforced system ($\bar{u} = 0$) and reads

$$\mathcal{E}(q, \dot{q}) = \frac{1}{2}\|\dot{q}\|_2^2 + \frac{1}{4}\cos 2q_1 \dot{q}_2^2 + a(\cos q_1 - 1). \tag{2.19}$$

Notice that the effective potential has been added because $\dot{q}_2$ is the another integral of motion. Therefore (2.18) it is a new *conservative* system, completely integrable [5].

2.2. Hamilton Equations

Formally the Hamilton equations can be derived through the Legendre transformation [5] defining the conjugate momenta as $p := \partial \mathcal{L}/\partial \dot{q}$. Thus, the Hamilton function becomes

$$\mathcal{H}(q, p) := \dot{q}^\top p - \mathcal{L}(q, p). \tag{2.20}$$

Notice that for a Lagrangian given by (2.7) then $p = M(q)\dot{q}$ with $\mathcal{L}(q, p)$ being the Lagrange function in the Hamiltonian set of coordinates $(q, p) \in \mathcal{S}^1 \times \mathcal{S}^1 \times \mathbb{R}^2$. Nevertheless, since the system is autonomous then the Hamiltonian function coincides with the internal energy of the system, and so it can be easily computed as

$$\mathcal{H}(q, p) = \frac{1}{2}p^\top M(q)^{-1} p + \mathcal{U}(q). \tag{2.21}$$

The Hamilton equations [5] are

$$\dot{q} = \frac{\partial \mathcal{H}}{\partial p},$$

$$\dot{p} = -\frac{\partial \mathcal{H}}{\partial q} + \begin{bmatrix} 0 \\ \tau \end{bmatrix}, \tag{2.22}$$

and for Furuta's pendulum become

$$\dot{q}_1 = \frac{1}{\Delta}\left[\left(\beta + \sin^2 q_1\right)p_1 - \alpha p_2 \cos q_1\right], \tag{2.23}$$

$$\dot{q}_2 = \frac{1}{\Delta}\left[p_2 - \alpha p_1 \cos q_1\right], \tag{2.24}$$

$$\dot{p}_1 = \frac{1}{2\Delta^2}\mathcal{I}\left[\left(\beta + \sin^2 q_1\right)p_1^2 + p_2^2 - 2\alpha p_1 p_2 \cos q_1\right] \cdot \left[\left(1 + \alpha^2\right)\sin 2q_1\right]$$
$$+ \mathcal{I}\omega_0^2 \sin q_1 - \frac{1}{2\Delta}\left[p_1^2 \sin 2q_1 + 2\alpha p_1 p_2 \sin q_1\right], \tag{2.25}$$

$$\dot{p}_2 = \mathcal{I}\gamma\omega_0^2 u, \tag{2.26}$$

with $\Delta(q_1) = \mathcal{I}(\beta + \sin^2 q_1 - \alpha^2 \cos^2 q_1)$.

Table 1: Physical parameter values of the laboratory pendulum.

Parameter	Value	Unit
Mass of the pendulum	0.0679	Kg
Length of the pendulum	0.14	m
Length of the arm	0.235	m
Mass of the arm	0.2869	Kg
Inertia of the motor	0.0012	Kg m^2
Constant torque	7.4	N m/v

2.3. Non-Conservative Friction Torques

In the previous sections a *conservative* mechanical model was derived using Classical Mechanics. In this section, we add the *non-conservative* torques due to friction in the real system. To this end, we focus the analysis in our available laboratory Furuta's pendulum, which was shown in **Figure** 1. The physical values of the parameters of this pendulum are shown in **Table** 1 (see [8] for details).

The friction phenomena in this pendulum come up in the joints. The pendulum joint is composed by a small ball bearing in which friction can be well modelled by a damping constant. Otherwise, in the arm joint a DC-motor and a slip ring are directly coupled (see Appendix A), and therefore the friction is not a damping-like torque, moreover it is strongly non-linear. In what follows of this section we derive the friction models designed in both joints.

Pendulum Joint

The *non-conservative* torques due to a natural damping is introduced through Rayleigh's dissipation function, namely $\mathfrak{D}(q)$. So, calling $\bar{c}_p$ to the damping constant then, the *non-conservative* torque, namely τ_d, becomes

$$\tau_d = \frac{d\mathfrak{D}(\dot{q}_1)}{d\dot{q}_1} = \frac{d}{d\dot{q}_1}\left(\frac{1}{2}\bar{c}_p\dot{q}_1^2\right) = \bar{c}_p\dot{q}_1. \tag{2.27}$$

This constant is easily estimated measuring the damping of a free-motion experiment yielding $\bar{c}_p = 0.0000226$. The small value is due to the small ball bearing and so it could be neglected.

Arm Joint

A carefully experimental study reveals that the friction torque in the motor shaft is not only a natural damping [8]. In fact, static and dynamic friction torques have to be taken into account in order to have a good estimation of the friction torque in this joint. To this end, a dynamic

friction model, called LuGre (from <u>L</u>und Institute and <u>Gre</u>noble Laboratory) model [9–11], is used. The dynamical equations of this model are

$$\frac{dz}{dt} = v - \frac{\sigma_0}{g(v)} z |v|, \tag{2.28}$$

$$g(v) = \tau_c + (\tau_s - \tau_c) e^{-(v/v_s)^2}, \tag{2.29}$$

$$\tau_r = \sigma_0 z + \sigma_1 \frac{dz}{dt} + \sigma_2 v, \tag{2.30}$$

where the parameters are defined as

τ_r: estimated friction torque,

τ_c: Coulomb friction torque,

τ_s: static friction torque,

$\sigma_{0,1}$: internal model parameters,

σ_2: dynamic friction constant,

v_s: Stribeck's velocity,

z: internal state of the model,

v: velocity ($\dot{q}_2 = \dot{\varphi}$).

Next, we give all the necessary steps to identify the parameters of the model (2.28)–(2.30). The physical meaning of these parameters is thoroughly described in [9]. Other kind of dynamical models for friction compensation can be also considered as in [12, 13]. Nevertheless, as we will show further, this model gave us a precise estimation. This good estimation allows to compensate the friction in this joint and, therefore, it gives us the chance of designing control strategies as if it was a *conservative* system. To this end, in [9] some identification methods are suggested but, with the base given in [9] and our experience we have designed our own identification methodology. Thus, the identification of these parameters can be made in four steps corresponding to four different experiments, respectively. Obviously, to identify all the parameters the pendulum is decoupled from the system, i.e. it only remains the arm plus motor. We briefly comment the four experiments designed.

Step 1 (identification of τ_c and σ_2). We use here a standard least-square algorithm to identify τ_c and σ_2. We consider the first-order system given by

$$\mathcal{J}_a \frac{dv}{dt} = \tau - \tau_c \operatorname{sgn}(v) - \sigma_2 v. \tag{2.31}$$

A side-effect parameter $\mathcal{J}_a$ can also be estimated with this experiment. Notice that, (2.31) becomes linear in the parameters. Thus, we introduce in the input τ a square signal (open-loop) and then we measure the angular position of the arm, as shown in Figure 3 (top-left). The results after applying a least-square algorithm to (2.31) are shown in Table 2. Notice, from Table 2, the asymmetry which means that the identification is different for positive and negatives velocities. The MATLAB code for the positive case of $v := \dot{q}_2$ is given in Appendix B.1.
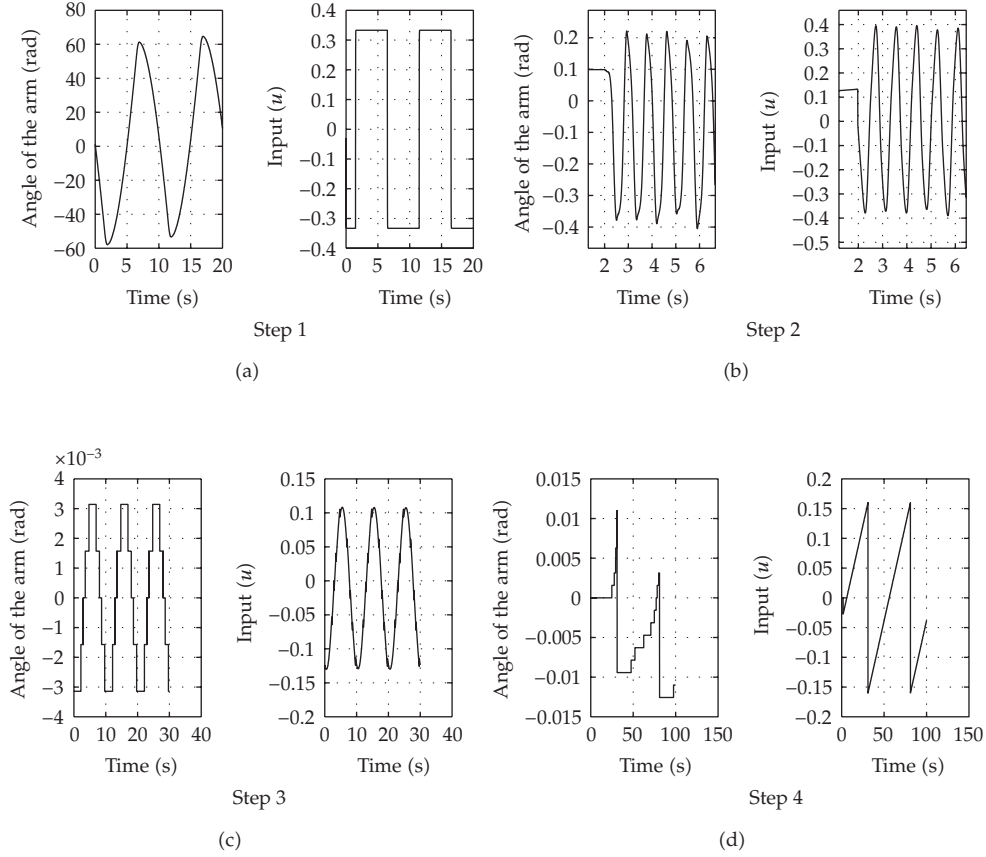
Figure 3: Experiments for Steps 1–4.

Step 2 (identification of τ_s). The static friction is identified forcing a limit cycle in the system. To this end, we feedback the angle of the motor with a slight proportional integral (PI) action, of the form $u = k_1 e + k_2 \int_0^t e(\tau) d\tau$, with $k_{1,2}$ gains and where e is the error between the measured state and the desired reference. The values are directly identified from Figure 3 (top-right) and the results are in Table 2. Notice again the asymmetric behaviour of the system. The MATLAB code is given in Appendix B.2.

Step 3 (identification of v_s). To identify Stribeck's velocity we first need to know the parameters $\mathcal{Q}_a$, τ_c, and τ_s, which are known after steps above. The experiment consists in forcing a very low speed in the motor shaft (open-loop). Thus, the fictitious state z can be approximated, from (2.28), by

$$\frac{dz}{dt} \approx 0 \implies z \approx \frac{g(v)}{\sigma_0} \frac{v}{|v|}, \qquad (2.32)$$

Table 2: Parameters of the LuGre model identified.

Experiment	Parameter	(+)	(−)
	$\mathcal{J}_a$	0.0003	0.0004
Step 1	τ_c/K	0.2238	0.2308
	σ_2	0.0022	0.0020
Step 2	τ_s/K	0.3960	0.3789
Step 3	v_s	0.9297	0.9725
Step 4	σ_0	30	
	σ_1	0.2389	

and then, from (2.30) the torque can be approximated by

$$\tau_r \approx \left(\tau_c + (\tau_s - \tau_c)e^{-(v/v_s)^2} \right) \operatorname{sgn}(v), \qquad (2.33)$$

which additionally allows to linearise the equation

$$\mathcal{J}_a \frac{dv}{dt} = \tau - \tau_r, \qquad (2.34)$$

as

$$v^2(t) = v_s^2 \ln \gamma(t), \qquad (2.35)$$

where γ is given by

$$\gamma = \frac{\tau_s - \tau_c}{(\tau - \mathcal{J}_a(dv/dt)) \operatorname{sgn}(v) - \tau_c}. \qquad (2.36)$$

The approximation is only valid if $\tau_s > \tau_c$ and $\gamma \geqslant 1$. A sinusoidal input torque (open-loop) allows us to estimate the value of γ and so v_s. The experiment is shown on Figure 3 (bottom-left) and the results are shown in Table 2 yielding, again, an asymmetric behaviour. The MATLAB code for positive values of $v := \dot{q}_2$ is given in Appendix B.3.

Step 4 (identification of σ_0 and σ_1). The parameter σ_0 is estimated forcing a ramp in the input (open-loop) such that the angle cross by zero. The maximal and minimal values of the ramp should be around $\pm 0.8 \, \tau_s$ and the slope small enough. The value of σ_0 can be estimated with $\sigma_0 \approx \Delta\tau/\Delta\varphi$. The experiment is shown in Figure 3 (bottom-right).

To estimate σ_1 we assume small displacements and $|z| \ll g(v)$. Thus, $\dot{z} = v = \dot{q}_2$ from (2.28), and then we use a linear second-order model given by

$$\mathcal{J}_a \ddot{q}_2 + (\sigma_1 + \sigma_2)\dot{q}_2 + \sigma_0 (q_2 + q_2(0)) = \tau. \qquad (2.37)$$
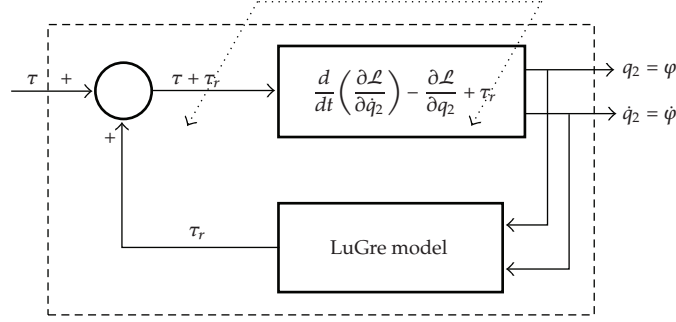
Figure 4: Friction compensation loop.

Therefore, since it is a second-order linear system the value of σ_1 is easily estimated imposing

$$\sigma_1 = 2\zeta\sqrt{\sigma_0 \mathcal{J}_a} - \sigma_2, \tag{2.38}$$

where ζ is the damping constant of the linear model and σ_2 and $\mathcal{J}_a$ are known from **Step 1**. The MATLAB code is given in **Appendix B.4**.

2.4. The Complete Model

Once the friction torques have been estimated we add them to the model in the Lagrangian framework, being dual for the Hamiltonian case. Thus, the Euler-Lagrange equations with the *non-conservative* torques included become

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_1}\right) - \frac{\partial \mathcal{L}}{\partial q_1} + \tau_d = 0, \tag{2.39}$$

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_2}\right) - \frac{\partial \mathcal{L}}{\partial q_2} + \tau_r = \tau, \tag{2.40}$$

where again we normalise dividing by $\mathcal{J}$ with $c_p := \bar{c}_p/\mathcal{J}$. Then, the differences from (2.10) are, on one hand, the matrix $C(q, \dot{q})$ which turns into

$$C(q, \dot{q}) = \mathcal{J}\begin{bmatrix} c_p & -\dot{q}_2 \sin q_1 \cos q_1 \\ -\alpha\dot{q}_1 \sin q_1 + \dot{q}_2 \sin q_1 \cos q_1 & \dot{q}_1 \sin q_1 \cos q_1 \end{bmatrix}, \tag{2.41}$$

and on the other hand the torque is reduced in $\tau_r = K u_r$, where τ_r is estimated by equations (2.28)–(2.30) with $v := \dot{q}_2$.

After adding the *non-conservative* friction torques the model is a better approximation of reality. Nevertheless, to easy the computations of the non-linear control laws, we use the friction models to compensate the friction instead of modelling it. In fact, this is the success of our model in practise. However, remind that the system has two degrees of freedoms and

just one actuation torque, which means that we can compensate the friction torques only in the actuated coordinate q_2. This can be easily seen in (2.39)-(2.40). Thus, in (2.40) the friction torque τ_r can be compensated adding it to the input torque (positive feedback). Otherwise, in the unactuated equation (2.39) we cannot compensate anything at all, since we do not have an independent input. Recall that our objective was to approximate in practise as well as possible to the *conservative* model given by the equations (2.9), or equivalently (2.22). In Figure 4 we show how the added torque compensates the friction. In the pendulum joint, as commented above, we cannot compensate τ_d in (2.39). Nevertheless, by construction the friction torque can be neglected, i.e. $\tau_d \approx 0$, because we use a small enough ball bearing for the joint (recall that $\bar{c}_p \approx 2.26 \times 10^{-5}$).

Finally, after compensating the *non-conservative* torque then, the completed model is a good approximation to the *conservative* model given by (2.9). The reader will realise this good approximation in the next section, where most of all the non-linear controllers were designed with (2.9), or equivalently (2.22). Moreover, in most of the controllers *conservative* quantities were used for the design stage.

3. A Survey of Non-Linear Control Strategies

In this section we survey all of non-linear strategies designed and experimentally tested in the available laboratory pendulum shown in Figure 1. We recall that the paper is focused on the design and experimental results on our available Furuta's pendulum. Thus, other results will be referenced and briefly discussed because they were not experimentally tested on our pendulum.

3.1. Swing-Up Strategies

As commented in the introduction by swing-up we mean the non-linear control problem to swing the pendulum up from the rest to the upright position, the so-called inverted position. The control strategy proposed in [14–16] was based on the Speed-Gradient method, in fact just the finite form of the Speed-Pseudogradient Algorithm proposed in [17, 18] was used. Briefly, we comment the basic idea. Consider the time-invariant, affine-in-control system $\dot{x} = f(x) + g(x)u$ (in our system either (2.15)-(2.16) or (2.23)-(2.26)) with output $y = h(x)$, where $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^l$, and the input $u(t) \in \mathbb{R}^m$. Consider also the control objective $y(t) \rightarrow 0$ when $t \rightarrow \infty$; this control objective can be written as an objective function $Q(x) = (1/2)\|h(x)\|^2$. For this objective function the Speed-Pseudogradient Algorithm in finite form reads

$$u \propto -\left(\frac{\partial h(x)}{\partial x}g(x)\right)^{\top} h(x), \tag{3.1}$$

which can be derived calculating $\dot{Q}(x)$ along trajectories of the system and noting that $h(x)$ are independent invariants for the unforced system ($u = 0$). Stability properties of algorithm (3.1) are described in [17]. In essence, the objective $Q(x)$ plays the role of a Lyapunov function with a closed trajectory as a limit set and then, roughly speaking, if $\dot{Q}(x) \leqslant 0$ the goal will be achieved. Some examples of the application can be seen in [17, 19, 20]. In

(a)

(b)

(c)

(d)
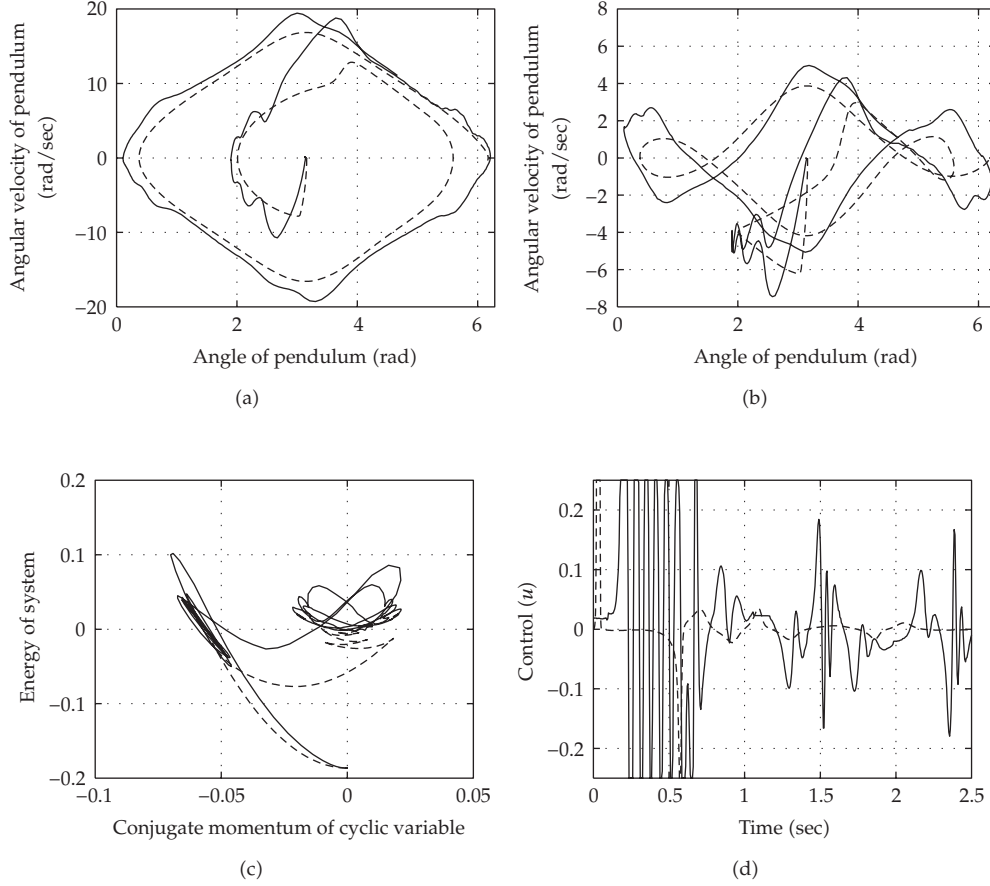
Figure 5: "Solid" experiment. "Dashed" simulation.

the case of study, selecting the objective function as the square of the total system energy $Q := Q_1 = 1/2(\gamma_1 \mathcal{H})^2$ then, the trajectories of the mechanical system will tend towards the surface (limit set) $\{(q_1, p_1, p_2) \in \mathcal{S}^1 \times \mathbb{R}^2 \mid \mathcal{H} = 0\}$, as in [18]. This aim does not guarantee that the pendulum will pass through the origin. The objective function was modified adding a new objective function Q_2 so that the trajectories tend towards the closed trajectory (limit set) $\{Q_1 = 0\} \cap \{Q_2 = 0\}$. The Hamiltonian structure of the system helped us to find the Q_2 function needed. In fact, since the system is symmetric w.r.t. the angle q_2, the conjugate momentum p_2 is a constant of motion for the unforced system, which is straightforward from (2.26). Recall that q_2 is a cyclic coordinate, which is closely related with Noether's theorem [5]. Thus, it is reasonable to choose the objective function as $Q := 1/2(\gamma_1 \mathcal{H})^2 + 1/2(\gamma_2 p_2)^2$, where $\gamma_{1,2}$ are positive control gains and p_2 is the conjugate momentum of the coordinate q_2. The output vector reads $y = h(x) = (\gamma_1 \mathcal{H}, \gamma_2 p_2)^\top$. The objective closed trajectory $\{Q_1 = 0\} \cap \{Q_2 = 0\}$ is the *homoclinic* orbit for Furuta's pendulum [21]. Thus, the objective function for Fradkov's Speed-Gradient method includes not only the energy but also the arm momentum, which are both conserved quantities for the unforced system. To show the fidelity of the model proposed we show Figure 5 from [15, 16], where we compare

the experiment (solid line) with the simulation (dashed line) with the controller proposed. Another proof of the fidelity of the *conservative* model was shown in [14] where we developed the strategy through Routh's function [7], which embedded both *conservative* quantities given before. A video of the experiment can be displayed from http://www.esi2.us.es/~jaar/investiga_e.htm.

Remark 3.1. We underscore that the success of the proposed strategy relies on the fact that the experimental model is *quasi-conservative*. In fact, the control strategy is based on both *conservative* quantities, energy and momentum. Without that *conservative* property the control strategy would fail.

Literature Review

The control problem of swinging the pendulum up was firstly solved by Åström and Furuta with a reduced system model [1, 4]. The control strategy was based on neglecting the reaction torques from the pendulum to the arm. This allows greatly simplify to a second-order model in $(q_1, \dot{q}_1)$-coordinates, and with the help of Fradkov's Speed-Gradient method the desired energy injection was easily computed [22]. However, we remind that it was based on a simplifying assumption. Moreover, the arm speed must be low for the success of any stabiliser at the switching time, but notice that this speed is not considered in the dimension 2 control law, i.e. designed with the second-order model. Fortunately, for standard pendula with the usual initial conditions (the lower position with no velocity) the dimension 2 control law behaves well. Nevertheless, if the reaction torques from the pendulum to the arm cannot be neglected, for example, either due to the mass of the pendulum or to the low friction on the arm, Åström and Furuta's law fails. Some conditions of this failed attempt case were shown in [23], where conditions to induce the approximation to fail, were tested. Moreover, our proposed control law keeps the nice properties of Åström-Furuta's one when compared with other swing-up control laws based on a fourth-order model, i.e. our swing-up control law is able to accomplish the goal with *arbitrary* small control signal magnitude (unlike [24, 25]) and, the control signal converges to zero as the *homoclinic* orbit is reached (unlike [26]).

3.2. *Non-Linear Stabilisers*

In the previous section we show the non-linear strategies to swing the pendulum up to the upright position. Those strategies stabilise the *homoclinic* orbit, which passes through the origin of the state space. Unfortunately, the origin is a saddle fixed point in closed loop and then, to solve the whole problem, that is, swing-up and stabilisation, a balancing strategy (stabiliser) has to be designed to "catch" the pendulum at the upright position. Fortunately, the swing-up strategy was designed such that the pendulum passes through the origin with $(q_1(t), \dot{q}_1(t), \dot{q}_2(t)) \rightarrow 0$, and $u(t) \rightarrow 0$. The latter allows to design the balancing stabiliser separately and then, to propose a switching strategy between both, i.e. the swing-up controller and the stabiliser. Again, the *conservative* model of the pendulum allowed to design non-linear stabilisers using invariants and/or conserved quantities. In this section we describe all the non-linear stabilisers designed using the *conservative* model and, moreover, successfully tested through experiments.

3.2.1. Energy Shaping

A *conservative* system allows, in a natural fashion, to be controlled by *energy-shaping* methods. In fact, a non-linear stabiliser was designed through the *energy-shaping* methodology called Interconnection and Damping Passivity-Based Control. This methodology was introduced in [27] to regulate the position of underactuated mechanical systems of the form (2.22) with total energy given by (2.21). The main idea is to force the closed-loop structure of the system to a desired one. This procedure is commonly called "matching." In fact, for mechanical systems the proposed closed-loop structure takes the Port-Controlled Hamiltonian form [6] given by

$$
\begin{bmatrix} \dot{q} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} 0 & M^{-1}M_d \\ -M_d M^{-1} & J_2 \end{bmatrix} \begin{bmatrix} \dfrac{\partial \mathcal{H}_d}{\partial q} \\ \dfrac{\partial \mathcal{H}_d}{\partial p} \end{bmatrix},
\tag{3.2}
$$

where in our case of study $(q, p) \in \mathcal{S}^1 \times \mathcal{S}^1 \times \mathbb{R}^2$, $M_d(q)$ is the desired inertia matrix, $\mathcal{V}_d(q)$ is the desired potential function and so, the total energy function of the closed loop reads $\mathcal{H}_d = (1/2)p^\top M_d^{-1}(q)p + \mathcal{V}_d(q)$. Subindex "d" denotes "desired" which means in closed-loop. Finally $J_2(q, p) \in \mathbb{R}^{2 \times 2}$ is a skew-symmetric matrix free for the designer. The design concludes after a damping injection stage. Stability properties require a positive definite M_d in a neighborhood of the desired equilibrium $q^\star$, with $q^\star = \arg \min \mathcal{V}_d(q)$, and then $(q^\star, 0)$ is a stable equilibrium point of the closed-loop system (3.2), with Lyapunov function $\mathcal{H}_d$.

Since the system is underactuated the "matching" procedure provides on one hand, a set of algebraic constraints (in the actuated part) and, on the other hand, a set of Partial Differential Equations (PDEs) on M_d and $\mathcal{V}_d$ (in the unactuated part). The PDEs arise for M_d are called kinetic energy PDEs and the ones for $\mathcal{V}_d$ are called potential energy PDEs. Clearly, the success of this methodology relies on the possibility of solving those PDEs, which is usually a daunting task. In fact, for the case of the Furuta's pendulum a new mathematical machinery was needed to solve them. First, the reparametrization of the free skew-symmetric matrix J_2 given in [28] allows to transform the kinetic energy PDE in a set of first-order ordinary differential equations depending only on the q_1-coordinate. Second, the change of coordinates proposed in [29, 30] allows to simplify, drastically, the *non-homogeneous* term of the kinetic energy PDEs. The complete solution was reported in [31] and the experimental test will be commented further in comparison with others strategies.

Remark 3.2. Again the success of the proposed strategy relies on the fact that the model for the design is *conservative*. The extension of this methodology considering natural damping was given in [32, 33]. Unfortunately, in Furuta's pendulum case and considering natural damping torques no solutions were found so far.

Literature Review

Other different theoretical solutions have been reported using *energy-shaping* approaches. In [2] the Controlled Lagrangians methodology was used for stabilisation, through just a kinetic energy shaping (see also [34]). In [35] a way to transform the PDEs was proposed providing another solution for Furuta's pendulum. Closely related it is the non-linear controller based

on a Lyapunov control approach given in [36]. All of these different solutions have not been experimentally tested, so far.

3.2.2. Forwarding

Cascade non-linear control techniques have attracted the interest of many researchers (see [37–39] and references therein), mainly because they provide great simplifications and/or reductions in the non-linear control design stage. *Forwarding* is a passivity approach applied to cascade non-linear systems with feedforward paths. In [38] the system is treated as a feedforward connection of two cascade subsystems through a non-linear vector function. Assuming that each subsystem has an associated Lyapunov function, separately, the main difficulty is the need to construct an additive cross term to get a Lyapunov function for the whole cascade. The cross term is constructed through the solution of a PDE. The solution of the PDE provides an invariant manifold decomposition. The manifold allows together with a change of coordinates to construct a compound scaled Lyapunov function solving the stabilisation problem in closed form [37]. So, to solve another PDE is needed. Finally, noting that the system is in Jurdjević-Quinn form [40], then a damping stabiliser solves the whole control problem. We apply this technique to the *conservative* model (2.18) and the design and experiments were reported in [41, 42]. It is worth noting that the main result of [41] was not the non-linear stabiliser designed through *Forwarding*. The main result was that the Lyapunov function associated to the *Forwarding* design was used together with the one of the swing-up controller, to design a Lyapunov-like switching strategy to solve the whole problem, i.e. the *almost* global stabilisation of the inverted position, both in theory and practise. The usual switching strategy to solve this global control problem is, to compound the swing-up controller with a Linear Quadratic Regulator (LQR) to stabilise the origin [8, 43], i.e. a state-feedback $u := Kx$ with state $x \in \mathbb{R}^n$ and a constant gain matrix K. In our available pendulum we realised that the Lyapunov function associated to the LQR cannot be used to design the switching strategy because the estimate of the Domain of Attraction (DoA) is too small. By too small we means that the level of the noise from the measurements is higher than the size of the largest admissible level set of the estimate of the DoA, and then the switching controller fails to commute.

Remark 3.3. The *Forwarding* needs the solution of a PDE. The solution of the PDE is an invariant which provides and invariant manifold decomposition for the closed-loop system. Thus, that conservative property was crucial again.

Literature Review

In [3] another solution was given using the non-linear cascade technique called *Backstepping* [44]. The approximated solution given relies on a nontrivial change of coordinates and it has not been experimentally tested so far.

3.2.3. Input-Output Feedback Linearisation

This technique was a step ahead after the development of the well-known *State Feedback Linearisation*, which linearises the state of the system from the control input through a

non-trivial change of coordinates and control. This change of coordinates comes up from the solution of a set of PDEs. The extension to this approach was made through the *Input-Output Feedback Linearisation* which linearises the input-output response of the system [45, 46]. In the cases in which it is not possible to linearise the whole system by a static state-feedback, a partial feedback linearisation can still be possible. This partial feedback linearisation transforms the system dynamics into two parts: the external dynamics, which become linear; the internal non-linear dynamics (locally characterised by the zero dynamics [45]). As proposed in [45] a possibility to control non-linear systems is to find an output, called "fictitious" such that the system linearised through that output is *minimum phase* [45]. Thus, in [47, 48] we proposed a nontrivial and constructive "fictitious" output through which it is possible to achieve the asymptotic stability with (locally) exponential convergence of the origin, for a class of underactuated mechanical systems with underactuation degree one. The generalisation to underactuation degree larger than one is given in [49]. The approach was based on classical *Input-Output Feedback Linearisation* and Lyapunov redesign [50], yielding excellent results in theory and experiments. In particular, for Furuta's pendulum the proposed output, namely y, reads

$$y := \dot{q}_2 + k_1 \cos q_1 \dot{q}_1 + k_2 \sin q_1, \tag{3.3}$$

where $k_{1,2}$ are positive control gains. The model used for the design was (2.18). In fact, the proposed output makes the system to be *exponential minimum-phase* [46], i.e. the equilibrium of the zero dynamics is hyperbolic. As it is well-known this property provides the system with robustness (to unmodelled dynamics, disturbances) [50]. The experiments in the available Furuta's pendulum equipment were reported in [51].

Remark 3.4. We would like to underscore that the *exponential minimum phase* property is not achieved by most of non-linear control design techniques. For example in *energy-shaping* and *Forwarding* techniques the closed-loop system is *weakly minimum phase* w.r.t. its output, inherent fact of those methodologies by their passivity properties (see [52] for further details).

Remark 3.5. As in the previous cases, a crucial *conservative* quantity allows to design the stabiliser. In this case the energy of the zero dynamics was the conserved quantity.

Literature Review

Close to *Input-Output Feedback Linearisation* is the *Flatness* approach. In the case of a system with a single control input then, it is said to be *differentially flat* if and only if it is fully linearisable via static state-feedback (see e.g. [53, 54]). A *Flatness* controller for a linearised model of Furuta's pendulum was given in [55] but it was based on a linear model, through the *Jacobian*, since the non-linear model is not *differentially flat* w.r.t. any output.

3.2.4. Singular Perturbations

A deep and fair review of the use of this technique to control systems with different time scales is given in [50] and references therein. A *singularly perturbed* system is a system in which its own structure makes different coordinates evolve in different time scales. In the
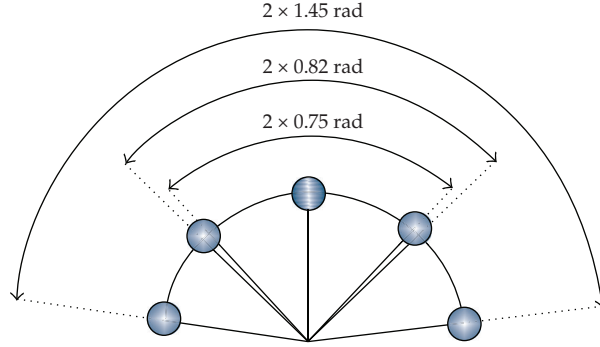
Figure 6: Experimental region of attraction.

beginning, it was a technique to analyse the stability of equilibria, because the time-scale separation is a property of the own structure of the system. Nevertheless, if the system has not this time-scale separation itself, we can be able to propose a stabiliser to force that separation-like property and then, to analyse the stability of the equilibrium. This is the case proposed in [56, 57] for Furuta's pendulum. In fact, we even make relaxations to the *Input-Output Feedback Linearisation* control law described above, and then we use the *Singular Perturbation* theory to analyse the stability. The theoretical and experimental result was reported in [56, 57], where a fair comparison with another techniques was also reported. Moreover, we designed an experiment to measure the largest attraction basin. In fact, to the best of our knowledge, this is the largest region of attraction experimental tested so far, to stabilise this kind of pendulum. We reproduce, and comment briefly, here Figure 6 from [57], where the results of the experiments were collected.

In the experiment we forced to the system to start from an initial condition as $(\theta(0), \dot{\theta}(0), \varphi(0), \dot{\varphi}(0)) = (\theta^\star, 0, 0, 0)$. The idea was to find out the largest $\theta^\star$ for which the origin (the upright position of the pendulum) was stabilised with a given controller. The controller design shows that the maximum theoretical value of $\theta^\star$ for the controller based on *singular perturbations* is defined through a condition in the controller gains, and therefore can be tuned (see [57]). In [2, 31] the maximum theoretical $\theta^\star$ depends on physical parameters of the own pendulum, and therefore the formulas are not tunable, giving rise in our pendulum to a maximum approximately the half of the value of the one given by the *singularly perturbed* approach. The best experimental results for each controller are collected in Figure 6 and give rise to $\theta^\star = 1.45$ rad with the *singular-perturbed* approach; $\theta^\star = 0.82$ rad with the *energy-shaping* approach; $\theta^\star = 0.75$ rad with a Linear Quadratic Regulator. Unfortunately, the system saturates and it was not possible to enlarge even more this practical region of attraction (see [57]). Some videos of the experiments can be displayed from http://www.esi2.us.es/~jaar/investiga_e.htm.

3.3. Global Stabiliser: A Challenge

In the previous sections we have shown the non-linear strategies to swing-up and stabilisation of the pendulum at the upright position. To solve the whole problem,

i.e. swing-up and stabilization, a switching strategy was designed to "catch" the pendulum at the upright position properly. As commented before, fortunately, the swing-up strategy was designed such that the pendulum passes through the origin of the state space. This is the practical engineering solution to the whole problem. But one might wonder, if it would be possible to solve the whole problem with only one control strategy. We cannot close this section without describing this mathematical and still *open* problem that has attracted the attention of many researchers. Thus, suppose that we are interested in solving the whole problem, designing only one smooth state-feedback controller. By smooth we means only with one control strategy without any switch between strategies. Of course, from a practical engineering point of view the problem is actually solved as it has been shown in the previous sections, but from a mathematical point of view it is a challenge problem. On one hand there is an obstacle inherent to the own configuration space of the system, remind that $q \in S^1 \times S^1$, which means that we cannot globally stabilise an isolated equilibrium in the whole state space. All we can do is to stabilise the upper equilibrium from *almost everywhere* in the state space, except possibly a set of zero Lebesgue measure. On the other hand, since the upper equilibrium of the unforced system is unstable then we have to design a static state-feedback that makes that equilibrium asymptotically stable and all the remaining equilibria unstable.

The latter idea, for an underactuated mechanical system, was proposed in [58] in a pendulum-on-a-cart system where *almost* global stability of the equilibrium was proved. For Furuta's pendulum case, we have found a promising controller that makes the work but it was not possible to prove the *almost* global stability of the equilibrium, so far. As it was reported in [8] the static state-feedback reads $u := k_1 \sin(2q_1) + k_2 \cos(q_1)\dot{q}_1 + k_3 \dot{q}_2$, with k_i, $i = 1, \ldots, 3$, control gains, and where we underscore the *unsigned definite* damping-like term. To prove *almost* global stability of the equilibrium, if so, is still a mathematical open problem. However we tested the proposed controller through experiments. In Figure 7 we show a successful experiment, where in the top figure we show the time histories of the position and the velocity of the pendulum and, in the bottom figure the time histories of the control input. Notice that there is no any switch. A video of the experiment can be displayed from http://www.esi2.us.es/~jaar/investiga_e.htm.

3.4. Autonomous Oscillations

Another field of interest for the control community is the generation of "artificial" and autonomous oscillations, i.e. by means of an appropriate control law. To accomplish this in [59] we proposed a methodology based on matching the open-loop system with a generalised hamiltonian system, which is able to exhibit robust oscillations. Thus, the closed-loop system displays robust oscillations associated with a limit cycle. The methodology mimics the *energy-shaping* approach but here, the objective is to reach a limit cycle (invariant limit set), which provides a stable oscillating behaviour, instead of a fixed-point attractor. The desired limit cycle is born through a *supercritical Hopf bifurcation* on an approximated second-order dynamical model and, therefore, for certain values of a control parameter the closed-loop system has a fixed-point attractor, changing to a limit cycle through other values of that control parameter. Successful experiments were reported in [59] and we reproduce in Figure 8 a sequence of snapshots of the experiment for a period of the oscillations of 0.5 second. A video of the experiment can be displayed from http://www.esi2.us.es/~jaar/investiga_e.htm.

(a)

(b)
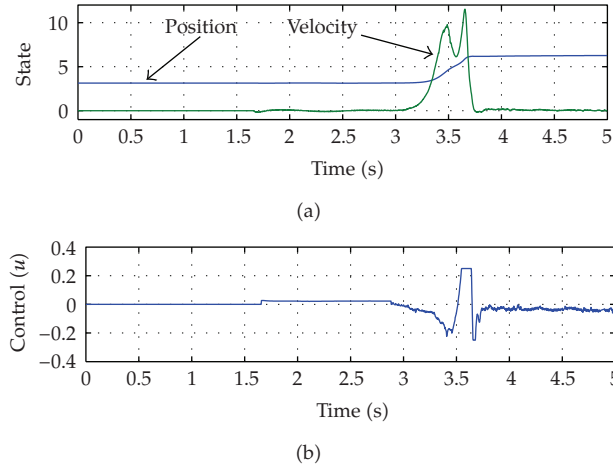
Figure 7: Experimental result with the challenging non-linear stabiliser.
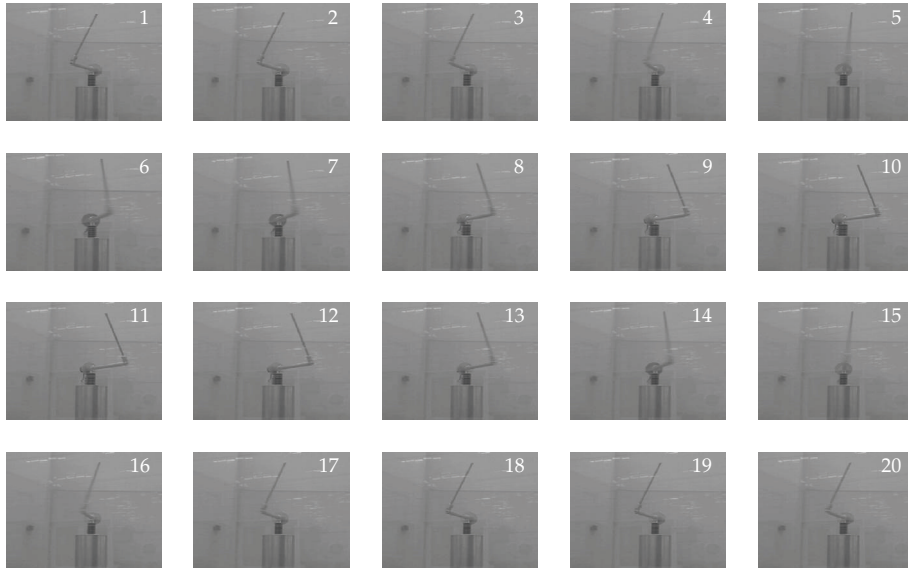
Figure 8: Sequence of snapshots showing one period of the oscillations.

Literature Review

Another approach to generate autonomous oscillations on Furuta's pendulum has been reported in [13, 60] based on virtual holonomic constraints. In our approach we design a closed-loop structure which associated Lyapunov function has a closed trajectory as its limit set (the limit cycle). In [13, 60] a closed trajectory is generated in closed loop and then only local exponential convergence can be guaranteed.

3.5. Bifurcations

This section is devoted to comment briefly some works regarding with bifurcations which have been also analysed and experimentally tested in our laboratory pendulum. The works commented here are mainly focused in the analysis of some non-linear behaviour of the controlled pendulum, rather than in control design. In fact, the controller used to analyse the behaviour is the linear one, i.e. a linear state-feedback stabiliser designed through the *Jacobian*. In [61, 62] a bifurcation map was constructed for Furuta's pendulum with a saturated linear controller, where the bifurcation parameters were the controller gains. Detailed analysis, simulations, and experiments in our available pendulum were reported in [63]. Finally, in [64] global bifurcation analysis was made for Furuta's pendulum. Nevertheless, we underscore that the dynamic friction model presented here was not used. Only a static model to compensate Coulomb's friction was used instead.

Literature Review

In [21] a complete bifurcation map for the Furuta's pendulum is given. Nevertheless, the bifurcation map is for the unforced pendulum and only through simulations with continuation methods.

4. Conclusion

A detailed dynamical model of the available laboratory Furuta's pendulum is provided. The completed model is based on Classical Mechanics augmented with a dynamical model that allows to compensate all those *non-conservatives* torques, arising to a *quasi-conservative* model in practise. A survey of all the non-linear controllers designed with that *quasi-conservative* model and successfully tested through experiments is also reported.

Appendices

A. Experimental Control System

The schematic representation of the full control system is depicted in Figure 9. The laboratory electromechanical system is composed by three subsystems: measurement, power, and control supervision. A brief description of them is given in this appendix.

Measurement Subsystem

Recall that the system has two degrees of freedom corresponding to the angles of rotation of the pendulum and the motor shaft, the angular coordinates θ and φ, respectively. These angles are measured by sensors position (*encoders*) of incremental nonhigh resolution (2000 and 1024 pulses per revolution, resp.). Velocities for each angle, $\dot{\theta}$ and $\dot{\varphi}$, are obtained from suitable filters to the respective angular positions. There is the possibility to obtain the velocity $\dot{\varphi}$ via a tachometer coupled to the motor shaft. The transmission of the signal position of the pendulum is through brushes using a slip-ring system.
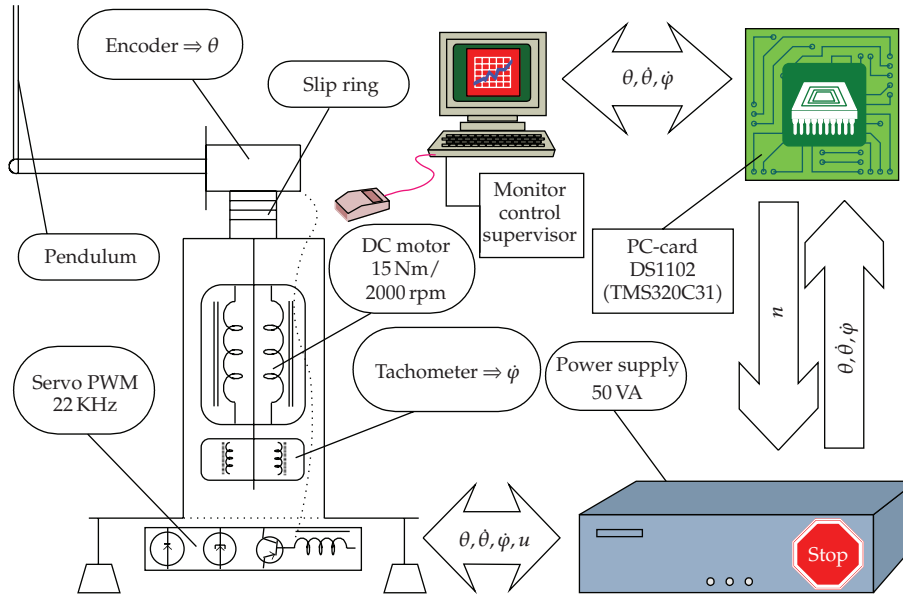
Figure 9: Schematic of the experimental control system.

To avoid problems of *aliasing*, mainly caused by noise, an analog filter (*antialiasing*) is used. The cutoff frequency is at 200 Hz, to filter the signal for the speed sensor (variable $\dot{\varphi}$) before being sampled.

Power Subsystem

This subsystem provided the necessary power to drive the motor. The DC-motor provides a maximum torque of 15 N m at 2000 r.p.m., with a regulated source of 50 VA. The motor is controlled by current through a PWM servo-amplifier at 22 KHz (see Figure 9).

Control and Supervision Subsystem

This subsystem closed the loop. It is composed by a PC and a data card *Dspace* DS1102, based on a Digital Signal Processors (DSP) $TMS320C31$. Among others, We have interfaz AD/DA, digital inputs/ouputs, counters, *incremental encoders* interface, digital filters, and communication interface. The real-time software is provided by *Dspace* running on *Matlab/Simulink*. The minimal sample time allowed is 1 msec.

B. MATLAB

For the sake of completeness, in this appendix we collect the detailed MATLAB code for identification of the dynamic friction model (LuGre).

B.1. Code for Step 1

```
function theta = step1p(x, u, sample)
% 1st experiment of LuGre model
% theta = step1p(position, control, sample)
% Out: theta = [J sigma2 Fc+]
v_min=1; n=length(sample); k=0;        % Initialization
t = sample';                           % Time
u = u'; x = x';                        % Control, Position
T = mean(diff(t));                     % Sample time
v = diff(x)/T; v(n) = v(n-1);          % Velocity
v_dot = diff(v)/T; v_dot(n)=v_dot(n-1); % Acceleration
% Filter
tau = 0.01;                            % Time constant
[num_d, den_d] = c2dm(1, [tau, 1], T, 'tustin');
u_f = filter(num_d, den_d, u);
v_f=filter(num_d, den_d, v);
v_dot_f = filter(num_d,den_d, v_dot);
% Data selection
for i=1:n;
   if v_f(i) > v_min
      k=k+1;
      v(k) = v_f(i);
      v_dot(k) = v_dot_f(i);
      u(k) = u_f(i);
   end;
end; clear u_f v_f v_dot_f
% Least square (Moore-Penrose)
phi_p = [v_dot(1:k),v(1:k),sign(v(1:k))];
theta_p = pinv(phi_p)*u(1:k);
theta = theta_p';                      % Out
```

B.2. Code for Step 2

```
function Fs = step2(u)
% 2nd experiment of LuGre model
% Fs = step2(control)
% Out: Fs = [Fs+,Fs-]
Fs = [max(u) min(u)];                  % Out
```

B.3. Code for Step 3

```
function vs_p = step3p(Fc_p, Fs_p, J, x, u, sample)
% 3rd experiment of LuGre model
% vs+ = step3p(Fc+, Fs+, position, control, sample)
% Out: vs_p
n = length(sample); k=0; q=0;           % Initialization
u=u'; x=x';                             % Control, Position
t = sample';                            % Time
T = mean(diff(t));                      % Sample time
v=diff(x)/T; v(n)=v(n-1);               % Velocity
v_dot=diff(v)/T; v_dot(n)=v_dot(n-1);   % Acceleration
% Data selection
for i=1:n;
    if u(i)>0
        k=k+1;
        v(k)=v(i);
        v_dot(k)=v_dot(i);
        u(k)=u(i);
    end;
end;
% gamma>=1
for i=1:k
    gamma(i)=((Fs_p-Fc_p)/(u(i)-J*v_dot(i))*sign(v(i))-Fc_p);
    if gamma(i)>1
        q=q+1;
        gamma(q)=gamma(i);
        v(q)=v(i)^2;
    end;
end;
% Least square (Moore-Penrose)
phi_p = [log(gamma(1:q))'];
theta_p = pinv(phi_p)*v(1:q);
vs_p = sqrt(theta_p);                   % Out
```

B.4. Code for Step 4

```
function sigma = step4(x, u)
% 4th experiment of LuGre model
% sigma = step4(position, control)
% Out: sigma
global Inertia sigma2
delta = input('Microscopic damping:');  % Initialization
sigma0 = (u(2)-u(1))/(x(2)-x(1));
sigma1 = 2*delta*sqrt(sigma0*Inertia)-sigma2;
sigma = [sigma0 sigma1];                 % Out
```

Acknowledgment

References

[1] K. J. Åström and K. Furuta, "Swinging up a pendulum by energy control," *Automatica*, vol. 36, no. 2, pp. 287–295, 2000.

[2] A. M. Bloch, N. E. Leonard, and J. E. Marsden, "Stabilization of the pendulum on a rotor arm by the method of controlled lagrangians," in *Proceedings of the 39th IEEE Conference on Robotics and Automation*, vol. 1, pp. 500–505, Detroit, Mich, USA, May 1999.

[3] R. Olfati-Saber, "Fixed point controllers and stabilization of the cart-pole system and the rotating pendulum," in *Proceedings of the 38th IEEE Conference on Decision and Control*, pp. 1174–1181, Phoenix, Ariz, USA, December 1999.

[4] M. Wiklund, A. Kristenson, and K. J. Åström, "A new strategy for swinging up an inverted pendulum," in *Proceedings of the 12th IFAC World Congress*, vol. 9, pp. 151–154, Sydney, Australia, 1993.

[5] H. Goldstein, *Classical Mechanics*, Addison-Wesley, New York, NY, USA, 2nd edition, 1980.

[6] A. J. Van der Schaft, L_2-*gain and Passivity Techniques in Nonlinear Control*, Springer, New York, NY, USA, 1989.

[7] D. R. Merkin, *Introduction to the Theory of Stability*, vol. 24 of *Texts in Applied Mathematics*, Springer, New York, NY, USA, 1996.

[8] J. Á. Acosta, *Underactuated nonlinear control systems*, Ph.D. thesis, Department of Automatic Control and Systems Engineering, University of Seville, Seville, Spain, 2004.

[9] F. Altpeter, *Friction modeling, identification and compensation*, Ph.D. thesis, École Polythechnique Fédérale de Lausanne, 1999.

[10] C. Canudas de Wit, "Control of system with dynamic friction," Tech. Rep. INPG-ENSIEG-CNRS, Laboratory of Automatic Control of Grenoble, Grenoble, France, 1999.

[11] C. Canudas de Wit, H. Olsson, K. J. Åström, and P. Lischinsky, "A new model for control of systems with friction," *IEEE Transactions on Automatic Control*, vol. 40, no. 3, pp. 419–425, 1995.

[12] M. Gafvert, "Dynamic model based friction compensation on the Furuta pendulum," in *Proceedings of the IEEE Conference on Control Applications*, vol. 2, pp. 1260–1265, Kohala Coast, Hawaii, USA, 1999.

[13] A. S. Shiriaev and A. Sörensen, "Friction compensation in the Furuta pendulum for stabilizing rotational modes," in *Proceedings of the IEEE Conference on Decision and Control*, vol. 4, pp. 3772–3777, Orlando, FLa, USA, February 2001.

[14] J. Á. Acosta, J. Aracil, and F. Gordillo, "On swinging the Furuta pendulum up based on Routh's reduction and SG method," in *Proceedings of the 9th Mediterranean Conference on Control and Automation*, vol. 1, Dubrovnik, Croatia, June 2001.

[15] J. Á. Acosta, F. Gordillo, and J. Aracil, "A new SG law for swinging the Furuta pendulum up," in *Proceedings of the 5th IFAC Symposium on Nonlinear Control Systems*, pp. 818–823, Saint-Petersbourg, Russia, July 2001.

[16] F. Gordillo, J. Á. Acosta, and J. Aracil, "A new swing-up law for the Furuta pendulum," *International Journal of Control*, vol. 76, no. 8, pp. 836–844, 2003.

[17] A. L. Fradkov and A. Yu. Pogromsky, *Introduction to Control of Oscillations and Chaos*, vol. 35, World Scientific, River Edge, NJ, USA, 1998.

[18] A. L. Fradkov, O. P. Tomchina, and O. L. Nagibina, "Swinging control of rotating pendulum," in *Proceedings of the 3rd IEEE Mediterranean Symposium on Control and Automation*, pp. 347–351, Limassol, Cyprus, July 1995.

[19] A. L. Fradkov, "Swinging control of nonlinear oscillations," *International Journal of Control*, vol. 64, no. 6, pp. 1189–1202, 1996.

[20] A. P. Konjukhov, O. L. Nagibina, and O. P. Tomchina, "Energy based double pendulum control in periodic and chaotic mode," in *Proceedings of the 3rd International Conference on Motion and Vibration Control*, pp. 99–104, Chiba, Japan, 1996.

[21] F. J. Muñoz-Almaraz, E. Freire, and J. Galán-Vioque, "Bifurcation behavior of the Furuta pendulum," *International Journal of Bifurcation and Chaos*, vol. 17, no. 8, pp. 2571–2578, 2000.

[22] P. Andrievskii, P. Guzenko, and A. L. Fradkov, "Control of nonlinear vibrations of mechanical systems via the method of velocity gradient," *Automation and Remote Control*, vol. 57, pp. 456–467, 1996.

[23] J. Á. Acosta, F. Gordillo, and J. Aracil, "Swinging up the Furuta pendulum by speed gradient method," in *Proceedings of the European Control Conference*, pp. 469–474, Porto, Portugal, September 2001.

[24] R. Olfati-Saber, "Cascade normal forms for underactuated mechanical systems," in *Proceedings of the 39th IEEE Conference on Decision and Control*, Sydney, Australia, December 2000.

[25] A. Shiriaev, A. Pogromsky, H. Ludvigsen, and O. Egeland, "On global properties of passivity-based control of an inverted pendulum," *International Journal of Robust and Nonlinear Control*, vol. 10, no. 4, pp. 283–300, 2000.

[26] I. Fantoni and R. Lozano, "Stabilization of the Furuta pendulum around its homoclinic orbit," *International Journal of Control*, vol. 75, no. 6, pp. 390–398, 2002.

[27] R. Ortega, M. W. Spong, F. Gómez-Estern, and G. Blankenstein, "Stabilization of underactuated mechanical systems via Interconnection and damping assignment passivity based control," *IEEE Transactions on Automatic Control*, vol. 47, no. 8, pp. 1218–1233, 2002.

[28] J. Á. Acosta, R. Ortega, A. Astolfi, and A. M. Mahindrakar, "Interconnection and damping assignment passivity—based control of mechanical systems with underactuation degree one," *Transactions on Automatic Control*, vol. 50, no. 12, pp. 1936–1955, 2005.

[29] G. Viola, R. Ortega, R. N. Banavar, J. Á. Acosta, and A. Astolfi, "Total energy shaping control of mechanical systems: simplifying the matching equations via coordinate changes," in *Proceedings of the 3rd IFAC Workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control*, vol. 3, pp. 87–92, Nagoya, Japan, July 2006.

[30] G. Viola, R. N. Banavar, J. Á. Acosta, and A. Astolfi, "Some remarks on interconnection and damping assignment passivity-based control of mechanical systems," in *Taming Heterogeneity and Complexity of Embedded Control*, pp. 721–735, ISTE, 2007.

[31] G. Viola, R. Ortega, R. Banavar, J. Á. Acosta, and A. Astolfi, "Total energy shaping control of mechanical systems: simplifying the matching equations via coordinate changes," *IEEE Transactions on Automatic Control*, vol. 52, no. 6, pp. 1093–1099, 2007.

[32] F. Gómez-Estern and A. J. Van der Schaft, "Physical damping in IDA-PBC controlled underactuated mechanical systems," *European Journal of Control*, vol. 10, no. 5, pp. 451–468, 2004.

[33] F. Gómez-Estern, A. J. van der Schaft, and J. Á. Acosta, "Passivation of underactuated systems with physical damping," in *Proceedings of the 6th IFAC Symposium on Nonlinear Control Systems*, pp. 1–3, Stuttgart, Germany, September 2004.

[34] A. M. Bloch, N. E. Leonard, and J. E. Marsden, "Controlled lagrangians and the stabilization of mechanical systems I: the first matching theorem," *IEEE Transactions on Automatic Control*, vol. 45, no. 12, pp. 2253–2270, 2000.

[35] D. Auckly, L. Kapitanski, and W. White, "Control of nonlinear underactuated systems," *Communications on Pure and Applied Mathematics*, vol. 3, pp. 354–369, 2000.

[36] C. Aguilar-Ibáñez, O. F. Gutierrez, and H. Sossa-Azuela, "Control of the Furuta pendulum by using a Lyapunov function," in *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 6128–6132, San Diego, Calif, USA, December 2006.

[37] F. Mazenc and L. Praly, "Adding integrations, saturated controls, and stabilization for feedforward systems," *IEEE Transactions on Automatic Control*, vol. 41, no. 11, pp. 1559–1578, 1996.

[38] R. Sepulchre, M. Janković, and P. Kokotović, *Constructive Nonlinear Control*, Communications and Control Engineering Series, Springer, London, UK, 1997.

[39] A. R. Teel, "A nonlinear small gain theorem for the analysis of control systems with saturation," *IEEE Transactions on Automatic Control*, vol. 41, no. 9, pp. 1256–1270, 1996.

[40] V. Jurdjevic and J. P. Quinn, "Controllability and stability," *Journal of Differential Equations*, vol. 28, no. 3, pp. 381–389, 1978.

[41] J. Aracil, J. Á. Acosta, and F. Gordillo, "A controller for swinging-up and stabilizing the inverted pendulum," in *Proceedings of the 17th IFAC World Congress*, Seoul, Korea, July 2008.

[42] J. Aracil, J. Á. Acosta, and F. Gordillo, "A nonlinear hybrid controller for swinging-up and stabilizing the Furuta pendulum," submitted.

[43] J. Á. Acosta, J. Aracil, and F. Gordillo, "Nonlinear control strategies for the Furuta pendulum," *Journal in Control and Intelligent Systems*, vol. 29, pp. 101–107, 2001.

[44] M. Krstić, I. Kanellakopoulos, and P. Kokotović, *Nonlinear Adaptative Control Design*, John Wiley & Sons, New York, NY, USA, 1995.

[45] A. Isidori, *Nonlinear Control Systems*, Communications and Control Engineering Series, Springer, Berlin, Germany, 3rd edition, 1995.

[46] S. Sastry, *Nonlinear Systems: Analysis, Stability and Control*, vol. 10 of *Interdisciplinary Applied Mathematics*, Springer, New York, NY, USA, 1999.

[47] J. Á. Acosta and M. López-Martínez, "Constructive feedback linearization of mechanical systems with friction and underactuation degree one," in *Proceedings of the 45th IEEE European Control Conference*, vol. 1, pp. 4255–4260, Kos, Greece, July 2007.

[48] M. López-Martínez and J. Á. Acosta, "Linealización por realimentación constructiva de sistemas mecánicos con grado de subactuación 1 inestables con fricción," *Revista Iberoamericana de Automática e Informática Industrial*, vol. 4, no. 4, pp. 70–79, 2007.

[49] J. Á. Acosta and M. López-Martínez, "Input—output linearization on underactuated mechanical systems," submitted.

[50] H. K. Khalil, *Nonlinear Systems*, Prentice-Hall, Upper Saddle River, NJ, USA, 2002.

[51] J. Á. Acosta and M. López-Martínez, "Constructive feedback linearization of underactuated mechanical systems with 2-DOF," in *Proceedings of the IEEE Conference on Decision and Control and European Control Conference*, pp. 4909–4914, Seville, Spain, December 2005.

[52] C. I. Byrnes, A. Isidori, and J. C. Willems, "Passivity, feedback equivalence, and the global stabilization of minimum phase nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 36, no. 11, pp. 1228–1240, 1991.

[53] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, "A lie-backlund approach to equivalence and flatness of nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 44, no. 5, pp. 922–937, 1999.

[54] M. V. Nieuwstadt, M. Rathinam, and R. M. Murray, "Differential flatness and absolute equivalence," in *Proceedings of the 33rd Conference on Decision and Control*, Lake Buena Vista, Fla, USA, 1995.

[55] C. Aguilar-Ibáñez and H. Sira-Ramírez, "A linear differential flatness approach to controlling the Furuta pendulum," *IMA Journal of Mathematical Control and Information*, vol. 24, no. 1, pp. 31–45, 2007.

[56] J. Á. Acosta and M. López-Martínez, "On singular perturbations of unstable underactuated mechanical systems with underactuation degree ≥ 1," in *Proceedings of the 17th IFAC World Congress*, Seoul, Korea, July 2008.

[57] J. Á. Acosta and M. López-Martínez, "A nonlinear strategy to control unstable underactuated mechanical systems with underactuation ≥ 1. Applications to control augmentations," *Automation & Control Systems Journal*, vol. 2, pp. 13–20, 2009.

[58] D. Angeli, "Almost global stabilization of the inverted pendulum via continuous state feedback," *Automatica*, vol. 37, pp. 1103–1108, 2001.

[59] J. Aracil, F. Gordillo, and J. Á. Acosta, "Stabilization of oscillation in the inverted pendulum," in *Proceedings of the 15th IFAC World Congress*, Barcelona, Spain, July 2002.

[60] L. B. Freidovich, A. Robertsson, R. Johansson, A. Sandberg, and A. S. Shiriaev, "Virtual-holonomic-constraints-based design of stable oscillations of Furuta pendulum: theory and experiments," *IEEE Transactions on Robotics*, vol. 23, no. 4, pp. 827–832, 2007.

[61] D. Pagano, L. Pizarro, and J. Aracil, "Modern applied mathematics techniques in circuits, systems and control," in *Bifurcation Analysis in a Rotational Inverted Pendulum*, pp. 144–149, World Scientific, Singapore, 1999.

[62] D. Pagano, L. Pizarro, and J. Aracil, "Local bifurcation analysis in the Furuta pendulum via normal forms," *International Journal of Bifurcation and Chaos*, vol. 10, no. 5, pp. 981–995, 2000.

[63] D. Pagano, *Bifurcaciones en los sistemas de control*, Ph.D. thesis, Escuela Superior de Ingenieros. Universidad de Sevilla, Sevilla, Spain, 1999.

[64] J. Aracil, K. J. Åström, and D. J. Pagano, "Global bifurcations in the Furuta pendulum," in *Proceedings of the Nonlinear Control Systems Design Symposium*, vol. 1, pp. 35–40, Enschede, The Netherlands, July 1998.

Review Article

A Comparative Analysis of Recent Identification Approaches for Discrete-Event Systems

Ana Paula Estrada-Vargas,[1,2] Ernesto López-Mellado,[1] and Jean-Jacques Lesage[2]

[1] *CINVESTAV, Unidad Guadalajara, Avenue Científica 1145, Colonia El Bajío, 45010 Zapopan Jal, Mexico*
[2] *LURPA, ENS de Cachan. 61 Avenue du Président Wilson, 94235 Cachan Cedex, France*

Correspondence should be addressed to Ernesto López-Mellado, elopez@gdl.cinvestav.mx

Analogous to the identification of continuous dynamical systems, identification of discrete-event systems (DESs) consists of determining the mathematical model that describes the behaviour of a given ill-known or eventually unknown system from the observation of the evolution of its inputs and outputs. First, the paper overviews identification approaches of DES found in the literature, and then it provides a comparative analysis of three recent and innovative contributions.

1. Introduction

Analogous to the identification of continuous dynamical systems, identification of discrete-event systems (DESs) consists of determining the mathematical model that describes the behaviour of a given DES from the observation of the evolution of inputs and outputs and possibly from other knowledge about the system behaviour.

The automated building of discrete-event models from external observation of system behaviour interests applications such as reverse engineering for (partially) unknown systems, fault diagnosis, or system verification. The first results that constitute the theoretical basis of DES identification approaches were called grammars inference [1]; the aim was the building of a finite automaton from positive samples of accepted words. Later several techniques that synthesise context-free grammars or Petri nets (PNs) have been proposed.

During the current decade, the interest on the DES identification problem grew yielding methods oriented to industrial systems for discovering or rediscovering the functioning of legacy control/management systems. Based on diverse approaches, these methods obtain mathematical models expressed as finite automata (FA) or PN, from inputs and/or outputs sequences observed in a passive way during the operation of the system
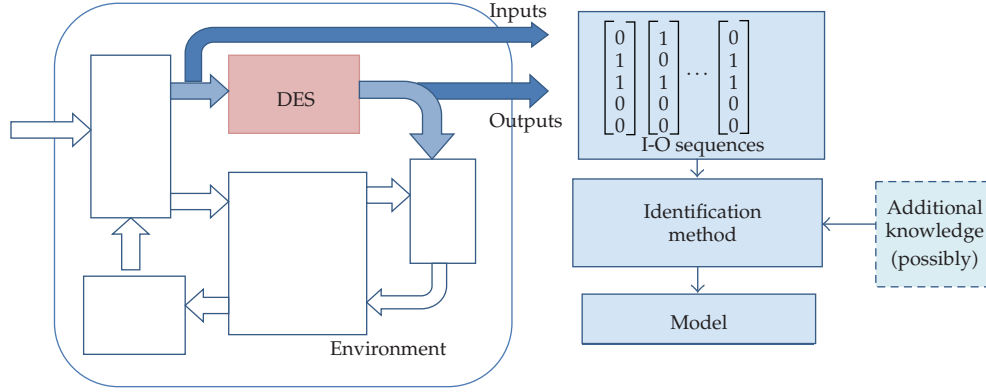
Figure 1: Passive identification of a DES during its operation.

within its environment (see Figure 1). The obtained models are close approximations to the actual system behaviour.

The paper presents a comparative study of identification approaches of DES. It focuses on three different approaches described in recent publications: (i) a progressive identification approach proposed by Meda-Campaña [2] in which several algorithms have been proposed allowing the online identification of concurrent DES, (ii) an offline input-output approach, proposed by Klein [3, 4] in which, through an efficient technique oriented to fault diagnosis, it is obtained a nondeterministic FA representing exactly the observed behaviour, (iii) and an offline approach based on an integer linear programming (ILP) technique initially proposed by Giua and Seatzu [5] and extended by several works like those of Cabasino et al. in [6] and Dotoli et al. in [7] which leads to free-labelled PN models representing observed sequences.

The remainder of this paper is organised as follows. In Section 2 the earlier works on the matter are overviewed. In Section 3, a summarised description of the abovementioned approaches to DES identification is presented. Then, in Section 4 the comparative analysis is developed. Finally concluding remarks and future trends are given.

2. Overview of Identification Techniques

2.1. Original Approaches from Computer Sciences

The first identification methods appeared in the field of theoretical computer sciences as a problem of obtaining a language representation from sets of accepted words; such methods are considered as learning techniques.

Gold's method for identification in the limit [1] processes positive samples: an infinite sequence of examples such that the sequences contain all and only all of the strings in the language to learn.

The Probably Approximately Correct (PAC) learning technique in [8] learns from random examples and studies the effect of noise on learning from queries.

The query learning model proposed in [9] considers a learning protocol based on a "minimally adequate teacher"; this teacher can answer two types of queries: membership query and equivalence query.

Several works adopted state machines as representation model, allowing description of the observed behaviour.

In [10] a method to model a language as Moore or Mealy machine is presented. The system under investigation is placed within a test bed and connected to a so-called experimenter, which generates the input signals and records the output signals of the system. The identification can be started considering a very few number of states. If, at some point of the experiment, it is impossible to find a correct machine with the assumed number of states, then the identification is started again considering a machine with one more state.

The method proposed in [11] obtains models representing Mealy machines. The presented method does not require any a priori knowledge of the system, and only a single observed sequence is available. The algorithm lists all reduced machines which may produce the input-output sequence given. The construction principle is the merging of equivalent states.

In [12] a method to identify nondeterministic Moore machines based on a set of input-output sequences is presented. All of the sequences start in the same initial state. The identification principle is the reduction of an initial machine represented as a tree.

In [13] it is presented a method manipulating simultaneously a sample of sequences to produce a convergent series of Mealy machines such that the behaviour of every new machine includes the behaviour of the previous one. The automaton is built step by step. At each step, the already available machine is examined and completed by adding transitions and possibly new states.

Later, in [14] an algorithm to identify a unique Moore machine generating the behaviour observed during m sequences starting at the same initial state is proposed. The learning procedure operates in three steps: induction, contradiction, and discrimination. A state can never be deleted, and only transitions between states can be modified. This method is improved in [15]; it proposes two algorithms to identify multiple systems as well as systems that may not be initialized between two records.

The identification problem for context-free grammars (CFGs) needs, beside given examples, some additional structural information for the inference algorithm [16].

The study in [17] has investigated a subclass of CFGs called simple deterministic grammars and gave a polynomial time algorithm for exactly identifying it using equivalence and membership queries in terms of general CFGs.

In [18] it has been shown that the inference problem for even linear grammars can be solved by reducing it to one for deterministic finite automata (DFA); a polynomial time algorithm for the reduction of the DFA has been presented.

Other works use as description formalism Petri net models. In [19] an algorithm for synthesising Petri net models is presented. The proposed algorithm has two phases. In the first phase, the language of the target system is identified in the form of DFA. In the second phase, the algorithm guesses from the DFA the structure of a Petri net that accepts the obtained language.

2.2. DES Identification Approaches

In recent years, model identification methods are oriented towards the description of (partially) unknown DES. The observed sequences of DES' outputs and/or inputs are processed for obtaining a model that describes its behaviour.

In [20] an identification method based on the least square estimator has been presented; later, several extensions to this work [2, 21–25] provided solutions to the updating of a model from the continuous recording of output sequences.

Another recent method [3, 4], which has been extended to distributed identification [26, 27], allows to build a non deterministic FA from a set of input-output sequences measured from the DES initial condition of functioning. The method was proposed for obtaining exact models adapted for fault detection in a model-based diagnosis approach [28].

In [5] an approach to build a free-labelled PN from a finite set of transitions strings is presented. The approach is based on the solution of (ILP) an Integer Linear Programming Problem. The obtained PN generates exactly the given strings thanks to the creation of examples and counter examples during the procedure. Several identification techniques have been derived from this seminal work [6, 7, 29–34] for dealing with diverse aspects of DES identification.

The first methods mentioned above deal with the modelling of a given language using different representations, whilst the recent methods addressed the problem of automated modelling of DES from observed behaviour based on model identification techniques. These works are represented in Figure 2 according to a classification, discussed in Section 4.

3. Recent Approaches of DES Identification

3.1. Choice of the Considered Approaches

In this section we overview three different approaches adopted in recent publications addressing the specific problem of DES identification; they have been selected because of the soundness of their results. The first approach deals with unknown partially measurable DES exhibiting cyclic behaviour; overviewed results were reported in [2, 21–25]. The second approach is offline; it is oriented to obtain models devoted to model-based fault diagnosis; literature of this approach can be found in [3, 4, 26–28]. The third approach was initially defined as offline and later extended to be online executed; it deals with DES that does not necessarily have binary outputs; the review is presented from some of the representative works among those in [5–7, 29–34].

3.2. Progressive Identification

Problem

The problem addressed in this work is to build a model for a DES as it evolves from the observation of its output signals [2]. This work can be considered as a basis for verification of systems, hardware or software, or it can be extended to address problems of reverse engineering.

Approach

The identification procedure computes an Interpreted Petri Net (IPN) model describing the behaviour of the unknown DES. Some assumptions are considered on the type of systems
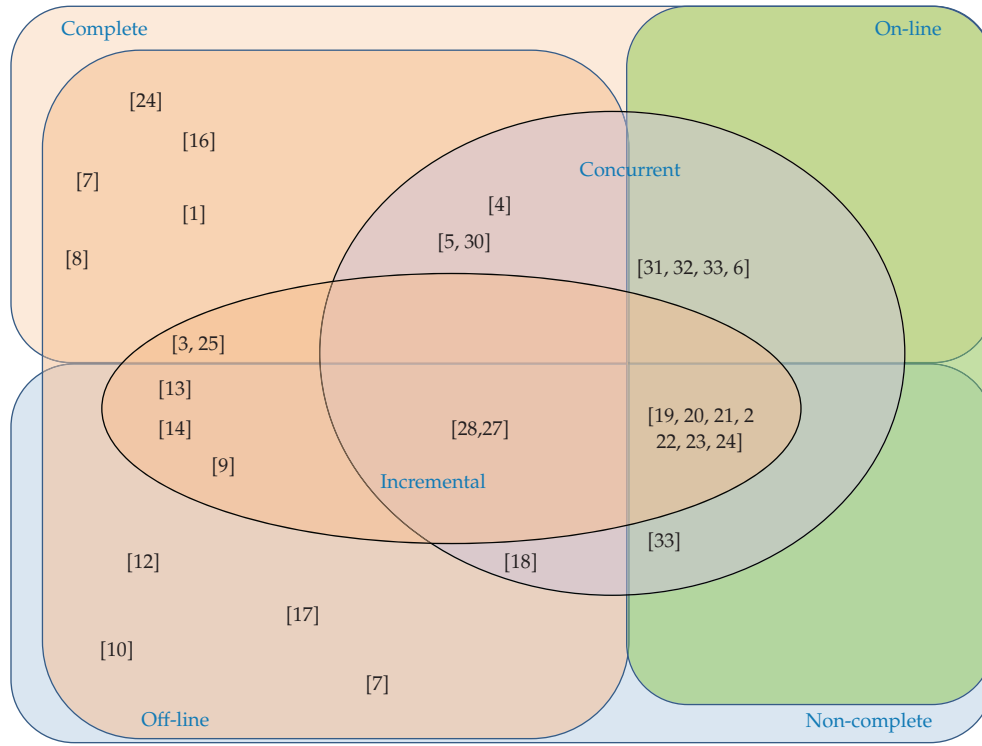
Figure 2: Classification of identification techniques.
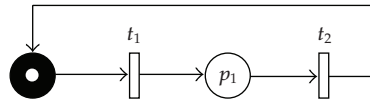
Figure 3: t-component associated with $m_1 = t_1 t_2$.

to be identified: they can be described by a live, safe, cyclic, without self-loops, and event-detectable IPN Q whose transitions are not fired simultaneously.

Methodology

A sequence of models is built in such a way that the current model acquires more details than the previous one approaching to the actual model of the system.

The algorithm receives a sequence of output signals obtained from observations during the system operation. These output signals must be binary vectors representing the current state of every one of the sensors measuring the output behaviour of the system.

The procedure returns an IPN whose measurable places represent the sensors of the system and nonmeasurable places represent internal states. Every reachable marking of the Petri net represents the current state of the system at each moment.

The strategy of the identification is based on the reconstruction of the cyclic components of the system model, by processing cyclic sequences of transitions (called *m*-words) computed from the observed output symbols.
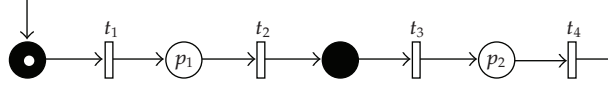
Figure 4: t-semiflow inferred $W_1 = m_1 m_2$.
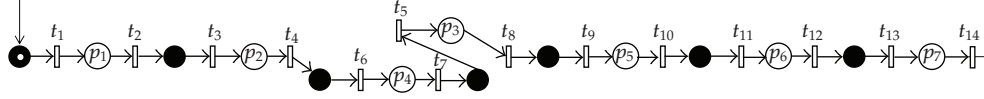
Figure 5: t-semiflow inferred $W_1 = m_1 m_2 m_3 m_4 m_5 m_6 m_7$.

The model identification procedure performs mainly two tasks: the computation of the measurable part of the system and the inference of the nonmeasurable part of the system.

Algorithm

Progressive identification is conducted as follows

(1) Read the vectors of output symbols generated by the system.

(2) Detect an output word (*m-word*) when the first and last output symbols are the same.

(3) For every two consecutive output symbols, compute a transition representing the output change.

(4) Compute an *m*-word adding each computed transition in the step above.

(5) Compute non measurable (dark) places

    (a) to constrain the firing order of the transitions to the order in which they were computed,

    (b) to compute the *t*-component associated with the *m*-word.

(6) Update the IPN model allowing firing of all computed *m*-words and inferring *t*-semiflows by

    (a) computing new measurable places and transitions,

    (b) removing or adding dependencies (possibly merging places) updating the *t*-semiflows.

Example 3.1. In order to illustrate the method, we take from [24] the following example of a system with 7 output signals. For sake of brevity, only main steps are shown.

Step 1. First output symbols are

$$o_1 = [0000000]^T, \quad o_2 = [1000000]^T, \quad o_3 = [0000000]^T = o_1, \quad o_4 = \cdots. \qquad (3.1)$$

Step 2. The first cyclic observed sequence is $o_1 o_2 o_1$.
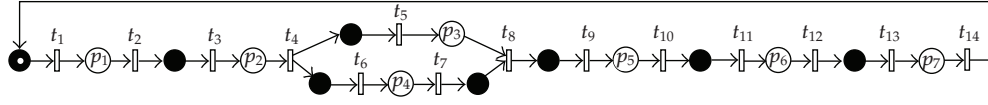
Figure 6: Complete t-semiflow $W_1 = m_1 m_2 m_3 m_4 m_5 m_6 m_7$ and inferred t-semiflow $W_2 = m_1 m_2 m_{3\text{-}4}$.
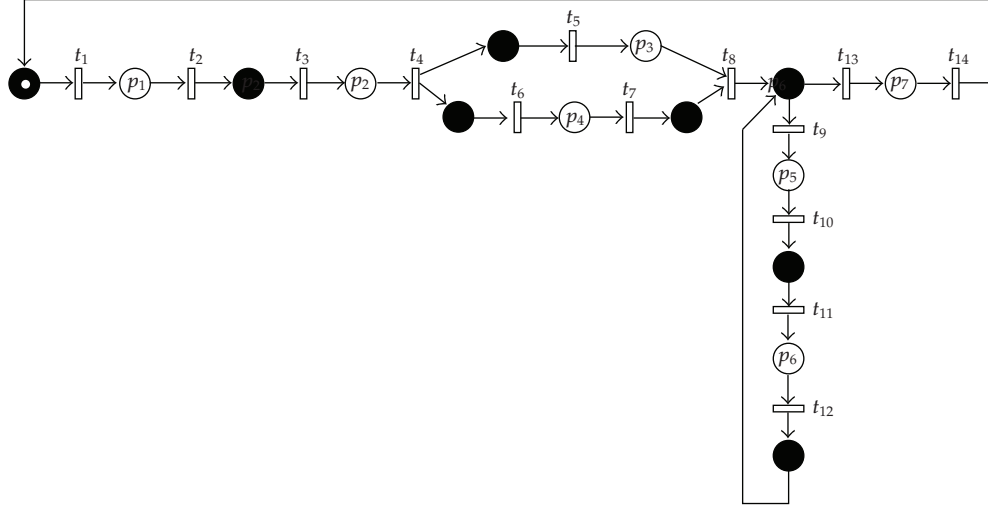
Figure 7: Final model obtained by the progressive identification.

Step 3. t_1 will represent the transition from o_1 to o_2, and t_2 will represent the transition from o_2 to o_1.

Step 4. The m-word resulting is $m_1 = t_1 t_2$.

Step 5. The t-component associated with the m-word $t_1 t_2$ is shown in Figure 3.

Step 6. The first t-semiflow inferred is $W_1 = m_1$.

Then the next output word is treated with Steps 1–4; the m-word $m_2 = t_3 t_4$ is obtained. Its respective t-component associated is added to infer a new t-semiflow $W_1 = m_1 m_2$ in Step 6, as shown in Figure 4.

After computing the m-words $m_3 = t_6 t_7$, $m_4 = t_5 t_8$, $m_5 = t_9 t_{10}$, $m_6 = t_{11} t_{12}$, and $m_7 = t_{13} t_{14}$, it is inferred in Step 6 the t-semiflow $W_1 = m_1 m_2 m_3 m_4 m_5 m_6 m_7$ shown in Figure 5.

The detection of the m-word $m_1 = t_1 t_2$ is the first one of $W_1 = m_1 m_2 m_3 m_4 m_5 m_6 m_7$. Then, it is supposed that W_1 has been completely observed and a new t-semiflow $W_2 = m_1$ is inferred. Observed m-words $m_2 = t_3 t_4$ and $m_{3\text{-}4} = t_5 t_6 t_7 t_8$ in Step 4 are added to the t-semiflow W_2, and the model is updated in Step 6 to allow the firing of all of them, as shown in Figure 6.

The last m-word $m_7 = t_{13} t_{14}$ is observed. It is made a merging of places to allow the firing of the m-words observed in the order they appeared. A new t-semiflow $W_3 = m_5 m_6$ is inferred and t-semiflows $W_1 = m_1 m_2 m_3 m_4 m_7$ and $W_2 = m_1 m_2 m_{3\text{-}4} m_7$ are updated. The final model is shown in Figure 7.

Complexity

The proposed algorithms to update the non measurable places have linear complexity on the number of the transitions computed and the *m*-words detected. Then, the complete algorithm to update a model that includes all of the updating procedures of non measurable places is executed also in polynomial time.

Limitations

In some cases, the obtained model may represent an exceeding behaviour with respect to the observed one from the system. Furthermore, in this approach only outputs of the DES are observed. Consequently, the state evolution of the systems that does not provoke an output evolution cannot be identified. Additionally, the behaviour represented by structures such as self-loops, shared resources in mutual exclusion, and implicit nonmeasurable places cannot be identified using this methodology

3.3. Parametric Automata Construction

Problem

In this work a method for building finite automaton from a set of inputs and outputs sequences measured during the system evolution is presented [3, 4]. The method was proposed for obtaining models adapted for fault detection in a model-based approach [28].

Approach

The identification approach proposes to compute a nondeterministic finite automaton with outputs (NDAAO) describing the behaviour of the unknown DES. The definition of the NDAAO will be presented below. The system to be identified is a compound system (controller + plant) running in a closed loop considered as an event generator.

Methodology

The algorithm receives a set of observed sequences obtained from the system to be identified. Each observed sequence is an ordered series of input/output (I/O) binary vectors exchanged between controller and plant during operation. As a consequence, observed sequences do not necessarily have the same length; however, the first and last I/O vectors of all sequences are identical (cyclic functioning).

　　　The procedure yields a Nondeterministic Autonomous Automaton with Output (NDAAO). Each state of the NDAAO gives as output a binary vector representing every one of the observed I/O signals of the system.

　　　The first step of the construction of the NDAAO is to fix a parameter k that represents the maximal length of the words (or sequences of I/O vectors) that will be generated by the constructed NDAAO. Basically, the principle of the algorithm is to create states that represent observed words of length k, which are connected through transitions in the order the words have been observed.

Algorithm. A nondeterministic autonomous automaton with output, denoted as NDAAO, is a five-tuple NDAAO = $(X, \Omega, r, \lambda, x_0)$, where X is a finite set of states, Ω is an output alphabet, $r : X \rightarrow 2^X$ is a nondeterministic transition relation, $\lambda : X \rightarrow \Omega$ is an output function, and $x_0 \in X$ is the initial state. The algorithm operates in six steps as follows.

(1) For each observed sequence of I/O vectors σ_i, construct k-length sequences of vectors $u_i(t)$, where k is the a priori fixed parameter.

(2) Construct the NDAAO.

(3) Rename the output function.

(4) Reduct the last state.

(5) Merge the equivalent states.

(6) Close the automaton.

Example 3.2. Let us consider the example of an elementary plant with a controller having two inputs and one output [3]. The observed sequences of I/O vectors are

$$
\sigma_1 = \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right),
$$

$$
\sigma_2 = \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right).
$$

(3.2)

For the sake of readability every I/O vector is coded using a symbol, namely, A, B, C, D, and E representing the observed alphabet. Then the observed sequences are: σ_1 = (A, B, C, D, E, A) and σ_2 = (A, C, B, C, D, A).

Step 1. After choosing a parameter k value (k = 2 in the example), construction of vector sequences of length k is given as

$$
\sigma_1^2 = ((A, A), (A, B), (B, C), (C, D), (D, E), (E, A), (A, A)),
$$

$$
\sigma_2^2 = ((A, A), (A, C), (C, B), (B, C), (C, D), (D, A), (A, A)).
$$

(3.3)

Step 2. Construction of the NDAAO. The identification principle is to associate each different word to a single state. This step is illustrated in Figure 8.

Step 3. Renaming of the Output Function. Each state of the NDAAO corresponds to a unique and stable value of the input and output signals. This value is described by the last letter of each k-length sequence.

Step 4. Reduction of the Last State. The last k states of each branch ending in x_f are labelled with the same letter. These states can be reduced through a procedure that has to be iterated $k - 1$ times. First, merge the prestates of x_f; second, redefine this new state as the final state x_f and delete the former x_f from the set of states. Steps 3 and 4 are illustrated in Figure 9.
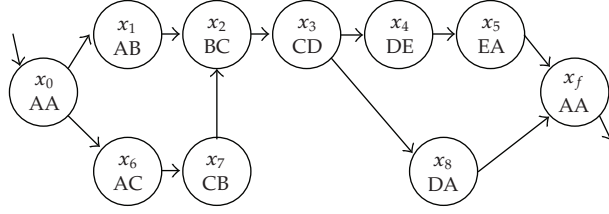
Figure 8: Association of words with states of the NDAAO.
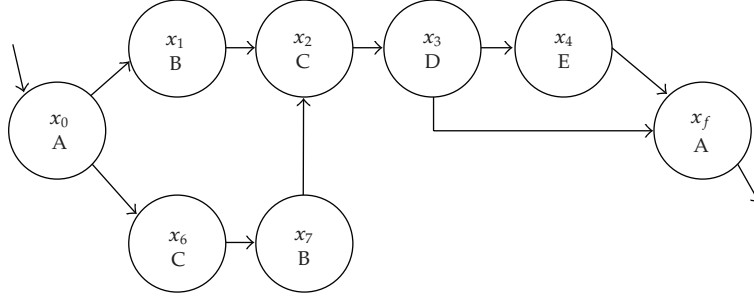
Figure 9: Reduction of the last state.

Step 5. Merging of Equivalent States. Two states are equivalent if and only if

    (1) they are associated with the same output,

    (2) they have the same set of posterior states.

    It has been proved that the merging of equivalent states does not affect the languages accepted by the NDAAO.

Step 6. Closure of the Automaton. Assuming that each observed sequence corresponds to a single production cycle, the states x_0 and x_f of the NDAAO identified are identical. Thus, the NDAAO can be closed resulting in a strongly connected NDAAO. Execution of Steps 5 and 6 can be observed in Figure 10.

Complexity

The time required to build different models is very low and the application of the identification method is efficient. However, the reduction of the NDAAO requires more time than the identification of the model but is not damming.

    If new information is available, the time required for the identification of the NDAAO is reduced. However, this gain is not very important since the reduction must be performed again.

Limitations

For a given value of the identification parameter k, the identified NDAAO is $(k+1)$-complete [35], this means that the NDAAO identified for a given value of the parameter k represents exactly the set of observed words of length lower or equal to $k + 1$.
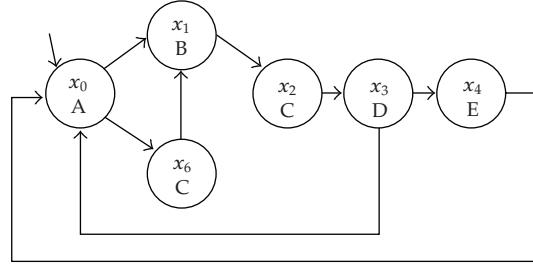
Figure 10: Final model for Example 3.2.

Concurrency cannot be explicitly represented in the obtained automaton, but recent extensions of this work [26] allow performing distributed automation based on an optimal partitioning of I/O that aims at minimizing concurrency between the subsystems. Nevertheless, this technique is dedicated to fault detection and isolation [27].

3.4. Integer Linear Programming Approach

Problem. Several extensions to the original technique presented in [5] have been proposed. We present here only one of the most recent works based on Integer Linear Programming [7]. The problem to be solved is the DES identification by computing a Petri Net model using the observation of events and the available output vectors.

Approach

The identification approach proposes to compute an IPN model such that the observed sequence of events belongs to the language accepted by the IPN. The method considers several hypotheses as follows

(A1) All of the DES events can be detected, distinguished, and not silent.

(A2) The DES can be (partially) observed.

(A3) The DES can be modelled by an IPN system with λ-free labelling function.

(A4) The set of measurable places has a priori a given cardinality q.

(A5) There is an upper bound on the number of places of the IPN.

Methodology

The algorithm receives sequences of events with their corresponding output vectors and the a priori upper bound of the number of nonmeasurable places.

The algorithm returns an IPN with places representing the sensors of the system and labelled transitions representing the observed events. It is also possible that the algorithm returns a 0 (zero) when there is no possible solution of the problem with the given input.

The strategy of the algorithm is to generate an Integer Linear Programming (ILP) problem adding one linear algebraic constraint for every one of the restrictions on the IPN. For selecting among several solutions, it is minimized a performance index. Such an index generally involves arcs weights and number of tokens in the initial marking of the Petri net.

Algorithm. First, we present some definitions taken from [7].

The PN set is given as $D = \{\text{PN} = (P, T, \textbf{Pre}, \textbf{Post}) : \textbf{Pre} \in \mathbb{N}^{mxn}, \textbf{Post} \in \mathbb{N}^{mxn}\}$.

$\mathcal{L}^E(\text{PN}, \textbf{M}_0)$ is the λ-free language of the Petri net PN in E^*, given the initial marking $\textbf{M}_0$.

Let us consider a DES with event set E and language $\mathcal{L}$ verifying assumptions (A1), (A2), and (A3). Let us observe an event sequence $\omega \in \mathcal{L}$ and the corresponding output vectors $\textbf{y} \in \mathbb{N}^q$. The identification problem consists in determining a place set P and its cardinality m, a transition set T and its cardinality n, as well as a λ-free labelling function λ and a PN system $\{\text{PN}, \textbf{M}_0\}$ satisfying assumptions (A4) and (A5) such that PN $\in D$, $\textbf{M}_0 \in \mathbb{N}^m$ and $w \in \mathcal{L}^E(\text{PN}, \textbf{M}_0)$.

A net system is a solution of the identification problem if and only if it satisfies the following set of linear algebraic constraints:

$$\xi(w, \mathbb{Y}, \lambda, T, m) = \begin{cases} \text{Pre}, \text{Post} \in \mathbb{N}^{m \times n}, \\[2mm] M_i \in \mathbb{N}^m \quad \text{with } i = 0, \ldots, h, \\[2mm] \text{Post}^T \vec{1}_{m \times 1} + \text{Pre}^T \vec{1}_{m \times 1} \geq \vec{1}_{n \times 1}, \\[2mm] \text{Post } \vec{1}_{n \times 1} + \text{Pre } \vec{1}_{n \times 1} \geq \vec{1}_{m \times 1}, \\[2mm] \forall t_{\beta_i}^{\alpha_i} \in \sigma \quad \text{with } \lambda(\sigma) = \omega, \quad \text{Pre } \vec{t}_{\beta_i}^{\alpha_i} \leq M_{i-1}, \\[2mm] \forall t_{\beta_i}^{\alpha_i} \in \sigma \quad \text{with } \lambda(\sigma) = \omega, \quad (\text{Post} - \text{Pre})\vec{t}_{\beta_i}^{\alpha_i} = M_i - M_{i-1}. \end{cases} \quad (3.4)$$

The first two constraints are derived from the definitions of markings and Pre and Post incidence matrices. Third and fourth linear algebraic constraints avoid isolated transitions and places, respectively. Constraints five and six are related with enabling and firing of transitions.

Some constraints can be added if additional structural properties are given. For example, if there is no place without successor transitions, then, it can be added $\text{Pre} \cdot \vec{1}_{n \times 1} \geq \vec{1}_{m \times 1}$ and if there are no source transitions, it can be added $\text{Pre}^T \cdot \vec{1}_{m \times 1} \geq \vec{1}_{n \times 1}$.

Since there is not always only one PN satisfying the constraint set, it is used a performance index as follows.

$$\phi(\text{Pre}, \text{Post}, M_0) = \vec{a}^T \text{Pre } \vec{b} + \vec{c}^T \text{Post } \vec{d} + \vec{e}^T M_0. \quad (3.5)$$

Now, the basis of the algorithm that solves the identification problem stated above is presented. The complete algorithm and a more accurate explanation of the solution are included in [7].

(1) Initiate the algorithm variables.

(2) Wait until a new vector and its corresponding output vector are observed.

(3) Associate a transition to the event as follows.

(3.1) If the event occurs for the first time, a new transition is created

(3.2) If the event occurred previously consider the following.

(3.2.1) If a transition related to the event has the same observed output change, then take such a transition and associate it to the event

(3.2.2) Otherwise, a new transition is created.

(4) Solve the ILP problem

$$\min \phi(\text{Pre}, \text{Post}, M_0), \quad \text{s.t. } \xi(w, \mathbb{Y}, \lambda_w, T_w, m'). \tag{3.6}$$

Starting with m' equal to the number of measurable places and incrementing its value, until a solution is found or until m' is equal to the upper bound of the number of places.

Example 3.3. The following example is taken from [7]. Let us consider a DES with $y \in \mathbb{N}^5$ and $\overline{m} = q = 5$. Assume that the initial output is $y_0 = [00102]^T$ and the observed sequence is $w = e_{\alpha_1}, e_{\alpha_2}, e_{\alpha_3}, e_{\alpha_4} = e_1, e_2, e_2, e_1$ with the corresponding outputs $y_1 = [40101]^T$, $y_2 = [31001]^T$, $y_3 = [01011]^T$ and $y_4 = [00102]^T$. At each event occurrence, the identification algorithm is applied, adding constraints to obtain a PN without neither transitions nor places without successors. However, no solution is provided until the occurrence of the last event. The ILP solved is

Minimise

$$[1 \ 1 \ 1 \ 1 \ 1](\text{Pre} + \text{Post}) \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + [1 \ 1 \ 1 \ 1 \ 1] M_0, \tag{3.7}$$

subject to

(1) Pre, Post $\in \mathbb{N}^{5 \times 4}$,

(2) $M_i \in \mathbb{N}^5$ with $i = 0, \ldots, h$,

(3) $\text{Post}^T \vec{1}_{5 \times 1} + \text{Pre}^T \vec{1}_{5 \times 1} \geq \vec{1}_{4 \times 1}$,

(4) $\text{Post} \, \vec{1}_{4 \times 1} + \text{Pre} \, \vec{1}_{4 \times 1} \geq \vec{1}_{5 \times 1}$,

$$\text{Pre} \, \vec{t}_{\beta_i}^{\alpha_i} \leq M_{i-1} \text{Pre}(t_1^1) \leq \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 2 \end{bmatrix}, \qquad \text{Pre}(t_1^2) \leq \begin{bmatrix} 4 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix},$$

$$\text{Pre}(t_2^2) \leq \begin{bmatrix} 3 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \qquad \text{Pre}(t_2^1) \leq \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}. \tag{3.8}$$

(5) for all $t_{\beta_i}^{\alpha_i} \in \sigma$ with $\lambda(\sigma) = w$

Figure 11: Solution for identification problem of Example 3.3.

(6) for all $t_{\beta_i}^{\alpha_i} \in \sigma$ with $\lambda(\sigma) = w$, $(\text{Post} - \text{Pre})\vec{t}_{\beta_i}^{\alpha_i} = M_i - M_{i-1}$

$$
(\text{Post} - \text{Pre})\left(t_1^1\right) = \begin{bmatrix} 4 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 2 \end{bmatrix}, \qquad (\text{Post} - \text{Pre})\left(t_1^2\right) = \begin{bmatrix} 3 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 4 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix},
$$

$$
(\text{Post} - \text{Pre})\left(t_2^2\right) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 3 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \qquad (\text{Post} - \text{Pre})\left(t_2^1\right) \le \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}.
$$

$$(3.9)$$

The IPN obtained is illustrated in Figure 11.

Complexity

In small-size examples, an optimal solution is obtained in a short time implementing and solving the ILP problem on a computer equipped with a standard solver of optimization problems.

In order to apply the identification algorithm online, the dynamics of the DES has to be slow with respect to the time required to solve the ILP problem at each occurrence.

Limitations

It is necessary to fix a-priori the upper bound on the number of places. The statement of ILP problem from a set of observed sequences is exponential. ILP is nondeterministic polynomial, and a solution is not always found.

4. Comparative Analysis of DES Identification Approaches

In this section a comparative analysis of the approaches mentioned above is provided. First, a set of criteria are introduced; then the analysis of every one of the methods regarding the given features is presented. Finally, the analysis is summarized in a comparative table.

4.1. Methods Characteristics

Several features have been considered in [4]; some others are added to have a more complete scope during comparative analysis. Considered characteristics are structured into 4 categories: those characterizing the DES to be identified, those describing the identification process, those qualifying the identified model, and those considering general algorithm features.

DES Characteristics

   (i) *Type of inputs/outputs*. In the general case, inputs and outputs of DES to be identified are *discrete* (they can take a finite number of values). If all inputs and outputs can only take two values (on/off), the DES is called *logic.*

  (ii) *Iterative behaviour*. A DES is called *cyclic* if it iteratively reaches the initial state during its operation. If it iterates over the same behaviour revisiting a state that is not the initial one, then it is called *repetitive.*

Identification Process Characteristics

   (i) *Operation Mode*. If the input sequences cannot be modified, identification is *passive*. Otherwise, the identification is *active*; it is allowed to force input sequences to the actual system to explore behaviours that may not be included in the observed functioning of the system. Since all identification methods considered here are passive, this criterion is not taken into account for comparison aspects.

  (ii) *A Priori Information*. If there is no available knowledge about the DES other than its inputs and outputs evolution, then the identification is *absolute* (commonly called black-box). Otherwise, the identification is *relative*.

 (iii) *Model Updating*. When the model construction is *incremental*, the method progressively updates the model from observed information; otherwise, the identification procedure is *global*: it must be executed on the whole of the observed sequences every time new sequences are collected.

Identified Model Characteristics

   (i) *Concurrency*. This feature considers whether the obtained model can represent *explicitly* concurrent behaviour observed from the system.

  (ii) *Accuracy*. This term is related with completeness of the identified model. If this model represents exactly the observed behaviour, then it is *complete*.

Algorithm Characteristics

(i) *Considered Data*. The identification algorithm constructs an identified model starting from experimental data that can be *inputs* and/or *outputs* of the observed system.

(ii) *Strategy*. If the identification algorithm returns all possible models representing the observed behaviour, the algorithm is called *enumerative*. If only one of the possible models is given, it is *constructive*.

(iii) *Execution*. If the construction of the model can be performed during the system operation by computing a new model from new measurements of the system inputs and/or outputs, the execution is made *online*. Otherwise, the execution is *offline*; the algorithm is not able to run at the same time than the system.

(iv) *Complexity*. This term refers to the computational complexity of the identification algorithm. *Polynomial* time procedures are better than *exponential* ones for coping with large systems exhibiting a large amount of input-output sequences.

4.2. Analysis of Methods

According to the abovementioned features, the identification techniques are analysed.

Progressive Identification

This approach only considers *logical* systems which are *not assumed to be cyclic*. Systems to identify are not assumed to be reinitialized: a single output sequence measured from an arbitrary instant is processed as input to the identification algorithms. However, for long sequence observation if the system exhibits iterative behaviour, this can be captured into the model.

The identification process is *passive*; inputs given to the system cannot be forced. It is considered that there is no information a-priori about the system; only the output sequences are taken into account; this means that the identification is *absolute*.

Since the observed behaviour of the system is progressively integrated to a model, the identification is *incremental*: every time a change on the outputs is observed, the model is updated computing observable part of the system and inferring internal states.

The system identification approach introduced obtains as model an interpreted Petri net. As a consequence of using Petri nets, the concurrency can be *explicitly* represented in the model.

All of the observed sequences are represented in the model, but this approach is *not complete* because the model could represent exceeding behaviour, that is, nonobserved sequences.

Only one solution is given; the solution strategy is *constructive*.

The algorithm provides procedures to be *online* executed, since it is supposed to construct a model while the system is working.

The algorithms are executed in *polynomial* time.

Parametric Automata Construction

This method works on *cyclic logical* systems. The identification procedure receives a set of recorded sequences whose first and last vectors are always the same.

The identification approach is *passive*: inputs given to the system cannot be manipulated. The DES to identify is considered a black-box: *absolute* identification is made.

When new cyclic sequences are provided, they are processed and the model can be updated. It is not necessary to process the whole set of sequences; thus the method is *incremental.*

Due to the limitations of the NDAAO, the concurrency *cannot be explicitly* expressed within the model structure.

Obtained model represents all and only all the observed sequences of a given length; thus the algorithm is *complete*.

Algorithm receives a set of I/O vector sequences and an identification parameter. It returns a unique solution. Thus, it is *constructive.*

I/O sequences are recorded *offline*. The algorithm works in *polynomial* time.

Integer Linear Programming Approach

Considered systems on this approach are *not necessarily cyclic*. Its outputs can be *discrete*(i.e., they can have a finite number of values).

The identification method is *passive*. Since it is required to fix an upper bound for the number of places of the PN, this method is not considered as black-box identification; then, it is *relative* (also-called gray-box identification).

Every time new information is observed, a new ILP is stated and solved; that is, the identification procedure is *global*.

Structure of the obtained model has the form of a PN, which can include *concurrency*.

All observed behaviour is represented on the language of the obtained IPN, but, since counterexamples are not considered in the statement of the ILP problem [5], nonobserved behaviour could be included in the obtained model: the methodology is *not complete*.

The identification algorithm constructs a model able to reproduce a single *event-output* sequence obtained from the system to identify.

The identification algorithm is *enumerative* if we do not consider a performance index. Otherwise, it is constructive, but there could be no solution for a given problem.

In order to apply the identification algorithm *online*, the dynamics of the DES has to be slow with respect to the time required to solve the ILP problem at each occurrence. If this condition is not fulfilled, then the algorithm must run *offline*.

The application is limited to small-length sequences because the ILP statement grows *exponentially*; besides, it is known that solution of ILP is computationally expensive.

4.3. Discussion

Main features of the analysed methods are summarized in Table 1. It can be observed that the progressive identification approach is well adapted for online identification, since it works incrementally in polynomial time. Nevertheless, since it is not a solid methodology, there could be more output sequences than the observed ones; that could be a problem dealing with some type of applications, such as fault diagnosis.

The parametric identification method is not conceived for online execution, but the current model can be incrementally updated when new behaviour is recorded. Although the synthesised model does not represent explicitly the concurrency, the observed input/output sequences of length $k+1$ are exactly represented.

Table 1: Main characteristics of identification approaches.

Comp. criteria	Identif. approach		
	Progressive approach	Parametric automata approach	Integer programming approach
DES to be identified characteristics			
Type of inputs/outputs	Logical	Logical	Discrete
Iterative behaviour	Repetitive	Cyclic	None
Identification process characteristics			
A-priori information	Absolute	Absolute	Relative
Model updating	Incremental	Incremental	Global
Identified model characteristics			
Concurrency	Explicit	Implicit	Explicit
Accuracy	Noncomplete	Complete	Noncomplete
Algorithm characteristics			
Considered data	Outputs	Inputs and outputs	Events and outputs
Strategy	Constructive	Constructive	Enumerative
Execution	Online	Offline	Offline/online
Complexity	Polynomial	Polynomial	Exponential

Despite the inefficiency inherent to integer linear programming, which limits the identification method to process small-size sequences, the ILP procedure yields concurrent models allowing nonbinary markings, when a solution is found.

A more detailed comparison of the methods cannot be made because they do not consider the same hypothesis, and the provided sequences representing the observed behaviour have different formats. Furthermore, the synthesised model is not expressed using the same formalism.

However, from this study we can point out several key features that an identification method should have for dealing with large and complex DESs.

First and foremost, the method must take into account both inputs and outputs of the system, and yield a model expressing explicitly concurrent behaviours; thus PN seems to be a more appropriate modelling formalism.

An efficient (polynomial time) technique based on a progressive strategy makes irrelevant the way the input/output sequences are collected; the sequences may be processed as they are obtained allowing updating the model, if necessary, during the system operation.

Regarding the accuracy of the obtained model, one may think that exceeding behaviour must be avoided. Nevertheless, it is not possible to assure that all of the behaviours have been exhibited by the system during the sequence observations; thus, it could be interesting to infer nonobserved behaviour from the collected sequences. The challenge is discerning among possible exceeding representations by determining plausible model structures; partial knowledge on system components and operations could be useful for this task (independent operations, mutual exclusions, etc.).

5. Summary and Perspectives

An overview of identification techniques for Discrete-Event Systems as well as of the theoretical results on which they are based has been given. Three of the most innovative contributions to this open problem have been outlined. Based on the analysis of their main characteristics, this analysis has been done regardless of any hypothesis on the technology of the systems to be identified (manufacturing, communication, management systems, etc.). These three approaches are very recent and must not be considered as completely finalised. Nevertheless, the study of published results allows perceiving their enormous potential. Currently the interest for DES identification is considerably increasing, and probably, as it is the case since a long time in the field of continuous systems, identification techniques will offer powerful alternatives to classical modelling techniques "by knowledge" for complex DES.

The proposed comparative study allows exhibiting the advantages and drawbacks of the reviewed methods. In some words, we can summarize that, even if it allows identifying systems with logical and discrete (i.e., taking a finite number of values) inputs and outputs, the ILP approach cannot be applied today to identify real complex DES, because of the computational complexity of the algorithm. The main advantage of the parametric method is to generate a complete identified model, but the obtained model cannot represent explicitly the concurrency. The progressive identification method is well adapted to deal with concurrent systems yielding an updated model as the systems evolve; however, the model represents more behaviour than that observed.

New approaches that combine the advantages of these pioneer ones have now to be explored. In [36] we describe the first results obtained in a recent project that aims providing an efficient method for building incrementally, as the DES evolves, a complete Petri net model capturing concurrency.

Acknowledgment

References

[1] E. M. Gold, "Language identification in the limit," *Information and Control*, vol. 10, no. 5, pp. 447–474, 1967.

[2] M. E. Meda-Campaña, *On-line identification of discrete event systems: fundamentals and algorithms for the synthesis of Petri net model*, Ph.D. thesis, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Unidad Guadalajara, Mexico, November 2002.

[3] S. Klein, L. Litz, and J.-J. Lesage, "Fault detection of discrete event systems using an identification approach," in *Proceedings of the 16th IFAC World Congress*, p. 6, Praha, Czech Republic, July 2005, CDROM, paper no 02643.

[4] S. Klein, *Identification of discrete event systems for fault detection purposes*, Ph.D. thesis, Ecole Normale Supérieure de Cachan, Paris, France, October 2005.

[5] A. Giua and C. Seatzu, "Identification of free-labeled Petri nets via integer programming," in *Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference (CDC-ECC '05)*, pp. 7639–7644, Seville, Spain, December 2005.

[6] M. P. Cabasino, A. Giua, and C. Seatzu, "Identification of deterministic Petri nets," in *Proceedings of the 8th International Workshop on Discrete Event Systems (WODES '06)*, pp. 325–331, Ann Arbor, Mich, USA, July 2006.

[7] M. Dotoli, M. P. Fanti, and A. M. Mangini, "Real time identification of discrete event systems using Petri nets," *Automatica*, vol. 44, no. 5, pp. 1209–1219, 2008.

[8]   L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[9]   D. Angluin, "Queries and concept learning," *Machine Learning*, vol. 2, no. 4, pp. 319–342, 1988.

[10] T. L. Booth, *Sequential Machines and Automata Theory*, John Wiley & Sons, New York, NY, USA, 1967.

[11] J. Kella, "Sequential machine identification," *IEEE Transactions on Computers*, vol. 20, no. 3, pp. 332–338, 1971.

[12] A. W. Biermann and J. A. Feldman, "On the synthesis of finite-state machines from samples of their behavior," *IEEE Transactions on Computers*, vol. 21, no. 6, pp. 592–597, 1972.

[13] L. P. J. Veelenturf, "Inference of sequential machines from sample computations," *IEEE Transactions on Computers*, vol. 27, no. 2, pp. 167–170, 1978.

[14] L. P. J. Veelenturf, "An Automata-theoretical approach to developing learning neural networks," *Cybernetics and Systems*, vol. 12, no. 1-2, pp. 179–202, 1981.

[15] M. Richetin, M. Naranjo, and P. Luneau, "Identification of automata by sequential learning," *Pattern Recognition Letters*, vol. 2, no. 6, pp. 379–385, 1984.

[16] L. S. Levy and A. K. Joshi, "Skeletal structural descriptions," *Information and Control*, vol. 39, no. 2, pp. 192–211, 1978.

[17] H. Ishizaka, "Polynomial time learnability of simple deterministic languages," *Machine Learning*, vol. 5, no. 2, pp. 151–164, 1990.

[18] Y. Takada, "Grammatical inference for even linear languages based on control sets," *Information Processing Letters*, vol. 28, no. 4, pp. 193–199, 1988.

[19] K. Hiraishi, "Construction of a class of safe Petri nets by presenting firing sequences," in *Proceedings of the 13th International Conference on Application and Theory of Petri Nets*, vol. 616 of *Lectures Notes in Computer Sciences*, pp. 244–262, Sheffield, UK, June 1992.

[20] M. E. Meda-Campaña, "DES identification using interpreted Petri nets," in *Proceedings of the International Symposium on Robotics and Automation*, pp. 353–357, Saltillo, Mexico, December 1998.

[21] M. E. Meda-Campaña, A. Ramírez-Treviño, and E. López-Mellado, "Asymptotic identification of discrete event systems," in *Proceedings of the 39th IEEE Confernce on Decision and Control*, pp. 2266–2271, Sydney, Australia, December 2000.

[22] M. E. Meda-Campaña and E. López-Mellado, "A passive method for on-line identification of discrete event systems," in *Proceedings of the 40th IEEE Conference on Decision and Control (CDC '01)*, pp. 4990–4995, Orlando, Fla, USA, December 2001.

[23] M. E. Meda-Campaña and E. López-Mellado, "Incremental synthesis of Petri nets models for identification of discrete event systems," in *Proceedings of the 41st IEEE Conference on Decision and Control*, pp. 805–810, Las Vegas, Nev, USA, December 2002.

[24] M. E. Meda-Campaña and E. López-Mellado, "Required event sequences for identification of discrete event s ystems," in *Proceedings of the 42nd IEEE Conference on Decision and Control (CDC '03)*, vol. 4, pp. 3778–3783, Maui, Hawaii, USA, December 2003.

[25] M. E. Meda-Campaña and E. López-Mellado, "Identification of concurrent discrete event systems using Petri nets," in *Proceedings of the 17th IMACS World Congress on Computational and Applied Mathematics*, pp. 11–15, Paris, France, July 2005.

[26] M. Roth, J.-J. Lesage, and L. Litz, "Distributed identification of concurrent discrete event systems for fault detection purposes," in *Proceedings of the European Control Conference (ECC '09)*, Budapest, Hungary, August 2009.

[27] M. Roth, J.-J. Lesage, and L. Litz, "Black-box identification of discrete event systems with optimal partitioning of concurrent subsystems," in *Proceedings of the American Control Conference (ACC '10)*, Baltimore, Md, USA, June 2010.

[28] M. Roth, J.-J. Lesage, and L. Litz, "An FDI method for manufacturing systems based on an identified model," in *Proceedings of the 13h IFAC Symposium on Information Control Problems in Manufacturing (INCO '09)*, pp. 1389–1394, Moscow, Russia, June 2009.

[29] M. P. Cabasino, A. Giua, and C. Seatzu, "Computational complexity analysis of a Petri net identification procedure," in *Proceedings of the International Symposium on Nonlinear Theory and Its Applications (NOLTA '08)*, Bologna, Italy, September 2006.

[30] M. P. Cabasino, A. Giua, and C. Seatzu, "Identification of unbounded Petri nets from their coverability graph," in *Proceedings of the 45th IEEE Conference on Decision and Control (CDC '06)*, pp. 434–440, San Diego, Calif, USA, December 2006.

[31] M. Dotoli, M. P. Fanti, and A. M. Mangini, "An optimization approach for identification of Petri Nets," in *Proceedings of the 8th International Workshop on Discrete Event Systems (WODES '06)*, pp. 332–337, Ann Arbor, Mich, USA, July 2006.

[32] M. Dotoli, M. P. Fanti, and A. M. Mangini, "On-line identification of discrete event systems: a case study," in *Proceedings of the IEEE International Conference on Automation Science and Engineering (CASE '07)*, pp. 405–410, Shangai, China, October 2006.

[33] M. Dotoli, M. P. Fanti, and A. M. Mangini, "On line identification of discrete event systems via Petri nets: an application to monitor specification," in *Proceedings of the 3rd Annual IEEE International Conference on Automation Science and Engineering (CASE '07)*, pp. 893–898, Scottsdale, Ariz, USA, September 2007.

[34] M. P. Fanti and C. Seatzu, "Fault diagnosis and identification of discrete event systems using Petri nets," in *Proceedings of the 9th International Workshop on Discrete Event Systems (WODES '08)*, pp. 432–435, Göteborg, Sweden, May 2008.

[35] T. Moor, J. Raisch, and S. O'Young, "Supervisory control of hybrid systems via l-complete approximations," in *Proceedings of the 4th IEEE Workshop on Discrete Event Systems (WODES '98)*, pp. 426–431, Cagliari, Italy, August 1998.

[36] A.P. Estrada-Vargas, E. López-Mellado, and J.-J. Lesage, "Off-line identification of concurrent discrete event systems exhibiting cyclic behaviour," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pp. 181–186, San Antonio Tex, USA, October 2009.

Review Article

A Survey of Some Recent Results on Nonlinear Fault Tolerant Control

Mouhacine Benosman

Temasek Laboratories, National University of Singapore, 5A, Engineering Drive 1, 9-02, Singapore 117411

Correspondence should be addressed to Mouhacine Benosman, tslmb@nus.edu.sg

Fault tolerant control (FTC) is the branch of control theory, dealing with the control of systems that become faulty during their operating life. Following the systems classification, as linear and nonlinear models, FTC can be classified in two different groups, linear FTC (LFTC) dealing with linear models, and the one of interest to us in this paper, nonlinear FTC (NFTC), which deals with nonlinear models. We present in this paper a survey of some of the results obtained in these last years on NFTC.

1. Introduction

Due to the complexity of modern engineering systems, it is increasingly important to ensure their reliability. This has motivated researchers to concentrate on FTC, which is primarily meant to ensure safety, that is, the stability of a system after the occurrence of a fault in the system. There are two approaches to synthesize controllers that are tolerant to system faults. One approach, known as *passive FTC*, aims at designing a controller which is *a priori* robust to some given *expected faults*. Another approach, known as *active FTC*, relies on the availability of a fault detection and diagnosis (FDD) block that gives, in real-time, information about the nature and the intensity of the fault. This information is then used by a control reconfiguration block to adjust online the control effort in such a way to maintain stability and to optimize the performance of the faulty system.

Passive FTC has the drawback to be reliable only for the class of faults expected and taken into account in the design of the passive FTC. Furthermore, the performances of the closed-loop are not optimized for each fault scenario. However, it has the advantage to avoid the time delay due to online diagnosis of the faults and reconfiguration of the controller, required in active FTC [1, 2], which is very important in practical situations where the time windows during which the system stays stabilizable is very short, for example, the unstable

double inverted pendulum example [3]. In practical applications passive FTCs complement active FTC schemes. Indeed, passive FTCs are necessary during the fault detection and estimation phases [4], where passive FTCs are used to ensure the stability of the faulty system, before switching to active FTCs, that recover some performance after the fault is detected and estimated. Another scenario where passive FTC is used as a complement of active FTC is in the switching-based active FTC, where the active FTC switches between different passive FTC, each controller being designed off-line to cope with a finite number of expected faults and stored in a controller bank; see for example, [5]. Several passive FTC methods have been proposed, mainly based on robust theory, for example, multiobjective linear optimization and LMIs techniques [6], QFT method [7, 8], H_∞ [3, 9], absolute stability theory [10], nonlinear regulation theory [11, 12], Lyapunov reconstruction [13, 14], and passivity-based FTC [15]. As for active FTC, many methods have been proposed for active LFTC, for example, [16–21], as well as for NFTC, for example, [14, 22–35]. As said before, this paper aim is to present some of the recent results on NFTC thus we will not further present LFTC here, and we refer the reader to other survey papers for linear systems; see for example, [36–38].

In [11, 12] the nonlinear regulation theory was used to solve the NFTC problem for particular practical examples, that is, robot manipulators in [12] and induction motors in [11]. The faults treated were modelled as additive actuator faults. In [13, 39] Lyapunov reconstruction techniques were used to solve the problem of loss of actuator effectiveness for nonlinear models affine in the control. The main drawback of this scheme is that it is based on the apriori knowledge of a stabilizing feedback for the nominal safe model and the knowledge of the associated Lyapunov function in closed form. Furthermore, the problem of inputs saturation has not been solved in this work. In [15] the authors studied the case of nonlinear systems with multiplicative actuator faults, and considered the case of systems with inputs saturation. In active NFTC field we also quote [24], where the authors studied the NFTC problem for a particular class of continuous nonlinear models, that is, linear in the control, and proposed a new adaptive fault estimation module, complemented with a control reconfiguration block. In [14], the authors study a specific problem of active FTC, namely, the problem of graceful performance degradation. This problem aims to define online new performances for the faulty system, these performances having to be feasible by the faulty system within its states/actuators limits. Indeed, after the occurrence of a fault the faulty system is expected to be unable to perform the tasks required and planned initially for the safe system. Therefore, new tasks, less demanding, have to be generated online for the faulty system. The idea used there is based on two main stages. The first stage concerns online trajectory planning or reshaping, using online optimization scheme that generates online the closest trajectory to the nominal one, but without violating the new constraints of the faulty system. The second stage concerns the control reallocation problem, using nonlinear model predictive control (NMPC). This scheme deals with nonminimum phase nonlinear models affine in the control. We also refer to [28, 29] where uncertain nonlinear models with constrained inputs, were considered.

An important part of FTC is the one specializing in actuator faults. Indeed, FTCs dealing with actuator faults are relevant for practical application and have already been the subject of many publications [13, 15, 18, 19, 25, 33, 39–49]. The nonlinear case has been studied in [25], where active FTC with respect to additive actuator faults was studied for nonlinear systems affine in the control. Constrained actuators were considered, and state-feedback as well as output-feedback FDDs/FTCs were proposed. In [4, 50], an active NFTC has been proposed for the class of SISO nonlinear systems, with incipient faults. The structure

of the FTC was based on three controllers: a nominal controller for the safe system, that guarantees the system trajectories boundedness until the fault is detected. Then, the NFTC was reconfigured to the second controller that recovers some control performances before the fault is isolated. After the isolation of the fault, a third controller was used based on the faulty model, to improve the control performances. The reconfiguration of the controllers was based on adaptive backstepping approaches. In [14, 33, 48] the authors used model predictive controllers (linear case in [48], and nonlinear case [14, 33]) to reconfigure the controller online after the isolation and estimation of the fault. Finally, we quote [51], where a class of delayed nonlinear systems, modelled with linear terms added to Lipschitz-like nonlinearities with delay terms have been studied. The authors proposed an adaptive LMI-based active NFTC to ensure the stability of the faulty model as well as some optimal performances.

We do not pretend here to present in details all the work quoted above, instead, we will concentrate on some of these results and point out pros and cons of each scheme. We also underline, that we will not report the proofs of the results here, the reader will be refereed to the corresponding paper for the detailed proofs.

This paper is structured as follows. In Section 2, we introduce some notations, and recall some definitions that will be used throughout the paper. Section 3, concerns passive NFTC, followed by active NFTC in Section 4. Finally, we conclude the paper in Section 5, by pointing out some open problems in NFTC.

2. Preliminaries

Throughout the paper we will use the L_2 norm denoted by $|\cdot|$, that is, for $x \in \mathbb{R}^n$ we define $|x| = \sqrt{x^T x}$. The notation $L_f h$ denotes the standard Lie derivative of a scalar function $h(\cdot)$ along a vector function $f(\cdot)$. We also denote by $\tanh(\cdot)$ the hyperbolic tangent function and by $h^{(l)}$ the lth-order-derivative of the scalar function h. Let us introduce now some definitions from [52], that will be frequently used in the sequel.

Definition 2.1 (see [52, page 45]). The solution $x(t, x_0)$ of the system $\dot{x} = f(x)$, $x \in \mathbb{R}^n$, f locally Lipschitz, is stable conditionally to Z, if $x_0 \in Z$, and for each $\epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that

$$|\tilde{x}_0 - x_0| < \delta, \quad \tilde{x}_0 \in Z \Longrightarrow |x(t, \tilde{x}_0) - x(t, x_0)| < \epsilon, \quad \forall t \geq 0. \tag{2.1}$$

If, furthermore, there exist $r(x_0) > 0$, s.t. $|x(t, \tilde{x}_0) - x(t, x_0)| \Rightarrow 0$, for all $|\tilde{x}_0 - x_0| < r(x_0)$ and $\tilde{x}_0 \in Z$, the solution is asymptotically stable conditionally to Z. If $r(x_0) \to \infty$, the stability is global.

Definition 2.2 (see [52, page 48]). Consider the system $H : \dot{x} = f(x, u)$, $y = h(x, u)$, $x \in \mathbb{R}^n$, $u, y \in \mathbb{R}^m$, with zero inputs, that is, $\dot{x} = f(x, 0)$, $y = h(x, 0)$, and let $Z \subset \mathbb{R}^n$ be its largest positively invariant set contained in $\{x \in \mathbb{R}^n \mid y = h(x, 0) = 0\}$. We say that H is globally zero-state detectable (GZSD) if $x = 0$ is globally asymptotically stable conditionally to Z. If $Z = \{0\}$, the system H is zero-state observable (ZSO).

Definition 2.3 (see [52, page 27]). We say that H is dissipative in $X \subset \mathbb{R}^n$ containing $x = 0$, if there exists a function $S(x)$, $S(0) = 0$ such that for all $x \in X$

$$S(x) \geq 0, \quad S(x(T)) - S(x(0)) \leq \int_0^T \omega(u(t), y(t)) dt, \tag{2.2}$$

for all $u \in U \subset \mathbb{R}^m$ and all $T > 0$ such that $x(t) \in X$, for all $t \in [0, T]$, where the function $\omega : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$, called the supply rate, is locally integrable for every $u \in U$, that is, $\int_{t_0}^{t_1} |\omega(u(t), y(t))| dt < \infty$, $\forall t_0 \leq t_1$. S is called the storage function. If the storage function is differentiable, the previous conditions write as

$$\dot{S}(x(t)) \leq \omega(u(t), y(t)). \tag{2.3}$$

The system H is said to be passive if it is dissipative with the supply rate $w(u, y) = u^T y$.

Remark 2.4. The definitions of (ZSD) and (ZSO) are simply an extension to the nonlinear case of the classical notions of detectability and observability for linear systems; see for example, [53].

We will also need the following definition to study the case of time-varying faults in Section 3.

Definition 2.5 (see [54]). A function $\overline{x} : [0, \infty) \rightarrow \mathbb{R}^n$ is called a limiting solution of the system $\dot{x} = f(t, x)$, and f a smooth vector function, with respect to an unbounded sequence t_n in $[0, \infty)$, if there exist a compact $\kappa \subset \mathbb{R}^n$ and a sequence $\{x_n : [t_n, \infty) \rightarrow \kappa\}$ of solutions of the system such that the associated sequence $\{\hat{x}_n :\rightarrow x_n(t + t_n)\}$ converges uniformly to $\overline{x}$ on every compact subset of $[0, \infty)$.

Definition 2.6 (see [55, page 144]). A continuous function $\alpha : [0, a) \rightarrow [0, \infty)$ is said to belong to class $\mathcal{K}$ if it is strictly increasing and $\alpha(0) = 0$. A continuous function $\beta : [0, a) \times [0, \infty) \rightarrow [0, \infty)$ is said to belong to class $\mathcal{KL}$ if for each fixed s the mapping $\beta(r, s)$ belongs to class $\mathcal{K}$ with respect to r and for each fixed r the mapping $\beta(r, s)$ is decreasing with respect to s and $\beta(r, s) \rightarrow 0$ as $s \rightarrow \infty$.

Definition 2.7. A system is said of nonminimum phase, if it has internal dynamics, and their associated zero dynamics are unstable in the Lyapunov sense.

Also, throughout this paper it is said that a statement $P(t)$ holds a.e. if the Lebesgue measure of the set $\{t \in [0, \infty) \mid P(t) \text{ is false}\}$ is zero [54]. We also mean by semiglobal stability of the equilibrium point x^0 for the autonomous system $\dot{x} = f(x)$, $x \in \mathbb{R}^n$ with f a smooth function, that for each compact set $K \subset \mathbb{R}^n$ containing x^0, there exist a locally Lipschitz state feedback, such that x^0 is asymptotically stable, with a basin of attraction containing K (see [56, Definition 3, page 1445]).

3. Passive NFTC

Let us start first with some passive NFTC algorithms. As we said before, these types of FTCs are not expected to "do all the job alone", since in practice they have to be associated with some active FTCs to obtain an efficient controller tolerant to faults.

3.1. Lyapunov-Reconstruction-Based Passive NFTC

We first consider nonlinear systems of the form

$$\dot{x} = f(x) + g(x)u, \tag{3.1}$$

where $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ represent, respectively, the state and the input vectors. The vector fields f, columns of g are supposed to satisfy the classical smoothness assumptions, with $f(0) = 0$. We also assume the system (3.1), locally reachable (in the sense of [57, Definition 5, page 400]). Adding to the previous classical assumptions, we need also the following to hold.

Assumption 3.1. *We assume the existence of a nominal closed-loop control $u_{nom}(t, x)$, such that the solutions of the closed-loop system*

$$\dot{x} = f(x) + g(x)u_{nom}(t, x) \tag{3.2}$$

satisfy $|x(t)| \leq \beta(|x(t_0)|, t - t_0),\ \forall x_{t_0} \in D,\ \forall t \geq t_0$, where $D = \{x \in \mathbb{R}^n \mid |x| < r_0\}, r_0 > 0$ and β is a class $\mathcal{KL}$ function.

Assumption 3.2. *We assume here two types of actuator faults.*

(i) *Firstly, one considers faults that enter the system in an additive way; that is, the faulty model writes as*

$$\dot{x} = f(x) + g(x)(u + F(t, x)), \tag{3.3}$$

where F represents the actuator fault and s.t. $|F(t, x)| \leq b(t, x)$, where $b : [0, \infty) \times D \rightarrow \mathbb{R}$ is a nonnegative continuous function.

(ii) *Secondly, one considers loss of actuator effectiveness, represented by a multiplicative matrix α as*

$$\dot{x} = f(x) + g(x)\alpha u, \tag{3.4}$$

where $\alpha \in \mathbb{R}^{m \times m}$ is a diagonal continuous time variant matrix, with the diagonal elements $\alpha_{ii}(t),\ i = 1, \ldots, m$ s.t. $0 < \epsilon_1 \leq \alpha_{ii}(t) \leq 1$.

The authors in [39] proved the following propositions.

Proposition 3.3. *The control law*

$$u(t,x) = u_{nom}(t,x) - \text{sgn}\left(\left(\frac{\partial V}{\partial x}g\right)^T\right)(b(t,x)+\epsilon), \quad \epsilon > 0, \tag{3.5}$$

where $u_{nom}(t,x)$ is s.t. Assumption 3.1 is satisfied, V is the associated Lyapunov function, $b(t,x)$ is defined in Assumption 3.2, and $\text{sgn}(v)$ denotes the vector sign function, s.t. $\text{sgn}(v)(i) = \text{sgn}(v(i))$; ensures that the equilibrium point $x = 0$ is locally UAS in D for the closed-loop system (3.3) and (3.5).

Proposition 3.4. *The control law*

$$u(t,x) = u_{nom}(t,x) - \text{sgn}\left(\left(\frac{\partial V}{\partial x}g\right)^T\right)\left(|u_{nom}| + \frac{|u_{nom}|}{\epsilon_1}\beta_1\right), \quad \beta_1 \geq 1, \tag{3.6}$$

where $u_{nom}(t,x)$ s.t. Assumption 3.1 is satisfied, V is the associated Lyapunov function, and $\text{sgn}(\cdot)$ denotes the sign function; ensures that the equilibrium point $x = 0$ is locally UAS in D for the closed-loop system (3.4) and (3.6).

These two controllers ensure robust stabilization with respect to additive as well as multiplicative actuators' faults; however, they are discontinuous; that is, due to the *sign* function, therefore the authors in [39] proposed the following two "continuous" versions of the previous propositions.

Proposition 3.5. *The control law*

$$u(t,x) = u_{nom}(t,x) - \text{sat}\left(\left(\frac{\partial V}{\partial x}g\right)^T\right)(b(t,x)+\epsilon), \quad \epsilon > 0 \tag{3.7}$$

ensures that the solutions of the closed-loop system (3.3) and (3.7) satisfy

$$\forall x(t_0) \quad s.t. \ |x(t_0)| \leq \alpha_2^{-1}(\alpha_1(r_0)),$$

$$\exists T \geq 0, \quad s.t. \begin{cases} |x(t)| \leq \beta(|x(t_0)|, t-t_0), & \forall t_0 \leq t \leq t_0+T, \\ |x(t)| \leq \alpha_2^{-1}(\alpha_1(\tilde{x})), & \forall t \geq t_0+T, \end{cases} \tag{3.8}$$

where, for a vector v,

$$sat(v) = \begin{cases} \dfrac{v(i)}{\widetilde{\epsilon}}, & \text{if } |v(i)| \leq \widetilde{\epsilon}, \\ \text{sgn}(v(i)), & \text{if } |v(i)| > \widetilde{\epsilon}, \end{cases}$$

$$\widetilde{x} = \alpha_3^{-1}(2m\widetilde{\epsilon}b_{\max}) \leq \alpha_2^{-1}(\alpha_1(r_0)),$$

$$b(t,x) \leq b_{\max}, \quad \forall t, \forall x \in D, \tag{3.9}$$

and α_1, α_2, and α_3 are class $\mathcal{K}$ functions in D and β is class $\mathcal{KL}$.

Proposition 3.6. *The control law*

$$u(t,x) = u_{nom}(t,x) - sat\left(\left(\frac{\partial V}{\partial x}g\right)^T\right)\left(|u_{nom}| + \frac{|u_{nom}|}{\epsilon_1}\beta_1\right), \quad \beta_1 \geq 1 \tag{3.10}$$

ensures that the solutions of the closed-loop system (3.4) and (3.10) satisfy

$$\forall x(t_0), \quad s.t. \ |x(t_0)| \leq \alpha_2^{-1}(\alpha_1(r_0)),$$

$$\exists T \geq 0, \quad s.t. \begin{cases} |x(t)| \leq \beta(|x(t_0)|, t - t_0), & \forall t_0 \leq t \leq t_0 + T, \\ |x(t)| \geq \alpha_2^{-1}(\alpha_1(\widetilde{x})), & \forall t \geq t_0 + T, \end{cases} \tag{3.11}$$

where, for a vector v,

$$sat(v) = \begin{cases} \dfrac{v(i)}{\widetilde{\epsilon}}, & \text{if } |v(i)| \leq \widetilde{\epsilon}, \\ \text{sgn}(v(i)), & \text{if } |v(i)| > \widetilde{\epsilon}, \end{cases}$$

$$\widetilde{x} = \alpha_3^{-1}(2m\widetilde{\epsilon}u_{nom-\max}) \leq \alpha_2^{-1}(\alpha_1(r_0)),$$

$$|u_{nom}| \leq u_{nom-\max}, \quad \forall t, \tag{3.12}$$

and α_1, α_2, and α_3 are class $\mathcal{K}$ functions in D and β is class $\mathcal{KL}$.

The two continuous controllers (3.7) and (3.10) and do not guarantee the local UAS anymore. However, they guarantee that the closed-loop trajectories are bounded by a class $\mathcal{K}$ function, and that this bound can be made as small as desired by choosing a small $\widetilde{\epsilon}$ in the definition of the function sat. The passive NFTC recalled above is in closed form and thus easy to implement. However, they have two main drawbacks. Firstly, they are based on

the availability of the closed-from expression of the Lyapunov function associated with the nominal stabilizing law, and secondly, they do not consider input saturations in the control design. Therefore, trying to overcome these limitations, other controllers have been proposed and are recalled hereinafter.

3.2. Passivity-Based NFTC

In [15], the passivity theory has been used to develop some new NFTC dealing with actuator multiplicative faults. These results are reported hereinafter.

Theorem 3.7. *Consider the closed-loop system that consists of the faulty system* (3.4)*, with constant unknown matrix α, and the dynamic state feedback:*

$$
\begin{aligned}
\dot{u} &= -L_g W(x)^T - k\xi, \quad u(0) = 0, \\
\dot{\xi} &= \epsilon_1 \left(-(L_g W(x))^T - k\xi \right), \quad \xi(0) = 0,
\end{aligned}
\tag{3.13}
$$

where W is a C^1 radially unbounded, positive semidefinite function, s.t. $L_f W \leq 0$, and $k > 0$. Consider the fictitious system

$$
\begin{aligned}
\dot{x} &= f(x) + g(x)\xi, \\
\dot{\xi} &= \epsilon_1 \left(-(L_g W)^T + \tilde{v} \right), \\
y &= h(\xi) = \xi.
\end{aligned}
\tag{3.14}
$$

If the system (3.14) *is (G)ZSD with the input $\tilde{v}$ and the output y, then the closed-loop system* (3.4) *with* (3.13) *admits the origin $(x, \xi) = (0,0)$ as (globally) asymptotically stable ((G)AS) equilibrium point.*

In Theorem 3.7, one of the necessary conditions is the existence of $W \geq 0$, s.t. the uncontrolled part of (3.3) satisfies $L_f W \leq 0$. To avoid this condition that may not be satisfied for some practical systems, the authors proposed the following Theorem.

Theorem 3.8. *Consider the closed-loop system that consists of the faulty system* (3.4)*, with constant unknown matrix α, and the dynamic state feedback:*

$$
\begin{aligned}
\dot{u} &= \frac{1}{\epsilon_1} \left(-k(\xi - \beta K(x)) - \beta L_g W^T + \beta \frac{\partial K}{\partial x}(f + g\xi) \right), \\
\beta &= \mathrm{diag}(\beta_{11}, \ldots, \beta_{mm}), \quad 0 < \frac{\tilde{\epsilon}_1}{\epsilon_1} \leq \beta_{ii} \leq 1, \\
\dot{\xi} &= -k(\xi - \beta K(x)) - \beta L_g W^T + \beta \frac{\partial K}{\partial x}(f + g\xi), \quad \xi(0) = 0, \quad u(0) = 0,
\end{aligned}
\tag{3.15}
$$

where $k > 0$ and the C^1 function $K(x)$ s.t. a C^1 radially unbounded, positive semidefinite function W satisfying

$$\frac{\partial W}{\partial x}\big(f(x) + g(x)\beta K(x)\big) \leq 0, \quad \forall x \in \mathbb{R}^n, \quad \forall \beta = \mathrm{diag}\big(\beta_{11}, \ldots, \beta_{mm}\big), \quad 0 < \widetilde{\epsilon}_1 \leq \beta_{ii} \leq 1.$$

(3.16)

Consider the fictitious system

$$\dot{x} = f(x) + g(x)\xi,$$

$$\dot{\xi} = \beta\frac{\partial K}{\partial x}\big(f + g\xi\big) - \beta L_g W^T + \widetilde{\widetilde{v}},$$

$$\widetilde{y} = \xi - \beta K(x).$$

(3.17)

If (3.17) is (G)ZSD with the input $\widetilde{\widetilde{v}}$ and the output $\widetilde{y}$, for for all β s.t. β_{ii}, $i = 1, \ldots, m$, $0 < \widetilde{\epsilon}_1 \leq \beta_{ii} \leq 1$. Then, the closed-loop system (3.4) with (3.15) admits the origin $(x, \xi) = (0, 0)$ as (G)AS equilibrium point.

The previous theorems may guaranty *global* AS. However, the conditions required may be difficult to satisfy for some systems. Thus, the authors in [15] introduced the following control law that ensures, under less demanding conditions, semiglobal stability instead of global stability.

Theorem 3.9. *Consider the closed-loop system that consists of the faulty system (3.4), with constant matrix α, and the dynamic state feedback:*

$$\dot{u} = -k(\xi - u_{nom}(x)), \quad k > 0,$$

$$\dot{\xi} = -k\epsilon_1(\xi - u_{nom}(x)), \quad \xi(0) = 0, \quad u(0) = 0,$$

(3.18)

where the nominal controller $u_{nom}(x)$ achieves semiglobal asymptotic and local exponential stability of $x = 0$ for the safe system (3.1). Then, the closed-loop (3.4) with (3.18) admits the origin $(x, \xi) = (0, 0)$ as semiglobal AS equilibrium point.

In [15], the practical problem of input saturation has been studied, and the following result on general nonlinear models, nonnecessarily affine on u, has been proposed.

Theorem 3.10. *Consider the closed-loop system that consists of the faulty system:*

$$\dot{x} = f(x) + g(x, \alpha u)\alpha u$$

(3.19)

for $\alpha \in [\epsilon_1, 1]$, and the static state feedback:

$$u(x) = -\lambda(x)G(x,0)^T,$$

$$G(x,0) = \frac{\partial W(x)}{\partial x}\epsilon_1 g(x,0),$$

$$\lambda(x) = \frac{2\overline{u}}{\left(1 + \gamma_1\left(|x|^2 + 4\overline{u}^2|G(x,0)|^2\right)\right)\left(1 + |G(x,0)|^2\right)} > 0,$$

$$\gamma_1 = \int_0^{2s} \frac{\overline{\gamma}_1(s)}{1 + \overline{\gamma}_1(1)} ds,$$

$$\overline{\gamma}_1(s) = \frac{1}{s}\int_s^{2s}\left(\widetilde{\gamma}_1(t) - 1\right)dt + s,$$

$$\widetilde{\gamma}_1(s) = \max_{\{(x,u)|x|^2+|u|^2 \le s\}}\left\{1 + \int_0^1 \frac{\partial W(x)}{\partial x}\frac{\partial g(x,\tau\epsilon_1 u)}{\partial u}d\tau\right\},$$

(3.20)

where W is a C^2 radially unbounded, positive semidefinite function, s.t. $L_f W \le 0$. Consider the fictitious system:

$$\dot{x} = f(x) + g(x, \epsilon_1 u)\epsilon_1 u,$$

$$y = \frac{\partial W(x)}{\partial x}\epsilon_1 g(x, \epsilon_1 u).$$

(3.21)

If (3.21) is (G)ZSD, then the closed-loop system (3.19) with (3.20) admits the origin as (G)AS equilibrium point. Furthermore $|u(x)| \le \overline{u}$, for all x.

For the particular case of affine nonlinear systems, that is, $g(x,u) = g(x)$, we have the following proposition, which is a direct consequence of Theorem 3.10.

Proposition 3.11. *Consider the closed-loop system that consists of the faulty system (3.4), with constant unknown matrix α, and the static state feedback:*

$$u(x) = -\lambda(x)G(x)^T,$$

$$G(x) = \frac{\partial W(x)}{\partial x}\epsilon_1 g(x),$$

$$\lambda(x) = \frac{2\overline{u}}{1 + |G(x)|^2}.$$

(3.22)

where W is a C^2 radially unbounded, positive semidefinite function, s.t. $L_f W \leq 0$. Consider the fictitious system:

$$\dot{x} = f(x) + g(x)\epsilon_1 u,$$
$$y = \frac{\partial W(x)}{\partial x}\epsilon_1 g(x). \tag{3.23}$$

If (3.23) is (G)ZSD, then the closed-loop system (3.4) with (3.22) admits the origin as (G)AS equilibrium point. Furthermore $|u(x)| \leq \overline{u}, \ \forall x$.

The time-varying versions, that is, for time-varying faults, of the previous results have also been proven in [15] and are recalled hereinafter.

Theorem 3.12. *Consider the closed-loop system that consists of the faulty system (3.4) with the dynamic state feedback:*

$$\dot{u} = -L_g W(x)^T - k\xi, \quad k > 0, \quad u(0) = 0$$
$$\dot{\xi} = \tilde{\alpha}(t)\left(-(L_g W(x))^T - k\xi\right), \quad \xi(0) = 0, \tag{3.24}$$

where $\tilde{\alpha}(t)$ is a C^1 function, s.t. $0 < \epsilon_1 \leq \tilde{\alpha}(t) \leq 1, \ \forall t$, and W is a C^1, positive semidefinite function, such that

(1) $L_f W \leq 0$;

(2) *the system $\dot{x} = f(x)$ is AS conditionally to the set $M = \{x \mid W(x) = 0\}$;*

(3) *for all $(\overline{x}, \overline{\xi})$ limiting solutions for the system*

$$\dot{x} = f(x) + g(x)\xi,$$
$$\dot{\xi} = \alpha(t)\left(-(L_g W)^T - k\xi\right), \tag{3.25}$$
$$y = h(x, \xi) = \xi,$$

with respect to unbounded sequence $\{t_n\}$ in $[0, \infty)$, then if $h(\overline{x}, \overline{\xi}) = 0$, a.e., then either $(\overline{x}, \overline{\xi})(t_0) = (0, 0)$ for some $t_0 \geq 0$ or $(0, 0)$ is a ω-limit point of $(\overline{x}, \overline{\xi})$, that is, $\lim_{t \to \infty}(\overline{x}, \overline{\xi})(t) \to (0, 0)$.

Then the closed-loop system (3.4) with (3.24) admits the origin $(x, \xi) = (0, 0)$ as UAS equilibrium point.

Theorem 3.13. *Consider the closed-loop system that consists of the faulty system:*

$$\dot{x} = f(x) + g(x, \alpha(t)u)\alpha(t)u, \tag{3.26}$$

for $\alpha \in [\epsilon_1, 1]$, $\forall t$, with the static state feedback:

$$u(x) = -\lambda(x)G(x,0)^T,$$

$$G(x,0) = \frac{\partial W(x)}{\partial x}g(x,0),$$

$$\lambda(x) = \frac{2\overline{u}}{\left(1 + \gamma_1\left(|x|^2 + 4\overline{u}^2|G(x,0)|^2\right)\right)\left(1 + |G(x,0)|^2\right)} > 0,$$

$$\gamma_1 = \int_0^{2s} \frac{\overline{\gamma}_1(s)}{1 + \overline{\gamma}_1(1)} ds,$$ (3.27)

$$\overline{\gamma}_1(s) = \frac{1}{s}\int_s^{2s} \left(\widetilde{\gamma}_1(t) - 1\right)dt + s,$$

$$\widetilde{\gamma}_1(s) = \max_{\{(x,u)|x|^2 + |u|^2 \le s\}}\left\{1 + \int_0^1 \frac{\partial W(x)}{\partial x}\frac{\partial g(x, \tau\epsilon_1 u)}{\partial u}d\tau\right\},$$

where W is a C^2, positive semidefinite function, such that

(1) $L_f W \le 0$;

(2) *the system $\dot{x} = f(x)$ is AS conditionally to the set $M = \{x \mid W(x) = 0\}$;*

(3) *for all $\overline{x}$ limiting solutions for the system*

$$\dot{x} = f(x) + g(x, \epsilon_1 u(x))\left(-\lambda(x)\alpha(t)\frac{\partial W}{\partial x}(x)g(x,0)\right)^T,$$

$$y = h(x) = \lambda(x)^{0.5}\left|\frac{\partial W}{\partial x}(x)g(x,0)\right|,$$ (3.28)

with respect to unbounded sequence $\{t_n\}$ in $[0, \infty)$, then if $h(\overline{x}) = 0$, a.e., then either $\overline{x}(t_0) = 0$ for some $t_0 \ge 0$ or 0 is a ω-limit point of $\overline{x}$.

Then the closed-loop system (3.26) with (3.27) admits the origin $x = 0$ as UAS equilibrium point. Furthermore $|u(x)| \le \overline{u}$, $\forall x$.

Proposition 3.14. *Consider the closed-loop system that consists of the faulty system (3.4) with the static state feedback:*

$$u(x) = -\lambda(x)G(x)^T,$$

$$G(x) = \frac{\partial W(x)}{\partial x}g(x),$$ (3.29)

$$\lambda(x) = \frac{2\overline{u}}{1 + |G(x)|^2}.$$

where W is a C^2, positive semidefinite function, such that

(1) $L_f W \leq 0$;

(2) *the system $\dot{x} = f(x)$ is AS conditionally to the set $M = \{x \mid W(x) = 0\}$;*

(3) *for all $\overline{x}$ limiting solutions for the system*

$$\dot{x} = f(x) + g(x)\left(-\lambda(x)\alpha(t)\frac{\partial W}{\partial x}(x)g(x)\right)^T,$$

$$y = h(x) = \lambda(x)^{0.5}\left|\frac{\partial W}{\partial x}(x)g(x)\right|,$$

(3.30)

with respect to unbounded sequence $\{t_n\}$ in $[0, \infty)$, then if $h(\overline{x}) = 0$, a.e., then either $\overline{x}(t_0) = 0$ for some $t_0 \geq 0$ or 0 is a ω-limit point of $\overline{x}$.

Then the closed-loop system (3.4) with (3.29) admits the origin $x = 0$ as UAS equilibrium point. Furthermore $|u(x)| \leq \overline{u}$, $\forall x$.

These passive NFTC schemes are valid for a large class of nonlinear systems, not necessarily affine in the control, and take into account input saturations; however, the conditions to satisfy might be difficult to check when dealing with models having a large number of states.

4. Active NFTC

As we have explained in the introduction, passive FTCs cannot cope with the fault alone, they have to be associated with active FTCs. Indeed, passive FTCs first ensure, at least the stability of the faulty system, during the time period when the FDD is estimating the fault, then active FTC takes over the passive FTC and, using the estimated faulty model they try to optimize the performances of the faulty system. We present in this section some active NFTC schemes.

4.1. Optimization-Based Active NFTC

In [14], the authors studied the problem of *graceful performance degradation* for affine nonlinear systems. The method is an optimization-based scheme, that gives a constructive way to re-shape online the output reference for the postfault system, and explicitly take into account the actuators and states saturations. The online output reference reshaping is associated with an online, MPC-based, controller reconfiguration, that forces the postfault system to track the new output reference.

The model considered are affine in the control:

$$\dot{x} = f(x) + g(x)u,$$

$$y = h(x),$$

(4.1)

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^{n_a}$, and $y \in \mathbb{R}^m$ represent respectively the state, the input and the controlled output vectors. The vector fields f, columns of g, and function h are supposed to satisfy the following classical assumptions.

Assumption 4.1. $f : \mathbb{R}^n \to \mathbb{R}^n$ *and the columns of* $g : \mathbb{R}^n \to \mathbb{R}^{n \times n_a}$ *are smooth vector fields on a compact set X of* $\mathbb{R}^n$ *and* $h(x)$ *is a smooth function on X with* $f(0) = 0, h(0) = 0$.

Assumption 4.2. *System (3.1) has a well-defined (vector) relative degree* $\{r_1, \ldots, r_m\}$ *at each point* $x^0 \in X$ *(see e.g., [58]).*

Assumption 4.3. *The system is fully or over-actuated, in the sense that the number of actuators is at least equal to the number of controlled outputs, that is,* $n_a \geq m$.

Assumption 4.4. *We assume that assumptions 4.1–4.3 above, are preserved after the occurrence of a fault in the system.*

Assumption 4.5. *We assume additionally that the desired nominal trajectory is feasible by the nominal (safe) system, within its input/state limits.*

The control objective is then, that to find a controller u s.t. the nominal as well as the faulty systems' output vector y tracks asymptotically a desired smooth feasible trajectory $y_d(t)$, while satisfying the actuators and states constraints:

$$
\begin{aligned}
u \in \Omega &\triangleq \left\{ u = (u_1, u_2, \ldots, u_{n_a})^T \mid u_i^- \leq u_i \leq u_i^+, \ i = 1, 2, \ldots, n_a \right\}, \\
x \in X &\triangleq \left\{ x = (x_1, x_2, \ldots, x_n)^T \mid x_i^- \leq x_i \leq x_i^+, \ i = 1, 2, \ldots, n \right\},
\end{aligned}
\tag{4.2}
$$

where $u^- = (u_1^-, u_2^-, \ldots, u_{n_a}^-)^T$, $u^+ = (u_1^+, u_2^+, \ldots, u_{n_a}^+)^T$ and $x^- = (x_1^-, x_2^-, \ldots, x_n^-)^T$, $x^+ = (x_1^+, x_2^+, \ldots, x_n^+)^T$ are vectors of lower/upper actuators and states limits, respectively. To do so the authors formulate the problem as the following optimization problem:

$$
\min_{(a, t_{2F})} J = \min_{(a, t_{2F})} \int_{t_{1F}}^{t_{2F}} (y_{\text{nom}}(t) - y_d(t))^T Q_1 (y_{\text{nom}}(t) - y_d(t)) dt + \int_{t_{1F}}^{t_{2F}} u(t)^T Q_2 u(t) dt, \tag{4.3}
$$

under the constraints

$$
\dot{x} = f_F(x) + g_F(x)u,
$$

$$
y_d(t, a, t_{2F}) = h(x),
$$

$$
u^- \leq u \leq u^+,
$$

$$
x^- \leq x \leq x^+,
$$

$$
y^{(k)}(t_{1F}) \triangleq \left(y_1^{(k)}(t_{1F}), \ldots, y_m^{(k)}(t_{1F}) \right)^T = y_{\text{nom}}^{(k)}(t_{1F}) \triangleq \left(y_{\text{nom}_1}^{(k)}(t_{1F}), \ldots, y_{\text{nom}_m}^{(k)}(t_{1F}) \right)^T,
$$

$$
y^{(k)}(t_{2F}) \triangleq \left(y_1^{(k)}(t_{2F}), \ldots, y_m^{(k)}(t_{2F}) \right)^T = y_{\text{nom}}^{(k)}(t_{2\text{nom}}) \triangleq \left(y_{\text{nom}_1}^{(k)}(t_{2\text{nom}}), \ldots, y_{\text{nom}_m}^{(k)}(t_{2\text{nom}}) \right)^T,
$$

$$
k = 0, \ldots, s,
$$

$$
t_{2F} \geq t_{2\text{nom}},
$$

$$
\tag{4.4}
$$

where $y_d(t) = (\sum_{i=1}^{i=l+1} a_{i1}((t-t_{1F})/(t_{2F}-t_{1F}))^{(i-1)}, \ldots, \sum_{i=1}^{i=l+1} a_{im}((t-t_{1F})/(t_{2F}-t_{1F}))^{(i-1)})^T s \in \mathbb{N}^+$, $Q_1 \in \mathbb{R}^{m \times m}$, $Q_2 \in \mathbb{R}^{n_a \times n_a}$ are positive definite weight matrices, $a = (a_{(1)1}, \ldots, a_{(l+1)1}, \ldots, a_{(1)m}, \ldots, a_{(l+1)m})^T \in \mathbb{R}^{m(l+1)}$ is the vector of the polynomials coefficients, t_{2F} is the final motion time for the optimal trajectory vector $y_d(t)$, and $t_{2\text{nom}}$ is the final motion time for the nominal trajectory vector $y_{\text{nom}}(t) \overset{\triangle}{=} (y_{\text{nom}_1}(t), \ldots, y_{\text{nom}_m}(t))^T$, and where f_F, g_F hold for the modified vector field f and matrix g after the occurrence of the fault. The existence of solutions and the computation scheme was then studied for different cases, that is, without internal dynamics, with internal dynamics for minimum phase and with internal dynamics for nonminimum phase systems. The authors did not consider in this paper explicitly FDD synthesis for nonlinear systems. Instead they assumed the availability of an FDD module and they studied both cases: first where FDD provides a precise postfault model and, second the realistic case, where FDD gives a delayed imprecise postfault model. This optimization-based scheme, can deal with the general class of nonlinear models affine in the control, with state and input constraints, and include a stable inversion part to deal with nonminimum phase systems, however, the necessary online computation can be time consuming for large models.

4.2. Learning-Based Active NFTC

We report here the results presented in [59], where the author used a learning scheme to modify the feedback control so as to stabilize the system in the presence of a fault.

The author considers systems of the form

$$\dot{x} = f(x) + G(x)[u + \eta(x, t) + \beta(t - T)\xi(x)] \tag{4.5}$$

where, $x \in \mathbb{R}^n$, and $u \in \mathbb{R}^m$ are the state and control vectors, respectively, and $G = [g_1, g_2, \ldots, g_m]$ is an $n \times m$ matrix function, f, $g_i : \mathbb{R}^n \to \mathbb{R}^n$ $i = 1, \ldots, m$ are known smooth vector fields representing the nominal system dynamics, $\beta(t - T)$ is a step function representing an abrupt fault occurring at an unknown time T, $\eta(x, t)$ represents the time-varying model uncertainties, and $\xi(x)$ is the vector of state-dependent faults. The author assumes the existence of a nominal controller $u_N(x)$ that guarantees uniform stabilization of the nominal system:

$$\dot{x} = f(x) + G(x)u. \tag{4.6}$$

The scheme assumes also the availability of the closed form Lyapunov function V_N associated with the nominal stable feedback system:

$$\dot{x} = f(x) + G(x)u_N(x). \tag{4.7}$$

The author proposes then the NFTC:

$$
\begin{aligned}
u &= u_N(x) + \phi\left(x, \widehat{\theta}, \overline{\theta}\right), \\
\phi\left(x, \widehat{\theta}, \overline{\theta}\right) &= -\Omega(x)^T \widehat{\theta} - \overline{\theta}\omega(x), \\
\dot{\widehat{\theta}} &= \Gamma\Omega(x)p(x), \\
\dot{\overline{\theta}} &= \gamma\omega(x)^T p(x), \\
p(x) &= \left(\frac{\partial V_N}{\partial x}G(x)\right)^T, \\
\omega_i(x) &= \tanh\left(\frac{p_i(x)}{\epsilon}\right), \quad \epsilon > 0, \quad i = 1, \dots, m,
\end{aligned}
\tag{4.8}
$$

where $\Omega(x)$ is a $q \times m$ and represents the basis function for the neural network approximation of the fault f by $\widehat{f}(x, \widehat{\theta}) = \Omega(x)^T \widehat{\theta}$, $\widehat{\theta} \in \mathbb{R}^q$. Then, under the assumption of matching conditions, that is, η, ξ are in the range space of G, the author proves that the feedback controller (4.8) stabilize the faulty system (4.5). However, this control law, is based on the knowledge of the full state vector, and might lead to chattering effect if the parameter ϵ is chosen too small.

4.3. Adaptive Backstepping-Based Active NFTC

The scheme presented here is based on the results of [4, 50]. The systems studied are of the form

$$
\begin{aligned}
\dot{x}_i &= x_{i+1} + \phi_i(\overline{x}_i) + \eta_i(x, u, t) + \beta_i(t - T_0)\xi_i(\overline{x}_i), \quad i = 1, \dots, n-1, \\
\dot{x}_n &= \phi_0(x)u + \phi_n(x) + \eta_n(x, u, t) + \beta_n(t - T_0)\xi_n(x), \\
y &= x_1
\end{aligned}
\tag{4.9}
$$

where $x \in \mathbb{R}^n$ is the state vector, $\overline{x}_i = (x_1, \dots, x_i)^T$, $u \in \mathbb{R}$, $y \in \mathbb{R}$ are the input and the output, respectively. The function ϕ_0 is a nonzero smooth function, and ϕ_i, η_i, f_i, $i = 1, \dots, n$ are smooth functions. The control goal is to force the output y to track a desired trajectory $y_r(t)$, where $y_r^{(l)}$, $l = 0, \dots, n$ are known, piecewise continuous and bounded. As in Section 4.2, η_i, ξ_i, $i = 1, \dots, n$ represent the model uncertainties and the expected faults, respectively, and β_i, $i = 1, \dots, n$, represent the time profile of the faults. Then, based on assumption of the availability of a FDD module that detects and estimates the fault, the authors propose the following three-stage controller:

$$
u = \begin{cases}
u_0(x, y_d, t), & t < T_d, \\
u_D(x, y_d, t), & T_d \le t < T_{\text{isol}}, \\
u_I(x, y_d, t), & T \ge T_{\text{isol}},
\end{cases}
\tag{4.10}
$$

where T_d, and T_{isol} are the time of the fault detection and fault isolation, respectively. Based on the adaptive-backstepping approach, the authors propose the following expression for the three controllers.

(i) First for $t < T_d$

$$u_0(t) = \frac{\alpha_n + y_r^{(n)}}{\phi_0(x)} \tag{4.11}$$

with

$$\alpha_0 = 0,$$

$$\alpha_1 = -c_1 z_1 - c_2 z_1 - \phi_1,$$

$$\alpha_i = -c_1 z_i - z_{i-1} - \phi_i - c_2 z_i \sum_{j=1}^{i-1} \left(\frac{\partial \alpha_{i-1}}{\partial x_j} \right)^2 + \sum_{j=0}^{i-2} \frac{\partial \alpha_{i-1}}{\partial y_r^{(j)}} y_r^{(j+1)} \tag{4.12}$$

$$+ \sum_{j=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial x_j} (x_{j+1} + \phi_j), \quad i = 2, \ldots, n,$$

$$z_i = x_i - \alpha_{i-1} - y_r^{(i-1)}, \quad i = 1, \ldots, n.$$

(ii) Second for $T_d \leq t < T_{\text{isol}}$

$$u_D(t) = \frac{\alpha_n + y_r^{(n)}}{\phi_0(x)} \tag{4.13}$$

with

$$\alpha_0 = 0,$$

$$\alpha_1 = -c_1 z_1 - \phi_1 - \widehat{\theta}_1^T \varphi_1 + \rho_1 \left(y, \widehat{\theta}_1, \widehat{\psi}, y_r \right),$$

$$\alpha_i = -z_{i-1} - c_i z_i - \phi_i - \widehat{\theta}_i^T \varphi_i(\overline{x}_i) + \sum_{k=1}^{i-1} \left[\frac{\partial \alpha_{i-1}}{\partial x_k} \left(x_{k+1} + \phi_k + \widehat{\theta}_k^T \varphi_k(\overline{x}_k) \right) \right] \tag{4.14}$$

$$+ \sum_{k=1}^{i-1} \left[\frac{\partial \alpha_{i-1}}{\partial y_r^{(k-1)}} y_r^{(k)} + \frac{\partial \alpha_{i-1}}{\partial \widehat{\theta}_k} \tau_{ki} \right] + \sum_{k=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial x_k} \varphi_k(\overline{x}_k)^T \Gamma_k \sum_{l=k}^{i-2} \left(\frac{\partial \alpha_l}{\partial \widehat{\theta}_k} \right)^T z_{l+1},$$

$$+ \rho_i \left(\overline{x}_i, \overline{\theta}_i, \widehat{\psi}, \overline{y}_r^{(i-1)} \right), \quad i = 2, \ldots, n.$$

and the parameter adaptive laws are

$$\dot{\widehat{\theta}}_k(t) = \tau_{kn}, \quad 1 \leq k \leq n,$$

$$\dot{\widehat{\psi}} = \Gamma_\psi \left[\sum_{k=1}^n z_k \omega_k - \sigma \left(\widehat{\psi} - \psi^0 \right) \right], \quad \psi^0 \geq 0, \Gamma_\psi > 0, \sigma > 0, \tag{4.15}$$

with

$$z_i = x_i - \alpha_{i-1} - y_r^{(i-1)}, \quad i = 1, \ldots, n,$$

$$\tau_{11} = \Gamma_1 \left[\varphi_1 \left(x_1 z_1 - \sigma \left(\widehat{\theta}_1 - \theta_1^0 \right) \right) \right], \quad \sigma > 0, \Gamma_1 > 0,$$

$$\tau_{ki} = \tau_{k(i-1)} - \Gamma_k z_i \frac{\partial \alpha_{i-1}}{\partial x_k} \varphi_k(\overline{x}_k), \quad \Gamma_k > 0, 1 \le k \le i - 1, i = 2, \ldots, n, \tag{4.16}$$

$$\tau_{ii} = \Gamma_i \left[\varphi_i(\overline{x}_k) z_i - \sigma \left(\widehat{\theta}_i - \theta_i^0 \right) \right], \quad \Gamma_i > 0, i = 2, \ldots, n.$$

where φ_i, $i = 1, \ldots, n$ are the basis functions of the linear approximation for the unknown fault function, that is, $\widehat{\xi}_i(\overline{x}_i, \widehat{\theta}_i) = (\widehat{\theta}_i)^T \varphi_i(\overline{x}_i)$, and ρ_i, $i = 1, \ldots, n$ are given bounding control functions.

(iii) Third for $t \ge T_{\text{isol}}$

$$u_I(t) = \frac{\alpha_n + y_r^{(n)}}{\phi_0(x)},$$

$$\dot{\widehat{\theta}}_k(t) = \tau_{kn}^I, \quad 1 \le k \le n, \tag{4.17}$$

$$\dot{\widehat{\psi}} = \Gamma_\psi \left[\sum_{k=1}^{n} z_k \omega_k - \sigma(\widehat{\psi} - \psi^0) \right], \quad \psi^0 \ge 0, \Gamma_\psi > 0, \sigma > 0,$$

associated with the same update laws (4.16), except that the basis functions and the bounding control functions are different from the previous case, that is, for u_D, since in this case they are specific to the isolated fault. Then, the authors proved that under the assumption of bounded uncertainties $\eta_i(x, u, t)$ and bounded fault approximation-error, that is, $\xi_i(\overline{x}_i) - \widehat{\xi}_i(\overline{x}_i, \widehat{\theta}_i)$ bounded $\forall i$, that all the signals and parameter estimates are uniformly bounded, that is, $z(t)$, $\widehat{\theta}(t)$, $\widehat{\psi}(t)$, and $x(t)$ are bounded $\forall t$. However, this approach is based on the special structure of the faulty model (4.9), and assumes the availability of the measurements of state vector for the feedback control. Eventually, the FDD and FTC presented here are based on the assumption of the fault being part of an *apriori* known set of expected fault's models.

4.4. Switched Control-Based Active NFTC

We report here the schemes introduced in [25, 26], where the authors consider both problems of FDD and FTC for a class of nonlinear systems, with input constraints. The model studied are of the form

$$\dot{x} = f(x) + G_{k(t)}(x)\left(u_{k(t)}(y) + \widetilde{u}_{k(t)(t)}\right),$$

$$y(x) = h(x), \quad u_k(t) \in U_k, \quad u_{k(t)}(y) + \widetilde{u}_{k(t)(t)} \in U_k, \tag{4.18}$$

$$k(t) \in K = \{1, \ldots N\}, \quad N < \infty, \quad U_k = \left\{ u \in \mathbb{R}^m : |u| \le u_k^{\max} \right\}, \quad u_k^{\max} > 0 \quad \forall k,$$

where $x \in \mathbb{R}^n$ is the vector of state variables, $y \in \mathbb{R}^m$ is the vector of measurable variables, and $u_k(y) \in \mathbb{R}^m$ denotes the control vector under the kth configuration. The additive actuator faults are modelled by $\widetilde{u}_k$. The vector function f and the matrices $G_k(x)$, $\forall k$ are assumed

to be sufficiently smooth on their domains of definition. For each value of $k \in K$ the system is controlled via a different set of manipulated inputs, which defines a given control configuration. The nonlinear model (4.18) is associated with the following assumption.

Assumption 4.6. *Consider the system* (4.18) *in configuration k under state-feedback. Then for every input $u_{j,k}$, $j = 1, \ldots, m$, there exists a unique state $x_{i,k}$, $i = 1, \ldots, n$, such that with $x_{i,k}$ as output, the relative degree of $x_{i,k}$ with respect to $u_{j,k}$ and only with respect to $u_{j,k}$ is equal to 1.*

This assumption means that each actuator is the only one influencing at least some state. This implies that the effect of a specific actuator on the system evolution is completely distinguishable, which allows fault isolation in that specific actuator. This sufficient fault detection/isolation condition, can be relaxed if the input enters the model in an "upper-triangular" or "lower-triangular" form (refer to [25, Remark 3]). The authors introduced a nonlinear FDD in the following theorem.

Theorem 4.7. *Consider the model* (4.18) *in configuration k which satisfies Assumption 4.6, under the control law:*

$$u_k = -\omega_k(x, u_k^{\max})(L_{G_k} V_k(x))^T,$$

$$\omega_k(x, u_k^{\max}) = \begin{cases} \dfrac{\alpha_k(x) + \sqrt{\alpha_k^2(x) + \left(u_k^{\max}|b_k^T(x)|\right)^4}}{|b_k^T(x)|^2 \left(1 + \sqrt{1 + \left(u_k^{\max}|b_k^T(x)|^2\right)^2}\right)}, & b_k^T(x) \neq 0, \\ 0, & b_k^T(x) = 0, \end{cases}$$

$$\alpha_k(x) = L_{f_k} V_k(x) + \rho_k V_k(x), \quad \rho_k > 0,$$

$$b_k(x) = L_{G_k} V_k(x).$$

(4.19)

assuming that the set $\Phi_k(u_k^{\max}) = \{x, s.t. \ L_{f_k} V_k(x) + \rho_k V_k(x) \leq u_k^{\max}|(L_{G_k} V_k(x))^T|\}$, contains the origin and a neighborhood of the origin.

Let the fault detection and isolation filter for the jth manipulated input in the kth configuration be described by

$$\dot{\tilde{x}}_{i,k} = f_i(x_1, \ldots, \tilde{x}_{i,k}, \ldots, x_n) + g_{j,k}[i](x_1, \ldots, \tilde{x}_{i,k}, \ldots, x_n) \times u_{j,k}(x_1, \ldots, \tilde{x}_{i,k}, \ldots, x_n)$$

$$e_{i,k} = \tilde{x}_{i,k} - x_i,$$

(4.20)

where $g_{j,k}[i]$ denotes the ith element of the vector $g_{j,k}$, $\tilde{x}_{i,k}(0) = x_i(0)$ and the subscripts i, k refer to the ith state under the kth control configuration. Let $T_{j,k}^f$ be the earliest time for which $\tilde{u}_{j,k} \neq 0$, then the fault detection and isolation filter of (4.20) ensures that $\lim_{t \to T_{j,k}^{f+}} e_{i,k}(t) \neq 0$. Also, $e_{i,k}(t) \neq 0$ only if $\tilde{u}_{j,k}(s) \neq 0$, $0 \leq s < t$.

Then, the NFTC has been introduced in the following theorem.

Theorem 4.8. *Consider the closed-loop system* (4.18)*,* (4.19)*, and let* $x(0) \in \Omega_{k_0}$ *for some* $k_0 \in K$*, with* Ω_k *being defined as:* $\Omega_k(u_k^{\max}) = \{x \in \mathbb{R}^n : V_k(x) \leq c_k^{\max}\} \subset \Phi_k$*,* $c_k^{\max} > 0$ *is a level set of* V_k*. Let* T_{j,k_0} *be the earliest time such that* $e_{i,k_0} \neq 0$ *for some* i *corresponding to a manipulated input* u_{j,k_0} *in* (4.20)*. Then, the following switching rule:*

$$k(t) = \begin{cases} k_0, & 0 \leq t < T_{j,k_0}, \\ q \neq k_0, & t \geq T_{j,k_0}, \quad x(T_{j,k_0}) \in \Omega_q, \quad u_{j,k_0} \notin u_q, \end{cases} \tag{4.21}$$

guarantees asymptotic stability of the origin of the closed-loop system (4.18) *and* (4.19)*.*

This active NFTC are applicable for the general class of nonlinear models affine in the control, and are based on a state-feedback (the authors proposed in the same papers an extension to the case of output feedback). However, they require Assumption 4.6 to hold to be able to detect and isolate the actuator fault. Another point is that this scheme does not consider multiplicative actuator faults.

4.5. Predictive Control-Based Active NFTC

The authors in [28, 29] study the problem of NFTC for nonlinear models affine in the control, with input constraints and uncertainties. The nature of faults treated is actuator faults, under the assumption of controllability of the faulty system. Let us recall below the main result of these work.

The models considered are of the form

$$\dot{x} = f(x) + G_k(x)u_k + W_k(x)\theta_k(t), \quad u_k \in U_k, \theta_k \in \Theta_k,$$
$$k \in \{1, \dots, N\}, \quad N < \infty, \tag{4.22}$$

where $x \in \mathbb{R}^n$ denotes the vector of state variables, $u \in U_k \subset \mathbb{R}^m$, $U_k = \{u \in \mathbb{R}^m$, s.t. $|u| \leq u_k^{\max}\}$, and $u_k^{\max} > 0$ $\forall k$ denotes the vector of constrained inputs. The vector $\theta_k(t) = [\theta_k^1 \cdots \theta_k^q]^T \in \Theta_k \subset \mathbb{R}^q$ denotes the vector of time-varying uncertainties but bounded variables taking values in a nonempty compact convex subset of $\mathbb{R}^q$. The vector $f(x)$ (s.t. $f(0) = 0$), the matrices $G_k(x) = [g_k^1(x) \cdots g_k^m(x)]$, $g_k^i \in \mathbb{R}^n$, $i = 1 \cdots m$, and $W(x) = [w_k^1(x) \cdots w_k^q(x)]$, $w_k^i \in \mathbb{R}^n$, $i = 1 \cdots q$, are assumed to be sufficiently smooth on their domain of definition. For each value of the index k the process is controlled via a different manipulated input, which defines a given control configuration. Switching between the available N control configuration is controlled by a higher-level supervisor, which ensures that only one control configuration is active at any given time, and allows only finite number of switches over any finite time interval of time. The main idea of this work is that the authors assume that after the occurrence of a fault, the system will be associated with one of the N configuration and then they build off-line a bank of N nonlinear model-predictive stabilizing controllers, and based on the value of the state vector at the time of fault occurrence, they switch among these controllers to ensure the stability of the faulty system. To make the presentation of the NFTC clear, we follow the same structure of the paper [28] and present first a Lyapunov-based switched controller, then we present the associated nonlinear model predictive controller (NMPC) and finally we present the NFTC based on this algorithm.

The Lyapunov-based controller associated with the system (4.22) is given by the bounded state feedback:

$$u_k^b = -\frac{\alpha_k(x) + \sqrt{\alpha_{1,k}(x)^2 + (u_k^{\max}\beta_k(x))^4}}{\beta_k(x)^2\left(1 + \sqrt{1 + (u_k^{\max}\beta_k(x))^2}\right)}(L_{G_k V_k})^T, \tag{4.23}$$

where, V_k is a robust control Lyapunov function (RCLF) (as defined in [60, page 49]), $\alpha_k(x) = L_f V_k + (\rho_k\|x\| + \chi_k\theta_k^b\|L_{W_k}V_k\|)(\|x\|/(\|x\| + \phi_k))$, $\alpha_{1,k}(x) = L_f V_k + \rho_k\|x\| + \chi_k\theta_k^b\|L_{W_k}V_k\|$, $\beta_k(x) = \|L_{G_k}V_k\|$, $L_{G_k}V_k = [L_{g_k^1}V_k \cdots L_{g_k^m}V_k]$, $L_{W_k}V_k = [L_{W_k^1}V_k \cdots L_{W_k^q}V_k]$, $\theta_k^b > 0$, s.t. $\|\theta_k(t)\| \leq \theta_k^b$, $\forall t$, and $\rho_k > 0$, $\chi_k > 1$, $\phi_k > 0$.

The following convergence result has been reported in [28] and proven in [61]: let $\Pi_k(\theta_k^b, u_k^{\max}) = \{x \in \mathbb{R}^n : \alpha_{1,k}(x) \leq u_k^{\max}\beta_k(x)\}$ and assume that $\Omega_k = \{x \in \mathbb{R}^n : V_k(x) \leq c_k^{\max}\} \subseteq \Pi_k(\theta_k^b, u_k^{\max})$, for some $c_k^{\max} > 0$. Then, given any positive real number, d_k^r, s.t.: $\mathbb{D}_k^r = \{x \in \mathbb{R}^n : \|x\| \leq d_k^r\} \subset \Omega_k$ and $\forall x_0 \in \Omega_k$, $\exists \epsilon_k^{r*} > 0$, s.t. if $\phi_k/(\chi_k - 1) < \epsilon_k^{r*}$ the solutions of the closed-loop system (4.22) and (4.23) satisfy $x(t) \in \Omega_k$, $\forall t$ and $\limsup_{t\to\infty}\|x(t)\| \leq d_k^r$.

We also need to recall a convergence result from [28], that characterizes the behavior of the solutions of (4.22) and (4.23), when the continuous controller (4.23) is implemented in discrete time. The result is as follows: consider the system (4.22) for a fixed k with $\theta_k = 0$, $\forall t$, associated with the controller (4.23). Let $u_k(t) = u_k^b(j\Delta_k)$, $j\Delta_k \leq t < (j + 1)\Delta_k$, $j = 0,\ldots,\infty$. Then, $\forall d_k > 0$, $\exists \Delta_k^* > 0$, $\delta_k' > 0$, $\epsilon_k^* > 0$ s.t. if $\Delta_k \in (0, \Delta_k^*]$ and $x(0) \in \Omega_k$ then $x(t) \in \Omega_k$ $\forall t$ and $\limsup_{t\to\infty}\|x(t)\| \leq d_k$. Also, if $V_k(x(0)) \leq \delta_k'$ then $V_k \leq \delta_k'$ $\forall \tau \in [0, \Delta_k)$ and if $\delta_k' < V_k(x(0)) \leq c_k^{\max}$, then $\dot{V}_k(x(\tau)) \leq -\epsilon_k^*$ $\forall \tau[0, \Delta_k)$.

Next we report the Lyapunov-based predictive control associated with the Lyapunov-based controller (4.23). The following result has been reported in [28] and proven in [62]:

Consider the system (4.22), for a fixed value of k, with $\theta_k(t) = 0$, $\forall t$, associated with the following NMPC controller:

$$\min\{J(x, t, u_k), u_k \in S_k\},$$

$$J(x, t, u_k) = \int_t^{t+T}\left(\|x^u(s, x, t)\|_{Q_k}^2 + \|u_k\|_{R_k}^2\right)ds, \quad Q_k \geq 0, R_k > 0,$$

$$\text{s.t.} \quad \dot{x} = f_k(x) + G_k(x)u_k, \tag{4.24}$$

$$\dot{V}_k(x(\tau)) \leq -\epsilon_k \quad \text{if } V_k(x(t)) > \delta_k', \quad \tau[t, t + \Delta k],$$

$$V_k(x(\tau)) \leq \delta_k' \quad \text{if } V_k(x(t)) \leq \delta_k', \quad \tau \in [t, t + \Delta k],$$

where ϵ_k, δ_k' are as defined above, S_k is the family of piecewise continuous functions with period Δ_k mapping $[t, t + T]$ into U_k, $T > 0$ is the horizon of the optimization, and V_k is RCLF that yields a stability region Ω_k, under continuous implementation of the controller (4.23), with a fixed $\rho_k > 0$. Then, $\forall d_k > 0$, $\exists \Delta_k^* > 0$, and $\delta_k' > 0$, s.t., if $x(0) \in \Omega_k$ and $\Delta \in (0, \Delta_k^*]$, then $x(t) \in \Omega_k$, $\forall t$ and $\limsup_{t\to\infty}\|x(t)\| \leq d_k$.

Finally, we can report the predictive control-based NFTC as follows: Consider the system (4.22), for which the bounded controllers (4.23) and Lyapunov-based MPCs (4.24) have been designed and the stability regions Ω_j, $j = 1,\ldots,N$, under the Lyapunov-based MPCs have been explicitly characterized. Let $d_{\max} = \max_{j=1,\ldots,N}d_j$, d_j as defined above, and

let $\Omega_U = \cup_{j=1}^{j=N} \Omega_j$. Define $J_j(t) = \int_t^{t+T_j} (\|x^u(s,x,t)\|_{Q_k}^2 + \|u_k^b\|_{R_k}^2) ds$, where $t + T_j \geq t$ is the earliest time at which the state of the closed-loop system under bounded controller enters the level set defined by $V_j(x) = \delta_j'$. Then, let $k(0) = i$ for some index $i \in \{1, \ldots, N\}$ and $x(0) \in \Omega_i$. Let T_i^f be the earliest time at which a fault occurs. Furthermore, let $f = \{j : \text{s.t.} \quad j \neq i, \ x(T_i^f) \in \Omega_j\}$, and let l be such that $J_l = \min_{j \in f} J_j$. Then, the following switching rule

$$k(t) \begin{cases} i, & 0 \leq t < T_i^f, \\ l, & t \geq T_i^f \end{cases} \tag{4.25}$$

guarantees that $x(t) \in \Omega_U, \ \forall t \geq 0$ and $\limsup_{t \to \infty} \|x(t)\| \leq d_{\max}$.

To avoid further overload this paper with long equations, we have reported here only the NFTC in the case without uncertainties, that is, $\theta_k(t) = 0, \ \forall k, \ \forall t$. The interested reader may refer to the references [28, 29] for the uncertain case.

This active NFTC, based on the computation off-line of a bank of robust nonlinear controllers, is valid for general nonlinear models affine in the control, however, it is based on the availability of a robust control Lyapunov function in closed-from, which is usually not easily accessible [60]. We can also point out, that in these work [28, 29], the authors assumed the availability of a FDD bloc, and did not consider the problems of fault isolation and estimation delays as well as FDD uncertainties.

5. Conclusion

In the last decades there have been a myriad of results on FTC. Many of those work concentrated initially on linear FTC, and more and more researches started focussing on the nonlinear FTC problems, the later being more challenging than the linear FTC because of the difficulties intrinsic to nonlinear systems. However, many encouraging results have been obtained. We wanted to summarize in this paper the results obtained recently on NFTC. We recalled in the introduction most of the FTC work on nonlinear models. We reported the detailed controllers of some of these results. Unfortunately, it was not possible to report in details all the available results. Our choice was mainly motivated by the degree of the "model-nonlinearities", and we reported the work that, in our opinion, treated some general degree of nonlinearities. Although many interesting results have been obtained so far, we believe that work treating *together* both problems of nonlinear FDD and nonlinear FTC in an *effective applicable* methods, are still missing. Real-life applications of those NFTC theories are also a missing part of the recent work. To conclude, the case of infinite dimension nonlinear models, that is, nonlinear partial derivative equations-based models, has yet to be studied, some recent results in this directions are presented in [63–66].

References

[1] Y. Zhang and J. Jiang, "Issues on integration of fault diagnosis and reconfigurable control in active fault-tolerant control systems," in *Proceedings of the 6th IFAC Symposium of Fault Detection Supervision and Safety for Technical Processes*, pp. 1513–1524, Beijing, China, August-September 2006.

[2] M. Staroswiecki, H. Yang, and B. Jiang, "Progressive accommodation of aircraft actuator faults," in *Proceedings of the 6th IFAC Symposium of Fault Detection Supervision and Safety for Technical Processes*, pp. 877–882, Beijing, China, August-September 2006.

[3] H. Niemann and J. Stoustrup, "Passive fault tolerant control of a double inverted pendulum—a case study," *Control Engineering Practice*, vol. 13, no. 8, pp. 1047–1059, 2005.

[4] X. Zhang, T. Parisini, and M. M. Polycarpou, "Adaptive fault-tolerant control of nonlinear uncertain systems: an information-based diagnostic approach," *IEEE Transactions on Automatic Control*, vol. 49, no. 8, pp. 1259–1274, 2004.

[5] A. Ingimundarson and R. Sanchez-Pena, "Using the unfalsified control concept to achieve fault tolerance," in *Proceedings of the 17th IFAC Word Congress*, pp. 1236–1242, Seoul, Korea, July 2008.

[6] F. Liao, J. L. Wang, and G.-H. Yang, "Reliable robust flight tracking control: an LMI approach," *IEEE Transactions on Control Systems Technology*, vol. 10, no. 1, pp. 76–89, 2002.

[7] S. Wu, M. Grimble, and W. Wei, "QFT based robust/fault tolerantflight control design for a remote pilotless vehicle," in *Proceedings of the IEEE International Conference on Control Applications*, vol. 1, pp. 57–62, Kohala Coast, Hawaii, USA, August 1999.

[8] S.-F. Wu, M. J. Grimble, and W. Wei, "QFT-based robust/fault-tolerant flight control design for a remote pilotless vehicle," *IEEE Transactions on Control Systems Technology*, vol. 8, no. 6, pp. 1010–1016, 2000.

[9] H. Niemann and J. Stoustrup, "Reliable control using the primary and dual Youla parametrization," in *Proceedings of the 41st IEEE Conference on Decision and Control*, vol. 4, pp. 4353–4358, 2002.

[10] M. Benosman and K.-Y. Lum, "Application of absolute stability theory to robust control against loss of actuator effectiveness," *IET Control Theory & Applications*, vol. 3, no. 6, pp. 772–788, 2009.

[11] C. Bonivento, A. Isidori, L. Marconi, and A. Paoli, "Implicit fault-tolerant control: application to induction motors," *Automatica*, vol. 40, no. 3, pp. 355–371, 2004.

[12] C. Bonivento, L. Gentili, and A. Paoli, "Internal model based fault tolerant control of a robot manipulator," in *Proceedings of the 43rd IEEE Conference on Decision and Control*, vol. 5, pp. 5260–5265, 2004.

[13] M. Benosman and K.-Y. Lum, "Passive actuators' fault-tolerant control for affine nonlinear systems," in *Proceedings of the 17th IFAC Word Congress*, pp. 14229–14234, Seoul, Korea, July 2008.

[14] M. Benosman and K.-Y. Lum, "Online references reshaping and control reallocation for nonlinear fault tolerant control," *IEEE Transactions on Control Systems Technology*, vol. 17, no. 2, pp. 366–379, 2009.

[15] M. Benosman and K.-Y. Lum, "Application of passivity and cascade structure to robust control against loss of actuator effectiveness," to appear in *International Journal of Robust and Nonlinear Control*.

[16] Y. Zhang and J. Jiang, "Design of integrated fault detection, diagnosis and reconfigurable control systems," in *Proceedings of the 38th IEEE Conference on Decision and Control*, pp. 3587–3592, Phoenix, Ariz, USA, 1999.

[17] Y. Zhang and J. Jiang, "Integrated design of reconfigurable fault-tolerant control systems," *Journal of Guidance, Control, and Dynamics*, vol. 24, no. 1, pp. 133–136, 2001.

[18] G. Tao, S. Chen, and S. M. Joshi, "An adaptive actuator failure compensation controller using output feedback," *IEEE Transactions on Automatic Control*, vol. 47, no. 3, pp. 506–511, 2002.

[19] N. E. Wu, Y. Zhang, and K. Zhou, "Detection, estimation, and accommodation of loss of control effectiveness," *International Journal of Adaptive Control and Signal Processing*, vol. 14, no. 7, pp. 775–795, 2000.

[20] B. Jiang and F. N. Chowdhury, "Fault estimation and accommodation for linear MIMO discrete-time systems," *IEEE Transactions on Control Systems Technology*, vol. 13, no. 3, pp. 493–499, 2005.

[21] M. Mahmoud, J. Jiang, and Y. Zhang, "Stabilization of active fault tolerant control systems with imperfect fault detection and diagnosis," *Stochastic Analysis and Applications*, vol. 21, no. 3, pp. 673–701, 2003.

[22] M. V. Basin and M. A. Pinsky, "Stability impulse control of faulted nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 43, no. 11, pp. 1604–1608, 1998.

[23] B. Jiang and F. N. Chowdhury, "Parameter fault detection and estimation of a class of nonlinear systems using observers," *Journal of the Franklin Institute*, vol. 342, no. 7, pp. 725–736, 2005.

[24] B. Jiang, M. Staroswiecki, and V. Cocquempot, "Fault accommodation for nonlinear dynamic systems," *IEEE Transactions on Automatic Control*, vol. 51, no. 9, pp. 1578–1583, 2006.

[25] P. Mhaskar, C. McFall, A. Gani, P. D. Christofides, and J. F. Davis, "Isolation and handling of actuator faults in nonlinear systems," *Automatica*, vol. 44, no. 1, pp. 53–62, 2008.

[26] P. Mhaskar, A. Gani, N. H. El-Farra, C. McFall, P. D. Christofides, and J. F. Davis, "Integrated fault-detection and fault-tolerant control of process systems," *AIChE Journal*, vol. 52, no. 6, pp. 2129–2148, 2006.

[27] A. Fekih, "Effective fault tolerant control design for nonlinear systems: application to a class of motor control system," *IET Control Theory & Applications*, vol. 2, no. 9, pp. 762–772, 2008.

[28] P. Mhaskar, A. Gani, and P. D. Christofides, "Fault-tolerant control of nonlinear processes: performance-based reconfiguration and robustness," *International Journal of Robust and Nonlinear Control*, vol. 16, no. 3, pp. 91–111, 2006.

[29] P. Mhaskar, "Robust model predictive control design for fault-tolerant control of process systems," *Industrial and Engineering Chemistry Research*, vol. 45, no. 25, pp. 8565–8574, 2006.

[30] H.-J. Ma and G.-H. Yang, "FTC synthesis for nonlinear systems: sum of squares optimization approach," in *Proceedings of the IEEE Conference on Decision and Control*, pp. 2645–2650, New Orleans, La, USA, 2007.

[31] M. A. Demetriou, K. Ito, and R. C. Smith, "Adaptive monitoring and accommodation of nonlinear actuator faults in positive real infinite dimensional systems," *IEEE Transactions on Automatic Control*, vol. 52, no. 12, pp. 2332–2338, 2007.

[32] P. G. de Lima and G. G. Yen, "Accommodating controller malfunctions through fault tolerant control architecture," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 2, pp. 706–722, 2007.

[33] M. Benosman and K. Lum, "On-line references reshaping and control re-allocation for nonminimum phase nonlinear fault tolerant control," in *Proceedings of the 17th IFAC Word Congress*, pp. 2563–2569, Seoul, Korea, July 2008.

[34] A. Shumsky, "Algebraic approach to the problem of fault accommodation in nonlinear systems," in *Proceedings of the 17th IFAC Word Congress*, pp. 1884–1889, Seoul, Korea, July 2008.

[35] H. Yang, V. Cocquempot, and B. Jiang, "Robust fault tolerant tracking control with application to hybrid nonlinear systems," *IET Control Theory & Applications*, vol. 3, no. 2, pp. 211–224, 2009.

[36] Y. Zhang and J. Jiang, "Bibliographical review on reconfigurable fault-tolerant control systems," in *Proceeding of the 5th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes*, pp. 265–276, Washington, DC, USA, 2003.

[37] R. Patton, "Fault-tolerant control systems: the 1997 situation," in *IFAC Symposium on Fault Detection Supervision and Safety for Technical Processes*, pp. 1033–1055, Hull, UK, 1997.

[38] J. Lunze and J. H. Richter, "Reconfigurable fault-tolerant control: a tutorial introduction," *European Journal of Control*, vol. 14, no. 5, pp. 359–386, 2008.

[39] M. Benosman and K.-Y. Lum, "Passive actuators' fault tolerant control for affine nonlinear systems," to appear in *IEEE Transactions on Control Systems Technology*.

[40] Q. Zhao and J. Jiang, "Reliable state feedback control system design against actuator failures," *Automatica*, vol. 34, no. 10, pp. 1267–1272, 1998.

[41] P. S. Maybeck, "Multiple model adaptive algorithms for detecting and compensating sensor and actuator/surface failures in aircraft flight control systems," *International Journal of Robust and Nonlinear Control*, vol. 9, no. 14, pp. 1051–1070, 1999.

[42] V. Dardinier-Maron, F. Hamelin, and H. Noura, "A fault-tolerant control design against major actuator failures: application to a three-tank system," in *Proceedings of the 38th IEEE Conference on Decision and Control*, vol. 4, pp. 3569–3574, Phoenix, Ariz, USA, 1999.

[43] G. Tao, S. M. Joshi, and X. Ma, "Adaptive state feedback and tracking control of systems with actuator failures," *IEEE Transactions on Automatic Control*, vol. 46, no. 1, pp. 78–95, 2001.

[44] M. Maki, J. Jiang, and K. Hagino, "A stability guaranteed active fault-tolerant control system against actuator failures," *International Journal of Robust and Nonlinear Control*, vol. 14, no. 12, pp. 1061–1077, 2004.

[45] J. Lunze and T. Steffen, "Control reconfiguration after actuator failures using disturbance decoupling methods," *IEEE Transactions on Automatic Control*, vol. 51, no. 10, pp. 1590–1601, 2006.

[46] W. Chen and M. Saif, "Adaptive actuator fault detection, isolation and accommodation in uncertain systems," *International Journal of Control*, vol. 80, no. 1, pp. 45–63, 2007.

[47] F. Bateman, H. Noura, and M. Ouladsine, "A fault tolerant control strategy for an unmanned aerial vehicle based on a sequential quadratic programming algorithm," in *Proceedings of the 47th IEEE Conference on Decision and Control*, pp. 423–428, Cancun, Mexico, 2008.

[48] T. Miksch, A. Gambier, and E. Badreddin, "Real-time implementation of fault-tolerant control using model predictive control," in *Proceedings of the 17th IFAC Word Congress*, pp. 11136–11141, Seoul, Korea, July 2008.

[49] J. Richter and J. Lunze, "Reconfigurable control of Hammerstein systems after actuator faults," in *Proceedings of the 17th IFAC Word Congress*, pp. 3210–3215, Seoul, Korea, July 2008.

[50] X. Zhang, M. M. Polycarpou, and T. Parisini, "Integrated design of fault diagnosis and accommodation schemes for a class of nonlinear systems," in *Proceedings of the 40th IEEE Conference on Decision and Control*, vol. 2, pp. 1448–1453, Orlando, Fla, USA, 2001.

[51] D. Ye and G. H. Yang, "Adaptive fault-tolerant control for a class of nonlinear systems with time delay," *International Journal of Systems Science*, vol. 39, no. 1, pp. 43–56, 2008.

[52] R. Sepulchre, M. Janković, and P. V. Kokotović, *Constructive Nonlinear Control*, Communications and Control Engineering Series, Springer, Berlin, Germany, 1997.

[53] C. I. Byrnes, A. Isidori, and J. C. Willems, "Passivity, feedback equivalence, and the global stabilization of minimum phase nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 36, no. 11, pp. 1228–1240, 1991.

[54] T.-C. Lee and Z.-P. Jiang, "A generalization of Krasovskii-LaSalle theorem for nonlinear time-varying systems: converse results and applications," *IEEE Transactions on Automatic Control*, vol. 50, no. 8, pp. 1147–1163, 2005.

[55] H. Khalil, *Nonlinear Systems*, Prentice-Hall, Upper Saddle River, NJ, USA, 3rd edition, 2002.

[56] A. Teel and L. Praly, "Tools for semiglobal stabilization by partial state and output feedback," *SIAM Journal on Control and Optimization*, vol. 33, no. 5, pp. 1443–1488, 1995.

[57] M. Vidyasagar, *Nonlinear Systems Analysis*, Prentice-Hall, Upper Saddle River, NJ, USA, 2nd edition, 1993.

[58] A. Isidori, *Nonlinear Control Systems*, Communications and Control Engineering Series, Springer, Berlin, Germany, 2nd edition, 1989.

[59] M. M. Polycarpou, "Fault accommodation of a class of multivariable nonlinear dynamical systems using a learning approach," *IEEE Transactions on Automatic Control*, vol. 46, no. 5, pp. 736–742, 2001.

[60] R. A. Freeman and P. V. Kokotović, *Robust Nonlinear Control Design: State-Space and Lyapunov Technique*, Systems & Control: Foundations & Applications, Birkhäuser, Boston, Mass, USA, 1996.

[61] N. H. El-Farra and P. D. Christofides, "Coordinating feedback and switching for control of hybrid nonlinear processes," *AIChE Journal*, vol. 49, no. 8, pp. 2079–2098, 2003.

[62] P. Mhaskar, N. H. El-Farra, and P. D. Christofides, "Predictive control of switched nonlinear systems with scheduled mode transitions," *IEEE Transactions on Automatic Control*, vol. 50, no. 11, pp. 1670–1680, 2005.

[63] A. Armaou and M. A. Demetriou, "Robust detection and accommodation of incipient component and actuator faults in nonlinear distributed processes," *AIChE Journal*, vol. 54, no. 10, pp. 2651–2662, 2008.

[64] N. H. El-Farra, M. A. Demetriou, and P. D. Christofides, "Actuator and controller scheduling in nonlinear transport-reaction processes," *Chemical Engineering Science*, vol. 63, no. 13, pp. 3537–3550, 2008.

[65] N. H. El-Farra, "Integrated fault detection and fault-tolerant control architectures for distributed processes," *Industrial and Engineering Chemistry Research*, vol. 45, no. 25, pp. 8338–8351, 2006.

[66] N. H. El-Farra and S. Ghantasala, "Actuator fault isolation and reconfiguration in transport-reaction processes," *AIChE Journal*, vol. 53, no. 6, pp. 1518–1537, 2007.

Review Article

Advances in Structural Control in Civil Engineering in China

Hongnan Li and Linsheng Huo

*State Key Laboratory of Coastal and Offshore Engineering, Dalian University of Technology,
Dalian 116024, China*

Correspondence should be addressed to Hongnan Li, hnli@dlut.edu.cn

In the recent years, much attention has been paid to the research and development of structural control techniques with particular emphasis on alleviation of wind and seismic responses of buildings and bridges in China. Structural control in civil engineering has been developed from the concept into a workable technology and applied into practical engineering structures. The aim of this paper is to review a state of the art of researches and applications of structural control in civil engineering in China. It includes the passive control, active control, hybrid controland semiactive control. Finally, the possible future directions of structural control in civil engineering in China are presented.

1. Introduction

Civil engineering structures located in environments where earthquakes or large wind forces are common will be subjected to serious vibrations during their lifetime. These vibrations can range from harmless to severe with the later resulting in serious structural damage and potential structural failure. The traditional method of antiseismic technique is to increase the stiffness of structures by enlarging the section of columns, beams, shear walls, or other elements, which will enhance the seismic load because of the added mass to structures. As a result, although the cost of structures with traditional antiseismic technique is increased a lot, the safety level of structures is less improved. Another disadvantage of the traditional antiseismic technique is that it focuses on the protection of the structure but neglects the facilities inside the structure. Hence, it cannot be used in some structures whose facilities inside them are very important, such as hospitals, city lifeline engineering, nuclear plants, museum buildings, and the buildings with precise instruments.

Even though engineers cannot design a building which is damage-proof during earthquakes and strong winds, the structural control is promising in reducing the vibration of structures. Different from the traditional antiseismic method, the structural control technique suppresses the structural vibration by installing some devices, mechanisms, substructures in the structure to change or adjust the dynamic performance of the structure. The structural control system is commonly classified by its device type resulting in four general control types: passive, active, hybrid, and semi-active control. An active control system is the one in which an external source power control actuators apply forces to the structure in a prescribed manner including active tendon system (ATS) and active mass damper (AMD). A passive control system does not require an external power source, such as base isolation method, energy dissipation devices, tuned mass damper (TMD), and tuned liquid damper (TLD). The hybrid control implies the combined use of active and passive control systems. Semiactive control systems are a class of active systems in which only small magnitude of external energy is needed to change the parameters of control system, such as active variable stiffness (AVS) system and active variable damper (AVD) system.

In the recent years, serious efforts have been undertaken in structural control and fruitful achievements have been made in China. Structural control has been developed from theoretical analysis and experimental research into engineering applications. A set of techniques including base isolation, energy dissipation, tuned mass dampers, tuned liquid dampers, active and semi-active control methods have been used in newlybuilt structures or existing buildings, bridges, facilities, and other structures in China. In this paper, the state of art of vibration control techniques, including theoretical and experimental studies, practices in civil engineering are reviewed. The possible future directions of structural control in civil engineering is discussed.

2. Passive Control

2.1. Base Isolation

The base isolation system was developed as one of remarkable technologies to reduce the seismic load of building and equipment. There are five kinds of materials that have been used for isolators in China, including sand layer, graphite lime mortar layer, slide friction layer, roller and rubber bearing. The rubber bearing is the one used mostly in China. Relatively easy to be manufactured, isolation bearings are made by vulcanization bonding of sheets of the rubber to thin steel reinforcing plates. The bearings are very stiff in the vertical direction, while quite flexible in the horizontal direction. Under the excitation of horizontal earthquakes, the fundamental vibration period of an isolated structure can be shifted from a short-period range to a long-period range. Thus, the horizontal seismic response of the isolated structure is significantly less than the unisolated structure. The significant advantages of structures with rubber bearing isolators are summarized as follows [1].

(1) The isolated structures are safer in strong earthquake. The isolators are very effective to reduce the seismic response of structures and can prevent the structures from damage or collapse. Compared to the traditional antiseismic structures, the responses of isolated structures can be reduced to 1/2~1/8 of the one of traditional structures, according to the testing results and the records in real earthquake.
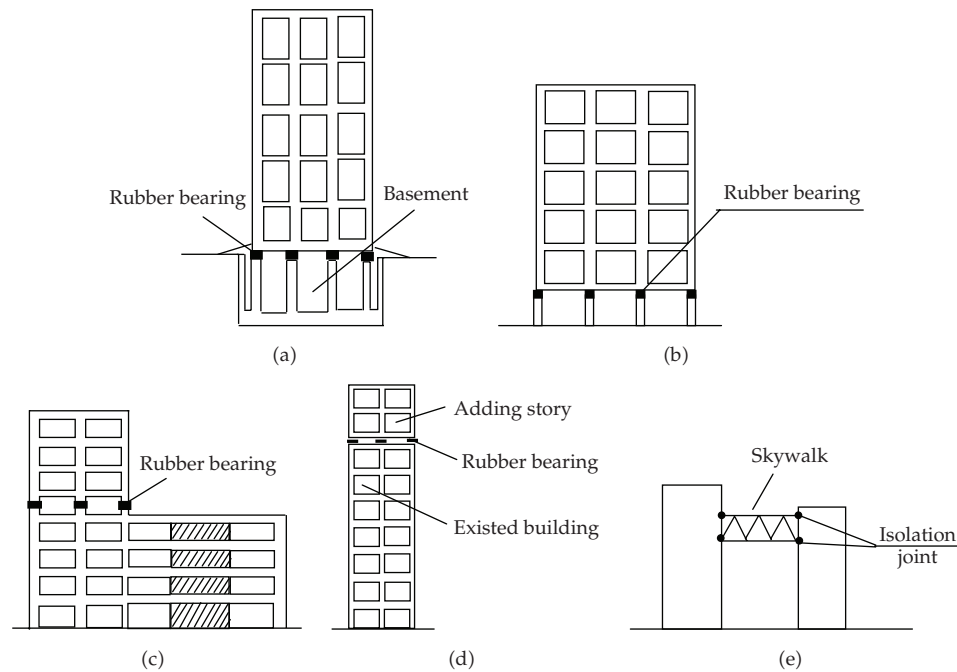
Figure 1: Different locations of Isolation Layer in Structure. (a) Basement isolation; (b) Story isolation with isolators on the top of the first story; (c) Story isolation with isolators on certain story of supper structures; (d) Top isolation; (e) Skywalk linking isolation.

(2) The building cost of isolation structures can be saved 3% ~ 15% of the cost of the general buildings according to the final statistics results of 30 buildings with rubber bearings completed in southern, western, and northern China.

(3) The seismic isolation rubber bearing system has wide ranges of application, both in newly designed structures and existing structures, important buildings and civil buildings especially for house buildings, for protecting the building structures, and for protecting the facilities inside the building.

(4) The safely working life of rubber bearings is over 70–100 years according to the permanent testing and investigation, which is larger than the working life of structure itself.

In order to solve some problems for more wide application of isolation systems, a great number of tests for rubber bearings and shaking table tests for building models were carried out by Zhou and Xian [2]. The test of mechanical characteristics for isolators includes the compression tests, compression with shear cycle loading tests. The test of durability for isolators includes low cycle fatigue tests, creep tests, and ozone aging tests. The test of structural system includes shaking table tests for large-scale structural model.

There are five kinds of locations of layer with rubber bearings in China [3]. Isolation layer is located on the base of building, which is the most common method. The isolation layer can be located on the certain story of the basement, which is used in many structures as shown in Figure 1(a). The isolation can also be located on the top of the first story (Figure 1(b)) or certain story of superstructure (Figure 1(c)). The isolation layer located

on the top of the building (Figure 1(d)) is always used to add 1-2 stories on the top of existed building for seismic retrofit. Isolation layer located at the joints between skywalk and buildings (Figure 1(e)) is used to separate the different mode shapes of buildings connected by skywalk.

In the case of a seismic isolated building with a large height-to-width ratio (HWR), the overturning moment at the level of seismic isolated layer may exceed the overturning resistance supplied by the gravity, which will result in the disconnection of bearings from the superstructure and produce an internal damages of rubber layers, even lead to a destroy of the entire building. The limits of the HWR for isolated building with the laminated rubber bearings under different conditions subjected to earthquake excitations were investigated by Li and Wu [4]. According to the numerical results, it has been found that the effects of site soil conditions on the HWR limit values are very important, in which the softer the site is, the smaller the HWR limit value is under different seismic intensities. The predominant period of isolated building also plays a considerable role in the HWR limit value namely, the isolated building with a longer period may have a relatively large HWR value. The stiffness of superstructure has little effects on the HWR limit value.

Most structures are simplified as planar models in base isolation design. However, a lot of structures are eccentric due to plan irregularity, mass eccentricity, elevation irregularity, and so forth The earthquake is essentially multidimensional and so is the structural response excited by earthquake, which will result in the torsionally coupled vibration that cannot be neglected. Based on many numerical analyses on different soil sites during the action of earthquakes, simplified formulate of torsional seismic actions for base-isolated eccentric structures is presented [5]. It is shown that the proposed formulations have good accuracy compared with others. It provides a convenient and simplified method for the structural design of base-isolated eccentric buildings with rubber pads.

Design principles of base isolation and energy dissipation systems are also included in Chinese Code for Seismic Design of Buildings published by the agency of Ministry of Construction in 2001. The design codes for base isolated highway bridges and base isolated railway bridges were also published by the agencies of Ministry of Transportation and Ministry of Railway. Currently, there are three different sets of technical codes on seismic isolation in China [3].

(1) *Technical Specification for Seismic Isolation with Laminated Rubber Bearing Isolators* (CECS 126: 2001). This is the national code for design and construction of buildings and bridges with seismic isolation in China.

(2) *Standard of Laminated Rubber Isolators* (JB 118-2000). This is the national standard of isolators for laminated rubber bearing in China.

(3) *Seismic Isolation and Energy Dissipation for Building Design* (Chapter 12 in the *code for seismic design of buildings*, GB 50011-2001). This is a part of national code in China for seismic design of buildings.

Some main introductions for all these three codes (standards) on seismic isolation in China are described as below [3].

(1) Provide the design methods of seismic isolation for buildings, bridges, special structures, and industry facilities.

(2) Provide the design methods of seismic isolation for new structures and the retrofit of existed structures.

(3) Allow to follow three design levels depending on the importance of structures and requirements of owners in the areas with different economic situation in China. Level 1, for general structures, using isolation will save building cost about 3% ~ 15%. Level 2, for important structures, using isolation will increase building cost 3% ~ 5%. Level 3, for especially important structures, using isolation will increase building cost 5% ~ 7%. But the isolation buildings designed by any level will increase the seismic safety about 2 ~ 8 times compared with the traditional antiseismic buildings.

(4) Provide two methods of structural analysis for seismic isolation of structures, equivalent shear method, and time-history analysis. The equivalent shear method is achieved by static analysis and only applicable for structures with height less than 40 m or 10 stories, regular configuration, and predominant shear deformations. Whereas, the time-history analysis can be used for all kinds of structures.

(5) Allow reducing the seismic shear force for designing superstructure, which can save the building cost of general civil buildings, especially in some poor economic areas.

(6) Control the maximum horizontal shear displacement $D_{\max}$ of isolation layer. The value of $D_{\max}$ will not be larger than 0.55 times of diameter of bearing and 300% shear strain deformation of bearings. $D_{\max}$ should be the total displacement including both translation and torsion of structural system.

The first multistory house building with rubber bearing in China was completed in 1993, as shown in Figure 2 [3]. It is an 8-story building supported by 23 rubber bearings with different diameters from 60 cm to 80 cm. The structure has undergone two strong earthquakes in 1994 and 1996 and is still employed now without any damage. The isolated buildings with the largest area in the world are named Isolation House Buildings on Subway Hub (IHBSH) as shown in Figure 3, which is located nearly the center of Beijing [3]. There is a very large platform composed by an RC frame with two sorties, which is used to put all equipment and facilities for railway hub in it and cover the noise from railway hub trains. The size of platform is 1500 m wide and 2000 m long. There are 50 house buildings built on the top floor of the platform. The floor area of all isolation house buildings is approximately 480000 m^2 which is the largest area using seismic isolation in the world. The rubber bearing layer is located on the top floor of the platform, as shown in Figure 4. The great benefits are achieved by using story isolation for this building complex including: increasing the safety of structures to 4 times of traditional design schemes, saving 25% of the construction cost, raising the number of stories from 6 to 9, and solving the environmental problems of railway vibration and noise in the city center.

The application of seismic isolators in China has proved that the isolation system is more safe, economic, and reasonable than the traditional structural system. More than 500 full-scale implementations of base isolated buildings and bridges have been accomplished to alleviate the earthquake and traffic induced response so far [6].

2.2. Energy Dissipation

An energy dissipation system is formed by adding some energy dissipaters into the structure. The energy dissipaters provide the structure with large amounts of damping which will dissipate most vibration energy from vibration sources and ensure the safety of structures in
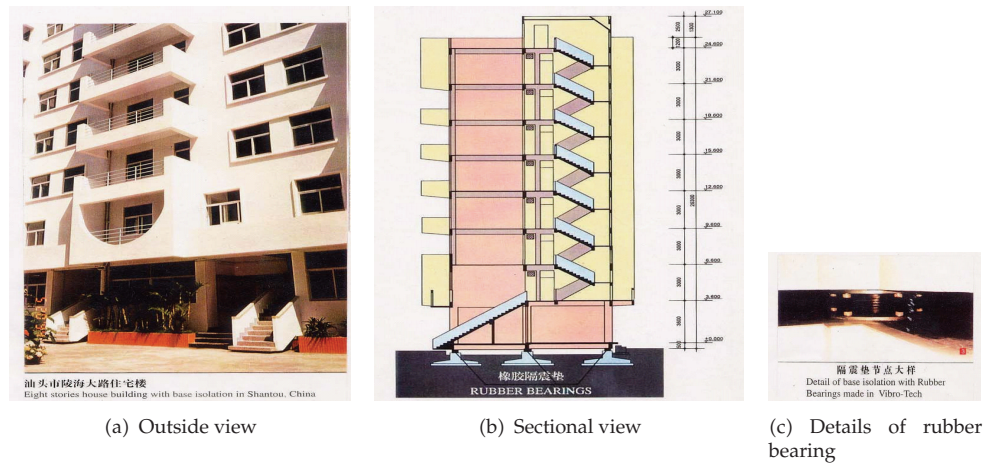
(a) Outside view   (b) Sectional view   (c) Details of rubber bearing

Figure 2: The first multistory house building with rubber bearing in China.

Figure 3: Front elevation of IHBSH.

earthquake or the comfort level in wind. Many kinds of dampers, including friction dampers, viscoelastic dampers, viscous fluid dampers, lead dampers and metallic dampers have been developed in China.

Friction provides an excellent mechanism of energy dissipation and has been used for many years in automotive brakes to dissipate kinetic energy of motion. This kind of dampers are made of ordinary materials with simple mechanical configurations; hence they are applicable for multi-story and high-rise buildings. The behaviors of two different energy dissipation dampers based on friction as shown in Figure 5 were studied by Zhou and Liu [7]. The results showed that the friction energy dissipation dampers can dissipate a great deal of earthquake input energy and the hysteretic behavior is stable. To satisfy the energy dissipation requirements for earthquake in different seismic intensities of China, double-phase friction dampers were proposed by Zhang and Yang [8]. The experimental results showed that the proposed double phase friction damper can reduce the structural vibrations both in small and strong earthquake. Wu and Ou [9] presented a kind of pseudofriction damper to reduce the effects of additional stiffness in Pall friction dampers. Based on the geometry nonlinearity of four-link mechanism of Pall-typed friction damper, the influence of frame displacement, brace stiffness, damper slip force of a novel pseudoviscous frictional damper as shown in Figure 6, is analyzed by Wu et al. [10]. The results showed that the novel frictional damper has great hysteretic characteristics with remarkably smaller tension force of brace than that of pall-typed friction damper, which is beneficial for frame columns to improve its ductility. A damper with innovative composite friction material has been developed by Xian et al. [11]. A full-scale test of this innovative friction damper and the
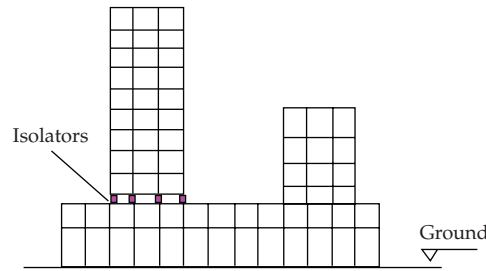
Figure 4: Stories Isolation used in IHBSH.

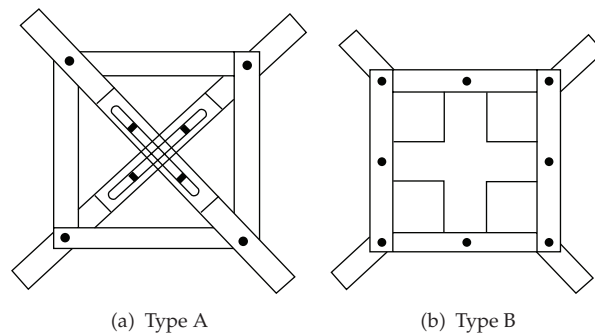(a) Type A   (b) Type B

Figure 5: Two types of friction dampers.

shaking table test for structural model with the damper were implemented, which showed that the seismic responses of structures with dampers can be reduced 30% ~ 40%. The friction dampers have been applied in engineering structures in China. For example, 134 passive friction dampers with 100 kN to 200 kN sliding force capacity have been implemented in Shenyang Government Office Buildings in 1997 to retrofit this building, as shown in Figure 7. The pseudodynamic tests of a 1/3 scale model of the building with and without friction dampers were carried out to examine the seismic capacity of the retrofitted buildings [12].

Lead dampers have ideal elastic-plastic property with rectangular hysteresis curve. The yielding limit of lead dampers cannot be adjusted after designed. Hence, in a strong earthquake, the energy dissipation capacity of dampers will be low because of large deformation. Whereas, in a mild earthquake, the dampers cannot yield to dissipate energy. Innovative bend-shear and shear-type lead dampers which can attenuate the vibration of buildings both in mild and strong earthquakes were developed by Zhou et al. [13].

Viscoelastic dampers have the merits of dissipating energy at all deformation levels. Therefore viscoelastic dampers can be applied in both wind and seismic protection. Some commercial products of viscoelastic dampers in China are shown in Figure 8. The mechanical model, design technique, and optimum method for viscoelastic dampers have been developed in China [14, 15]. Viscoelastic dampers were also applied to engineering structures for suppressing the structural vibration caused by wind or earthquake in China, such as Suqian Communication Building (Figure 9) and Chaoshan Xinhe Building (Figure 10) [16, 17].

Fluids can also be used to dissipate energy; hence many kinds of viscous fluid dampers have been proposed. Commercial viscous fluid dampers with different capacities were
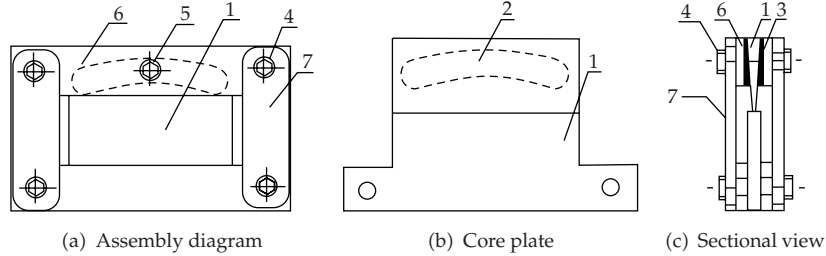
(a) Assembly diagram   (b) Core plate   (c) Sectional view

Figure 6: Pseudoviscous frictional damper with convex frictional surface.

(a)   (b)

Figure 7: Shenyang Government Office Building retrofitted by friction dampers.

also available in China (Figure 11). Viscous dampers have been used in some engineering structures in China, including Beijing Yingtai Center, Yizhenyuan building of Huaiyin City, Beijing Exhibition Building, an office building of Suqian City, Construction Building of Suqian City, an office building of Xi'an City, and Beijing Restaurant. The centre tower building of Beijing Yintai Center is a 60-story steel structure with 248 m in height and susceptible to wind, so 73 sets of viscous fluid dampers were installed in it (Figure 12) [18].

Metallic dampers dissipate energy input to a structure from an earthquake through inelastic deformation of metals. Based on the experimental study and theoretical analysis of the fatigue properties of X-shape steel plate energy dissipaters, the nonlinear method of steel plate of the energy dissipater was established and the strain-fatigue parameters are determined, which provide an important base for the design of mild steel energy dissipaters [19]. In most cases, Metallic dampers dissipate energy through the nonlinear property of steel plate after yielding out of plane. Dual-functional mild steel dampers are presented by G. Li and H. Li [20], as shown in Figures 13 and 14, which is to strengthen the initial stiffness by extending the steel plate in its own plane and to increase the energy-dissipating ability by changing the geometric shape of the steel plate. Results from theoretical analysis and quasistatic experiments showed that these types of mild metallic dampers not only provide certain stiffness in normal use, but also are of good ability of the seismic energy dissipation. Dual-functional mild metallic dampers have been installed in No. 3 Experiment Building of Dalian University of Technology (DUT) in China to increase the stiffness of the first floor and suppress the earthquake induced vibration [21], as shown in Figure 15.

For some old designed buildings based on the dated building design codes in which the seismic protection was seldom taken into account; both the land developers and engineers have to determine whether to demolish them and then set up new ones or reuse the original buildings through making some innovations to meet the current need. Due to economy, many
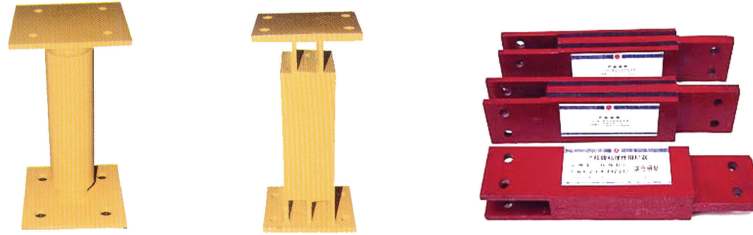
Figure 8: Some commercial products of viscoelastic dampers in China.

Figure 9: Viscoelastic dampers in Suqian Communication Building.

old existing buildings have been increased with new stories on the top of them in China. A new type of energy-dissipated structural system for the existing building with the story-increased frame was presented [22], in which the sliding-friction layer between the lowest increased floor of outer frame structure and roof of original building was applied, and energy-dissipated dampers were used for the connections between the columns of outer frame and each floor of original building, as shown in Figure 16. The experimental results showed that friction and energy-dissipated devices are very effective in reducing the seismic response and dissipating the input energy of model structure. A four-story office building, built up in 1950s without the seismic design along Beiling Street in Shenyang City of China because there was no any seismic codes at that time in China, was designed and analyzed with an increased four-story outer frame structure on its top by using the passive control method presented. The existing building is composed of brick masonry structure; its design is not in compliance with the current Seismic Code of China. The seismic protection intensity on this building site is VII degree, and the site soil belongs to the type II in Chinese Code. The practical building story-increased structure after construction is shown in Figure 17. The results showed that the good seismic-reduction effectiveness on displacements and accelerations can be achieved by installing the friction and energy-dissipated devices in the story-increased structures.

Figure 10: Viscoelastic dampers in Chaoshan Xinhe Building.
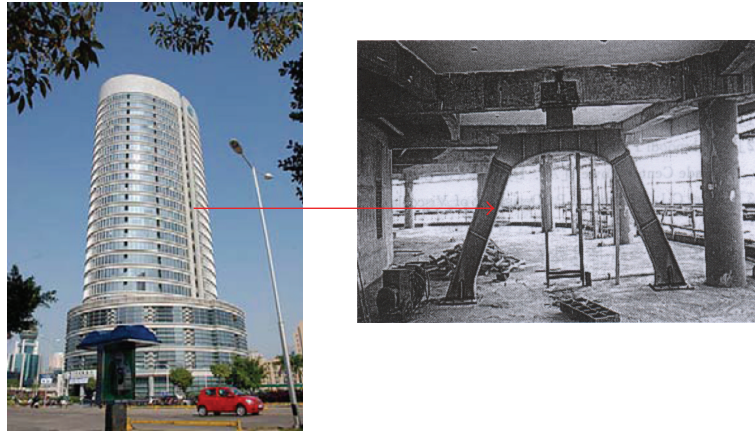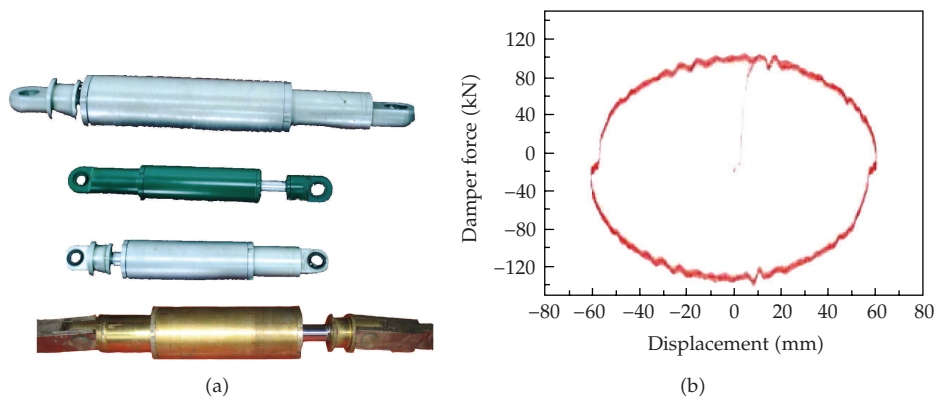

(a)   (b)

Figure 11: Some commercial viscous dampers and hysteresis curve.
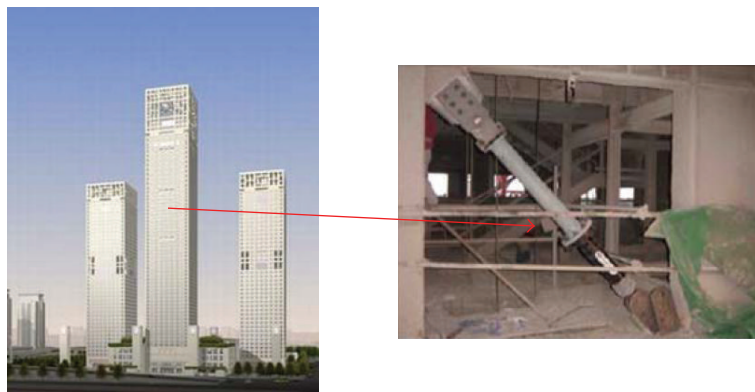
Figure 12: Viscous fluid dampers installed in Beijing Yintai Center.

(a)

(b)
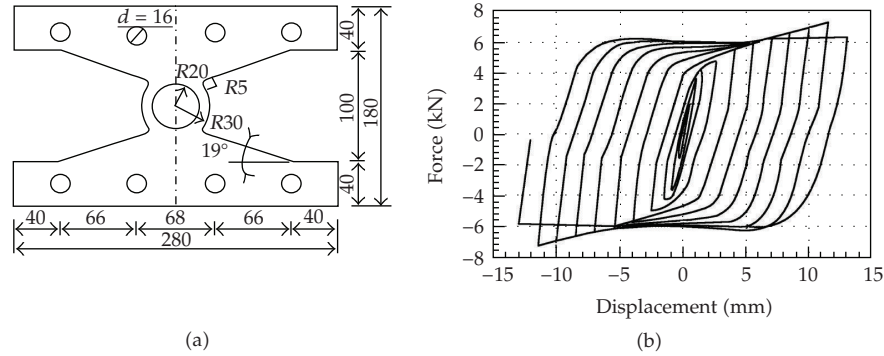
Figure 13: Circle hole type mild steel damper and its hysteresis curve.

(a)

(b)
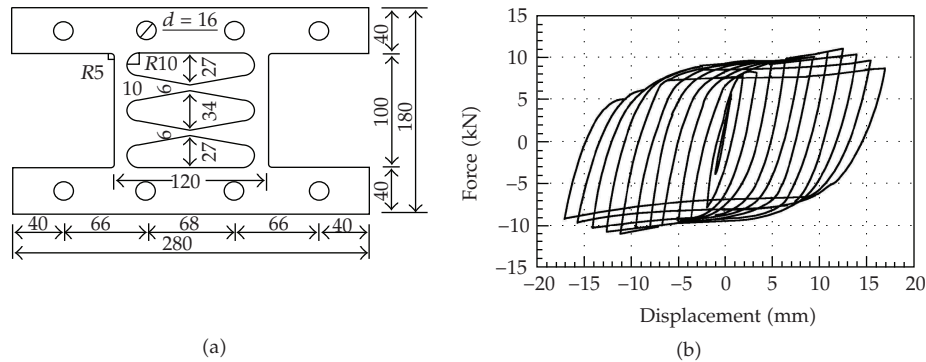
Figure 14: Double X type mild steel damper and its hysteresis curve.

2.3. Tuned Mass Damper

The concept of the tuned mass damper (TMD) consists of a secondary mass with properly tuned spring and damping elements, providing a frequency-dependent hysteresis that increases damping in the primary structure. The frequency of TMD should be tuned to the same as the one of the controlled structures to get the best control results. During an earthquake, TMD will move against the direction of main structural vibration and an inertia force will be acted on the structure to reduce the response of the structure. One TMD can attenuate only the first mode response of a structure with its frequency tuned to the fundamental frequency of the structure. However, the first several modes of a highrise structure are primary and the anticipated response reduction cannot be achieved if only the first mode is controlled. Li and Wang [23] presented the method of using multiple TMDs to control multiple modes of structures and got obvious results of vibration reduction. In China, the TMD system may be formed by adding one or more stories supported by rubber bearings on the roof of main building structure, as shown in Figure 18 [3]. TMD control system was applied in Zhengzhou International Conference and Exhibition Center, as shown in Figure 19. Tuned Mass dampers were also used to suppress the human induced vibration of pedestrian bridges [24–26]. Six sets of TMD system were installed in the box beam of a 52 m-span pedestrian bridge in Shenyang City of China, as shown in Figure 20.

(a)   (b)   (c)
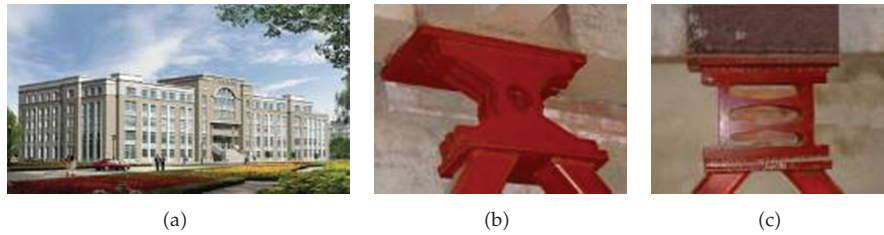
Figure 15: Double function mild metallic dampers in the No. 3 Experiment Building of DUT.
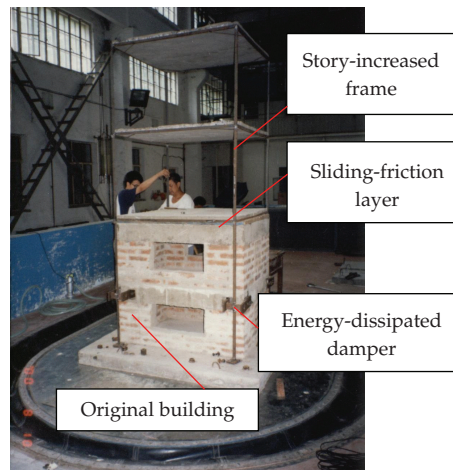
Figure 16: Test model of story-increased system.

2.4. Tuned Liquid Damper

Similar in concept a TMD, the tuned liquid damper (TLD) and tuned liquid column damper (TLCD) impart indirect damping to the system and thus improve structural performance. A TLD absorbs structural energy by means of viscous actions of the fluid and wave breaking. In a TLCD, energy is dissipated by the passage of liquid through an orifice with inherent head loss characteristics.

TLDs are mostly used to reduce the wind induced vibration of structures in engineering. Zhang et al. [27] carried out shaking table test and verified the vibration control result with TLD subjected to earthquake. The research shows that TLD is applicable for long period structure. However, it is not enough to consider only the first mode of structural response. Li et al. [28] presented the multimodal vibration control of structures using multiple TLDs and carried out experimental verification. The experimental results showed that reduction effectiveness of the seismic responses for controlling two modes is better than controlling only a single mode.

The frequency of TLCD is related to the length of liquid in the TLCD tube. The length of TLCD maybe too large to exceed the available space of the building for structures with low frequencies. Yan and Li [29] presented the adjustable frequency tuned liquid column damper by adding springs to the TLCD system, which can modify the frequency of TLCD and expended its application ranges (Figure 21).

Figure 17: The photo of story-increased building.
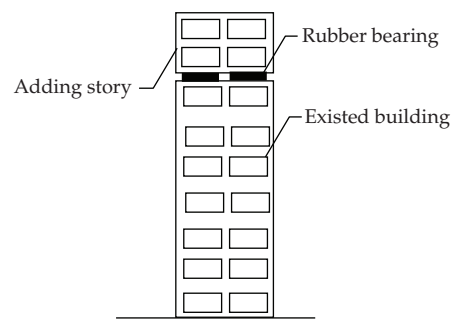
(a)   (b)

Figure 18: Adding story of top of building.

(a)   (b)

Figure 19: TMD system in Zhengzhou Conference and Exhibition.

Figure 20: TMD system in a large span pedestrian bridge.
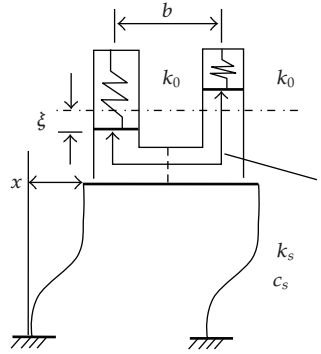
Figure 21: Adjustable frequency TLCD.

To suppress the torsionally coupled responses of structures, Huo and Li [30] analyzed the control performance of circular tuned liquid damper (CTLCD). Subsequently, TLCD and CTLCD were used to control the torsionally coupled vibration of structures subjected to multidimensional earthquake excitations (Figure 22). The optimal parameters of the liquid dampers are designed for the structures excited by bi-directional seismic based on Genetic Algorithms [31].

Recently, TLD control scheme was proposed to reduce wind-induced vibration of Dalian International Trade Center in China, as shown in Figure 23, which is of 81 stories (including one story basement) with the size of 339 m high and 77.7 m long in the east-west direction and 44 m wide in south-north direction. Since the basic wind pressure in the Dalian region is $0.75 \, \text{KN}/\text{m}^2$ and the mansion is slender, that is, the ratio of height over width is 6.7, it is relatively more flexible to large wind-vibration action in the horizontal direction. According to numerical results, the designed TLDs in the building can efficiently reduce the structural vibration, and the effects of vibration reduction are going to be better as the increase of liquid mass in tanks, in which the reduction rates may be as high as 19% for structural acceleration response that influences the human comfort.

3. Active Control

Active control systems are used to control the response of structures to internal or external excitation, such as machinery or traffic noise, wind or earthquakes, where the safety or comfort level of the occupants is of concern. Active control may employ hydraulic, pneumatic, electromagnetic, or motor driven ball-screw actuation. The desirable performance of structural systems can be achieved through active control, whereby actuators excite the structure based on the structural response information measured by sensors. Control of structures involves a number of disciplines, including structural dynamics, control theory, and material engineering. The practice of active structural control is still in doubt, therefore, resulting in the slow shift to application from experimental and theoretical study stage.

Active mass damper (AMD) is one of the most common control device employed in full-scale civil engineering. AMD system is designed and installed in the Nanjing Communication Tower in Nanjing, China (Figure 24). The physical size of the damper was constrained to a ring-shaped floor area with inner and outer radius of 3 m and 6.1 m, respectively. In addition, the damper was elevated off the floor by steel supports with Teflon
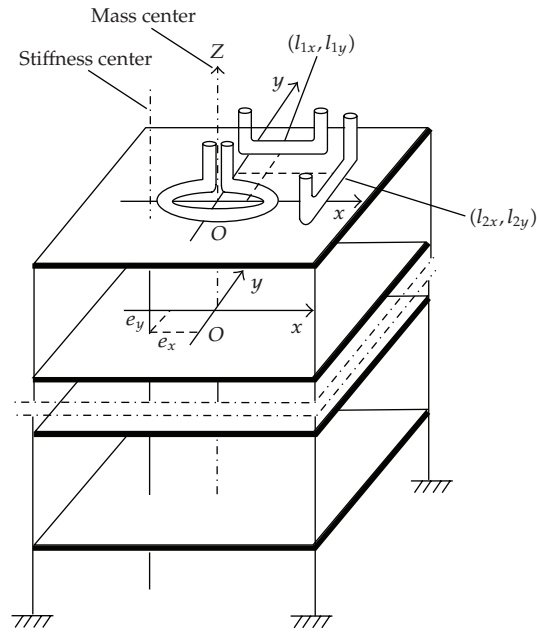
Figure 22: Eccentric structure with liquid dampers.
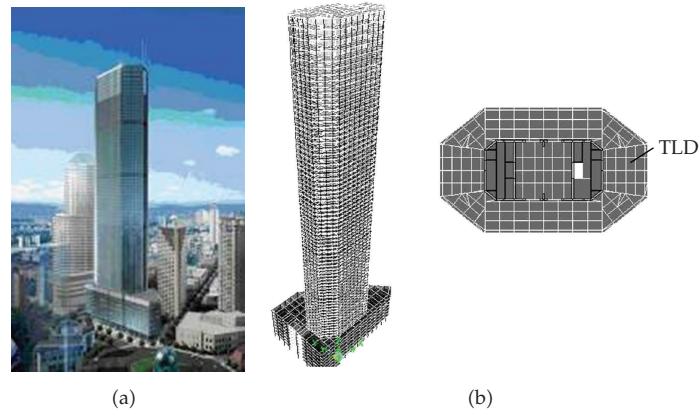
Figure 23: TLD control scheme for Dalian International Trade Center of China.

bearings to allow free access to the floor area. The final design of the active mass damper is shown in Figure 24 [32]. The AMD control methods for eccentric building structures using artificial network, genetic algorithm, and fuzzy logic were presented by Li and Jin [33] and Jin and Zhou [34].

4. Hybrid Control and Semiactive Control

A hybrid control may use active control to supplement and improve the performance of a passive control scheme. Alternatively, passive control may be added to an active control
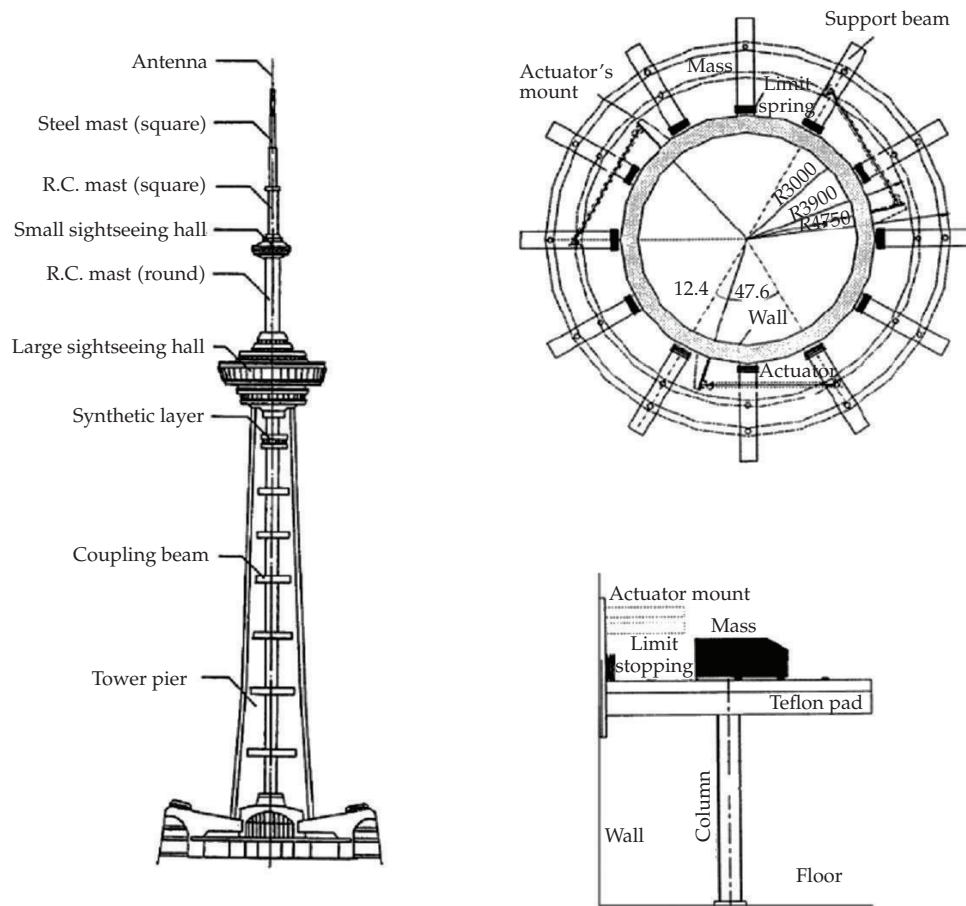
Figure 24: Nanjing Communication Tower with AMD.

scheme to decrease its energy requirements. Hybrid control schemes can sometimes alleviate some of the limitations that exist for either a passive or an active control acting alone, thus leading to an improved solution. One of the hybrid systems is combination of TMD or TLD and active control devices. TMD or TLD is used to control mild responses of structure and active control devices are used to alleviate the large amplitude responses of structures. Hence, the energy required to drive hybrid system is reduced a lot, and the applicability and reliability of the system are increased. According to the characteristics of earthquake response of high-rise buildings with spacious first story, a kind of hybrid control system with hydraulic damper system (HDS) and AMD is proposed. The experimental results of shaking table and simulation analysis indicate that the hybrid control system HDS-AMD has great potentialities for vibration control of civil engineering structures [35]. Guangzhou TV and Sightseeing Tower is a tower currently under construction in Guangzhou City in China, which is total 610 m high including the 450 m-high main tower and a 160 m-high antenna. The tower is a very flexible structure with the first period about 10.03 and susceptible to wind, hence hybrid TMD control system is proposed for the vibration reduction, in which two water tanks are used as mass block, as shown in Figure 25 [36].
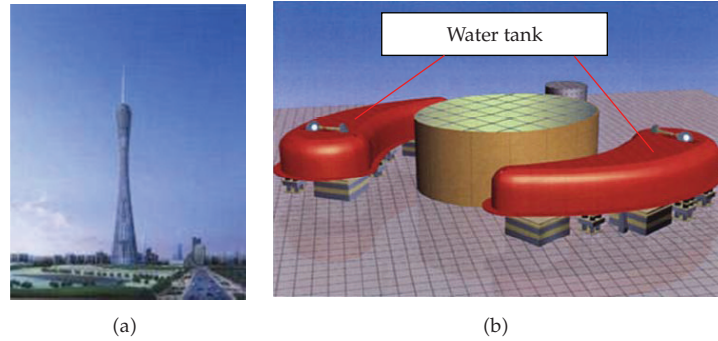
Figure 25: Hybrid TMD control system for Guangzhou TV Tower.

One of the most attractive features of semi-active control systems is that the external energy required for them to operate is usually orders of magnitude smaller than for active systems. The semi-active control system is much more reliable and simpler than general active control system, and more effective to reduce the structural response than other passive control system. Examples of semi-active control device include variable orifice fluid dampers, variable stiffness devices, semi-active tuned liquid dampers, and controllable friction devices. The behavior of active variable stiffness system is experimentally studied by Li and Liu [37]. Both of the structural analysis and a small-scaled model shaking table test indicated that the variable stiffness system not only changed frequencies of controlled structures but also increased damping ratio. A hydraulic actuator with a controllable orifice was developed by Sun [38]. Li et al. [39] experimentally investigated the electromechanical behavior, which is dependent on exciting frequency and applied voltage. To improve the control performance of passive TLCD, Li et al. proposed the semi-active TLCD control strategy based on Artificial Neural Network (ANN) and fuzzy logic [40, 41].

5. Structural Vibration Control with Smart Materials

Smart materials refer to materials that are "responsive." Often the response is the conversion of one form of energy into another in useful quantities. Commonly used smart materials in vibration control include Magneto-rheological (MR) fluids, piezoelectric ceramics, and shape memory alloy. Magnetorheological (MR) dampers are one of the most promising devices for structural control of civil engineering structures. MR dampers have the capability to provide large controllable damping forces and to dynamically change their properties to accommodate varying loading conditions and usage patterns. X.-L. Li and H.-N. Li [42] proposed a new MR mechanical model which is called double-sigmoid. The outstanding feature of the proposed model is a fact that magnitude of control current and a wide range of excitation conditions are under consideration. The identification of parameters is relatively easy and the physical concept of the model is obvious through this method. The proposed double-sigmoid MR mechanical model is as follows:

$$f_d = f_y \frac{1 - e^{-k(\dot{x} + \dot{x}_h)}}{1 + e^{-k(\dot{x} + \dot{x}_h)}} + C_b \cdot \dot{x}, \tag{5.1}$$
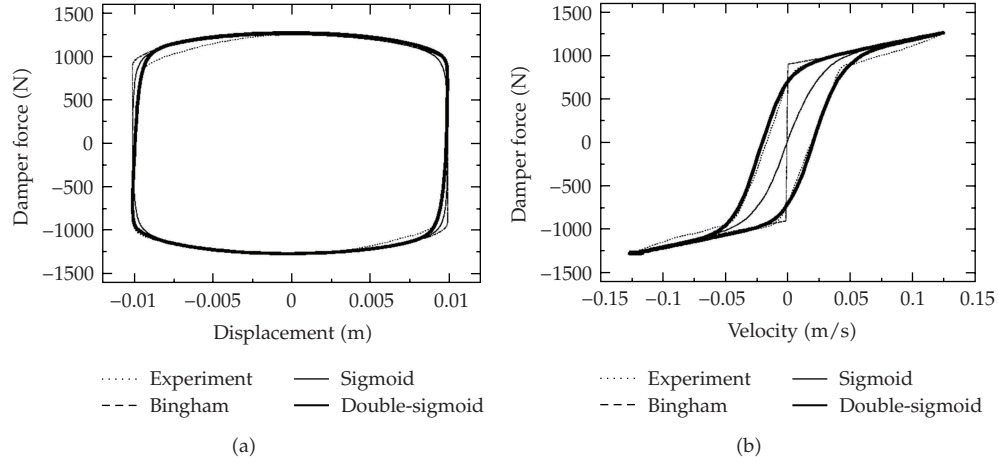
Figure 26: Force-displacement and force-velocity curves of different models.

Figure 27: Eccentric vibration control experiment with MR damper.

where f_d is the force provided by MR dampers, f_y is the yielding force of MR dampers, k is a constant, $\dot{x}$ is the velocity of MR dampers, C_b is the damping coefficient of MR dampers, $\dot{x}_h$ is traversing velocity of force-velocity curve. To verify the accuracy of proposed MR damper model, force-displacement and force-velocity curves for three different models are computed. Comparison of theoretical and experimental results is shown in Figure 26. It can be seen that the phenomena of the hysteresis nonlinear properties in higher velocity region and the saturation characteristic in lower velocity region of MR dampers can be described accurately using proposed model.

MR damper was proposed to reduce the torsional response of eccentric buildings. The shaking table experiment of a 2-story frame-shear wall structure model with an asymmetric stiffness distribution was carried out, as shown in Figure 27 [43]. The experimental results showed that the coupled translation and torsion response were significantly mitigated.

256 MR dampers with 2.26 kN capacity at the cables of the Dongting Lake Cable-stayed Bridge of China to suppress the wind-rain induced dramatic vibration, as shown in Figure 28

(a)

(b)
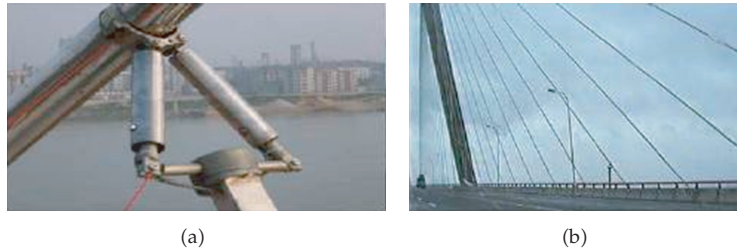
Figure 28: Dongting Lake Bridge with MR dampers.

(a)

(b)

Figure 29: Binzhou Yellow River Highway Bridge with MR dampers.

[44]. 40 MR dampers with 8 kN capacity at the cables of Shandong Binzhou Yellow River Highway Bridge, as shown in Figure 29 [45].

The variable friction that dampers with a PZT actuator has been proposed by Ou [45] and other researchers. Although the specific configuration of the damper is different, the mechanism to adjust sliding force is similar, that is, the nominal pressure force can be adjusted through applying voltage to the PZT layer due to its reverse piezoelectric property. A new design for piezoelectric friction damper was presented by Li et al. [46]. This piezoelectric friction damper was developed by combining the existing slot bolted connector design and the piezoceramic actuator. The schematic of this design is shown in Figure 30. It is composed of tube piezoceramic stack actuators, load cells, preload bolts, an upper plate, a sliding plate, and a lower plate. Three plates are slot-bolted together so that sliding takes place among the sliding plate, upper plate, and lower plate. This type of semiactive piezoelectric can be connected to the beams or floors using chevron braces through the connecting plates. A preload is applied to the piezoceramic actuators to avoid slack and constraining the deformation of actuators.

Shape memory alloys (SMAs) are multifunctional materials as they have self-sensing, self-actuating and energy dissipation properties. Many researchers focus on developing SMA-based isolators with self-centering performance. Two kinds of self-sensing, SMA dampers are developed by Li et al. [47]. The self-sensing SMA dampers have following characteristics: (5.1) the SMA wires are always elongated during vibration no matter which direction of the controlled structure move in; (2) the SMA dampers can self-sense deformation, and thus resulting in a potential way to quantificational assess the safety and damage of the controlled structures post-earthquake hazards; (3) one of SMA dampers can dissipate much more energy because the configuration of the SMA damper with amplifying function deform the SMA wires in the damper larger than the corresponding drift of the controlled structure. Ren et al. presented mechanical model of X type SMA damper and

1. Piezoceramic actuator   5. Connecting plate
2. Preloaded bolt   6. Sliding plate
3. Loadcell   7. Upper plate
4.Beam or floor   8.Lower plate
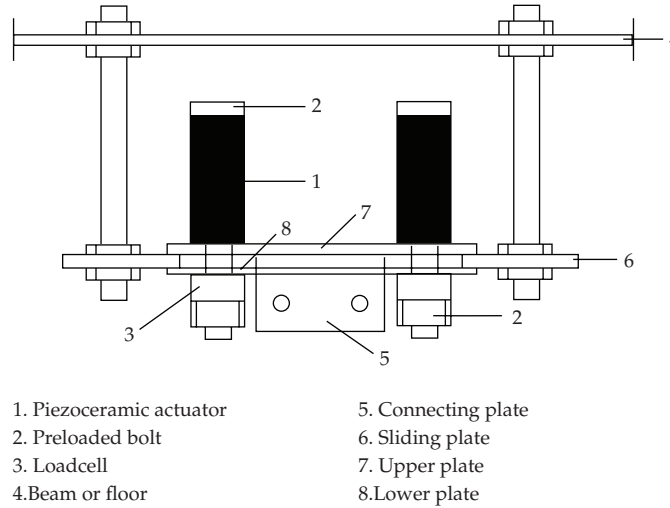
Figure 30: Schematic of a new piezoelectric friction damper.

Restoration SMA wire

Energy dissipation SMA wire

(a)   (b)
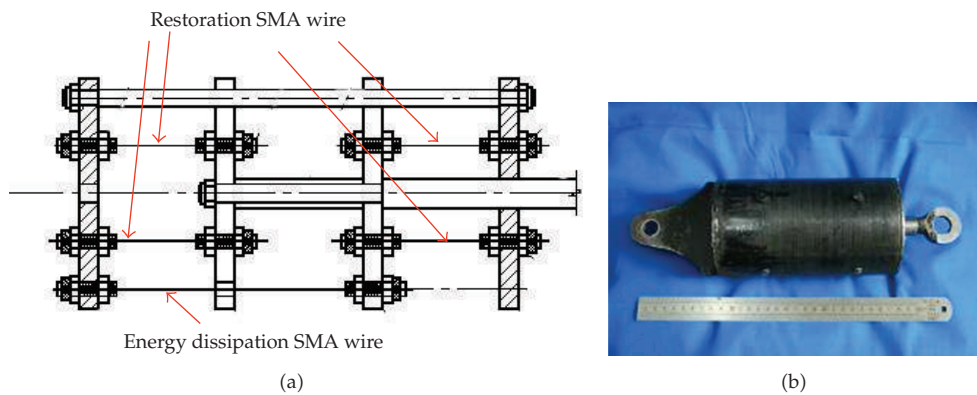
Figure 31: Energy dissipation and recentering SMA dampers.

a one-dimensional strain-rate dependent constitutive model for SMA [48, 49]. A new type of SMA damper was proposed utilizing the energy dissipating and recentering features of superelastic SMA, as shown in Figure 31 [50]. The research results showed that the innovative SMA damper has both the energy dissipating and recentering features with the hysteretic loop under cyclic loading-unloading. The hysteretic behaviors of the damper can be modified to best fit the needs for passive structural control applications by adjusting the pretension of the Nitinol wires. An innovative multidirectional shape memory alloy damper was developed by Ren et al., as shown in Figure 32 [51], which can dissipate energy on the any direction of horizontal plane and recenter on vertical and torsional direction.

6. Conclusions

The researches on structural vibration control have made great achievements, and some control devices have been applied to full-scale engineering structures in the recent years in

Plan view

(a)

Site view

(b)
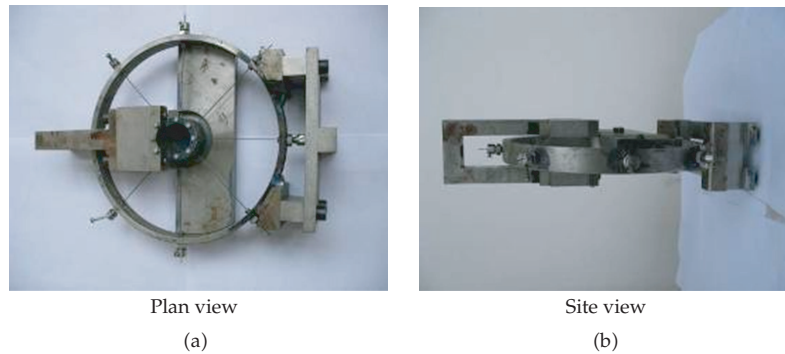
Figure 32: Multidimensional SMA damper.

China. However, most researches focus on linear vibration of structures. Studies on control theories for nonlinear systems are still very limited so far and should be paid more attention in the future. Performance-based method for all kinds of control systems should be developed to satisfy the engineering requirements. More smart materials and new technologies should be used to control the structural vibration subjected to dynamic load.

Acknowledgments

References

[1] F. L. Zhou, *Seismic Control of Structures*, Chinese Seismic, Beijing, China, 1997.

[2] F. L. Zhou and Q. L. Xian, "Recent development on seismic isolation, energy dissipation, passive and semi-active control of structures in P. R. China," in *Proceedings of the China-U.S. Millennium Symposium on Earthquake Engineering: Earthquake Engineering Frontiers in the New Millennium*, Beijing, China, November 2000.

[3] F. L. Zhou, P. Tan, Q. L. Xian, X. Y. Huang, and Z. Yang, "Research and application of seismic isolation system for building structures," *Journal of Architecture and Civil Engineering*, vol. 12, no. 2, pp. 1–8, 2006.

[4] H.-N. Li and X.-X. Wu, "Limitations of height-to-width ratio for base-isolated buildings under earthquake," *Structural Design of Tall and Special Buildings*, vol. 15, no. 3, pp. 277–287, 2006 (Chinese).

[5] H. N. Li and X. X. Wu, "Simplified formulate of torsional seismic actions for base-isolated eccentric structures," *Journal of the Harbin Institute of Technology*, vol. 26, no. 1, pp. 52–54, 2004 (Chinese).

[6] J. P. Ou, "Some recent advances of structural vibration control in mainland China," in *Proceedings of the 3rd International Conference on Earthquake Engineering*, Nanjing, China, October 2004.

[7] Y. Zhou and J. Liu, "Comparative experimental study on the behavior of two different energy dissipation brace based on friction," *Earthquake Engineering and Engineering Vibration*, vol. 17, no. 1, p. 48, 1997 (Chinese).

[8] W. Y. Zhang and W. B. Yang, "Experimental investigation on the braced frame with double phase friction attenuation controller under low frequency cyclic loading," *Building Science*, vol. 13, no. 4, pp. 3–7, 1997 (Chinese).

[9] B. Wu and J. P. Ou, "Behavior of pseudo-viscous friction energy dissipators," *World Information on Earthquake Engineering*, vol. 15, no. 11, pp. 1–11, 1999 (Chinese).

[10] B. Wu, J.-G. Zhang, and J.-P. Ou, "Numerical analysis of hysteretic behavior and brace force of pseudo-viscous frictional dampers," *Journal of Harbin Institute of Technology*, vol. 35, no. 7, pp. 834–843, 2003 (Chinese).

[11] Q. L. Xian, F. L. Zhou, C. M. Zhang, and G. H. Yu, "Design and test and analysis of the mix friction energy dissipation bracings used for high-rise buildings," *World Information on Earthquake Engineering*, vol. 12, no. 1, pp. 1–9, 1996 (Chinese).

[12] B. Wu, H. Li, and L. Y. Lin, "Retrofitting of Shenyang Government Office Building by friction dampers," *Journal of Architecture and Buildings*, vol. 19, no. 5, pp. 28–36, 1998 (Chinese).

[13] Y. Zhou, F. L. Zhou, and X.-S. Deng, "Research on lead damper and its applications," *World Information on Earthquake Engineering*, vol. 15, no. 1, pp. 53–61, 1999 (Chinese).

[14] Z. Xu, Z. Zhou, H. Zhao, and Y. Shen, "A new method on viscoelastic dampers," *Engineering Mechanics*, vol. 18, no. 6, pp. 88–92, 2001 (Chinese).

[15] Q. Zhang, W. Lou, and Y. Chen, "Parameter design and location optimum practical method of viscoelastic dampers," *Structural Engineers*, vol. 19, no. 3, pp. 39–44, 2003 (Chinese).

[16] W. Cheng, J. Sui, Y. Chen, et al., "Design and earthquake absorption using viscoelastic damper in transportation building of Suqian city," *Journal of Building Structures*, vol. 21, no. 3, pp. 30–35, 2000 (Chinese).

[17] Y. Yin, Y. Zhou, and L. B. Mei, "Transient dynamic finite element analysis about energy dissipation structure of Chaoshan-xinghe building," *Earthquake Resistant Engineering and Retrofitting*, vol. 27, no. 3, pp. 35–40, 2005 (Chinese).

[18] G. P. Zhao, Y. Lu, P. B. Li, H. J. Han, Z. C. Lu, and J. Huang, "Using viscous dampers to control wind-induced vibration for steel tower of Beijing Yintai Center," *Building Structures*, vol. 37, no. 11, pp. 8–10, 2007 (Chinese).

[19] B. Wu and J. P. Ou, "Fatigue properties and design criteria of mild steel yielding energy dissipaters," *World Information on Earthquake Engineering*, vol. 12, no. 4, pp. 8–13, 1996 (Chinese).

[20] G. Li and H. Li, "Study on vibration reduction of structure with a new type of mild metallic dampers," *Journal of Vibration and Shock*, vol. 25, no. 3, pp. 66–72, 2006 (Chinese).

[21] H.-N. Li and G. Li, "Experimental study of structure with "dual function" metallic dampers," *Engineering Structures*, vol. 29, no. 8, pp. 1917–1928, 2007.

[22] H.-N. Li, Y.-W. Yin, and S.-Y. Wang, "Studies on seismic reduction of story-increased buildings with friction layer and energy-dissipated devices," *Earthquake Engineering and Structural Dynamics*, vol. 32, no. 14, pp. 2143–2160, 2003.

[23] H. N. Li and S. Y. Wang, "Seismic vibration control of high-rise buildings using TMD," *Journal of Engineering Mechanics*, vol. 14, no. A03, pp. 136–140, 1997 (Chinese).

[24] Z. Q. Ye, A. Q. Li, and Y. L. Ding, "The vibration energy dissipation design of a pedestrian bridge (1)," *Special Structures*, vol. 20, no. 1, pp. 68–70, 2003 (Chinese).

[25] Y. L. Ding, A. Q. Li, and Z. Q. Ye, "The vibration energy dissipation design of a pedestrian bridge (2)," *Special Structures*, vol. 20, no. 1, pp. 71–73, 2003 (Chinese).

[26] L. S. Huo and H. N. Li, "Theoretical analysis of vibration control of large-span pedestrian bridge by use of MTMD," *Journal of Disaster Prevention and Mitigation Engineering*, vol. 28, no. 3, pp. 298–302, 2008 (Chinese).

[27] M. Z. Zhang, S. W. Ding, and X. Guo, "Study on structure control using tuned sloshing damper," *Earthquake Engineering & Engineering Vibration*, vol. 13, no. 1, pp. 40–48, 1993 (Chinese).

[28] H. Li, L. Zhang, and Y. S. Yang, "A study on vibration reduction in high-rise structures by multi-TLDs," *Earthquake Engineering and Engineering Vibration*, vol. 17, no. 1, p. 23, 1997 (Chinese).

[29] S. Yan and H. N. Li, "Vibration control parameters research on system with adjustable frequency tuned liquid column damper," *Earthquake Engineering and Engineering Vibration*, vol. 18, no. 4, pp. 96–101, 1998 (Chinese).

[30] L.-S. Huo and H.-N. Li, "Torsion-coupled response control of structures using circular tuned liquid column dampers," *Journal of Engineering Mechanics*, vol. 22, no. 2, pp. 124–131, 2005 (Chinese).

[31] H.-N. Li and L.-S. Huo, "Optimal design of liquid dampers for structural torsion coupled vibration," *Chinese Journal of Computational Mechanics*, vol. 22, no. 2, pp. 129–134, 2005 (Chinese).

[32] B. F. Spencer and T. T. Song, "New applications and development of active, semi-active and hybrid control techniques for seismic and non-seismic vibration in the USA," in *Proceedings of the International Post-SMiRT Seminar on Seismic Isolation, Passive Energy Dissipation and Active Control of Vibration of Structures*, Cheju, Korea, 1999.

[33] H.-N. Li and Q. Jin, "AMD active control for irregular buildings using GA-BP neural network," *Earthquake Engineering and Engineering Vibration*, vol. 23, no. 2, pp. 134–142, 2003 (Chinese).

[34] Q. Jin and J. Zhou, "Discrete-time variable structure control method for AMD-eccentric buildings with time delay in control," *Journal of Harbin Institute of Technology*, vol. 37, no. 6, pp. 790–794, 2005 (Chinese).

[35] A. Qi, J. Liu, and H. Li, "Hybrid control research on earthquake responses of high-rise building with spacious first story," *Journal of Vibration Engineering*, vol. 12, no. 2, pp. 256–261, 1999 (Chinese).

[36] J. Teng, Z.-X. Lu, Y.-Q. Xiao, A.-Z. Yan, and C.-T. Li, "Vibration control of high-rising buildings with TMD considering contact non-linear damping effect," *Journal of Vibration and Shock*, vol. 28, no. 3, pp. 90–97, 2009 (Chinese).

[37] M. X. Li and J. Liu, "Semi-active control for structural vibration using variable stiffness with nonlinear damping," *Journal of Vibration Engineering*, vol. 11, no. 3, pp. 333–339, 1998 (Chinese).

[38] Z. Y. Sun, *Experimental study on variable orifice dampers*, Ph.D. dissertation, Harbin Institute of Technology, Harbin, China, 1998.

[39] H. Li, X. S. Yuan, and B. Wu, "Variable orifice dampers: experiments and analysis," *Journal of Vibration Engineering*, vol. 15, no. 1, pp. 15–30, 2002.

[40] H.-N. Li, L.-S. Huo, and S. Yan, "TLCD semi-active control for irregular buildings using neural networks," *Earthquake Engineering and Engineering Vibration*, vol. 21, no. 4, pp. 135–141, 2001 (Chinese).

[41] H. Li and Q. Jin, "TLCD semi-active control for irregular buildings using fuzzy neural network based on Takagi-Sugeno model," *Chinese Journal of Computational Mechanics*, vol. 20, no. 5, pp. 523–529, 2003 (Chinese).

[42] X.-L. Li and H.-N. Li, "Double-sigmoid model for magnetorheological damper and corresponding experiment verification," *Journal of Vibration Engineering*, vol. 19, no. 2, pp. 168–172, 2006 (Chinese).

[43] X.-L. Li and H.-N. Li, "Experimental study of semi-active control of coupled translation and torsion response for frame-shear wall eccentric structure," *Journal of Dalian University of Technology*, vol. 48, no. 5, pp. 691–697, 2008 (Chinese).

[44] J. M. Ko, Y. Q. Ni, Z. Q. Chen, and B. F. Spencer, "Implementation of MR dampers to Dongting Lake Bridge for cable vibration mitigation," in *Proceeding of the 3rd World Conference on Structural Control*, Como, Italy, 2002.

[45] J. P. Ou, *Structural Vibration Control-Active, Semi-Active and Smart Control Systems*, Press of Science, Beijing, China, 2003.

[46] J. Li, H. N. Li, and G. Song, "Semi-active vibration suppression using piezoelectric friction dampers based on sub-optimal bang-bang control laws," in *Proceedings of the 3rd China-Japan-US Symposium on Structural Health Monitoring and Control*, Dalian, China, 2004.

[47] H. Li, C. X. Mao, and J. P. Ou, "Shape memory alloy-based smart building," in *Proceedings of the 8th US National Conference on Earthquake Engineering*, San Francisco, Calif, USA, 2006.

[48] W. Ren, H. Li, and G. Song, "Mechanical model of X type SMA plate damper," *Journal of Vibration and Shock*, vol. 25, no. 4, pp. 53–57, 2006 (Chinese).

[49] W. Ren, H. Li, and G. Song, "A one-dimensional strain-rate-dependent constitutive model for superelastic shape memory alloys," *Smart Materials and Structures*, vol. 16, no. 1, pp. 191–197, 2007.

[50] H.-N. Li, H. Qian, G.-B. Song, and D.-W. Gao, "Type of shape memory alloy damper: design, experiment and numerical simulation," *Journal of Vibration Engineering*, vol. 21, no. 2, pp. 179–184, 2008.

[51] W. Ren, H. Li, and G. Song, "Design and numerical evaluation of an innovative multi-directional shape memory alloy damper," in *Active and Passive Smart Structures and Integrated Systems 2007*, vol. 6525 of *Proceedings of SPIE*, San Diego, Calif, USA, March 2007.

Review Article

Time Domain Simulation of a Target Tracking System with Backlash Compensation

Maurício Gruzman,[1] Hans Ingo Weber,[1] and Luciano Luporini Menegaldo[2]

[1] *Departamento de Engenharia Mecânica, Pontifícia Universidade Católica do Rio de Janeiro,*
R. Marquês de São Vicente 225, 22453-900 Rio de Janeiro, RJ, Brazil
[2] *Departamento de Engenharia Mecânica, Instituto Militar de Engenharia, Praça General Tibúrcio 80,*
Praia Vermelha, 22290-270 Rio de Janeiro, RJ, Brazil

Correspondence should be addressed to Maurício Gruzman, maurgruzman@hotmail.com

Received 15 July 2009; Accepted 23 September 2009

Academic Editor: José Balthazar

This paper presents a model of a target tracking system assembled in a moving body. The system is modeled in time domain as a nonlinear system, which includes dry friction, backlash in gear transmission, control input tensions saturation, and armature current saturation. Time delays usually present in digital controllers are also included, and independent control channels are used for each motor. Their inputs are the targets angular errors with respect to the system axial axis and the outputs are control tensions for the motors. Since backlash in gear transmission may reduce the systems accuracy, its effects should be compensated. For that, backlash compensation blocks are added in the controllers. Each section of this paper contains a literature survey of recent works dealing with the issues discussed in this article.

1. Introduction

In many military and civilian applications we can find equipments assembled into moving bodies, which must follow a target. They need to be isolated from the rotational motion of the body where they are mounted, called body 0 in this work (see Figure 1). This can be achieved with a system composed by a Cardan suspension (introducing therefore bodies 1 and 2), two motors connected to each axis, and two rate gyros mounted in body 2 to measure the angular absolute velocities in two directions perpendicular to the axial axis, defined in body 2. In literature, this Cardan suspension assembly is often called double gimbaled system. With the information provided by the rate gyros, a controller commands the motors in order to compensate any perturbation that is transmitted to body 2. At the same time, the equipment
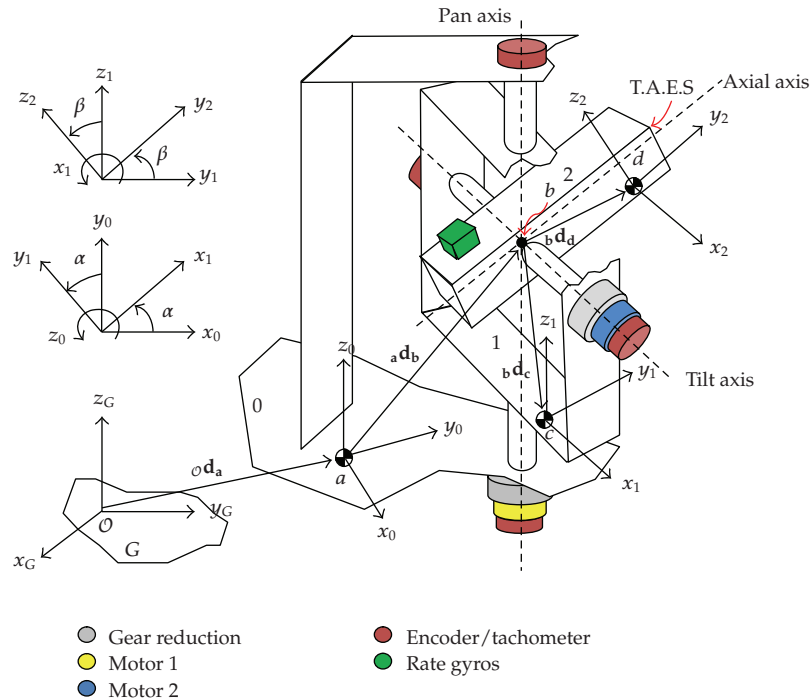
Figure 1: Target tracking system assembled in a moving body (0).

must also follow a target. Therefore some sensor (as radars, infrared-sensor, or vision sensors) can be mounted in body 2 to provide information about targets angular errors with respect to the axial axis. In this work this sensor will be called TAES (target angular errors sensor). To point precisely to the target the controller must command the motors in order to keep the angular errors equal to zero. Apparently, a contradictory task is sought at the same time: the controller needs to stabilize the equipment with respect to inertial frame; however it must also move the equipment to follow a target. This contradiction is only apparent, since both actions can be combined efficiently in a two-loop controller. With information received from TAES, the outer loop calculates the desired absolute angular velocities for body 2, in order to point to the target. The inner loop pursues to keep body 2 absolute angular velocities in the desired values, by comparing the information from the outer loop and the rate gyros. This approach has been discussed in [1–4], although backlash in gear transmission has not been considered in the numerical models.

In the present work, the controller described in Gruzman et al. [3] will be employed, where the outer loop consists of a Fuzzy Logic Controller [5] and the inner loop is a PI (proportional and integral) controller with antiwindup [6]. Since the mentioned control approach is developed for systems without backlash in gear transmission, an extra controller will be added to compensate problems caused by backlash. This controller provides compensation values for the control tensions whenever the suspension axes pass through a backlash gap in gear reduction.

In this work, actuators are permanent magnet DC motors controlled by armature voltage. Sensor errors and noise are not considered, and the motion of body 0 is prescribed. Control signals (voltages) saturation and armature current limiters are included in the model.

Many authors, as Masten [2], Downs et al. [7], and Lee and Yoo [8], model gimbaled target tracking systems in frequency domain. Such approach is convenient for control design purposes but may often result in excessive simplification of the dynamic equations that are usually highly nonlinear. In this work the differential equations are not linearized and are integrated with respect to time. Besides, adequate dry friction torque models, gear backlash models, armature current saturation, and controllers with inner and outer loops, having different time delays, can be included with no further difficulties. The main contributions of this work are, therefore, the presentation of the following:

(i) a sophisticated time domain target tracking system model,

(ii) a backlash compensation block that is added to the controllers previously developed for systems without backlash in Gruzman et al. [3] in order to reduce delays, oscillations, and inaccuracy caused by backlash.

2. Equations of Motion of the Device

The equations of motion of the device can be obtained by Lagrange formulation [9]. It is not the aim of this work to study the dynamics of body 0; therefore it is assumed that its motion is prescribed. The position vector of its center of mass (point a in Figure 1) with respect to the inertial frame origin will be called $_O\mathbf{d_a}$. The position vectors of bodies 1 and 2 centers of mass, with respect to the center of the Cardan suspension (point b), are, respectively, $_b\mathbf{d_c}$ and $_b\mathbf{d_d}$. The orientation coordinates of body 0 can be given by the angles of successive rotations called pitch (δ), yaw (ψ), and roll (γ), presented in Figure 2.

It can be seen in Figure 1 that center-of-mass fixed reference frames are used. They are chosen in such a way that z_0 and z_1 remain parallel to the pan axis, x_1 and x_2 to the tilt axis, and y_2 to the axial axis. The device has four generalized coordinates:

Writinng the Lagrange equations for the device, one has

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\alpha}}\right) - \frac{\partial L}{\partial \alpha} = \xi_\alpha, \qquad \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\alpha}_m}\right) - \frac{\partial L}{\partial \alpha_m} = \xi_{\alpha_m},$$

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\beta}}\right) - \frac{\partial L}{\partial \beta} = \xi_\beta, \qquad \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\beta}_m}\right) - \frac{\partial L}{\partial \beta_m} = \xi_{\beta_m}.$$

(2.1)

(i) α: rotation angle of body 1 (outer gimbal) relative to body 0 measured around the pan axis;

(ii) α_m: rotation angle of motor 1 rotor relative to body 0 measured around the pan axis;

(iii) β: rotation angle of body 2 (comprising the inner gimbal and the TAES) relative to body 1 measured around the tilt axis;

(iv) β_m: rotation angle of motor 2 rotor relative to body 1 measured around the tilt axis.

The system's Lagrangian is given by the sum of bodies 1, 2 and rotor 1 and rotor 2 Lagrangians, each containing the potential and kinetic energy of the bodies, which are assumed to be rigid in this work. The right-hand side terms of (2.1) correspond to nonconservative generalized torques. They include the viscous and dry friction torques
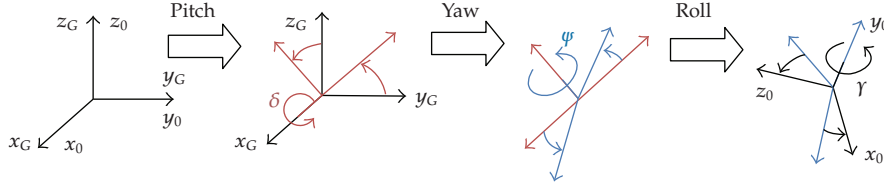
Figure 2: Orientation coordinates of body 0 with respect to inertial frame (G).

(T_{vf} and T_{df}), the electromotive torques due to the motors (T_m), and the transmission torques (T_s) between rotor 1 and body 1 and between rotor 2 and body 2:

$$
\begin{aligned}
\xi_\alpha &= T_{vf,\alpha} + T_{df,\alpha} + T_{s,\alpha}, \\
\xi_{\alpha_m} &= T_{vf,\alpha_m} + T_{df,\alpha_m} + T_{s,\alpha_m} + T_{m,\alpha_m}, \\
\xi_\beta &= T_{vf,\beta} + T_{df,\beta} + T_{s,\beta}, \\
\xi_{\beta_m} &= T_{vf,\beta_m} + T_{df,\beta_m} + T_{s,\beta_m} + T_{m,\beta_m}.
\end{aligned}
\tag{2.2}
$$

2.1. Torques due to Viscous Friction

The viscous friction torques at body 1, body 2, rotor 1, and rotor 2 axes are, respectively, given by

$$
\begin{aligned}
T_{vf,\alpha} &= -c_\alpha \cdot \dot\alpha, \\
T_{vf,\beta} &= -c_\beta \cdot \dot\beta, \\
T_{vf,\alpha_m} &= -c_{\alpha_m} \cdot \dot\alpha_m, \\
T_{vf,\beta_m} &= -c_{\beta_m} \cdot \dot\beta_m,
\end{aligned}
\tag{2.3}
$$

where c_α, c_β, c_{α_m}, and c_{β_m} are the resultant viscous friction coefficients at body 1, body 2, rotor 1, and rotor 2 axes, respectively.

2.2. Torques due to the Motors

The electromotive torque in a rotor of a permanent magnet DC motor is given by (2.4), written with rotors 1 generalized coordinate subscript:

$$
T_{m,\alpha_m} = k_{m,\alpha_m} \cdot i_{\alpha_m},
\tag{2.4}
$$

where k_{m,α_m} is the motor torque constant and i_{α_m} the armature current, whose value is obtained by [10]

$$
\frac{di_{\alpha_m}}{dt} = \frac{u_{\alpha_m} - R_{\alpha_m} \cdot i_{\alpha_m} - k_{b,\alpha_m} \cdot \dot\alpha_m}{l_{\alpha_m}},
\tag{2.5}
$$

where R, l, and k_b are, respectively, the armature resistance, armature inductance, and back-emf constant. The tension provided to the motor (u) is the control variable and its calculations will be discussed in Section 4.

By a similar way, (2.4) and (2.5) can be written for rotor 2 as

$$T_{m,\beta_m} = k_{m,\beta_m} \cdot i_{\beta_m}, \tag{2.6}$$

$$\frac{di_{\beta_m}}{dt} = \frac{u_{\beta_m} - R_{\beta_m} \cdot i_{\beta_m} - k_{b,\beta_m} \cdot \dot{\beta}_m}{l_{\beta_m}}. \tag{2.7}$$

Armature current limiters are usually employed to avoid damages to the motors and circuits. They keep the current between a maximum value ($i_{\max}$) and a minimum value ($i_{\min}$). If the system has limiters, then the first derivative of the currents with respect to time, given by (2.5) and (2.7), is set to zero whenever

$$
\begin{aligned}
&(i_{\alpha_m} \geq i_{\max,\alpha_m}), && \left(\frac{u_{\alpha_m} - k_{b,\alpha_m} \cdot \dot{\alpha}_m}{R_{\alpha_m}} \geq i_{\max,\alpha_m} \right), \\
&\text{or,} \quad (i_{\alpha_m} \leq i_{\min,\alpha_m}), && \left(\frac{u_{\alpha_m} - k_{b,\alpha_m} \cdot \dot{\alpha}_m}{R_{\alpha_m}} \leq i_{\min,\alpha_m} \right), \\
&(i_{\beta_m} \geq i_{\max,\beta_m}), && \left(\frac{u_{\beta_m} - k_{b,\beta_m} \cdot \dot{\beta}_m}{R_{\beta_m}} \geq i_{\max,\beta_m} \right), \\
&\text{or,} \quad (i_{\beta_m} \leq i_{\min,\beta_m}), && \left(\frac{u_{\beta_m} - k_{b,\beta_m} \cdot \dot{\beta}_m}{R_{\beta_m}} \leq i_{\min,\beta_m} \right).
\end{aligned}
\tag{2.8}
$$

2.3. Torques between Rotor 1 and Body 1 and between Rotor 2 and Body 2

The torque between the rotor and the body driven by that rotor (body 1 or body 2) is affected by backlash present at the transmission. When the body is transversing the backlash gap, this torque will be equal to zero. The approach adopted in Nordin and Gutman [11] and Lagerberg and Egardt [12] will be used to model flexibility and backlash in each set comprising the rotor, gear reduction, and the body. Only two elements with nonzero inertia are considered in each set. The first has the inertia of the rotor and the second, that will be called load, has the inertia of the driven body. The inertias of the gears are usually disregarded but can be included in rotors inertia. For that, it is necessary to calculate the equivalent inertia of each gear in the rotor. There is also a shaft without inertia between the gear reduction and the load. This shaft has coefficients of structural rigidity k_s and structural damping c_s. A backlash with amplitude of 2η exists between this shaft and the load. The parameters k_s, c_s, and 2η correspond to resultant values for the entire set. Baek et al. [13] show how to obtain the resultant backlash and coefficient of structural rigidity of a system comprising an actuator, gear reduction, and a load. The following equations will be presented for the set rotor 1-body 1 (therefore the variable α, that corresponds to body 1 generalized coordinate, is used after the " ," in the subscripts).
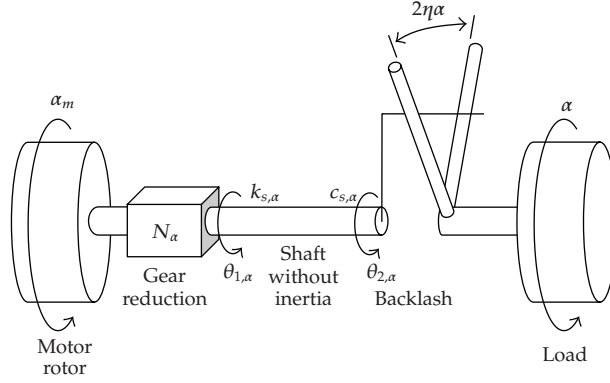
Figure 3: Backlash model (between rotor 1 and body 1).

The gear reduction ratio is given by

$$N_\alpha = \frac{\alpha_m}{\theta_{1,\alpha}}. \tag{2.9}$$

The torque at the shaft without inertia is

$$T_{e,\alpha} = k_{s,\alpha}(\theta_{d,\alpha} - \theta_{b,\alpha}) + c_{s,\alpha}(\dot\theta_{d,\alpha} - \dot\theta_{b,\alpha}), \tag{2.10}$$

where $\theta_{b,\alpha} = \theta_{2,\alpha} - \alpha$ and $\theta_{d,\alpha} = \theta_{1,\alpha} - \alpha = \alpha_m/N_\alpha - \alpha$.

The backlash angle is θ_b and its first derivative is given by the following ordinary differential equation (ODE):

$$\dot\theta_{b,\alpha} = \begin{cases} \max\left(0, \dot\theta_{d,\alpha} + \dfrac{k_{s,\alpha}}{c_{s,\alpha}}(\theta_{d,\alpha} - \theta_{b,\alpha})\right) & \text{when } \theta_{b,\alpha} = -\eta_\alpha, \\[3mm] \dot\theta_{d,\alpha} + \dfrac{k_{s,\alpha}}{c_{s,\alpha}}(\theta_{d,\alpha} - \theta_{b,\alpha}) & \text{when } |\theta_{b,\alpha}| < \eta_\alpha, \\[3mm] \min\left(0, \dot\theta_{d,\alpha} + \dfrac{k_{s,\alpha}}{c_{s,\alpha}}(\theta_{d,\alpha} - \theta_{b,\alpha})\right) & \text{when } \theta_{b,\alpha} = \eta_\alpha. \end{cases} \tag{2.11}$$

A new state variable is introduced in this model (θ_b). In order to obtain θ_b, (2.11) must be integrated. The torques transmitted to the rotor 1 and the load (body 1) are, respectively, given by

$$T_{s,\alpha_m} = -\frac{T_{e,\alpha}}{N_\alpha}, \\ T_{s,\alpha} = T_{e,\alpha}. \tag{2.12}$$

For the set rotor 2-body 2, the equations are

$$T_{e,\beta} = k_{s,\alpha}(\theta_{d,\beta} - \theta_{b,\beta}) + c_{s,\beta}(\dot{\theta}_{d,\beta} - \dot{\theta}_{b,\beta}),$$ (2.13)

where $\theta_{b,\beta} = \theta_{2,\beta} - \alpha$ and $\theta_{d,\beta} = \theta_{1,\beta} - \beta = \beta_m/N_\beta - \beta$.

The backlash angle is θ_b and its first derivative is given by the following ordinary differential equation (ODE):

$$\dot{\theta}_{b,\beta} = \begin{cases} \max\left(0, \dot{\theta}_{d,\beta} + \dfrac{k_{s,\beta}}{c_{s,\beta}}(\theta_{d,\beta} - \theta_{b,\beta})\right) & \text{when } \theta_{b,\beta} = -\eta_\beta, \\[2ex] \dot{\theta}_{d,\beta} + \dfrac{k_{s,\beta}}{c_{s,\beta}}(\theta_{d,\beta} - \theta_{b,\beta}) & \text{when } |\theta_{b,\beta}| < \eta_\beta, \\[2ex] \min\left(0, \dot{\theta}_{d,\beta} + \dfrac{k_{s,\beta}}{c_{s,\beta}}(\theta_{d,\beta} - \theta_{b,\beta})\right) & \text{when } \theta_{b,\beta} = \eta_\beta, \end{cases}$$ (2.14)

$$T_{s,\beta_m} = -\frac{T_{e,\beta}}{N_\beta},$$ (2.15)

$$T_{s,\beta} = T_{e,\beta}.$$ (2.16)

2.4. Torques due to Dry Friction

Dry friction torques may be found in the system; therefore they should be included in the model. There are many numerical models for dry friction available in literature [14–20], but they depend of empiric parameters that are not easy to be estimated. Therefore, the procedure presented by Piedbœuf et al. [21] will be used to calculate the resultant dry friction torques at body 1, body 2, rotor 1, and rotor 2 axes. With this approach the only parameters that need to be known are the dynamic dry friction torque (T_{din}) and the maximum value of the static dry friction torque (T_{limit}) at each axis, which can be obtained with simple experiments, as shown in [21]. Before each integration step, it is verified if stiction due to dry friction is happening in any generalized coordinate, as will be explained next.

The dynamic equations of the system (obtained from the Lagrange equations) in a matrix form is given by (2.17), where the components of inertia matrix ($J_{1,1}, J_{1,2}, \ldots$) and f_1, f_2, f_3, and f_4 are nonlinear functions of the generalized coordinates:

$$\begin{bmatrix} J_{1,1} & J_{1,2} & J_{1,3} & J_{1,4} \\ J_{2,1} & J_{2,2} & J_{2,3} & J_{2,4} \\ J_{3,1} & J_{3,2} & J_{3,3} & J_{3,4} \\ J_{4,1} & J_{4,2} & J_{4,3} & J_{4,4} \end{bmatrix} \begin{bmatrix} \ddot{\alpha} \\ \ddot{\alpha}_m \\ \ddot{\beta} \\ \ddot{\beta}_m \end{bmatrix} = \begin{bmatrix} T_{vf,\alpha} \\ T_{vf,\alpha_m} \\ T_{vf,\beta} \\ T_{vf,\beta_m} \end{bmatrix} + \begin{bmatrix} T_{df,\alpha} \\ T_{df,\alpha_m} \\ T_{df,\beta} \\ T_{df,\beta_m} \end{bmatrix} + \begin{bmatrix} T_{s,\alpha} \\ T_{s,\alpha_m} \\ T_{s,\beta} \\ T_{s,\beta_m} \end{bmatrix} + \begin{bmatrix} 0 \\ T_{m,\alpha_m} \\ 0 \\ T_{m,\beta_m} \end{bmatrix} + \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix}.$$ (2.17)

Calling a vector $\mathbf{F} = \mathbf{T_{vf}} + \mathbf{T_s} + \mathbf{T_m} + \mathbf{f}$, where $\mathbf{F} = [F_1 \quad F_2 \quad F_3 \quad F_4]^T$, (2.17) can be rewritten as

$$
\begin{bmatrix} J_{1,1} & J_{1,2} & J_{1,3} & J_{1,4} \\ J_{2,1} & J_{2,2} & J_{2,3} & J_{2,4} \\ J_{3,1} & J_{3,2} & J_{3,3} & J_{3,4} \\ J_{4,1} & J_{4,2} & J_{4,3} & J_{4,4} \end{bmatrix} \begin{bmatrix} \ddot{\alpha} \\ \ddot{\alpha}_m \\ \ddot{\beta} \\ \ddot{\beta}_m \end{bmatrix} = \begin{bmatrix} T_{df,\alpha} \\ T_{df,\alpha_m} \\ T_{df,\beta} \\ T_{df,\beta_m} \end{bmatrix} + \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{bmatrix}. \tag{2.18}
$$

In the following, we will concentrate our analysis in the generalized coordinate α. The extension to the other generalized coordinates is straightforward.

If $\dot{\alpha}$ is not zero, it means that body 1 is in slip regime, and the generalized dry friction torque at this body will be

$$
T_{df,\alpha} = -T_{\mathrm{din},\alpha} \cdot \mathrm{sign}(\dot{\alpha}). \tag{2.19}
$$

If it is equal to zero and the modulus of the sum of all torques ($T_{\mathrm{TEST},\alpha}$), including torques due to inertia but excluding dry friction torque, is lower than $T_{\mathrm{limit},\alpha}$ it means that body 1 is at the stick regime. Thus, the dry friction torque should be calculated according to (2.20):

$$
T_{df,\alpha} = -T_{\mathrm{TEST},\alpha} \quad \text{when} - T_{\mathrm{limit},\alpha} < T_{\mathrm{TEST},\alpha} < T_{\mathrm{limit},\alpha}. \tag{2.20}
$$

It should be stressed that it is not convenient to consider $\dot{\alpha} = 0$ as a transition condition between stick and slip, since during the numerical resolution of the equations of motion the true zero may not be achieved. Besides, it can be observed in experiments that sliding abruptly stops under a minimum angular velocity (v_{min}). Therefore it will be considered that the body is in stick regime whenever the modulus of $\dot{\alpha}$ is lower than $v_{\mathrm{min},\alpha}$ and the modulus of $T_{\mathrm{TEST},\alpha}$ is lower than $T_{\mathrm{limit},\alpha}$. Next, a numerical damping term must be included to the static dry friction torque to eliminate the residual angular relative velocity of the body, as it will be seen in the following example.

Consider that at a certain instant during the numerical resolution of the equations system (2.18), the modulus of $\dot{\alpha}$ is smaller than $v_{\mathrm{lim},\alpha}$. It is assumed that the first derivatives of the other generalized coordinates are larger than the respective minimum angular velocities of transition between slip and stick. Then, it must be checked if the modulus of $T_{\mathrm{TEST},\alpha}$, obtained from the system of equations below, is higher or lower than $T_{\mathrm{limit},\alpha}$:

$$
\begin{bmatrix} 1 & J_{1,2} & J_{1,3} & J_{1,4} \\ 0 & J_{2,2} & J_{2,3} & J_{2,4} \\ 0 & J_{3,2} & J_{3,3} & J_{3,4} \\ 0 & J_{4,2} & J_{4,3} & J_{4,4} \end{bmatrix} \begin{bmatrix} T_{\mathrm{TEST},\alpha} \\ \ddot{\alpha}_m \\ \ddot{\beta} \\ \ddot{\beta}_m \end{bmatrix} = \begin{bmatrix} 0 \\ -T_{\mathrm{din},\alpha_m} \cdot \mathrm{sign}(\dot{\alpha}_m) \\ -T_{\mathrm{din},\beta} \cdot \mathrm{sign}(\dot{\beta}) \\ -T_{\mathrm{din},\beta_m} \cdot \mathrm{sign}(\dot{\beta}_m) \end{bmatrix} + \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{bmatrix}. \tag{2.21}
$$

If $|T_{\text{TEST},\alpha}| > T_{\text{limit},\alpha}$ then body 1 is in slip regime and $\ddot{\alpha}$, $\ddot{\alpha}_m$, $\ddot{\beta}$, and $\ddot{\beta}_m$ are obtained with (2.22). Otherwise (2.23) should be used:

$$
\begin{bmatrix} \ddot{\alpha} \\ \ddot{\alpha}_m \\ \ddot{\beta} \\ \ddot{\beta}_m \end{bmatrix} = \begin{bmatrix} J_{1,1} & J_{1,2} & J_{1,3} & J_{1,4} \\ J_{2,1} & J_{2,2} & J_{2,3} & J_{2,4} \\ J_{3,1} & J_{3,2} & J_{3,3} & J_{3,4} \\ J_{4,1} & J_{4,2} & J_{4,3} & J_{4,4} \end{bmatrix}^{-1} \left(\begin{bmatrix} -T_{\text{din},\alpha} \cdot \text{sign}(\dot{\alpha}) \\ -T_{\text{din},\alpha_m} \cdot \text{sign}(\dot{\alpha}_m) \\ -T_{\text{din},\beta} \cdot \text{sign}(\dot{\beta}) \\ -T_{\text{din},\beta_m} \cdot \text{sign}(\dot{\beta}_m) \end{bmatrix} + \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{bmatrix} \right), \tag{2.22}
$$

$$
\begin{bmatrix} \ddot{\alpha} \\ \ddot{\alpha}_m \\ \ddot{\beta} \\ \ddot{\beta}_m \end{bmatrix} = \begin{bmatrix} J_{1,1} & J_{1,2} & J_{1,3} & J_{1,4} \\ J_{2,1} & J_{2,2} & J_{2,3} & J_{2,4} \\ J_{3,1} & J_{3,2} & J_{3,3} & J_{3,4} \\ J_{4,1} & J_{4,2} & J_{4,3} & J_{4,4} \end{bmatrix}^{-1} \left(\begin{bmatrix} -T_{\text{TEST},\alpha} - b_{\text{numeric},\alpha} \cdot \dot{\alpha} \\ -T_{\text{din},\alpha_m} \cdot \text{sign}(\dot{\alpha}_m) \\ -T_{\text{din},\beta} \cdot \text{sign}(\dot{\beta}) \\ -T_{\text{din},\beta_m} \cdot \text{sign}(\dot{\beta}_m) \end{bmatrix} + \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{bmatrix} \right). \tag{2.23}
$$

As previously discussed, a numerical damping term should be added to dry friction static torque because body 1 is in stick regime. An expression proposed in [21] for the numerical damping coefficient is

$$
b_{\text{numeric},\alpha} = \frac{\mu_\alpha}{v_{\text{lim},\alpha}} \cdot T_{\text{limit},\alpha}, \tag{2.24}
$$

where μ_α is a constant value chosen such that $J_{1,1}/b_{\text{numeric},\alpha}$ is reasonably larger than the minimum integration step size of the ordinary differential equations (ODEs) integrator. In the example presented, stiction occurred only in α, but the method can also be used for stiction in the other generalized coordinates, simultaneously or not. When (2.21) is used for the verification of stiction in a generalized coordinate i (in a system with n generalized coordinates), the following rule should be applied:

$$
\text{if} \begin{cases} j \neq i \implies J_{j,i} = 0, \ T_{df,i} = 0, \\ \\ j = i \implies J_{j,i} = 1, \end{cases} \quad \text{for } j = 1, \dots, n. \tag{2.25}
$$

3. Resolution of the Equations Considering Control Signal Time Delay

The resultant system has four first-order ODEs, given by (2.5), (2.7), (2.11), and (2.14), and four second-order ODEs, given by (2.17). To obtain only first-order equations the following coordinate transformation can be applied: $q_1 = \alpha$, $q_2 = \alpha_m$, $q_3 = \beta$, $q_4 = \beta_m$, $q_5 = \dot{\alpha}$, $q_6 = \dot{\alpha}_m$,

$q_7 = \dot{\beta}$, and $q_8 = \dot{\beta}_m$ and (2.17) becomes

$$\dot{q}_1 = q_5,$$

$$\dot{q}_2 = q_6,$$

$$\dot{q}_3 = q_7,$$

$$\dot{q}_4 = q_8,$$

$$\begin{bmatrix} \dot{q}_5 \\ \dot{q}_6 \\ \dot{q}_7 \\ \dot{q}_8 \end{bmatrix} = \begin{bmatrix} J_{1,1} & J_{1,2} & J_{1,3} & J_{1,4} \\ J_{2,1} & J_{2,2} & J_{2,3} & J_{2,4} \\ J_{3,1} & J_{3,2} & J_{3,3} & J_{3,4} \\ J_{4,1} & J_{4,2} & J_{4,3} & J_{4,4} \end{bmatrix}^{-1} \left(\begin{bmatrix} T_{vf,\alpha} \\ T_{vf,\alpha_m} \\ T_{vf,\beta} \\ T_{vf,\beta_m} \end{bmatrix} + \begin{bmatrix} T_{df,\alpha} \\ T_{df,\alpha_m} \\ T_{df,\beta} \\ T_{df,\beta_m} \end{bmatrix} + \begin{bmatrix} T_{s,\alpha} \\ T_{s,\alpha_m} \\ T_{s,\beta} \\ T_{s,\beta_m} \end{bmatrix} + \begin{bmatrix} 0 \\ T_{m,\alpha_m} \\ 0 \\ T_{m,\beta_m} \end{bmatrix} + \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} \right).$$

$$(3.1)$$

The components of the vector $\mathbf{T_{vf}}$ are given by (2.3). The components of the vector $\mathbf{T_{df}}$ are calculated, before each integration step, according to the method presented in Section 2.4. The components of the vector $\mathbf{T_m}$ are given by (2.4) and (2.6), which depend on the armature current of motor 1 and motor 2. Those currents are obtained by the first-order ODE system, given by (2.5) and (2.7), which should be solved simultaneously with (3.1). Equations (2.3), (2.5), and (2.7) can be rewritten with the new coordinates introduced in this section:

$$T_{vf,\alpha} = -c_\alpha \cdot q_5, \qquad T_{vf,\beta} = -c_\beta \cdot q_7, \qquad T_{vf,\alpha_m} = -c_{\alpha_m} \cdot q_6, \qquad T_{vf,\beta_m} = -c_{\beta_m} \cdot q_8, \qquad (3.2)$$

$$\frac{di_{\alpha_m}}{dt} = \frac{u_{\alpha_m} - R_{\alpha_m} \cdot i_{\alpha_m} - k_{b,\alpha_m} \cdot q_6}{l_{\alpha_m}}, \qquad \frac{di_{\beta_m}}{dt} = \frac{u_{\beta_m} - R_{\beta_m} \cdot i_{\beta_m} - k_{b,\beta_m} \cdot q_8}{l_{\beta_m}}. \qquad (3.3)$$

The components of the vector $\mathbf{T_s}$ are given by (2.12), (2.15), and (2.16), computed after calculating (2.10) and (2.13). To calculate (2.10) and (2.13) two first-order ODEs, given by (2.11) and (2.14), must be solved simultaneously with (3.1)–(3.3). With the initial conditions

$$q_1(0), q_2(0), q_3(0), q_4(0), q_5(0), q_6(0), q_7(0), q_8(0), i_{\alpha_m}(0), i_{\beta_m}(0), \theta_{b,\alpha}(0), \theta_{b,\beta}(0) \qquad (3.4)$$

and the prescribed motion of the body 0, a numerical algorithm for calculation of the approximate solution of first-order differential equations system can be used [22]. The voltages provided to the motors (u) are the control variables and should not be updated on the right side of (3.3) at each integration step; otherwise the time delay Δt that corresponds to the delay that occurs in a real system due to data acquisition by the sensors, data processing, and calculation of the control tensions would be ignored. Thus, the control voltages should be updated during the numerical integration of the differential equations by the following way:

(i) with data provided by the sensors, at an instant t, the control voltages are calculated;

(ii) the ODEs are integrated from t to $(t + \Delta t)$, using constant voltages (zero order hold) calculated at $(t - \Delta t)$, (remark the delay of Δt);
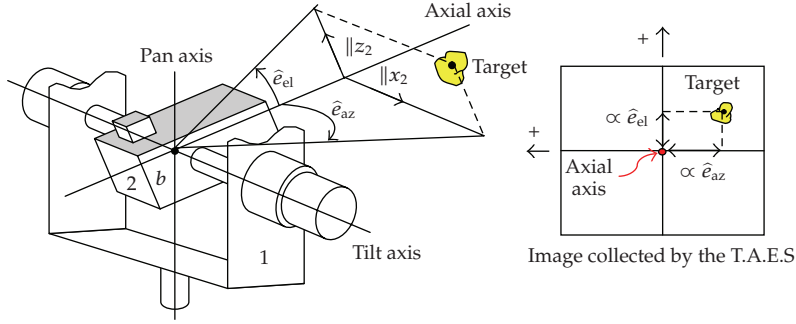
Figure 4: Angular errors of azimuth and elevation.

(iii) at instant $(t + \Delta t)$, with data provided by the sensors, new control voltages are calculated;

(iv) the integration process continues from $(t + \Delta t)$ to $(t + 2\Delta t)$, using constant voltages calculated at t (again a delay of Δt);

(v) then, the process is repeated until the final simulation instant is reached.

4. Control Signals

The controllers need to provide the adequate voltages to the motors to keep the axial axis pointed to the target. This means that the angular errors of azimuth ($\hat{e}_{az}$) and elevation ($\hat{e}_{el}$) with respect to the axial axis should be kept as near as zero as possible.

It will be assumed that the TAES provides the targets centroid coordinates in the image plane, as shown in Figure 4. These coordinates with respect to the image plane center are proportional to $\hat{e}_{az}$ and $\hat{e}_{el}$. The time to calculate the targets centroid coordinates is usually larger than the control signal update period Δt. As the controller must also provide compensation to body 0 angular motion, it is necessary to adopt a control strategy where the voltages provided to the motors can be updated faster than the information from the TAES. A method that is commonly employed and discussed in [1–4] consists in the use of independent controllers, with an external loop, also called tracking loop, and an internal, or stabilization loop, for each motor. One of the controllers has the objective to keep body 2 absolute angular speed in x_2 direction at some desired value, while the other seeks to do the same for body 2 in z_2 direction. With information about the angular errors, the desired values are calculated at the external loops. In the inner loop, working at higher sampling rate, the voltage provided to the motor in order to keep the absolute angular speed in x_2 direction (or z_2) at the desired value is calculated and updated at every time interval Δt. The input for this loop is an error signal equal to the desired absolute angular speed minus the value measured by one of the rate gyros mounted in body 2. In this work the approach presented by Gruzman et al. [3] will be used, where the internal loops have PI (proportional and integral) controllers with antiwindup [6, 10] and the outer loops incremental Fuzzy Logic Controllers (FLCs). The FLC output is an increment for the desired angular absolute speed for body 2 in x_2 direction (or z_2). The inputs for the FLC are the angular errors provided by the TAES, their derivatives with respect to time, and a variable called LoSu. This variable corresponds to the saturation level of the control signal and is used to avoid increments for the desired absolute angular speed

if the control signal is near its superior saturation limit (or decrements if it is near the inferior saturation limit). One controller provides the voltage to the tilt motor (motor 2) to maintain body 2 absolute angular velocity in x_2 direction at the desired value. The other controller provides the voltage to the pan motor (motor 1) to keep the desired angular absolute speed in z_2 direction. It is important to remember that the pan motor axis is parallel to z_1 and not z_2; therefore the error comprising body 2 absolute angular speed in z_2 direction minus the absolute angular velocity in this direction, measured by the rate gyro, should be divided by $\cos \beta$. The relative angle β can be measured by an encoder mounted at the tilt axis (see Figure 1). If the systems working space is limited to $-90° < \beta < +90°$, there is no risk to have a division by zero. In Figure 4 the pan motor general control structure is presented. For the tilt motor the structure is similar, but there is no division by $\cos \beta$ at the stabilization loop. The outer loop time delay is $\Delta t'$. The inner loop time delay is Δt and $\Delta t' > \Delta t$.

4.1. Inner Loop (Stabilization Loop)

In the inner loops PI controllers are used. The inputs are the errors comprising the desired absolute angular speeds minus the absolute angular speeds measured with the rate gyros. If digital controllers are used, the control signals provided to the pan and tilt motors are given, respectively, by the following equations:

$$u_{\alpha_m} = k_{P,\alpha_m} \cdot e_{\alpha_m} + k_{I,\alpha_m} \cdot \sum e_{\alpha_m} \cdot \Delta t, \qquad u_{\beta_m} = k_{P,\beta_m} \cdot e_{\beta_m} + k_{I,\beta_m} \cdot \sum e_{\beta_m} \cdot \Delta t, \tag{4.1}$$

where k_P and k_I are the gains of the proportional and integral terms, respectively. The errors are given by

$$e_{\alpha_m} = \frac{\omega_{z_2\text{desired}} - \omega_{z_2}}{\cos \beta}, \qquad e_{\beta_m} = \omega_{x_2\text{desired}} - \omega_{x_2}, \tag{4.2}$$

where $\omega_{z^2\text{desired}}$ and $\omega_{x^2\text{desired}}$ are body 2 desired absolute angular speeds in z_2 and x_2 directions. They are updated by the outer loop at every time interval $\Delta t'$. The two rate gyros assembled in body 2 provide new measurements of ω_{z^2} and ω_{x^2} in a faster rate. Therefore the voltages can be updated at every time interval Δt.

If the absolute angular speed takes a long time to reach the desired value, the sum in (4.1) may become excessively large and will keep increasing even if the control signal is saturated. The control signal will then remain saturated even when the error changes sign and it may take a long time before the integrator and the controller output come inside saturation range. This problem is called integrator windup and can be avoided if an antiwindup action is included at control. A common antiwindup strategy is the back-calculation [6] that is efficient but has a gain that needs to be adjusted. Therefore, in this work a simpler method will be used where the sum in (4.1) is interrupted if the control signal is at the maximum saturation limit and the error is positive, or, if the control signal is at the minimum saturation limit and the error is negative. In [3] the controllers do not use any model of the target tracking system; therefore the PI gains are obtained by Ziegler-Nichols frequency response method [6, 10].
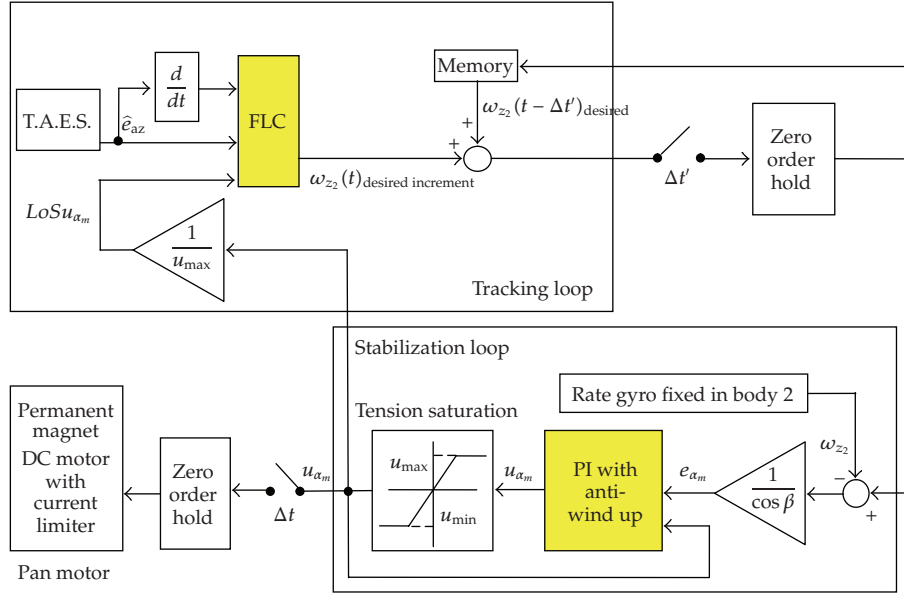
Figure 5: Pan motor general control structure.

4.2. Outer Loop (Tracking Loop)

In the outer loop, the TAES provides information about the targets elevation and azimuth angular errors, as shown in Figure 5, to a Fuzzy Logic Controller [5] whose output is an increment to the desired absolute angular speed, used as reference in the inner loop. In this work FLCs were chosen for the external loops because they are more flexible than traditional controllers, like PID, due to the possibility of implementing nonlinear mappings between inputs and outputs. It is not the objective of this work to study how the angular errors of azimuth and elevation are obtained. For the model used in this work the only necessary information is the time duration $\Delta t'$ of the whole process that occurs in the tracking loop. This time delay, from the instant when the TAES notices the target up to moment when the desired absolute angular speed is provided to the inner loop, may be relatively large when compared with Δt.

The inputs for the FLC used at the pan motor control are the angular azimuth error ($\hat{e}_{az}$), the derivative of this error with respect to time, and the control voltage saturation level ($LoSu_{\alpha_m}$). The $LoSu_{\alpha_m}$ is used to avoid increments for the desired absolute angular speeds if the control signal is near its superior saturation limit (or decrements if it is near the inferior limit). The pan motors FLC output is an increment for body 2 desired absolute angular speed in z_2 direction ($\omega_{z^2\text{desired increment}}$). This increment is added to the previous value of the referred angular speed (stored in memory), resulting in the value that will be provided to the internal loop ($\omega_{z^2\text{desired}}$). The inputs for the tilt motor FLC are the elevation angular error ($\hat{e}_{el}$), the derivative of this error with respect to time, and the control voltage saturation level ($LoSu_{\beta_m}$). The output is an increment for body 2 desired absolute angular speed in x_2 direction ($\omega_{x^2\text{desired increment}}$). The angular errors derivatives in a certain instant t can be approximated according to (4.3). The control signal saturation levels are given by (4.4), where u_{SAT} is the value of the maximum voltage that can be provided to

Figure 6: Membership functions for the FLC with three inputs ($\widehat{e}$, $\dot{\widehat{e}}$, and *LoSu*) and one output ($w_{\text{2desired increment}}$).

the motor:

$$\dot{\widehat{e}}_{\text{az}}(t) \approx \frac{\widehat{e}_{\text{az}}(t) - \widehat{e}_{\text{az}}(t - \Delta t')}{\Delta t'}, \qquad \dot{\widehat{e}}_{\text{el}}(t) \approx \frac{\widehat{e}_{\text{el}}(t) - \widehat{e}_{\text{el}}(t - \Delta t')}{\Delta t'}, \tag{4.3}$$

$$LoSu_{\alpha_m} = \frac{u_{\alpha_m}}{u_{\text{SAT},\alpha_m}}, \qquad LoSu_{\beta_m} = \frac{u_{\beta_m}}{u_{\text{SAT},\beta_m}}. \tag{4.4}$$

Both FLCs use the input and output membership functions shown in Figure 6 and the rule-base of Table 1, presented in [3]. The algorithm used for inference is max-min method and for defuzzification is the center of area method [5].

The membership functions and rule-base presented were adjusted through simulations of systems with outer loop time delay ($\Delta t'$) varying between 10 milliseconds and 20 milliseconds. For $\Delta t'$ out of that range a new adjustment may be necessary.

4.3. Backlash Compensation Block

4.3.1. Overview

The controller described in previous sections is projected for systems without backlash. Its performance will be significantly worse if large backlash angles are present in gear transmission, since backlash causes delays, oscillations, and inaccuracy. To avoid backlash, some designers, as Arambel et al. [23] and Borrello [24], prefer to use direct drives, but this solution is not always possible, since a stronger actuator, that may be too large or heavy for the system, will be required. Other design techniques can be employed to avoid the harmful effects of backlash in systems with gear transmission, as the drive/antidrive concept, discussed by Haider et al. [25] or the auxiliary brakes presented by Gruzman and Weber [26]. Despite the effectiveness of those techniques, they increase the complexity and weight of the target tracking system, since extra components will have to be assembled in the Cardan

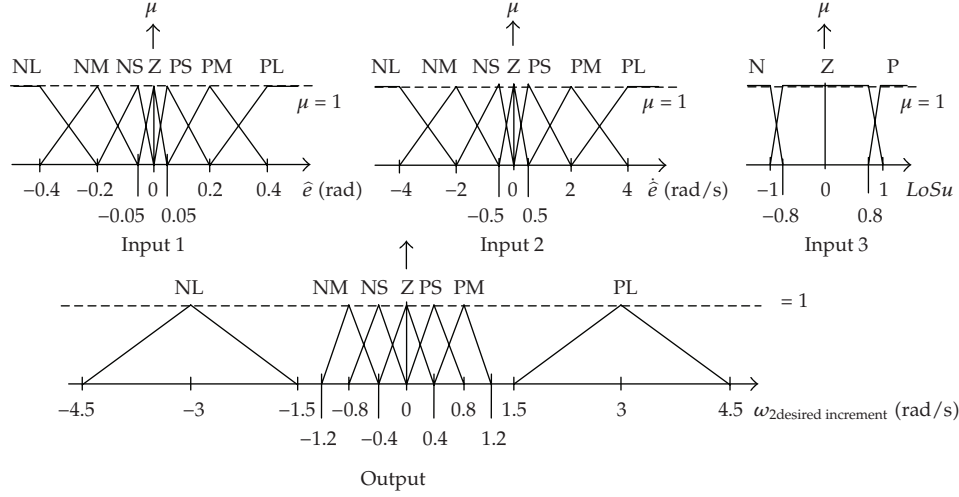**Table 1:** Rule-base for the FLC with three inputs ($\hat{e}$, $\dot{\hat{e}}$, and $LoSu$) and one output ($\omega_{2\text{desired increment}}$).

			$\hat{e}$						
			NL	NM	NS	Z	PS	PM	PL
		NL	Z	Z	Z	Z	Z	PS	PM
		NM	Z	Z	Z	Z	Z	PS	PM
		NS	Z	Z	Z	Z	PS	PM	PM
	$LoSu = $ N	Z	Z	Z	Z	Z	PM	PM	PL
		PS	Z	Z	Z	PS	PM	PL	PL
		PM	Z	Z	Z	PS	PL	PL	PL
		PL	Z	Z	Z	PM	PL	PL	PL
		NL	NL	NL	NL	NM	Z	PS	PM
		NM	NL	NL	NL	NS	Z	PS	PM
		NS	NL	NL	NM	NS	PS	PM	PM
$\dot{\hat{e}}$	$LoSu = $ Z	Z	NL	NM	NM	Z	PM	PM	PL
		PS	NM	NM	NS	PS	PM	PL	PL
		PM	NM	NS	Z	PS	PL	PL	PL
		PL	NM	NS	Z	PM	PL	PL	PL
		NL	NL	NL	NL	NM	Z	Z	Z
		NM	NL	NL	NL	NS	Z	Z	Z
		NS	NL	NL	NM	NS	Z	Z	Z
	$LoSu = $ P	Z	NL	NM	NM	Z	Z	Z	Z
		PS	NM	NM	NS	Z	Z	Z	Z
		PM	NM	NS	Z	Z	Z	Z	Z
		PL	NM	NS	Z	Z	Z	Z	Z

suspension. The approach presented in [25] requires two motors for each axis while the approach presented in [26] requires braking systems for each axis. Some authors as Rzasa [27] mention the use of dry friction in the drive train to eliminate limit cycles, but this technique reduces the efficiency of the actuator, since extra power will be required to compensate this additional torques. Therefore, in this work, the effects of backlash will be compensated by the controllers. For that, a backlash compensation block is added in the stabilization loop of each motors controller. This block should actuate whenever body 1 (or 2) transverses a backlash gap in gear transmission. To reduce the effects of backlash, a possible strategy consists to transverse the backlash region quickly in order to restore the controllability. In literature this approach is often called as "strong action in backlash gap" and it is more concerned with the lost motion at the load. However, problems related with dynamic stability can occur, because reengagement may happen with large relative velocity (undesired collisions will occur). To avoid this problem some authors prefer to implement the called the "weak action in backlash gap." It avoids problems with dynamic stability, but tracking performance may be poor, since the backlash region is transversed with reduced speed. In Nordin and Gutman [28] an extensive literature survey of both methods is presented. Some authors as Menon and Krishnamurthy [29] and Moscrop [30] uses a combination of the "weak" and "strong" actions while transversing the backlash gap, because it exploits the advantages of both (smooth reengagement is ensured at the same time that the controller tries to transverse quickly the backlash region). In the present work a combination of both actions will be considered, but
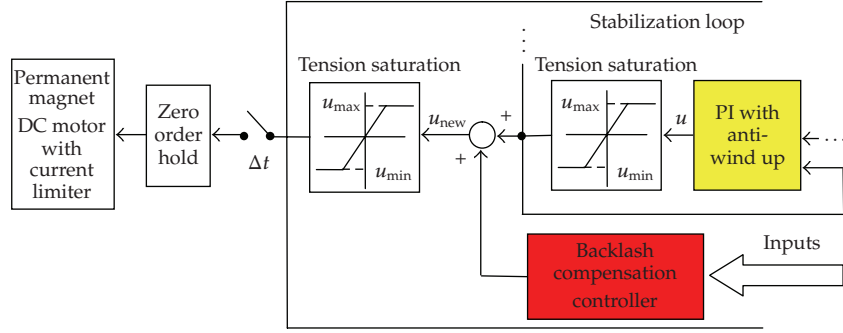
Figure 7: Backlash compensation controller in the stabilization loop.

the compensation signal will be provided to the output of the main controller (in this work corresponds to the PI controller of the stabilization loop) instead of its input, as is done in [29–34]. The reason for providing the compensation signal to the output of the main controller will be discussed further in this section.

The backlash compensation block is added in the stabilization loop, according to Figure 7, to provide an extra tension (positive or negative) for the motor.

Two tension saturation blocks should be used: the first before the backlash compensation block and the second after it. If there is no tension saturation block before the backlash compensation controller, the compensation tension may be in some cases ignored, as in the following example.

Example 4.1. A target tracking system is supplied by a battery that provides tensions between −24 V and 24 V; therefore the tension saturation block should not allow control tensions out of this range. If the output of the PI controller is a 30 V tension, the output of the backlash compensation controller is a −4 V tension, and there is no tension saturation block after the output of the PI controller, then the tension saturation block, that is after the sum of the PI and backlash compensation block outputs, will have an input of 26 V and an output of 24 V, since 26 V is beyond its saturation level. As it can be noticed, the backlash compensation controller had no influence at the output of the stabilization loop. With two tension saturation block, as shown in Figure 7, this tension would be reduced to 20 V, because the first block would reduce the PI controller output to 24 V. Therefore, in this case the backlash compensation controller influenced the output of the stabilization loop.

The idea behind the backlash compensation approach used in this work can be explained with the schematic draw of Figure 8, where A is the driving mechanism, B the driven mechanism, x_A the driving mechanism position, x_B the driven mechanism position, v_A the driving mechanism speed, v_B the driven mechanism speed, and l the total backlash size. Initially consider that B is being driven to the left ($v_B < 0$) as in Figure 8(a). To keep this motion, a negative tension (u) is being provided to driving mechanism, which is in contact with the left teeth of B. If in a certain instant it is desired an increase in v_B, a positive tension will be provided to A, since it will have to act in the right teeth of B, as shown in Figure 8(b). For that, A must pass through the backlash gap. As the backlash region should be quickly transposed, an additional positive tension ($u_{compensation}$) is provided to the driving mechanism by the backlash compensation controller, as shown in Figure 8(c). Just before
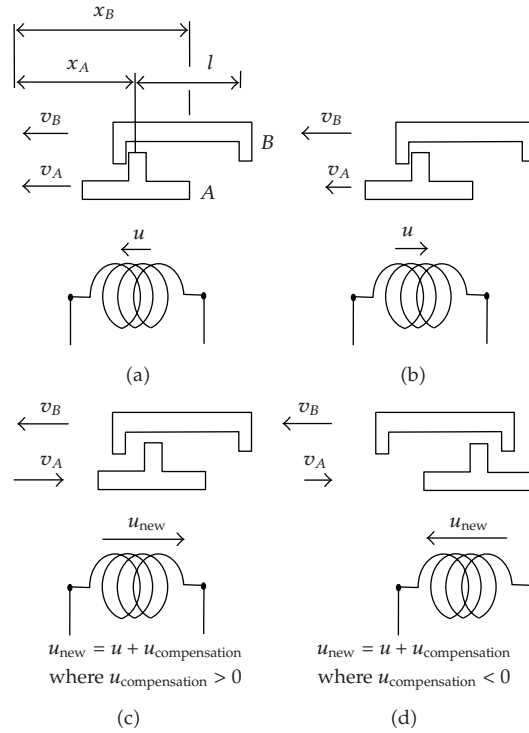
Figure 8: Schematic of a drive system with backlash and the compensation approach.

A touches the right teeth of B, the compensation tension becomes negative, in order to reduce the speed of A and smooth the collision between A and the right teeth of B. As soon as the contact is reestablished, the compensation tension becomes zero, as in the beginning of the motion.

In order to have the desired relative motion between A and B when transposing the backlash gap, the value of the compensation tension should be adequately calculated and depends of u. If u is negative, A must contact the left teeth of B; otherwise it must contact the right teeth of B. The relative position between A and B (x_{rel}) is given by (4.5) and the relative velocity between A and B (v_{rel}) by (4.6):

$$x_{rel} = x_B - x_A, \tag{4.5}$$

$$v_{rel} = v_B - v_A. \tag{4.6}$$

The theory presented for the linear system of Figure 8 can be applied to the target tracking system of this work, but the correspondent generalized coordinates of the target tracking system must be used, according to Table 2.

Since in most of the practical applications many parameters of the system may be unknown, the backlash compensation block developed in this work consists of a controller that does not require a model of the system. Besides, it differs from other backlash compensation techniques that do not use a model of the system, because they provide compensation to the inputs instead of the output of the main controller, as demonstrated

Table 2: Correspondent generalized coordinates.

Variable	Rotor 1 – body 1 set	Rotor 2 – body 2 set
$x_{\text{rel}} = x_B - x_A$	$\Delta_\alpha = \alpha - \dfrac{\alpha_m}{N_\alpha}$	$\Delta_\beta = \beta - \dfrac{\beta_m}{N_\beta}$
$v_{\text{rel}} = v_B - v_A$	$\dot\Delta_\alpha = \dot\alpha - \dfrac{\dot\alpha_m}{N_\alpha}$	$\dot\Delta_\beta = \dot\beta - \dfrac{\dot\beta_m}{N_\beta}$
l	$2 \cdot \eta_\alpha$	$2 \cdot \eta_\beta$

(a) Traditional approach
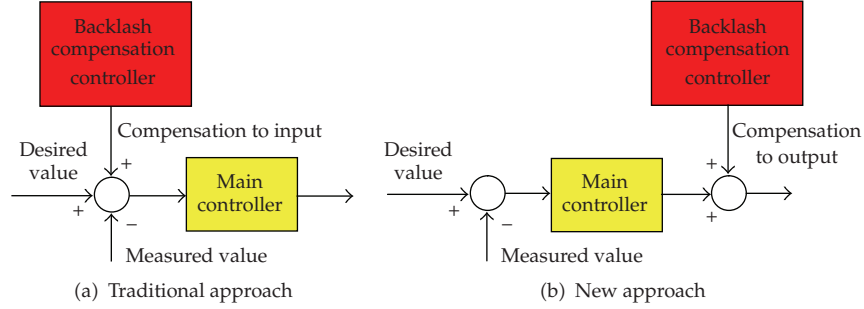
(b) New approach

Figure 9: Schematic of traditional backlash compensation and the new approach.

in Figure 9. In this work it corresponds to the PI controller of the stabilization loop, which is the last controller before the actuator.

Some examples of these backlash compensation techniques can be found in Kim et al. [31], Mokhtari and Barati [32], and Santos and Vieira [33]. In those works a compensation value is added to a desired position command for the driven mechanism. The new desired position command is the input to some specific controller. In Woo et al. [34] a compensation value is added to a desired velocity command for the driven mechanism. The new desired velocity command is the input to a PD controller.

The flexibility is an advantage of the backlash compensation approach proposed in this work, when compared to the traditional techniques, since it is not designed for a specific type of main controller. In fact it can be a PID, FLC, state feedback controller, LQR, and so forth. Besides, it can be implemented for main controllers working with different feedback variables as joint angles, joint speeds, absolute angular speeds, absolute angular accelerations, and so forth. This is a desired factor for backlash compensation controllers used in gimbaled tracking systems, because not every system has a controller that pursues to keep body 2 absolute angular speeds in desired values. Several examples of gimbaled tracking systems where the main controller pursues to keep the joint angles in desired values can be found in the works presented by Skoglar [35], Liu [36], and Gruzman et al. [37]. Systems where the main controller uses absolute angular acceleration feedback are also found in literature, as in Algrain and Quinn [38].

The driven axes and rotors angular positions and velocities are required for the backlash compensation controller proposed in this work. It is assumed that sensors are assembled after the rotors and the axes in order to measure the generalized coordinates (α, α_m, β, and β_m) and its first derivatives, as presented in Figure 1. If the cost of using sensors after the rotors and load becomes prohibitive, they can be assembled only after the axes, but state observers [10] such as Kalman filter [6] will be required to estimate the rotors
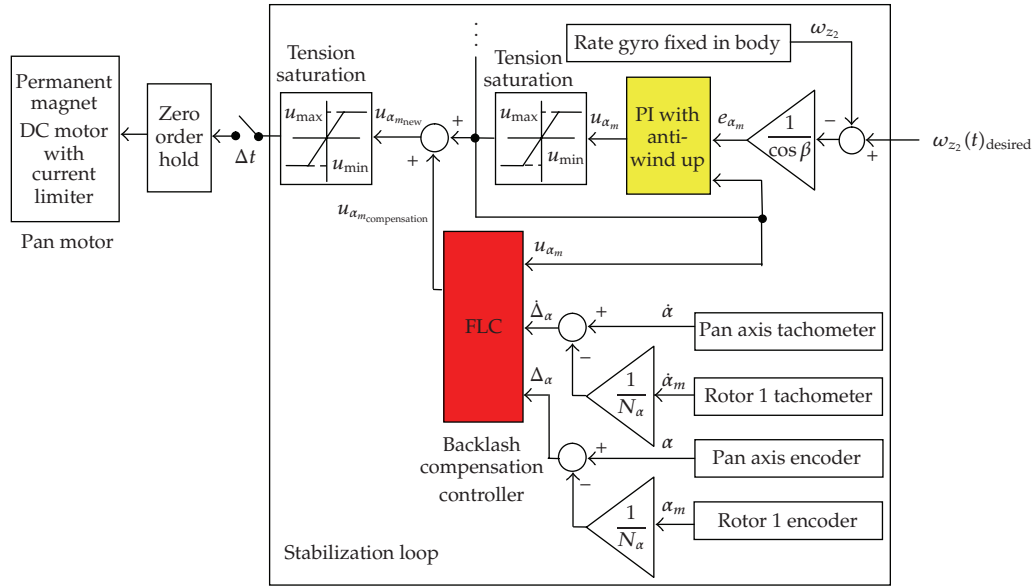
Figure 10: Pan motor stabilization loop with backlash compensation.

generalized coordinates and its first derivatives. The inconvenience of state estimation is that the parameters of the DC motors are required. Frequently they are not available, and additional experiments, as presented in [39, 40], may be necessary.

The total backlash angles in each gear are required for the compensation block proposed. Therefore, it assumed in this work that this parameter is known. Simple experiments can be done for a good estimation of the total backlash angle in a transmission train: the rotor is kept fixed and the driven axis is turned manually until no further motion is allowed, and the angular displacement corresponds to the total backlash angle. In addition, the total backlash can be estimated online by adaptive controllers [33, 34, 41].

4.3.2. Controller Used in the Backlash Compensation Block

The strategy to compensate the backlash nonlinearity consists of a heuristic approach, as presented in Figure 8; therefore a fuzzy controller will be used in the backlash compensation block, because it provides a natural framework to incorporate heuristics. Three inputs are necessary for the FLC backlash compensation block: the tension calculated in the main controller (u), the relative angular position between the driven axis (divided by the gear ratio) and the rotor ($\triangle$), and the relative angular velocity between the driven axis (divided by the gear ratio) and the rotor ($\dot{\triangle}$), both given by the expressions in Table 2. The output is an extra control tension ($u_{\text{compensation}}$) that will be added to the tension calculated by the main controller, as shown in Figure 10, where the pan motor stabilization loop with the backlash compensation block is presented.

In Figure 10 it is considered that tachometers are used to obtain the first derivatives of the generalized coordinates, but it is also possible to obtain finite differences approximations if only encoders are available.
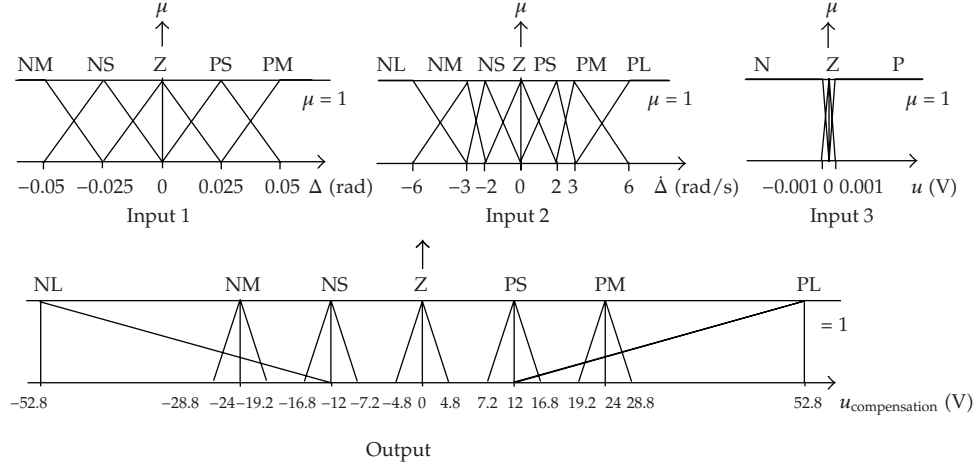
Figure 11: Membership functions for the FLC with three inputs (Δ, $\dot{\Delta}$, and u) and one output ($u_{\text{compensation}}$).

Table 3: Rule-base for the FLC with three inputs (Δ, $\dot{\Delta}$, and u) and one output ($u_{\textbf{compensation}}$).

			Δ				
			PM	PS	Z	NS	NM
		PL	PH	PH	PH	PM	PM
		PM	PH	PM	PM	PM	PS
		PS	PH	PS	PS	PS	PS
	$u = P$	Z	PM	PS	PS	PS	Z
		NS	PM	PS	PS	Z	NM
		NM	PS	PS	Z	NM	NH
		NL	PS	Z	NM	NH	NH
		PL	Z	Z	Z	Z	Z
		PM	Z	Z	Z	Z	Z
		PS	Z	Z	Z	Z	Z
$\dot{\Delta}$	$u = Z$	Z	Z	Z	Z	Z	Z
		NS	Z	Z	Z	Z	Z
		NM	Z	Z	Z	Z	Z
		NL	Z	Z	Z	Z	Z
		PL	PH	PH	PM	Z	NS
		PM	PH	PM	Z	NS	NS
		PS	PM	Z	NS	NS	NM
	$u = N$	Z	Z	NS	NS	NS	NM
		NS	NS	NS	NS	NS	NH
		NM	NS	NM	NM	NM	NH
		NL	NM	NM	NH	NH	NH

The FLC backlash compensation block for the pan and tilt motor control use the input and output membership functions shown in Figure 11 and the rule-base of Table 3, adjusted

(a)

(b)

Figure 12: Angular errors of elevation (a) and azimuth (b).

for total backlash angles (2η) of 0.1 rad and saturation tensions of ±24 V. The algorithm used for inference is max-min method and for defuzzification is the center of area method [5].

The membership functions and rule-base for the backlash compensation controller are chosen to reduce delays, oscillations, and inaccuracies, by providing motors overdrives such that backlash region is quickly, but with smooth reengagement, transversed, according to the strategy presented in Figure 8.
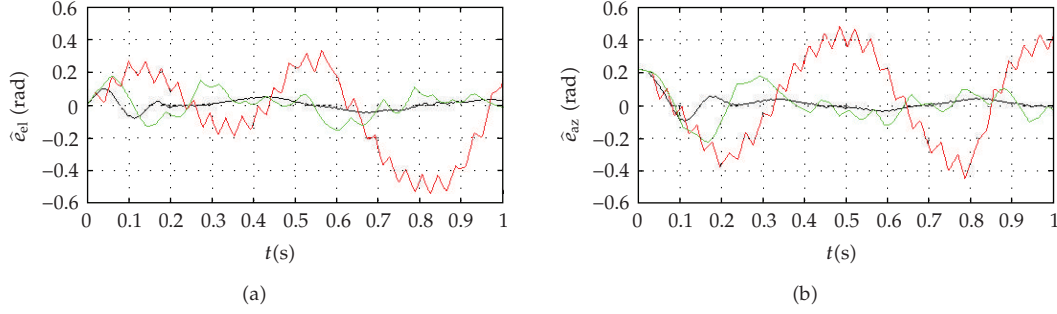
5. Simulation Results

In this section two different cases will be simulated. In the first the target will move in a circle five meters ahead body 0, that is fixed in inertial frame. In the second case, body 0 and the target will move. The objective of the controllers is to keep the angular errors of azimuth and elevation equal to zero. The curves in black are the angular errors of a system without backlash compensation controllers and $\eta = 0.0005$ rad. This value was chosen to simulate a system with negligible backlash. The curves in red correspond to a system with large backlash in gear train ($\eta = 0.05$ rad) and also without the backlash compensation controller. The curves in green correspond to a system with $\eta = 0.05$ rad and backlash compensation controllers for each motor. The root mean square (rms) of elevation and azimuth angular errors of each simulation are also presented. Measurement errors and noise in sensors were not considered. The algorithm used for calculating the approximate solution of the ODE was a fourth-order Runge-Kutta with variable step-size.

The following data were used in all simulation.

General Data

Gravity acceleration (g): 9,81 m/s². Initial conditions are

$$
\begin{aligned}
q_1(0) = q_2(0) &= q_3(0) = q_4(0) = q_5(0) = q_6(0) = q_7(0) = q_8(0) \\
&= i_{\alpha_m}(0) = i_{\alpha_m}(0) = \theta_{b,\alpha}(0) = \theta_{b,\beta}(0) = 0.
\end{aligned}
\tag{5.1}
$$

Resultant coefficient of rigidity in the drive trains (k_s) is 3000 Nm/rad. Resultant damping coefficient in the drive trains (c_s) is 2 Nms/rad.

Body 1 and Body 2 Data

Body 1 mass (m_1) is 0.3 kg. Body 2 mass (m_2) is 0.4 kg. Viscous friction coefficients at bodies 1 and 2 axes are $c_\alpha = c_\beta = 0.01$ kg m^2/rad s. Dry torques friction at bodies 1 and 2 axes is disregarded. Position vector of the center of the Cardan suspension (point b) with respect to body 0 center of mass (in coordinates of body 0 fixed frame), position vector of body 1 center of mass with respect to point b (in coordinates of body 1 fixed frame), and position vector of body 2 center of mass with respect to point b (in coordinates of body 2 fixed frame) are given, respectively, by

$$\mathbf{{}^0_a d_b} = \begin{bmatrix} 0 & 0.5 & 0 \end{bmatrix}^T \text{m}, \qquad \mathbf{{}^1_b d_c} = \begin{bmatrix} 0 & 0 & -0.02 \end{bmatrix}^T \text{m}, \qquad \mathbf{{}^2_b d_d} = \begin{bmatrix} 0.01 & 0.04 & 0.025 \end{bmatrix}^T \text{m}. \quad (5.2)$$

Bodies 1 and 2 inertia tensors, expressed in coordinates of body 1 and 2 frames, are respectively,

$$^1 II_1 = \begin{bmatrix} 2.59 & -0.44 & 0.14 \\ -0.44 & 4.69 & -0.69 \\ 0.14 & -0.69 & 2.72 \end{bmatrix} \cdot 10^{-4} \text{ kg m}^2, \qquad ^2 II_2 = \begin{bmatrix} 9.76 & -1.14 & -0.32 \\ -1.14 & 4.67 & -1.51 \\ -0.32 & -1.51 & 9.57 \end{bmatrix} \cdot 10^{-4} \text{ kg m}^2.$$

$$(5.3)$$

Motors Data

Rotors 1 and 2 inertia tensors, expressed in coordinates of body 1 and 2 frames, are respectively,

$$^1 II_{r_1} = \begin{bmatrix} 0.3 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & 0.3 \end{bmatrix} \cdot 10^{-4} \text{ kg m}^2, \qquad ^2 II_{r_2} = \begin{bmatrix} 0.3 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & 0.3 \end{bmatrix} \cdot 10^{-4} \text{ kg m}^2. \quad (5.4)$$

Viscous friction coefficients at the rotors axes are $c_{\alpha_m} = c_{\beta_m} = 0.0004$ kg m^2/rad s. Dry friction dynamic torques at the rotors axes are $T_{\text{din},\alpha_m} = T_{\text{din},\beta_m} = 0.013$ Nm. Maximum dry friction static torques at the rotors axes are $T_{\text{limit},\alpha_m} = T_{\text{limit},\beta_m} = 0.017$ Nm. Gear reduction ratio is $N_\alpha = N_\beta = 30$. Armature resistance is $R_{\alpha_m} = R_{\beta_m} = 2.3\,\Omega$. Armature inductance is $l_{\alpha_m} = l_{\beta_m} = 0.003$ H. Torque and back-emf constants are $k_{b,\alpha_m} = k_{b,\beta_m} = k_{m,\alpha_m} = k_{m,\beta_m} = 0,045$ V/rad/s. Maximum and minimum current limits at the motor armatures are $i_{\max} = 10$ A and $i_{\min} = -10$ A. Maximum control voltages are $+24$ V. Minimum control voltages are -24 V.

Controllers Data

Tracking loop (FLC) is the membership functions and rule-base presented in Section 4.3.

Stabilization loop (PI control): the following gains were obtained with Ziegler-Nichols tuning in [3] for a system without backlash in gear transmission: $k_{P,\alpha_m} = k_{P,\beta_m} = 17.41$; $k_{I,\alpha_m} = k_{I,\beta_m} = 2176.88$.

Backlash compensation block (FLC): membership functions and rule-base presented in Section 4.3.

Table 4: Root mean square (rms) of elevation and azimuth angular errors.

Simulation	Correspondent curve in Figure 12	rms $\hat{e}_{el}$	rms $\hat{e}_{az}$
$\eta = 0.0005$ rad and no backlash compensation	black	0.0338 rad	0.0525 rad
$\eta = 0.05$ rad and no backlash compensation	red	0.2465 rad	0.2656 rad
$\eta = 0.05$ rad and backlash compensation	green	0.0766 rad	0.1 rad

Outer loop time delay ($\Delta t'$) is 15 milliseconds. Inner loop time delay (Δt): 1 milliseconds.

(a) First Case: Body 0 Fixed in Inertial Frame and Target with Prescribed Motion

The results presented in Figure 12 correspond to simulations where body 0 was fixed within inertial frame, with its center of mass (point a in Figure 1) coinciding with inertial frame origin ($\mathcal{O}$). The target (tt) position vector with respect to $\mathcal{O}$, in inertial frame coordinates (G), had the following prescribed motion:

$$\,^{G}_{\mathcal{O}}\mathbf{d_{tt}}(t) = \begin{bmatrix} \cos(4\pi t) & 5 & \text{sen}(4\pi t) \end{bmatrix}^{T} \text{m}. \tag{5.5}$$

According to the results presented in Figure 12 it can be observed a significant loss in precision when large backlashes are present in gear transmission, by comparing the red curves ($\eta = 0.05$ rad and no backlash compensation) and the black curves ($\eta = 0.0005$ rad and no backlash compensation). There is also a significant increase in rms of the elevation and azimuth angular errors when backlash angles are larger, as shown in Table 4. When the backlash compensation controller is used in a system with $\eta = 0.05$ rad (green curves), there is an improvement in performance, even if less accurate than the very small backlash case without compensation (black curves), as it can be observed in the curves of Figure 12 and in the rms values of Table 4.

(b) Second Case: Body 0 and Target with Prescribed Motion

The results presented in Figure 13 correspond to simulations where body 0 and the target (tt) had an oscillatory prescribed motion. Their motion could represent, for example, a missile (body 0) chasing an airplane (target). The position vector of the target with respect to inertial frame origin ($\mathcal{O}$) in inertial frame (G) coordinates, the position vector of body 0 center of mass with respect to $\mathcal{O}$ in G coordinates, and body 0 orientation coordinates with respect to G are given, respectively, by

$$\,^{G}_{\mathcal{O}}\mathbf{d_{tt}}(t) = \begin{bmatrix} \cos(3.5\pi t) & 5 + 300t & \text{sen}(2.7\pi t) \end{bmatrix}^{T} \text{m},$$

$$\,^{G}_{\mathcal{O}}\mathbf{d_{a}}(t) = \begin{bmatrix} \sin(2\pi t) & 300t & \sin(3\pi t) \end{bmatrix}^{T} \text{m}, \tag{5.6}$$

$$\delta(t) = 0.2\sin(6\pi t) \text{ rad}, \qquad \psi(t) = 0.2\cos(6\pi t) \text{ rad}, \qquad \gamma(t) = 0.$$

Once again, by the results presented in Figure 13 and Table 5, it can be observed a significant loss in precision when large backlashes are present in gear transmission (red curves). When the backlash compensation controller is used in a system with large backlash
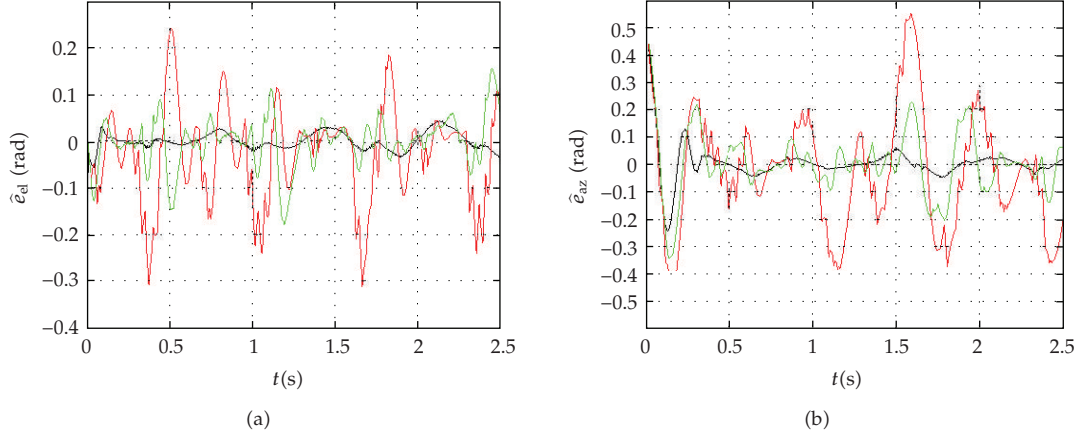
(a)

(b)

Figure 13: Angular errors of elevation (a) and azimuth (b).

Table 5: Root mean square (rms) of elevation and azimuth angular errors.

Simulation	Correspondent curve in Figure 12	rms $\hat{e}_{el}$	rms $\hat{e}_{az}$
$\eta = 0.0005$ rad and no backlash compensation	black	0.0183 rad	0.0696 rad
$\eta = 0.05$ rad and no backlash compensation	red	0.1051 rad	0.2122 rad
$\eta = 0.05$ rad and backlash compensation	green	0.0539 rad	0.1173 rad

(green curves), a performance improvement can be observed, even if it is less accurate than the very small backlash case without compensation (black curves).

6. Conclusions

This paper presented the modeling and control of a target tracking system assembled into a moving body. The system is modeled in time domain, instead of the most traditional frequency domain approach. This allowed a more realistic model of the system, since the inertial parts of the equations of motion were not linearized and other nonlinearities such as dry friction, backlash in gear transmission, control tensions saturation, and armature current saturation were included in the model. Different time delays in each control loop were also included in the model. The presentation of the model was done in several steps, each of them containing a literature survey of recent works where similar problems were analyzed. Initially, the generalized coordinates used in the model were defined and the Lagrange equations for the system were presented. Despite the assumption that the system is composed of rigid bodies, the Lagrange equations allowed to consider flexibility in gear transmissions that were included in the nonconservative generalized torques of the equation. In this torques, the motors electromotive force, viscous friction, backlash, and dry friction were also included. The model does not include body flexibility and aerodynamic effects, which may increase inaccuracy in the system performance. Such effects should appear more drastically in the cases that the TAES is large and more flexible as an antenna. Further developments of the model might be pointed toward this direction.

Next, a methodology was presented for the numerical integration of the equations of motion simultaneously with the update of the control signals with time delays. For finding of

the control signals, independent controllers, with two loops each, were used for each motor. The outer loop, that contains a fuzzy logic controller, used information about the targets angular errors of azimuth and elevation and calculates the desired absolute angular speeds for body 2, to point the axial axis to the target. The inner loop, that is faster, uses information from the other sensors in order to keep body 2 angular absolute speeds in the desired values. A PI controller with antiwindup action is used in this loop. Tuning the gains of PI controller through Ziegler-Nichols method requires obtaining an absolute angular velocity response of the device with its actual parameters. This step could be avoided if other controller design techniques are used, such as robust control.

Afterwards, special attention was given to the problems caused by backlash in gear transmission. To reduce its effects a backlash compensation block was included in the stabilization loop. This compensation block consisted of a fuzzy logic controller whose inputs are the tension obtained in the PI controller and the relative angular position and velocity between the driven axis and the rotor (divided by the gear ratio). The output is an extra tension that is added to the tension calculated in the PI controller, assuring that the backlash region is quickly transverse, but with smooth reengagement. The backlash compensation controller can be used in combination with different principal controllers, which in this work corresponds to the inner loop PI controller. It provides compensation signal to the output of the main controller, instead of changing its inputs, as usual in traditional backlash compensation techniques. By last, results of numerical simulation were presented. It is shown the improvement in performance of a system that has large backlash in gear transmission, but whose controller has been previously tuned to work without backlash. A suggestion for future work is the inclusion of noise and sensors inaccuracy in the model.

Acknowledgment

References

[1] M. A. Brdyś and J. J. Littler, "Fuzzy logic gain scheduling for non-linear servo tracking," *International Journal of Applied Mathematics and Computer Science*, vol. 12, no. 2, pp. 209–219, 2002.

[2] M. K. Masten, "Inertially stabilized platforms for optical imaging systems," *IEEE Control Systems Magazine*, vol. 28, no. 1, pp. 47–64, 2008.

[3] M. Gruzman, H. I. Weber, and L. L. Menegaldo, "Control of a target tracking system embedded in a moving body," in *Proceedings of the 20th International Congress of Mechanical Engineering (COBEM '09)*, Gramado, Brazil, 2009.

[4] P. J. Kennedy and R. L. Kennedy, "Direct versus indirect line of sight (LOS) stabilization," *IEEE Transactions on Control Systems Technology*, vol. 11, no. 1, pp. 3–15, 2003.

[5] K. M. Pasino and S. Yurkovich, *Fuzzy Control*, Addison-Wesley Longman, Menlo Park, Calif, USA, 1998.

[6] K. J. Åström and R. M. Murray, *Feedback Systems: An Introduction for Scientists and Engineers*, Princeton University Press, Princeton, NJ, USA, 2008.

[7] J. Downs, S. Smith, J. Schwickert, and L. Stockum, "High performance gimbal control for self-protection weapon systems," in *Proceedings of the 7th Conference on Aqcuisition, Tracking and Pointing*, vol. 3365 of *Proceedings of SPIE*, pp. 77–86, Orlando, Fla, USA, April 1998.

[8] H.-P. Lee and I.-E. Yoo, "Robust control design for a two-axis gimbaled stabilization system," in *Proceedings of the IEEE Aerospace Conference Proceedings (AC '08)*, pp. 1–7, Big Sky, Mont, USA, March 2008.

[9] S. H. Cradall, D. C. Karnopp, F. J. Kurtz, and D. C. Pridemorebrown, *Dinamics of Mechanical and Eletromechanical Systems*, McGraw-Hill, New York, NY, USA, 1968.

[10] K. Ogata, *Engenharia de Controle Moderno*, Prentice-Hall, São Paulo, Brazil, 4th edition, 2003.

[11] M. Nordin and P.-O. Gutman, "Non-linear speed control of elastic systems with backlash," in *Proceedings of the 39th IEEE Conference on Decision and Control (CDC '00)*, vol. 4, pp. 4060–4065, Sydney, Australia, December 2000.

[12] A. Lagerberg and B. Egardt, "Evaluation of control strategies for automotive powertrains with backlash," in *Proceedings of the 6th International Symposium on Advanced Vehicle Control (AVEC '02)*, Hiroshima, Japan, 2002.

[13] J. H. Baek, Y. K. Kwak, and S. H. Kim, "Backlash estimation of a seeker gimbal with two-stage gear reducers," *International Journal of Advanced Manufacturing Technology*, vol. 21, no. 8, pp. 604–611, 2003.

[14] P. Dupont, V. Hayward, B. Armstrong, and F. Altpeter, "Single state elastoplastic friction models," *IEEE Transactions on Automatic Control*, vol. 47, no. 5, pp. 787–792, 2002.

[15] V. Lampaert, J. Swevers, and F. Al-Bender, "Experimental comparison of different friction models for accurate low-velocity tracking," in *Proceedings of the 10th Mediterranean Conference on Control and Automation (MED '02)*, Lisbon, Portugal, July 2002.

[16] V. Lampaert, J. Swevers, and F. Al-Bender, "Comparison of model and non-model based friction compensation techniques in the neighbourhood of pre-sliding friction," in *Proceedings of the American Control Conference (AAC '04)*, vol. 2, pp. 1121–1126, Boston, Mass, USA, June-July 2004.

[17] C. Makkar, W. E. Dixon, W. G. Sawyer, and G. Hu, "A new continuosly differentiable friction model for control systems design," in *Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM '05)*, pp. 600–605, Monterey, Calif, USA, July 2005.

[18] B. D. Nguyen, *Modelling of frictional contact conditions in structures*, M.S. Dissertation, Georgia Institute of Technology, Atlanta, Ga, USA, 2005.

[19] S. V. Duin, *Impulse control systems for servomechanisms with nonlinear friction*, Ph.D. thesis, School of Electrical, Computer and Telecomunications Engineering, University of Wollongong, Sydney, Australia, 2006.

[20] S. Andersson, A. Söderberg, and S. Björklund, "Friction models for sliding dry, boundary and mixed lubricated contacts," *Tribology International*, vol. 40, no. 4, pp. 580–587, 2007.

[21] J.-C. Piedbœuf, J. de Carufel, and R. Hurteau, "Friction and stick-slip in robots: simulation and experimentation," *Multibody System Dynamics*, vol. 4, no. 4, pp. 341–354, 2000.

[22] M. A. Ruggiero and V. L. R. Lopes, *Cálculo Numérico—Aspectos Teóricos e Computacionais*, Makron Books do Brasil, São Paulo, Brazil, 2nd edition, 1996.

[23] P. O. Arambel, R. K. Mehra, B. Bradley, et al., "New generation high speed turret and pseudo bang-bang controller," in *Proceedings of the American Control Conference (AAC '01)*, vol. 4, pp. 2561–2566, Arlington, Va, USA, June 2001.

[24] M. Borrello, "A multi stage pointing acquisition and tracking (PAT) control system approach for air to air laser communications," in *Proceedings of the American Control Conference (AAC '05)*, vol. 6, pp. 3975–3980, Portland, Ore, USA, June 2005.

[25] Z. Haider, F. Habib, M. H. Mukhtar, and K. Munawar, "Design, control and implementation of 2-DOF motion tracking platform using drive-anti drive mechanism for compensation of backlash," in *Proceedings of the IEEE International Workshop on Robotic and Sensor Environments (ROSE '07)*, pp. 57–62, Ottawa, Canada, October 2007.

[26] M. Gruzman and H. I. Weber, "Simulation of a positioning system with backlash, flexibility and a proportional controller aided by a brake," in *Proceedings of the 15th International Workshop on Dynamics and Control*, Barcelona, Spain, 2009.

[27] J. R. Rzasa, *Design and application of pan and tilt servo gimbals in pointing, acquisition and tracking*, M.S. Dissertation, University of Maryland, College Park, Md, USA, 2007.

[28] M. Nordin and P.-O. Gutman, "Controlling mechanical systems with backlash—a survey," *Automatica*, vol. 38, no. 10, pp. 1633–1649, 2002.

[29] K. Menon and K. Krishnamurthy, "Control of low velocity friction and gear backlash in a machine tool feed drive system," *Mechatronics*, vol. 9, no. 1, pp. 33–52, 1999.

[30] J. W. Moscrop, *Modelling, analysis and control of linear feed axes in precision machine tools*, Ph.D. thesis, School of Electrical, Computer and Telecomunications Engineering, University of Wollongong, Sydney, Australia, 2008.

[31] N. H. Kim, U. Y. Huh, and J. G. Kim, "Fuzzy position control of motor plant with backlash," in *Proceedings of the 30th Annual Conference of the IEEE Industrial Electronics Society (IECON '04)*, vol. 3, pp. 3190–3195, Busan, South Korea, November 2004.

[32] H. Mokhtari and F. Barati, "A new scheme for a mechanical load position control driven by a permanent magnet DC motor and a nonzero backlash gearbox," in *Proceedings of the IEEE International Symposium on Industrial Electronics (ISIE '06)*, vol. 3, pp. 2052–2057, Montréal, Canada, July 2006.

[33] T. S. Santos and F. H. T. Vieira, "Uma proposta de controle adaptativo para sistemas discretos no tempo com folga desconhecida," in *Poceedings of the Congresso Bras leiro de Automática*, Juiz de Fora, Brazil, 2008.

[34] K. T. Woo, L.-X. Wang, F. L. Lewis, and Z. X. Li, "A Fuzzy system compensator for backlash," in *Proceedings of the IEEE International Conference on Robotics and Automation*, vol. 1, pp. 181–186, Leuven, Belgium, May 1998.

[35] P. Skoglar, *Modelling and Control of IR/EO-Gimbal for UAV Surveillance Applications*, Department of Electrical Engineering, Linköping Institute of Technology, Linköping, Sweden, 2002.

[36] T. L. Liu, *The development of a target-lockup optical remote sensing system for unmanned aerial vehicle*, M.S. Dissertation, Institute of Aeronautics and Astronautics, National Cheng Kung University, Tainan, Taiwan, 2004.

[37] M. Gruzman, H. I. Weber, and L. L. Menegaldo, "Modeling of a pointing and target tracking system," in *Proceedings of the 13th International Symposium on Dynamic Problems of Mechanics*, Angra dos Reis, Brazil, 2009.

[38] M. C. Algrain and J. Quinn, "Accelerometer based line-of-sight stabilization approach for pointing and tracking systems," in *Proceedings of the 2nd IEEE Conference on Control Applications*, vol. 1, pp. 159–163, Vancouver, Canada, September 1993.

[39] J. C. Basilio and M. V. Moreira, "State-space parameter identification in a second control laboratory," *IEEE Transactions on Education*, vol. 47, no. 2, pp. 204–210, 2004.

[40] M. Hadef, A. Bourouina, and M. R. Mekideche, "Parameter identification of a DC motor using moments method," *International Journal of Electrical and Power Engineering*, vol. 1, no. 2, pp. 210–214, 2007.

[41] J. O. Jang, H. T. Chung, and I. S. Lee, "Backlash compensation of discrete time systems using fuzzy logic," in *Proceedings of the 40th IEEE Conference on Decision and Control*, vol. 4, pp. 3956–3961, Orlando, Fla, USA, December 2001.

Review Article

Robust State-Derivative Feedback LMI-Based Designs for Linear Descriptor Systems

Flávio A. Faria, Edvaldo Assunção, Marcelo C. M. Teixeira, and Rodrigo Cardim

Department of Electrical Engineering, Faculdade de Engenharia de Ilha Solteira,
São Paulo State University (UNESP), 15385-000 Ilha Solteira, SP, Brazil

Correspondence should be addressed to Flávio A. Faria, flaviof15@yahoo.com.br

Techniques for stabilization of linear descriptor systems by state-derivative feedback are proposed. The methods are based on Linear Matrix Inequalities (LMIs) and assume that the plant is a controllable system with poles different from zero. They can include design constraints such as: decay rate, bounds on output peak and bounds on the state-derivative feedback matrix K, and can be applied in a class of uncertain systems subject to structural failures. These designs consider a broader class of plants than the related results available in the literature. The LMI can be efficiently solved using convex programming techniques. Numerical examples illustrate the efficiency of the proposed methods.

1. Introduction

The Linear Matrix Inequalities (LMIs) formulation has emerged recently as a useful tool for solving a great number of practical control problems [1–10]. Furthermore, LMI can be solved with polynomial convergence time, by convex optimization algorithms [1, 11–13].

Recently, LMI has been used for the study of descriptor systems [14–17]. Descriptor systems can be found in various applications, for instance, in electrical systems, or in robotics [18]. The proportional and derivative feedback ($u = Lx(t) - K\dot{x}(t)$, where $x(t)$ is the plant state vector) has been studied by many authors to design controllers in the following problems: stabilization and regularizability of linear descriptor systems [19, 20], feedback control of singular systems [21], nonlinear control with exact feedback linearization [22], $\mathcal{L}_\infty$-control of continuous-time systems with state delay [23], and design of PD observers [24]. In [18, 25] some properties of this type of feedback and its applications to pole placement were presented.

There exist few researches using only derivative feedback ($u = -K\dot{x}(t)$). In some practical problems the state-derivative signals are easier to obtain than the state signals, for instance, in the following applications: suppression of vibration in mechanical systems [26], control of car wheel suspension systems [27], vibration control of bridge cables [28], and vibration control of landing gear components [29]. The main sensors used in these problems are accelerometers. In this case, from the signals of the accelerometers it is possible to reconstruct the velocities with a good precision but not the displacements [26]. Defining the velocities and displacement as the state variables, then one has available for feedback only the state-derivative signals. Procedures for solving the pole-placement problem for linear systems using state-derivative feedback were proposed in [26, 30, 31]. In [28, 32] a Linear Quadratic Regulator (LQR) controller design scheme for standard state space systems was presented. The results were obtained in Reciprocal State Space (*RSS*) framework. Robust state-derivative feedback LMI-based designs for linear time-invariant systems were recently proposed in [33, 34]. These results considered only standard linear systems, and they can be applied to uncertain systems, with or without, structural failures.

Structural failures appear naturally in systems, for physical wear of the equipment, or for short circuit of electronic components. Recent researches on structural failures (or faults), have been presented in LMI framework [35–38].

In this paper, we will show that it is possible to extend the presented results in [33], for applications in a class of descriptor systems, subject to structural failures in the plant. The procedure can include some specifications: decay rate, bounds on output peak and bounds on the state-derivative feedback matrix K, which can make easier the practical implementation of the controllers. These methods allow new specifications, and also to consider a broader class of plants that the related results are available in the literature [19, 25, 31, 39]. Two examples illustrate the efficiency of the proposed method.

2. Statement of the Problem

Consider a controllable linear descriptor system described by

$$E\dot{x}(t) = Ax(t) + Bu(t), \tag{2.1}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $E \in \mathbb{R}^{n \times n}$, $A \in \mathbb{R}^{n \times n}$, and $B \in \mathbb{R}^{n \times m}$. It is known that the stability problem for descriptor systems is more complicated than for standard systems, because it requires considering not only stability, but also regularity [15, 25]. In the next sections, LMI conditions for asymptotic stability of descriptor system (2.1) using state-derivative feedback, are proposed. The problem is defined as follows.

Problem 1. Find a constant matrix $K \in \mathbb{R}^{m \times n}$, such that the following conditions hold:

(1) $(E + BK)$ has a full rank;

(2) the closed-loop system (2.1) with the state-derivative feedback control

$$u(t) = -K\dot{x}(t), \tag{2.2}$$

is regular and asymptotically stable (in this work, a descriptor system is regular if it has uniqueness in the solutions and avoid impulsive responses).

Remark 2.1. In [25, 39] the authors assure that $(E + BK)$ has a full rank (nonsingular matrix) only if the following equation holds:

$$\text{rank}[E, B] = n. \tag{2.3}$$

Unfortunately, there exist several practical problems that not satisfy (2.3). In that way, the input control (2.2) can only be applied in descriptor systems (2.1), when (2.3) holds. Some authors have been using the state-derivative and state feedback $(u = Lx(t) - K\dot{x}(t))$ to solve (2.1), when (2.3) does not hold [18, 20]. However, usually these designs are more complex than the design procedures with only state or state-derivative feedback.

Assuming that $(E + BK)$ has a full rank, then from (2.2) it follows that (2.1) can be rewrite such as a standard linear system, given by

$$E\dot{x}(t) = Ax(t) - BK\dot{x}(t) \Longleftrightarrow \dot{x}(t) = (E + BK)^{-1}Ax(t). \tag{2.4}$$

3. LMI-Based Stability Conditions for State-Derivative Feedback

Necessary and sufficient conditions for asymptotic stability of standard linear system (2.4) are proposed in the next theorems.

Theorem 3.1. *Assuming that* (2.3) *holds, the necessary and sufficient condition for the solution of Problem 1 is the existence of matrices* $Q = Q'$, $Q \in \mathbb{R}^{n \times n}$ *and* $Y \in \mathbb{R}^{m \times n}$, *such that,*

$$AQE' + EQA' + BYA' + AY'B' < 0, \tag{3.1}$$

$$Q > 0. \tag{3.2}$$

Furthermore, when (3.1) *and* (3.2) *hold, then a state-derivative feedback matrix that solves Problem 1 can be given by*

$$K = YQ^{-1}. \tag{3.3}$$

Proof. Observe that for any nonsymmetric matrix M $(M \neq M')$, $M \in \mathbb{R}^{n \times n}$, if $M + M' < 0$, then M has a full rank. Now, defining $Q = P^{-1}$ and $Y = KQ$, the following equations are equivalents:

$$AQE' + EQA' + BYA' + AY'B' = AQ(E + BK)' + (E + BK)QA' < 0 \tag{3.4}$$

$$\Longleftrightarrow P(E + BK)^{-1}A + A'\left[(E + BK)^{-1}\right]'P < 0, \tag{3.5}$$

From (3.4) one has the matrix $(E + BK)QA'$ has full rank, and so, $(E + BK)$ also has a full rank, as required in Problem 1, and (3.5) was obtained after premultiplying by $P(E + BK)^{-1}$ and posmultiplying by $\left[(E + BK)^{-1}\right]'P$ in both sides of (3.4).

System (2.4) is globally asymptotically stable only if there exists $P = P' > 0$ (that is equivalent to $Q = Q' = P^{-1} > 0$) such that (3.4) or (3.5) holds. $\qquad\square$

Remark 3.2. Note that from (3.4) it follows that matrix A must have a full rank, and so, all its eigenvalues are different from zero. This condition was also considered in other papers [26, 28, 33] for linear systems.

Equations (3.1) and (3.2) are LMI. When (3.1) and (3.2) are feasible, they can be easily solved using available software, such as LMISol [40], that is a free software, or MATLAB [11]. The algorithms have polynomial time convergence.

Usually, only the stability of the control systems is insufficient to obtain a suitable performance. In the design of control systems, the specification of the decay rate can also be very useful.

3.1. Decay Rate in State-Derivative Feedback

Consider, for instance, the controlled system (2.4). According to [1], the decay rate is defined as the largest real constant γ, $\gamma > 0$, such that,

$$\lim_{t \to \infty} e^{\gamma t} \|x(t)\| = 0 \tag{3.6}$$

holds, for all trajectories $x(t)$, $t \geq 0$.

Theorem 3.3. *Assuming that* (2.3) *holds, the closed-loop system given by* (2.4)*, in Problem 1, has decay rate greater or equal to γ if there exist matrices $Q = Q'$ and Y, where $Q \in \mathbb{R}^{n \times n}$ and $Y \in \mathbb{R}^{m \times n}$, such that:*

$$\begin{bmatrix} AQE' + EQA' + BYA' + AY'B' & EQ + BY \\ QE' + Y'B' & -\dfrac{Q}{2\gamma} \end{bmatrix} < 0, \tag{3.7}$$

$$Q > 0. \tag{3.8}$$

Furthermore, when (3.7) *and* (3.8) *hold, then a state-derivative feedback matrix can be given by:*

$$K = YQ^{-1}. \tag{3.9}$$

Proof. Stability corresponds to positive decay rate, $\gamma > 0$. One can use the quadratic Lyapunov function $V(x(t)) = x'(t)Px(t)$ to impose a lower bound on the decay rate with $\dot{V}(x(t)) < -2\gamma V(x(t))$, as described in [1]. Note that, from (2.4),

$$\begin{aligned} \dot{V}(x(t)) &= \dot{x}'(t)Px(t) + x'(t)P\dot{x}(t) \\ &= x'(t)A'\left[(E + BK)^{-1}\right]' Px(t) + x'(t)P(E + BK)^{-1}Ax(t). \end{aligned} \tag{3.10}$$

Then, from $\dot{V}(x(t)) < -2\gamma V(x(t))$ it follows that,

$$x'(t)A'\left[(E + BK)^{-1}\right]' Px(t) + x'(t)P(E + BK)^{-1}Ax(t) < -2\gamma x'(t)Px(t), \tag{3.11}$$

or

$$A'\left(P^{-1}E' + P^{-1}K'B'\right)^{-1} + \left(EP^{-1} + BKP^{-1}\right)^{-1}A < -2\gamma P. \tag{3.12}$$

After premultiplying by $(EP^{-1} + BKP^{-1})$ and posmultiplying by $(P^{-1}E' + P^{-1}K'B')$ in both sides of (3.12), observe that (3.12) holds if and only if

$$\begin{aligned}
\left(EP^{-1} + BKP^{-1}\right)A' &+ A\left(P^{-1}E' + P^{-1}K'B'\right) \\
&< \left(EP^{-1} + BKP^{-1}\right)(-2\gamma P)\left(P^{-1}E' + P^{-1}K'B'\right)
\end{aligned} \tag{3.13}$$

and so

$$\begin{aligned}
-\left(EP^{-1} + BKP^{-1}\right)A' &- A\left(P^{-1}E' + P^{-1}K'B'\right) \\
&- (-1)\left(EP^{-1} + BKP^{-1}\right)(2\gamma P)(-1)\left(P^{-1}E' + P^{-1}K'B'\right) > 0.
\end{aligned} \tag{3.14}$$

Now, using the Schur complement [1], the equation above is equivalent to

$$\begin{bmatrix} -\left(EP^{-1} + BKP^{-1}\right)A' - A\left(P^{-1}E' + P^{-1}K'B'\right) & -\left(EP^{-1} + BKP^{-1}\right) \\ -\left(P^{-1}E' + P^{-1}K'B'\right) & \dfrac{P^{-1}}{(2\gamma)} \end{bmatrix} > 0. \tag{3.15}$$

Therefore, defining $Q = P^{-1}$ and $Y = KP^{-1}$, then it follows the expression (3.7). If $P > 0$ then $Q > 0$, as specified in (3.8). So, when (3.7) and (3.8) hold, a state-derivative feedback matrix K is given by (3.9). □

The next section shows that it is possible to extend the presented results, for the case where there exist polytopic uncertainties or structural failures in the plant. A fault-tolerant design is proposed.

4. Robust Stability Condition for State-Derivative Feedback

In this work, structural failure is defined as a permanent interruption of the system's ability to perform a required function under specified operating conditions [41]. Systems subject to structural failures can be described by uncertain polytopic systems.

Consider the linear time-invariant uncertain polytopic descriptor system, with or without structural failures, described as convex combinations of the polytope vertices:

$$\sum_{i=1}^{r_e} e_i E_i \dot{x}(t) = \sum_{j=1}^{r_a} a_j A_j x(t) + \sum_{k=1}^{r_b} b_k B_k u(t), \tag{4.1}$$

$$e_i \geq 0, \quad i = 1, \ldots, r_e, \quad \sum_{i=1}^{r_e} e_i = 1,$$

$$a_j \geq 0, \quad j = 1, \ldots, r_a, \quad \sum_{i=1}^{r_a} a_j = 1, \tag{4.2}$$

$$b_k \geq 0, \quad k = 1, \ldots, r_b, \quad \sum_{j=1}^{r_b} b_k = 1,$$

where r_e, r_a, and r_b are the numbers of polytope vertices of E, A, and B, respectively. In (4.2), e_i, a_j, and b_k, are constant and unknown real numbers for all index i, j, k. The next theorem solves Problem 1, replacing system (2.1) by the uncertain system (4.1).

Theorem 4.1. *A sufficient condition for the solution of Problem 1 for the uncertain system* (4.1) *is the existence of matrices $Q = Q'$ and Y, where $Q \in \mathbb{R}^{n \times n}$ and $Y \in \mathbb{R}^{m \times n}$, such that,*

$$A_j Q E_i' + E_i Q A_j' + B_k Y A_j' + A_j Y' B_k' < 0, \tag{4.3}$$

$$Q > 0, \tag{4.4}$$

where $i = 1, 2, \ldots, r_e$, $j = 1, 2, \ldots, r_a$, and $k = 1, 2, \ldots, r_b$. Furthermore, when (4.3) *and* (4.4) *hold, then a state-derivative feedback matrix can be given by,*

$$K = YQ^{-1}. \tag{4.5}$$

Proof. From (4.2) and (4.3) it follows that

$$\sum_{i=1}^{r_e} e_i \sum_{j=1}^{r_a} a_j \sum_{k=1}^{r_b} b_k \left[A_j Q E_i' + E_i Q A_j' + B_k Y A_j' + A_j Y' B_k' \right]$$

$$= \left(\sum_{j=1}^{r_a} a_j A_j \right) Q \left(\sum_{i=1}^{r_e} e_i E_i' \right) + \left(\sum_{i=1}^{r_e} e_i E_i \right) Q \left(\sum_{j=1}^{r_a} a_j A_j' \right) \tag{4.6}$$

$$+ \left(\sum_{k=1}^{r_b} b_k B_k \right) Y \left(\sum_{j=1}^{r_a} a_j A_j' \right) + \left(\sum_{j=1}^{r_a} a_j A_j \right) Y' \left(\sum_{k=1}^{r_b} b_k B_k' \right) < 0.$$

Therefore, condition (3.1) of Theorem 3.1 holds for the uncertain system (4.1), where $E = e_1 E_1 + \cdots + e_{r_e} E_{r_e}$, $A = a_1 A_1 + \cdots + a_{r_a} A_{r_a}$, and $B = b_1 B_1 + \cdots + b_{r_b} B_{r_b}$. Now, conditions (4.4) and (4.5) are equivalent to conditions (3.2) and (3.3). Finally, from Theorem 3.1, the existence of matrices $Q = Q'$ and Y such that (4.3) and (4.4) hold is a sufficient condition for the solution of Problem 1.   □

Theorem 4.2. *A sufficient condition for the decay rate of the robust closed-loop system given by (2.2) and (4.1) to be greater or equal to γ is the existence of matrices $Q = Q'$ and Y, $Q \in \mathbb{R}^{n \times n}$, $Y \in \mathbb{R}^{m \times n}$, such that:*

$$
\begin{bmatrix}
A_j Q E_i' + E_i Q A_j' + B_k Y A_j' + A_j Y' B_k' & E_i Q + B_k Y \\
Q E_i' + Y' B_k' & -\dfrac{Q}{2\gamma}
\end{bmatrix} < 0, \quad \forall i, j,
$$

(4.7)

$$
Q > 0.
$$

Furthermore, when (4.7) *hold, then a robust state-derivative feedback matrix can be given by*

$$
K = Y Q^{-1}.
$$

(4.8)

Proof. It follows directly from the proofs of Theorems 3.3 and 4.1. □

Due to limitations imposed in the practical applications of control systems, many times it should be considered output constraints in the design.

5. Bounds on Output Peak

Consider that the output of the system (2.1) is given by

$$
y(t) = C x(t),
$$

(5.1)

where $y(t) \in \mathbb{R}^p$ and $C \in \mathbb{R}^{p \times n}$. Assume that the initial condition of (2.1) and (5.1) is $x(0)$. If the feedback system (2.1), (2.2), and (5.1) is asymptotically stable, one can specify bounds on output peak as described in:

$$
\max \|y(t)\|_2 = \max \sqrt{y'(t) y(t)} < \xi_0,
$$

(5.2)

for $t \geq 0$, where ξ_0 is a known positive constant. From [1], (5.2) is satisfied when the following LMI holds:

$$
\begin{bmatrix}
1 & x(0)' \\
x(0) & Q
\end{bmatrix} > 0,
$$

$$
\begin{bmatrix}
Q & Q C' \\
C Q & \xi_0^2 I
\end{bmatrix} > 0,
$$

(5.3)

and the LMI that guarantees stability (Theorem 3.1 or Theorem 4.1), or stability and decay rate (Theorem 3.3 or Theorem 4.2).

An interesting method for specification of bounds on the state-derivative feedback matrix K was recently proposed in [33]. The result is presented below.

Lemma 5.1. *Given a constant $\mu_0 > 0$, then the specification of bounds on the state-derivative feedback matrix K can be described finding the minimum of β, $\beta > 0$, such that $KK' < \beta I / \mu_0^2$. The optimal value of β can be obtained by the solution of the following optimization problem:*

$$\min \beta$$

$$s.t. \quad \begin{bmatrix} \beta I & Y \\ Y' & I \end{bmatrix} > 0,$$

$$Q > \mu_0 I,$$

$$(Set\ of\ LMI),$$

(5.4)

where the Set of LMI can be equal to (3.1), (3.2) or (3.7), (3.8) or (4.3), (4.4) or (4.7), with or without the LMI (5.3).

Proof. See [33].   □

In the following section, Example 6.1 illustrates the efficiency of this optimization procedure that can reduce the practical difficulties in the implementation of the controllers.

6. Examples

The effectiveness of the proposed LMI designs is demonstrated by simulation results.

Example 6.1. A simple electrical circuit, can be represented by the linear descriptor system below [25]:

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t),$$

(6.1)

where x_1 is the current and the x_2 is the potential of the capacitor.

Suppose the output of the system is given by $y(t) = x_1$. So it is a Single-Input/Single-Output (SISO) system, with $n = 2$, $m = 1$ and $p = 1$. Consider as specification an output peak bound $\xi_0 = 10$ and an initial condition equal to $x(0) = [1 \quad 0]'$. Then, using the package "LMI control toolbox" from MATLAB [11] to solve the LMI (3.1) and (3.2) from Theorem 3.1, and (5.3), one feasible solution was obtained

$$Q = \begin{bmatrix} 59.366 & -16.491 \\ -16.491 & 98.944 \end{bmatrix},$$

$$Y = \begin{bmatrix} -98.944 & -49.472 \end{bmatrix}.$$

(6.2)

A state-derivative feedback matrix was calculated using (3.3)

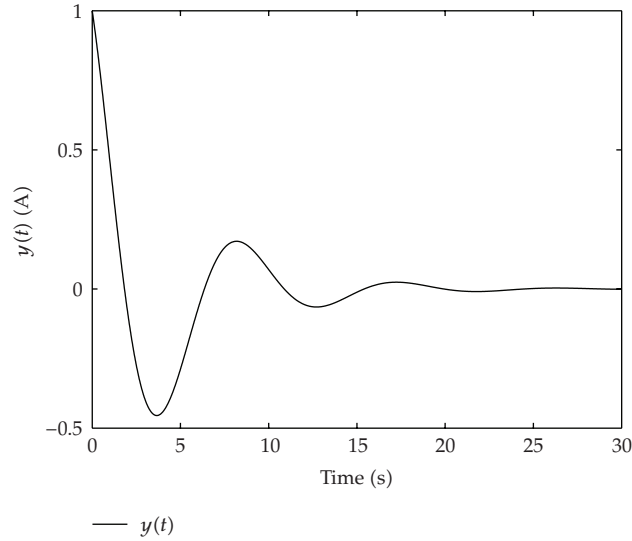$$K = \begin{bmatrix} -1.8932 & -0.81553 \end{bmatrix}.$$

(6.3)

Figure 1: The response of the signal $y(t)$ of the controlled system (2.4).

Note that, as discussed before, the obtained solution K is such that $\det(E + BK) \neq 0$ (it is equal to 1.8932).

For the initial condition $x(0)$ given above, the simulation results of the controlled system are presented in Figure 1. From Figure 1, the settling time of the controlled system is approximately 25 seconds and $\max \sqrt{y'(t)y(t)}$ is equal to $1 < \xi_0 = 10$. The specification for the controlled system was satisfied using the designed controller. Note by Figure 1 that only the stability of the controlled system can be insufficient to obtain a suitable performance. Specifying a lower bound for the decay rate equal $\gamma = 2$, to obtain a faster transient response and using the LMI (3.7) and (3.8) from Theorem 3.3, and (5.3) from Section 5, one feasible solution was obtained

$$
Q = \begin{bmatrix} 90.071 & -22.22 \\ -22.22 & 10.662 \end{bmatrix},
$$

$$
Y = \begin{bmatrix} 5.4955 & -3.8158 \end{bmatrix}.
$$

(6.4)

A state-derivative feedback matrix was calculated using (3.9)

$$
K = \begin{bmatrix} -0.056149 & -0.47492 \end{bmatrix}.
$$

(6.5)

For the solution (6.5) one has $\det(I + BK) = 0.056149$, and the simulation result of the controlled system for the same initial condition $x(0)$, is presented in Figure 2. Note that in Figure 2, the settling time was approximately equal to 1 second and $\max \sqrt{y'(t)y(t)}$ is equal to $1 < \xi_0 = 10$. Then, the specifications were satisfied by using the designed controller.
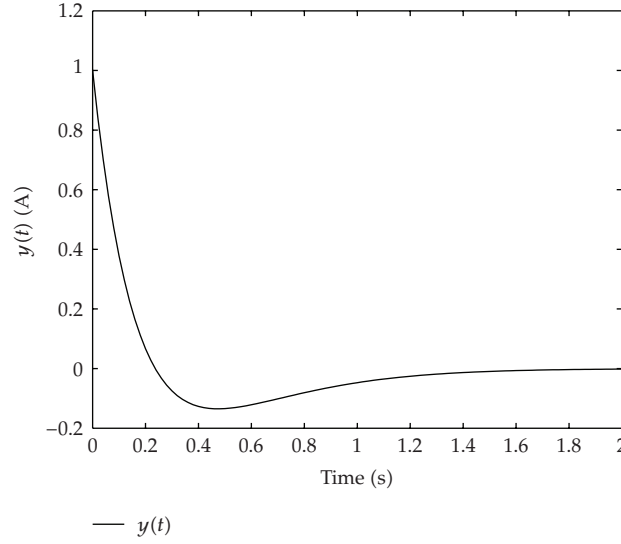
Figure 2: The response of the signal $y(t)$ of the controlled system (2.4), with bound on the decay rate.

Table 1

Stability	Stability with decay rate ($\gamma = 2$)
$Q = \begin{bmatrix} 99.846 & -5.7193 \\ -5.7193 & 1.3314 \end{bmatrix}$,	$Q = \begin{bmatrix} 99.9 & -5.8266 \\ -5.8266 & 1.3436 \end{bmatrix}$,
$Y = \begin{bmatrix} -0.0088332 & -0.076523 \end{bmatrix}$,	$Y = \begin{bmatrix} -0.0096954 & -0.091559 \end{bmatrix}$,
$K = \begin{bmatrix} -0.004484 & -0.076735 \end{bmatrix}$,	$K = \begin{bmatrix} -0.0054497 & -0.091774 \end{bmatrix}$,
$\beta = 0.005934$.	$\beta = 0.0084843$.

To facilitate the implementation of the controller, the specification of bounds on the state-derivative feedback matrix K can be done using the optimization procedure stated in Lemma 5.1, with $\mu_0 = 1$. The optimal values, obtained with the "LMI control toolbox" are given in Table 1.

Note that the absolute values of the entries of K are smaller than the obtained without optimization method, given in (6.3) and (6.5), respectively.

This procedure can also be applied to the control design of uncertain systems subject to failures.

Example 6.2. Consider the linear uncertain descriptor system represented by matrices:

$$E = \begin{bmatrix} 0 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & e_{33} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad A = \begin{bmatrix} a_{11} & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ -1 & 0 & 3 & 0 \\ 0 & 1 & 0 & -2 \end{bmatrix}, \tag{6.6}$$

where $0.8 \leq e_{33} \leq 1.2$ and $5.4 \leq a_{11} \leq 6.4$.

A fail in the actuator is described by:

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & b_{32} \\ 0 & 1 \end{bmatrix}, \tag{6.7}$$

where $b_{32} = 1$ without fail, or $b_{32} = 0$ with fail of the actuator. Then, the vertices of the polytope are given by triple: $(E_i, A_j, B_k) = \{(E_1, A_1, B_1), (E_1, A_1, B_2), (E_1, A_2, B_1), (E_1, A_2, B_2), (E_2, A_1, B_1), (E_2, A_1, B_2), (E_2, A_2, B_1), (E_2, A_2, B_2)\}$, where

$$E_1 = \begin{bmatrix} 0 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad E_2 = \begin{bmatrix} 0 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 1.2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} 5.4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ -1 & 0 & 3 & 0 \\ 0 & 1 & 0 & -2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 6.4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ -1 & 0 & 3 & 0 \\ 0 & 1 & 0 & -2 \end{bmatrix}, \tag{6.8}$$

$$B_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

And the example was solved considering stability with decay rate. It was specified a lower bound for the decay rate equal to $\gamma = 2$, an output peak bound $\xi_0 = 10$, and an initial condition $x(0) = [0.3 \quad 0.1 \quad 0 \quad 0]'$. Using LMI control toolbox for solving the set of LMI (4.7) from Theorem 4.2 with (5.3), a feasible solution was the following:

$$Q = \begin{bmatrix} 21.496 & -1.7143 & -24.031 & 5.9229 \\ -1.7143 & 5.2937 & -1.282 & -20.904 \\ -24.031 & -1.282 & 75.634 & 5.0044 \\ 5.9229 & -20.904 & 5.0044 & 268.7 \end{bmatrix}, \tag{6.9}$$

$$Y = \begin{bmatrix} 43.512 & 3.359 & -135.59 & -7.6619 \\ 2.3436 & -7.8481 & 2.1942 & 18.933 \end{bmatrix}.$$
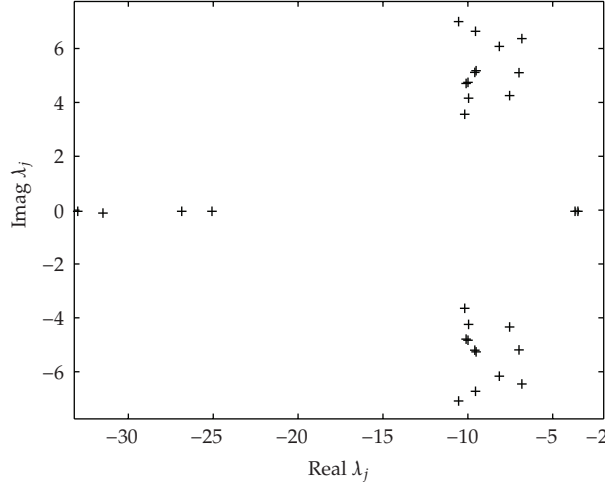
Figure 3: The eigenvalues location of the vertices from robust controlled uncertain system (4.1) and (2.2), subject to failures.

A robust state-derivative feedback matrix is obtained using (4.8)

$$K = \begin{bmatrix} 0.068019 & 0.34647 & -1.7672 & 0.029854 \\ -0.012215 & -1.7426 & -1.1708 \times 10^{-4} & -0.064834 \end{bmatrix}. \tag{6.10}$$

The locations in the s-plane of the eigenvalues, for the vertices (E_i, A_j, B_k), of the robust controlled system, are plotted in Figure 3. There exist eight vertices, and four eigenvalues for each vertice.

Considering that the output system is

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \tag{6.11}$$

the responses of the controlled system with parameter $e_{33} = 0.8$, and $a_{11} = 6.4$ for uncertain matrices E and A respectively, are showed in Figure 4. Note that with (*dotted line*) or without (*solid line*) fail of the actuator the controlled system has fast transient responses.

Now, solving the optimization procedure stated in Lemma 5.1, with LMI (4.7), (5.3), and $\mu_0 = 1$, the optimal values, obtained with the "LMI control toolbox" were the following:

$$Q = \begin{bmatrix} 2.166 & -0.79427 & -1.9412 & 4.435 \\ -0.79427 & 1.7839 & 0.4143 & -10.039 \\ -1.9412 & 0.4143 & 7.6281 & 0.75171 \\ 4.435 & -10.039 & 0.75171 & 7.0245 \times 10^6 \end{bmatrix},$$

$$Y = \begin{bmatrix} 3.4619 & -0.22593 & -12.759 & -1.4344 \\ 0.98266 & -2.662 & -1.5705 & 10.009 \end{bmatrix}, \tag{6.12}$$
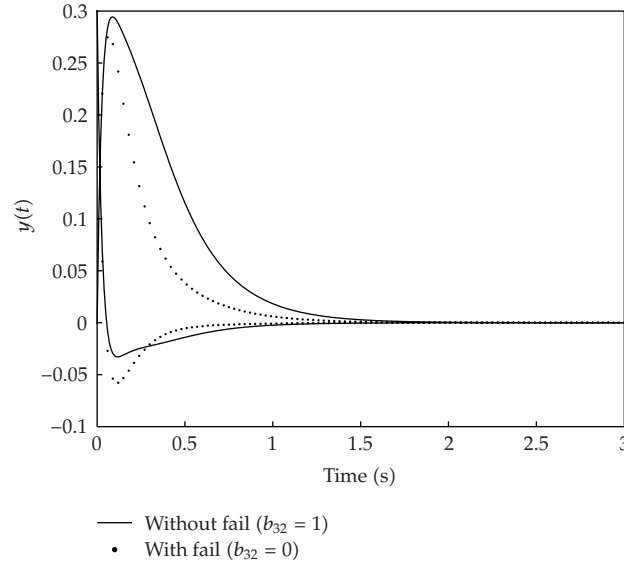
$$\beta = 178.26,$$

Figure 4: The response of the signal $y(t)$ of the controlled system, with and without fail of the actuator.

$$K = \begin{bmatrix} 0.28356 & 0.37604 & -1.6208 & 3.2763 \times 10^{-7} \\ -0.3028 & -1.5813 & -0.19705 & -6.2275 \times 10^{-7} \end{bmatrix}. \tag{6.13}$$

Note that some absolute values of the entries of K in (6.13) are greater than the obtained in first design, given in (6.10). However, the norm of matrix K obtained in first design is $\|K\| = 1.939$ and one obtained from optimization procedure is $\|K\| = 1.7655$. Therefore the optimization procedure was able to control problem with a smaller norm of the state-derivative feedback matrix K.

7. Conclusions

Necessary and sufficient stability conditions based on LMI for state-derivative feedback of linear descriptor systems, were proposed. We can include in the LMI-based control design, the specification of the decay rate, bounds on output peak, and bound on the state-derivative feedback matrix K. The plant can be linear time-invariant SISO or MIMO, and can also have polytopic uncertainties in its parameters or be subject to structural failures. In this case, one obtains a fault-tolerant design. Therefore, the new design methods allow a broader class of plants and performance specifications, than the related results available in the literature, for instance in [19, 25, 39]. The proposed methods are LMI-based designs that, when feasible, can be efficiently solved by convex programming techniques. Theoretical analysis and numerical simulations illustrate these results.

Acknowledgments

References

[1] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, vol. 15 of *Studies in Applied and Numerical Mathematics*, SIAM, Philadelphia, Pa, USa, 2nd edition, 1994.

[2] E. Assunção and P. L. D. Peres, "A global optimization approach for the $\mathcal{H}_2$-norm model reduction problem," in *Proceedings of the 38th IEEE Conference on Decision and Control (CDC '99)*, vol. 2, pp. 1857–1862, Phoenix, Ariz, USA, 1999.

[3] D. D. Šiljak and D. M. Stipanović, "Robust stabilization of nonlinear systems: the LMI approach," *Mathematical Problems in Engineering*, vol. 6, no. 5, pp. 461–493, 2000.

[4] M. C. M. Teixeira, E. Assunção, and R. G. Avellar, "On relaxed LMI-based designs for fuzzy regulators and fuzzy observers," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 5, pp. 613–623, 2003.

[5] R. M. Palhares, M. B. Hell, L. M. Durães, et al., "Robust $\mathcal{H}_\infty$ filtering for a class of state-delayed nonlinear systems in an LMI setting," *International Journal Of Computer Research*, vol. 12, no. 1, pp. 115–122, 2003.

[6] M. C. M. Teixeira, E. Assunção, and R. M. Palhares, "Discussion on: $\mathcal{H}_\infty$ output feedback control design for uncertain fuzzy systems with multiple time scales: an LMI approach," *European Journal of Control*, vol. 11, no. 2, pp. 167–169, 2005.

[7] E. Assunção, C. Q. Andrea, and M. C. M. Teixeira, "$\mathcal{H}_2$ and $\mathcal{H}_\infty$-optimal control for the tracking problem with zero variation," *IET Control Theory Applications*, vol. 1, no. 3, pp. 682–688, 2007.

[8] E. Assunção, H. F. Marchesi, M. C. M. Teixeira, and P. L. D. Peres, "Global optimization for the $\mathcal{H}_\infty$-norm model reduction problem," *International Journal of Systems Science*, vol. 38, no. 2, pp. 125–138, 2007.

[9] M. C. M. Teixeira, M. R. Covacic, and E. Assunção, "Design of SPR systems with dynamic compensators and output variable structure control," in *Proceedings of the International Workshop on Variable Structure Systems (VSS '06)*, pp. 328–333, Alghero, Italy, 2006.

[10] R. Cardim, M. C. M. Teixeira, E. Assunção, and M. R. Covacic, "Variable-structure control design of switched systems with an application to a DC-DC power converter," *IEEE Transactions on Industrial Electronics*, vol. 56, no. 9, pp. 3505–3513, 2009.

[11] P. Gahinet, A. Nemirovski, A. J. Laub, and M. Chilali, *LMI Control Toolbox for Use with Matlab*, The Math Works, Natick, Mass, USA, 1995.

[12] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11, no. 1–4, pp. 625–653, 1999.

[13] L. El Ghaoui and S. Niculescu, *Advances in Linear Matrix Inequality Methods in Control*, vol. 2 of *SIAM Advances in Design and Control*, SIAM, Philadelphia, Pa, USA, 2000.

[14] T. Taniguchi, K. Tanaka, and H. O. Wang, "Fuzzy descriptor systems and nonlinear model following control," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 4, pp. 442–452, 2000.

[15] S. Xu and J. Lam, "Robust stability and stabilization of discrete singular systems: an equivalent characterization," *IEEE Transactions on Automatic Control*, vol. 49, no. 4, pp. 568–574, 2004.

[16] M. Chaabane, O. Bachelier, M. Souissi, and D. Mehdi, "Stability and stabilization of continuous descriptor systems: an LMI approach," *Mathematical Problems in Engineering*, vol. 2006, Article ID 39367, 15 pages, 2006.

[17] K. Tanaka, H. Ohtake, and H. O. Wang, "A descriptor system approach to fuzzy control system design via fuzzy Lyapunov functions," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 3, pp. 333–341, 2007.

[18] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols, "Feedback design for regularizing descriptor systems," *Linear Algebra and Its Applications*, vol. 299, no. 1–3, pp. 119–151, 1999.

[19] G.-R. Duan and X. Zhang, "Regularizability of linear descriptor systems via output plus partial state derivative feedback," *Asian Journal of Control*, vol. 5, no. 3, pp. 334–340, 2003.

[20] Y.-C. Kuo, W.-W. Lin, and S.-F. Xu, "Regularization of linear discrete-time periodic descriptor systems by derivative and proportional state feedback," *SIAM Journal on Matrix Analysis and Applications*, vol. 25, no. 4, pp. 1046–1073, 2004.

[21] H. Jing, "Eigenstructure assignment by proportional-derivative state feedback in singular systems," *Systems & Control Letters*, vol. 22, no. 1, pp. 47–52, 1994.

[22] T. K. Boukas and T. G. Habetler, "High-performance induction motor speed control using exact feedback linearization with state and state derivative feedback," *IEEE Transactions on Power Electronics*, vol. 19, no. 4, pp. 1022–1028, 2004.

[23] E. Fridman and U. Shaked, "$\mathcal{H}_\infty$-control of linear state-delay descriptor systems: an LMI approach," *Linear Algebra and Its Applications*, vol. 351, no. 1, pp. 271–302, 2002.

[24] A.-G. Wu and G.-R. Duan, "Design of PD observers in descriptor linear systems," *International Journal of Control, Automation and Systems*, vol. 5, no. 1, pp. 93–98, 2007.

[25] A. Bunse-Gerstner, V. Mehrmann, and N. K. Nichols, "Regularization of descriptor systems by derivative and proportional state feedback," *SIAM Journal on Matrix Analysis and Applications*, vol. 13, no. 1, pp. 46–67, 1992.

[26] T. H. S. Abdelaziz and M. Valášek, "Pole-placement for SISO linear systems by state-derivative feedback," *IEE Proceedings—Control Theory and Applications*, vol. 151, no. 4, pp. 377–385, 2004.

[27] E. Reithmeier and G. Leitmann, "Robust vibration control of dynamical systems based on the derivative of the state," *Archive of Applied Mechanics*, vol. 72, no. 11-12, pp. 856–864, 2003.

[28] Y. F. Duan, Y. Q. Ni, and J. M. Ko, "State-derivative feedback control of cable vibration using semiactive magnetorheological dampers," *Computer-Aided Civil and Infrastructure Engineering*, vol. 20, no. 6, pp. 431–449, 2005.

[29] S. K. Kwak, G. Washington, and R. K. Yedavalli, "Acceleration feedbackbased active and passive vibration control of landing gear components," *Journal of Aerospace Engineering*, vol. 15, no. 1, pp. 1–9, 2002.

[30] T. H. S. Abdelaziz and M. Valášek, "Direct algorithm for pole placement by state-derivative feedback for multi-input linear systems—nonsingular case," *Kybernetika*, vol. 41, no. 5, pp. 637–660, 2005.

[31] R. Cardim, M. C. M. Teixeira, E. Assunção, and F. A. Faria, "Control designs for linear systems using state-derivative feedback," in *Systems, Structure and Control*, pp. 1–28, In-Tech, Vienna, Austria, 2008.

[32] S.-K. Kwak, G. Washington, and R. K. Yedavalli, "Acceleration-based vibration control of distributed parameter systems using the "reciprocal state-space framework"," *Journal of Sound and Vibration*, vol. 251, no. 3, pp. 543–557, 2002.

[33] E. Assunção, M. C. M. Teixeira, F. A. Faria, N. A. P. da Silva, and R. Cardim, "Robust state-derivative feedback LMI-based designs for multivariable linear systems," *International Journal of Control*, vol. 80, no. 8, pp. 1260–1270, 2007.

[34] F. A. Faria, E. Assunção, M. C. M. Teixeira, R. Cardim, and N. A. P. da Silva, "Robust state-derivative pole placement LMI-based designs for linear systems," *International Journal of Control*, vol. 82, no. 1, pp. 1–12, 2009.

[35] M. Zhong, S. X. Ding, J. Lam, and H. Wang, "An LMI approach to design robust fault detection filter for uncertain LTI systems," *Automatica*, vol. 39, no. 3, pp. 543–550, 2003.

[36] J. Liu, J. L. Wang, and G.-H. Yang, "An LMI approach to minimum sensitivity analysis with application to fault detection," *Automatica*, vol. 41, no. 11, pp. 1995–2004, 2005.

[37] D. Ye and G.-H. Yang, "Adaptive fault-tolerant tracking control against actuator faults with application to flight control," *IEEE Transactions on Control Systems Technology*, vol. 14, no. 6, pp. 1088–1096, 2006.

[38] S. S. Yang and J. Chen, "Sensor faults compensation for MIMO faulttolerant control systems," *Transactions of the Institute of Measurement and Control*, vol. 28, no. 2, pp. 187–205, 2006.

[39] G. R. Duan, G. W. Irwin, and G. P. Liu, "Robust stabilization of descriptor linear systems via proportional-plus-derivative state feedback," in *Proceedings of the American Control Conference (ACC '99)*, vol. 2, pp. 1304–1308, San Diego, Calif, USA, 1999.

[40] M. C. de Oliveira, D. P. Farias, and J. C. Geromel, "LMISol, User's guide," UNICAMP, Campinas-SP, Brazil, 1997, http://www.dt.fee.unicamp.br/~mauricio/software.html.

[41] R. Isermann and P. Ballé, "Trends in the application of model-based fault detection and diagnosis of technical processes," *Control Engineering Practice*, vol. 5, no. 5, pp. 709–719, 1997.

Research Article

Applied Model-Based Analysis and Synthesis for the Dynamics, Guidance, and Control of an Autonomous Undersea Vehicle

Kangsoo Kim[1] and Tamaki Ura[2]

[1] *NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation,*
3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan
[2] *Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro, Tokyo 153-8505, Japan*

Correspondence should be addressed to Kangsoo Kim, kskim@idea.brl.ntt.co.jp

Model-based analysis and synthesis applied to the dynamics, guidance, and control of an autonomous undersea vehicle are presented. As the dynamic model for describing vehicle motion mathematically, the equations of motion are derived. The stability derivatives in the equations of motion are determined by a simulation-based technique using computational fluid dynamics analysis. The dynamic model is applied to the design of the low-level control systems, offering model-based synthetic approach in dynamics and control applications. As an intelligent navigational strategy for undersea vehicles, we present the optimal guidance in environmental disturbances. The optimal guidance aims at the minimum-time transit of a vehicle in an environmental flow disturbance. In this paper, a newly developed algorithm for obtaining the numerical solution of the optimal guidance law is presented. The algorithm is a globally working procedure deriving the optimal guidance in any deterministic environmental disturbance. As a fail-safe tactic in achieving the optimal navigation in environments of moderate uncertainty, we propose the quasi-optimal guidance. Performances of the optimal and the quasi-optimal guidances are demonstrated by the simulated navigations in a few environmental disturbances.

1. Introduction

In this article, we present model-based analysis and synthesis applied to the dynamics, guidance, and control of an autonomous undersea vehicle (AUV). The vehicle dynamics is one of the most important concerns in designing and developing an AUV, while the guidance and control are the key issues in achieving the desired vehicle performance. Our approach deals with these individual but closely interrelated issues in a consistent way based on the model-based simulations.

In our research, as the dynamic model of an AUV, we employ a set of equations of motion describing the coupled six-degree-of-freedom (6-DOF) behaviour in three-dimensional (3D) space. In the linearized form of the equations of motion, to complete the dynamic model of an AUV, we have to determine the so-called stability derivatives or hydrodynamic coefficients. There are many well-established approaches for determining the stability derivatives of aerial vehicles [1, 2] or marine vehicles [3], which are based on either experiment or theoretical prediction. While the experimental approach allows direct measurement of the fluid dynamic forces and moments acting on the vehicle, it requires a large amount of time, labour, expense, as well as an experimental facility. On the other hand, a few state-of-the-art techniques are now available for predicting the stability derivatives theoretically [4–6]. Most of them, however, are specialized for deriving the stability derivatives for the dynamics of conventional airplanes [4, 5] or ships [6], making them hard to directly apply to the modelling problems related to the dynamics of a specific AUV. In this respect, we introduce a general-purpose technique for deriving the dynamic model of an undersea vehicle, primarily depending on the computational fluid dynamics (CFD) analysis.

The derived dynamic model is directly applied to the model-based design of the motion control systems of an AUV. Two proportional-integral-derivative (PID) type low-level controllers are employed to make a vehicle follow the desired trajectories in the longitudinal and lateral planes, represented as time sequences of the depth (altitude) and heading.

As an intelligent high-level control of AUVs, a strategy of optimal guidance is presented. The optimal guidance proposed in this research is the minimum-time guidance in sea current environments, allowing a vehicle to reach a destination with the minimum travel time. When the power consumption of an AUV is controlled to be constant throughout the navigation, the navigation time is directly proportional to the total energy consumption. Released from the umbilical cable, an AUV has to rely on restricted energy stores during an undersea mission. Therefore, for an AUV, minimizing navigation time offers an enhanced potential for vehicle safety and mission success rate. We present a newly developed numerical procedure for deriving the optimal heading reference, by tracking which vehicle achieves the minimum-time navigation in a given sea current disturbance. The proposed procedure is systematic and seeks the solution in a global manner in any deterministic current field, whether stationary or time-varying. Moreover, unlike other path-finding algorithms, such as dynamic programming (DP) or generic algorithms (GAs) [7–9], our procedure does not require a computation time increase for the time-varying problems.

In real environments of AUV navigation, there are some factors that can cause failure in realizing the proposed optimal guidance strategy [10, 11]. Some examples are environmental uncertainties, severe sensor noises, or temporally faulty actuators. Though these risk factors significantly affect the realization of optimality in actual sea navigation, they have not been seriously treated in most of related literatures. In this article, we present the concept of quasioptimality as a fail-safe strategy for realizing the proposed optimal navigation.

2. An AUV "R-One"

In this article, we practice our strategy in dynamics, guidance, and control on the AUV "R-One," a long-range cruising type AUV, developed by the Institute of Industrial Science (IIS), the University of Tokyo [12]. Figure 1 shows the overall layout of R-One.
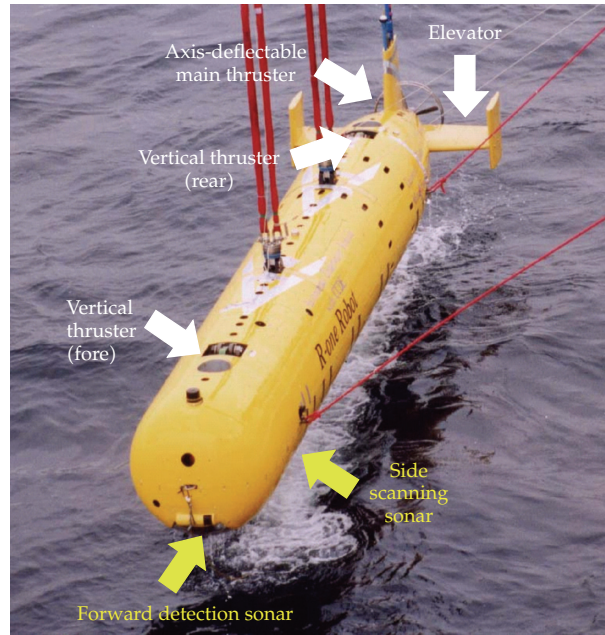
Figure 1: Overall layout of the long-range cruising type AUV R-One.

(a)

(b)
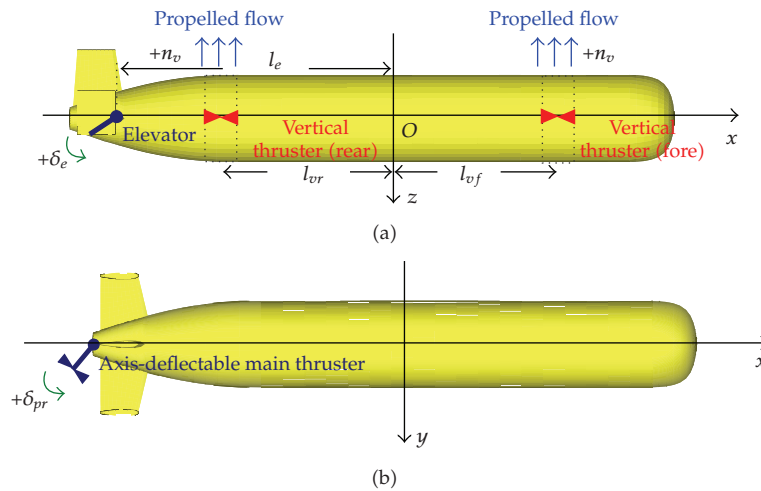
Figure 2: Coordinate system and actuator actions in describing the dynamics of R-One. The coordinate system takes its origin at the center of gravity of the vehicle. The n_v is the rpm of fore and rear vertical thrusters. The δ_e is the elevator deflection. The δ_{pr} is the deflected amount of the main thruster axis.

Figure 2 shows the coordinate system and the actions of the actuators installed in the R-One. The axis-deflectable main thruster keeps or changes the vehicle's kinematic states in the horizontal plane. Two elevators and two vertical thrusters play the same role in the vertical plane.

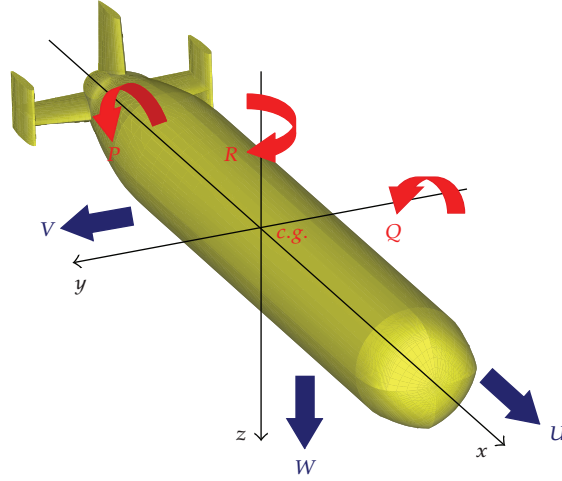**Figure 3:** Body-fixed coordinate system with linear and angular velocity components.

3. Modelling Vehicle Dynamics

3.1. Equations of Motion for Vehicle Dynamics.

The equations of motion describing the vehicle motion mathematically can be derived from the conservation law of the linear and the angular momenta with respect to the inertial frame of [1, 2]. The equations of motion (3.1) describing the 6-DOF motion of an AUV are defined with respect to the body-fixed frame of reference shown in Figure 3, in which the origin is taken at the vehicle's center of gravity. The procedures for deriving the equations of motion of an aerial vehicle which is quite similar to (3.1) are found in [1, 2]. It should be noted here however, that in deriving (3.1) by referring to the equations of motion for the aerial vehicles shown in [1, 2], the hydrostatic loads which do not appear in the flight dynamics have to be additionally involved

$$m(\dot{U} + QW - RV) = -(m - \rho\nabla)g\sin\Theta + X,$$

$$m(\dot{V} + RU - PW) = (m - \rho\nabla)g\cos\Theta\sin\Phi + Y,$$

$$m(\dot{W} + PV - QU) = (m - \rho\nabla)g\cos\Theta\cos\Phi + Z,$$

$$I_{xx}\dot{P} - I_{xz}\dot{R} - I_{xz}PQ + (I_{zz} - I_{yy})QR = \rho\nabla gz_B\cos\Theta\sin\Phi + L,$$

$$I_{yy}\dot{Q} + (I_{zz} - I_{zz})RP + I_{xz}(P^2 - R^2) = \rho\nabla gz_B\sin\Theta + M,$$

$$-I_{xz}\dot{P} + I_{zz}\dot{R} + (I_{yy} - I_{xx})PQ + I_{xz}QR = N.$$

$$(3.1)$$

In (3.1), U, V, W, and P, Q, R are the x, y, z components of linear and angular velocities. ∇, m, and I represent volume, mass, and mass moments or products of inertia of a vehicle, and ρ and g are constants expressing water density and gravitational acceleration. Hydrodynamic forces and moments are represented by X, Y, Z, and L, M, N, each of which is the component in the direction of x, y, z. Φ, Θ and Ψ are so-called Euler angles to be

defined in the coordinate transformation between the body-fixed and the inertial frames of reference. The z_B is the z-directional displacement of the buoyancy center of the vehicle. The equations of motion are frequently linearized for use in stability and control analysis as mentioned in [1–3]. The following equations are the linearized forms of (3.1), in which u, v, w, p, q, r, ϕ, θ, and ψ denote small amounts of velocities, and angular velocities and displacements, perturbed from their reference values, which are expressed by their uppercase letters

$$m(\dot{u} + qW_0) = -\theta(m - \rho\nabla)g\cos\theta_0 + X,$$

$$m(\dot{v} + rU_0 - pW_0) = \phi(m - \rho\nabla)g\cos\theta_0 + Y,$$

$$m(\dot{w} - qU_0) = -\theta(m - \rho\nabla)g\sin\theta_0 + Z,$$

$$I_{xx}\dot{p} - I_{xz}\dot{r} = \phi\rho\nabla gz_B\cos\theta_0 + L,$$

$$I_{yy}\dot{q} = \theta\rho\nabla gz_B\cos\theta_0 + M, \tag{3.2}$$

$$-I_{xz}\dot{p} + I_{zz}\dot{r} = N,$$

$$\dot{\phi} = p + r\tan\theta_0,$$

$$\dot{\theta} = q,$$

$$\dot{\psi} = r\sec\theta_0.$$

In general, to complete the linearized equations of motion for use in stability and control analysis, hydrodynamic loads are expanded and linearized on the assumption that they are functions of the instantaneous values of the perturbed velocities, accelerations, and control inputs. Thus, the expanded expressions of the hydrodynamic loads are obtained in the form of a Taylor series in these variables, which is linearized by discarding all the higher-order terms. For example, X is expanded as

$$X = X_u u + X_w w + X_{\dot{u}}\dot{u} + X_{n_m}n_m, \tag{3.3a}$$

where

$$X_u = \left(\frac{\partial X}{\partial u}\right)_0, \qquad X_w = \left(\frac{\partial X}{\partial w}\right)_0, \dots \tag{3.3b}$$

The subscript zero in (3.3a) indicates a reference condition where the derivatives are evaluated. In (3.3a) and (3.3b), derivatives such as X_u or X_w are called stability derivatives [1–5]. By expanding all the external hydrodynamic loads introducing stability derivatives

of their dynamic correlations, the equations of motion (3.2) are expressed by means of the stability derivatives as

$$(m - X_{\dot{u}})\dot{u} + mW_0\dot{\theta} - X_u u - X_w w + \theta(\rho\nabla - m)g\cos\theta_0 = X_{n_m}n_m,$$

$$(m - Z_{\dot{w}})\dot{w} - Z_{\dot{q}}\dot{q} - Z_u u - Z_w w - (mU_0 + Z_q)q - \theta(\rho\nabla - m)g\sin\theta_0 = Z_{n_v}n_v + Z_{\delta e}\delta_e,$$

$$-M_{\dot{w}}\dot{w} + (I_{yy} - M_{\dot{q}})\dot{q} - M_u u - M_w w - M_q q - \theta\rho\nabla g z_B\cos\theta_0 = -Z_{n_v}l_{vf}n_v + Z_{n_v}l_{vr}n_v + Z_{\delta e}l_e\delta_e,$$

$$(m - Y_{\dot{v}})\dot{v} - Y_{\dot{r}}\dot{r} - Y_v v - (mW_0 + Y_p)p + (mU_0 - Y_r)r = Y_{\delta_{pr}}\delta_{pr},$$

$$-L_{\dot{v}}\dot{v} + (I_{xx} - L_{\dot{p}})\dot{p} - (I_{xz} + L_{\dot{r}})\dot{r} - L_v v - L_p p - L_r r - \phi\rho\nabla g z_B\cos\theta_0 = L_{\delta_{pr}}\delta_{pr},$$

$$-N_{\dot{v}}\dot{v} + (I_{xz} + N_{\dot{p}})\dot{p} + (I_{zz} - N_{\dot{r}})\dot{r} - N_v v - N_p p - N_r r = N_{\delta_{pr}}\delta_{pr},$$

$$\dot{\phi} = p + r\tan\theta_0,$$

$$\dot{\theta} = q,$$

$$\dot{\psi} = r\sec\theta_0,$$

$$(3.3)$$

where n_m represents the rpm of the main thruster.

3.2. Evaluation of Stability Derivatives by CFD Analyses.

As noticeable in (3.3), within the framework of small perturbation theory, constructing the dynamic model is, in effect, reduced to the determination of the stability derivatives defined in the linearized equations of motion. The most commonly and widely employed approaches for evaluating the stability derivatives are the wind tunnel test for aerial vehicles and the towing tank test for marine vehicles [1–6]. These experimental approaches, however, require a huge experimental facility and a large workforce, which makes them expensive and laborious, even when the test is for a single model. In this article, we present a model-based approach for evaluating the stability derivatives. In the approach, dominant stability derivatives are evaluated from the hydrodynamic loads which are obtained by CFD analyses. When we are to evaluate the value of X_u in (3.3) defined at a reference speed of U_0, for example, we conduct CFD analyses repeatedly at the cruising speeds of $U_0(1 \pm \eta)$, where U_0 is the reference cruising speed and η is the perturbation ratio of U_0. By taking central difference approximation of X with respect to u by using the X values obtained at $U_0(1 \pm \eta)$, we can derive X_u defined at U_0. However, while the majority of dominant stability derivatives are able to be evaluated by this technique, there are other stability derivatives which are not. For such stability derivatives, the simplified estimation formulae proposed in the field of flight dynamics [1, 2] are modified and applied.

In our CFD analyses, we used a commercial fluid dynamics solver called "Star-CD," developed by CD-adapco [13]. The Star-CD is a Navier-Stokes solver based on the finite difference numerical scheme. Like other famous commercial CFD solvers such as FLUENT or ANSYS, Star-CD also has shown numerous field application results that it replicates experimental model results with acceptably fine accuracy [14, 15]. The Star-CD derives
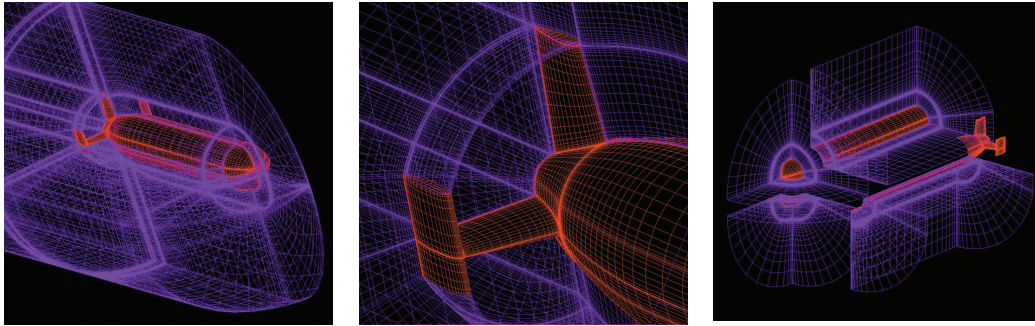
Figure 4: Grid system for CFD analyses of flow field around the R-One. The entire grid system is completed by assembling individually generated subgrid blocks.

the numerical solution by pressure-implicit split-operator (PISO) algorithm, which is a well-known, robust scheme with predictor-corrector steps. In our CFD analyses, the problem-specific high Reynolds number (Re) requires a proper turbulence model. In using Star-CD, we selected the Reynolds-Averaged Navier-Stokes (RANS) turbulence model, which is one of the most widely used turbulence model in engineering applications of moderate turbulent conditions [16].

Figure 4 shows the grid system for evaluating the hydrodynamic loads by CFD analyses. To generate a computationally robust, structured grid system adapting to the complicated aftbody geometry of the vehicle, we employed a grid generation technique called the multiblock method [17].

Not only estimating the drag force of acceptable accuracy is the primary concern in our CFD analyses for deriving the stability derivatives, it also serves as the most fundamental measure to evaluate a CFD solver [14–16]. After completing the hull structure, drag forces acting on the R-One at three cruising speeds were investigated by means of the towing tank tests [12]. In Figure 5, two drag curves, obtained by CFD calculations and tank tests, are shown together. The drag curves shown are quadratic interpolations of the raw data set of drags, calculated and measured at the cruising speeds of 1.03, 1.54, and 2.06 m/s, respectively. Validity of the quadratic interpolation is based on the fact that within the small Re interval, drag of an immersed body has quadratic dependency on its advance speed [18]. In Figure 5, the gradients of two drag curves are also expressed. As seen in the figure, drags obtained by CFD calculations are more or less excessive than the ones by tank tests. However, it is noted that the gradients of drags show close similarity between CFD analyses and tank tests, which advocates our approach to evaluating the stability derivatives principally by means of the CFD analyses.

Figure 6 shows the pressure distribution with a few selected streamlines along the body surface of R-One. By integrating the pressure over the entire body surface, hydrodynamic loads are obtained.

It is generally known and also noticeable from (3.3) that, according to the coupling relation, linearized equations of motion are split into two independent groups: longitudinal equations for surge, heave, and pitch, and lateral equations for sway, roll, and yaw [1–3]. In Tables 1(a) and 1(b), the longitudinal and lateral stability derivatives appearing in (3.3) are summarized. By substituting all stability derivatives in (3.3) with their corresponding numerical values in the tables, the dynamic model of R-One is completed.
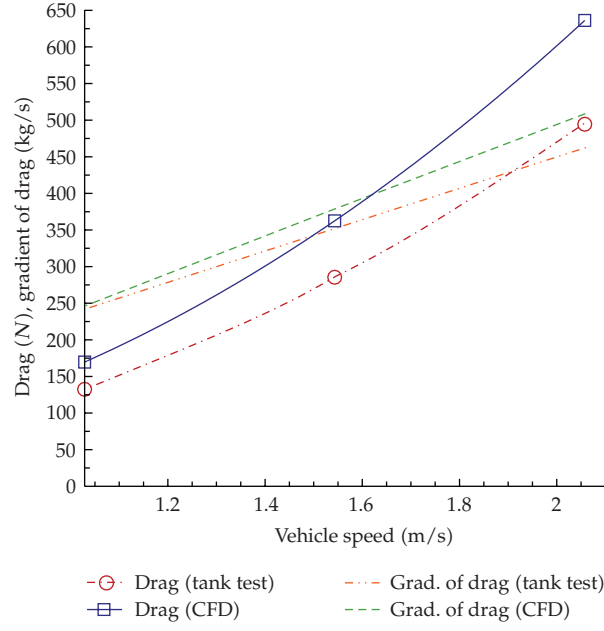
Figure 5: Curves of drags and gradient of drags.

Table 1: (a) Stability derivatives in the longitudinal equations of motion for R-One. (b) Stability derivatives in the lateral equations of motion for R-One.

(a)

$X_{\dot{u}}$	(kg)	−237.65	Z_u	(kg/s)	−64.41
$Z_{\dot{w}}$	(kg)	−2152.10	Z_w	(kg/s)	−2819.64
$Z_{\dot{q}}$	(kg·m)	−25547.00	Z_q	(kg·m/s)	−11360.06
$M_{\dot{u}}$	(kg·m)	0.00	M_u	(kg·m/s)	0.00
$M_{\dot{w}}$	(kg·m)	−1928.80	M_w	(kg·m/s)	870.36
$M_{\dot{q}}$	(kg·m^2)	−153400.00	M_q	(kg·m^2/s)	−39351.25
X_u	(kg/s)	−364.37	Z_{δ_e}	(kg·m/s^2)	−3168.10
X_w	(kg/s)	64.72	M_{δ_e}	(kg·m^2/s^2)	−10974.31
X_q	(kg·m/s)	0.00			

(b)

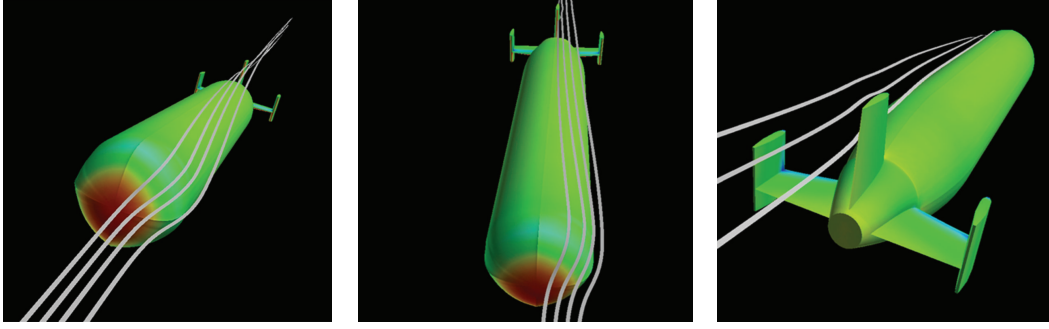|					
$Y_{\dot{v}}$	(kg)	−4653.77	Y_r	(kg·m/s)	3931.06
$Y_{\dot{r}}$	(kg)	746.74	L_v	(kg·m/s)	−515.99
$L_{\dot{v}}$	(kg·m)	−1.31	L_p	(kg·m^2/s)	−1165.25
$L_{\dot{p}}$	(kg·m^2)	3.81	L_r	(kg·m^2/s)	1500.45
$L_{\dot{r}}$	(kg·m^2)	−3.35	N_v	(kg·m/s)	−4054.37
$N_{\dot{v}}$	(kg·m)	−554.21	N_p	(kg·m^2/s)	−1.02
$N_{\dot{p}}$	(kg·m^2)	−1.19	N_r	(kg·m^2/s)	−13704.63
$N_{\dot{r}}$	(kg·m^2)	−15934.27	$Y_{\delta_{\mathrm{pr}}}$	(kg·m/s^2)	−399.04
Y_v	(kg/s)	−1809.97	$L_{\delta_{\mathrm{pr}}}$	(kg·m^2/s^2)	0.00
Y_p	(kg·m/s)	0.00	$N_{\delta_{\mathrm{pr}}}$	(kg·m^2/s^2)	1677.38

Figure 6: Visualized results of a CFD analysis.

3.3. Vehicle Motion Simulation

State-space forms of the longitudinal and the lateral equations of motion for R-One, completed by assigning the numerical values in Table 1 to corresponding stability derivatives in (3.3), are represented as follows.

(i) *Longitudinal equations of motion for R-One:*

$$
\begin{bmatrix} \dot{u} \\ \dot{w} \\ \dot{q} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} -0.0786 & 0.0140 & 0 & 0.0145 \\ -0.0103 & -0.4725 & 0.2465 & 0.0610 \\ 0.0001 & 0.0108 & -0.2420 & -0.0156 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} u \\ w \\ q \\ \theta \end{bmatrix} + \begin{bmatrix} 0.0554 & 0 & 0 \\ 0 & 0.0027 & -0.2169 \\ 0 & -0.0001 & -0.0684 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} n_m \\ n_v \\ \delta_e \end{bmatrix}.
$$

(3.5a)

(ii) *Lateral equations of motion for R-One:*

$$
\begin{bmatrix} \dot{v} \\ \dot{p} \\ \dot{r} \\ \dot{\phi} \end{bmatrix} = \begin{bmatrix} -0.2097 & 0.0053 & -0.5388 & 0.0112 \\ -4.7444 & -11.2192 & 16.1215 & -23.6516 \\ -0.1185 & 0.0643 & -1.1931 & 0.1357 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} v \\ p \\ r \\ \phi \end{bmatrix} + \begin{bmatrix} -0.0388 \\ -0.0948 \\ 0.0634 \\ 0 \end{bmatrix} \delta_{pr}.
$$

(3.5b)

By solving (3.5a) and (3.5b) in the time domain with appropriate initial conditions and actuator inputs, motion responses of the R-One are computed. In the inertial navigation system (INS) installed in R-One, not only vehicle kinematics but also time sequences of the actuator inputs during an undersea mission are recorded. In Figure 7, simulated vehicle trajectories are compared with actual vehicle trajectories recorded during the Teisi knoll survey mission [19]. As noticeable from the figure, the dynamic model of R-One implemented by our model-based approach provides motion responses exhibiting sufficiently good agreement between the simulated and actual vehicle trajectories.
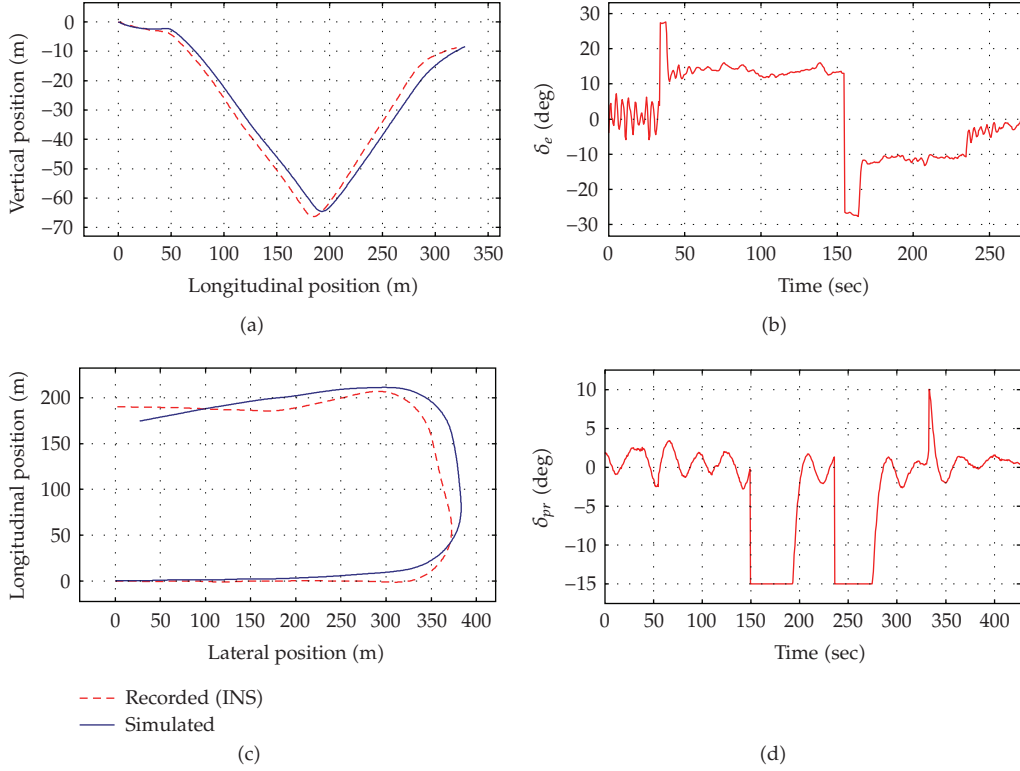
Figure 7: Simulated and actual vehicle trajectories (a and c) generated from the corresponding actuator inputs (b and d).

4. Tracking Control Design

The controller implemented for the motion control of R-One is based on the PID compensation. Needless to say that, PID-type controller is the most commonly and widely used controller for most artificial control systems. However, in designing a PID controller, precise plant dynamics is a key prerequisite to ensuring acceptably good control performance. Deriving a precise plant dynamics is not easy in some cases. For this reason, during the past three decades, a few significant attempts have been made to provide controller models that do not depend on a precise description of the plant model in its design [20–22]. Neural network (NN) controllers based on the self-organizing map or fuzzy logic controllers are the most famous ones in such attempts [21, 22]. In order to derive a practically useful controller by NN or fuzzy logic, however, we have to ensure huge random diversity in training data. This is a very difficult task in a real world problem, because, in general, we do not have any definitive guidelines for deciding whether the prepared training data is biased or not [22, 23].

To change or keep the kinematic states of the vehicle, two independent low-level controls were implemented in the R-One: the depth (altitude) control for the longitudinal motion and the heading control for the lateral motion. Configurations of the depth and the heading controls are depicted in Figures 8(a) and 8(b).

To build the mathematical models for the control systems shown in the Figure 8, transfer functions of $p(s)/\delta_e(s), w(s)/\delta_e(s)$, and $r(s)/\delta_{pr}(s)$ are extracted from (3.5a) and
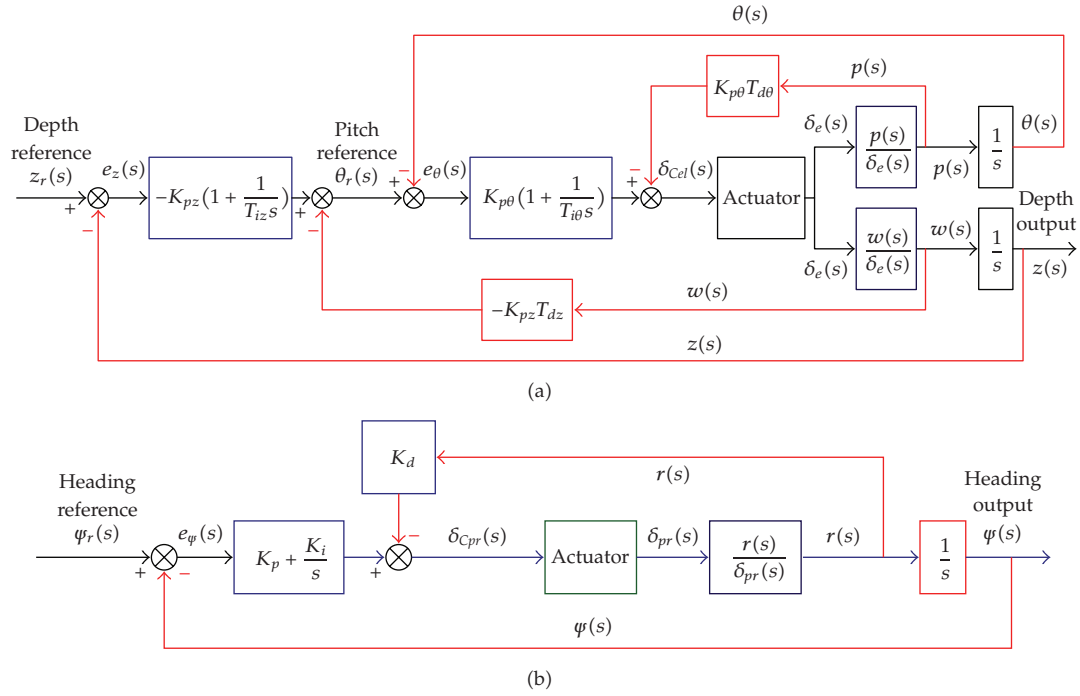
(a)

(b)

Figure 8: (a) Configuration of the depth (altitude) control system for R-One. (b) Configuration of the heading control system for R-One.

(a)

(b)
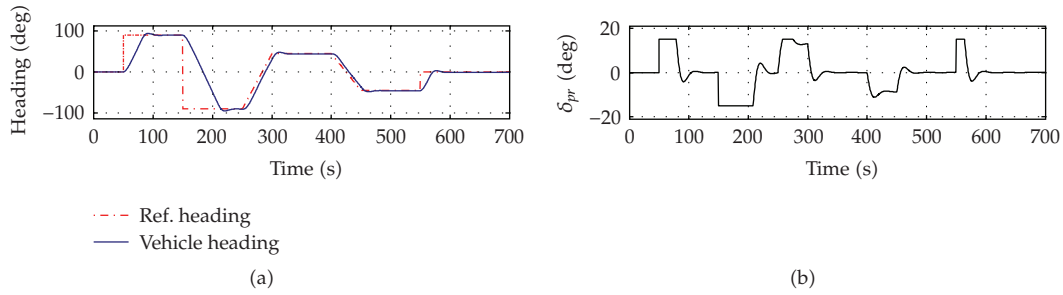
Figure 9: Simulated heading tracking control (a) with corresponding actuator input (b).

(3.5b). Then, PID-tuning is carried out to determine the optimal values of controller gains from the standpoint of system robustness and swiftness of response. In determining the optimal gain values, we used the model-based control system design tool called "SISO Design Tool," offered by the "Control System Toolbox" included in "Matlab" [24].

An example of the performance result for the designed control systems is shown in Figure 9, where it is clearly seen that the designed heading controller lets the vehicle follow the heading reference with sufficient swiftness and small overshoot.

5. Optimal Guidance of AUV

5.1. Background

The sea environment contains several disturbances, such as surface waves, wind, and sea currents. Among them, the sea current is known to be the most significant disturbance for the dynamics of an undersea vehicle as it directly interacts with the vehicle motion [3, 8, 11]. Considering the guidance problem to make a vehicle transit to a given destination in a region of sea current, it is quite natural that there arises a navigation time difference according to the selection of an individual navigation trajectory. The problem of the minimum-time vessel guidance in a region of current flow has interested people as long years ago as ancient Greece [25]. However, since the problem requires a minimization technique of the functionals, it had hardly been treated mathematically until the advent of the calculus of variations. On the basis of this mathematical tool, Bryson and Ho [26] derived the minimum-time guidance law of a surface vessel in a region of a surface current flow. Though the law is an optimal controller of explicit form, obtaining its solution is not easy since it actually is a so-called two-point boundary value problem. As an ad hoc approach for the minimum-time navigation problem in a linearly varying, shear flow-like current distribution presented by Lewis and Syrmos, a graphical solution finding technique has been presented [25]. As is naturally expected, however, such an approach is problem-specific and lacks universality in its applicability. Papadakis and Perakis [8] treated the minimum-time routing problem of a vessel moving in a wave environment. In their approach, by subdividing the navigation region into several subregions of different sea states, the path for the optimal routing is obtained by the DP approach. Aside from the difficulties in constructing a numerical solution procedure for their approach, it has a problem that the solution significantly depends on the features of regional subdivision. As a completely discrete and nonlinear approach, the cell mapping technique was applied to derive the minimum-time tracking trajectory to capture a moving target in a deterministic vortex field [27]. It, however, has the same problem of regional subdivision as is inherent in the approach by Papadakis and Perakis [8], which might lead to the divergence due to numerical instability.

In this research, we propose a newly developed procedure for obtaining the numerical solution of the optimal guidance law, which achieves the minimum-time navigation of a vehicle in a given current field. The algorithm of our solution procedure is simple but consistently applicable to any current field if only the distribution of which is deterministic. As a fault-tolerant strategy for putting the proposed optimal navigation into practice, the concept of quasioptimality is introduced. The basic idea of the quasioptimal navigation is quite simple and, in effect, consists of the on-site feedbacks of the optimal guidance revisions.

5.2. The Optimal Guidance Law

In our optimal guidance problem, we employed the guidance law presented by Bryson and Ho [26] as

$$\dot{\psi} = \sin^2\psi \frac{\partial v_c}{\partial x} + \frac{1}{2}\left(\frac{\partial u_c}{\partial x} - \frac{\partial v_c}{\partial y}\right)\sin 2\psi - \cos^2\psi \frac{\partial u_c}{\partial y}, \tag{5.1}$$

where ψ represents the vehicle heading as defined in Figure 10, and u_c, v_c are x, y components of the sea current velocity. The detailed procedure of deriving (5.1) is well explained in [11].
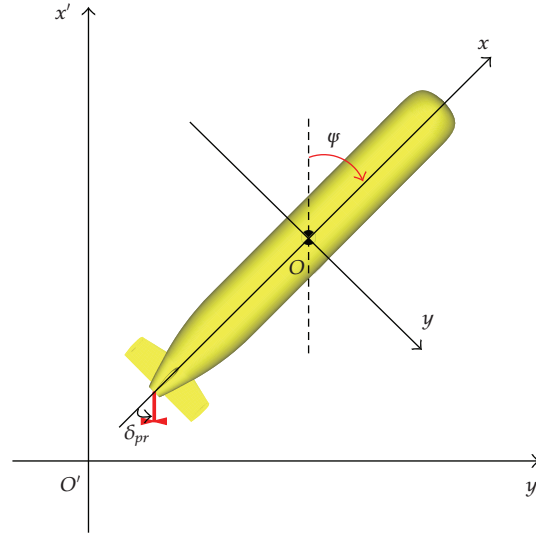
Figure 10: Coordinate system and definition of the heading for describing the optimal guidance problem. $O'\text{-}x'y'$ is the inertial frame of reference, while $O\text{-}xy$ is the body-fixed coordinate system.

Though Bryson and Ho derived (5.1) on the assumption of a stationary flow condition, we have shown that it is also valid for time-varying currents, like tidal flows [11]. Equation (5.1) is a nonlinear ordinary differential equation of an unspecified vehicle heading $\psi(t)$. Though it seems that the solution would be readily obtainable by using a suitable numerical scheme, such as Runge-Kutta, there still remains a significant shortfall: while (5.1) defines an initial value problem, its solution cannot be obtained with an arbitrarily assigned initial heading. If we solve (5.1) with an arbitrary initial value of ψ, a vehicle following the solution of (5.1) as the heading reference does not arrive at the destination. This is because (5.1) is, in fact, a two-point boundary value problem, the correct initial value of which constitutes a part of the solution.

5.3. Numerical Solution Procedure

To obtain the solution of the two-point boundary value problems, an iterative solution procedure is generally used, such as "shooting" or "relaxation" [28]. Starting from an initial guess, solutions generated by these schemes are repeatedly adjusted to eliminate the discrepancies between the estimated and the desired boundary conditions at both endpoints. These schemes, however, strongly rely on the initial guess, inappropriate assignment of which may lead to a local solution or divergence [28]. In this article, we present a numerical procedure to obtain the solution of (5.1), called "AREN," which stands for Arbitrary REference Navigation". In applying AREN, first we need to make a simulated navigation along any feasible trajectory, generated by an arbitrary guidance and terminating at the destination. We call this navigation the "reference navigation," because it is used as the reference in deriving the optimal navigation. In a reference navigation, the time it takes for the vehicle to reach the destination should be recorded, which we call the "reference navigation time," denoted by t_{f_ref}. As already mentioned, the only requirement for the reference navigation is letting the vehicle arrive at the destination. Therefore, a navigation following any trajectory shown in Figure 11 can be qualified as the reference navigation.
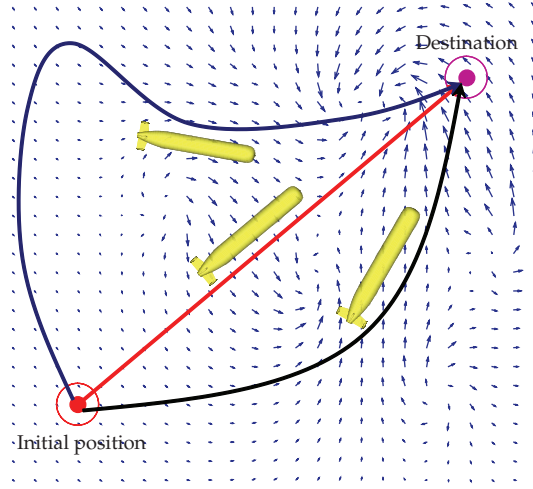
Figure 11: A few possible trajectories for the reference navigation.

In AREN, to search for the correct initial heading numerically, the interval of $0 \sim 2\pi$ is divided by equally spaced $N - 1$ subintervals as

$$\psi_0^{(i)} = i\Delta\psi \quad \text{for } i = 0, 1, \dots, N - 1, \tag{5.2}$$

where $\Delta\psi = 2\pi/N$. In (5.2), $\psi_0^{(i)}$ is the (i)th initial heading trial and $\Delta\psi$ is its increment, that is, the interval of initial heading trials. Next, for an initial heading trial $\psi_0^{(i)}$, we solve (5.1) in the time domain using an appropriate time marching scheme, which produces a simulated navigation starting from $\psi_0^{(i)}$. The navigation produced here is called the (i)th "trial navigation" adjoining $\psi_0^{(i)}$. Once the trajectory produced by a trial navigation passes through vicinity of the destination, it can be considered as a potential optimal navigation since the optimal guidance law (5.1) with the correct initial heading lets a vehicle reach the destination. Therefore, N trial navigations starting from N initial headings given in (5.2) are all possible candidates for the optimal navigation. In practice, however, discretization error in initial heading trials causes convergence error at the destination, so that the optimal navigation should be identified in an approximate manner. We define the "minimum distance," as the shortest distance between the destination and the trajectory generated by trial navigation. In Figure 12, $l_{\min}^{(k-1)}, l_{\min}^{(k)}$, and $l_{\min}^{(k+1)}$ represent the minimum distances between the destination and the trajectories generated by $(k-1)$th, (k)th, and $(k+1)$th trial navigations, respectively. If the minimum distance of the (k)th trial navigation is smaller than any other minimum distance and thus satisfies (5.3), we choose it as the optimal navigation because in the (k)th trial navigation, the vehicle approaches the destination with the smallest deviation among all trial navigations, that is, the optimal navigation candidates.

$$l_{\min}^{(k)} \leq l_{\min}^{(i)}, \quad \text{for } i = 0, 1, \dots, N - 1. \tag{5.3}$$

However, in choosing the optimal navigation among the trial ones, there still remains a serious drawback: we have no idea how long we have to continue a trial navigation so as to
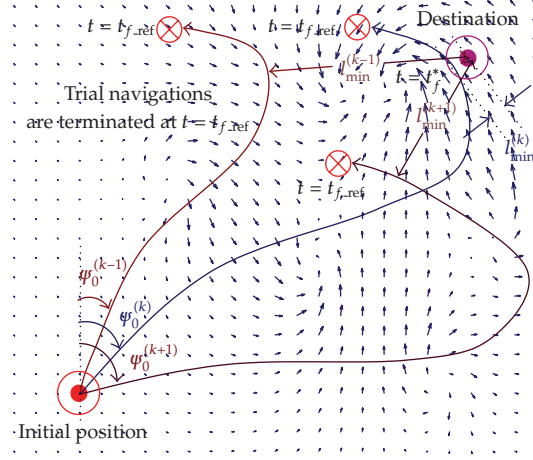
Figure 12: Trial navigations corresponding to a few discrete initial heading trials. The approximate optimal navigation converging to the destination is determined on the basis of the minimum distances. Note that all trial navigations are continued until $t = t_{f\_ref}$.

determine its true minimum distance. This is where the reference navigation time prepared beforehand comes into play. It is apparent that the reference navigation is nonoptimal, since it is based on an arbitrary guidance law, only achieving the vehicle's arrival at the destination. Therefore, the reference navigation time must be larger or equal to that of the optimal navigation as

$$O < t_f^* \leq t_{f\_ref}, \tag{5.4}$$

where t_f^* represents navigation time with the optimal guidance. It should be noted here that, according to the minimum principle, once we have started a trial navigation with an initial heading close to the optimal one, a vehicle should pass by the vicinity of the destination at a time surely smaller than t_{f_ref}. In other words, the reference navigation time qualifies to be the upper limit of the simulation times of the trial navigations, in order to identify an optimal navigation among the trial ones. In Figure 12, among all trial navigations continued until $t = t_{f\_ref}$, (k)th trial navigation marks the smallest minimum distance, satisfying (5.3). Therefore, the (k)th trial navigation is determined as the optimal navigation.

6. Optimal Navigation Examples

6.1. The Reference Navigation

As mentioned in the previous section, to implement the optimal guidance for an AUV navigation by AREN, it is necessary to make a reference navigation beforehand. The simplest guidance satisfying the vehicle's arrival at the destination may be proportional navigation (PN) [25, 26]. In PN, vehicle heading is continuously adjusted so that the line of sight (LOS) is directed toward the target point. In our work, we employ PN as the reference navigation.

6.2. Optimal Navigation in a Shearing Flow

The first numerical example in this research is an optimal navigation in a current disturbance of the linear shear flow, taken from Bryson and Ho [26]. The current velocity in this problem is described by

$$u_c(x, y) = 0,$$
$$v_c(x, y) = -\frac{U_c x}{h},$$

(6.1)

where U_c and h are set to be 1.544 m/s and 100 m, respectively. Starting from the initial position at $(x_0, y_0) = (-186\,\text{m}, 366\,\text{m})$, the vehicle is directed to move toward the destination at the origin in this example. Due to its simplicity, the current distribution of (6.1) allows derivation of the analytic optimal guidance law expressed as

$$\frac{x}{h} = \csc \psi - \csc \psi_f,$$

$$\frac{y}{h} = \frac{1}{2}\left[\csc \psi_f (\cot \psi - \cot \psi_f) + \cot \psi (\csc \psi_f - \csc \psi) + \log \frac{\csc \psi_f - \cot \psi_f}{\csc \psi - \cot \psi}\right],$$

(6.2)

where ψ_f is the vehicle heading at the final state.

Navigation trajectories are shown in Figure 13. In the reference navigation by PN, significant adverse drift happens at the initial stage, since within the region of $|x| > 100\,\text{m}$ current flow speed exceeds the vehicle speed relative to the water. The optimal guidance detours the vehicle across the upper half plane, taking advantage of the favorable current flow. Navigation times by PN and optimal guidance are 353.7 and 739.2 s, respectively, indicating a 52% decrease in navigation time by the optimal guidance proposed.

6.3. Optimal Navigation in a Time-Varying Flow

The next numerical example is an optimal navigation in a time-varying current flow. In actual sea environments, for a lot of currents the direction and the magnitude of their velocities change continuously like tidal flows. As mentioned previously, we have proved that the optimal guidance law (5.1) is also valid for time-varying currents as well as for stationary ones. Therefore, once the flow velocity distribution in a navigation region is described as a function of the position and time, our numerical scheme is expected to be effective and thus realize the minimum-time navigation in a time-varying flow.

Navigation trajectories in an artificially made time-varying current are depicted in Figure 14. As shown in the figure, near the middle of the navigation region, the vehicle following the guidance of PN temporarily fails to proceed toward the destination due to severe drift caused by strong local current flow of adverse direction. The occurrence of such a disadvantage is prevented in the optimal navigation. By following the optimal guidance, the vehicle proceeds taking advantage of the favorable flows and avoids passing through the region in which the current flow is gradually changing to the strongly adverse one. In Figure 14(b), it should be noted that at 623.0 s, having been released from the severe drift,
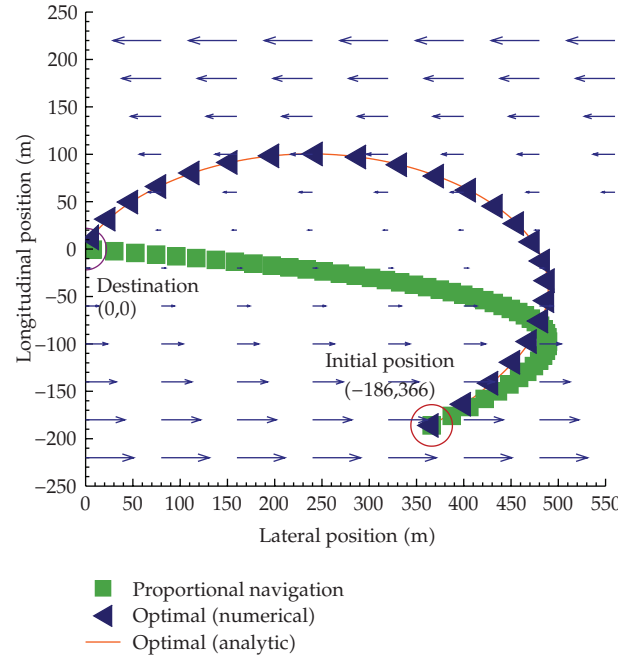
Figure 13: Navigation trajectories in a linear shearing current flow.

the vehicle under the guidance of PN is about to restart toward the destination. At the same instant, however, the optimal guidance has already made the vehicle arrive at the destination.

6.4. QuasiOptimal Navigation

The quasioptimal navigation is a fail-safe strategy introduced to react to failure in realizing the optimal navigation, due to environmental uncertainties or temporal malfunctions in sensors and actuators. The basic idea of the quasioptimal navigation is quite simple. The optimal heading reference during the quasioptimal navigation is revised repeatedly in response to the on-site request to prevent failure in on-going optimal navigation. According to Bellman's principle of optimality [25, 26], once we have failed in tracking the optimal trajectory, the best policy we can take from then on is to construct and follow a revised optimal trajectory starting at the present state. To derive the revised optimal trajectory, the optimal guidance law (5.1) has to be applied and solved again by using the proposed numerical solution procedure, that is, AREN, taking the present vehicle position as the new initial position. The whole trajectory obtained here is not optimal, since it already has included the past nonoptimal interval. Nonetheless, it is evidently the best trajectory we can take under the present situation, so that we call the corresponding navigation the quasioptimal navigation. Optimal and quasioptimal navigation trajectories in a time-varying current flow are shown in Figure 15.

The current distribution in this example is the same one that we took in the previous example. In this example, however, while the optimal navigation is performed with the exact information about the current flow distribution, assuming a situation of incorrect localization due to sensor failure, mismatched current flow information is fed to the vehicle guidance
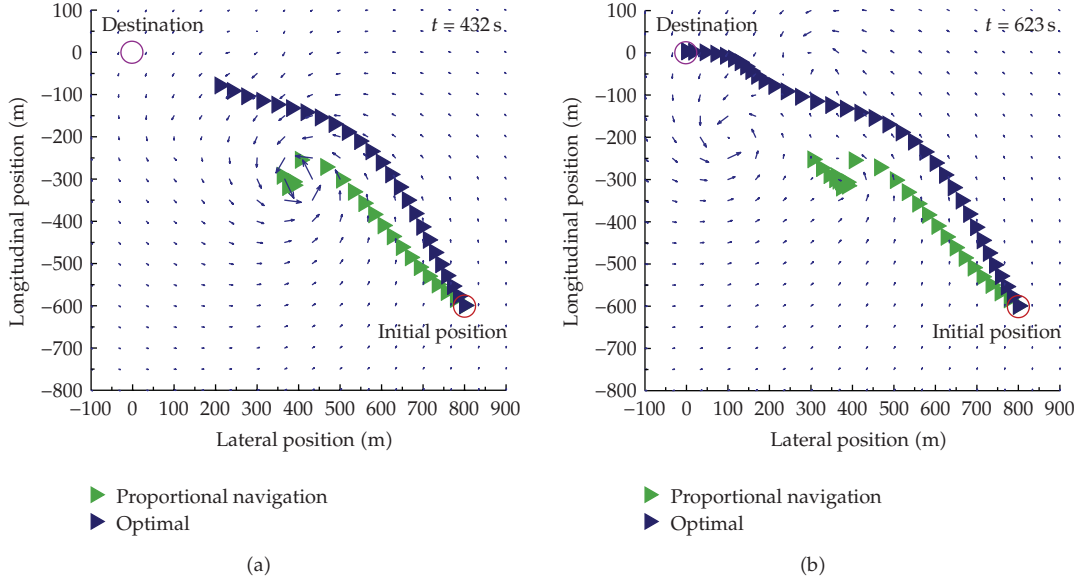
Figure 14: Navigation trajectories in a time-varying current flow at (a) $t = 432$ s and (b) $t = 623$ s.

controller in the quasioptimal case. The time interval during which mismatched information is taken is $0.0 \sim 252.0$ s. Starting at 252.0 s, optimal guidance revised on the basis of the correct current flow information achieves the quasioptimal navigation. Figure 16 shows the time sequence of the vehicle headings during the optimal and the quasioptimal navigations.

As expected naturally, the performance of the quasioptimal navigation is not as high as that of the optimal one. While the optimal guidance completes the navigation at 623 s, the quasioptimal one continues it until 702 s. Note that in Figure 16, an abrupt heading change occurs at 252.0 s during the quasioptimal navigation.

6.5. Optimal Navigation in Northwestern Pacific

In what follows, we try to accomplish the minimum-time navigation within a real sea environment. The sea region selected for this optimal navigation example is located in the Northwestern Pacific Ocean near Japan. The current field considered here is an actual measurement of the surface flow provided by the Japan Meteorological Agency, available at http://www.data.kishou.go.jp/db/kobe/db_kobe.html. The most notable environmental characteristic in this sea region is the current field dominated by the Kuroshio. The Kuroshio is a strong western boundary current in the Northwestern Pacific Ocean, flowing northeastward along the eastern coast of Japan [29]. As seen in Figure 18, Kuroshio-induced flows moving eastwards constitute the principal stream in this region.

In the current field data from the database of the Japan Meteorological Agency, current velocity is defined only on the predefined, large-scale grid nodes covering the sea region. As noticeable from (5.1), however, in order to derive the optimal heading reference, current velocity and its gradient at every vehicle position have to be available. As a remedy for this data deficiency, we estimate the current velocity and its gradient by interpolating the predefined values on grid nodes surrounding the present vehicle position. In applying
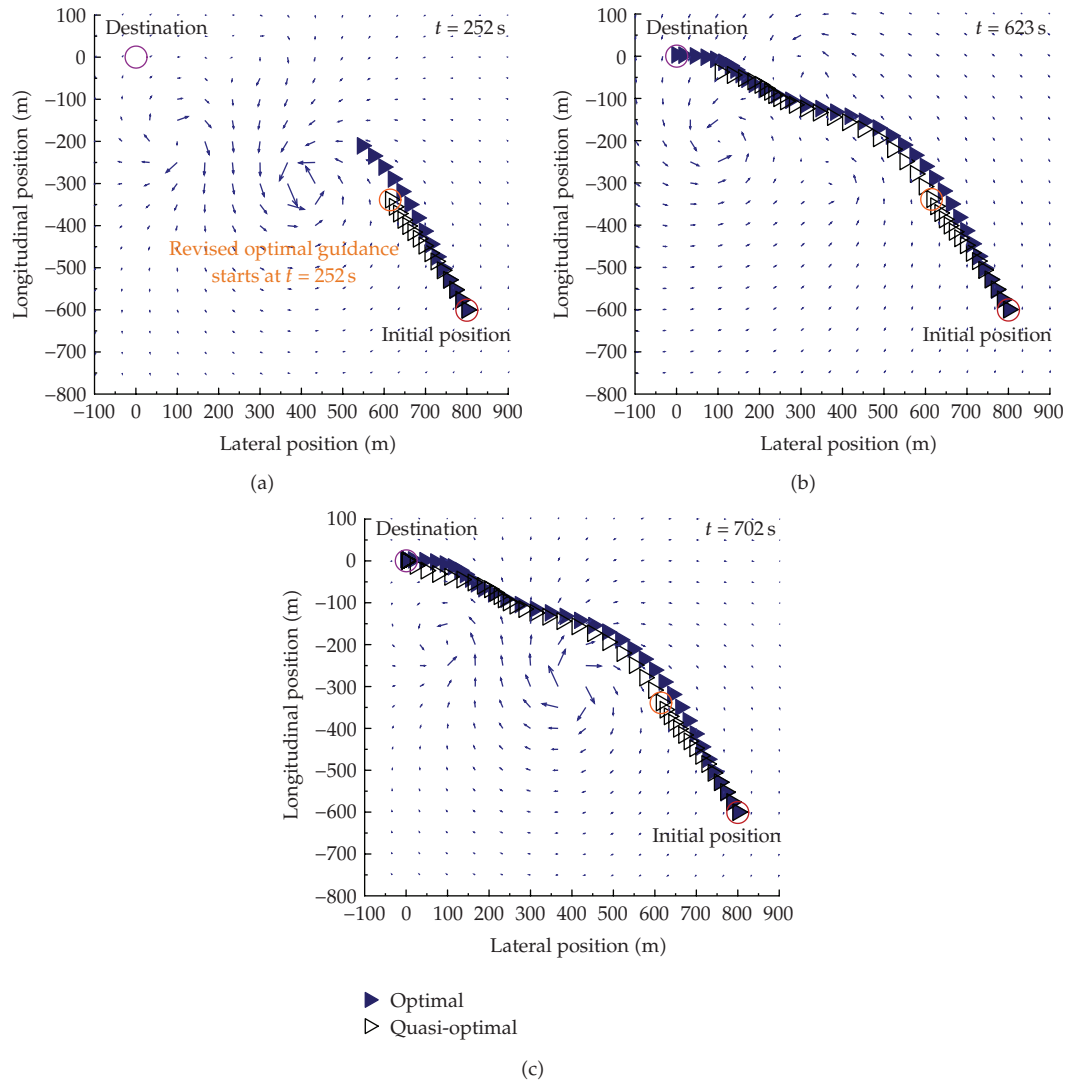
Figure 15: Navigation trajectories in a time-varying current flow at (a) $t = 252$ s, (b) $t = 623$ s, and (c) $t = 702$ s.

the interpolation, the nearest grid node to the present vehicle position has to be identified first. Then, the current velocity at the present vehicle position is estimated by 2D biquadratic interpolation utilizing values on the nearest node and surrounding eight nodes, as shown in Figure 17. Gradients of current velocities are obtained by the same manner. Since the velocity gradients are not provided from the database, however, prior to the interpolation, we calculate their nodal values by finite difference approximations.

The description of the navigation to be optimized here is as follows. Starting from an initial position, the vehicle is to transit to a destination in a mission-specified area, where an undersea survey mission is to be undertaken. Taking into account the inshore or harbor launch, the vehicle is made to start from the initial position off Minamiizu, the southern extreme of the Izu peninsula (Figure 18).
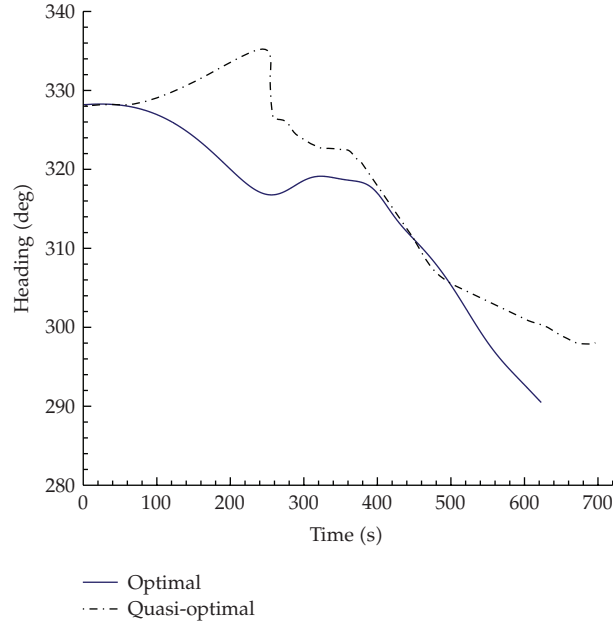
Figure 16: Time sequence of vehicle headings.
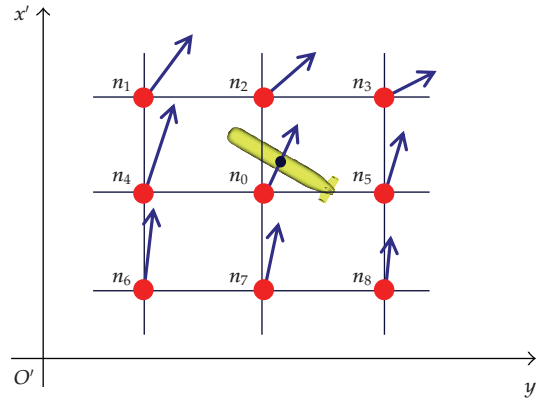
Figure 17: Cut sea area for interpolating current velocity and its gradient. When n_0 is the nearest grid node to the present vehicle position, current velocity and its gradient on that position are estimated using their values on n_0 to n_8.

Figure 18 shows the navigation trajectories achieved by the PN and optimal guidance. As shown in the figure, like the preceding examples in which exact values of current velocity and its gradients are available anywhere in the navigation region, the vehicle tracks the optimal reference trajectory with a negligibly small deviation. This indicates that our strategy of optimal navigation is also valid in the real sea current data, originally defined only on the coarsely defined discrete grid nodes.

In Figure 18, with the vehicle moving under PN, having reached the region of the mainstream of Kuroshio, its speed relative to the ground is remarkably reduced. This is

Figure 18: Navigation trajectories in a Northwestern Pacific Ocean region. The current data is built from actual measurements.

because in this region, for the vehicle following the guidance of PN, the direction of its advance velocity is placed out-of-phase with the direction of the mainstream of Kuroshio. In the optimal navigation, the former segment of the navigation trajectory is formed along the shoreline until the vehicle reache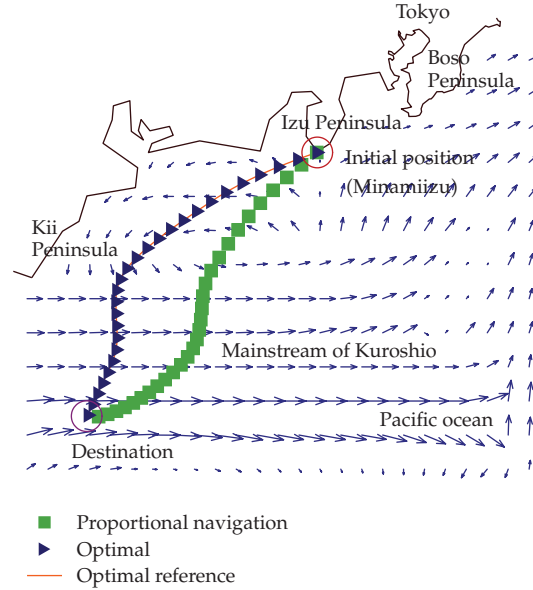s a point off the southern extreme of Kii peninsula. By taking this route, the vehicle attains a speed increase, riding the coastal current mainly flowing westwards. Note that upon the vehicle reaching a point off the southern extreme of the Kii peninsula, the optimal trajectory takes a large turn, slightly rolling back eastward from the destination. This slight rollback is the result of the optimal guidance's action to prevent the vehicle's advancing direction from being out-of-phase with that of the mainstream of Kuroshio, leading to the optimal navigation trajectory shown. The optimal trajectory obtained reveals one of the significant advantages of our approach over GA-based path planning which is not able to generate the optimal path with interim backward intervals by its nature [7, 9].

Navigation times by PN and optimal guidance are 232198 and 212006 s, respectively, indicating an 8.7% decrease in navigation time by the optimal guidance proposed.

7. Conclusions and Future Works

In this article, model-based analysis and synthesis to the following three research fields in AUV design and development have been presented.

(i) Dynamic system modelling of an AUV.

(ii) Motion control design and tracking control application.

(iii) Optimal guidance of an AUV in environmental disturbances.

In the dynamic system modelling of the AUV R-One, we evaluated the hydrodynamic loads by using CFD analyses. Then, by differentiating a hydrodynamic load with respect to the amount of a perturbation, corresponding stability derivatives were obtained. Using the stability derivatives evaluated, we built up the dynamic model of the R-One, which is characterized to be 6-DOF [3 longitudinal (surge, heave, pitch) + 3 lateral (sway, roll, yaw)], linear, and multiple-input multiple-output (MIMO).

Depth and heading control systems are designed by employing controller models based on the PID compensations. In the PID-tuning, model-based simulations for the depth and the heading controls are exploited in determining the optimal gains.

Concerning the guidance problem of AUVs moving in sea environmental disturbances, a newly developed procedure for obtaining the numerical solution of the optimal guidance law to achieve the minimum-time navigation has been presented. The optimal heading is obtained as the solution of the optimal guidance law, which is fed to the heading controller as the optimal reference. Reduced computational cost is one of the outstanding features of the proposed procedure. Numerical calculations of the optimal navigation examples presented in this article except for the last one are completed within 10 minutes on a single core 2.4 GHz windows XP platform. Moreover, unlike other path-finding algorithms such as DP or GAs, our procedure does not require a computation time increase for the time-varying problems.

As a fail-safe strategy for putting the proposed optimal navigation into execution, the concept of quasioptimal guidance has been proposed. The fact that there actually are several possible actions lessening the chance of achieving optimality emphasizes the practical importance of the quasioptimal navigation.

We have not considered the problem of unknown or nondeterministic currents. Our approach cannot be applied to an entirely unknown environment. For a sea region containing partially or coarsely defined currents, however, an estimated distribution can be built via interpolation and extrapolation, as shown in the last navigation example. The estimation possibly contains more or less uncertainty. Notably, however, it is the quasioptimal strategy that can cope with the environmental uncertainty. When the uncertainty in the estimation is significant, convergence may not be guaranteed.

Acknowledgments

References

[1] B. Etkin, *Dynamics of Flight—Stability and Control*, John Wiley & Sons, New York, NY, USA, 1996.

[2] D. McRuer, I. Ashkenas, and D. Graham, *Aircraft Dynamics and Automatic Control*, Princeton University Press, Princeton, NJ, USA, 1990.

[3] E. V. Lewis, Ed., *Principles of Naval Architecture. Vol III. Motions in Waves and Controllability*, The Society of Naval Architects and Marine Engineers, Jersey City, NJ, USA, 1989.

[4] M. A. Park, L. L. Green, R. C. Montgomery, and L. L. Raney, "Determination of stability and control derivatives using computational fluid dynamics and automatic differentiation," in *Proceedings of the 17th AIAA Applied Aerodynamics Conference*, Norfolk, Va, USA, June 1999, AIAA paper 99-3136.

[5] F. T. Johnson, E. N. Tinoco, and N. J. Yu, "Thirty years of development and application of CFD at Boeing commercial airplanes, Seattle," *Computers & Fluids*, vol. 34, no. 10, pp. 1115–1151, 2005.

[6] T. Ohmori, "Finite-volume simulation of flows about a ship in maneuvering motion," *Journal of Marine Science and Technology*, vol. 3, no. 2, pp. 82–93, 1998.

[7] A. Alvarez, A. Caiti, and R. Onken, "Evolutionary path planning for autonomous underwater vehicle in a variable ocean," *IEEE Journal of Oceanic Engineering*, vol. 29, no. 2, pp. 418–429, 2004.

[8] N. A. Papadakis and A. N. Perakis, "Deterministic minimal time vessel routing," *Operations Research*, vol. 38, no. 3, pp. 426–438, 1990.

[9] C. Hocaolglu and A. C. Sanderson, "Planning multi-paths using speciation in generic algorithms," in *Proceedings of IEEE International Conference on Evolutionary Computation (ICEC '97)*, pp. 378–383, Indianapolis, Ind, USA, April 1997.

[10] D. A. Smallwood and L. L. Whitcomb, "Model-based dynamic positioning of underwater robotic vehicles: theory and experiment," *IEEE Journal of Oceanic Engineering*, vol. 29, no. 1, pp. 169–186, 2004.

[11] K. Kim and T. Ura, "Optimal and quasi-optimal navigations of an AUV in current disturbances," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '08)*, pp. 3661–3667, Nice, France, September 2008.

[12] T. Ura, "Development of autonomous underwater vehicles in Japan," *Advanced Robotics*, vol. 16, no. 1, pp. 3–15, 2002.

[13] CD-adapco, *STAR-CD V3.2 User Guide Manuals*, CD-adapco, New York, NY, USA, 2002.

[14] G. Iaccarino, "Predictions of a turbulent separated flow using commercial CFD codes," *Journal of Fluids Engineering*, vol. 123, no. 4, pp. 819–828, 2001.

[15] M. Garbey and C. Picard, "A code-independent technique for computational verification of fluid mechanics and heat transfer problems," *Acta Mechanica Sinica*, vol. 24, no. 4, pp. 387–397, 2008.

[16] J. L. Lumley, "Computational modeling of turbulent flows," *Advances in Applied Mechanics*, vol. 18, pp. 123–176, 1978.

[17] J. F. Thompson, "A composite grid generation code for general 3-D regions—the Eagle code," *AIAA Journal*, vol. 26, no. 3, pp. 271–272, 1988.

[18] F. M. White, *Fluid Mechanics*, McGraw-Hill, New York, NY, USA, 1988.

[19] T. Ura, T. Obara, S. Takagawa, and T. Gamo, "Exploration of Teisi Knoll by autonomous underwater vehicle "R-One Robot"," in *Proceedings of MTS/IEEE OCEANS*, vol. 1, pp. 456–461, Honolulu, Hawaii, USA, November 2001.

[20] T. Takagi and M. Sugeno, "Fuzzy identifications of systems and its application to modelling and control," *IEEE Transactions on Systems Man and Cybernetics*, vol. 15, no. 1, pp. 116–132, 1985.

[21] K. S. Narendra and K. Parathasarathy, "Identification and control of dynamic system using neural network," *IEEE Transactions on Neural Network*, vol. 1, no. 1, pp. 4–27, 1990.

[22] S. Haykin, *Neural Networks*, Prentice Hall, Upper Saddle River, NJ, USA, 1999.

[23] H.-X. Li and S. Guan, "Hybrid intelligent control strategy," *IEEE Control Systems Magazine*, vol. 21, no. 3, pp. 36–48, 2001.

[24] The Mathworks Inc., *Control System Toolbox User's Guide*, The Mathworks Inc., Natick, Mass, USA, 2007.

[25] F. L. Lewis and V. L. Syrmos, *Optimal Control*, A Wiley-Interscience Publication, John Wiley & Sons, New York, NY, USA, 1986.

[26] A. E. Bryson and Y. C. Ho, *Applied Optimal Control*, Taylor & Francis, Levittown, NY, USA, 1975.

[27] L. G. Crespo and J. Q. Sun, "Optimal control of target tracking with state constraints via cell mapping," *Journal of Guidance, Control, and Dynamics*, vol. 24, no. 5, pp. 1029–1031, 2001.

[28] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in Fortran*, Cambridge University Press, Cambridge, UK, 2nd edition, 1992.

[29] S. Imawaki, H. Ichikawa, M. Ikeda, A. Isobe, and M. Kamachi, "Introduction to special section: Kuroshio observation, state estimation and prediction," *Journal of Oceanography*, vol. 60, no. 3, pp. 265–268, 2004.

Research Article

Control of Limit Cycle Oscillations of a Two-Dimensional Aeroelastic System

M. Ghommem, A. H. Nayfeh, and M. R. Hajj

Department of Engineering Science and Mechanics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Correspondence should be addressed to M. R. Hajj, mhajj@vt.edu

Linear and nonlinear static feedback controls are implemented on a nonlinear aeroelastic system that consists of a rigid airfoil supported by nonlinear springs in the pitch and plunge directions and subjected to nonlinear aerodynamic loads. The normal form is used to investigate the Hopf bifurcation that occurs as the freestream velocity is increased and to analytically predict the amplitude and frequency of the ensuing limit cycle oscillations (LCO). It is shown that linear control can be used to delay the flutter onset and reduce the LCO amplitude. Yet, its required gains remain a function of the speed. On the other hand, nonlinear control can be effciently implemented to convert any subcritical Hopf bifurcation into a supercritical one and to significantly reduce the LCO amplitude.

1. Introduction

The response of an aeroelastic system is governed by a combination of linear and nonlinear dynamics. When combined, the nonlinearities (geometric, inertia, free-play, damping, and/or aerodynamics) lead to different behavior [1–3], including multiple equilibria, bifurcations, limit cycles, chaos, and various types of resonances (internal and super/subharmonic) [4]. A generic nonlinear system that has been used to characterize aeroelastic behavior and dynamic instabilities is a two-dimensional rigid airfoil undergoing pitch and plunge motions [5, 6]. As the parameters of this system (e.g., freestream velocity) are varied, changes may occur in its behavior. Of particular interest is its response around a bifurcation point. Depending on the relative magnitude and type of nonlinearity, the bifurcation can be of the subcritical or supercritical type. Hence, one needs to consider the combined effects of all nonlinearities to predict the system's response. Furthermore, the nonlinearities provide an opportunity to implement a combination of linear and nonlinear control strategies to delay the occurrence

of the bifurcation (i.e., increase the allowable flight speed) and avoid catastrophic behavior of the system (subcritical Hopf bifurcation) by suppressing or even alleviating the large-amplitude LCO and eliminating LCO that may take place at speeds lower than the nominal flutter speed.

Different methods have been proposed to control bifurcations and achieve desirable nonlinear effects in complex systems. Abed and Fu [7, 8] proposed a nonlinear feedback control to suppress discontinuous bifurcations of fixed points, such as subcritical Hopf bifurcations, which can result in loss of synchronism or voltage collapse in power systems. For the pitch-plunge airfoil, Strganac et al. [9] used a trailing edge flap to control a two-dimensional nonlinear aeroelastic system. They showed that linear control strategies may not be appropriate to suppress large-amplitude LCO and proposed a nonlinear controller based on partial feedback linearization to stabilize the LCO above the nominal flutter velocity. Librescu et al. [10] implemented an active flap control for 2D wing-flap systems operating in an incompressible flow field and exposed to a blast pulse and demonstrated its performances in suppressing flutter and reducing the vibration level in the subcritical flight speed range. Kang [11] developed a mathematical framework for the analysis and control of bifurcations and used an approach based on the normal form to develop a feedback design for delaying and stabilizing bifurcations. His approach involves a preliminary state transformation and center manifold reduction.

In this work, we present a methodology to convert subcritical bifurcations of aeroelastic systems into supercritical bifurcations. This methodology involves the following steps: (i) reduction of the dynamics of the system into a one-dimensional dynamical system using the method of multiple scales and then (ii) designing a nonlinear feedback controller to convert subcritical to supercritical bifurcations and reduce the amplitude of any ensuing LCO.

2. Representation of the Aeroelastic System

The aeroelastic system, considered in this work, is modeled as a rigid wing undergoing two-degree-of-freedom motions, as presented in Figure 1. The wing is free to rotate about the elastic axis (pitch motion) and translate vertically (plunge motion). Denoting by h and α the plunge deflection and pitch angle, respectively, we write the governing equations of this system as [4, 9]

$$\begin{pmatrix} m_T & m_W x_\alpha b \\ m_W x_\alpha b & I_\alpha \end{pmatrix} \begin{pmatrix} \ddot{h} \\ \ddot{\alpha} \end{pmatrix} + \begin{pmatrix} c_h & 0 \\ 0 & c_\alpha \end{pmatrix} \begin{pmatrix} \dot{h} \\ \dot{\alpha} \end{pmatrix} + \begin{pmatrix} k_h(h) & 0 \\ 0 & k_\alpha(\alpha) \end{pmatrix} \begin{pmatrix} h \\ \alpha \end{pmatrix} = \begin{pmatrix} -L \\ M \end{pmatrix}, \quad (2.1)$$

where m_T is the total mass of the wing and its support structure, m_W is the wing mass alone, I_α is the mass moment of inertia about the elastic axis, b is the half chord length, $x_\alpha = r_{cg}/b$ is the nondimensionalized distance between the center of mass and the elastic axis, c_h and c_α are the plunge and pitch structural damping coefficients, respectively, L and M are the aerodynamic lift and moment about the elastic axis, and k_h and k_α are the structural stiffnesses for the plunge and pitch motions, respectively. These stiffnesses are approximated
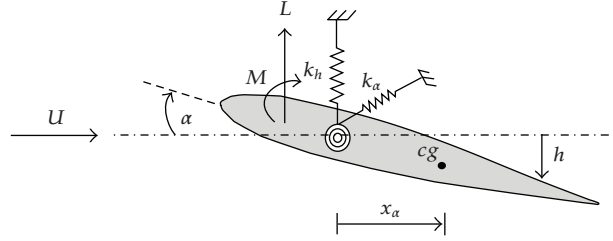
Figure 1: Sketch of a two-dimensional airfoil.

in polynomial form by

$$k_\alpha(\alpha) = k_{\alpha 0} + k_{\alpha 1}\alpha + k_{\alpha 2}\alpha^2 + \cdots,$$
$$k_h(h) = k_{h0} + k_{h1}h + k_{h2}h^2 + \cdots. \tag{2.2}$$

The aerodynamic loads are evaluated using a quasi-steady approximation with a stall model [9] and written as

$$L = \rho U^2 b c_{l_\alpha}\left(\alpha_{\text{eff}} - c_s \alpha_{\text{eff}}^3\right),$$
$$M = \rho U^2 b^2 c_{m_\alpha}\left(\alpha_{\text{eff}} - c_s \alpha_{\text{eff}}^3\right), \tag{2.3}$$

where U is the freestream velocity, c_{l_α} and c_{m_α} are the aerodynamic lift and moment coefficients, and c_s is a nonlinear parameter associated with stall. The effective angle of attack due to the instantaneous motion of the airfoil is given by [9]

$$\alpha_{\text{eff}} = \alpha + \frac{\dot{h}}{U} + \left(\frac{1}{2} - a\right)b\frac{\dot{\alpha}}{U}, \tag{2.4}$$

where a is the nondimensionalized distance from the midchord to the elastic axis.

For the sake of simplicity, we define the state variables

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = \begin{pmatrix} h \\ \alpha \\ \dot{h} \\ \dot{\alpha} \end{pmatrix}, \tag{2.5}$$

and write the equations of motion in the form

$$\dot{\mathbf{Y}} = F(\mathbf{Y}, U), \tag{2.6}$$

Table 1: System variables.

$$d = m_T I_\alpha - m_W^2 x_\alpha^2 b^2$$

$$k_1 = (I_\alpha \rho b c_{l_\alpha} + m_W x_\alpha \rho b^3 c_{m_\alpha})/d$$

$$k_2 = -(m_W x_\alpha \rho b^2 c_{l_\alpha} + m_T \rho b^2 c_{m_\alpha})/d$$

$$c_1 = [I_\alpha (c_h + \rho U b c_{l_\alpha}) + m_W x_\alpha \rho U b^3 c_{m_\alpha}]/d$$

$$c_2 = [I_\alpha \rho U b^2 c_{l_\alpha}(1/2 - a) - m_W x_\alpha b c_{l_\alpha} + m_W x_\alpha \rho U b^4 c_{m_\alpha}(1/2 - a)]/d$$

$$c_3 = [-m_W x_\alpha b(c_h + \rho U b c_{l_\alpha}) - m_T x_\alpha \rho U b^2 c_{m_\alpha}]/d$$

$$c_4 = [m_T(c_\alpha - \rho U b^3 c_{m_\alpha}(1/2 - a)) - m_W x_\alpha \rho U b^3 c_{l_\alpha}(1/2 - a)]/d$$

$$p_\alpha(\mathbf{Y}) = -m_W x_\alpha b k_\alpha(\mathbf{Y})/d$$

$$q_\alpha(\mathbf{Y}) = m_T k_\alpha(\mathbf{Y})/d$$

$$p_h(\mathbf{Y}) = I_\alpha k_h(\mathbf{Y})/d$$

$$q_h(\mathbf{Y}) = -m_W x_\alpha b k_h(\mathbf{Y})/d$$

$$g_{NL1}(\mathbf{Y}) = (c_s \rho U^2 b)(c_{l_\alpha} I_\alpha + m_W x_\alpha b^2 c_{m_\alpha}) \alpha_{\text{eff}}^3(\mathbf{Y})/d$$

$$g_{NL2}(\mathbf{Y}) = -(c_s \rho U^2 b^2)(c_{l_\alpha} m_W x_\alpha + m_T c_{m_\alpha}) \alpha_{\text{eff}}^3(\mathbf{Y})/d$$

where

$$F(\mathbf{Y}, U) = \begin{pmatrix} Y_3 \\ Y_4 \\ -p_h(Y_1)Y_1 - (k_1 U^2 + p_\alpha(Y_2))Y_2 - c_1 Y_3 - c_2 Y_4 + g_{NL1}(Y) \\ -q_h(Y_1)Y_1 - (k_2 U^2 + q_\alpha(Y_2))Y_2 - c_3 Y_3 - c_4 Y_4 + g_{NL2}(Y) \end{pmatrix}. \tag{2.7}$$

The set of new variables that are used in (2.7) in terms of physical parameters is provided in Table 1. The original system, (2.6), is then rewritten as

$$\dot{\mathbf{Y}} = A(U)\mathbf{Y} + Q(\mathbf{Y}, \mathbf{Y}) + C(\mathbf{Y}, \mathbf{Y}, \mathbf{Y}), \tag{2.8}$$

where $Q(\mathbf{Y}, \mathbf{Y})$ and $C(\mathbf{Y}, \mathbf{Y}, \mathbf{Y})$ are, respectively, the quadratic and cubic vector functions of the state variables collected in the vector $\mathbf{Y}$.

To determine the system's stability, we consider the linearized governing equations, which are written in a first-order differential form as

$$\dot{\mathbf{Y}} = A(U)\mathbf{Y}, \tag{2.9}$$

where

$$A(U) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\dfrac{I_\alpha k_{h0}}{d} & -\left(k_1 U^2 - \dfrac{m_W x_\alpha b k_{\alpha 0}}{d}\right) & -c_1 & -c_2 \\ \dfrac{m_W x_\alpha b k_{h0}}{d} & -\left(k_2 U^2 + \dfrac{m_T k_{\alpha 0}}{d}\right) & -c_3 & -c_4 \end{pmatrix}. \tag{2.10}$$
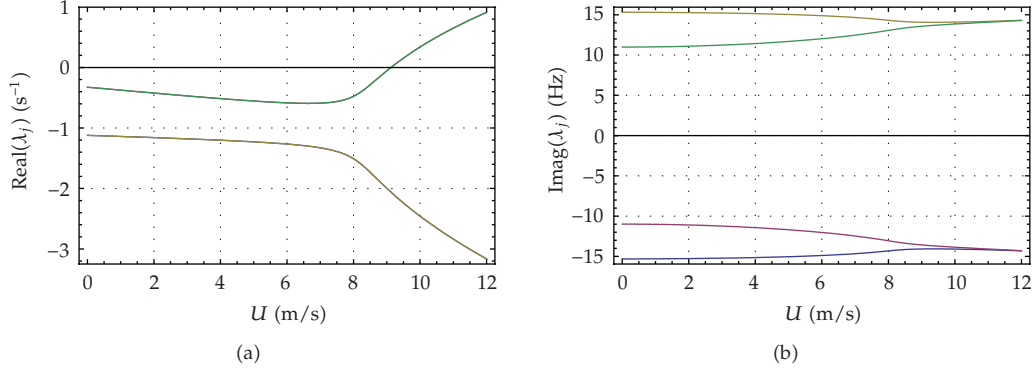
(a)

(b)

Figure 2: Variations of (a) damping ($\mathrm{Real}(\lambda_j)$) and (b) frequencies ($\mathrm{Imag}(\lambda_j)$) with the freestream velocity U.

The 4×4 matrix $A(U)$ has a set of four eigenvalues, $\{\lambda_j, \; j = 1, 2, \ldots, 4\}$. These eigenvalues determine the stability of the trivial solution of (2.6). If the real parts of all of the λ_j are negative, the trivial solution is asymptotically stable. On the other hand, if the real part of one or more eigenvalues is positive, the trivial solution is unstable. The flutter speed U_f, for which one or more eigenvalues have zero real parts, corresponds to the onset of linear instability. For the specific values given in [9], Figures 2(a) and 2(b) show, respectively, variations of the real and imaginary parts of the λ_j with U, which, respectively, correspond to the damping and frequencies of the plunge and pitch motions. We note that the damping of two modes becomes positive at $U_f = 9.1242 \, \mathrm{m/s}$, which corresponds to the flutter speed at which the aeroelastic system undergoes a Hopf bifurcation.

3. Static Feedback Control

To manage the Hopf bifurcation and achieve desirable nonlinear dynamics, we follow Nayfeh and Balachandran [12] and use a static feedback control. To the system given by (2.6), we add a static feedback $\mathbf{u}(\mathbf{Y})$, which includes linear, $L_u \mathbf{Y}$, quadratic $Q_u(\mathbf{Y}, \mathbf{Y})$, and cubic $C_u(\mathbf{Y}, \mathbf{Y}, \mathbf{Y})$ components; that is,

$$\mathbf{u}(\mathbf{Y}) = L_u \mathbf{Y} + Q_u(\mathbf{Y}, \mathbf{Y}) + C_u(\mathbf{Y}, \mathbf{Y}, \mathbf{Y}). \tag{3.1}$$

Hence, the controlled system takes the form

$$\dot{\mathbf{Y}} = F(\mathbf{Y}, U) + \mathbf{u}(\mathbf{Y}). \tag{3.2}$$

3.1. Normal Form of Hopf Bifurcation

To compute the normal form of the Hopf bifurcation of (3.2) near $U = U_f$, we follow Nayfeh and Balachandran [12] and introduce a small nondimensional parameter ϵ as a book keeping

parameter. Defining the velocity perturbation as a ratio of the flutter speed $\sigma_U U_f$, we write $U = U_f + \epsilon^2 \sigma_U U_f$ and seek a third-order approximate solution of (3.2) in the form

$$\mathbf{Y}(t, \sigma_U, \sigma_\alpha, \sigma_h) = \epsilon \mathbf{Y}_1(T_0, T_2) + \epsilon^2 \mathbf{Y}_2(T_0, T_2) + \epsilon^3 \mathbf{Y}_3(T_0, T_2) + \cdots , \tag{3.3}$$

where the time scales $T_m = \epsilon^m t$. In terms of these scales, the time derivative d/dt is written as

$$\frac{d}{dt} = \frac{\partial}{\partial T_0} + \epsilon^2 \frac{\partial}{\partial T_2} + \cdots = D_0 + \epsilon^2 D_2 + \cdots . \tag{3.4}$$

Scaling L_u as $\epsilon^2 L_u$, substituting (3.3) and (3.4) into (3.2), and equating coefficients of like powers of ϵ, we obtain

 Order (ϵ),

$$D_0 \mathbf{Y}_1 - A(U_f)\mathbf{Y}_1 = 0, \tag{3.5}$$

 Order (ϵ^2),

$$D_0 \mathbf{Y}_2 - A(U_f)\mathbf{Y}_2 = Q(\mathbf{Y}_1, \mathbf{Y}_1) + Q_u(\mathbf{Y}_1, \mathbf{Y}_1), \tag{3.6}$$

 Order (ϵ^3),

$$D_0 \mathbf{Y}_3 - A(U_f)\mathbf{Y}_3 = -D_2 \mathbf{Y}_1 + \sigma_U B \mathbf{Y}_1 + L_u \mathbf{Y}_1 + 2[Q(\mathbf{Y}_1, \mathbf{Y}_2) + Q_u(\mathbf{Y}_1, \mathbf{Y}_2)]$$
$$+ C(\mathbf{Y}_1, \mathbf{Y}_1, \mathbf{Y}_1) + C_u(\mathbf{Y}_1, \mathbf{Y}_1, \mathbf{Y}_1), \tag{3.7}$$

where

$$B = -2k_1 U_f^2 I_1 - 2k_2 U_f^2 I_2, \quad I_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad I_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{3.8}$$

The general solution of (3.5) is the superposition of four linearly independent solutions corresponding to the four eigenvalues: two of these eigenvalues have negative real parts and the other two are purely imaginary $(\pm i\omega)$. Because the two solutions corresponding to the two eigenvalues with negative real parts decay as $T_0 \rightarrow \infty$, we retain only the nondecaying solutions and express the general solution of the first-order problem as

$$\mathbf{Y}_1(T_0, T_2) = \eta(T_2)\mathbf{p}e^{i\omega T_0} + \overline{\eta}(T_2)\overline{\mathbf{p}}e^{-i\omega T_0}, \tag{3.9}$$

where $\eta(T_2)$ is determined by imposing the solvability condition at the third-order level and $\mathbf{p}$ is the eigenvector of $A(U_f)$ corresponding to the eigenvalue $i\omega$; that is,

$$A(U_f)\mathbf{p} = i\omega\mathbf{p}. \tag{3.10}$$

Substituting (3.9) into (3.6) yields

$$D_0\mathbf{Y}_2 - A(U_f)\mathbf{Y}_2 = [Q(\mathbf{p},\mathbf{p}) + Q_u(\mathbf{p},\mathbf{p})]\eta^2 e^{2i\omega T_0} + 2[Q(\mathbf{p},\overline{\mathbf{p}}) + Q_u(\mathbf{p},\overline{\mathbf{p}})]\eta\overline{\eta}$$
$$+ [Q(\overline{\mathbf{p}},\overline{\mathbf{p}}) + Q_u(\overline{\mathbf{p}},\overline{\mathbf{p}})]\overline{\eta}^2 e^{-2i\omega T_0}. \tag{3.11}$$

The solution of (3.11) can be written as

$$\mathbf{Y}_2 = (\zeta_2 + \zeta_{2u})\eta^2 e^{2i\omega T_0} + 2(\zeta_0 + \zeta_{0u})\eta\overline{\eta} + \left(\overline{\zeta}_2 + \overline{\zeta}_{2u}\right)\overline{\eta}^2 e^{-2i\omega T_0}, \tag{3.12}$$

where

$$[2i\omega I - A(U_f)]\zeta_2 = Q(\mathbf{p},\mathbf{p}), \qquad [2i\omega I - A(U_f)]\zeta_{2u} = Q_u(\mathbf{p},\mathbf{p}),$$
$$A(U_f)\zeta_0 = -Q_u(\mathbf{p},\overline{\mathbf{p}}), \qquad A(U_f)\zeta_{0u} = -Q_u(\mathbf{p},\overline{\mathbf{p}}). \tag{3.13}$$

Substituting (3.9) and (3.12) into (3.7), we obtain

$$D_0\mathbf{Y}_3 - A(U_f)\mathbf{Y}_3 = -\Big[D_2\eta\mathbf{p} - (\sigma_U B + L_u)\eta\mathbf{p}$$
$$- (4Q(\mathbf{p},\zeta_0) + 2Q(\overline{\mathbf{p}},\zeta_2)$$
$$+ 3C(\mathbf{p},\mathbf{p},\overline{\mathbf{p}}) + 4Q_u(\mathbf{p},\zeta_0) + 2Q_u(\overline{\mathbf{p}},\zeta_2)$$
$$+ 3C_u(\mathbf{p},\mathbf{p},\overline{\mathbf{p}}) + 4Q(\mathbf{p},\zeta_{0u}) + 2Q(\overline{\mathbf{p}},\zeta_{2u}))\eta^2\overline{\eta}\Big]e^{i\omega T_0} + \text{cc} + \text{NST}, \tag{3.14}$$

where cc stands for the complex conjugate of the preceding terms and NST stands for terms that do not produce secular terms. We let $\mathbf{q}$ be the left eigenvector of $A(U_f)$ corresponding to the eigenvalue $i\omega$; that is,

$$A(U_f)^T\mathbf{q} = i\omega\mathbf{q}. \tag{3.15}$$

We normalize it so that $\mathbf{q}^T\mathbf{p} = 1$. Then, the solvability condition requires that terms proportional to $e^{i\omega T_0}$ in (3.14) be orthogonal to $\mathbf{q}$. Imposing this condition, we obtain the following normal of the Hopf bifurcation:

$$D_2\eta = \widehat{\beta}\eta + \widehat{\Lambda}\eta^2\overline{\eta}, \tag{3.16}$$

where

$$\widehat{\beta} = \beta + \beta_u, \tag{3.17}$$

with

$$\beta = \mathbf{q}^T \sigma_U B \mathbf{p}, \qquad \beta_u = \mathbf{q}^T L_u \mathbf{p},$$
$$\widehat{\Lambda} = \Lambda + \Lambda_u, \tag{3.18}$$

with

$$\Lambda = 4\mathbf{q}^T Q(\mathbf{p}, \zeta_0) + 2\mathbf{q}^T Q(\overline{\mathbf{p}}, \zeta_2) + 3\mathbf{q}^T C(\mathbf{p}, \mathbf{p}, \overline{\mathbf{p}}),$$
$$\Lambda_u = 4\mathbf{q}^T Q_u(\mathbf{p}, \zeta_0) + 2\mathbf{q}^T Q_u(\overline{\mathbf{p}}, \zeta_2) + 3\mathbf{q}^T C_u(\mathbf{p}, \mathbf{p}, \overline{\mathbf{p}}) \tag{3.19}$$
$$+ 4\mathbf{q}^T Q(\mathbf{p}, \zeta_{0u}) + 2\mathbf{q}^T Q(\overline{\mathbf{p}}, \zeta_{2u}).$$

Letting $\eta = (1/2)a \exp(i\theta)$ and separating the real and imaginary parts in (3.16), we obtain the following alternate normal form of the Hopf bifurcation:

$$\dot{a} = \widehat{\beta}_r a + \frac{1}{4} \widehat{\Lambda}_r a^3, \tag{3.20}$$

$$\dot{\theta} = \widehat{\beta}_i + \frac{1}{4} \widehat{\Lambda}_i a^2, \tag{3.21}$$

where $(\cdot)_r$ and $(\cdot)_i$ stand for the real and imaginary parts, respectively, a is the amplitude and $\dot{\theta}$ is the frequency of the oscillatory motion associated with the Hopf bifurcation.

We note that, because the a component is independent of θ, the system's stability is reduced to a one-dimensional dynamical system given by (3.20). Assuming that $\widehat{\Lambda}_r \neq 0$, a admits three steady-state solutions, namely,

$$a = 0, \quad a = \pm \sqrt{\frac{-4\widehat{\beta}_r}{\widehat{\Lambda}_r}}. \tag{3.22}$$

The trivial fixed point of (3.20) corresponds to the fixed point $(0,0)$ of (3.2), and a nontrivial fixed point (i.e., $a \neq 0$) of (3.20) corresponds to a periodic solution of (3.2). The origin is asymptotically stable when $\widehat{\beta}_r < 0$, unstable when $\widehat{\beta}_r > 0$, unstable when $\widehat{\beta}_r = 0$ and $\widehat{\Lambda}_r > 0$, and asymptotically stable when $\widehat{\beta}_r = 0$ and $\widehat{\Lambda}_r < 0$. On the other hand, the nontrivial fixed points exist when $-\widehat{\beta}_r \widehat{\Lambda}_r > 0$. They are stable when $\widehat{\beta}_r > 0$ and $\widehat{\Lambda}_r < 0$ (supercritical Hopf bifurcation) and unstable when $\widehat{\beta}_r < 0$ and $\widehat{\Lambda}_r > 0$ (subcritical Hopf bifurcation). We note that a stable nontrivial fixed point of (3.20) corresponds to a stable periodic solution of (3.2). Likewise, an unstable nontrivial fixed point of (3.20) corresponds to an unstable periodic solution of (3.2).

Therefore, to delay the occurrence of Hopf bifurcation (i.e., stabilize the aeroelastic system at speeds higher than the flutter speed), one needs to set the real part of $\widehat{\beta}$ to a negative value by appropriately managing the linear control represented by L_u in (3.2). To eliminate subcritical instabilities and limit LCO amplitudes to small values at speeds higher than the flutter speed (supercritical Hopf bifurcation which is a favorable instability for such

systems), the nonlinear feedback control given by $Q_u(\mathbf{Y}, \mathbf{Y}) + C_u(\mathbf{Y}, \mathbf{Y}, \mathbf{Y})$ should be chosen so that $\text{Real}(\Lambda + \Lambda_u) < 0$.

3.2. Case Study

To demonstrate the linear and nonlinear control strategies, we consider an uncontrolled case (i.e., $\mathbf{u}(\mathbf{Y}) = 0$) in which only the pitch structural nonlinearity is taken into account; that is, $k_{h1} = k_{h2} = c_s = 0$, $k_{\alpha1} = 9.9967$, and $k_{\alpha2} = 167.685$. The hysteretic response as a function of the freestream velocity, obtained through the numerical integration of (2.6) for these parameters, is presented in Figure 3. The onset of flutter takes place at $U_f = 9.1242\,\text{m/s}$ and is characterized by a jump to a large-amplitude LCO when transitioning through the Hopf bifurcation. As the speed is increased beyond the flutter speed, the LCO amplitudes of both of the pitch and plunge motions increase. Furthermore, LCO take place at speeds lower than U_f if the disturbances to the system are sufficiently large. Clearly, this configuration exhibits a subcritical instability $(\text{Real}(\Lambda) > 0)$.

For linear control, we consider the matrix L_u defined in (3.2) in the form of

$$L_u = \begin{pmatrix} -k_l & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{3.23}$$

where k_l is the linear feedback control gain. Then, for the specific values of the system parameters given in [9], we obtain

$$\hat{\beta} = \left(0.433891\sigma_U U_f - 0.110363k_l\right) + \iota\left(0.323942\sigma_U U_f - 0.303804k_l\right). \tag{3.24}$$

To guarantee damped oscillations of the airfoil at speeds higher than the flutter speed, one needs to set k_l to a value such that $\text{Real}(\hat{\beta}) < 0$. Using a gain of 10, we plot in Figure 4 the plunge and pitch displacements for a freestream velocity of $U = 10\,\text{m/s}$ with and without linear control. Clearly, linear control damps the LCO of the uncontrolled system. We note that, by increasing the linear feedback control gain k_l, the amplitudes of pitch and plunge decay more rapidly.

Although linear control is capable of delaying the onset of flutter in terms of speed and reducing the LCO amplitude, the system would require higher gains at higher speeds. Furthermore, it maintains its subcritical response. To overcome these difficulties and convert the subcritical instability to a supercritical one, we introduce the following nonlinear feedback control law:

$$\mathbf{u}^T = -\left(k_{\text{nl1}} \quad k_{\text{nl2}} \quad k_{\text{nl3}} \quad k_{\text{nl4}}\right)\dot{\alpha}^3, \tag{3.25}$$

where the k_{nli} are the nonlinear feedback control gains. For the specific airfoil's geometry given in [9], we obtain

$$\hat{\Lambda}_r = 0.866899 - (132.844k_{\text{nl1}} + 13.3643k_{\text{nl2}} + 2.9858k_{\text{nl3}} + 1.32415k_{\text{nl4}}). \tag{3.26}$$

(a) Plunge motion

Low IC
High IC

(b) Pitch motion
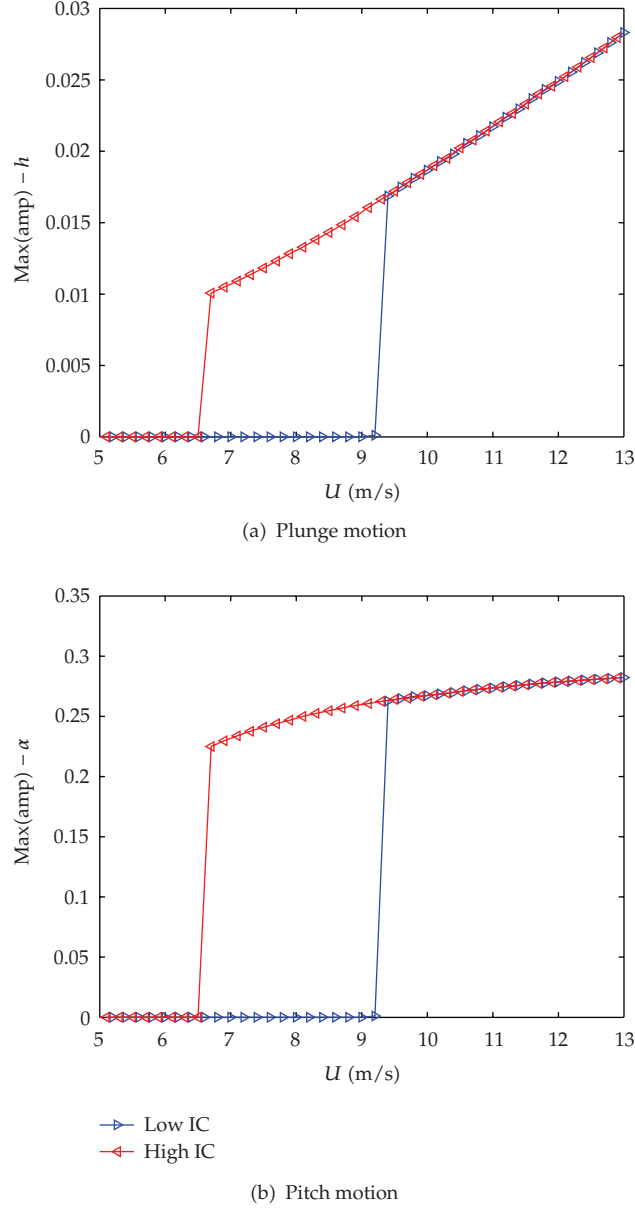
Figure 3: Hysteretic response of the aeroelastic system (subcritical instability). The steady-state amplitudes are plotted as a function of U.

This equation shows that applying gain to the plunge displacement is more effective than applying it to the pitch displacement or plunge velocity or pitch velocity.

The subcritical instability takes place for positive values of $\widehat{\Lambda}_r$. As such it can be eliminated by forcing $\widehat{\Lambda}_r$ to be negative. This can be achieved by using nonlinear control gains $k_{\mathrm{nl1}} = 0.02$ ($k_{\mathrm{nl2}} = k_{\mathrm{nl3}} = k_{\mathrm{nl4}} = 0$). The results are presented in Figure 5. The subcritical Hopf bifurcation at U_f observed in Figure 3 has been transformed into the supercritical Hopf bifurcation of Figure 5. A comparison of the two figures shows that the unstable limit

(a) Plunge motion

(b) Pitch motion

Figure 4: Measured pitch and plunge responses:: −, without linear control, ∗, with linear control.

(a) Plunge motion
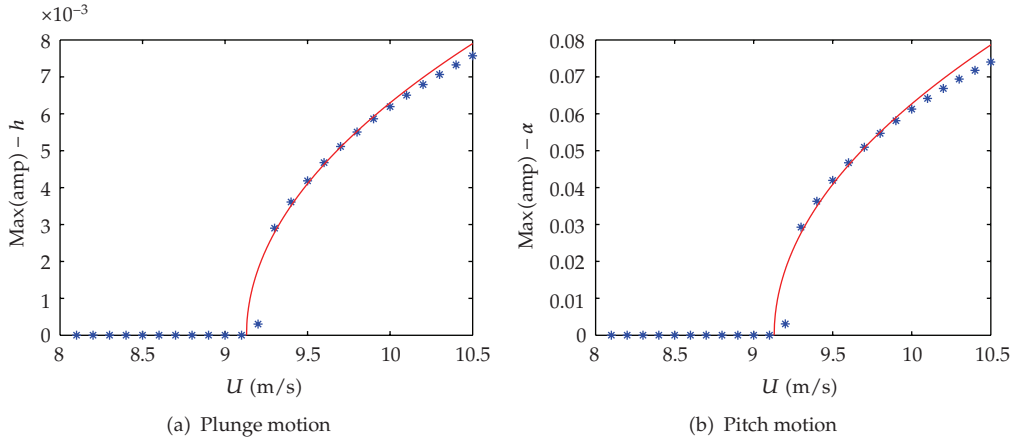
(b) Pitch motion
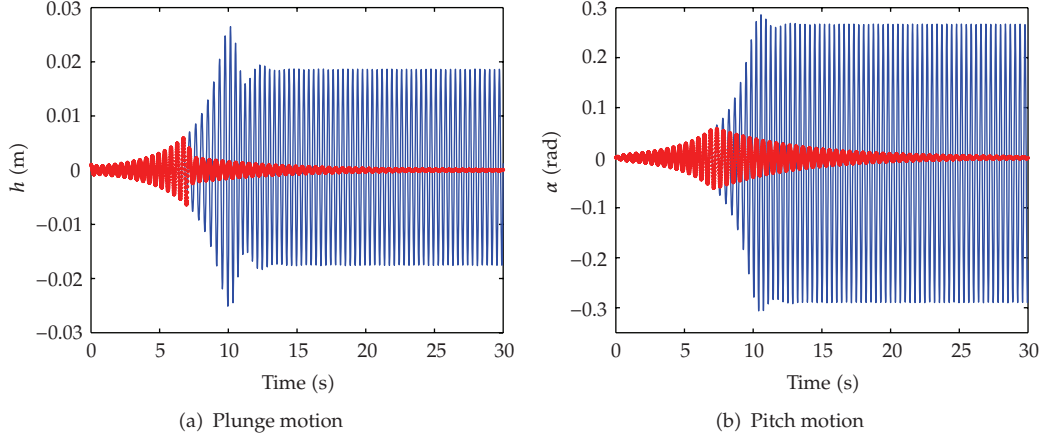
Figure 5: LCO amplitudes of plunge and pitch motions (controlled configuration): −, analytical prediction, ∗, numerical integration.

cycles observed over the freestream speed between $U = 6.5\,\text{m/s}$ and $U = 9.12\,\text{m/s}$ have been eliminated. Furthermore, the exponentially growing oscillations predicted by the linear model are limited to a periodic solution whose amplitude increases slowly with increasing freestream velocity. Moreover, increasing the value of k_{nl1} reduces the amplitude of the limit cycles created due to Hopf bifurcation.

In order to check the accuracy of the analytical formulation given by the normal form in predicting the amplitude of LCO near the Hopf bifurcation, we consider the first-order solution given by (3.9). The amplitude of plunge and pitch LCO, A_h and A_α, respectively, are given by

$$A_h = a\sqrt{\mathbf{p1}_r^2 + \mathbf{p1}_i^2},$$

$$A_\alpha = a\sqrt{\mathbf{p2}_r^2 + \mathbf{p2}_i^2},$$

(3.27)

where $\mathbf{pj}_r$ and $\mathbf{pj}_i$ denote the real and imaginary parts of the jth component of the vector $\mathbf{p}$, respectively. In Figures 5(a) and 5(b), we plot the LCO amplitudes for both pitch and plunge motions obtained by integrating the original system and those predicted by the normal form. The results show good agreement in the LCO amplitudes only near the bifurcation.

4. Conclusions

Linear and nonlinear controls are implemented on a rigid airfoil undergoing pitch and plunge motions. The method of multiple scales is applied to the governing system of equations to derive the normal form of the Hopf bifurcation near the flutter onset. The linear and nonlinear parameters of the normal form are used to determine the stability characteristics of the bifurcation and efficiency of the linear and nonlinear control components. The results show that linear control can be used to delay the flutter onset and dampen LCO. Yet, its required gains remain a function of the speed. On the other hand, nonlinear control can be efficiently implemented to convert subcritical to supercritical Hopf bifurcations and to significantly reduce LCO amplitudes.

Acknowledgment

References

[1] E. H. Dowell and D. Tang, "Nonlinear aeroelasticity and unsteady aerodynamics," *AIAA Journal*, vol. 40, no. 9, pp. 1697–1707, 2002.

[2] A. Raghothama and S. Narayanan, "Non-linear dynamics of a two-dimensional air foil by incremental harmonic balance method," *Journal of Sound and Vibration*, vol. 226, no. 3, pp. 493–517, 1999.

[3] L. Liu, Y. S. Wong, and B. H. K. Lee, "Application of the centre manifold theory in non-linear aeroelasticity," *Journal of Sound and Vibration*, vol. 234, no. 4, pp. 641–659, 2000.

[4] H. C. Gilliatt, T. W. Strganac, and A. J. Kurdila, "An investigation of internal resonance in aeroelastic systems," *Nonlinear Dynamics*, vol. 31, no. 1, pp. 1–22, 2003.

[5] B. H. K. Lee, L. Y. Jiang, and Y. S. Wong, "Flutter of an airfoil with a cubic restoring force," *Journal of Fluids and Structures*, vol. 13, no. 1, pp. 75–101, 1999.

[6] C. C. Chabalko, M. R. Hajj, D. T. Mook, and W. A. Silva, "Characterization of the LCO response behaviors of the NATA model," in *Proceedings of the 47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, vol. 5, pp. 3196–3206, Newport, RI, USA, May 2006, AIAA paper no.2006-1852.

[7] E. H. Abed and J.-H. Fu, "Local feedback stabilization and bifurcation control. I. Hopf bifurcation," *Systems & Control Letters*, vol. 7, no. 1, pp. 11–17, 1986.

[8] E. H. Abed and J.-H. Fu, "Local feedback stabilization and bifurcation control. II. Stationary bifurcation," *Systems & Control Letters*, vol. 8, no. 5, pp. 467–473, 1987.

[9] T. W. Strganac, J. Ko, D. E. Thompson, and A. J. Kurdila, "Identification and control of limit cycle oscillations in aeroelastic systems," in *Proceedings of the 40th AIAA/ASME/ASCE/AHS/ASC Structrures, Structural Dynamics, and Materials Conference and Exhibit*, vol. 3, pp. 2173–2183, St. Louis, Mo, USA, April 1999, AIAA paper no. 99-1463.

[10] L. Librescu, S. Na, P. Marzocca, C. Chung, and M. K. Kwak, "Active aeroelastic control of 2-D wing-flap systems operating in an incompressible flowfield and impacted by a blast pulse," *Journal of Sound and Vibration*, vol. 283, no. 3–5, pp. 685–706, 2005.

[11] W. Kang, "Bifurcation control via state feedback for systems with a single uncontrollable mode," *SIAM Journal on Control and Optimization*, vol. 38, no. 5, pp. 1428–1452, 2000.

[12] A. H. Nayfeh and B. Balachandran, *Applied Nonlinear Dynamics. Analytical, Computational, and Experimental Methods*, Wiley Series in Nonlinear Science, John Wiley & Sons, New York, NY, USA, 1995.

Research Article

Numerical Investigation of Aeroelastic Mode Distribution for Aircraft Wing Model in Subsonic Air Flow

Marianna A. Shubov, Stephen Wineberg, and Robert Holt

Department of Mathematics and Statistics, University of New Hampshire, Durham, NH 03824, USA

Correspondence should be addressed to Marianna A. Shubov, marianna.shubov@euclid.unh.edu

In this paper, the numerical results on two problems originated in aircraft wing modeling have been presented. *The first problem* is concerned with the approximation to the set of the aeroelastic modes, which are the eigenvalues of a certain boundary-value problem. The affirmative answer is given to the following question: can the leading asymptotical terms in the analytical formulas be used as reasonably accurate description of the aeroelastic modes? The positive answer means that these leading terms can be used by engineers for practical calculations. *The second problem* is concerned with the flutter phenomena in aircraft wings in a subsonic, incompressible, inviscid air flow. It has been shown numerically that there exists a pair of the aeroelastic modes whose behavior depends on a speed of an air flow. Namely, when the speed increases, the distance between the modes tends to zero, and at some speed that can be treated as the flutter speed these two modes merge into one double mode.

1. Introduction and Formulation of Problem

The main goal of this paper is to present the numerical results for two problems arising in aircraft wing modeling. *The first one* is related to the numerical approximation of the discrete spectrum of a certain matrix differential operator that represents the structural part of an aircraft wing model in a *subsonic incompressible inviscid* air flow. More precisely, it is shown that the asymptotic approximation for the set of aeroelastic modes has a very regular structure, that is, there are two distinct branches which are called the *β-branch* and the *δ-branch* (see formulas (2.2), (2.3)). Each branch is described by leading asymptotical terms and remainder terms. An affirmative answer is given to the following question: can the leading asymptotical terms be used as a reasonably accurate description of the aeroelastic modes? The question can be rephrased as: if one discretizes the main matrix differential

operator with spectral methods, can the spectrum of the discretized finite-dimensional operator be described by the leading asymptotical terms of the continuous operator? The answer is affirmative, and one can see almost perfect spectral branches for the finite-dimensional approximation (see Figure 2).

The second problem is related to the entire model involving both the structural and aerodynamical parts. In the second problem, the authors provide a numerical justification of a certain fact well-known in the aircraft engineering community. Namely, it is shown that there exists a pair of the aeroelastic modes whose behavior depends on a speed of an air flow in the following way: the distance between these aeroelastic modes tends to zero with increasing speed. At some speed that can be treated as "the flutter speed", the two modes merge in one double mode. If a speed continues to increase, the double aeroelastic mode splits up into two simple modes that are moving apart. Such a dynamics clearly indicates that there exists a specific speed ("the flutter speed") that causes an appearance of a double aeroelastic mode leading to a flutter instability. In this study, the authors deal with a linear wing model, which is obviously a significant simplification of a realistic situation. In a more precise setting, when the speed is near what is called "the flutter speed", one should use a nonlinear wing model, which results in limit cycle oscillations rather than flutter. One of the important conclusions of this investigation is that even though the model studied in the paper is linear, it is capable of capturing the flutter instability, which is definitely supported by numerical simulations. Before the numerical analysis results, a mathematical statement of the problem and a description of those analytical results that are relevant to the numerical analysis will be given.

An ultimate goal of an aircraft wing modeling is to find an approach to flutter control. Flutter is a structural dynamical instability, which consists of violent vibrations of the solid structure with rapidly increasing amplitude [1, 2]. The physical reason for this phenomenon is that under special conditions, the energy of air flow is rapidly absorbed by the structure and transformed into the energy of mechanical vibrations. In engineering practice, flutter must be avoided either by design of the structure or by introducing control mechanism capable to suppress harmful vibrations. Flutter is an inherent feature of fluid-structure interaction and, thus, it cannot be eliminated completely. However, the critical conditions for the flutter onset can be shifted to the safe range of the operating parameters. This is exactly the goal for designing flutter control mechanisms.

Flutter is an in-flight event, which happens beyond some speed-altitude combinations. High-speed aircrafts are most susceptible for flutter although no speed regime is truly immune from flutter. Flutter instabilities occur in a variety of different engineering and even biomedical situations. Namely in aeronautic engineering, flutter of helicopter, propeller, and turbine blades is a serious problem. It also affects electric transmission lines, high-speed hard disk drives, and long-spanned suspension bridges. Flutter of cardiac tissue and blood vessel walls is of a special concern to medical practitioners.

At the present moment, there exist only a few models of fluid-structure interaction involving flutter development, for which precise mathematical formulations are available. The authors' main objects of interest are analytical and numerical results on such models, which can be used, in particular, for flutter explanation. It is certainly important for designing flutter control mechanisms.

Ideally, a complete picture of a fluid-structure interaction should be described by a system of partial differential equations, a system which contains both the equations governing the vibrations of an elastic structure and the hydrodynamic equations governing the motion of gas or fluid flow. The system of equations of motion should be supplied
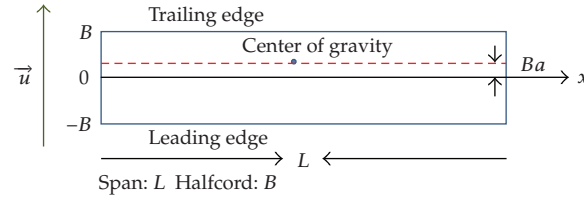
Figure 1: Wing structure beam model.

with appropriate boundary and initial conditions. The structural and hydrodynamic parts of the system must be coupled in the following sense. The hydrodynamic equations define a pressure distribution on the elastic structure. This pressure distribution in turn defines the so-called aerodynamic loads, which appear as forcing terms in structural equations. On the other hand, the parameters of the elastic structure enter the boundary conditions for the hydrodynamic equations.

In the present paper it is assumed that the model describes a wing of high-aspect ratio (i.e., the length of a wing is much greater than its width, though both quantities are finite) in a *subsonic, inviscid, incompressible* air flow. The hydrodynamic equations have been solved explicitly and aerodynamic loads are represented as forcing terms in the structural equations as time convolution-type integrals with complicated kernels. Thus, the model is described by a system of integro-differential equations. Analytical results on this model include the following. The system of equations of motion is treated as a single evolution-convolution equation in the Hilbert state space of the model. (The integral convolution part of this equation vanishes if a speed of an air stream is zero, and the equations of motion describe the so-called ground vibrations.) After applying a Laplace transformation with respect to the time variable to both sides of the evolution equation, one obtains a matrix differential equation involving the complex parameter λ. For this new equation, the following results have been shown.

(a) The representation of the solution of the original initial boundary-value problem in the frequency domain has been given in terms of *the generalized resolvent*, which is an analytic operator-valued function of the spectral parameter λ. The aeroelastic modes are defined as the poles of the generalized resolvent; the corresponding mode shapes are defined in terms of the residues at these poles [3–5].

(b) Explicit asymptotic formulas for the aeroelastic modes have been derived [4, 6]. (To the best of the authors' knowledge, these are the first such formulas in the literature on aeroelasticity.) The entire set of aeroelastic modes splits into two branches, which are asymptotically close to the eigenvalues of the structural part of the system [3, 5, 7].

(c) It has been shown that the set of the mode shapes forms a nonorthogonal basis (Riesz basis) of the state space of the system. (The set of the generalized eigenvectors of the structural part of the system has a similar property [4, 8].)

1.1. Statement of Problem. Operator Setting in Energy Space

The so-called "Goland model" [1, 6, 9] is considered, that is, the simplest structural model— a uniform rectangular beam (Figure 1) with two types of motion, plunge and pitch. To

formulate a mathematical model, the following dynamical variables are introduced [3, 4, 9, 10]:

$$X(x,t) = \begin{pmatrix} h(x,t) \\ \alpha(x,t) \end{pmatrix}, \quad -L \le x \le 0, \ t \ge 0, \tag{1.1}$$

where $h(x,t)$ is the bending, $\alpha(x,t)$ is the torsion angle, and x is the span variable. The model can be described by the following linear system:

$$(M_s - M_a)\ddot{X}(x,t) + (D_s - uD_a)\dot{X}(x,t) + \left(K_s - u^2 K_a\right)X(x,t) = \begin{bmatrix} f_1(x,t) \\ f_2(x,t) \end{bmatrix}, \tag{1.2}$$

where the "overdot" denotes the differentiation with respect to time t. The subscripts "s" and "a" are use to distinguish the structural and aerodynamical parameters, respectively. All 2×2 matrices in (1.2) are given by the following formulas:

$$M_s = \begin{bmatrix} m & S \\ S & I \end{bmatrix}, \quad M_a = (-\pi\rho)\begin{bmatrix} 1 & -a \\ -a & \left(a^2 + \dfrac{1}{8}\right) \end{bmatrix}, \quad D_s = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \tag{1.3}$$

where m is the density of the flexible structure, S is the mass moment, I is the moment of inertia, ρ is the density of air, a is the linear parameter of the structure, and $-1 \le a \le 1$ (a is a relative distance between the elastic axis of a model wing and its line of center of gravity). Even though in the current paper the matrix D_s has only trivial entries, it is preferable to keep it in (1.2) since the problem with a nontrivial structural damping D_s will be studied in the authors' forthcoming paper:

$$D_a = (-\pi\rho)\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad K_s = \begin{bmatrix} E\dfrac{\partial^4}{\partial x^4} & 0 \\ 0 & -G\dfrac{\partial^2}{\partial x^2} \end{bmatrix}, \quad K_a = (-\pi\rho)\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}, \tag{1.4}$$

where E is the bending stiffness, and G is the torsion stiffness. The parameter u in (1.2) denotes the stream speed. The right hand side of system (1.2) can be represented as

the following system of two convolution-type integral operations:

$$f_1(x,t) = -2\pi\rho \int_0^t \left[uC_2(t-\sigma) - \dot{C}_3(t-\sigma)\right]g(x,\sigma)d\sigma \equiv \int_0^t \tilde{C}_1(t-\sigma)g(x,\sigma)d\sigma, \tag{1.5}$$

$$f_2(x,t) = -2\pi\rho \int_0^t \left[\frac{1}{2}C_1(t-\sigma) - auC_2(t-\sigma) + a\dot{C}_3(t-\sigma)\right.$$

$$\left. + uC_4(t-\sigma) + \frac{1}{2}\dot{C}_5(t-\sigma)\right]g(x,\sigma)d\sigma \equiv \int_0^t \tilde{C}_2(t-\sigma)g(x,\sigma)d\sigma, \tag{1.6}$$

$$g(x,t) = u\dot{\alpha}(x,t) + \ddot{h}(x,t) + \left(\frac{1}{2} - a\right)\ddot{\alpha}(x,t). \tag{1.7}$$

Explicit formulas for the aerodynamical functions C_i, $i = 1,\ldots,5$, can be found in [11]. It is known that the self-straining control actuator action can be modeled by the following boundary conditions [5, 9, 12–14]:

$$Eh''(0,t) + \beta\dot{h}'(0,t) = 0, \qquad h'''(0,t) = 0, \qquad G\alpha'(0,t) + \delta\dot{\alpha}(0,t) = 0, \qquad \beta,\delta \in \mathbb{C}^+ \cup \{\infty\}, \tag{1.8}$$

where $\mathbb{C}^+$ is the closed right half-plane. The boundary conditions at $x = -L$ are

$$h(-L,t) = h'(-L,t) = \alpha(-L,t) = 0. \tag{1.9}$$

In (1.8) and (1.9) and below, the prime designates derivative with respect to x. The initial state of the system is given as follows:

$$h(x,0) = h_0(x), \qquad \dot{h}(x,0) = h_1(x), \qquad \alpha(x,0) = \alpha_0(x), \qquad \dot{\alpha}(x,0) = \alpha_1(x). \tag{1.10}$$

The solution of the problem given by (1.2) and conditions (1.8)–(1.10) are given in the energy space $\mathcal{H}$. It is assumed that the parameters satisfy the following two conditions:

$$\det\begin{bmatrix} m & S \\ S & I \end{bmatrix} > 0, \qquad 0 < u \leq \frac{\sqrt{2G}}{L\sqrt{\pi\rho}}. \tag{1.11}$$

The second condition in (1.11) means that the flow speed must be below the "divergence" or static aeroelastic instability speed. The state space $\mathcal{H}$, which is a Hilbert space, is described as the set of 4-component vector-valued functions $\Psi = (h, \dot{h}, \alpha, \dot{\alpha})^T \equiv (\psi_0, \psi_1, \psi_2, \psi_3)^T$ ("T" stands for transposition) obtained as a closure of smooth functions satisfying the boundary conditions

$$\psi_0(-L) = \psi_0'(-L) = \psi_2(-L) = 0 \tag{1.12}$$

in the following energy norm, which is well-defined under conditions (1.11):

$$
\|\Psi\|_{\mathcal{H}}^2 = \frac{1}{2}\int_{-L}^0 \Big[E|\psi_0''(x)|^2 + G|\psi_2'(x)|^2 + \tilde{m}|\psi_1(x)|^2 + \tilde{I}|\psi_3(x)|^2
$$
$$
+ \tilde{S}(\psi_3(x)\overline{\psi}_1(x) + \overline{\psi}_3(x)\psi_1(x)) - \pi\rho u^2|\psi_2(x)|^2\Big]dx,
$$
(1.13)

where the following notations have been used:

$$
\tilde{m} = m + \pi\rho, \qquad \tilde{S} = S - a\pi\rho, \qquad \tilde{I} = I + \pi\rho\Big(a^2 + \frac{1}{8}\Big), \qquad \Delta = \tilde{m}\tilde{I} - \tilde{S}^2.
$$
(1.14)

The initial-boundary value problem defined by (1.2) and conditions (1.8)–(1.10) can be represented in the following form

$$
\dot{\Psi} = i\mathcal{L}_{\beta\delta}\Psi + \tilde{\mathcal{F}}\Psi, \qquad \Psi = (\psi_0,\psi_1,\psi_2,\psi_3)^T, \qquad \Psi|_{t=0} = \Psi_0.
$$
(1.15)

$\mathcal{L}_{\beta\delta}$ is a matrix differential operator in $\mathcal{H}$ given by the following expression:

$$
\mathcal{L}_{\beta\delta} = -i\begin{bmatrix} 0 & 1 & 0 & 0 \\[2mm] -\dfrac{E\tilde{I}}{\Delta}\dfrac{d^4}{dx^4} & -\dfrac{\pi\rho u\tilde{S}}{\Delta} & -\dfrac{\tilde{S}}{\Delta}\Big(G\dfrac{d^2}{dx^2}+\pi\rho u^2\Big) & -\dfrac{\pi\rho u\tilde{I}}{\Delta} \\[2mm] 0 & 0 & 0 & 1 \\[2mm] \dfrac{E\tilde{S}}{\Delta}\dfrac{d^4}{dx^4} & \dfrac{\pi\rho u\tilde{m}}{\Delta} & \dfrac{\tilde{m}}{\Delta}\Big(G\dfrac{d^2}{dx^2}+\pi\rho u^2\Big) & \dfrac{\pi\rho u\tilde{S}}{\Delta} \end{bmatrix}
$$
(1.16)

defined on the domain

$$
\mathfrak{D}(\mathcal{L}_{\beta\delta}) = \Big\{\Psi\in\mathcal{H}: \psi_0\in H^4(-L,0),\ \psi_1\in H^2(-L,0),\ \psi_2\in H^2(-L,0),
$$
$$
\psi_3\in H^1(-L,0);\ \psi_1(-L)=\psi_1'(-L)=\psi_3(-L)=0;\ \psi_0'''(0)=0;
$$
(1.17)
$$
E\psi_0''(0)+\beta\psi_1'(0)=0,\ G\psi_2'(0)+\delta\psi_3(0)=0\Big\}.
$$

$\tilde{\mathcal{F}}$ is a linear integral operator in $\mathcal{H}$ given by the formula

$$
\tilde{\mathcal{F}} = \frac{1}{\Delta}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & [\tilde{I}(\tilde{C}_1*)-\tilde{S}(\tilde{C}_2*)] & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & [-\tilde{S}(\tilde{C}_1*)+\tilde{m}(\tilde{C}_2*)] \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & u & \big(\frac{1}{2}-a\big) \\ 0 & 0 & 0 & 0 \\ 0 & 1 & u & \big(\frac{1}{2}-a\big) \end{bmatrix}.
$$
(1.18)

In (1.18), the star "$*$" stands for the convolution operation and the kernels $\tilde{C}_1$ and $\tilde{C}_2$ are defined in (1.5), and (1.6).

Remark 1.1. It is important to emphasize that (1.15) is not an *evolution equation*. It does not have a dynamics generator and does not define any semigroup in the standard sense. By applying the Laplace transformation to both sides of (1.15), one obtains the following Laplace transform representation of the solution:

$$\widehat{\Psi}(\lambda) = -\left(\lambda I - i\mathcal{L}_{\beta\delta} - \lambda\widehat{\mathfrak{F}}(\lambda)\right)^{-1}\left(I - \widehat{\mathfrak{F}}(\lambda)\right)\Psi_0. \tag{1.19}$$

To find the solution in the space-time domain, one has to "calculate" the inverse Laplace transform of $\widehat{\Psi}$ by contour integration in the complex λ-plane. In this connection, the properties of the "generalized resolvent operator"

$$\mathcal{R}(\lambda) = \left(\lambda I - i\mathcal{L}_{\beta\delta} - \lambda\widehat{\mathfrak{F}}(\lambda)\right)^{-1} \tag{1.20}$$

are of crucial importance. It has been proved that $\mathcal{R}(\lambda)$ has a countable set of poles which we called the eigenvalues or *the aeroelastic modes*. The residues of $\mathcal{R}(\lambda)$ at these poles are precisely the projectors on the corresponding generalized eigenspaces. The differential part and the role of the convolution part of the system have been analyzed in [3–8, 15]. In particular, it has been shown that the convolution part does not "destroy" the main characteristics of the discrete spectrum, which is produced by the differential part of the problem. Namely, it has been proved that the aeroelastic modes are asymptotically close to the discrete spectrum of the operator $i\mathcal{L}_{\beta\delta}$, and the rate at which the aeroelastic modes approach that spectrum has been calculated.

2. Asymptotic and Spectral Results on Matrix Differential and Integral Operators

2.1. Matrix Differential Operator $\mathcal{L}_{\beta\delta}$

Theorem 2.1. (a) $\mathcal{L}_{\beta\delta}$ *is a closed linear operator in* $\mathcal{H}$*, whose resolvent is compact, and therefore, the spectrum is discrete [3, 4, 16].*

(b) *Operator* $\mathcal{L}_{\beta\delta}$ *is nonselfadjoint unless* $\mathfrak{R}\,\beta = \mathfrak{R}\,\delta = 0$. *If* $\mathfrak{R}\,\beta \geq 0$ *and* $\mathfrak{R}\,\delta \geq 0$, *then this operator is dissipative, that is,* $\mathfrak{I}(\mathcal{L}_{\beta\delta}\Psi, \Psi) \geq 0$ *for* $\Psi \in \mathfrak{D}(\mathcal{L}_{\beta\delta})$. *The adjoint operator* $\mathcal{L}_{\beta\delta}^*$ *is given by the matrix differential expression* (1.16) *on the domain obtained from* (1.17) *by replacing the parameters* β *and* δ *with* $(-\bar{\beta})$ *and* $(-\bar{\delta})$, *respectively.*

Theorem 2.2. (a) *The operator* $\mathcal{L}_{\beta\delta}$ *has a countable set of complex eigenvalues. If*

$$\delta \neq \sqrt{G\widetilde{I}}, \tag{2.1}$$

then the set of eigenvalues is located in a strip parallel to the real axis.

(b) *The entire set of eigenvalues asymptotically splits into two different subsets. We call them the* β-*branch and the* δ-*branch and denote these branches by* $\{v_n^{\beta}\}_{n\in\mathbb{Z}}$ *and* $\{v_n^{\delta}\}_{n\in\mathbb{Z}}$, *respectively.*

If $\Re \beta \geq 0$ and $\Re \delta > 0$, then each branch is asymptotically close to its own horizontal line in the closed upper half-plane. If $\Re \beta > 0$ and $\Re \delta = 0$, then both horizontal lines coincide with the real axis. If $\Re \beta = \Re \delta = 0$, then the operator $\mathcal{L}_{\beta\delta}$ is selfadjoint and, thus, its spectrum is real.

(c) *The following asymptotics are valid for the β-branch of the spectrum as $|n| \to \infty$:*

$$\nu_n^\beta = (\operatorname{sgn} n)\left(\frac{\pi^2}{L^2}\right)\sqrt{\frac{E\widetilde{I}}{\Delta}}\left(|n| - \frac{1}{4}\right)^2 + \kappa_n(\omega), \quad \omega = |\delta|^{-1} + |\beta|^{-1}, \tag{2.2}$$

with Δ being defined in (1.14). *A complex-valued sequence $\{\kappa_n\}$ is bounded above in the following sense:* $\sup_{n \in \mathbb{Z}}\{|\kappa_n(\omega)|\} = C(\omega), \quad C(\omega) \to 0 \ as \ \omega \to 0.$

(d) *The following asymptotics are valid for the δ-branch of the spectrum:*

$$\nu_n^\delta = \frac{\pi n}{L\sqrt{\widetilde{I}/G}} + \frac{i}{2L\sqrt{\widetilde{I}/G}}\ln\frac{\delta + \sqrt{G\widetilde{I}}}{\delta - \sqrt{G\widetilde{I}}} + O\left(|n|^{-1/2}\right), \quad |n| \longrightarrow \infty. \tag{2.3}$$

In (2.3), *"ln" means the principal value of the logarithm. If β and δ stay away from zero, that is, $|\beta| \geq \beta_0 > 0$ and $|\delta| \geq \delta_0 > 0$, then the estimate $O(|n|^{-1/2})$ in* (2.3) *is uniform with respect to both parameters.*

(e) *There may be only a finite number of multiple eigenvalues of a finite multiplicity each. Therefore, only a finite number of the associate vectors may exist.*

The theorem below deals with the Riesz basis property of the generalized eigenvectors of the structural operator $\mathcal{L}_{\beta\delta}$. The Riesz basis is a mild modification of an orthonormal basis, namely a linear isomorphism of an orthonormal basis.

Theorem 2.3. *The set of generalized eigenvectors of the operator $\mathcal{L}_{\beta\delta}$ forms a Riesz basis in the energy space $\mathcal{H}$.*

The next statement deals with the asymptotic distribution of the aeroelastic modes [4, 7, 15].

Theorem 2.4. (a) *The set of the aeroelastic modes (which are the poles of the generalized resolvent operator) is countable and does not have accumulation points on the complex plane $\mathbb{C}$. There might be only a finite number of multiple poles of a finite multiplicity each. There exists a sufficiently large $R > 0$ such that all aeroelastic modes, whose distance from the origin is greater than R, are simple poles of the generalized resolvent. The value of R depends on the speed u of an airstream, that is, $R = R(u)$.*

(b) *The set of the aeroelastic modes splits asymptotically into two series, which we call the β-branch and the δ-branch. Asymptotical distribution of the β-and the δ-branches of the aeroelastic modes can be obtained from asymptotical distribution of the spectrum of the operator $\mathcal{L}_{\beta\delta}$. Namely if $\{\lambda_n^\beta\}_{n \in \mathbb{Z}}$ is the β-branch of the aeroelastic modes, then $\lambda_n^\beta = i\widehat{\lambda}_n^\beta$ and the asymptotics of the set $\{\widehat{\lambda}_n^\beta\}_{n \in \mathbb{Z}}$ is given by the right-hand side of formula* (2.2). *Similarly, if $\{\lambda_n^\delta = i\widehat{\lambda}_n^\delta\}_{n \in \mathbb{Z}}$ is the δ-branch of the aeroelastic modes, then the asymptotical distribution of the set $\{\widehat{\lambda}_n^\delta\}_{n \in \mathbb{Z}}$ is given by the right-hand side of formula* (2.3).

2.2. Structure and Properties of the Matrix Integral Operator

Important information on the Laplace transform of the convolution-type matrix integral operator (1.18) is collected below.

Lemma 2.5. *Let $\widehat{\mathfrak{F}}$ be the Laplace transform of the kernel of matrix integral operator* (1.18). *The following formula is valid for $\widehat{\mathfrak{F}}$:*

$$
\widehat{\mathfrak{F}}(\lambda) = \begin{bmatrix} 0 & 0 & 0 & 0 \\[2mm] 0 & \mathcal{L}(\lambda) & u\mathcal{L} & \left(\frac{1}{2} - a\right)\mathcal{L}(\lambda) \\[2mm] 0 & 0 & 0 & 0 \\[2mm] 0 & \mathcal{N}(\lambda) & u\mathcal{N}(\lambda) & \left(\frac{1}{2} - a\right)\mathcal{N}(\lambda) \end{bmatrix},
\tag{2.4}
$$

where

$$
\mathcal{L}(\lambda) = -\frac{2\pi\rho}{\Delta}\frac{u}{\lambda}\left\{-\frac{\widetilde{S}}{2} + \left[\widetilde{I} + \left(a + \frac{1}{2}\right)\widetilde{S}\right]T\left(\frac{\lambda}{u}\right)\right\},
$$
$$
\mathcal{N}(\lambda) = -\frac{2\pi\rho}{\Delta}\frac{u}{\lambda}\left\{\frac{\widetilde{m}}{2} - \left[\widetilde{S} + \left(a + \frac{1}{2}\right)\widetilde{m}\right]T\left(\frac{\lambda}{u}\right)\right\},
\tag{2.5}
$$

and T is the Theodorsen function defined by the formula

$$
T(z) = \frac{K_1(z)}{K_0(z) + K_1(z)};
\tag{2.6}
$$

K_0 and K_1 are the modified Bessel functions [17].

2.3. Properties of the Theodorsen Function

The Theodorsen function is a bounded analytic function on the complex plane with the branch-cut along the negative real semi-axis. As $|z| \to \infty$, the following asymptotic representation holds:

$$
T(z) = \frac{1}{2} + O\left(\frac{1}{1+|z|}\right), \quad \text{as } |z| \longrightarrow \infty.
\tag{2.7}
$$

Taking into account that $z = \lambda/u$, one can write $\lambda \widehat{\mathfrak{F}}(\lambda)$ as the following sum: $\lambda \widehat{\mathfrak{F}}(\lambda) = \mathfrak{M} + \mathfrak{N}(\lambda)$, where the matrix $\mathfrak{M}$ is defined by the formula

$$
\mathfrak{M} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & A & uA & \left(\dfrac{1}{2} - a\right)A \\ 0 & 0 & 0 & 0 \\ 0 & B & uB & \left(\dfrac{1}{2} - a\right)B \end{bmatrix}
\tag{2.8}
$$

with A and B being given by

$$
A = -\pi \rho u \Delta^{-1} \left[\tilde{I} + \left(a - \frac{1}{2}\right)\tilde{S}\right], \qquad B = \pi \rho u \Delta^{-1}\left[\tilde{S} + \left(a - \frac{1}{2}\right)\tilde{m}\right].
\tag{2.9}
$$

The matrix-valued function $\mathfrak{N}(\lambda)$ is defined by the formula

$$
\mathfrak{N}(\lambda) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & A_1(\lambda) & uA_1(\lambda) & \left(\dfrac{1}{2} - a\right)A_1(\lambda) \\ 0 & 0 & 0 & 0 \\ 0 & B_1(\lambda) & uB_1(\lambda) & \left(\dfrac{1}{2} - a\right)B_1(\lambda) \end{bmatrix},
\tag{2.10}
$$

where $A_1(\lambda) = -2\pi \rho u \Delta^{-1}(T(\lambda/u) - 1/2)[\tilde{I} + (a + 1/2)\tilde{S}]$, and $B_1(\lambda) = 2\pi \rho u \Delta^{-1}(T(\lambda/u) - 1/2)[\tilde{S} + (a+1/2)\tilde{m}]$. For each λ, $\mathfrak{N}(\lambda)$ is a bounded operator in $\mathcal{H}$ with the following estimate for its norm:

$$
\|\mathfrak{N}\|_{\mathcal{H}} \leq C(1 + |\lambda|)^{-1},
\tag{2.11}
$$

where C is an absolute constant the precise value of which is immaterial for us. Therefore, the generalized resolvent (1.20) can be written in the form

$$
\mathcal{R}(\lambda) = \mathfrak{S}^{-1}(\lambda), \quad \text{where } \mathfrak{S}(\lambda) = \lambda I - i\mathcal{L}_{\beta\delta} - \mathfrak{M} - \mathfrak{N}(\lambda).
\tag{2.12}
$$

Theorem 2.6. $\mathfrak{M}$ *is a bounded linear operator in $\mathcal{H}$. The operator $\mathcal{K}_{\beta\delta}$ defined by*

$$
\mathcal{K}_{\beta\delta} = \mathcal{L}_{\beta\delta} - i\,\mathfrak{M}
\tag{2.13}
$$

is an unbounded nonselfadjoint operator in $\mathcal{H}$ with compact resolvent. The spectral asymptotics of $\mathcal{K}_{\beta\delta}$ coincide with the spectral asymptotics of $\mathcal{L}_{\beta\delta}$ (see Theorem 2.2). In contrast to $\mathcal{L}_{\beta\delta}$, the operator $\mathcal{K}_{\beta\delta}$ is not dissipative for any boundary control gains. However, $\mathcal{K}_{\beta\delta}$ is also a Riesz spectral operator, that is, the set of its generalized eigenvectors forms a Riesz basis of $\mathcal{H}$.

3. Numerical Results for Two-Branch Discrete Spectrum of Operator $\mathcal{L}_{\beta\delta}$

The first question is related to the accuracy of the asymptotic approximations of the eigenvalues of the operator $\mathcal{L}_{\beta\delta}$. By their nature, asymptotic formulas (2.2) and (2.3) should be understood in the following way.

Formula (2.2) for the β-branch eigenvalues means that there exist a positive number N_1 and a small constant $0 < \varepsilon \ll 1$ such that for all $|n| \geq N_1$, the β-branch eigenvalues (2.2) satisfy the estimate

$$\left| v_n^\beta - (\operatorname{sgn} n)\left(\frac{\pi^2}{L^2}\right)\sqrt{\frac{E\widetilde{I}}{\Delta}}\left(|n| - \frac{1}{4}\right)^2 \right| \leq \varepsilon. \tag{3.1}$$

In other words, for $|n| \geq N_1$, all eigenvalues $\{v_n^\beta\}$ are located in the ε-vicinity of the points $\{\overset{\circ}{v}_n^\beta\}$, given by the leading asymtotical term of (2.2). Obviously, ε can be chosen as small as desired by manipulating the control parameters β and δ.

Formula (2.3) for the δ-branch means that there exists $N_2 > 0$ and $\widetilde{\varepsilon} > 0$ such that for all $|n| \geq N_2$ the δ-branch eigenvalues satisfy the estimate

$$\left| v_n^\delta - \frac{\pi n}{L\sqrt{\widetilde{I}/G}} - \frac{\imath}{2L\sqrt{\widetilde{I}/G}}\ \ln\frac{\delta + \sqrt{G\widetilde{I}}}{\delta - \sqrt{G\widetilde{I}}} \right| \leq \frac{\widetilde{\varepsilon}}{\sqrt{|n|}}. \tag{3.2}$$

This formula means that for $|n| \geq N_2$ each eigenvalue v_n^δ is located in a small circle of radius that tends to zero at the rate $|n|^{-1/2}$ and is centered at the points $\{\overset{\circ}{v}_n^\beta\}$; each center coincides with the leading asymptotical term from (2.3).

Regarding this description, the following important question holds: from which numbers N_1 and N_2 can the eigenvalues be approximated by the leading asymptotical terms with acceptable accuracy? In other words, can one claim that asymptotical formulas (2.2) and (2.3) are valuable for practitioners or are they just important analytical results of the spectral analysis?

As is well known, from practical applications only the first dozen of the lowest eigenfrequencies are important for engineers. The results of numerical simulations below show that the asymtotical formulas are indeed quite accurate, that is, if one places on the complex plane the numerically produced β-branch and δ-branch eigenvalues, then the theoretically predicted branches can be seen practically immediately (see Figure 2).

Figure 2 means that the leading terms from asymptotics (2.2) and (2.3) can be used by practitioners as good approximations for the eigenvalues.

The numerical procedure which is quite nontrivial, is briefly outlined below. The numerical approximation is based on Chebyshev polynomials. First, a finite-dimensional approximation for the main differential operator $\mathcal{L}_{\beta\delta}$, (denoted by $L_{\beta\delta}$) is described. Let $\mathcal{H}_N$ be an N-dimensional subspace of polynomials of degree $(N-1)$; obviously $\mathcal{H}_N \subset \mathcal{H}$, where $\mathcal{H}$ is the main Hilbert space with norm (1.13). Each polynomial is uniquely determined by its
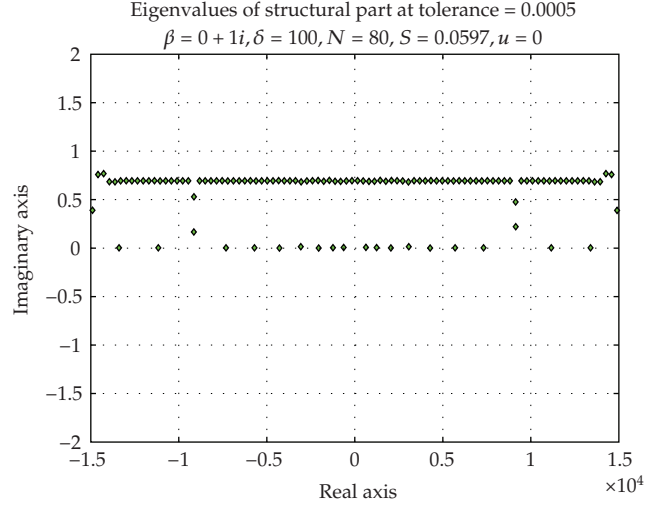
Figure 2: Distinct β-and δ-branches.

values at N points. These N points are the N extremal points of the degree $(N-1)$ Chebyshev polynomial $T_{N-1}(x)$ on $[-1,1]$, defined by

$$T_{N-1}(y) = \cos(N-1)\theta, \quad y = \cos(\theta), \ 0 \leq \theta \leq \pi, \ -1 \leq y \leq 1. \tag{3.3}$$

These N extrema are $y_k = \cos(k\pi/(N-1))$, $k = 0,\ldots,N-1$. Transforming them to $[-L,0]$ with $x_{k+1} = (L/2)(y_k - 1)$ yields

$$-L = x_N < x_{N-1} < \cdots < x_j < \cdots < x_2 < x_1 = 0. \tag{3.4}$$

Next, the action of $L_{\beta\delta}$ is considered on $\mathcal{H}_N$, defined to be the subspace of $\mathcal{H}$ consisting of 4-component vectors, each of whose components is a polynomial of degree $N-1$:

$$\overrightarrow{\Psi}(x) \equiv \begin{bmatrix} \psi_0(x) \\ \psi_1(x) \\ \psi_2(x) \\ \psi_3(x) \end{bmatrix}, \tag{3.5}$$

where ψ_0, ψ_1, ψ_2, and ψ_3 are polynomials of degree $N-1$ on $-L \leq x \leq 0$.

The boundary conditions are imposed on the finite dimensional system by restricting the polynomials $\psi_i(x)$, $i = 0, 1, 2, 3$, to a subspace $\mathcal{X}_N \subset \mathcal{H}_N$ of the functions satisfying the boundary conditions as follows:

$$\overrightarrow{\Psi}(x) \equiv \begin{bmatrix} \psi_0(x) \\ \psi_1(x) \\ \psi_2(x) \\ \psi_3(x) \end{bmatrix} \quad \text{such that} \quad \begin{cases} \psi_1(-L) = \psi_1'(-L) = \psi_3(-L) = 0, \\ \psi_0'''(0) = 0, \\ E\psi_0''(0) + \beta\psi_1'(0) = 0, \\ G\psi_2'(0) + \delta\psi_3(0) = 0, \\ \psi_0(-L) = \psi_0'(-L) = \psi_2(-L) = 0. \end{cases} \quad \text{Aeroelastic controls} \tag{3.6}$$

Here, derivative(s) at the boundary points means derivative(s) of the relevant polynomial component at the boundary points. In the matrix representation of $L_{\beta\delta}$, the derivatives d^2/dx^2 and d^4/dx^4 are computed with differentiation matrices, represented by D^2 and D^4, respectively, [18].

The finite dimensional eigenvalue problem, with polynomials represented as vectors of nodal values, then becomes the following. Find

$$\overrightarrow{\Psi} \equiv \begin{bmatrix} \overrightarrow{\psi}_0 \\ \overrightarrow{\psi}_1 \\ \overrightarrow{\psi}_2 \\ \overrightarrow{\psi}_3 \end{bmatrix} \in \mathcal{X}_N \tag{3.7}$$

such that $\imath L_{\beta\delta}\overrightarrow{\Psi} = \lambda\overrightarrow{\Psi}$, that is,

$$\imath L_{\beta\delta}\overrightarrow{\Psi} = \begin{bmatrix} 0 & I & 0 & 0 \\ -\dfrac{E\tilde{I}}{\Delta}D^4 & -\dfrac{\pi\rho u\tilde{S}}{\Delta} & -\dfrac{\tilde{S}}{\Delta}(GD^2 + \pi\rho u^2) & -\dfrac{\pi\rho u\tilde{I}}{\Delta} \\ 0 & 0 & 0 & I \\ \dfrac{E\tilde{S}}{\Delta}D^4 & \dfrac{\pi\rho u\tilde{m}}{\Delta} & -\dfrac{\tilde{m}}{\Delta}(GD^2 + \pi\rho u^2) & \dfrac{\pi\rho u\tilde{S}}{\Delta} \end{bmatrix} \cdot \begin{bmatrix} \psi_0 \\ \psi_1 \\ \psi_2 \\ \psi_3 \end{bmatrix} = \lambda \begin{bmatrix} \psi_0 \\ \psi_1 \\ \psi_2 \\ \psi_3 \end{bmatrix}. \tag{3.8}$$

This system's eigenvalues are so sensitive to roundoff errors that conventional approaches to solving the eigenvalue problem are numerically intractable, even in double precision. Fortunately, there is a better way to calculate the eigenvalues of this system, which we now briefly describe. (For a more detailed explanation of these details, and the numerical difficulties, see [19].)

Conventional collocation methods would impose boundary values by replacing some rows in the matrix representation of $\imath L_{\beta\delta}$ with rows that impose the appropriate restrictions on the vector components. To overcome the numerical difficulties, an alternative method of imposing boundary values is used—namely, representing $\mathcal{X}_N$ as the kernel of a linear operator defined as follows.

The nine boundary conditions can be described by a mapping of $\Lambda : \mathcal{H}_N \rightarrow \mathbb{R}^9$ given by

$$
\Lambda \vec{\Psi} \equiv \Lambda \begin{bmatrix} \psi_0 \\ \psi_1 \\ \psi_2 \\ \psi_3 \end{bmatrix} = \begin{bmatrix} \psi_0(-L) \\ \psi_0'(-L) \\ \psi_2(-L) \\ G\psi_2'(0) + \delta\psi_3(0) \\ E\psi_0''(0) + \beta\psi_1'(0) \\ \psi_0'''(0) \\ \psi_1(-L) \\ \psi_1'(-L) \\ \psi_3(-L) \end{bmatrix}, \quad \text{so } \Lambda\vec{\Psi} = 0 \text{ means} \begin{cases} \psi_0(-L) = 0 \\ \psi_0'(-L) = 0 \\ \psi_2(-L) = 0 \\ G\psi_2'(0) + \delta\psi_3(0) = 0 \\ E\psi_0''(0) + \beta\psi'_1(0) = 0 \\ \psi_0'''(0) = 0 \\ \psi_1(-L) = 0 \\ \psi_1'(-L) = 0 \\ \psi_3(-L) = 0 \end{cases} \quad (3.9)
$$

which are exactly the nine boundary conditions. In other words, $\mathcal{K}_N$ is the kernel of the operator Λ. It can easily be seen that

$$
k = \dim(\text{Ker}(\Lambda)) = \dim(\mathcal{H}_N) - \text{rank}(\Lambda) = 4N - 9, \qquad \mathbb{R}^k \cong \mathcal{K}_N. \quad (3.10)
$$

Let B represent a $4N \times k$ matrix whose columns form an orthonormal basis for $\text{Ker}(\Lambda)$, so

$$
B : \mathbb{R}^k \longrightarrow \mathcal{K}_N = \text{Ran}(B), \qquad B^T B = \text{identity on } \mathbb{R}^k, \qquad BB^T = \text{identity on } \mathcal{K}_N. \quad (3.11)
$$

Finally, transform $L_{\beta\delta}$ with the basis change B to an operator with the domain $\mathbb{R}^k$ rather than $\mathcal{K}_N$, according to the diagram on **Figure 3**: the relationship between the eigenvalues of $B^T L_{\beta\delta} B$ and $L_{\beta\delta}$ is as follows:

$$
\begin{aligned}
&\text{If } \vec{\Psi} \in \mathcal{K}_N, \quad L_{\beta\delta}\vec{\Psi} = \lambda\vec{\Psi}, \quad \text{then } \vec{\Psi} = B\vec{v}, \quad \text{for some } \vec{v} \in \mathbb{R}^k, \\
&B^T L_{\beta\delta} B\vec{v} = \lambda B^T B\vec{v} = \lambda\vec{v} \\
&\text{which means } \begin{cases} \text{eigenvalues of } L_{\beta\delta} \\ \text{restricted to } \mathcal{K}_N \end{cases} \subset \begin{cases} \text{eigenvalues} \\ \text{of } B^T L_{\beta\delta} B \end{cases}.
\end{aligned} \quad (3.12)
$$

However, the reverse inclusion is not necessarily true. This means that after solving the standard eigenvalue problem $B^T L_{\beta\delta} B\vec{v} = \lambda\vec{v}$, one has to use some criterion to select those λ such that $L_{\beta\delta} B\vec{v} = \lambda B\vec{v}$. The chosen criterion is to calculate both eigenvalues and eigenvectors $\vec{v}$ and choose those λ such that $\|L_{\beta\delta} B\vec{v} - \lambda B\vec{v}\| < \varepsilon$. This is done in two steps.

(1) Use a QZ algorithm to calculate eigenvalue-eigenvector pairs, and calculate individual eigenvalue condition numbers (e.g., with MATLAB's package "condeig").
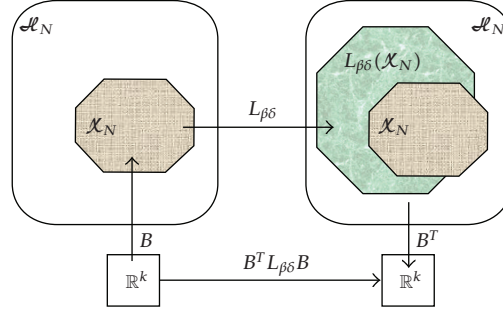
Figure 3: Graphical representation of basis change.

(2) Use an inverse power iteration on each eigenvalue-eigenvector pair from QZ algorithm, which improves the accuracy of the eigenvalues. The transformation of the eigenvalue problem $L_{\beta\delta}\vec{\psi} = \lambda\vec{\psi}$ to $B^T L_{\beta\delta} B \vec{v} = \lambda\vec{v}$ produces the spectrum with the asymptotic distribution predicted by Theorem 2.2.

4. Numerical Results on Flutter Modes

The second problem is related to the effect of the integral operator. As it is shown in the first round of calculations, the spectrum of the structural part of the entire aeroelastic dynamics generator splits into two branches. It is shown that the leading asymptotical terms of the β-and δ-branches can be used as practically acceptable approximations for the eigenvalues. However, analysis of "structural eigenvalues" was restricted to the ground vibration model. It is certainly important to clarify how the integral part of the problem (i.e., the aerodynamics) affects the distribution of the eigenvalues.

Obviously, one needs to select an acceptable approximation for the matrix-valued function $\widehat{\widehat{\mathfrak{F}}}(\lambda)$, or to find out which approximation would be the best for the Theodorsen function $T(\lambda)$. It is well-known that the problem of finding a good approximation for this function is an important component of computational aeroelasticity. Typically, a rational function is the standard approximation for $T(\lambda)$. In this work a different approximation, which appears naturally, is suggested. Namely, the asymptotic approximation for the Theodorsen function as its argument λ tends to infinity is used. Splitting the Theodorsen function yields corresponding splitting for the matrix-valued function $\widehat{\widehat{\mathfrak{F}}}(\lambda)$.

So, the main assumption is that the influence of the aerodynamics takes place through the matrix $\mathfrak{M}$ whose entries depend on the wind speed u. The goal of the numerical simulations below is to control the aeroelastic modes' response to increasing u, and/or changing the mass moment, S. In particular, one hopes to reach a specific speed, the flutter speed. As is known, flutter instability can be caused by one of the following two reasons. The first reason is the existence of an "unstable aeroelastic mode," that is, the mode whose real part is positive. The second reason is that for some speed u, the two aeroelastic modes initiated by two different branches merge together creating one double mode. For speeds that are close to the critical speed (or the flutter speed), one should use the modified model problem (most likely, a nonlinear model).
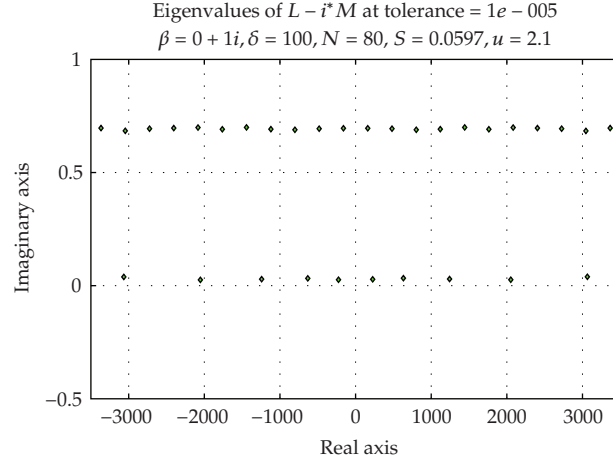
Eigenvalues of $L - i^* M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 80, S = 0.0597, u = 2.1$

Figure 4: Smaller eigenvalues.

Below, a summary of numerical findings is presented.

The graphs below show the distribution of the eigenvalues of a nonselfadjoint operator $\mathcal{L}_{\beta\delta}$ defined in (1.16) and (1.17). However, to find approximations for the aeroelastic modes, one should rotate each picture by 90 degrees in a counterclockwise direction.

As explained earlier, numerical calculations are based on spectral methods using Chebyshev polynomials. For a typical run from 60 to 80 Chebyshev nodes were used. As many as 120 nodes were used to see whether more nodes yielded significantly better accuracy, but there actually was higher numerical error with 120 nodes than with 80 or 100 nodes. This was apparently due to the extremely large fourth-order derivatives in the matrix differential operator at nodes near the boundaries.

Figure 4 shows a typical display of the lowest eigenvalues of the system. The diagram shows the clear presence of the projected β-and δ-branches, when the speed u is relatively low.

Changing the parameters of the system causes the eigenvalues to shift positions. It was put forth in [1, 2] that the presence of double eigenvalues would create the flutter condition. In the search for these double eigenvalues, the only parameters of the system that have been changed were the wind speed, u, and the mass moment of the wing, S.

The first significant change in eigenvalue position was observed while examining the system with $u = 2.1$ and S on the order of 0.2. The eighth and ninth eigenvalues with positive real parts were seen to jump branches as S changed. En route in this transition, the eigenvalues are clearly near-double. In addition, the effects of fixing S and changing u were examined. It was noticed that the same eigenvalues were near-double. It should be noted that this near-double condition is also true of the mirrored eighth and ninth eigenvalues, but only the set with positive real components was examined numerically. The near-double nature of these eigenvalues is subsequently examined.

The following five plots (Figures 5, 6, 7) show the progression of two pairs of near-double eigenvalues with changing mass moment, S.

The next five plots (Figures 8, 9, 10) show the progression of two pairs of near-double eigenvalues with changing wind speed, u.
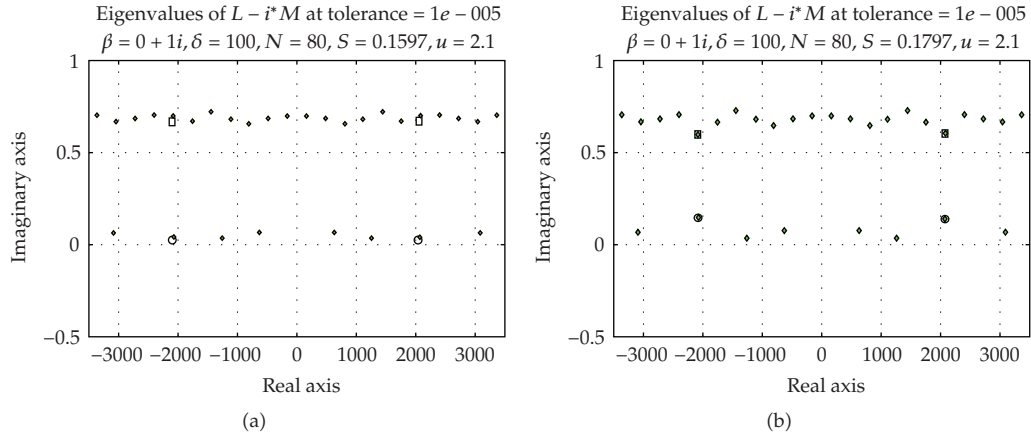
Eigenvalues of $L - i^*M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 80, S = 0.1597, u = 2.1$

Eigenvalues of $L - i^*M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 80, S = 0.1797, u = 2.1$

(a)

(b)

Figure 5: Progression of near-double eigenvalues with changing mass moment.

Eigenvalues of $L - i^*M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 80, S = 0.1997, u = 2.1$

Eigenvalues of $L - i^*M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 80, S = 0.2197, u = 2.1$

(a)

(b)
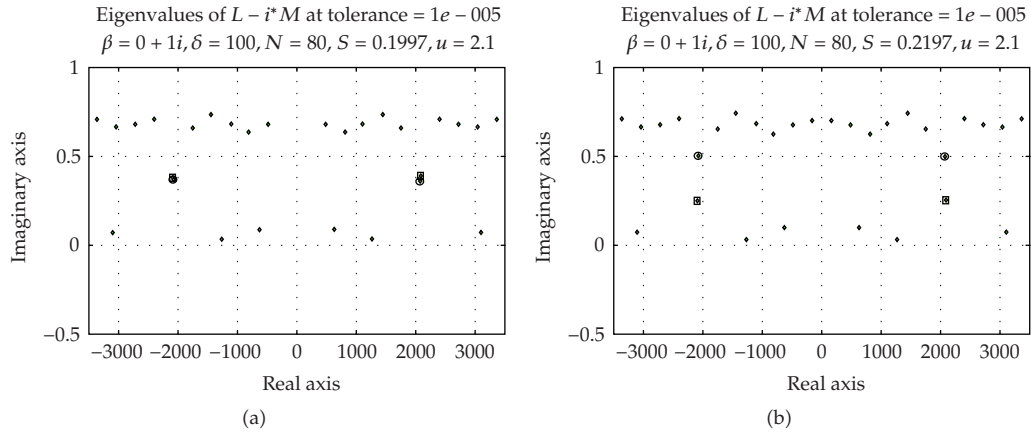
Figure 6: Progression of near-double eigenvalues with changing mass moment.

Eigenvalues of $L - i^*M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 80, S = 0.2397, u = 2.1$
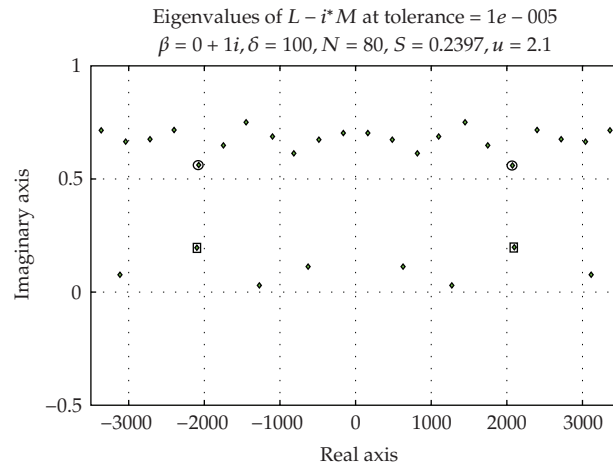
Figure 7: Progression of near-double eigenvalues with changing mass moment.

Eigenvalues of $L - i^*M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 80, S = 0.165, u = 87.5$

Eigenvalues of $L - i^*M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 80, S = 0.165, u = 90$

(a)

(b)
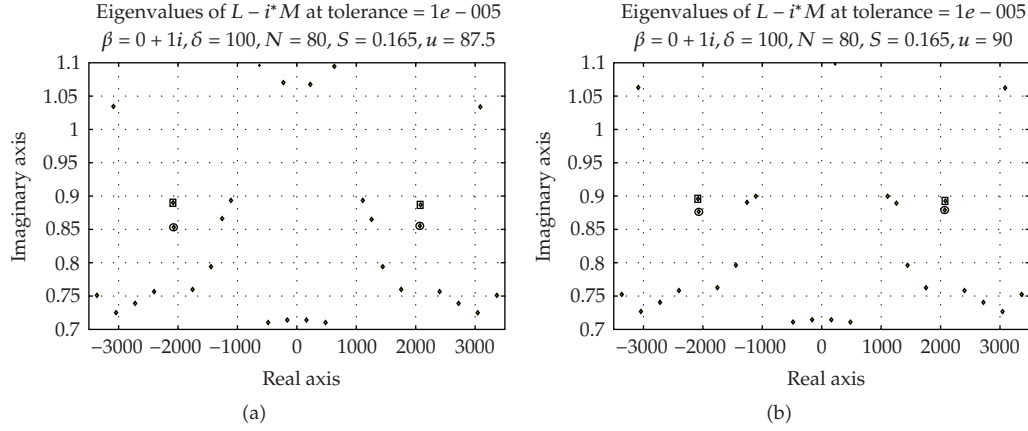
Figure 8: Progression of near-double eigenvalues with changing wind speed.

Eigenvalues of $L - i^*M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 80, S = 0.165, u = 92.5$

Eigenvalues of $L - i^*M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 80, S = 0.165, u = 95$

(a)

(b)
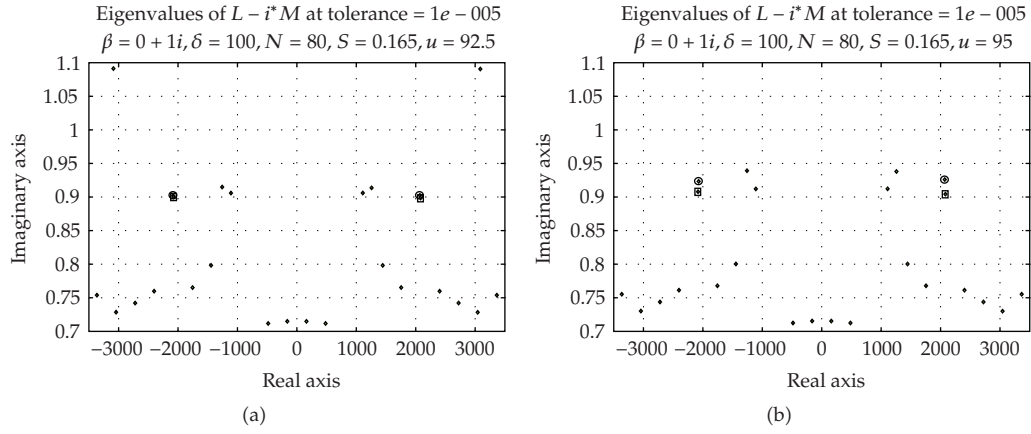
Figure 9: Progression of near-double eigenvalues with changing wind speed.

Eigenvalues of $L - i^*M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 80, S = 0.165, u = 97.5$

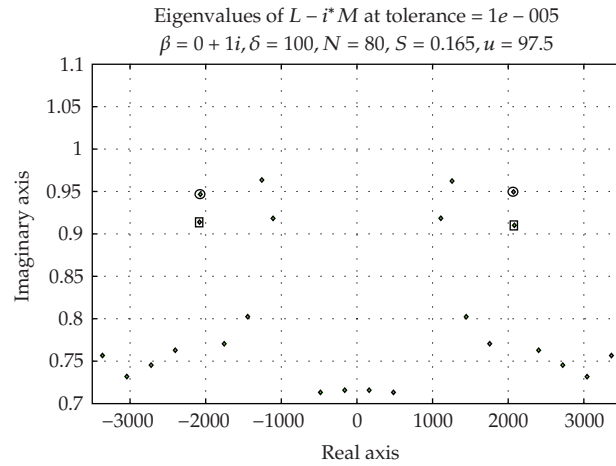**Figure 10:** Progression of near-double eigenvalues with changing wind speed.

Absolute distance between first set of close
Eigenvalues on positive real half plane
Eigenvalues of $L - i^* M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 60, S = 0.1672, u = [0, 735]$

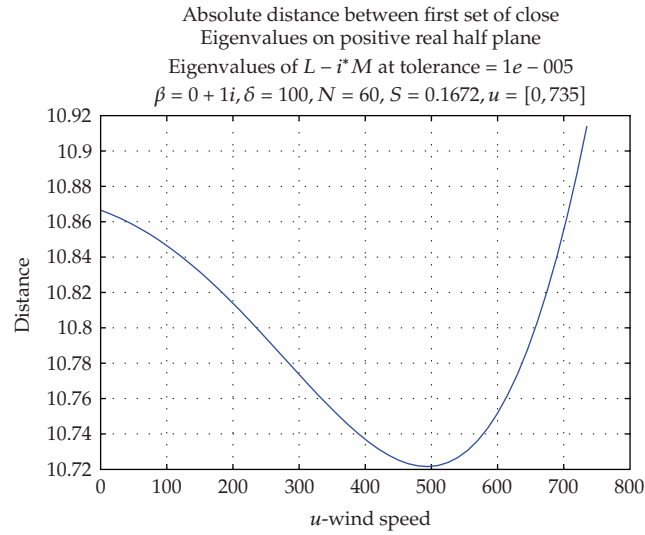

Figure 11: Absolute distance between first set of close eigenvalues.

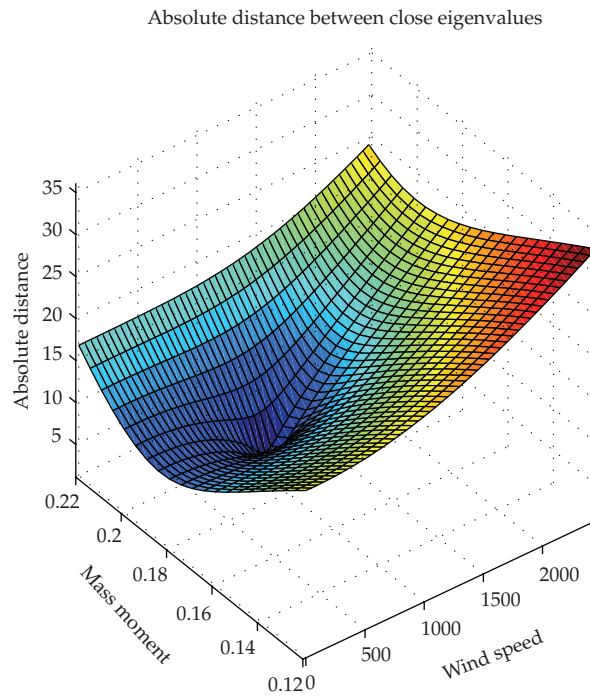Absolute distance between close eigenvalues

Figure 12: 3D Plot of absolute distance between near-double eigenvalues.

To examine the conditions under which these eigenvalues are the closest, one can fix S and calculate the absolute distance between these near-double eigenvalues as a function of u. Figure 11 shows a typical example of such a function. Note the presence of a distinct minimum distance.
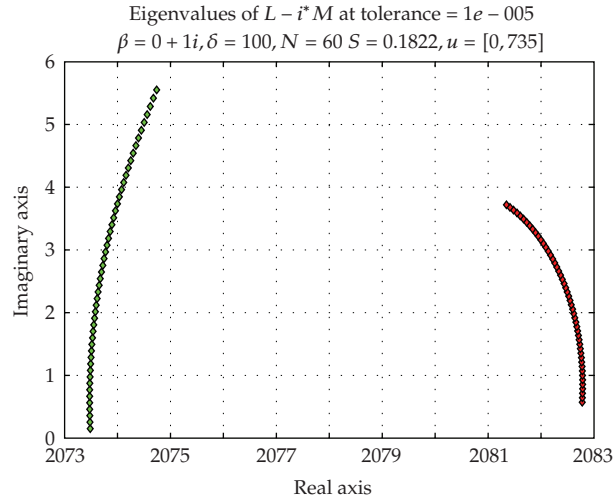
Eigenvalues of $L - i^*M$ at tolerance $= 1e - 005$
$\beta = 0 + 1i, \delta = 100, N = 60\ S = 0.1822, u = [0, 735]$

Figure 13: Typical paths of close eigenvalues.

Many production runs with fixed S have been performed, and it was found that both u and the absolute distance at the minimum were not unique. That is, the separation of these near-double eigenvalues is not only a function of u, but also of S. It becomes clear that in order to get a full handle on this problem, it is most useful to develop a fully three-dimensional plot. Figure 12 shows the absolute distance between these eigenvalues of note as a function of both S and u. It was interesting to see that there is a distinct low point in this figure, which occurs in the area of $u = 950$ and $S = 0.19$. Because eigenvalues didn't cross over into the negative imaginary half plane, and because there is such a well defined low point in this figure, one can assume that it is in this region that flutter will be observed.

It is important to note that these near-double eigenvalues were never observed to become exact double eigenvalues. Because the eigenvalues are numerical approximations, one cannot say that two eigenvalues have the same exact value, and a close examination of the path these eigenvalues take when passing close to each other reveals that they are indeed distinct. Figure 13 shows characteristic paths that always have clear separation. This confirms Figure 11, which illustrates that the absolute distance between these near-double eigenvalues is never zero.

5. Concluding Remarks

The present paper is concerned with numerical investigation of two problems arising in the area of theoretical aeroelasticity. Namely, it has been shown that analytical formulas representing the asymptotical distribution of aeroelastic modes for a specific aircraft wing model can be used by practitioners. Specifically, the leading terms in the spectral asymptotics represent vibrational frequencies of a wing quite accurately. The second problem is to clarify the nature of such dynamical instabilities as flutter. It has been shown that the model can capture the flutter phenomenon. In particular, it was observed that two different aeroelastic modes moved towards each other to create a double mode when the speed of the aircraft approached a certain value called the "flutter speed". This direction of research could be extended as follows. It is worthwhile to do a full study on the path of the near-double

eigenvalues. The paths that they take while moving past each other may indicate the conditions in which flutter is most rapidly approached. This information may prove useful when an aircraft changes the angle of attack to circumvent flutter. A more important future research will involve the examination of other near-double eigenvalues of this system. It is shown that this near-double condition is not unique to the eigenvalues examined in this paper. However, we do not yet know whether these additional eigenvalues can be trusted numerically to be near-double. Also, it is not yet known what their minimum absolute separation is, and if it is small enough to produce flutter.

Another future step in this project will be to include the full integral operator. This would be a more complete physical model of the system, and we expect that it will yield more accurate results. However, it will no doubt be more demanding computationally.

Acknowledgment

References

[1] R. L. Bisplinghoff, H. Ashley, and R. L. Halfman, *Aeroelasticity*, Dover, New York, NY, USA, 1996.

[2] Y. C. Fung, *An Introduction to the Theory of Aeroelasticity*, Dover, New York, NY, USA, 1993.

[3] J. Lutgen, "A note on Riesz bases of eigenvectors of certain holomorphic operator-functions," *Journal of Mathematical Analysis and Applications*, vol. 255, no. 1, pp. 358–373, 2001.

[4] M. A. Shubov, "Riesz basis property of root vectors of non-self-adjoint operators generated by aircraft wing model in subsonic airflow," *Mathematical Methods in the Applied Sciences*, vol. 23, no. 18, pp. 1585–1615, 2000.

[5] A. V. Balakrishnan and M. A. Shubov, "Asymptotic and spectral properties of operator-valued functions generated by aircraft wing model," *Mathematical Methods in the Applied Sciences*, vol. 27, no. 3, pp. 329–362, 2004.

[6] M. A. Shubov, "Asymptotic analysis of aircraft wing model in subsonic air flow," *IMA Journal of Applied Mathematics*, vol. 66, no. 4, pp. 319–356, 2001.

[7] M. A. Shubov, A. V. Balakrishnan, and C. A. Peterson, "Spectral properties of nonselfadjoint operators generated by coupled Euler-Bernoulli and Timoshenko beam model," *Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 84, no. 2, 2004.

[8] M. A. Shubov, "Asymptotic representations for root vectors of nonselfadjoint operators and pencils generated by an aircraft wing model in subsonic air flow," *Journal of Mathematical Analysis and Applications*, vol. 260, no. 2, pp. 341–366, 2001.

[9] A. V. Balakrishnan, "Subsonic flutter suppression using self-straining actuators," *Journal of the Franklin Institute*, vol. 338, no. 2-3, pp. 149–170, 2001.

[10] A. V. Balakrishan, "Aeroelastic control with self-straining actuators: continuum models," in *Smart Structures and Materials, Mathematics Control in Smart Structures*, V. Vasunradan, Ed., vol. 3323 of *Proceedings of SPIE*, pp. 44–54, 1998.

[11] A. V. Balakrishnan and J. W. Edwards, "Calculation of the transient motion of elastic airfoils forced by control surface motion and gusts," Tech. Rep. NASA TM 81351, 1980.

[12] A. V. Balakrishnan, "Damping performance of strain actuated beams," *Computational and Applied Mathematics*, vol. 18, no. 1, pp. 31–86, 1999.

[13] C.-K. Lee, W.-W. Chiang, and T. C. O'Sullivan, "Piezoelectric modal sensor/actuator pairs for critical active damping vibration control," *Journal of the Acoustical Society of America*, vol. 90, no. 1, pp. 374–384, 1991.

[14] M. A. Shubov and C. A. Peterson, "Asymptotic distribution of eigenfrequencies for coupled Euler–Bernoulli/Timoshenko beam model," NASA NASA/CR-2003-212022, NASA Dryden Center, November 2003.

[15] M. A. Shubov, "Asymptotics of aeroelastic modes and basis property of mode shapes for aircraft wing model," *Journal of the Franklin Institute*, vol. 338, no. 2-3, pp. 171–185, 2001.

[16] I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, vol. 18 of *Translations of Mathematical Monographs*, American Mathematical Society, Providence, RI, USA, 1969.

[17] M. Abramowitz and I. Stegun, Eds., *Handbook of Mathematical Functions*, Dover, New York, NY, USA, 1972.

[18] J. C. Mason and D. C. Handscomb, *Chebyshev Polynomials*, Chapman & Hall/CRC, Boca Raton, Fla, USA, 2003.

[19] M. A. Shubov and S. Wineberg, "Asymptotic distribution of the aeroelastic modes for wing flutter problem: numerical analysis," under review.

Research Article

Nonlinear and Dynamic Aerodynamic Models for Commercial Transport Aircraft with Adverse Weather Effects

Ray C. Chang,[1] Cun-En Ye,[1] C. Edward Lan,[2] and Yannian Lee[3]

[1] *Department of Aviation Mechanical Engineering, China University of Science and Technology, Hen-Shan (312), Taiwan*
[2] *Department of Aeronautics and Astronautics, National Cheng Kung University, Taiwan*
[3] *Accident Investigation Division, Aviation Safety Council, Taiwan*

Correspondence should be addressed to Ray C. Chang, raychang@cc.hc.cust.edu.tw

The flight dynamic equations in mathematics for aircraft response to the adverse weathers, such as wind shear, atmospheric turbulence, and in-flight icing, are nonlinear and unsteady. To effectively analyze the performance degradation and variations in stability of commercial aircraft that encountered these weather hazards, the nonlinear and dynamic (i.e., time dependent) aerodynamic models based on flight data would be needed. In the present paper, a numerical modeling method based on a fuzzy-logic algorithm will be presented to estimate the aerodynamic models for a twin-jet transport by using the flight data from the flight data recorder (FDR). The aerodynamic models having the capability to generate continuous stability derivatives, especially for sensitivity study of unknown factors in adverse weather conditions, will be demonstrated in this paper.

1. Introduction

Aircrafts in flight are frequently subject to atmospheric disturbances. The hazards due to these disturbances may be in the form of wind shear, turbulence (both clear-air and convective), thunderstorms, in-flight icing, and so forth. The severity of aircraft response to the disturbance is related to the dynamic aerodynamics that results from the instantaneous changes of aircraft flight attitudes. To provide the mitigation concepts and promote the understanding of aerodynamic responses of the commercial transport aircraft in adverse weather conditions, the nonlinear and dynamic (i.e., time-dependent) aerodynamic models based on flight data would be needed.

Regarding the flight data of commercial transport aircraft, the flight data recorder (FDR) is ICAO-regulated devices. The FDR, popularly referred to as a "black box", is

a mandatory device used to record specific aircraft parameters for event investigations. Although the data resolution and accuracy in the FDR are not as well defined as those in flight test, the predicted aerodynamic derivatives correlated well with the nominal values of transport aircraft in normal flight [1]. Therefore, the FDR data will be used in the present analysis.

In setting up the aerodynamic models to predict the required derivatives in aerodynamics, the traditional method of using the tabulated data derived from such methods as the maximum likelihood method (MMLE) [2], the least-square or the stepwise regression method [3], is very difficult to correlate complex functional relations among numerous parameters. In earlier development of the fuzzy-logic algorithm, Zadeh [4] used the fuzzy sets to simulate physical parameters with membership functions. The disadvantage in this approach is that the predicted curves in the functional approximation tend to be piecewise continuous. Therefore, it is not suitable when derivatives of the function are needed. In 1985, Takagi and Sugeno [5] used the internal functions, instead of the fuzzy sets, in developing the fuzzy-logic algorithm. In this approach, the main features are the internal functions, membership functions, and the output cells. The output curves in the prediction are smooth. The present development will be based on this second approach. This second approach, to be called the Fuzzy-Logic Modeling (FLM) technique [6, 7], is capable of correlating multiple parameters without assuming explicit functional relations.

Among the different adverse weathers, clear-air turbulence is all the more important in flight safety since it is hard to detect and predict. Clear-air turbulence is the leading cause of serious personal injuries in nonfatal accidents of commercial aircraft. One main type of motion that causes flight injuries in clear-air turbulence is the sudden plunging motion with the abrupt change in altitude. This paper presents FLM technique to establish unsteady aerodynamic models with six aerodynamic coefficients based on the datasets from the flight data recorder (FDR) of a twin-jet transport. The aerodynamic derivatives extracted from these aerodynamic models are then used to evaluate the variations in stability. The sensitivity study of unknown factors during the sudden plunging motion in severe clear-air turbulence will be demonstrated in this paper.

2. Theoretical Development

The general idea of the FLM technique is to set up the relations between its input and output variables of the whole system. The internal functions, membership functions, and outputs are three basic elements for the FLM approach. Two main tasks are involved in the FLM process. One is the identification of the coefficients of the internal functions, which is called parameter identification. The other one is structure identification to identify the optimal structure of fuzzy cells of the model. Details of the FLM technique are described in the followings [6–8].

2.1. Internal Functions

The fuzzy-logic model uses many internal functions to cover the defined ranges of the influencing parameters. Although the form of internal functions is very simple, it can represent a highly nonlinear relationship between the input and output for the whole system. These internal functions are assumed to be linear functions of input variables [8] as follows:

$$P^i = y_i(x_1, x_2, \ldots, x_r, \ldots, x_k) = p_0^i + p_1^i x_1 + \cdots + p_r^i x_r + \cdots p_k^i x_k, \tag{2.1}$$

where p_r^i, $r = 0, 1, 2, \ldots, k$, are the coefficients of internal functions y_i, and k is number of input variables; $i = 1, 2, \ldots, n$, and n is the total number of fuzzy cells.

The recorded data in FDR, such as flight altitude (h), calibrated airspeed (CAS), angle of attack (α), accelerometer readings (a_x, a_y, and a_z), Euler angles (θ, ϕ, and ψ), and so forth, is chosen as the input variables for a specific fuzzy model. In the present paper, y_i is denoted as an estimated aerodynamic coefficient of force or moment, and x_r are the variables of the input data. The numbers of the internal functions (i.e., cell's numbers) are quantified by the membership functions.

2.2. Membership Functions

The values of each fuzzy variable, such as the angle of attack, are divided into several ranges; each of which represents a membership function with $A(x_r)$ as its membership grade. One membership function from each variable constitutes a fuzzy cell. For the ith cell, the corresponding membership grades are represented by $A_r^i(x_r)$, $r = 1, 2, \ldots, k$. In other words, the membership functions allow the membership grades of the internal functions for a given set of input variables to be calculated. For a given system with input variables $x_1, x_2, \ldots, x_r, \ldots x_k$, the recorded values of each input variables are normalized by using $(x_r - x_{r,\min})/(x_{r,\max} - x_{r,\min})$ to transform them into the ranges of $[0, 1]$. The membership grading also ranges from 0 to 1, with "0" meaning no effect from the corresponding internal function and "1" meaning a full effect. Generally, overlapped straight lines, triangles, or trapezoids are frequently the shapes used to represent the grades. In this paper, the overlapped triangular membership function is used to represent the grades of internal functions.

The membership functions partition the input space into many fuzzy subspaces, which are called the fuzzy cells. The total number of fuzzy cells is $n = N_1 \times N_2 \times \cdots \times N_r \times \cdots \times N_k$. For a variable x_r, the number of membership function is N_r. Each fuzzy cell is in a different combination from others formed by taking one membership function from each input variable.

In the present application, triangular membership functions are used throughout. Let N be the number of membership functions and let j be the index for the jth membership functions. Then the membership grades can be described as follows [9]:

(1) $N = 2$, $A(x_r) = x_r$, $j = 1$,

$$A(x_r) = 1 - x_r, \quad j = 2, \tag{2.2}$$

(2) $j = 3$ to m, where $m = \max(0, (N - 2)/2)$:

$$A(x_r) = \frac{x_r}{d_u}, \quad 0 \leq x_r \leq d_u,$$

$$A(x_r) = \frac{(1 - x_r)}{(1 - d_u)}, \quad d_u \leq x_r \leq 1, \tag{2.3}$$

where $d_u = \Delta x_1^*(j - 2)$, and $\Delta x_1 = 1.0/(N - m - 1)$,

(3) $j \geq N - m$,

$$A(x_r) = \frac{(d_d - x_r)}{d_d}, \quad 0 \leq x_r \leq d_d,$$

$$A(x_r) = \frac{(d_d - x_r)}{(d_d - 1)}, \quad d_d \leq x_r \leq 1,$$

$$(2.4)$$

where $d_d = \Delta x_2^*(j - N + m)$, and $\Delta x_2 = 1.0/(m+1)$.

2.3. Fuzzy Rule and Output

A fuzzy cell is formed by taking one membership function from each variable. The total number of cells is the number of possible combinations by taking one membership function from each input variable. For every cell, it has a fuzzy rule to guide the input and output relations. The rule of the ith cell is stated [8] as follows: *if x_1 is $A_1^i(x_1)$, if x_2 is $A_2^i(x_2)$, ..., and if x_k is $A_k^i(x_k)$, then the cell output is*

$$P^i(x_1, x_2, \ldots, x_r, \ldots x_k) = p_0^i + p_1^i x_1 + \cdots + p_r^i x_r + \cdots p_k^i x_k,$$

$$(2.5)$$

where $i = 1, 2, \ldots, n$ is the index of the cells, n is the total number of cells of the model; $P^i(x_1, x_2, \ldots, x_r, \cdots x_k)$ is the internal function with parameters $p_0^i, p_1^i, \ldots, p_r^i, \ldots p_k^i$ to be determined, and $A_k^i(x_k)$ denotes the membership grade for x_k. Each function covers a certain range of input variables.

In each fuzzy cell, the contribution to the outcome (i.e., the cell output) is based on the internal function (2.5). The final prediction of the outcome is the weighted average of all cell outputs after the process of reasoning algorithm. Because of this weighting among many factors over large ranges of possibilities, the word "fuzzy" is derived to describe the method. However, its prediction is never "fuzzy". The output estimated by the fuzzy-logic algorithm corresponding to the jth input $(x_{1,j}, x_{2,j}, \ldots, x_{r,j}, \ldots, x_{k,j})$ is as follows:

$$\widehat{y}_j = \frac{\sum_{i=1}^{n} \text{product}\left[A^i(x_{1,j}), \ldots, A^i(x_{r,j}), \ldots, A^i(x_{k,j})\right] p^i}{\sum_{i=1}^{n} \text{product}\left[A^i(x_{1,j}), \ldots, A^i(x_{r,j}), \ldots, A^i(x_{k,j})\right]}.$$

$$(2.6)$$

In (2.6) product$[A^i(x_{1,j}), \ldots, A^i(x_{r,j}), \ldots, A^i(x_{k,j})]$ is the weighted value of the ith cell, and the index j of the data set, where $j = 1, 2, \ldots, m$, m is the total number of the data record, and the "product" stands for product operator of its elements in this paper.

2.4. Parameter Identification

Given a set of membership functions for each input variable, the unknown coefficients of the internal functions are adjusted by using the Newton gradient-descent method. The accuracy

of the established aerodynamic model through the fuzzy-logic algorithm is estimated by the sum of squared errors (SSEs) and the multiple correlation coefficients (R^2):

$$\text{SSE} = \sum_{j=1}^{m} (\widehat{y}_j - y_j)^2,$$ (2.7)

$$R^2 = 1 - \frac{\left\{\sum_{j=1}^{m} (\widehat{y}_j - y_j)^2\right\}}{\left\{\sum_{j=1}^{m} (\overline{y} - y_j)^2\right\}}.$$ (2.8)

In (2.7) and (2.8), where $\widehat{y}_j$, the output of the fuzzy-logic model at point j, is estimated by (2.6), y_j is the data point used for the model training at point j, $\overline{y}$ is the mean of the sample data, and m is the total number of data points. The model training is to determine the unknown coefficients of the internal functions, p_r^i, by maximizing the value of R^2. These coefficients are determined by the following iterative formula to minimize the sum of squared error (2.7):

$$p_{r,t+1}^i = p_{r,t}^i - \alpha_r \frac{\partial(\text{SSE})}{\partial p_r^i},$$ (2.9)

where α_r is the convergence factor or the step size in the gradient method; subscript index t denotes the iteration sequence.

In usual, a large value of α_r is chosen, such as 10, at the beginning to accelerate the process of convergence; the small value of less then 1 is taken to obtain more precise unknown coefficients of the internal functions, p_r^i, in model training process. After simplification, (2.9) becomes as follows.

For $r = 0$,

$$p_{0,t+1}^i = p_{0,t}^i - 2\alpha_0 (\widehat{y}_j - y_j) \times \frac{\text{product}\left[A_1^i(x_1), \ldots, A_r^i(x_{r,j}), \ldots, A_k^i(x_{k,j})\right]}{\sum_{s=1}^{n} \text{product}\left[A_1^s(x_{1,j}), \ldots, A_r^s(x_{r,j}), \ldots, A_k^s(x_{k,j})\right]},$$ (2.10a)

and for $r = 1, 2, \ldots, k$,

$$p_{r,t+1}^i = p_{r,t}^i - 2\alpha_r (\widehat{y}_j - y_j) \times \frac{\text{product}\left[A_1^i(x_1), \ldots, A_r^i(x_{r,j}), \ldots, A_k^i(x_{k,j})\right] x_{r,j}}{\sum_{s=1}^{n} \text{product}\left[A_1^s(x_{1,j}), \ldots, A_r^s(x_{r,j}), \ldots, A_k^s(x_{k,j})\right]}.$$ (2.10b)

The iteration during the search sequence stops when one of the following three criteria [6, 9] is satisfied:

(1) Cost = $\text{SSE}_t < \varepsilon_1$,

(2) RER = $(\text{SSE}_t - \text{SSE}_{t-1})/\text{SSE}_t < \varepsilon_2$,

(3) $t = t_{\max}$.

In the above criteria, Cost = SSE_t is the sum of squared errors (SSEs) in current iteration to be denoted by "Cost" and RER = (cost_current − cost_previous)/cost_current to

be denoted by "RER" (i.e., the relative error) for simplicity in descriptions; ε_1 and ε_2 are the required precision criteria, and $t_{\max}$ is a specified maximum iteration number.

Given membership functions and the training data, this parameter identification procedure can be applied to establish a fuzzy-logic model. Although the procedure is well understood, to obtain a good model with the appropriate membership functions that are quite challenging, it requires numerical experimentation.

2.5. Model Structure Identification

In the fuzzy-logic model, the model structure is indicated by the number of membership functions. For a fuzzy-logic model with multiple variables, the structure is the combination of the numbers and forms of the membership functions assigned to all input variables. Since the sequence defines the one-to-one relationship between the numbers and the forms for each variable, the structure can be uniquely described by the numbers.

The model structure is optimized by maximizing (2.8). A search forward algorithm has been employed for the identification. At each search stage, there may be many fuzzy-logic models with different structure combinations. The search stage numbers are denoted by N_s. Out of all the possible intermediate fuzzy-logic models at each search stage, for an efficient search, only some structures are developed and evaluated. Two selection criteria, to be given below, are used to choose these structures. With the incremental sequence and the selection criteria, the search forward algorithm can be shown graphically in Figure 1.

The step-by-step searching flow is summarized as follows [6].

(1) Specify the input variables x_r, $r = 1, 2, \ldots, k$ and the output variable y.

(2) Assume an initial structure, also called parent structure, as $(N_{10}, N_{20}, \ldots, N_{r0}, \ldots, N_{k0})$.

(3) Begin at the search stage number $N_s = l$ form all possible structures starting from the parent structure by adding one more membership function a time only to one input variable. Those all possible structures are called child structures as $(N_{10} + 1, N_{20}, \ldots, N_{r0}, \ldots, N_{k0})$, $(N_{10}, N_{20} + 1, \ldots, N_{r0}, \ldots, N_{k0}), \ldots$, and $(N_{10}, N_{20}, \ldots, N_{r0}, \ldots, N_{k0} + 1)$. Do the identification of internal coefficients in (2.1) for each child structure and then calculate its R^2 by using (2.8).

(4) Select the top 5 child structures among all calculated values of R^2 as new parent structures for next search step $N_s = N_s + l$.

(5) Go back to step (2) starting from the new parent structures and repeat the same procedures in steps (2) and (3) until the best structure is identified;

(6) Pick out the maximum value of R^2 among the child structures in each searching stage as $R^2_{\max}$. The structure of the largest $R^2_{\max}$ corresponding to the all picking values is the optimal structure within a sensible N_s.

In the structure identification, parameter identification to determine the p-parameters according to (2.9) is also needed. But the number of iteration to determine the p-parameters is limited to 2000, so that the best structure is decided on a relative basis. After this last step, (2.9) is applied iteratively until both values of R^2 and RER reach the requirements in the final parameter identification
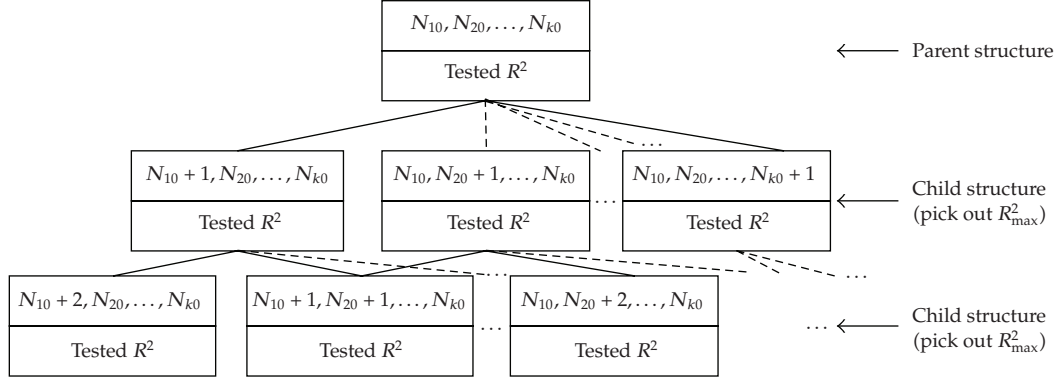
Figure 1: Best model structure searching flow.

3. Application to Aerodynamic Modeling

3.1. Flight Data

The twin-jet transport of the present study encountered clear-air turbulence in cruise flight at the altitude around 10,050 m. As a result, several passengers and cabin crews sustained injuries, because of which this event was classified as an accident. The present study was initiated to examine possible concepts of accident prevention in the future. The datasets used for the modeling are extracted from the FDR during turbulence encounter lasting for 92 seconds.

The main aircraft geometric and inertial characteristics are taken to be

W (take-off) = 1,431,800 N (321900 lb),

$S = 260 \, \text{m}^2$ (2798.7 ft^2), $\bar{c} = 6.608$ m (21.68 ft), and $b = 44.827$ m (147.08 ft),

$I_{xx} = 10,710,000 \, \text{kg-m}^2$ (7,899,900 slugs-ft^2), and $I_{yy} = 14,883,800 \, \text{kg-m}^2$ (10,978, 000 slugs-ft^2),

$I_{zz} = 25,283,271 \, \text{kg-m}^2$ (18,648,470 slugs-ft^2) and $I_{xz} = 0.0 \, \text{kg-m}^2$.

The required operational parameters in FDR dataset for generating aerodynamic data files are time (t), CAS, pressure altitude (h), roll attitude (ϕ), pitch attitude (θ), magnetic heading (ψ), longitudinal acceleration (a_x), lateral acceleration (a_y), vertical acceleration (a_z), angle of attack (α), aileron deflection (δ_a), elevator (δ_e), rudder (δ_r), stabilizer (δ_s), engine EPR, outside air temperature, wind speed, wind direction, and fuel flow rate. Since only the normal acceleration is recorded in 8 Hz resolution (i.e., 8 points per second), all other parameters are interpolated with a monotone cubic spline to the same sampling rate.

3.2. Compatibility Analysis

Typically, the longitudinal, lateral, and vertical accelerations (a_x, a_y, a_z) along the (x, y, z)-body axes of aircraft, angle of attack α, and the Euler angles (ϕ, θ, and ψ) as well as all control deflections are available and recorded in the FDR of all transport aircraft. Since the

recorded flight data may contain errors (or called biases), compatibility analysis is performed to remove them by satisfying the following kinematic equations:

$$\dot{\phi} = p + q \sin\phi \tan\theta + r \cos\phi \tan\theta, \tag{3.1}$$

$$\dot{\theta} = q \cos\phi - r \sin\phi, \tag{3.2}$$

$$\dot{\psi} = (q \sin\phi + r \cos\phi)\sec\theta, \tag{3.3}$$

$$\dot{V} = (a_x - g \sin\theta) \cos\alpha \cos\beta + (a_y + g \sin\phi \cos\theta) \sin\beta$$
$$+ (a_z + g \cos\phi \cos\theta) \sin\alpha \cos\beta, \tag{3.4}$$

$$\dot{\alpha} = \frac{[(a_z + g \cos\theta \cos\phi) \cos\alpha - (a_x - g \sin\theta) \sin\alpha]}{(V \cos\beta)}$$
$$+ q - \tan\beta(p \cos\alpha + r \sin\alpha), \tag{3.5}$$

$$\dot{\beta} = \frac{\cos\beta(a_y + g \cos\theta \sin\phi)}{V} + p \sin\alpha - r \cos\alpha$$
$$- \frac{\sin\beta[(a_z + g \cos\theta \cos\phi) \sin\alpha - (a_x - g \sin\theta) \cos\alpha]}{V}, \tag{3.6}$$

where g is acceleration due to gravity, V is flight speed, β is sideslip angle, p is roll rate, q is pitch rate, and r is yaw rate in (3.1)–(3.6). Let the biases be denoted by $b_{a_x}, b_{a_y}, b_{a_z}, b_p, b_q, b_r, b_V, b_\alpha, b_\beta, b_\theta, b_\phi$, and b_ψ, respectively, for a_x, a_y, a_z, and so forth. These biases are estimated by minimizing the squared sum of the differences between the two sides of the above equations. These equations in vector form can be written as

$$\dot{\vec{z}} = \vec{f}(x) = \vec{f}(x_m - \Delta x), \tag{3.7}$$

where

$$\vec{z} = (V, \alpha, \beta, \theta, \phi, \psi)^T, \tag{3.8}$$

$$\vec{x}_m = (a_x, a_y, a_z, p, q, r, V, \alpha, \beta, \theta, \phi, \psi)^T, \tag{3.9}$$

$$\Delta \vec{x} = \left(b_{a_x}, b_{a_y}, b_{a_z}, b_p, b_q, b_r, b_V, b_\alpha, b_\beta, b_\theta, b_\phi, b_\varphi\right)^T, \tag{3.10}$$

where the subscript "m" indicates the measured or recorded values. The cost function is defined as

$$J = \frac{1}{2}\left(\dot{\vec{z}} - \vec{f}\right)^T Q\left(\dot{\vec{z}} - \vec{f}\right), \tag{3.11}$$

where Q is a weighting diagonal matrix with elements being 1.0 except the one for the slowly varying flight speed being 10.0 and $\dot{\vec{z}}$ is calculated with a central difference scheme with $\dot{\vec{z}}_m$,

which is the measured value of $\vec{z}$. The steepest descent optimization method is adopted to minimize the cost function. As a result of the analysis, variables not present in the FDR, such as β, p, q, and r, are also estimated.

The force and moment coefficients are obtained from the following flight dynamic equations [7, 10] about the airplane body axes:

$$ma_x = C_x \overline{q} S + T_x, \tag{3.12}$$

$$ma_y = C_y \overline{q} S + T_y, \tag{3.13}$$

$$ma_z = C_z \overline{q} S + T_z, \tag{3.14}$$

$$C_l \overline{q} S b = I_{xx} \dot{p} - I_{xz}(\dot{r} + pq) - (I_{yy} - I_{zz})qr, \tag{3.15}$$

$$C_m \overline{q} S \overline{c} = I_{yy} \dot{q} - I_{xz}(r^2 - p^2) - (I_{zz} - I_{xx})rp - T_m, \tag{3.16}$$

$$C_n \overline{q} S b = I_{zz} \dot{r} - I_{xz}(\dot{p} - qr) - (I_{xx} - I_{yy})pq, \tag{3.17}$$

where m is denoted as aircraft mass; $\overline{q}$ is denoted as dynamic pressure; S is denoted as wing reference area; C_x, C_z, and C_m are denoted as longitudinal aerodynamic force and moment coefficients; C_y, C_l, and C_n are denoted as lateral-directional aerodynamic force and moment coefficients; I_{xx}, I_{yy}, and I_{zz} are denoted as moments of inertia about x-, y-, and z-axes, respectively; I_{xy}, I_{xz}, and I_{yz} are denoted as products of inertia; T_x, T_y, T_z, and T_m are denoted as thrust terms about x-, y-, z-axes, and in equation of pitching moment, respectively, in (3.7)–(3.12).

The above equations are used to determine all aerodynamic coefficients based on accelerometer readings (a_x, a_y, and a_z), Euler angles (ϕ, θ, and ψ), angular rates (p, q, and r), and thrusts (T_x, T_y, T_z, and T_m). The angular rates are estimated through compatibility analysis. Since thrusts were not measured during flight for most flight vehicles, those values and the effects on the forces and pitching moments in equations of (3.7), (3.8), (3.9), and (3.11) should be predicted by a thrust model.

3.3. Equivalent Harmonic Motion

The reduced frequency is a parameter to indicate the degree of unsteadiness in unsteady aerodynamics and is estimated in this paper by fitting the local trajectory with a harmonic motion. In the static case, the reduced frequency is 0. Large values of the reduced frequency imply the importance of unsteady aerodynamic effect. For longitudinal aerodynamics, the equivalent harmonic motion is the one based on the angle-of-attack variation following the classical unsteady aerodynamic theory of Theodorsen [11]. For lateral-directional aerodynamics, it is based on the time variation of roll angle [8].

For the longitudinal motion, the time history of the angle of attack (α) and time rate of angle of attack ($d\alpha/dt$, or $\dot{\alpha}$) are fitted with one of a harmonic motion at any instant [7, 12] as follows:

$$\alpha(t) = \overline{\alpha} + \alpha \cos\left(\omega t + \overline{\phi}\right), \tag{3.18}$$

$$\dot{\alpha}(t) = -a\omega \sin\left(\omega t + \overline{\phi}\right), \tag{3.19}$$

where those terms on the left hand side of (3.13) and (3.14) are given and the unknowns are the local mean angle of attack ($\bar{\alpha}$), the local amplitude of the harmonic motion (a), the phase lag ($\bar{\phi}$), and the angular frequency (ω). These unknowns are calculated through an optimization method by minimizing the following cost function (least squares):

$$ J = \sum_{i=1}^{n} \left[\alpha_i - \left(\bar{\alpha} + a\cos\left(\omega t_i + \bar{\phi}\right)\right)\right]^2 + \left[\dot{\alpha}_i - \left(\bar{\alpha} + a\omega\sin\left(\omega t_i + \bar{\phi}\right)\right)\right]^2. \qquad (3.20) $$

In (3.15), α_i is the measured value at point i and n is the number of the data points used in the optimization. For the case in the present study, $n = 20$ is found to be the best value. The 20 points preceding and including the current time are employed in (3.15). The least-square method is found to converge well and gives reasonably accurate results. The lateral-directional equivalent reduced frequency is computed in the same manner.

The local equivalent reduced frequency in the longitudinal motion is defined as

$$ k_1 = \frac{\omega \bar{c}}{V}, \qquad (3.21) $$

where $\bar{c}$ is the mean chord length of wing airfoil section.

The lateral-directional equivalent reduced frequency is defined as

$$ k_2 = \frac{\omega b}{V}, \qquad (3.22) $$

where b is the wing span.

3.4. Fuzzy-Logic Thrust Model

As shown before, the thrust terms appear in the force equations and the pitching moment equations (3.7)–(3.9) and (3.11). Since the values of thrust for aircraft in flight cannot be directly measured in the current state of the art, they are not recorded in the FDR. The manufacturers of airplanes and engines and safety agencies agreed that using such parameters as the Mach number, airspeed, flight altitude, temperature, the rpm of the pressure compressors, and engine pressure ratios is adequate to estimate the engine thrust. A realistic thrust model is quite complex and cannot be represented by any simple equation. Since such thrust model is not available for the present study, a realistic one tied to the recorded engine performance parameters will be developed with the fuzzy-logic algorithm.

For a commercial aircraft, most likely only the axial force and the pitching moment are affected by thrust. This assumption will be made in this paper. Theoretically, clear-air turbulence (i.e., random change in u, w (or α), and v (or β) affects the engine performance through its effects on static and dynamic distortions at the engine face. However, the effects of clear-air turbulence on the engine parameters and the thrust magnitude are not known and, therefore, are ignored.

For this purpose, data from the flight manual for the fuel flow rates ($\dot{m}_f$) at various altitudes (h), weights (W), Mach numbers (M), calibrated airspeed (CAS), engine pressure ratios (EPRs) in cruise flight are utilized. Note that the drag polar for a given aircraft is

generally not known to most researchers. To estimate it and, hence, the thrust magnitude in cruise, the assumption of a design lift-to-drag ratio (L/D) of 17.5 is made. This value of lift-to-drag in cruise is assumed based on the past design experience for twin-jet transports. In the flight manual, various weights, altitudes, Mach numbers, CAS, EPR, and fuel flow rates in cruise are tabulated. The lift coefficient can be calculated at each flight condition immediately. As a result, the drag coefficient can be estimated from the assumption of lift-to-drag ratio. Therefore, the design thrust in cruise at various Mach numbers can be estimated. For the Pratt & Whitney turbofan engines, thrust (T) is defined by EPR, so that the thrust model is set up as

$$T = f(h, W, M, \text{CAS}, \text{EPR}, \dot{m}_f). \tag{3.23}$$

For GE turbofan engines, the rpm of the low-pressure compressor (N_1) is used to set the level of thrust, so that the thrust model is set up as

$$T = f(h, W, M, \text{CAS}, N_1, \dot{m}_f). \tag{3.24}$$

In the present study, the P & W turbofan engines powering one twin-jet transport will be illustrated. The actual thrust in operation is obtained by using the recorded variables in the FDR, in particular the fuel flow rates.

In the present study, the P & W engines powering the twin-jet commercial aircraft will be illustrated.

The following climb equation [13] is to be satisfied in the least-square sense over a 5-second internal:

$$\frac{W}{g}\frac{dV}{dt} = T - D - W\sin\gamma, \tag{3.25}$$

$$\frac{D}{W} = \frac{D}{L}\cos\gamma. \tag{3.26}$$

All these equations are still valid in descent with negative climb angles (γ). The above equations are further employed for parameter identification in the process of modeling.

Once the thrust model is generated as a function of h, W, M, CAS, EPR, and $\dot{m}_f$ with the flight conditions of climbing, cruise, and descent, one can estimate the thrust magnitude by inserting these flight variables from the FDR into the model.

3.5. Fuzzy-Logic Unsteady Aerodynamic Models

Modeling means to establish the numerical relationship among certain variables of interest. In the fuzzy-logic model, more complete necessary influencing flight variables can be included to capture all possible effects on aircraft response to atmospheric disturbances. For longitudinal aerodynamics, the models are assumed to be of the form [6]

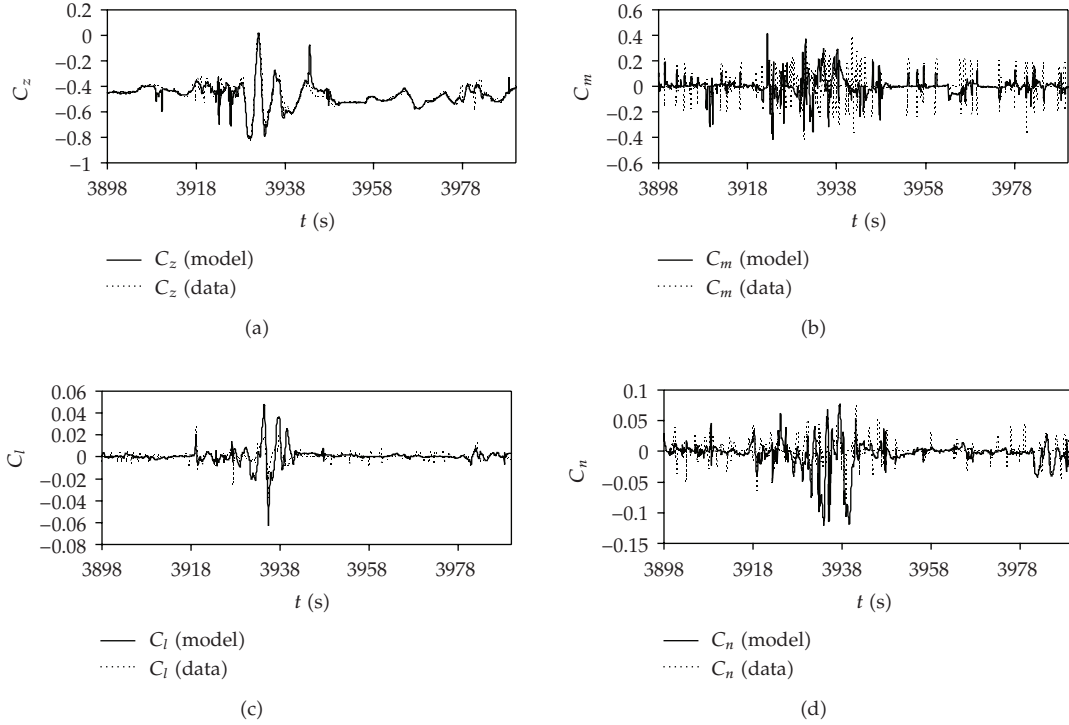$$C_x, C_z, C_m = f(\alpha, \dot{\alpha}, q, k_1, \beta, \delta_e, M, p, \delta_s, \bar{q}), \tag{3.27}$$

Figure 2: Predicted aerodynamic coefficients in normal force and moments for a twin-jet transport encountering severe clear-air turbulence at cruise altitudes around 10,050 m.

where the left-hand side represents the coefficients of axial force (C_x), normal force (C_z), and pitching moment (C_m), respectively. All variables on the right-hand side of (3.22) have been defined in the previous section. It should be noted that the stabilizer angle (δ_s) is included here, because it varies, though slowly, in flight to provide pitch trim (i.e., reducing the total static pitching moment to 0.0). The roll rate is included here because it is known that an aircraft under high aerodynamic loads at transonic speeds may have its longitudinal stability derivatives affected when additional disturbance due to roll rate is imposed.

For the lateral-directional aerodynamics [8],

$$C_y, C_l, C_n = f(\alpha, \beta, \phi, p, r, k_2, \delta_a, \delta_r, M, \dot{\alpha}, \dot{\beta}), \tag{3.28}$$

where the left-hand side represents the coefficients of side force (C_y), rolling moment (C_l), and yawing moment (C_n), respectively.

4. Numerical Results and Discussions

In the present study, the accuracy of the established unsteady aerodynamic models with six aerodynamic coefficients by using FLM technique is estimated by the sum of squared errors (SSEs) and the square of multiple correlation coefficients (R^2). Figure 2 presents the aerodynamic coefficients of normal force C_z, pitching moment C_m, rolling moment C_l, and yawing moment C_n predicted by the unsteady aerodynamic models. The predicted data by

the final models have good agreement with the flight data. The C_m-data scattering is most likely caused by turbulence-induced buffeting on the structure, in particular on the horizontal tail. Once the aerodynamic models are set up, one can calculate all necessary derivatives to analyze the stability.

The fuzzy-logic aerodynamic models are capable of generating the continuous derivatives for the static and dynamic stability study of a twin-jet transport in turbulence response. Firstly, how the fuzzy-logic prediction is achieved will be illustrated with one numerical example in the C_z calculation. At first, the range for each variable is defined to be larger than what actually occurred in the present set of C_z data as follows: $[\alpha] = [-13, 12]$, $[\dot{\alpha}] = [54, 50]$, $[q] = [-20, 10]$, $[k_1] = [0, 0.6]$, $[\beta] = [-7, 3]$, $[\delta_e] = [-10, 6]$, $[M] = [0, 1.6]$, $[p] = [-24, 38]$, $[\delta_s] = [-3, 3]$, and $[\bar{q}] = [4.964, 21.746]$.

For the first cell $(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$, the coefficients in (2.1) after model training are found to be $p_k^1 = (2.61755, 1.26662, 1.42338, 2.07962, -0.44241, 2.78017, 1.78150, 1.30818, 1.82872, 1.67592, 1.13787)$.

Assume that in the following flight conditions C_z is to be predicted: $\alpha = 6.91015$ deg.; $\dot{\alpha} = 2.95510$ deg/sec; $q = 1.16609$ deg/sec; $k_1 = 0.01965$; $\beta = -1.55252$; $\delta_e = 0.68120$ deg; $M = 0.77279$; $p = -2.62359$ deg/sec; $\delta_s = -0.13930$ deg; $\bar{q} = 11.0545$ kpa.

These values of variables are converted to $[0, 1]$. For example, $x_\alpha = [6.91015 - (-13)]/[12 - (-13)] = 0.79641$.

Other variable values are converted in the same way. It follows that the cell internal function becomes $P^1 = 2.61755 + (1.26662) * (0.79641) + (1.42338) * (0.54764) + (2.07962) * (0.70554) - (0.44241) * (0.03275) + (2.78017) * (0.54475) + (1.7815) * (0.66758) + (1.30818) * (0.48299) + (1.82872) * (0.34478 + (1.67592) * (0.47678) + (1.13787) * (0.3730) = 11.04817$.

The membership functions for the first cell are exactly equal to x_r, being 0.79641, 0.54764, and so forth. Their product can be calculated to be $1.08536E-004$. Therefore, the contribution of the first cell to the total output is $11.04817 * 1.08536E - 004 = 1.19912E - 003$.

The total output from all cells can be calculated to be 5.9962; while the denominator in (2.6) is calculated to be 7.46459. Therefore, the final prediction is 0.8033. Comparing with data of 0.81038, this prediction has an error of -0.88%.

To examine the stability characteristics, it is imperative to understand the flight environment in detail. The corresponding flight data are presented in Figure 3. In general, one uses a_n instead of a_z for turbulence intensity study. Note a_n is the same as a_z in this paper. The variation of normal acceleration is presented in Figure 3(a), showing the highest a_n being 1.75 g around $t = 3930$ seconds and the lowest being 0.02 g around $t = 3932$ seconds. Figure 3(b) shows that α is approximately in phase with a_n. When a_n is the highest (around $t = 3930$ seconds), the aircraft rapidly plunging downward with the altitude (h) reaching the lowest as shown in Figure 3(c); and α is highest about 6.5 deg. in Figure 3(b). At the same time, M is around 0.77 in Figure 3(d). Since α reaches a value about 6.5 deg in transonic flight, compressibility effect is important. It should be noted that the turbulent vertical wind field was not measured or estimated in the FDR, but is included in the total α.

The aerodynamic derivatives extracted from the unsteady aerodynamic models can be calculated with two approaches, first one being the central difference method [7] and the other being the forced small-amplitude oscillation method [14]. The longitudinal stability derivative $(C_{m\alpha})$ is extracted from the model of C_m. It is evaluated with the central difference approach as follows:

$$C_{m\alpha} = \frac{[C_m(\alpha + \Delta\alpha, \ldots) - C_m(\alpha - \Delta\alpha, \ldots)]}{2\Delta\alpha}, \tag{4.1}$$
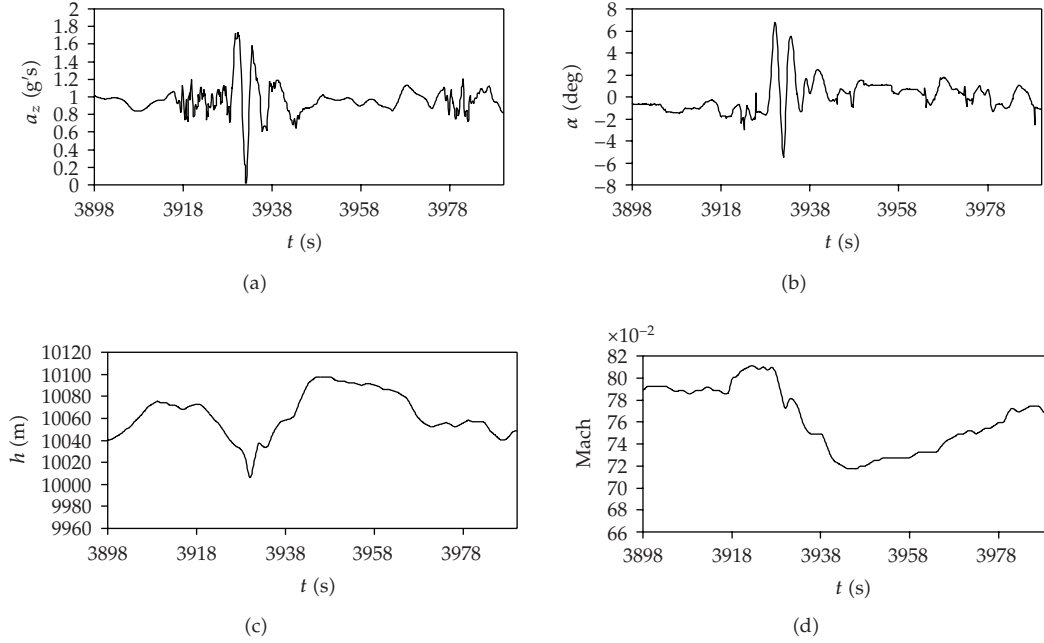
(a)

(b)

(c)

(d)

Figure 3: The time history of flight variables for a twin-jet transport in sever clear-air turbulence at the altitude around 10,050 m in transonic flight.

where $\Delta\alpha = 0.5$ degree represents that α is perturbed by 0.5 degree while keeping all other variables unchanged. The roll damping (C_{lp}) is extracted from the models of C_l with the central difference approach as follows:

$$C_{lp} = \frac{[C_l(\ldots, p + \Delta p, \ldots) - C_l(\ldots, p - \Delta p, \ldots)]}{2\Delta p}, \tag{4.2}$$

where Δp deg/sec. Similarly, all other aerodynamic derivatives are calculated by using the same method.

The time period between 3927.5 seconds and 3932.5 seconds is emphasized in evaluating the stability characteristics, because of the plunging motion that affects the flight safety the most. In order to evaluate the variations in characteristics, the units of all aerodynamic derivatives are converted to rad^{-1}. The main longitudinal and lateral-directional stability derivatives along the flight path are presented in Figure 4. It should be noted that these derivatives are evaluated at the instantaneous conditions, instead of about the trim conditions as have been traditionally done. From the point of view in static stability, initially, the configuration has longitudinal stability ($C_{z\alpha} < 0$ and $C_{m\alpha} < 0$) as shown in Figure 4(a), stable longitudinal damping ($C_{mq} < 0$) in Figure 4(b), lateral stability ($C_{l\beta} < 0$) and directional stability ($C_{n\beta} > 0$) in Figure 4(c), small roll damping ($C_{lp} < 0$) and insufficient directional damping (C_{nr} small or positive) in Figure 4(d). During the plunging motion, in the period between $t = 3928.5$ seconds. and $t = 3930.5$ seconds, $C_{m\alpha} > 0$ and $C_{l\beta} > 0$, so that the static stability becomes unstable. The aerodynamic instability is most likely caused by the motion that produces a time-dependent pressure distribution on the aircraft surface involving compressibility effects.
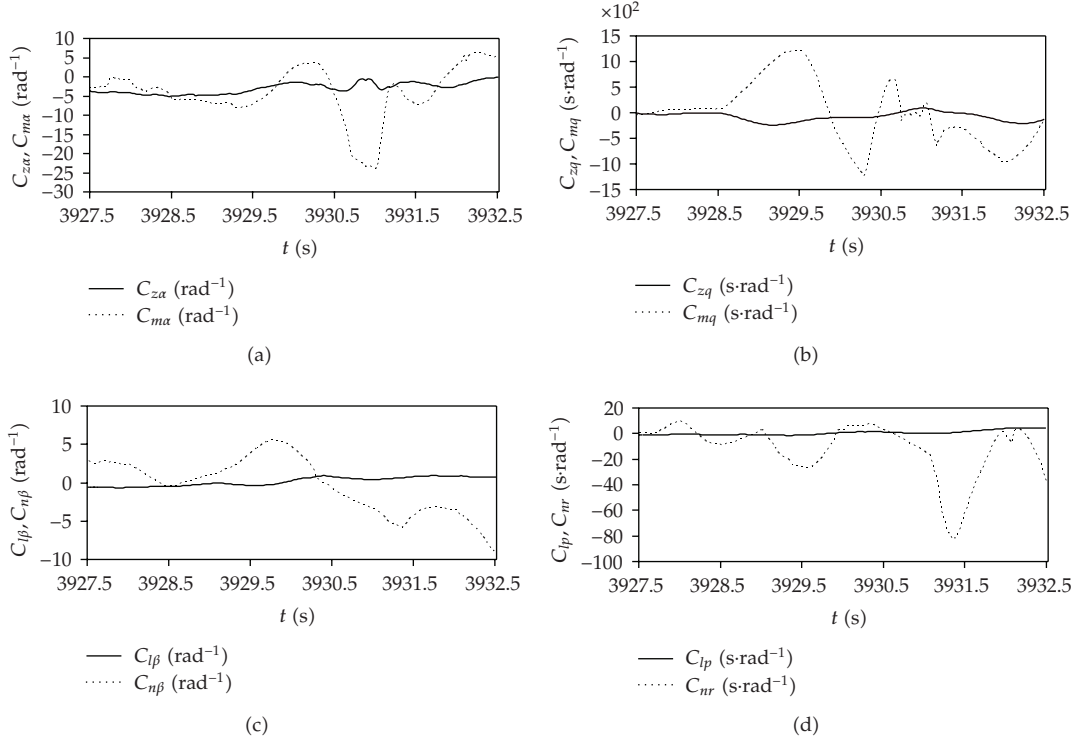
Figure 4: The time history of main longitudinal and lateral-directional of the static stability derivatives along the flight path.

Figure 5 presents the time history of main longitudinal and lateral-directional oscillatory derivatives along the flight path to associate with $\dot{\alpha}$ and $\dot{\beta}$-derivatives. Note that in Figure 5(a), the oscillatory derivatives are defined as

$$\left(C_{mq}\right)_{\text{osc}} = C_{mq} + C_{m\dot{\alpha}}, \tag{4.3}$$

$$\left(C_{zq}\right)_{\text{osc}} = C_{zq} + C_{z\dot{\alpha}} \tag{4.4}$$

In Figure 4(c), the oscillatory derivatives are defined as

$$\left(C_{lp}\right)_{\text{osc}} = C_{lp} + C_{l\beta} \sin\alpha, \tag{4.5}$$

$$\left(C_{nr}\right)_{\text{osc}} = C_{nr} - C_{n\beta} \cos\alpha. \tag{4.6}$$

During the plunging motion, the values have some differences between oscillatory and damping derivatives in Figures 5(a) and 5(c) due to the effects of the dynamic derivatives (i.e., $\dot{\alpha}$ and $\dot{\beta}$-derivatives). The effects of $\dot{\alpha}$-derivative on $\left(C_{zq}\right)_{\text{osc}}$, and $\dot{\beta}$-derivative on $\left(C_{lp}\right)_{\text{osc}}$ are small. However, the effect of $\dot{\alpha}$-derivative on $\left(C_{mq}\right)_{\text{osc}}$ is to improve the stability in pitch after $t = 3929.5$ seconds; while the effects of $\dot{\beta}$-derivative is to cause the directional stability more unstable (i.e., $\left(C_{nr}\right)_{\text{osc}}$ more positive). These results indicate that the turbulent crosswind has the effects on directional stability and damping. Although the
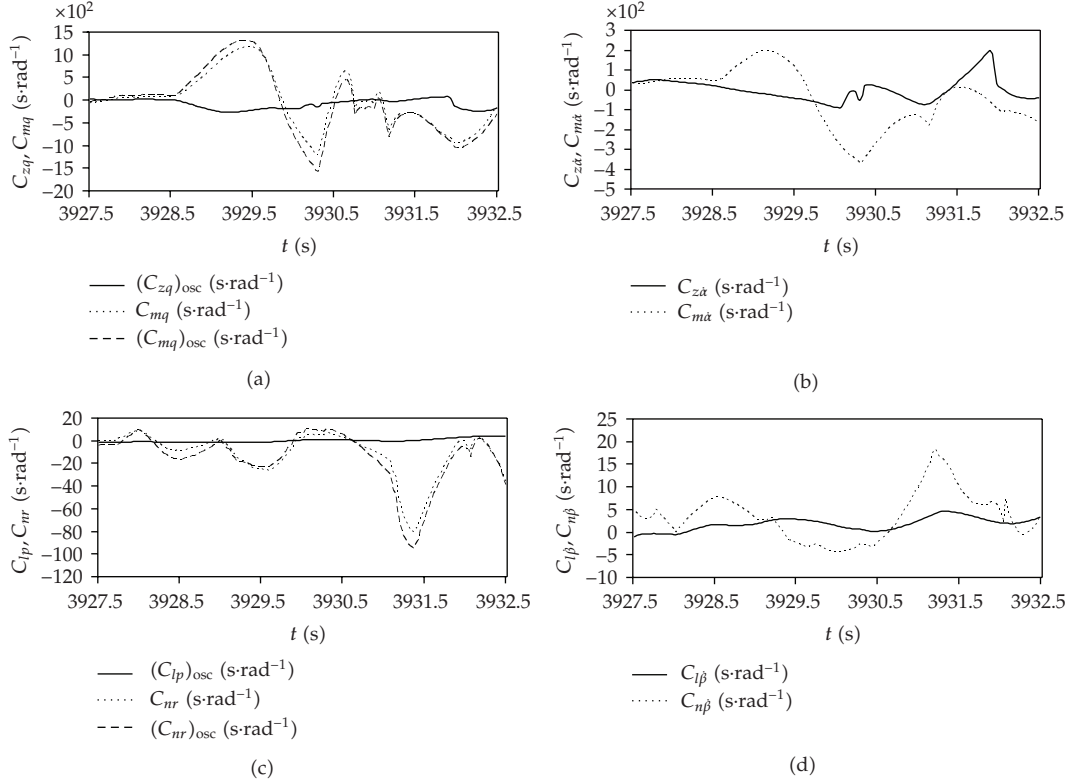
(a)

(b)

(c)

(d)

Figure 5: The time history of main longitudinal and lateral-directional oscillatory derivatives along the flight path.

dynamic derivatives tend to be small for the present configuration, these are much helpful to understand the unknown factors of instability characteristics. To be stable, $(C_{zq})_{\mathrm{osc}} < 0$, $(C_{mq})_{\mathrm{osc}} < 0$, $(C_{lp})_{\mathrm{osc}} < 0$, and $(C_{nr})_{\mathrm{osc}} < 0$. Physically, if it is unstable, the motion will be divergent in oscillatory motions.

5. Concluding Remarks

The main objective in this paper was to illustrate the nonlinear unsteady aerodynamic models based on the FLM technique having the capability to evaluate the variations in stability of commercial aircraft with adverse weather effects. It was shown that the FLM technique was capable of handling nonlinear and unsteady aerodynamic environment exhibited for a twin-jet transport in severe clear-air turbulence with sudden plunging motion in transonic flight. The predicted results showed that the models could produce relatively accurate aerodynamic coefficients and several derivatives for the assessment of stability characteristics, especially for the sensitive study of unknown factors in adverse weather conditions.

Acknowledgments

References

[1] C. T. Weng, C. E. Lan, and C. S. Ho, "Analysis of dynamic ground effect for a jet transport in crosswind," AIAA Paper 2004–5066, August 2004.

[2] R. E. Maine and K. W. Iliff, "User's manual for MMLE3, a general FORTRAN program for maximum likelihood parameter estimation," Tech. Rep. TP-1563, NASA, November 1980.

[3] V. Klein, J. G. Batterson, and P. C. Murphy, "Determination of airplane model structure from flight data by using modified stepwise regression," Tech. Rep. TP-1916, NASA, 1981.

[4] L. A. Zadeh, "Outline of new approach to the analysis of complex systems and decision processes," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, pp. 28–44, 1973.

[5] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 15, no. 1, pp. 116–132, 1985.

[6] Z. Wang, C. E. Lan, and J. M. Brandon, "Fuzzy logic modeling of nonlinear unsteady aerodynamics," AIAA Paper 98–4351, August 1998.

[7] Z. Wang, J. Li, C. E. Lan, and J. M. Brandon, "Estimation of unsteady aerodynamic models from flight test data," AIAA Paper 2001–4017, August 2001.

[8] Z. Wang, J. Li, C. E. Lan, and J. M. Brandon, "Estimation of lateral-directional unsteady aerodynamic models from flight test data," AIAA Paper 2002–4626, August 2002.

[9] Y. N. Lee and C. E. Lan, "Estimation of engine integrity through fuzzy logic modeling," AIAA Paper 2003–6817, November 2003.

[10] J. Roskam, *Airplane Flight Dynamics and Automatic Flight Controls—Part I*, DAR Corporation, Lawrence, Kan, USA, 2003.

[11] T. Theodorsen, "General theory of aerodynamic instability and the mechanism of flutter," NACA Report 496, 1935.

[12] C. T. Weng, *Aerodynamic analysis of a jet transport in windshear encounter during landing based on FDR data*, dissertation for Doctor of Philosophy, Institute of Aeronautics and Astronautics, National Cheng Kung University, Tainan, Taiwan, 2004.

[13] C. E. Lan and J. Roskam, *Airplane Aerodynamics and Performance*, DAR Corporation, Lawrence, Kan, USA, 1997.

[14] C. E. Lan, S. Bianchi, and J. M. Brandon, "Estimation of nonlinear aerodynamic roll models for identification of uncommanded rolling motions," *Journal of Aircraft*, vol. 45, no. 3, pp. 916–922, 2008.

Research Article

Limiting Phase Trajectories and Resonance Energy Transfer in a System of Two Coupled Oscillators

L. I. Manevitch,[1] A. S. Kovaleva,[2] and E. L. Manevitch[3]

[1] *N. N. Semenov Institute of Chemical Physics, Russian Academy of Sciences, 4, Kosygina street, Moscow 119991, Russia*
[2] *Space Research Institute, Russian Academy of Sciences, Moscow 117997, Russia*
[3] *Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA*

Correspondence should be addressed to L. I. Manevitch, lmanev@chph.ras.ru

We study a problem of energy exchange in a system of two coupled oscillators subject to 1:1 resonance. Our results exploit the concept of limiting phase trajectories (LPTs). The LPT, associated with full energy transfer, is, in certain sense, an alternative to nonlinear normal modes characterized by conservation of energy. We consider two benchmark examples. As a first example, we construct an LPT and examine the convergence to stationary oscillations for a Duffing oscillator subjected to resonance harmonic excitation. As a second example, we treat resonance oscillations in a system of two nonlinearly coupled oscillators. We demonstrate the reduction of the equations of motion to an equation of a single oscillator. It is shown that the most intense energy exchange and beating arise when motion of the equivalent oscillator is close to an LPT. Damped beating and the convergence to rest in a system with dissipation are demonstrated.

1. Introduction

The problem of passive irreversible transfer of mechanical energy (referred to as energy pumping) in oscillatory systems has been studied intensively over last decades; see, for example, [1, 2] for recent advances and references. In this case, the key role of transient process is evident, in contrast to great majority of conventional problems of nonlinear dynamics, in which the main attention has been given to nonlinear normal modes (NNMs), characterized by the conservation of energy. Recent studies [3, 4]; have shown that the NNM approach is effective in the case of weak energy exchange, while the concept of the limiting phase trajectories (LPT) can be used to describe intense energy exchange between weakly coupled oscillators or oscillatory chains.

The limiting phase trajectory (LPT) has been introduced [3] as a trajectory corresponding to oscillations with the most intensive energy exchange between weakly coupled oscillators or an oscillator and a source of energy; the transition from energy exchange to energy localization at one of the oscillators is associated with the disappearance of the LPT. Recently, the LPT ideas have been applied to the analysis of resonance-forced vibrations in a 1DOF and 2DOF nondissipative system [3–6]. An explicit expression of the LPT in a single oscillator [5, 6] is prohibitively difficult for practical utility. The purpose of the present paper is to show that explicit asymptotic solutions can be obtained for a class of problems that are associated with the maximum energy exchange, and introduce relevant techniques.

The paper is organized as follows. The first part is concerned with the construction of the LPT for the Duffing oscillator subject to 1 : 1 resonance harmonic excitation. In Section 2 we briefly reproduce the results of [5, 6]. We derive the averaged equations determining the slowly varying envelope and the phase of the nonstationary motion and then construct the LPT for different types of motion. Section 3 introduces the nonsmooth temporal transformations as an effective tool of the nonlinear analysis. Section 4 examines the transformations of the LPT into stationary motion in the weakly dissipated system.

In the remainder of the paper (Section 5), we analyse the dynamics of a 2DOF system. The system consists of a linear oscillator of mass M (the source of energy) coupled with a mass m (an energy sink) by a nonlinear spring with a weak linear component. Excitation is due to an initial impulse acting upon the mass M. It is shown that motion of the overall structure can be divided into two stages. The first stage is associated with the maximum energy exchange between the oscillators; here, motion is close to beating in a nondissipative system. At the second stage, trajectories of both masses in the damped system are approaching to rest. The task is to construct an explicit asymptotic solution describing both stages of motion for each oscillator. To this end, we reduce the equations of a 2DOF system to an equation of a single oscillator and then find beating oscillations characterizing the most intense energy exchange in a nondissipative system. A special attention is given to a difference between the dynamics of the Duffing oscillator and an equivalent oscillator in the presence of dissipation. While the Duffing oscillator is subjected to harmonic excitation, the 2DOF system is excited by an initial impulse applied to the linear oscillator. The response of the linear oscillator, exponentially vanishing at $t \to \infty$, stands for an external excitation for the nonlinear energy sink. We examine the transformation of beating to damped oscillations. We show that a solution of the system, linearized near the rest state, is sufficient to describe the second part of the trajectory.

2. Resonance Oscillations of the Duffing Oscillator

2.1. Main Equations

We investigate the transient response of the Duffing oscillator in the presence of resonance 1:1. The dimensionless equation of motion is

$$\frac{d^2u}{dt^2} + 2\varepsilon\gamma\frac{du}{dt} + u + 8\alpha\varepsilon u^3 = 2\varepsilon F \sin(1 + \varepsilon s)t, \tag{2.1}$$

where $\varepsilon > 0$ is a small parameter. We recall that maximum energy pumping from the source of excitation into the oscillator takes place if the oscillator is initially at rest; this corresponds to the initial conditions

$$t = 0^+, \qquad u = 0, \qquad \frac{du}{dt} = 0. \tag{2.2}$$

An orbit satisfying conditions (2.2) is said to be *the limiting phase trajectory* [3].

In order to describe the nonlinear dynamics, we invoke a complex-valued transformation [7–9]. Introduce the variables ψ and ψ^*, such that

$$u = \frac{1}{2i}(\psi - \psi^*), \qquad v = \frac{1}{2}(\psi + \psi^*),$$

$$\psi = v + iu, \qquad \psi^* = v - iu, \tag{2.3}$$

where $i = \sqrt{-1}$; the asterisk denotes complex conjugate. It will be shown that only one complex function is sufficient for a complete description of the dynamics. Inserting ψ, ψ^* from (2.3) into (2.1), a little algebra shows that (2.1) is equivalent to the following (still exact) equation of motion

$$\frac{d\psi}{dt} - i\psi + \varepsilon i\alpha(\psi - \psi^*)^3 + \varepsilon\gamma(\psi + \psi^*) - 2\varepsilon F \sin(1 + \varepsilon s)t = 0. \tag{2.4}$$

Applying the multiple scales method [10, 11], we construct an approximate solution of (2.4) as an expansion

$$\psi(t, \varepsilon) = \psi_0(\tau_0, \tau_1) + \varepsilon\psi_1(\tau_0, \tau_1) + \cdots,$$

$$\frac{d}{dt} = \frac{\partial}{\partial\tau_0} + \varepsilon\frac{\partial}{\partial\tau_1}, \frac{d^2}{dt^2} = \frac{\partial^2}{\partial\tau_0^2} + 2\varepsilon\frac{\partial^2}{\partial\tau_0\partial\tau_1} + \cdots, \tag{2.5}$$

where $\tau_0 = t$ and $\tau_1 = \varepsilon t$ are the fast and slow time-scales, respectively. A similar representation is valid for the function ψ^*. Then we substitute expressions (2.5) in (2.4) and equate the coefficients of like powers of ε. In the leading order approximation, we obtain

$$\frac{\partial\psi_0}{\partial\tau_0} - i\psi_0 = 0,$$

$$\psi_0(\tau_0, \tau_1) = \varphi_0(\tau_1)e^{i\tau_0} \tag{2.6}$$

A slow function $\varphi_0(\tau_1)$ will be found at the next level of approximation. Equating the coefficients of order ε leads to

$$\frac{\partial\psi_1}{\partial\tau_0} + \frac{d\varphi_0}{d\tau_1}e^{i\tau_0} - i\psi_1 + i\alpha(\psi_0 - \psi_0^*)^3 + \gamma(\psi_0 + \psi_0^*) + iF\left(e^{i(\tau_0 + s\tau_1)} - e^{-i(\tau_0 + s\tau_1)}\right) = 0, \tag{2.7}$$

In order to avoid the secular growth of $\psi_1(\tau_0, \tau_1)$ in τ_0, that is, avoid a response not uniformly valid with increasing time, we eliminate resonance terms from (2.7). This yields the following equation for φ_0:

$$\frac{d\varphi_0}{d\tau_1} + \gamma\varphi_0 - 3i\alpha|\varphi_0|^2\varphi_0 = -iFe^{is\tau_1}, \quad \varphi_0(0) = 0. \tag{2.8}$$

Next we introduce the polar representation

$$\varphi_0 = ae^{i\delta}, \tag{2.9}$$

where a and δ represent a real amplitude and a real phase of the process $\varphi_0(\tau_1)$. Inserting (2.9) into (2.8) and setting separately the real and imaginary parts of the resulting equations equal to zero, (2.8) is transformed into the system

$$\frac{da}{d\tau_1} + \gamma a = -F\sin\Delta,$$

$$a\frac{d\Delta}{d\tau_1} = -sa + 3\alpha a^3 - F\cos\Delta, \tag{2.10}$$

where $a > 0$, $\Delta = \delta - s\tau_1$. It now follows from (2.3), (2.9) that

$$u(t, \varepsilon) = a(\tau_1)\sin(t + \Delta(\tau_1)) + O(\varepsilon). \tag{2.11}$$

This means that the amplitude $a(\tau_1)$ and the phase $\Delta(\tau_1)$ completely determine the process $u(t, \varepsilon)$ (in the leading-order approximation). Note that $a = 0$ if the oscillator is not excited.

2.2. Critical Parameters and LPTs of the Undamped Oscillator

In this section, we recall main definitions and results concerning the dynamics of the nondissipative system. In the absence of damping, system (2.10) is rewritten as

$$\frac{da}{d\tau_1} = -F\sin\Delta,$$

$$a\frac{d\Delta}{d\tau_1} = -sa + 3\alpha a^3 - F\cos\Delta. \tag{2.12}$$

It is easy to prove that system (2.12) conserves the integral of motion

$$H = \frac{3}{4}\alpha a^4 - \frac{sa^2}{2} - Fa\cos\Delta = H_0, \tag{2.13}$$

where H_0 depends on initial conditions. In the phase plane, the LPT corresponds to the contour $H = 0$, as only in this case a system trajectory goes through the point $a = 0$. Taking $H_0 = 0$, we obtain the following expression:

$$H = a\left(\frac{3}{4}\alpha a^3 - \frac{s}{2}a - F\cos\Delta\right) = 0. \tag{2.14}$$

Formula (2.14) implies that the LPT has two branches, the first branch is $a = 0$; the second branch satisfies the cubic equation

$$\frac{3}{4}\alpha a^3 - \frac{s}{2}a - F\cos\Delta = 0. \tag{2.15}$$

Equality (2.15) determines the second initial condition $a(0^+) = 0$, $\cos\Delta(0^+) = 0$. We suppose that $da/d\tau1 > 0$ at $\tau_1 = 0^+$; under this assumption, $\Delta(0^+) = -\pi/2$. Hence the initial conditions for the LPT take the form

$$\tau_1 = 0^+, a(0^+) = 0, \Delta(0^+) = -\frac{\pi}{2}. \tag{2.16}$$

Throughout this paper, we write 0 instead of 0^+, except as otherwise noted.

Next we determine critical parameters of system (2.12). The steady states of (2.12) satisfies the equations

$$\frac{da}{d\tau_1} = 0, \qquad \frac{d\Delta}{d\tau_1} = 0. \tag{2.17}$$

The second equation is equivalent to the equality

$$-sa + 3\alpha a^3 - F\,\mathrm{sgn}(\cos\Delta) = 0, \tag{2.18}$$

where $\cos\Delta = \pm1$. We analyze the properties of (2.18) considering the properties of its discriminant D_1

$$D_1 = \frac{1}{9\alpha^2}\left(\frac{F^2}{4} - \frac{s^3}{81\alpha}\right). \tag{2.19}$$

If $D_1 < 0$, (2.18) has 3 different real roots; if $D_1 > 0$, (2.18) has a single real and two complex conjugate roots; if $D_1 = 0$, two real roots will merge; see, for example, [12]. The latter condition gives the first critical value of the parameter α

$$\alpha_1^* = \frac{4s^3}{81F^2}. \tag{2.20}$$

A straightforward investigation proves that, if $\alpha > \alpha_1^*$ (strong nonlinearity), there exists only a single stable centre C_+: $(0, a_+)$ (Figure 1); if $\alpha < \alpha_1^*$ (weak nonlinearity), there exist two
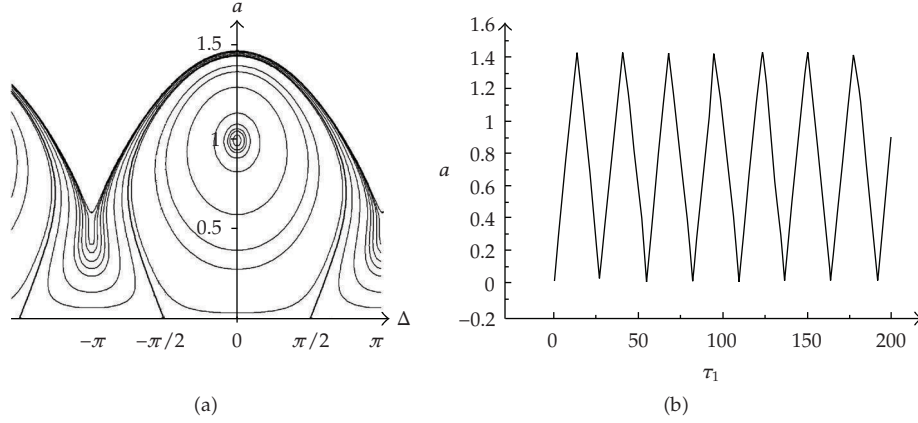
(a)   (b)

Figure 1: Phase portrait (a) and plot of $a(\tau_1)$ (b) for $s = 0.4$, $F = 0.13$, $\alpha = 0.187 = \alpha_1^*$.

stable centres C_-: $(-\pi, a_-)$, C_+: $(0, a_+)$, and an intermediate unstable hyperbolic point O: $(-\pi, a_0)$, see Figures 2 and 3.

We now suppose that $\alpha < \alpha_1^*$, that is the system may exhibit both types of oscillations, either near $\Delta = -\pi$, or near $\Delta = 0$ (Figures 2–4). In both cases, the LPT begins at $a = 0$, $\Delta = -\pi/2$ but the run of the LPT depends on the relationship between the parameters. In order to find a critical value $\alpha_2^* < \alpha_1^*$ ensuring the transition from small to large oscillations, we analyze (2.15). Consider the discriminant of (2.15)

$$D_2 = \frac{4}{\alpha^2}\left(F^2 - \frac{2s^3}{81\alpha}\right) \tag{2.21}$$

In the critical case $\alpha = \alpha_2^*$, an unstable hyperbolic point coincides with the maximum of the left branch of the LPT at $\Delta = -\pi$ (Figure 2). This means that $D_2 = 0$ at α_2^*, or

$$\alpha_2^* = \frac{2s^3}{81F^2} = \frac{\alpha_1^*}{2}, \tag{2.22}$$

which defines a boundary between small quasilinear ($\alpha < \alpha_2^*$) and large nonlinear ($\alpha > \alpha_2^*$) oscillations. In particular, for $s = 0.4$, $F = 0.13$ we obtain $\alpha_2^* = 0.0935$ (Figure 2).

Figures 3 and 4 are plotted for $\alpha < \alpha_2^*$ and $\alpha > \alpha_2^*$, respectively. In Figure 3(a), one can see the LPT encircling the center C_- of relatively small oscillations; Figure 4(a) shows the LPT encircling the centre C_+ of large oscillations; this case is associated with the maximum energy absorption. Figures 3(b) and 4(b) demonstrate the behavior of the function $a(\tau_1)$ corresponding to the respective branch of the LPT. Note that both branches of the LPT begin at the same point $a = 0$, $\Delta = -\pi/2$.

If $\alpha = \alpha_1^* = 2\alpha_2^*$, the above-mentioned coincidence of the stable and unstable points at $\Delta = -\pi$ results in the transformation of the phase portraits (Figure 1) and disappearance of the stable centre C_-. Figure 1(a) demonstrates a single stable fixed point at $\Delta = 0$.
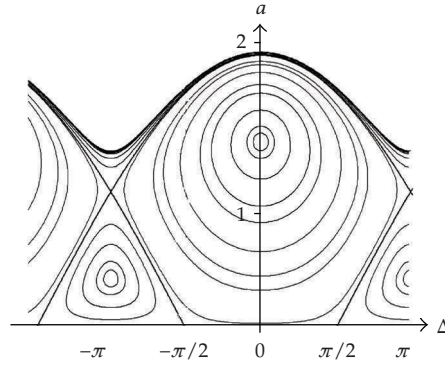
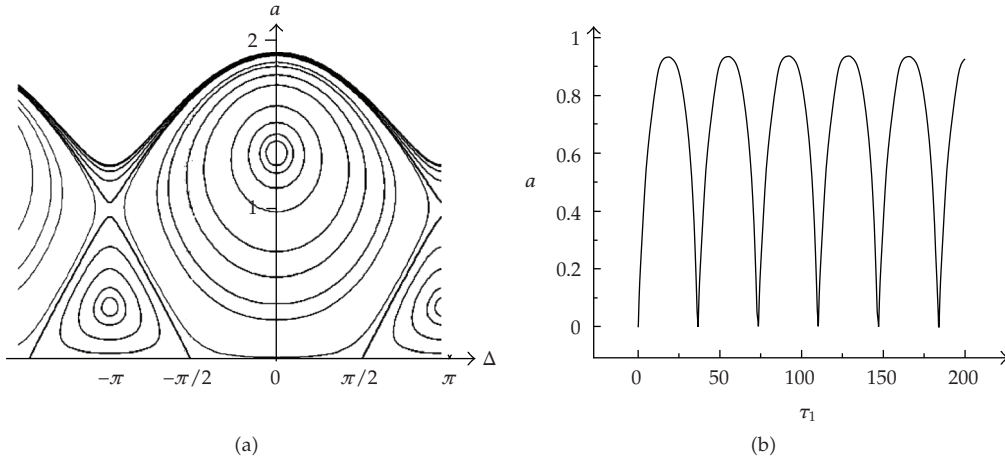**Figure 2:** Passage from small to large oscillations: $\alpha = 0.0935 = \alpha_2^*$.

(a)

(b)

Figure 3: Phase portrait (a) and plot of $a(\tau_1)$ (b) for quasilinear oscillations: $s = 0.4$, $F = 0.13$, $\alpha = 0.093 < \alpha_2^*$.
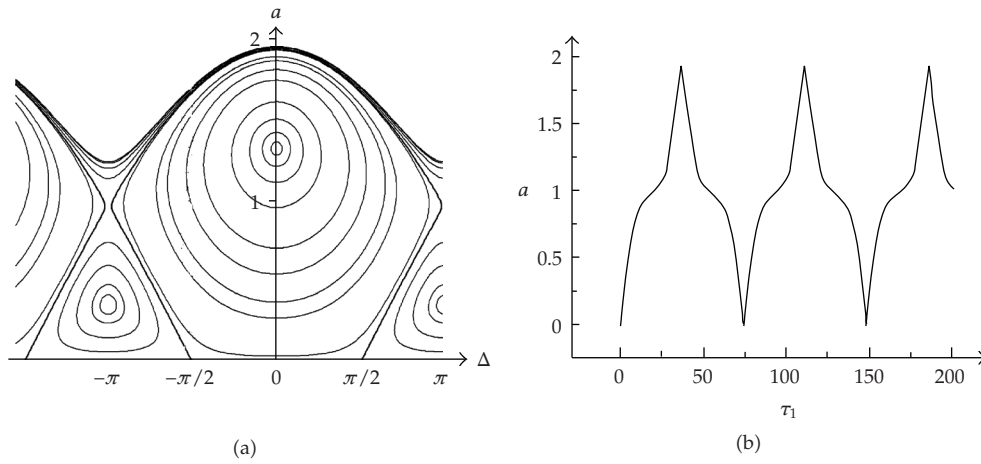

(a)

(b)

Figure 4: Phase portrait (a) and plot of $a(\tau_1)$ (b) for strongly nonlinear oscillations: $s = 0.4$, $F = 0.13$, $\alpha_2^* < \alpha = 0.094 < \alpha_1^*$.
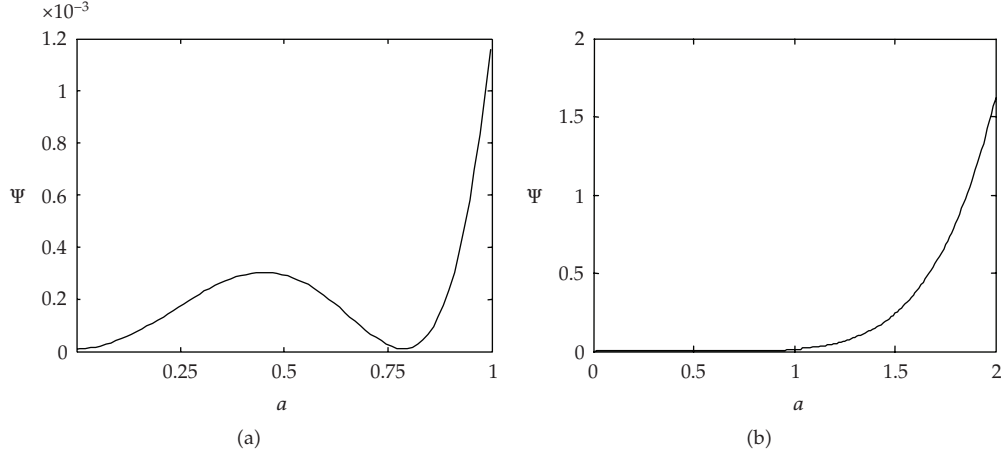
Figure 5: Potential $\Psi(a)$ in the small (a) and large (b) scales.

2.3. Reduction to the 2nd-Order Equation

For the further analysis, it is convenient to reduce the equation of the LPT to the second-order form. Using (2.15) to exclude Δ, we obtain the following equation:

$$\frac{d^2 a}{d\tau_1^2} + f(a) = 0,$$

$$a = 0, \qquad \frac{da}{d\tau_1} = F \quad \text{at } \tau_1 = 0,$$

(2.23)

where

$$f(a) = \frac{a}{4}\left(\frac{3}{2}\alpha a^2 - s\right)\left(\frac{9}{2}\alpha a^2 - s\right) = \frac{d\Psi(a)}{da},$$

$$\Psi(a) = \frac{a^2}{8}\left(\frac{3}{2}\alpha a^2 - s\right)^2.$$

(2.24)

Note that (2.23) can be treated as the equation of a conservative oscillator with potential $\Psi(a)$, yielding the integral of energy

$$E = \frac{1}{2}v^2 + \Psi(a) = \frac{1}{2}F^2, \quad v = \frac{da}{d\tau_1}.$$

(2.25)

Figure 5 depicts the potential $\Psi(a)$ in the small and large scales for system (3.1) with the parameters $\alpha = 1/3$, $s = 1/4$. The phase portrait is given in Figure 6.

The amplitude of oscillations A_0 is defined by the following equality (see Figure 6(b)):

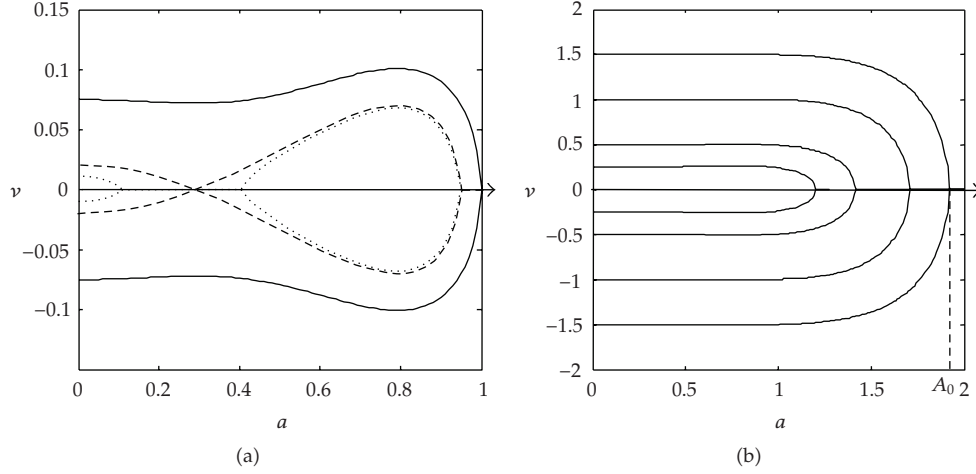$$E = \Psi(A_0) = \frac{1}{2}F^2.$$

(2.26)

(a)

(b)

Figure 6: Phase portraits of (2.23) in the small (a) and large (b) scales.

From (2.25) we have $v = da/d\tau_1 = \pm\sqrt{F^2 - 2\Psi(a)}$. Thus, the half-period of oscillations is defined as

$$T_1(A_0) = \int_0^{A_0} \frac{da}{\sqrt{F^2 - 2\Psi(a)}}. \tag{2.27}$$

It follows from Figures 1 and 6 that a high-energy system is weakly sensitive to the shape of the potential; the orbit is close to the straight line until it reaches the wall of the potential well. This implies that motion of system (2.23) is similar to the dynamics of a particle moving with constant velocity between two motion-limiters. A connection between the smooth and vibro-impact modes of motion in a smooth nonlinear oscillator has been revealed in [13, 14]; a detailed exposition of this approach can be found in [2, Chapter 6].

The vibro-impact hypothesis suggests that the time T_1^* to reach A_0 is calculated as

$$T_1^* = \frac{A_0}{F}. \tag{2.28}$$

Note that $T_1^* < T_1$, as the vibro-impact approximation ignores the deceleration of motion in the vicinity of A_0, when the velocity falls below the maximum level F. Formally, T_1^* can be found from (2.27) by letting $\Psi(a) = 0$.

3. Analysis of the Transient Dynamics

In what follows, we consider the dynamics of a weakly damped oscillator with strong nonlinearity ($\alpha > \alpha^*$); it is often the case of particular interest.

As seen in Figure 7, the damped system exhibits strongly nonlinear behavior on the time interval $[0, T_1^*]$; an instant T_1^* corresponds to the first maximum of the function $a(\tau_1)$. After that motion becomes similar to smooth oscillations about the stationary point. This allows separating the transient dynamics into two stages. While on the interval $0 \leq \tau_1 \leq T_1^*$
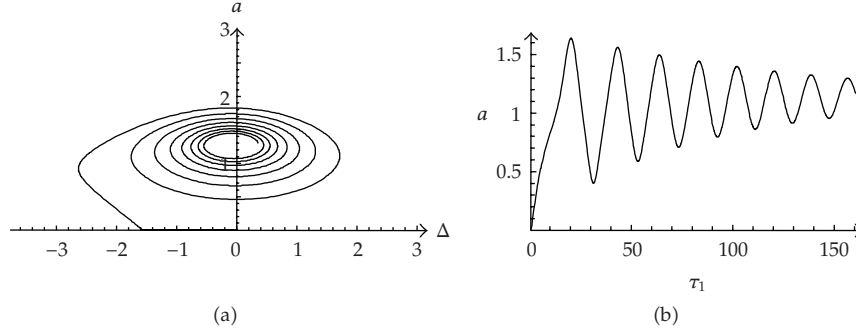
Figure 7: Phase portrait (a) and plot of $a(\tau_1)$ (b) for strongly nonlinear oscillations in a dissipative system.

motion is close to the LPT of the undamped system, at the second stage, $\tau_1 \geq T_1^*$, motion is similar to quasilinear oscillations.

In the remainder of this section we investigate a segment of the trajectory on the interval $[0, T_1^*]$. The task is to calculate an instant T_1^* and the values $a(T_1^*)$ and $\Delta(T_1^*)$ determining the starting point for the second interval of motion. For simplicity, we assume that dissipation is sufficiently small and may be ignored on the interval $[0, T_1^*]$. This allows one to approximate the first part of the trajectory by a corresponding segment of the LPT of the nondissipative system.

As mentioned above, the dynamics of a strongly nonlinear oscillator is similar to free motion of a particle moving with constant velocity between two motion-limiters. This allows us to employ the method of nonsmooth transformations [1, 2] in the study of strongly nonlinear oscillations.

At the first step, we introduce nonsmooth functions $\tau(\phi)$ and $e(\phi) = d\tau/d\phi$ defined as follows:

$$\tau = \tau(\phi) = \frac{2}{\pi}\left|\arcsin\left(\sin\frac{\pi\phi}{2}\right)\right|,$$

$$e(\phi) = 1, \quad 0 < \phi \leq 1, \quad e(\phi) = -1, \quad 1 < \phi \leq 2,$$

(3.1)

where $\phi = \Omega\tau_1$, the frequency Ω will be found below. Plots of functions (3.1) are given in Figure 8. In a general setting, the solution of (2.12) is constructed in the form

$$a(\tau_1) = X_1(\tau) + e(\phi)Y_1(\tau), \qquad \Delta(\tau_1) = X_2(\tau) + e(\phi)Y_2(\tau),$$

$$\frac{d}{d\tau_1} = \Omega\left(e\frac{\partial}{\partial\tau} + \frac{\partial}{\partial\phi}\right).$$

(3.2)

We recall that $\partial e/\partial\phi = \delta(\phi - n)$, where $\delta(\phi - n)$ is Dirac's delta-function, $n = 1, 2, \ldots$. We exclude δ-singularity by requiring

$$Y_{1,2} = 0 \quad \text{at } \tau = 1, 2, \ldots.$$

(3.3)

(a)

(b)

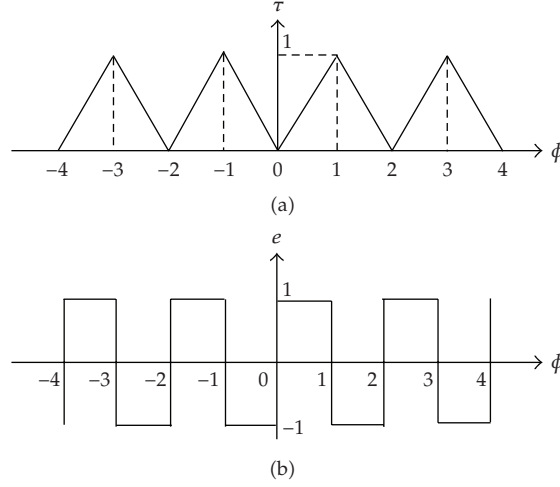**Figure 8:** Functions $\tau(\phi)$ and $e(\phi)$.

This implies that

$$\frac{da}{d\tau_1} = \Omega\left(e\frac{\partial X_1}{\partial \tau} + \frac{\partial Y_1}{\partial \tau}\right),$$

$$\frac{d\Delta}{d\tau_1} = \Omega\left(e\frac{\partial X_2}{\partial \tau} + \frac{\partial Y_2}{\partial \tau}\right) \tag{3.4}$$

provided $Y_{1,2} = 0$ at $\tau = 1, 2, \ldots$. To derive the equations for X_i, Y_i, $i = 1, 2$, we insert (3.4) into (2.12) and separate the terms with and without e. This yields the set of equations

$$\Omega\frac{\partial Y_1}{\partial \tau} = -F\sin X_2\cos Y_2,$$

$$\Omega\frac{\partial X_1}{\partial \tau} = -F\sin Y_2\cos X_2,$$

$$\Omega\left(X_1\frac{\partial Y_2}{\partial \tau} + Y_1\frac{\partial X_2}{\partial \tau}\right) + sX_1 - 3\alpha X_1^3 - 9\alpha X_1 Y_1^2 = -F\cos Y_2\cos X_2, \tag{3.5}$$

$$\Omega\left(X_1\frac{\partial X_2}{\partial \tau} + Y_1\frac{\partial Y_2}{\partial \tau}\right) = F\sin Y_2\sin X_2.$$

It is easy to prove that (3.5) are satisfied by $Y_1 = 0$, $X_2 = 0$. Under these conditions, the variables X_1 and Y_2 satisfy the equations similar to (2.12)

$$\Omega\frac{\partial X_1}{\partial \tau} + F\sin Y_2 = 0,$$

$$\Omega X_1\frac{\partial Y_2}{\partial \tau} + sX_1 - 3\alpha X_1^3 + F\cos Y_2 = 0, \tag{3.6}$$

with the initial conditions $X_1 = 0$, $Y_2 = -\pi/2$ at $\tau = 0$. System (3.6) is integrable, yielding the integral of motion similar to (2.14)

$$\frac{3}{4}\alpha X_1^3 - \frac{s}{2}X_1 - F\cos Y_2 = 0. \tag{3.7}$$

Then, it follows from (3.3) and (3.6) that

$$\frac{\partial X_1}{\partial \tau} = 0 \quad \text{at } Y_2 = 0, \ \tau = 1. \tag{3.8}$$

It is worth noting that equalities (3.8) have a clear physical meaning: *they represent the condition of maximum of* X_1 *at* $Y_2 = 0$.

For the further analysis, it is convenient to transfer (3.6) into the second-order form. Using (3.7) to exclude Y_2, the resulting equation and the initial conditions are written as

$$\Omega_2 \frac{\partial^2 X_1}{\partial \tau^2} + f(X_1) = 0,$$

$$X_1 = 0, \qquad \Omega\frac{dX_1}{d\tau} = F, \tag{3.9}$$

where $f(X_1)$ is defined by (2.24). A precise solution of (3.9), expressed in terms of elliptic functions, is prohibitively difficult for practical utility [6, 13]. In order to highlight the substantial dynamical features, the solution to (3.6), (3.9) is expressed in terms of successive approximations

$$X_1 = x_0 + x_1 + \cdots, \qquad Y_2 = y_0 + y_1 + \cdots, \qquad \Omega = \Omega_0(1 + \varepsilon_1 + \cdots), \tag{3.10}$$

where it is assumed that $|x_1(\tau| \ll |x_0(\tau)|, |y_1(\tau) \ll |y_0(\tau)|, \varepsilon_1 \ll 1$ on an interval of interest. The validity of this assumption will be tested below by numerical simulations. Since the vibro-impact approximation is insensitive to the presence of the potential, the function x_0 is chosen as the solution of the equation

$$\frac{\partial^2 x_0}{\partial \tau^2} = 0 \tag{3.11}$$

with the initial conditions $x_0 = 0$, $\Omega_0 \partial x_0/\partial\tau = F$ at $\tau = 0$. It follows from (3.11) that

$$x_0(\tau) = A_0\tau, \qquad A_0\Omega_0 = F. \tag{3.12}$$

From (3.1) and the maximum condition we have

$$\Omega_0 T_1^* = 1, \quad T_1^* = \Omega_0^{-1} = \frac{A_0}{F},$$

$$x_0(\tau) = F\tau_1, \quad 0 \le \tau_1 \le T_1^*, x_0(\tau(T_1^*)) = x_0(1) = A_0. \tag{3.13}$$

We now recall that system (3.9) is conservative; it possesses the integral of energy

$$E_1 = \frac{1}{2}V_1^2 + \frac{1}{\Omega^2}\Psi(X_1) = \frac{1}{2\Omega^2}F^2, \quad V_1 = \frac{dX_1}{d\tau}, \tag{3.14}$$

where $\Psi(X_1)$ is defined by (2.24). By analogy with (2.26), we obtain

$$\Psi(X_1) = \frac{1}{2}F^2 \quad \text{at } \tau = 1. \tag{3.15}$$

Inserting (3.10), (3.12) into (3.15) and ignoring small terms, we then have the equation to determine A_0

$$\Psi(A_0) = \frac{1}{2}F^2. \tag{3.16}$$

Given A_0, we obtain from $T_1^* = A_0/F$ (see (2.28)).

The approximation x_1 is governed by the following equation:

$$\Omega_0^2 \frac{\partial^2 x_1}{\partial \tau^2} = -f(x_0),$$

$$x_1(\tau) = -\Omega_0^{-2}\int_0^\tau (\tau - \xi)f(A_0\xi)d\xi, \tag{3.17}$$

Given A_0 and Ω_0, formula (3.17) yields

$$x_1(\tau) = -\frac{A_0\tau^3}{4\Omega_0^2}\left[\frac{9\alpha^2}{56}(A_0\tau)^4 - \frac{3s\alpha}{5}(A_0\tau)^2 + \frac{s^2}{6}\right]. \tag{3.18}$$

We now find the function $Y_2(\tau)$. Arguing as above, we construct $Y_2 = y_0 + y_1$, where y_0 can be found from the first equation (3.6), in which we let $X_1 = x_0$. This yields

$$y_0 = -\arcsin\left(\frac{F}{A_0\Omega_0}\right) = -\frac{\pi}{2}, \quad 0 < \tau < 1. \tag{3.19}$$

The term $y_1(\tau)$ is defined by the second equation (3.6). As before, we take $X_1 = x_0$ and exclude $\cos Y_2$ by (3.7) to get

$$\frac{\partial y_1}{\partial \tau} = \frac{1}{\Omega_0}\left(-\frac{s}{2} + \frac{9}{4}\alpha A_0^2\tau^2\right),$$

$$y_1(\tau, t_0) = \frac{1}{\Omega_0}\left(-\frac{s\tau}{2} + \frac{3}{4}\alpha A_0^2\tau^3\right). \tag{3.20}$$
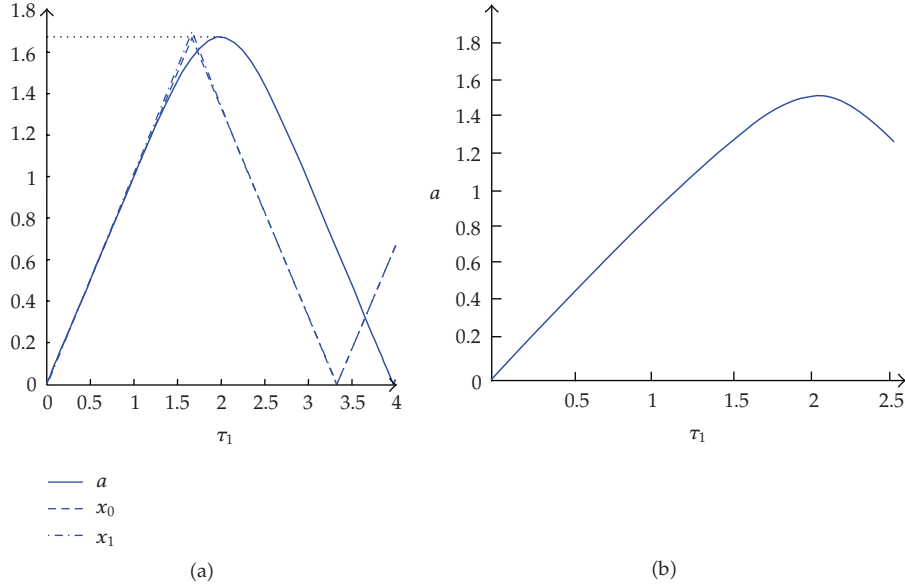
Figure 9: LPT of system (2.12) (a) and solution $a(\tau_1)$ of system (2.10) (b) solid line—numerical solution; dashed line—the leading-order approximation x_0; dot-and-dash line—the first order approximation $x_0 + x_1$.

As an example, we calculate the LPT for system (2.12) with the parameters

$$\gamma = 0, \qquad s = 0.2, \qquad \alpha = 0.333, \qquad F = 1 \tag{3.21}$$

and compare the results with the numerical solution for system (2.10), in which $\gamma = 0.04$.

Calculations by formulas (3.13), (3.16) give $A_0 = 1.67$, $T_1^* = \Omega_0^{-1} = 167$. Thus we have the maximum $M = X_1(T_1^*) \approx 1.67$ at $T_1^* \approx 1.67$ in the leading-order approximation and $M_1 \approx 1.67$ at $T_1^* \approx 2$ for the numerical solution $a(\tau_1)$ (Figure 9(a)); for system (2.10) with $\gamma = 0.05$ the numerical solution gives the maximum $M \approx 1.56$ at $T_1^* \approx 2.1$. This confirms that small dissipation can be ignored over the interval $0 \leq \tau \leq T^*$.

It is easy to check by a straightforward calculation that the correction x_1 is negligible. In a similar way, one can evaluate the small term y_1.

4. Quasilinear Oscillations

In this section, we examine quasilinear oscillations on the second interval of motion, $\tau > 1$. It is easy to see that an orbit of the dissipative system tends to its steady state as $\tau \to \infty$. The steady state O: (a_0, Δ_0) for system (2.10) is determined by the equality

$$a^2 \left[\left(s - 3\alpha a^2 \right)^2 + \gamma^2 \right] = F^2, \tag{4.1}$$

or, for sufficiently small γ,

$$\gamma a_0 = -F \sin \Delta_0, \qquad s a_0 - 3\alpha a_0^3 = -F \cos \Delta_0,$$

$$\Delta_0 \approx -\frac{\gamma a_0}{F} + O(\gamma^3), \qquad a_0 \left(s - 3\alpha a_0^2 \right) = -F + O(\gamma^2). \tag{4.2}$$

Let $\xi = a - a_0$, $\beta = \Delta - \Delta_0$ denote deviations from the steady state. In addition, we must impose the matching conditions

$$a_0 + \xi = x_0^*, \qquad \frac{d\xi}{d\tau_1} = 0 \quad \text{at } \tau_1 = T_1^*, \tag{4.3}$$

where $x_0^* = x_0(T_1^*)$, T_1^* is determined by (2.28).

We suppose that the contribution of nonlinear force in oscillations near O is relatively small. Under this assumption, one can consider the system linearized near O

$$\frac{d\xi}{d\tau_1} + F\beta = -\gamma\xi,$$

$$\frac{d\beta}{d\tau_1} - \frac{k_1}{a_0}\xi = -\gamma\beta, \tag{4.4}$$

where $k_1 = 9\alpha a_0^2 - s$. If $k_1 > 0$, the solution of system (4.4) takes the form

$$\xi(z) = c_0 e^{-\gamma(\tau_1 - T_1^*)} \cos \kappa(\tau_1 - T_1^*), \qquad \beta(z) = rc_0 e^{-\gamma(\tau_1 - T_1^*)} \sin \kappa(\tau_1 - T_1^*), \quad \tau_1 - T_1^* > 0, \tag{4.5}$$

where we denote $c_0 = x_0^* - a_0$, $\kappa^2 = Fk_1/a_0 > 0$, $r = \kappa/F$. In particular, taking the parameters (3.21) we find $x_0^* = 1.46$, $a_0 = 1.065$, $\Delta_0 = 0.1$, $k_1 = 3.2$, and, therefore, $c_0 = 0.395$, $\kappa = \sqrt{3}$.

Figure 10 demonstrates a good agreement between a numerical solution of (2.10) with parameters (3.21) (solid line) and an approximate solution found by matching the segment (3.12) (dot-and-dash) with the solution (4.5) of the linearized systems (dash) at the point T_1^*. Despite a certain discrepancy in the initial interval of motion, the numerical and analytic solutions approach closely as τ_1 increases. This implies that a simplified model (3.12), (3.16) matched with solution (4.5) suffices to describe a complicated near-resonance dynamics.

Arguing as above, one can obtain the solution in case $k_1 < 0$. Denoting $k_2 = F|k_1|/a_0$ and assuming $\gamma \ll k$, we find a solution similar to (4.5) with $\cosh(k(\tau_1 - T_1^*))$ and $\sinh(k(\tau_1 - T_1^*))$ in place of $\cos(\kappa(\tau_1 - T_1^*))$ and $\sin(\kappa(\tau_1 - T_1^*))$, respectively.

We now correlate numerical and analytic results. As seen in Figure 11, the first maximum of the slowly varying envelope of the process $u(t,\varepsilon)$ equals $M_1 \approx 1.5$; it is reached at the instant $t^* \approx 20$, or $T_1^* \approx 2$; the second maximum $M_2 \approx 1.4$ is at $t^* \approx 60$, $T_1^* \approx 6$, the third maximum $M_3 \approx 1.3$ is at $t^* \approx 100$, $T_1^* \approx 10$, and so forth. When these results are compared with that of Figure 10, it is apparent that the numerically constructed envelope is in a good agreement with the asymptotic approximations of the function $a(\tau_1)$.
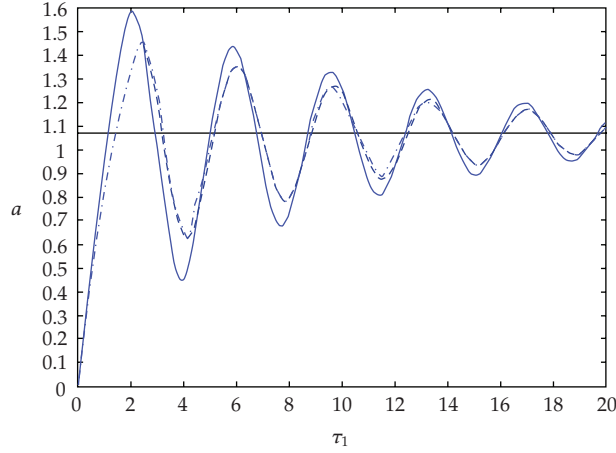
Figure 10: Transient dynamics of system (2.10): numerics (solid); segment (3.12) (dot-and-dash); solution (4.5) (dash).
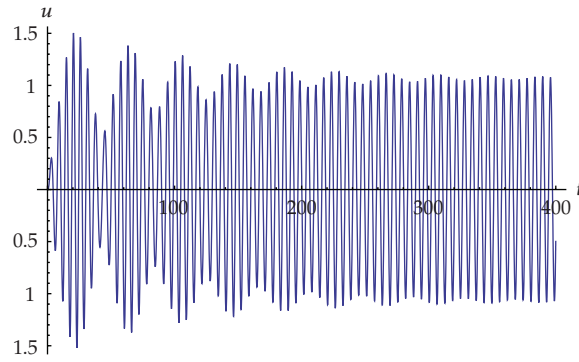

Figure 11: Numerical integration of (2.1): $\varepsilon = 0.1$, $s = 0.2$, $\alpha = 0.333$, $F = 1$, $\gamma = 0.05$.

5. Dynamics of a 2DOF System

5.1. Reduction of a 2DOF System to a Single Oscillator

In this section we present a reduction of the dynamical equations of a 2DOF system to an equation of a single oscillator. The system consists of a linear oscillator of mass M (the source of energy) coupled with a mass m (an energy sink) by a nonlinear spring with a weak linear component. For brevity, we consider the nonlinear spring with cubic nonlinearity. In this case, the equations of motion and the initial conditions have the following form:

$$
M\frac{d^2x_1}{dt^2} + \tilde{\gamma}\frac{dx_1}{dt} + k_1 x_1 + k_3(x_1 - x_2)^3 + D(x_1 - x_2) + \tilde{\eta}\left(\frac{dx_1}{dt} - \frac{dx_2}{dt}\right) = 0,
$$

$$
m\frac{d^2x_2}{dt^2} - k_3(x_1 - x_2)^3 - D(x_1 - x_2) - \tilde{\eta}\left(\frac{dx_1}{dt} - \frac{dx_2}{dt}\right) = 0,
$$

(5.1)

with the initial conditions

$$t = 0: \quad x_1 = x_2 = 0: \quad \frac{dx_1}{dt} = V_0 > 0, \quad \frac{dx_2}{dt} = 0. \tag{5.2}$$

Here x_1 and x_2 are the displacements of the masses M and m, respectively; $k_1 > 0$ is the stiffness of linear spring; $k_3 > 0$ is the coefficient of nonlinear coupling between the linear oscillator and the sink; D is the coefficient of linear coupling ($D < 0$ corresponds to a system with multiple states of equilibrium); the coefficients $\tilde{\gamma}$ and $\tilde{\eta}$ characterize dissipation in the linear oscillator and the coupling, respectively. For simplicity, we let $\tilde{\gamma} = 0$. Note that energy transfer cannot be activated in a nonexited system. In the absence of external forcing, it requires nonzero initial conditions for at least a single unit.

In what follows we assume that $m/M = \varepsilon^2 \ll 1$. Then, we introduce the dimensionless time variable $\tau_0 = \omega_0 t$, where $\omega_0 = \sqrt{k_1/M}$. In these notations, system (5.1) becomes

$$\frac{d^2 x_1}{d\tau_0^2} + x_1 + \varepsilon^2 c(x_1 - x_2)^3 + \varepsilon^3 d(x_1 - x_2) + \varepsilon^3 \eta \left(\frac{dx_1}{d\tau_0} - \frac{dx_2}{d\tau_0} \right) = 0,$$

$$\frac{d^2 x_2}{d\tau_0^2} - c(x_1 - x_2)^3 - \varepsilon d(x_1 - x_2) - \varepsilon \eta \left(\frac{dx_1}{d\tau_0} - \frac{dx_2}{d\tau_0} \right) = 0, \tag{5.3}$$

$$\tau_0 = 0: \quad x_1 = x_2 = 0; \quad v_1 = \varepsilon v_0, \quad v_2 = 0,$$

where we denote

$$\varepsilon^2 c = \frac{k_3}{k_1}, \qquad \varepsilon^3 d = \frac{D}{k_1}, \qquad \varepsilon^3 \eta = \frac{\tilde{\eta}}{\sqrt{k_1 M}}, \qquad \varepsilon v_0 = \frac{V_0}{\omega_0}, \qquad vi = \frac{dx_i}{d\tau_0}. \tag{5.4}$$

In addition, we consider the relative displacement $u = x_2 - x_1$, $du/d\tau_0 = v$ satisfying the equation

$$\frac{d^2 u}{d\tau_0^2} + \frac{d^2 x_1}{d\tau_0^2} + \varepsilon \eta \frac{du}{d\tau_0} + cu^3 + \varepsilon du = 0, \tag{5.5}$$

with the initial conditions $\tau_0 = 0$: $u = 0$, $v = -\varepsilon v_0$. Using (5.3), (5.5), the variable x_1 can be excluded. We recall that oscillations in the damped system vanish at rest O_1: ($x_1 = x_2 = 0$, $v_1 = v_2 = 0$) as $\tau_0 \to \infty$. This implies that the effect of dissipation, whatever small it might be, must be considered in the approximate solution; otherwise, the convergence to O_1 is ignored. Hence the solution of the first equation (5.3) should be written as

$$x_1(\tau_0) = \varepsilon v_0 h_\varepsilon(\tau_0) \sin \tau_0 - \varepsilon^2 c \int_0^{\tau_0} h_\varepsilon(\tau_0 - s) \sin(\tau_0 - s) u^3(s) ds + \varepsilon^3 \ldots,$$

$$h_\varepsilon(\tau_0) = e^{-(\varepsilon^3 \eta/2)\tau_0}, \tag{5.6}$$

and, by (5.6),

$$\frac{d^2x_1}{d\tau_0^2} = -\,\varepsilon v_0 h_\varepsilon(\tau_0)\sin\tau_0 - \varepsilon^2 c u^3(\tau_0) + \varepsilon^2 c \int_0^{\tau_0} h_\varepsilon(\tau_0 - s)\sin(\tau_0 - s)u^3(s)ds + \varepsilon^3 \cdots. \qquad (5.7)$$

Next, we insert (5.7) into (5.5) and ignore small terms insubstantial for the asymptotic analysis. As a result, we obtain the following equation:

$$\frac{d^2u}{d\tau_0^2} + \varepsilon\eta\frac{du}{d\tau_0} + cu^3 + \varepsilon du = \ \varepsilon v_0 h_\varepsilon(\tau_0)\sin\tau_0 - \varepsilon c_\varepsilon(\tau_0),$$

$$(5.8)$$

$$c_\varepsilon(\tau_0) = \varepsilon c I_\varepsilon, \quad I_\varepsilon = \int_0^{\tau_0} h_\varepsilon(\tau_0 - s)\sin(\tau_0 - s)u^3(s)ds,$$

with the initial conditions $u(0) = 0$, $v(0) = -\varepsilon\,v_0$. We will show that, under conditions of 1:1 resonance, $I_\varepsilon(\tau_0) \sim \varepsilon^{-1}$, $c_\varepsilon(\tau_0) \sim 1$; this means that the integral term should be taken into consideration in the asymptotic analysis.

Formula (5.8) represents a nonhomogeneous integro-differential equation with respect to u, that is above transformations reduce the original 2DOF system to a single oscillator of a more complicated structure. The initial condition $v(0) = -\varepsilon\,v_0$ implies that initially the system is close to rest, and the trajectory of system approaches the LPT of (5.8). Thus the task is to construct the LPT for the integro-differential equation (5.8).

5.2. Equations of the Resonance Dynamics

To study the system subject to 1:1 resonance, we rewrite (5.8) in the form

$$\frac{du}{d\tau_0} - v = \ 0,$$

$$\frac{dv}{d\tau_0} + (1 + 2\varepsilon\sigma)u + \varepsilon\mu\Big(cu^3 - u\Big) + \varepsilon\big(\eta v - v_0 h_\varepsilon(\tau_0)\sin\tau_0\big) + \varepsilon c_\varepsilon\ (\tau_0) = 0, \qquad (5.9)$$

$$u(0) = 0, \quad v(0) = -\,\varepsilon v_0.$$

In (5.9), we denote $\mu = 1/\varepsilon$, $\sigma = d/2$. The resonance conditions imply that the parenthetical expression with factor $\varepsilon\mu$ is relatively small compared to all other terms of order 1.

As in Section 2, we use complex-valued transformations (2.5) and the multiple time-scale method. Inserting (2.5) into (5.9), we then have

$$\frac{d\psi}{d\tau_0} - i(1 + \varepsilon\sigma)\psi + \varepsilon\mu i\left[\frac{c}{8}(\psi - \psi^*)^3 + \frac{1}{2}\psi - \psi^*\right]$$

$$+ \varepsilon\left(\frac{\eta}{2}(\psi + \psi^*) - i\sigma\psi^* - v_0 h_\varepsilon(\tau_0)\sin\tau_0\right) - \varepsilon C_\varepsilon(\tau_0) = 0, \qquad (5.10)$$

$$C_\varepsilon(\tau_0) = \varepsilon i\frac{c}{8}\int_0^{\tau_0} h_\varepsilon(\tau_0 - s)\sin(\tau_0 - s)(\psi - \psi^*)^3 ds$$

Then we construct an approximate solution of (5.10) in terms of expansions (2.5) with the slow and fast tine scales $\tau_0 = t$, $\tau_1 = \varepsilon t$, respectively.

For the resonance effect to be considered in a proper way, the leading-order equation and its solution should be

$$\frac{\partial \psi_0}{\partial \tau_0} - i(1 + \varepsilon\sigma)\psi_0 = 0,$$

$$\psi_0(\tau_0, \tau_1) = \varphi_0(\tau_1)e^{i(1+\varepsilon\sigma)\tau_0}.$$

(5.11)

The function $\varphi_0(\tau_1)$ can be found from the equation

$$\frac{\partial \psi_1}{\partial \tau_0} + \frac{d\varphi_0}{d\tau_1}e^{i(1+\varepsilon\sigma)\tau_0} - i(1 + \varepsilon\sigma)\psi_1 + i\mu\left[\frac{c}{8}(\psi_0 - \psi_0^*)^3 + \frac{1}{2}\psi_0 - \psi_0^*\right]$$

$$+ \eta\frac{\psi + \psi^*}{2} - i\sigma\varphi^* - v_0 h_\varepsilon(\tau_0)\sin\tau_0 + C_{0\varepsilon}(\tau_0) = 0,$$

(5.12)

where $C_{0\varepsilon}(\tau_0) = \varepsilon i(c/8)\int_0^{\tau_0} h_\varepsilon(\tau_0 - s)\sin(\tau_0 - s)(\varphi_0 - \varphi_0^*)^3 ds$.

To avoid secularity, we separate the resonance terms including $e^{i(1+\varepsilon\sigma)\tau_0}$ and then equate the sum to zero. First, we evaluate $C_{0\varepsilon}(\tau_0)$. To do so, we present the cubic term as $(\psi_0 - \psi_0^*)^3 = -3|\varphi_0|^2\varphi_0 e^{i(1+\varepsilon\sigma)\tau_0} + nonresonance\ terms$, and then write $C_{0\varepsilon}(\tau_0)$ in the form

$$C_{0\varepsilon}(\tau_0) = -\varepsilon\frac{3c}{16}|\varphi_0(\tau_1)|^2\varphi_0(\tau_1)\int_0^{\tau_0} e^{-(\varepsilon^3\eta/2)(\tau_0-s)}\left(e^{i(\tau_0-s)} - e^{-i(\tau_0-s)}\right)e^{i(1+\varepsilon\sigma)s} ds + \cdots$$

$$= -\varepsilon\frac{3c}{16}|\varphi_0(\tau_1)|^2\varphi_0(\tau_1)[S_{1\varepsilon}(\tau_0) - S_{2\varepsilon}(\tau_0)] + \cdots,$$

(5.13)

where the nonresonance terms are omitted. Here we denote

$$S_{1\varepsilon}(\tau_0) = \int_0^{\tau_0} e^{-(\varepsilon^3\eta/2)(\tau_0-s)}e^{i(\tau_0-s)}e^{i(1+\varepsilon\sigma)s} ds = \frac{e^{i(1+\varepsilon\sigma)\tau_0}}{\varepsilon(i\sigma + \varepsilon^2\eta)}\left(1 - e^{-[i\sigma+(\varepsilon^3\eta/2)]\tau_0}\right),$$

$$S_{2\varepsilon}(\tau_0) = \int_0^{\tau_0} e^{-(\varepsilon^3\eta/2)(\tau_0-s)}e^{-i(\tau_0-s)}e^{i(1+\varepsilon\sigma)s} ds = e^{[i-(\varepsilon^3\eta/2)]\tau_0}\int_0^{\tau_0} e^{\varepsilon[i(2+\sigma)+\varepsilon^2\eta/2]s} ds \sim O(1).$$

(5.14)

Ignoring S_2 compared with S_1, we calculate

$$C_{0\varepsilon}(\tau_0) = -i\left(\frac{3c}{16\sigma}\right)|\varphi_0(\tau_1)|^2\varphi_0(\tau_1)\left(1 - h_{1\varepsilon}(\tau_1)e^{-i\sigma\tau_1}\right)e^{i(1+\varepsilon\sigma)\tau_0} + nonresonance\ terms,$$

(5.15)

$$h_{1\varepsilon}(\tau_1) = e^{-(\varepsilon^2\eta/2)\tau_1}.$$

(5.16)

If we sum the resonance constituents in all other terms of (5.12) and then equate the total sum to zero, we obtain the equation

$$
\frac{\partial \varphi_0}{\partial \tau_1} + i\mu\left(-\frac{3c}{8}|\varphi_0|^2\varphi_0 + \frac{1}{2}\varphi_0\right) + \frac{\eta}{2}\varphi_0 + \frac{i}{2}v_0 h_{1\varepsilon}(\tau_1)e^{-i\sigma\tau_1}
$$
$$
- i\frac{3c}{16\sigma}|\varphi_0(\tau_1)|^2\varphi_0(\tau_1)\left(1 - h_{1\varepsilon}(\tau_1)e^{-i\sigma\tau_1}\right) = 0.
\tag{5.17}
$$

Then we insert the polar representation $\varphi_0 = ae^{i\delta}$ into (5.17) and set separately the real and imaginary parts of the resulting equations equal to zero. In these transformations, the last term in (5.17) can be omitted if $2\mu\sigma = \mu d > 1$. Under this assumption, we obtain

$$
\frac{da}{d\tau_1} = -\gamma a - h_{1\varepsilon}(\tau_1)F\sin\Delta,
$$
$$
a\frac{d\Delta}{d\tau_1} = \mu a\left(-s + \alpha a^2\right) - h_{1\varepsilon}(\tau_1)F\cos\Delta,
\tag{5.18}
$$

where $\Delta = \delta + \sigma\tau_1$, and $\sigma = d/2$, $\gamma = \eta/2$, $\alpha = 3c/8$, $s = (1/2)(1 - \varepsilon d)$, $F = v_0/2$.

By analogy with (2.16), we accept the initial conditions

$$
a(0) = 0, \qquad \Delta(0) = -\frac{\pi}{2}.
\tag{5.19}
$$

In the absence of dissipation ($\eta = 0$), system (5.18) takes the form

$$
\frac{da}{d\tau_1} = -F\sin\Delta,
$$
$$
a\frac{d\Delta}{d\tau_1} = \mu a\left(-s + \alpha a^2\right) - F\cos\Delta,
\tag{5.20}
$$

which is very similar to (2.10). Critical parameters, stable centers, and the LPT for (5.20) are derived in the same way as in Section 3. However, as mentioned above, steady state positions of dissipated systems (2.10) and system (5.18) are different. This reflects the fact that, while (2.10) is subjected to persistent harmonic excitation, the effect of an initial impulse exponentially decreases with time. Therefore, the first segment of the trajectory (5.18) can be approximated by a corresponding solution of (5.20) but the second segments convergences to zero as $\tau_1 = \infty$.

5.3. Dynamical Analysis of the Oscillator

5.3.1. Beating in a Nondissipative System

We now compare analytic and numerical results. We recall that numerical and experimental studies [3, 4] have shown the effective energy exchange for sets of parameters. Following [3, 4], we choose $\varepsilon = 0.316$ ($\varepsilon^2 = 0.1$), $c = 0.8$, $d = 0.632$ ($\varepsilon d = 0.2$), $v_0 = 2.215$ ($\varepsilon v_0 = 0.7$). Note that in this case $\mu d = 2$, and integral terms in (5.17) can be ignored.
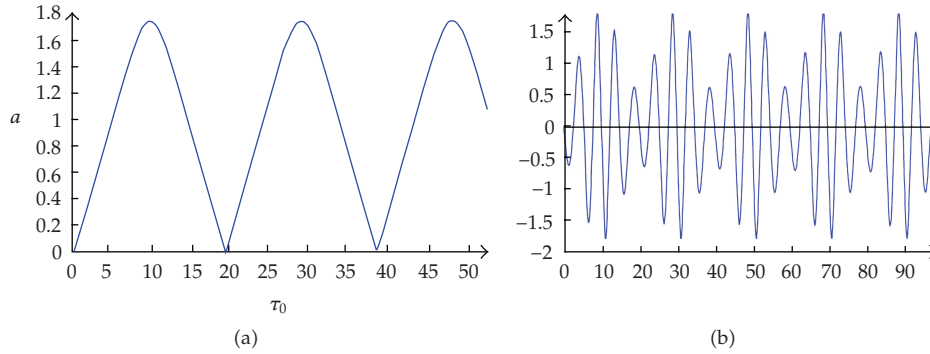
Figure 12: LPT of system (5.20) (a) and beating solution $u(\tau_0)$ of (5.8).

- —— x_1
- - - - x_2
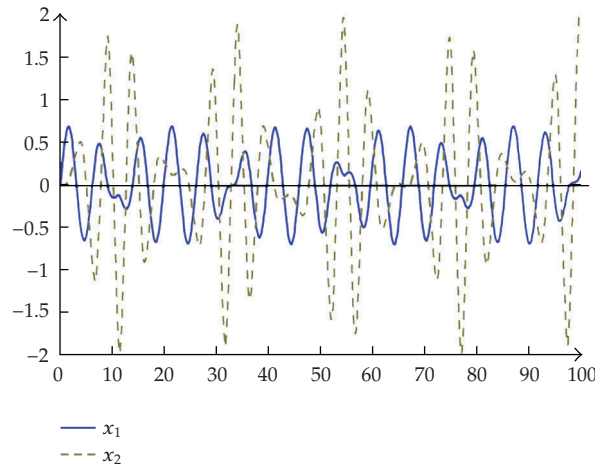
Figure 13: Plots of $x_1(\tau_0)$ (solid) and $x_2(\tau_0)$ (dash).

The LPT (Figure 12(a)) is calculated as a solution of (5.20) with initial conditions $a(0) = 0$, $\Delta(0) = -\pi/2$. One can see that the LPT approximates the envelope of the process $u_2(\tau_0)$ (Figure 13) with an error about 10%. Thus we may conclude that the observed intense energy exchange is associated with motion over the LPT.

We now calculate the critical parameter α_1^*. Here we have $\mu^2 = 10$, $F = 1.1075$, $\varepsilon d = 0.2$, $s = 0.4$, and, by (2.20)

$$\alpha_1^* = \frac{3}{8}, \qquad c_1^* = \frac{4\mu^2 s^3}{27F^2}, \qquad c_1^* = \frac{32\mu^2 s^3}{81F^2} = 0.205. \tag{5.21}$$

Since the accepted value c $0.8 > c_1^* =$, the resonance exchange takes place in the domain of large oscillations demonstrated in Figure 1.

Figure 14: Plots of $x_1(t)$ (solid) and $x_2(t)$ (dash) in the damped system.

5.3.2. The Transient Dynamics
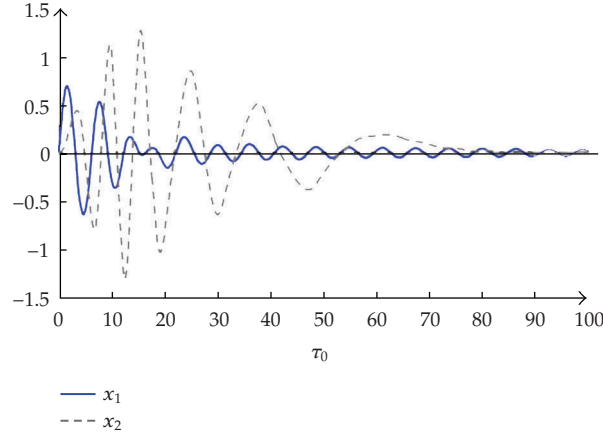
Here we consider the resonance dynamics of a weakly dissipated system. We recall that, if $\tau_1 \rightarrow \infty$, the transient process in system (5.18) does not turn into a stationary process; it vanishes at rest O: $\{a = 0\}$. To describe transient oscillations, we use the same arguments as in Section 3. The trajectory is separated into two parts; motion is assumed to be close to strongly nonlinear undamped oscillations over the interval $[0, \tau_1^*]$; then the orbit approaches the point O with an exponentially decreasing amplitude of oscillations.

We note that the contribution of nonlinear force in oscillations near O is negligible. In particular, this implies that the resonance construction of Section 5 is incorrect if τ_1 is large enough. In order to present the linearised dynamics, one can consider system (5.3) linearized near the point O_1: $(x_1 = x_2 = 0, v_1 = v_2 = 0)$. Let ξ_1 and ξ_2 denote small deviations of x_1, x_2 from O_1. A corresponding linearized system takes the form

$$\frac{d^2\xi_1}{d\tau_0^2} + \xi_1 + \varepsilon^3 \left[d(\xi_1 - \xi_2) + \eta \left(\frac{d\xi_1}{dt} - \frac{d\xi_2}{dt} \right) \right] = 0,$$

$$\frac{d^2\xi_2}{d\tau_0^2} - \varepsilon d(\xi_1 - \xi_2) - \varepsilon \eta \left(\frac{d\xi_1}{dt} - \frac{d\xi_2}{dt} \right) = 0. \tag{5.22}$$

In addition, we impose the matching constraints at $\tau_0 = \tau_0^* = \tau_1^*/\varepsilon$, namely,

$$\xi_i = x_i, \qquad \frac{d\xi_i}{d\tau_0} = \frac{dx_i}{d\tau_0}, \tag{5.23}$$

where x_i denotes the corresponding solution over $[0, \tau_0^*]$.

Formally, one can reproduce above transformations and reduce system (5.22) to a single integro-differential equation. However, a system of low dimensionality enables a straightforward investigation.

By definition, $x_1(\tau_0^*) \approx 0$ but $|x_2(\tau_0^*)|$ is close to maximum, that is, $dx_2/d\tau_0 = d\xi_2/d\tau_0 \approx 0$ at τ_0^*. The solution can be presented in the form

$$\xi_1(\tau_0) = \varepsilon v_0 e^{-\varepsilon^3 \eta(\tau_0 - \tau_0^*)/2} \sin(\tau_0 - \tau_0^*),$$

$$\xi_2(\tau_0) = C_0 e^{-\varepsilon\eta(\tau_0 - \tau_0^*)/2} \cos\left[\sqrt{\varepsilon d}(\tau_0 - \tau_0^*)\right] + \varepsilon \ldots, \tag{5.24}$$

where $\xi_1(\tau_0)$ is defined by (5.6) with the negligible integral term (see discussion after (5.17)); $\xi_2(\tau_0)$ is obtained as a solution of the second equation in (5.22) provided $\xi_1(\tau_0) \sim O(\varepsilon)$. It is seen from (5.24) that the amplitude of $\xi_1(\tau_0)$ is much lesser than the amplitude of $\xi_2(\tau_0)$ but the period and the rate of decay for $\xi_2(\tau_0)$ vastly exceed similar parameters for $\xi_1(\tau_0)$). Omitting detailed calculation, we note that the results of numerical simulation (Figure 14) agree with the conclusions following from (5.23), (5.24). The parameters of simulations are $\varepsilon^2 = 0.1$, $c = 0.8$, $\varepsilon d = 0.2$, $\varepsilon\eta = 0.2$, $\varepsilon v_0 = 0.7$.

6. Conclusion

In this paper, we have extended the concept of the limiting phase trajectories (LPTs) to dissipative oscillatory systems. Using this concept, we have constructed an approximate solution describing the maximum energy exchange between coupled oscillators. The solution consists of two parts: on an initial interval, the trajectory is close to the LPT of the undamped system; then motion becomes similar to quasilinear oscillations, approaching an asymptotically stable state. We have demonstrated a good agreement between numerical and approximate analytical solutions for two typical examples, namely, the Duffing oscillator with harmonic excitation and a system of two coupled oscillators excited by an initial impulse. In addition, we have developed a procedure for reducing a 2DOF system to a single oscillator. This allows us to obtain an approximate analytic solution describing the energy exchange and beating with complete energy transfer in a 2DOF system.

Acknowledgments

References

[1] A. F. Vakakis, O. Gendelman, L. A. Bergman, D. M. McFarland, G. McFarland, and Y. S. Lee, *Passive Nonlinear Targeted Energy Transfer in Mechanical and Structural Systems: I and II*, vol. 156 of *Solid Mechanics and Its Applications*, Springer, New York, NY, USA, 2008.

[2] A. F. Vakakis, L. I. Manevitch, Yu. V. Mikhlin, V. N. Pilipchuk, and A. A. Zevin, *Normal Modes and Localization in Nonlinear Systems*, Wiley Series in Nonlinear Science, John Wiley & Sons, New York, NY, USA, 1996.

[3] L. I. Manevitch, "New approach to beating phenomenon in coupled nonlinear oscillatory chains," *Archive of Applied Mechanics*, vol. 77, no. 5, pp. 301–312, 2007.

[4] L. I. Manevitch, E. L. Manevitch, and C.-H. Lamarque, "Toward the design of an optimal energetic sink in a strongly inhomogeneous two-degree-of-freedom system," *Journal of Applied Mechanics*, vol. 74, pp. 1078–1085, 2007.

[5] L. I. Manevitch, "Vibro-impact models for smooth non-linear systems," in *Vibro-Impact Dynamics of Ocean Systems*, R. A. Ibrahim, V. I. Babitsky, and M. Okuma, Eds., vol. 44 of *Lecture Notes in Applied and Computational Mechanics*, pp. 191–202, Springer, Berlin, Germany, 2009.

[6] L. I. Manevitch and A. I. Musienko, "Limiting phase trajectories and energy exchange between anharmonic oscillator and external force," *Nonlinear Dynamics*, vol. 58, no. 4, pp. 633–642, 2009.

[7] L. I. Manevitch, "Complex representation of dynamics of coupled nonlinear oscillators," in *Mathematical Models of Non-Linear Excitations, Transfer Dynamics and Control in Condensed Systems and Other Media*, L. Uvarova, A. Arinstein, and A. Latyshev, Eds., pp. 269–300, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[8] L. I. Manevitch, "The description of localized normal modes in a chain of nonlinear coupled oscillators using complex variables," *Nonlinear Dynamics*, vol. 25, no. 1–3, pp. 95–109, 2001.

[9] A. I. Manevich and L. I. Manevitch, *The Mechanics of Nonlinear Systems with Internal Resonances*, Imperial College Press, London, UK, 2005.

[10] J. Kevorkian and J. D. Cole, *Multiple Scale and Singular Perturbation Methods*, vol. 114 of *Applied Mathematical Sciences*, Springer, New York, NY, USA, 1996.

[11] A. H. Nayfeh, *Perturbation Methods*, Wiley Classics Library, John Wiley & Sons, New York, NY, USA, 2000.

[12] G. A. Korn and T. M. Korn, *Mathematical Handbook for Scientists and Engineers*, Dover, New York, NY, USA, 2nd edition, 2000.

[13] R. M. Rosenberg, "The $Ateb(h)$-functions and their properties," *Quarterly of Applied Mathematics*, vol. 21, pp. 37–47, 1963.

[14] R. M. Rosenberg, "On nonlinear vibrations of systems with many degrees of freedom," *Advances in Applied Mechanics*, vol. 9, pp. 155–242, 1966.

Research Article

A Wavelet Interpolation Galerkin Method for the Simulation of MEMS Devices under the Effect of Squeeze Film Damping

Pu Li[1] and Yuming Fang[2]

[1] *School of Mechanical Engineering, Southeast University, Jiangning, Nanjing 211189, China*
[2] *College of Electronic Science and Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*

Correspondence should be addressed to Pu Li, seulp@seu.edu.cn

This paper presents a new wavelet interpolation Galerkin method for the numerical simulation of MEMS devices under the effect of squeeze film damping. Both trial and weight functions are a class of interpolating functions generated by autocorrelation of the usual compactly supported Daubechies scaling functions. To the best of our knowledge, this is the first time that wavelets have been used as basis functions for solving the PDEs of MEMS devices. As opposed to the previous wavelet-based methods that are all limited in one energy domain, the MEMS devices in the paper involve two coupled energy domains. Two typical electrically actuated micro devices with squeeze film damping effect are examined respectively to illustrate the new wavelet interpolation Galerkin method. Simulation results show that the results of the wavelet interpolation Galerkin method match the experimental data better than that of the finite difference method by about 10%.

1. Introduction

Modeling and simulation of MEMS devices play an important role in the design phase for system optimization and for the reduction of design cycles. The performances of MEMS devices are represented by partial-differential equations (PDEs) and associated boundary conditions. In the past two decades, there have been extensive, and successful, works focused on solving the partial-differential equations of MEMS [1–15]. A detailed review of the works is available in [1]. In the previous works, Galerkin method was widely used to reduce the partial-differential equations to ordinary-differential equations (ODEs) in time and then solve the reduced equations either numerically or analytically. The previous works differ from each other in the choice of the basis functions.

The basis set can be chosen arbitrarily, as long as its elements satisfy all of the boundary conditions and are sufficiently differentiable. To enhance convergence, the basis set has to be chosen to resemble the behavior of the device. For example, two ways have been used to generate the basis set for the reduced-order models of MEMS devices [1]. The first way [4, 9] uses the undamped linear model shapes of the undeflected microstructure as basis functions. For simple structures with simple boundary conditions, the mode shapes are found analytically. For complex structures or complex boundary conditions, the linear mode shapes are obtained numerically using the finite element method. The second way [2] conducts experiments or solves the PDEs using FEM or FDM to generate snapshots under a training signal, then applies a modal analysis method (one of the variation of the proper orthogonal decomposition method [6]) to the time series to extract the mode shapes of the device structural elements.

In the past two decades also, a new numerical concept was introduced and is gaining increasing popularity [16–25]. The method is based on the expansion of functions in terms of a set of basis functions called wavelets. Indeed wavelets have many excellent properties such as orthogonality, compact support, exact representation of polynomials to a certain degree, and flexibility to represent functions at different levels of resolution. Indeed a complete basis can be generated easily by a signal function through dilatation and translation. The wavelet-based methods may be classified as wavelet-Galerkin method [19, 20], wavelet-collocation method [21, 22], and wavelet interpolation Galerkin method [23–25]. Among the three methods, the wavelet-Galerkin method is the most common one because of its implementation simplicity. The method is a Galerkin scheme using scaling or wavelet functions as the trial and weight functions. However, both scaling and wavelet functions do not satisfy the boundary conditions. Thus the treatment of general boundary conditions is a major difficulty for the application of the wavelet-Galerkin method, especially for the bounded region problems, even though different efforts [19, 20] have been made. For the wavelet-collocation method, boundary conditions can be treated in a satisfactory way [21]. In the method, trial functions are a class of interpolating functions generated by autocorrelation of the usual compactly supported Daubechies scaling functions. However, the method requires the calculation of higher-order derivatives (up to the second derivatives for second-order parabolic problems) of the wavelets. Due to the derivatives of compactly supported wavelets being highly oscillatory, it is difficult to compute the connection coefficients by the numerical evaluation of integral [18]. The wavelet interpolation Galerkin method is a Galerkin scheme that both trial and weight functions are a class of interpolating functions generated by autocorrelation of the usual compactly supported Daubechies scaling functions. For the method, the boundary conditions [24] can be treated easily and the formulations are derived from the weak form; thus only the first derivatives of wavelets (for second-order parabolic problems) are required.

Wavelets have proven to be an efficient tool of analysis in many fields including the solution of PDEs. However, few papers in MEMS area give attention to the wavelet-based methods. This paper presents a new wavelet interpolation Galerkin method for the numerical simulation of MEMS devices under the effect of squeeze film damping. To the best of our knowledge, this is the first time that wavelets have been used as basis functions for solving the PDEs of MEMS devices. As opposed to the previous wavelet-based methods that are all limited in one energy domain, the MEMS devices in the paper involve two coupled energy domains. The squeeze film damping effect on the dynamics of microstructures has already been extensively studied. We stress that our intention here is not to discover new physics to the squeeze film damping.

The outline of this paper is as follows. Section 2 presents a brief introduction to some major concepts and properties of wavelets. In Sections 3 and 4, two typical electrically actuated micro devices with squeeze film damping effect are examined respectively to illustrate the wavelet interpolation Galerkin method. Section 5 calculates the frequency responses and the quality factors using the present method, and compares the calculated results with those generated by experiment [26, 27], by the finite difference method, and by other published analytical models [15, 26]. Finally, a conclusion is given in Section 6.

2. Basic Concepts of Daubechies' Wavelets and Wavelet Interpolation

In this section, we shall give a brief introduction to the concepts and properties of Daubechies' wavelets. More detailed discussions can be found in [16–18, 21].

2.1. Daubechies' Orthonormal Wavelets

Daubechies [16, 17] constructed a family of orthonomal bases of compactly supported wavelets for the space of square-integrable funcntions, $L^2(R)$. Due to the fact that they possess several useful properties, such as orthogonality, compact support, exact representation of polynomials to a certain degree, and ability to represent functions at different levels of resolution, Daubechies' wavelets have gained great interest in the numerical solutions of PDEs [18–22].

Daubechies' functions are easy to construct [16, 17]. For an even integer L, we have the Daubechies' scaling function $\phi(x)$ and wavelet $\psi(x)$ satisfying

$$
\phi(x) = \sum_{i=0}^{L-1} \widehat{p}_i \phi(2x - i)
$$

$$
\psi(x) = \sum_{i=2-L}^{1} (-1)^i \widehat{p}_{1-i} \phi(2x - i).
$$

(2.1)

The fundamental support of the scaling function $\phi(x)$ is in the interval $[0, L-1]$ while that of the corresponding wavelet $\psi(x)$ is in the interval $[1 - L/2, L/2]$. The parameter L will be referred to as the degree of the scaling function $\phi(x)$. The coefficients $\widehat{p}_i$ are called the wavelet filter coefficients. Daubenchies [16, 17] established these wavelet filter coefficients to satisfy the following conditions:

$$
\sum_{i=0}^{L-1} \widehat{p}_i = 2,
$$

$$
\sum_{i=0}^{L-1} \widehat{p}_i \widehat{p}_{i-m} = \delta_{0,m},
$$

$$
\sum_{i=2-L}^{1} (-1)^i \widehat{p}_{1-i} \widehat{p}_{i-2m} = 0 \quad \text{for integer,}
$$

$$
\sum_{i=0}^{L-1} (-1)^i i^m \widehat{p}_i = 0, \quad m = 0, 1, \ldots, \frac{L}{2} - 1,
$$

(2.2)

where $\delta_{0,m}$ is the Kronecker delta function. Correspondingly, the constructed scanling function $\phi(x)$ and wavelet $\psi(x)$ have the following properties:

$$\int_{-\infty}^{\infty} \phi(x)\mathrm{d}x = 1,$$

$$\int_{-\infty}^{\infty} \phi(x-i)\phi(x-m)\mathrm{d}x = \delta_{i,m} \quad \text{for integers } i,m,$$

$$\int_{-\infty}^{\infty} \phi(x)\psi(x-m)\mathrm{d}x = 0 \quad \text{for integer } m,$$

$$\int_{-\infty}^{\infty} x^m \psi(x)\mathrm{d}x = 0, \quad m = 0,1,\ldots,\frac{L}{2}-1. \tag{2.3}$$

Denote by $L^2(R)$ the space of square-integrable functions on the real line. Let V_J and W_J be the subspace generated, respectively, as the L^2-closure of the linear spans of $\phi_{J,i}(x) = 2^{J/2}\phi(2^J x - i)$ and $\psi_{J,i}(x) = 2^{J/2}\psi(2^J x - i)$, $J,i \in Z$. $\mathbb{Z}$ denotes the set of integers. Then (2.3) implies that

$$V_{J+1} = V_J \oplus W_J, \quad V_0 \subset V_1 \subset \cdots V_J \subset V_{J+1}, \quad V_{J+1} = V_0 \oplus W_0 \oplus W_1 \oplus \cdots W_J, \tag{2.4}$$

Equation (2.4) presents the multiresolution properties of wavlets. Any function $f \in L^2(R)$, may be approximated by the multiresolution apparatus described above, by its projection $P_J^V f$ onto the subspace V_J

$$P_J^V f = \sum_{i \in Z} f_{J,i}\phi_{J,i}(x). \tag{2.5}$$

2.2. Wavelet Interpolation Scaling Function

For a given Daubechies' scaling function, its autocorrelation function $\theta(x)$ can be defined as follows [21]:

$$\theta(x) = \int_{-\infty}^{\infty} \phi(\tau)\phi(\tau - x)\mathrm{d}\tau. \tag{2.6}$$

The function satisfies the following interpolating property

$$\theta(k) = \delta_{0,k}, \quad k \in Z, \tag{2.7}$$

and has a symmetric support $[-(L-1),(L-1)]$. The derivative of the function $\theta(k)$ may be computed by differentiating the convolution product

$$\theta^{(s)}(k) = (-1)^{(s)}\int_{-\infty}^{\infty} \phi(\tau)\phi^{(s)}(\tau - k)\mathrm{d}\tau. \tag{2.8}$$

Let $\theta(x)$ act as the scaling function, we have

$$\theta_{J,k}(x) = \theta\left(2^J x - k\right), \quad k \in Z. \tag{2.9}$$

For a set of dyadic grids of the type $x_k^J \in R$: $x_k^J = 2^{-J} k$, where $k, J \in Z$, the $\theta_{J,k}(x)$ verifies the interpolation property at the dyadic points: $\theta_{J,k}(x_n^J) = \delta_{n,k}$. Let V_J^x be the linear span of the set $\{\theta(2^J x - k), k \in Z\}$. It can be proved that $\{V_J^x\}$ forms a multiresolution analysis, where $\theta_{J,k}(x)$ acts as the role of scaling function (the so-called interpolation scaling function), and the set $\{\theta(2^J x - k), k \in Z\}$ is a Riesz's basis for V_J^x. For a function $f \in H^1(R)$, an interpolation operator $I_J : H^1(R) \rightarrow V_J^x$ can be defined [21]:

$$I_J(f) = \sum_k f_k^J \theta\left(2^J x - k\right), \quad k \in Z, \tag{2.10}$$

where $f_k^J = f(x_k^J) = f(2^{-J} k)$. Thus, for a function $f(x)$ defined on $x \in [0, 1]$, $f(x)$ has the following approximation

$$
\begin{aligned}
f(x) &= \sum_{k=-(L-1)}^{2^J+(L-1)} f_{J,k} \theta\left(2^J x - k\right) \\
&= \sum_{k=-(L-1)}^{1} f_{J,k} \theta\left(2^J x - k\right) + \sum_{k=0}^{2^J} f_{J,k} \theta\left(2^J x - k\right) + \sum_{k=2^J+1}^{2^J+L-1} f_{J,k} \theta\left(2^J x - k\right).
\end{aligned}
\tag{2.11}
$$

In this paper, wavelet collocation scheme is applied on $x \in [0, 1]$, where $x_k^J = 2^{-J} k$ and $k = 0, 1, \ldots, 2^J$. Therefore, instead of the values of $f(x)$ at x_k^J, $k = -(L-1), \ldots, -1$ and $k = 2^J + 1, \ldots, (L-1)$, we may use some values which are extrapolated from the values in those dyadic points internal to the interval $x \in [0, 1]$. As described in [21, 22], we define

$$f(x) = \sum_{k=0}^{2^J} f_{J,k} \widehat{\theta}\left(2^J x - k\right), \tag{2.12}$$

where

$$\widehat{\theta}\left(2^J x - k\right) = \begin{cases} \theta(2^J x - k) + \displaystyle\sum_{n=-(L-1)}^{-1} a_{nk}\theta(2^J x - n), & k = 0, 1, \ldots, 2M-1 \\[2ex] \theta(2^J x - k), & k = 2M, \ldots, 2^J - 2M \\[2ex] \theta(2^J x - k) + \displaystyle\sum_{n=2^J+1}^{2^J+L-1} b_{nk}\theta(2^J x - n), & k = 2^J - 2M + 1, \ldots, 2^J, \end{cases} \tag{2.13}$$

where the coefficients a_{nk} and b_{nk} are defined by

$$a_{nk} = l_k^1\left(x_n^J\right), \qquad b_{nk} = l_k^2\left(x_n^J\right), \tag{2.14}$$

where $l_k^1(x)$ and $l_k^2(x)$ represent Lagrange interpolation polynomials, defined by

$$l_k^1(x) = \prod_{\substack{i=0 \\ i \neq k}}^{2M-1} \frac{x - x_i^J}{x_k^J - x_i^J}, \qquad l_k^2(x) = \prod_{\substack{i=2^J-2M+1 \\ i \neq k}}^{2^J} \frac{x - x_i^J}{x_k^J - x_i^J}. \tag{2.15}$$

An analogous manner can be given for two-dimensional problem. By using tensor products, it is then possible to define a multiresolution on the square $x, y \in [0, 1]$. The two-dimensional scaling function is defined by $\Theta_{k,k'}^J(x, y) = \sum_{k=0}^{2^J} \sum_{k'=0}^{2^J} f_{k,k'}^J \widehat{\theta}(2^J x - k)\widehat{\theta}(2^J y - k')$. Let $V_J^{xy} = V_J^x \otimes V_J^y$ be the linear span of the set $\{\theta(2^J x - k)\theta(2^J y - k'), J, k, k' \in Z\}$; thus the set $\{V_J^{xy}\}$ forms a multiresolution analysis and the set $\{\theta(2^J x - k)\theta(2^J y - k'), k, k' \in Z\}$ is a Riesz basis for $\{V_J^{xy}\}$. Therefore, for a function $f(x, y)$ defined on $x, y \in [0, 1]$, it has the following approximation:

$$f(x, y) = \sum_{k=0}^{2^J} \sum_{k'=0}^{2^J} f_{k,k'}^J \widehat{\theta}\left(2^J x - k\right)\widehat{\theta}\left(2^J y - k'\right). \tag{2.16}$$

3. Wavelet Interpolation Galerkin Method for a Parallel Plate Microresonator under the Effect of Squeeze Film Damping

3.1. Governing Equations

In this section, we examine the example of a rectangular parallel plate under the effect of squeeze film damping. As shown in Figure 1, the rectangular parallel plate is excited by a conventional voltage. The voltage is composed of a dc component V_0 and a small ac component $v(t)$, $V_0 \gg v(t)$. The plate is rigid. The displacement of the plate under the electric force is composed of a static component to the dc voltage, denoted by z_0, and a small dynamic component due to the ac voltage, denoted by $z(t)$, $z_0 \gg z(t)$, that is,

$$z_E(t) = z_0 + z(t). \tag{3.1}$$

The equation of motion that governs the displacement of the plate is written as

$$m_{\text{plate}} \ddot{z}_E + k_{\text{spring}} z_E = \frac{\varepsilon A_{\text{plate}}(V_0 + v)^2}{2\left(g_0 - z_E\right)^2} - f(t), \tag{3.2}$$

where m_{plate} is the mass of the plate, A_{plate} is the are of the plate, k_{spring} is the stiffness of the spring, g_0 is the zero-voltage air gap spacing, ε is the dielectric constant of the gap medium,

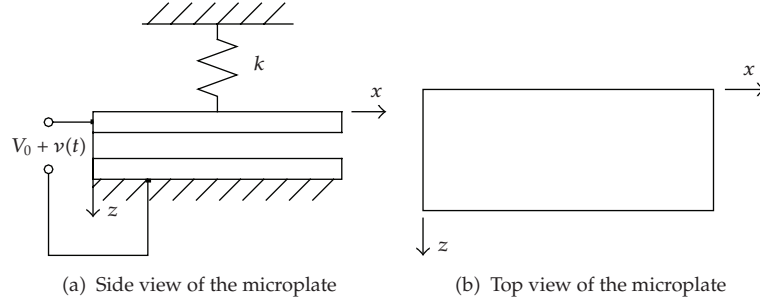(a) Side view of the microplate   (b) Top view of the microplate

Figure 1: A schematic drawing of an electrically actuated microplate under the effect of squeeze film damping.

$f(t)$ is the force acting on the plate owing to the pressure of the squeeze gas film between the plate and the substrate.

We expand (3.2) in a Taylor series around V_0 and z_0 up to first order and rewrite (3.2) as

$$m_{\text{plate}}\ddot{z} + k_E z = \frac{\varepsilon A_{\text{plate}} V_0}{\widehat{g}_0^2} v - f(t), \tag{3.3}$$

where $k_E = k_{\text{spring}} - (\varepsilon A_{\text{plate}} V_0^2 / (g_0 - z_0)^3)$, $\widehat{g}_0 = g_0 - z_0$. The force $f(t)$ acting on the plate owing to the pressure of the squeeze gas film is given by

$$f(t) = \int_0^{l_y} \int_0^{l_x} (p(x, y, t) - p_0) dx\, dy, \tag{3.4}$$

where l_x and l_y are the length and width of plate, $p(x, y, t)$ is the absolute pressure in the gap and p_0 is the ambient pressure. The pressure $p(x, y, t)$ is governed by the nonlinear Reynolds equation [3]

$$\frac{\partial}{\partial x}\left(h^3 p \frac{\partial p}{\partial x}\right) + \frac{\partial}{\partial y}\left(h^3 p \frac{\partial p}{\partial y}\right) = 12\eta_{\text{eff}}\left(h\frac{\partial p}{\partial t} + p\frac{\partial h}{\partial t}\right), \tag{3.5}$$

where $h(x, t) = g_0 - z_0 - z(t) = \widehat{g}_0 - z(t)$ and η_{eff} is the effective viscosity of the fluid in the gap. In this section, all edges of the rectangular plate are ideally vented; thus the pressure boundary conditions for the case in **Figure 1** are

$$p(x, 0, t) = p(x, l_y, t) = p(0, y, t) = p(l_x, y, t) = p_0. \tag{3.6}$$

For convenience, we introduce the nondimensional variables

$$X = \frac{x}{l_x}, \quad Y = \frac{y}{l_y}, \quad Z = \frac{z}{\widehat{g}_0}, \quad \widehat{P} = \frac{p}{p_0}, \quad T = \frac{t}{S}, \quad H = \frac{h}{\widehat{g}_0} = 1 - Z, \tag{3.7}$$

where T is a timescale, $S = \sqrt{m_{\text{plate}}/k_E} = 1/\omega_n$, ω_n is the nature frequency of the plate. Substituting (3.7) into (3.3)–(3.6), we obtain

$$\frac{\mathrm{d}^2 Z}{\mathrm{d}T^2} + Z = \alpha V_0 v - P_{\text{non}} \iint_0^1 \left(\widehat{P} - 1\right) \mathrm{d}X \, \mathrm{d}Y, \tag{3.8}$$

$$\frac{\partial}{\partial X}\left(H^3 \widehat{P} \frac{\partial \widehat{P}}{\partial X}\right) + \beta^2 \frac{\partial}{\partial Y}\left(H^3 \widehat{P} \frac{\partial \widehat{P}}{\partial Y}\right) = \frac{\sigma}{S}\left(H \frac{\partial \widehat{P}}{\partial T} + \widehat{P} \frac{\partial H}{\partial T}\right), \tag{3.9}$$

where $\alpha = \varepsilon A_{\text{plate}}/k_E \widehat{g}_0^3$, $P_{\text{non}} = p_0 l_x l_y/k_E \widehat{g}_0$, $\sigma = 12 \eta_{\text{eff}} l_x^2/\widehat{g}_0^2 p_0$, and $\beta = l_x/l_y$. The nondimensional boundary conditions are

$$\widehat{P}(X, 0, T) = \widehat{P}(X, 1, T) = \widehat{P}(1, Y, T) = \widehat{P}(0, Y, T) = 1. \tag{3.10}$$

As mentioned above, the microplate is under small oscillation around $\widehat{g}_0$ and therefore the pressure variation from ambient in the squeeze film is also small, $\widehat{P}(X, Y, T)$ is given by

$$\widehat{P}(X, Y, T) = \frac{p}{p_0} = 1 + P(X, Y, T), \tag{3.11}$$

where $|P(X, Y, T)| \ll 1$. Substituting (3.11) into (3.9), and linearizing the outcome around p_0 and $\widehat{g}_0$, we obtain

$$\frac{\partial^2 P}{\partial X^2} + \beta^2 \frac{\partial^2 P}{\partial Y^2} - \frac{\sigma}{S} \frac{\partial P}{\partial T} = -\frac{\sigma}{S} \frac{\partial Z}{\partial T}. \tag{3.12}$$

The boundary conditions for the case are

$$P(X, 0, T) = P(X, 1, T) = P(0, Y, T) = P(1, Y, T) = 0. \tag{3.13}$$

For a harmonic excitation, the ac component voltage v(t) is given by

$$v(T) = v_0 e^{j\omega TS}. \tag{3.14}$$

Usually, the excitation frequency ω is approximate to the natural frequency ω_n. The steady-state solution of (3.8) and (3.12) may be expressed by

$$Z(T) = A e^{j\omega TS}, \tag{3.15}$$

$$P(X, Y, T) = A \cdot P_A(X, Y) e^{j\omega TS}, \tag{3.16}$$

where A is the complex amplitude to be determined. Substituting (3.15) and (3.16) into (3.12), we obtain

$$\frac{\partial^2 P_A(X,Y)}{\partial X^2} + \beta^2 \frac{\partial^2 P_A(X,Y)}{\partial Y^2} - j\sigma\omega P_A(X,Y) = -j\sigma\omega. \tag{3.17}$$

The boundary conditions are

$$P_A(X,0) = P_A(X,1) = P_A(0,Y) = P_A(1,Y) = 0. \tag{3.18}$$

3.2. Wavelet Interpolation Method for Squeeze Film Damping Equations

3.2.1. Construction of Basis Functions

In this subsection, the approximate solution of $P_A(X,Y)$ is approximated by the following form:

$$
\begin{aligned}
P_A(X,Y) &\approx \sum_{k=0}^{2^J} \sum_{k'=0}^{2^J} p_{k,k'}^J \Theta_{k,k'}^J(X,Y) = \sum_{k=0}^{2^J} \sum_{k'=0}^{2^J} p_{k,k'}^J \widehat{\theta}_{J,k}(X)\widehat{\theta}_{J,k'}(Y) \\
&= \sum_{k=0}^{2^J} \sum_{k'=0}^{2^J} p_{k,k'}^J \widehat{\theta}\left(2^J X - k\right)\widehat{\theta}\left(2^J Y - k'\right), \quad k,k' \in Z,
\end{aligned}
\tag{3.19}
$$

where the unknowns $p_{k,k'}^J$ are the values of $P_A(X,Y)$ at the dyadic points $X = k2^{-J}$, and $Y = k'2^{-J}$. The unknowns $p_{k,k'}^J$ are complex.

For the application of Galerkin method, (3.19) should be able to satisfy the boundary conditions. Substituting (3.19) into (3.18), leads to

$$
\begin{aligned}
\sum_{k'=0}^{2^J} p_{0,k'}^J \widehat{\theta}(0)\widehat{\theta}\left(2^J Y - k'\right) = 0 &\Longrightarrow p_{0,k'}^J = 0, \quad \text{for } k' = 0,1,2,\ldots 2^J, \\
\sum_{k'=0}^{2^J} p_{2^J,k'}^J \widehat{\theta}(0)\widehat{\theta}\left(2^J Y - k'\right) = 0 &\Longrightarrow p_{2^J,k'}^J = 0, \quad \text{for } k' = 0,1,2,\ldots 2^J, \\
\sum_{k=1}^{2^J-1} p_{k,0}^J \widehat{\theta}\left(2^J X - k\right)\widehat{\theta}(0) = 0 &\Longrightarrow p_{k,0}^J = 0, \quad \text{for } k = 1,2,\ldots,\left(2^J - 1\right), \\
\sum_{k=1}^{2^J-1} p_{k,2^J}^J \widehat{\theta}\left(2^J X - k\right)\widehat{\theta}(0) = 0 &\Longrightarrow p_{k,2^J}^J = 0, \quad \text{for } k = 1,2,\ldots,(2^J - 1).
\end{aligned}
\tag{3.20}
$$

Thus (3.19) is rewritten as

$$P_A(X,Y) = \sum_{k=1}^{2^J-1} \sum_{k'=1}^{2^J-1} p_{k,k'}^J \Theta_{k,k'}^J(X,Y) = \sum_{k=1}^{2^J-1} \sum_{k'=1}^{2^J-1} p_{k,k'}^J \widehat{\theta}\left(2^J X - k\right)\widehat{\theta}\left(2^J Y - k'\right). \tag{3.21}$$

3.2.2. Discretion of the Boundary Value Problem

The weak form functional of (3.17) is

$$W(P_A) = \iint_\Omega \left\{ \frac{1}{2}\left[\left(\frac{\partial P_A}{\partial X}\right)^2 + \beta^2 \left(\frac{\partial P_A}{\partial Y}\right)^2 + j\sigma\omega P_A^2 \right] - j\sigma\omega P_A \right\} dX\, dY. \tag{3.22}$$

From the necessary conditions for the determination of the minimum W, we obtain

$$\delta W(P_A) = \iint_\Omega \left[\frac{\partial \delta P_A}{\partial X}\frac{\partial P_A}{\partial X} + \beta^2 \frac{\partial \delta P_A}{\partial Y}\frac{\partial P_A}{\partial Y} + j\sigma\omega \delta P_A P_A - j\sigma\omega \delta P_A \right] dX\, dY = 0. \tag{3.23}$$

Substituting (3.19) into (3.23), leads to

$$\sum_{k=1}^{2^J-1} \sum_{k'=1}^{2^J-1} \left\{ \iint_\Omega \left[\frac{\partial \Theta_{m,n}^J}{\partial X}\frac{\partial \Theta_{k,k'}^J}{\partial X} + \beta^2 \frac{\partial \Theta_{m,n}^J}{\partial Y}\frac{\partial \Theta_{k,k'}^J}{\partial Y} + j\sigma\omega \Theta_{m,n}^J \Theta_{k,k'}^J \right] dX dY \right\} p_{k,k'}^J$$

$$= j\sigma\omega \iint_\Omega \Theta_{m,n}^J dX\, dY, \quad \text{for } m,n = 1,2,\dots,\left(2^J-1\right). \tag{3.24}$$

This is a $\left(2^J-1\right)^2 \times \left(2^J-1\right)^2$ linear system

$$\mathbf{\Theta p} = j\sigma\omega\mathbf{E}, \tag{3.25}$$

where $\mathbf{p} = \left[p_{1,1}^J\, p_{1,2}^J\, \cdots\, p_{1,2^J-1}^J\, p_{2,1}^J\, p_{2,2}^J\, \cdots\, p_{2^J-1,2^J-1}^J \right]^T$ is an $\left(2^J-1\right)^2 \times 1$ unknown coefficients' vector, $\mathbf{E} = \left[\iint_\Omega \Theta_{1,1}^J dX d Y\, \iint_\Omega \Theta_{1,2}^J dX\, dY\, \cdots\, \iint_\Omega \Theta_{2^J-1,2^J-1}^J dX\, dY \right]^T$ is a $\left(2^J-1\right)^2 \times 1$ matrix, and $\mathbf{\Theta}$ is a $\left(2^J-1\right)^2 \times \left(2^J-1\right)^2$ matrix. The entries in $\mathbf{\Theta}$ are of the form

$$\widehat{\Theta}(i,j) = \widehat{\Theta}\left(\left[(m-1)\left(2^J-1\right)+n\right], \left[(k-1)\left(2^J-1\right)+k'\right] \right)$$

$$= \iint_\Omega \left[\frac{\partial \widehat{\theta}(2^J X - m)}{\partial X}\widehat{\theta}\left(2^J Y - n\right)\frac{\partial \widehat{\theta}(2^J X - k)}{\partial X}\widehat{\theta}\left(2^J Y - k'\right) \right.$$

$$+ \beta^2 \widehat{\theta}\left(2^J X - m\right)\frac{\partial \widehat{\theta}(2^J Y - n)}{\partial Y}\widehat{\theta}\left(2^J X - k\right)\frac{\partial \widehat{\theta}(2^J Y - k')}{\partial Y} \tag{3.26}$$

$$\left. + j\sigma\omega\widehat{\theta}\left(2^J X - m\right)\widehat{\theta}\left(2^J Y - n\right)\widehat{\theta}\left(2^J X - k\right)\widehat{\theta}\left(2^J Y - k'\right) \right] dX\, dY.$$

3.2.3. Squeeze Film Damping of the Parallel Plate

The numerical solution of (3.25) can be written as

$$\mathbf{p} = j\sigma w \Theta^{-1}\mathbf{E}. \tag{3.27}$$

The elements of $\mathbf{p}$ can be expressed as

$$p_{k,k'}^{J} = p_{k,k'}^{J,R} + jp_{k,k'}^{J,I} \quad \text{for } m,n = 1,2,\ldots,\left(2^{J}-1\right), \tag{3.28}$$

where $p_{k,k'}^{J,R}$ and $p_{k,k'}^{J,I}$ are the real and imaginary parts of $p_{k,k'}^{J}$, respectively. Using (3.21) and (3.28), the force acting on the plate owing to the pressure of the squeeze gas film can be rewritten as

$$
P_{\mathrm{non}}\iint_{0}^{1}\left(\widehat{P}-1\right)\mathrm{d}X\,\mathrm{d}Y
$$

$$
= Ae^{j\omega TS}\cdot P_{\mathrm{non}}\sum_{k=1}^{2^{J}-1}\sum_{k'=1}^{2^{J}-1}\left(p_{k,k'}^{J,R}+jp_{k,k'}^{J,I}\right)\iint_{0}^{1}\widehat{\theta}\left(2^{J}X-k\right)\widehat{\theta}\left(2^{J}Y-k'\right)\mathrm{d}X\,\mathrm{d}Y \tag{3.29}
$$

$$
= K_{a}\cdot Z(T) + C_{a}\cdot\frac{\mathrm{d}Z(T)}{\mathrm{d}T},
$$

where

$$
K_{a} = P_{\mathrm{non}}\sum_{k=1}^{2^{J}-1}\sum_{k'=1}^{2^{J}-1}p_{k,k'}^{J,R}\iint_{0}^{1}\widehat{\theta}\left(2^{J}X-k\right)\widehat{\theta}\left(2^{J}Y-k'\right)\mathrm{d}X\,\mathrm{d}Y,
$$

$$
C_{a} = \frac{P_{\mathrm{non}}}{wS}\sum_{k=1}^{2^{J}-1}\sum_{k'=1}^{2^{J}-1}p_{k,k'}^{J,I}\iint_{0}^{1}\widehat{\theta}\left(2^{J}X-k\right)\widehat{\theta}\left(2^{J}Y-k'\right)\mathrm{d}X\,\mathrm{d}Y \tag{3.30}
$$

$K_{a}\cdot Z(T)$ and $C_{a}\cdot(\mathrm{d}Z(T)/\mathrm{d}T)$ are the spring and damping components of the force. Substituting (3.29), (3.14) and (3.15) into (3.8), we obtain

$$\frac{\mathrm{d}^{2}Z}{\mathrm{d}T^{2}} + C_{a}\cdot\frac{\mathrm{d}Z}{\mathrm{d}T} + (K_{a}+1)Z(T) = \alpha V_{0}v(T),$$

$$Z(T) = Ae^{j\omega TS} = \frac{aV_{0}v_{0}}{K_{a}+1}\cdot\frac{1}{1-(\omega^{2}(S^{2}/(K_{a}+1)))+(j\omega(C_{a}S/(K_{a}+1)))}e^{j\omega TS}, \tag{3.31}$$

where $S = \sqrt{m_{\mathrm{plate}}/k_{E}} = 1/w_{n}$. The quality factor and the damped natural frequency are expressed as

$$Q_{\mathrm{squeeze}} = \frac{1}{2\xi} = \frac{\sqrt{K_{a}+1}}{C_{a}}, \qquad w_{\mathrm{squeeze}} = w_{n}\sqrt{K_{a}+1}. \tag{3.32}$$

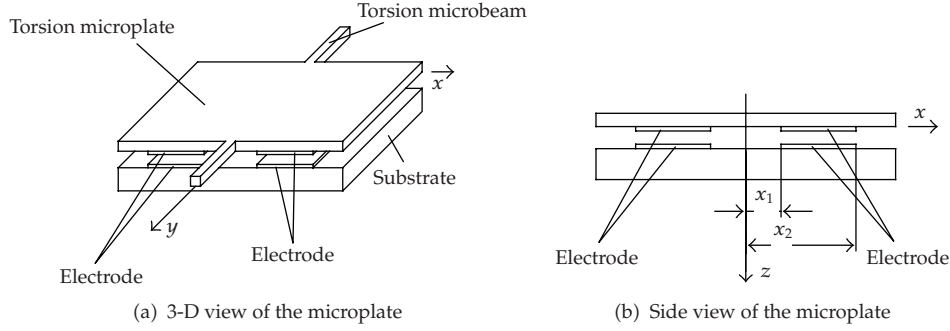(a) 3-D view of the microplate (b) Side view of the microplate

Figure 2: A schematic drawing of a torsion microplate under the effect of squeeze film damping.

4. Wavelet Interpolation Galerkin Method for a Torsion Microplate under the Effect of Squeeze Film Damping

A similar analysis as the one given for the parallel plate microresonator can be given for a torsion microplate.

4.1. Governing Equations

In this section, we examine the example of a rectangular torsion microplate under the effect of squeeze film damping. As shown in Figure 2, the microplate is suspended by two torsion microbeams. l_x, l_y and h_δ are the length, width and thickness of the plate. There are two pairs of electrodes between the microplate and the substrate. The locations of the two pairs of electrodes are symmetrical. x_1 and x_2 are the positions of the two pairs of electrodes. The thickness of the electrodes is neglected. On each pair of the electrodes, an equal dc voltage V_0 and an equal ac voltage $v(t)$ with opposite potential were applied. The rotation angle of the plate is composed of a static component to the dc voltage, denoted by γ_0, and a small dynamic component due to the ac voltage, denoted by $\gamma(t)$. In this case, $\gamma_0 = 0$; thus the equation of the plate around V_0 and γ_0 can be written as

$$J\ddot{\gamma} + k_{T-E}\gamma = -\frac{(x_2^2 - x_1^2)\varepsilon l_y V_0}{g_0^2}v + \int_{-l_x/2}^{l_x/2}\int_0^l (p(x,y,t) - p_0)x\,\mathrm{d}y\,\mathrm{d}x, \qquad (4.1)$$

where $k_{T-E} = [k_T - 2\varepsilon l_y V_0^2((x_2^3 - x_1^3)/3g_0^3)]$ and k_T is the stiffness of the two torsion microbeams. The pressure $p(x,y,t)$ is governed by (3.5), where $h(x,t) = g_0 + x\gamma(t)$. The pressure boundary conditions for the case in Figure 2 are

$$p(x,0,t) = p(x,l_y,t) = p\left(-\frac{l_x}{2},y,t\right) = p\left(\frac{l_x}{2},y,t\right) = p_0. \qquad (4.2)$$

For convenience, we introduce the nondimensional variables

$$X = \frac{x}{l_x} + \frac{1}{2}, \quad X_1 = \frac{x_1}{l_x} + \frac{1}{2}, \quad X_2 = \frac{x_2}{l_x} + \frac{1}{2}, \quad Y = \frac{y}{l_y}, \quad \vartheta = \frac{\gamma}{\gamma_{\max}}, \quad \gamma_{\max} = \frac{2g_0}{l_x},$$

$$\widehat{P} = \frac{p}{p_0}, \quad T = \frac{t}{S}, \quad H = \frac{h}{g_0} = 1 + 2\left(X - \frac{1}{2}\right)\vartheta, \tag{4.3}$$

where $S = \sqrt{J/(k_{T-E})} = 1/\omega_n$, ω_n is the nature frequency of the plate. Substituting (4.3) into (4.1), (3.5) and (4.2), we obtain

$$\ddot{\vartheta} + \vartheta = -\alpha V_0 v + P_{\text{non}} \iint_0^1 \left(\widehat{P}(X, Y, T) - 1\right)\left(X - \frac{1}{2}\right) dX\, dY, \tag{4.4}$$

$$\frac{\partial}{\partial X}\left(H^3 \widehat{P}\frac{\partial \widehat{P}}{\partial X}\right) + \beta^2 \frac{\partial}{\partial Y}\left(H^3 \widehat{P}\frac{\partial \widehat{P}}{\partial Y}\right) = \frac{\sigma}{S}\left(H\frac{\partial \widehat{P}}{\partial T} + \widehat{P}\frac{\partial H}{\partial T}\right), \tag{4.5}$$

where $\alpha = (x_2^2 - x_1^2)\varepsilon l_y / k_{T-E} g_0^2 \gamma_{\max}$, $P_{\text{non}} = p_0 l_x^2 l_y / k_{T-e}\gamma_{\max}$, $\sigma = 12\eta l_x^2 / g_0^2 p_0$, and $\beta = l_x/l_y$. The nondimensional boundary conditions are

$$\widehat{P}(X, 0, T) = \widehat{P}(X, 1, T) = \widehat{P}(0, Y, T) = \widehat{P}(1, Y, T) = 1. \tag{4.6}$$

As mentioned above, the microplate is under small torsion oscillation around $\gamma_0 = 0$ and therefore the pressure variation from ambient in the squeeze film is also small, $\widehat{P}(X, Y, T)$ is given by

$$\widehat{P}(X, Y, T) = \frac{p}{p_0} = 1 + P(X, Y, T), \tag{4.7}$$

where $|P(X, Y, T)| \ll 1$. Substituting (4.7) into (4.5), and linearizing the outcome around p_0 and γ_0, we obtain

$$\frac{\partial^2 P}{\partial X^2} + \beta^2 \frac{\partial^2 P}{\partial Y^2} - \frac{\sigma}{S}\frac{\partial P}{\partial T} = \frac{2\sigma}{S}\left(X - \frac{1}{2}\right)\frac{\partial \vartheta}{\partial T}. \tag{4.8}$$

The boundary conditions for the case are

$$P(X, 0, T) = P(X, 1, T) = P(0, Y, T) = P(1, Y, T) = 0. \tag{4.9}$$

For a harmonic excitation, the ac component voltage $v(T)$ is given by

$$v(T) = v_0 e^{j\omega T S}. \tag{4.10}$$

Correspondingly, the steady-state solution of (4.4) and (4.8) may be expressed by

$$\vartheta(T) = A e^{j\omega TS},$$

$$P(X, Y, T) = A \cdot P_A(X, Y) e^{j\omega TS}, \tag{4.11}$$

where A is the complex amplitude to be determined. Substituting (4.11) into (4.8), we obtain

$$\frac{\partial^2 P_A(X, Y)}{\partial X^2} + \beta^2 \frac{\partial^2 P_A(X, Y)}{\partial Y^2} - j\sigma\omega P_A(X, Y) = j2\sigma\omega\left(X - \frac{1}{2}\right). \tag{4.12}$$

The boundary conditions are

$$P_A(X, 0) = P_A(X, 1) = P_A(0, Y) = P_A(1, Y) = 0. \tag{4.13}$$

4.2. Wavelet Interpolation Method for Squeeze Film Damping Equations

In this section, the approximate solution of $P_A(X, Y)$ can be approximated by (3.21). The weak form functional of (4.12) is

$$W(P_A) = \iint_\Omega \left\{ \frac{1}{2}\left[\left(\frac{\partial P_A}{\partial X}\right)^2 + \beta^2\left(\frac{\partial P_A}{\partial Y}\right)^2 + j\sigma\omega P_A^2\right] + j2\sigma\omega\left(X - \frac{1}{2}\right)P_A \right\} dX\, dY. \tag{4.14}$$

From the necessary conditions for the determination of the minimum W, we obtain

$$\delta W(P_A) = \iint_\Omega \left[\frac{\partial \delta P_A}{\partial X}\frac{\partial P_A}{\partial X} + \beta^2\frac{\partial \delta P_A}{\partial Y}\frac{\partial P_A}{\partial Y} + j\sigma\omega\delta P_A P_A + j2\sigma\omega\left(X - \frac{1}{2}\right)\delta P_A\right] dX\, dY = 0. \tag{4.15}$$

Substituting (3.21) into (4.15), leads to

$$\sum_{k=1}^{2^J-1}\sum_{k'=1}^{2^J-1}\left\{\iint_\Omega\left[\frac{\partial\Theta_{m,n}^J}{\partial X}\frac{\partial\Theta_{k,k'}^J}{\partial X} + \beta^2\frac{\partial\Theta_{m,n}^J}{\partial Y}\frac{\partial\Theta_{k,k'}^J}{\partial Y} + j\sigma\omega\Theta_{m,n}^J\Theta_{k,k'}^J\right]dX\,dY\right\}p_{k,k'}^J$$

$$= -j2\sigma\omega\iint_\Omega\left(X - \frac{1}{2}\right)\Theta_{m,n}^J dX\, dY, \quad \text{for } m, n = 1, 2, \ldots, \left(2^J - 1\right). \tag{4.16}$$

This is a $\left(2^J - 1\right)^2 \times \left(2^J - 1\right)^2$ linear system

$$\Theta\mathbf{p} = -j\sigma\omega\mathbf{E}, \tag{4.17}$$

where $\mathbf{p} = \left[p_{1,1}^J\, p_{1,2}^J\, \cdots\, p_{1,2^J-1}^J\, p_{2,1}^J\, p_{2,2}^J\, \cdots\, p_{2^J-1,2^J-1}^J\right]^T$ is an $\left(2^J - 1\right)^2 \times 1$ unknown coefficients' matrix, $\mathbf{E} = 2\left[\iint_\Omega (X{-}1/2)\Theta_{1,1}^J dX\, dY\, \iint_\Omega (X{-}1/2)\Theta_{1,2}^J dX\, dY \cdots \iint_\Omega (X{-}(1/2))\Theta_{2^J-1,2^J-1}^J dXdY\right]^T$

is a $\left(2^J - 1\right)^2 \times 1$ matrix, and Θ is a $\left(2^J - 1\right)^2 \times \left(2^J - 1\right)^2$ matrix. The entries in Θ are of the form

$$
\begin{aligned}
\widehat{\Theta}(i,j) &= \widehat{\Theta}\left(\left[(m-1)\left(2^J - 1\right) + n\right], \left[(k-1)\left(2^J - 1\right) + k'\right]\right) \\
&= \iint_\Omega \left[\frac{\partial\widehat{\theta}(2^J X - m)}{\partial X}\widehat{\theta}\left(2^J Y - n\right)\frac{\partial\widehat{\theta}(2^J X - k)}{\partial X}\widehat{\theta}\left(2^J Y - k'\right) \right. \\
&\qquad\qquad + \beta^2 \widehat{\theta}\left(2^J X - m\right)\frac{\partial\widehat{\theta}(2^J Y - n)}{\partial Y}\widehat{\theta}\left(2^J X - k\right)\frac{\partial\widehat{\theta}(2^J Y - k')}{\partial Y} \\
&\qquad\qquad \left. + j\sigma\omega\widehat{\theta}\left(2^J X - m\right)\widehat{\theta}\left(2^J Y - n\right)\widehat{\theta}\left(2^J X - k\right)\widehat{\theta}\left(2^J Y - k'\right) \right] dX\, dY.
\end{aligned}
\tag{4.18}
$$

4.2.1. Squeeze Film Damping of the Torsion Plate

The numerical solution of (4.17) can be written as

$$
\mathbf{p} = -j\sigma\omega\Theta^{-1}\mathbf{E}.
\tag{4.19}
$$

The elements of $\mathbf{p}$ can be expressed as

$$
p_{k,k'}^J = -p_{k,k'}^{J,R} - jp_{k,k'}^{J,I} \quad \text{for } m,n = 1,2,\ldots,\left(2^J - 1\right),
\tag{4.20}
$$

where $p_{k,k'}^{J,R}$ and $p_{k,k'}^{J,I}$ are the real and imaginary parts of $p_{k,k'}^J$, respectively. Using (4.20) and (4.4), the force acting on the plate owing to the pressure of the squeeze gas film can be rewritten as

$$
P_{\text{non}} \iint_0^1 \left(\widehat{P} - 1\right)\left(X - \frac{1}{2}\right) dX\, dY = -K_a \cdot \vartheta(T) - C_a \cdot \frac{d\vartheta(T)}{dT},
\tag{4.21}
$$

where $K_a \cdot \vartheta(T)$ and $C_a \cdot (d\vartheta(T)/dT)$ are the spring and damping components of the force

$$
\begin{aligned}
K_a &= P_{\text{non}} \sum_{k=1}^{2^J-1} \sum_{k'=1}^{2^J-1} p_{k,k'}^{J,R} \iint_0^1 \widehat{\theta}\left(2^J X - k\right)\widehat{\theta}\left(2^J Y - k'\right)\left(X - \frac{1}{2}\right) dX\, dY \\
C_a &= \frac{P_{\text{non}}}{\omega S} \sum_{k=1}^{2^J-1} \sum_{k'=1}^{2^J-1} p_{k,k'}^{J,I} \iint_0^1 \widehat{\theta}\left(2^J X - k\right)\widehat{\theta}\left(2^J Y - k'\right)\left(X - \frac{1}{2}\right) dX\, dY
\end{aligned}
\tag{4.22}
$$

Table 1: The parameters of the accelerometer presented by Veijola et al. [26].

Parameters	Values
Mass m_{plate}	4.9×10^{-6} kg
Spring constant k_{spring}	212.1 N/m
Gap spacing g_0	3.95 μm
Length of the moving masss l_x	2 960 μm
Width of the moving masss l_y	1 780 μm
Ambient pressure p_0	11 Pa
Bias voltage V_0	9 V
Effective viscosity η_{eff}	10.2×10^{-9} N·s·m^{-2}

Substituting (4.21) into (4.4), leads to

$$\frac{d^2\vartheta}{dT^2} + C_a \cdot \frac{d\vartheta}{dT} + (K_a + 1)\vartheta = \alpha V_0 v(T),$$

$$\vartheta(T) = Ae^{j\omega TS} = \frac{aV_0 v_0}{K_a + 1} \cdot \frac{1}{1 - (\omega^2(S^2/(K_a + 1))) + \left(j\omega(C_a S/(K_a + 1))\right)} e^{j\omega TS}.$$

(4.23)

The quality factor and the damped natural frequency are given in(3.32)

5. Comparisons with Experiments

Veijola et al. [26] conducted experiments to measure the frequency response of an accelerometer under the effect of squeeze film damping. Minikes et al. [27] measured the quality factors of two torsion rectangular mirrors at low pressure. In this section, the experimental results presented by Veijola et al. [26] and Minikes et al. [27] were used to verify the wavelet interpolation Galerkin method.

5.1. Comparisons with the Experimental Results of Veijola et al. [26]

In [26], Veijola et al. simulated the frequency response of an accelerometer with a spring-mass- damper model with a parallel-plate electrostatic force. The damping coefficient was estimated by the Blech model [28]. The spring constants and the gas pressures were estimated by curve fitting the experimental measurements. They compared their simulations with experimental data and found good agreement. The parameters for the accelerometer are listed in Table 1.

In this subsection, we use the wavelet interpolation Galerkin method to predict the frequency response of the accelerometer. Various numerical tests have been conducted by changing the degree of the Daubechies wavelet L and the number J of the scale. Better accuracy can be achieved by increasing L and J. The higher L is, the smoother the scaling function becomes. The price for the high smoothness is that its supporting domain gets larger. The higher J is, the more accurate the solution becomes. The number of differential equations and the CPU time increase significantly as J increases. In this work, only the solutions for $L = 6$ and $J = 4$ are presented, as the results for higher resolutions are indistinguishable from the exact solution.

For comparison purpose, we give the frequency responses of the accelerometer calculated by Blech's model [28] and the finite difference method, respectively. Blech [28] expanded the air film pressure into an assumed double sine series and derived an analytical expression for the spring and damping forces. In this subsection, the number of terms for the double sine series is taken as (5, 5), which shows good convergence. For the finite difference method, we use the following approximate formulae for a node (i, j) on the microplate:

$$\left.\frac{\partial P_A}{\partial X}\right|_{i,j} = \frac{P_A(i+1, j) - P_A(i-1, j)}{2\Delta_X},$$

$$\left.\frac{\partial P_A}{\partial Y}\right|_{i,j} = \frac{P_A(i, j+1) - P_A(i, j-1)}{2\Delta_Y},$$

$$\left.\frac{\partial^2 P_A}{\partial X^2}\right|_{i,j} = \frac{P_A(i+1, j) - 2P_A(i, j) + P_A(i-1, j)}{\Delta_X^2},$$

$$\left.\frac{\partial^2 P_A}{\partial Y^2}\right|_{i,j} = \frac{P_A(i, j+1) - 2P_A(i, j) + P_A(i, j-1)}{\Delta_Y^2}.$$

(5.1)

In this subsection, we assume that $\Delta_X = \Delta_Y = 1/2^J = 1/2^6$; thus the element size of the finite difference method is equal to the wavelet interpolation Galerkin method. Substituting (5.1) into (3.17), we obtain

$$\frac{P_A(i+1, j) + P_A(i-1, j) + \beta^2 P_A(i, j+1) + \beta^2 P_A(i, j-1)}{\Delta_X^2} - \left(\frac{2 + 2\beta^2}{\Delta_X^2} + j\sigma\omega\right) P_A(i, j) = -j\sigma\omega.$$
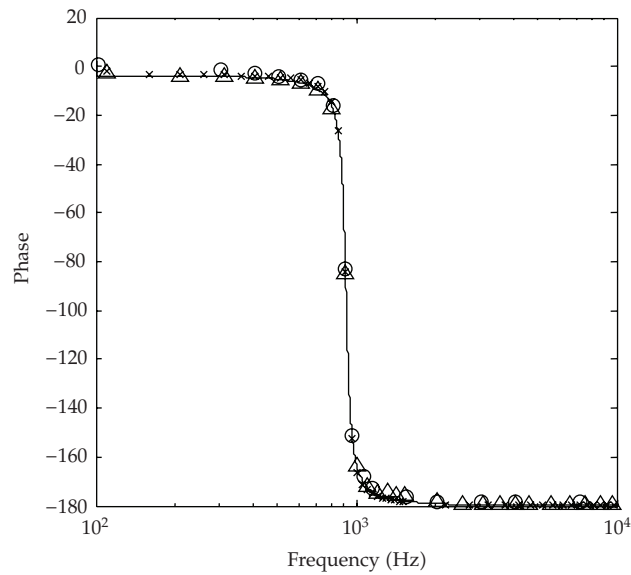
(5.2)

Figure 3 shows the comparisons of the frequency response obtained by different methods. As expected, the wavelet interpolation Galerkin method, Blech's model and the finite difference method give almost same results. The three results agree well with the experimental results [26] except for one data at resonance peak of the amplitude frequency response. The reason for this discrepancy is that the damping coefficient is slightly underestimated by the three methods, respectively. Table 2 shows the Comparison of the damping obtained by different methods. In the experiment [26], the squeeze film damping is dominant. Obviously, the accuracy of the finite difference method is less than the wavelet interpolation Galerkin method and Blech's model. The wavelet interpolation Galerkin method and the Blech model give almost identical results. Figure 4 shows the real part and the imaginary part of the air film pressure calculated by the wavelet interpolation Galerkin method.

5.2. Comparisons with the Experimental Results of Minikes et al. [27]

Minikes et al. [27] measured the quality factors of two rectangular torsion mirrors at low pressure and plotted the curves of the quality factors as a function of air pressure in the range from 10^{-2} torr to 10^2 torr. The structure of the two torsion mirrors is identical with the structure shown in Figure 2. The two torsion mirrors have similar dimensions in terms
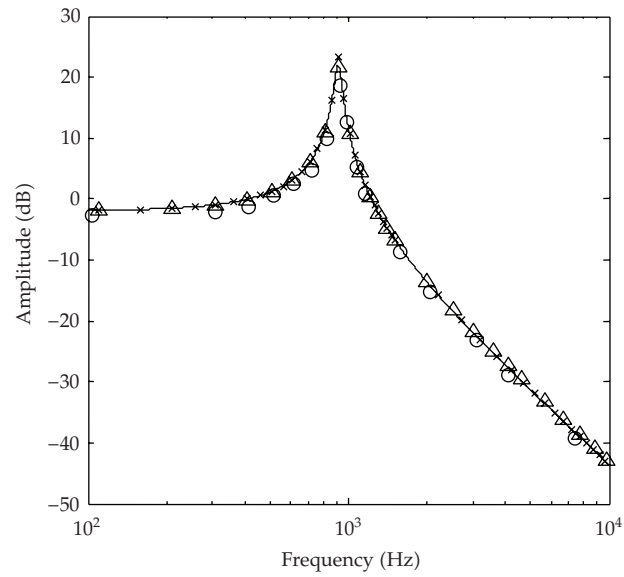
(a) Amplitude response

○   Experimental data [26]

△   Blech's model

✕   The finite difference method

—— The wavelet interpolation Galerkin method

(b) Phase response

Figure 3: Comparisons of the frequency response obtained by the wavelet Galerkin method, the Blech model and the experimental data of Veijola et al. [26].
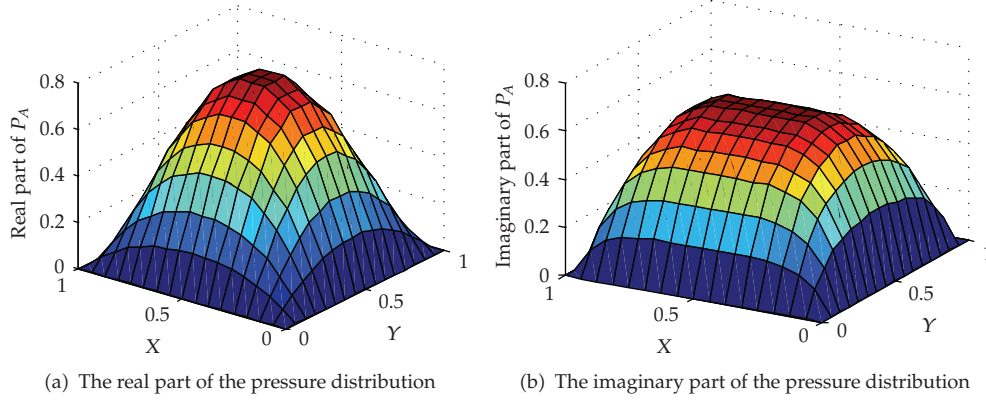
(a) The real part of the pressure distribution

(b) The imaginary part of the pressure distribution

Figure 4: The air film pressure distribution calculated by the wavelet interpolation Galerkin method.

Table 2: A Comparison of the damping ratios obtained by different methods with the experimental data [26].

Methods	Damping ratio	Peak value (dB)
Experimental data [26]	0.0239	18.3
Blech's model (error)	0.0159 (33.5%)	21.7
The wavelet interpolation Galerkin method (error)	0.0155 (35.1%)	21.9
The finite difference method (error)	0.0136 (43.1%)	23.2

of surface area and inertial moment, but have different gaps between the mirror and the actuation electrodes. The parameters for the two torsion mirrors are listed in Table 3.

In Table 3, the values of the two torsional natural frequencies are measured under the dc bias voltage. Based on the two torsional natural frequencies and two moments of inertia, we determined the torsional stiffness k_{T-E}. The extracted torsional stiffness k_{T-E} for mirror 1 and mirror 2 are 2.461×10^{-6} and 2.362×10^{-6} N·m/rad, respectively. Now we use the wavelet interpolation Galerkin method to predict the quality factors of the two torsion mirrors. Various numerical tests have been conducted by changing the degree of the Daubechies wavelet L and the number J of the scale. Only the solutions for $L = 6$ and $J = 4$ are presented, as the results for higher resolutions are indistinguishable from the exact solution.

For comparison purpose, we give the quality factors calculated by Pan's model [15] and the finite difference method, respectively. Pan et al. [15] expanded the air film pressure into an assumed double sine series and derived an analytical expression for the spring and damping torques. In this work, the number of terms for the double sine series is taken as (6, 5), which shows good convergence. For the finite difference method, we use (5.1) and (4.12) for a node (i, j) on the mirror; thus obtain

$$
\frac{P_A(i+1,j) + P_A(i-1,j) + \beta^2 P_A(i,j+1) + \beta^2 P_A(i,j-1)}{\Delta_X^2} - \left(\frac{2 + 2\beta^2}{\Delta_X^2} + j\sigma\omega \right) P_A(i,j)
$$

$$
= j2\sigma\omega\left(X_i - \frac{1}{2} \right).
$$

(5.3)

Table 3: The parameters of two torsion mirrors [27].

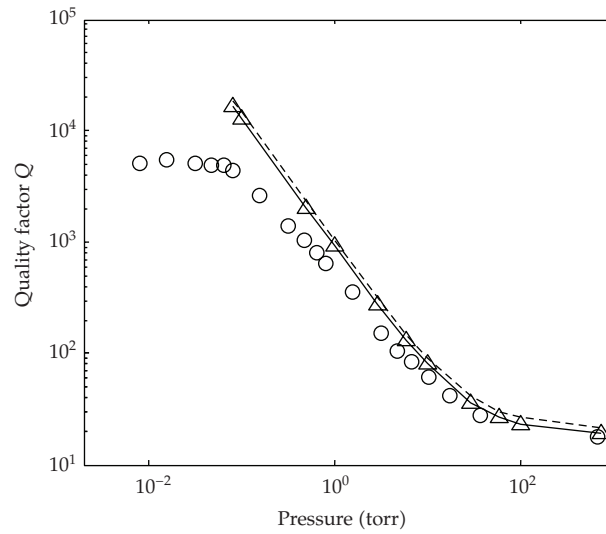|	Mirror 1	Mirror 2
Width of the mirror l_y	$500\,\mu$m	$500\,\mu$m
Length of the mirror l_x	$500\,\mu$m	$500\,\mu$m
Thickness of the mirror h_δ	$30\,\mu$m	$30\,\mu$m
Density of the mirror ρ	$2300\,\text{kg/m}^3$	$2300\,\text{kg/m}^3$
Gap spacing g_0	$28\,\mu$m	$13\,\mu$m
Torsional natural frequency f_n	13092.56 Hz	12824.87 Hz

Table 4: A comparison of quality factors obtained by Pan's model, the wavelet interpolation Galerkin method and the finite difference method for mirror 1.

p_0 (torr)	Pan's model	The wavelet interpolation Galerkin method (error)	The finite difference method (error)
0.08	5.784×10^4	5.987×10^4 (3.5%)	6.586×10^4 (13.9%)
0.10	4.483×10^4	4.627×10^4 (3.2%)	5.160×10^4 (15.1%)
0.50	7.070×10^3	7.330×10^3 (3.7%)	8.006×10^3 (13.2%)
1	3.306×10^3	3.388×10^3 (2.5%)	3.694×10^3 (11.7%)
3	1.061×10^3	1.080×10^3 (1.8%)	1.173×10^3 (10.6%)
6	5.680×10^2	5.895×10^2 (3.8%)	6.526×10^2 (14.9%)
10	3.979×10^2	4.052×10^2 (1.8%)	4.454×10^2 (11.9%)
30	2.451×10^2	2.505×10^2 (2.2%)	2.721×10^2 (11.0%)
60	2.074×10^2	2.160×10^2 (4.1%)	2.327×10^2 (12.2%)
100	1.994×10^2	2.035×10^2 (2.1%)	2.240×10^2 (12.3%)
760	1.850×10^2	1.896×10^2 (2.5%)	2.109×10^2 (14.0%)

In this subsection, we assume that $\Delta_X = \Delta_Y = 1/2^J = 1/2^6$; thus the finite difference method yields the same grids as the wavelet interpolation Galerkin method.

Tables 4 and 5 show the comparisons of quality factors obtained by the wavelet interpolation Galerkin method, the finite difference method and Pan's model for the mirror 1 and 2, respectively. As shown in Tables 4 and 5, the qualify factors obtained by the wavelet interpolation Galerkin method are almost 1% ~4% higher than Pan's model. However the qualify factors obtained by the finite difference method are almost 10% ~15% higher than Pan's model. The result of the wavelet interpolation Galerkin method matches the result of Pan's model better than that of the finite difference method.

Figure 5 shows the comparisons of quality factors obtained by different methods. As expected, the wavelet interpolation Galerkin method, the finite difference method and Pan's model give almost same results. Above $p_0 = 10$ torr, the viscous damping is dominant, the three methods give results in good agreement with the experimental results [27]. Below $p_0 = 10$ torr, the accuracy of the three methods decreases as the pressure decreases. The main reason for this trend are as follows. The three methods are based on Reynolds equation, which is derived from the Navier-Stokes equations and the continuity equation. The main assumption is that the gas in the gap can be treated as a continuum. Below $p_0 = 10$ torr, the Knudsen number Kn > 0.1, the gas in the gap cannot be treated as a continuum. Thus the three methods fail to give a good prediction. Below $p_0 = 0.1$ torr, the influence of squeeze film
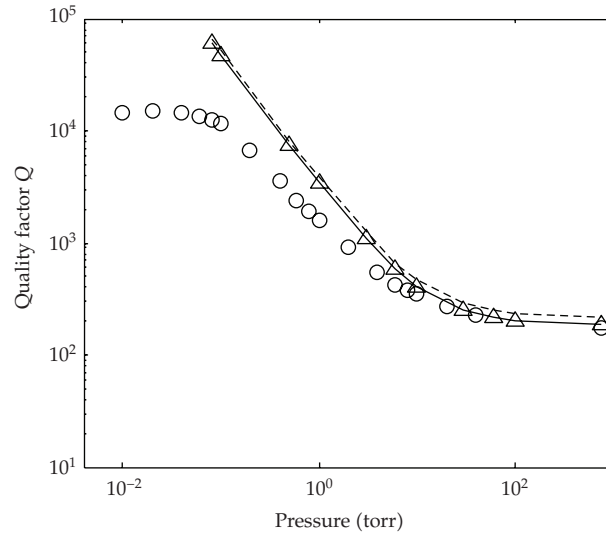
(a) Mirror 1 ($g_0 = 28\,\mu$m)

○   Experimental data [27]
△   Pan's model
- - -   The finite difference method
———   The wavelet interpolation Galerkin method

(b) Mirror 2 ($g_0 = 13\,\mu$m)

Figure 5: Comparison of the quality factors obtained by the wavelet Galerkin method, the Pan model and the experimental data of Minikes et al. [27].

damping begins to vanish and the qualify factors reaches a plateau that is dominated by the intrinsic damping.

Tables 6 and 7 list the comparisons of quality factors between $p = 10$ and 760 torr for mirrors 1 and 2, respectively. Obviously, the accuracy of the finite difference method is less than the wavelet interpolation Galerkin method and Pan's model. Figure 6 shows the real

(a) The real part of mirror 1

(b) The imaginary part of mirror 1

(c) The real part of mirror 2

(d) The imaginary part of mirror 2
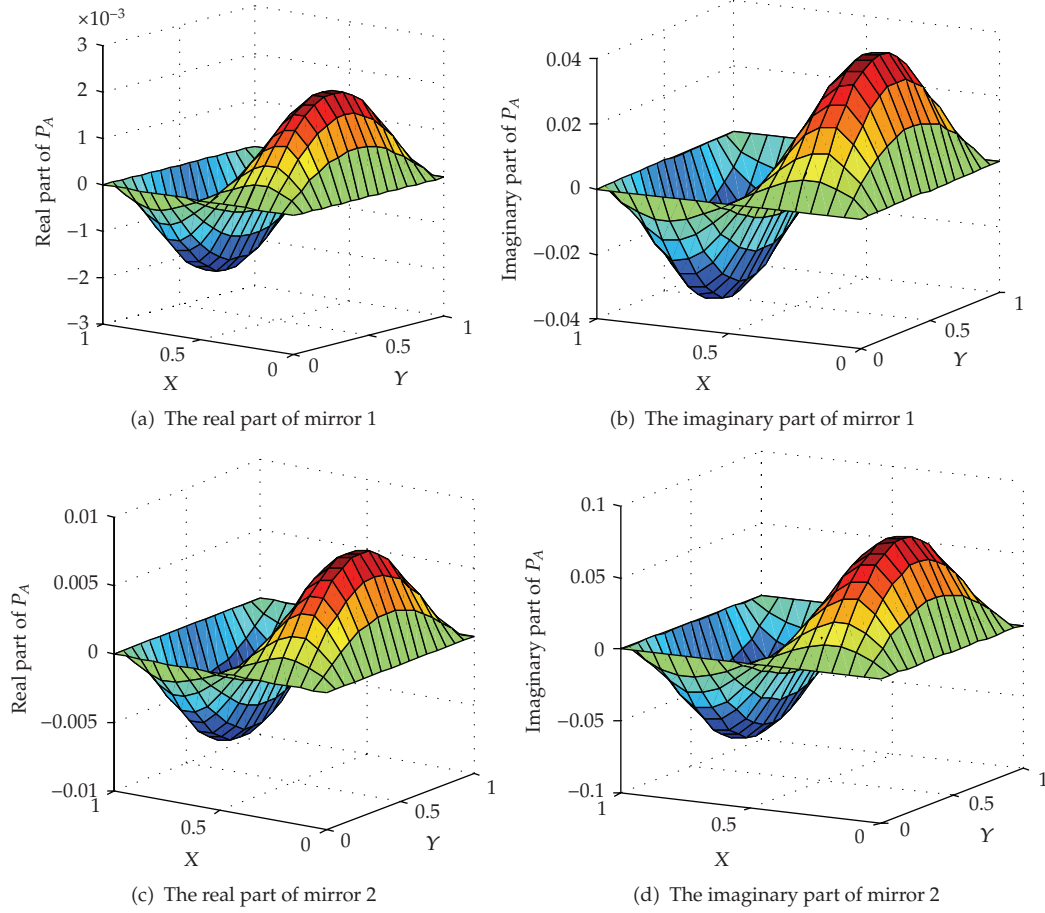
Figure 6: The air film pressure distribution calculated by the wavelet interpolation Galerkin method at p_0 = 1 torr.

Table 5: A comparison of quality factors obtained by Pan's model, the wavelet interpolation Galerkin method and the finite difference method for mirror 2.

p_0 (torr)	Pan's model	The wavelet interpolation Galerkin method (error)	The finite difference method (error)
0.08	1.645×10^4	1.679×10^4 (2.1%)	1.847×10^4 (12.3%)
0.10	1.268×10^4	1.297×10^4 (2.3%)	1.458×10^4 (15.0%)
0.50	2.135×10^3	2.208×10^3 (3.4%)	2.352×10^3 (10.2%)
1	8.919×10^2	9.195×10^2 (3.1%)	1.010×10^3 (13.2%)
3	2.681×10^2	2.712×10^2 (1.2%)	3.012×10^2 (12.3%)
6	1.282×10^2	1.318×10^2 (2.8%)	1.463×10^2 (14.1%)
10	8.004×10^1	8.118×10^1 (1.4%)	9.092×10^1 (13.6%)
30	3.461×10^1	3.599×10^1 (4.0%)	3.959×10^1 (14.4%)
60	2.574×10^1	2.628×10^1 (2.1%)	2.970×10^1 (15.4%)
100	2.255×10^1	2.278×10^1 (1.0%)	2.593×10^1 (15.0%)
760	1.820×10^1	1.900×10^1 (4.4%)	2.055×10^1 (12.9%)

Table 6: A comparison of quality factors between $p_0 = 10$ and 760 torr for mirror 1.

p_0 (torr)	Experiment data [27]	Pan's model (error)	The wavelet interpolation Galerkin method (error)	The finite difference method (error)
10	345	398 (15.4%)	405 (17.4%)	445 (29.0%)
20	275	289 (5.1%)	301 (9.4%)	311 (13.1%)
40	229	236 (3.1%)	241 (5.2%)	252 (10.0%)
760	179	185 (3.4%)	190 (6.1%)	211 (17.9%)

Table 7: A comparison of quality factors between $p_0 = 10$ and 760 torr for mirror 2.

p_0 (torr)	Expertiment data [27]	Pan's model (error)	The wavelet interpolation Galerkin method (error)	The finite difference method (error)
10	66	80.0 (21.2%)	81.2 (23.0%)	90.9 (37.7%)
20	40	48.8 (22.0%)	50.1 (25.3%)	54.0 (35.0%)
40	27	30.1 (11.5%)	31.0 (14.8%)	33.5 (24.1%)
760	18	18.2 (1.1%)	19.0 (5.6%)	20.6 (14.4%)

part and the imaginary part of the air film pressure calculated by the wavelet interpolation Galerkin method at $p_0 = 1$ torr. The air film pressure looks similar to the results calculated by Pan's model.

6. Summary and Conclusions

A new wavelet interpolation Galerkin method has been developed for the numerical simulation of MEMS devices under the effect of squeeze film damping. The air film pressure are expressed as a linear combination of a class of interpolating functions generated by autocorrelation of the usual compactly supported Daubechies scaling functions. To the best of our knowledge, this is the first time that wavelets have been used as basis functions for solving the PDEs of MEMS devices. As opposed to the previous wavelet-based methods that are all limited in one energy domain, the MEMS devices in the paper involve two coupled energy domains. Two typical electrically actuated micro devices with squeeze film damping effect are examined respectively to illustrate the wavelet interpolation Galerkin method. The method is validated by comparing its results with available theoretical and experimental results. The accuracy of the method is higher than the finite difference method.

In this paper, the wavelet interpolation Galerkin method is not suitable to solve problems defined on nonrectangular domains, since higher-dimensional wavelets are constructed by employing the tensor product of the one-dimensional wavelets and so their application is restricted to rectangular domains. In this paper, both trial and weight functions are a class of interpolating functions generated by autocorrelation of the first-generation wavelets. Future area of research is based on the second-generation wavelets [29]. The main

advantage of the second-generation wavelets is that the wavelets are constructed in the spatial domain and can be custom designed for complex domains. This work is currently under way.

Acknowledgments

References

[1] A. H. Nayfeh, M. I. Younis, and E. M. Abdel-Rahman, "Reduced-order models for MEMS applications," *Nonlinear Dynamics*, vol. 41, no. 1–3, pp. 211–236, 2005.

[2] E. S. Hung and S. D. Senturia, "Generating efficient dynamical models for microelectromechanical systems from a few finite-element simulation runs," *Journal of Microelectromechanical Systems*, vol. 8, no. 3, pp. 280–289, 1999.

[3] A. H. Nayfeh and M. I. Younis, "A new approach to the modeling and simulation of flexible microstructures under the effect of squeeze-film damping," *Journal of Micromechanics and Microengineering*, vol. 14, no. 2, pp. 170–181, 2004.

[4] M. I. Younis, E. M. Abdel-Rahman, and A. Nayfeh, "A reduced-order model for electrically actuated microbeam-based MEMS," *Journal of Microelectromechanical Systems*, vol. 12, no. 5, pp. 672–680, 2003.

[5] E. M. Abdel-Rahman, M. I. Younis, and A. H. Nayfeh, "Characterization of the mechanical behavior of an electrically actuated microbeam," *Journal of Micromechanics and Microengineering*, vol. 12, no. 6, pp. 759–766, 2002.

[6] Y. C. Liang, W. Z. Lin, H. P. Lee, S. P. Lim, K. H. Lee, and H. Sun, "Proper orthogonal decomposition and its applications—part II: model reduction for MEMS dynamical analysis," *Journal of Sound and Vibration*, vol. 256, no. 3, pp. 515–532, 2002.

[7] Y. C. Liang, W. Z. Lin, H. P. Lee, S. P. Lim, K. H. Lee, and D. P. Feng, "A neural-network-based method of model reduction for the dynamic simulation of MEMS," *Journal of Micromechanics and Microengineering*, vol. 11, no. 3, pp. 226–233, 2001.

[8] W. Z. Lin, K. H. Lee, S. P. Lim, and Y. C. Liang, "Proper orthogonal decomposition and component mode synthesis in macromodel generation for the dynamic simulation of a complex MEMS device," *Journal of Micromechanics and Microengineering*, vol. 13, no. 5, pp. 646–654, 2003.

[9] P. Li, R. Hu, and Y. Fang, "A new model for squeeze-film damping of electrically actuated microbeams under the effect of a static deflection," *Journal of Micromechanics and Microengineering*, vol. 17, no. 7, pp. 1242–1251, 2007.

[10] L. D. Gabbay, J. E. Mehner, and S. D. Senturia, "Computer-aided generation of nonlinear reduced-order dynamic macromodels—I: non-stress-stiffened case," *Journal of Microelectromechanical Systems*, vol. 9, no. 2, pp. 262–269, 2000.

[11] J. E. Mehner, L. D. Gabbay, and S. D. Senturia, "Computer-aided generation of nonlinear reduced-order dynamic macromodels—II: stress-stiffened case," *Journal of Microelectromechanical Systems*, vol. 9, no. 2, pp. 270–278, 2000.

[12] M. I. Younis and A. H. Nayfeh, "Simulation of squeeze-film damping of microplates actuated by large electrostatic load," *Journal of Computational and Nonlinear Dynamics*, vol. 2, no. 3, pp. 232–241, 2007.

[13] C. Zhang, G. Xu, and Q. Jiang, "Characterization of the squeeze film damping effect on the quality factor of a microbeam resonator," *Journal of Micromechanics and Microengineering*, vol. 14, no. 10, pp. 1302–1306, 2004.

[14] A. K. Pandey and R. Pratap, "Effect of flexural modes on squeeze film damping in MEMS cantilever resonators," *Journal of Micromechanics and Microengineering*, vol. 17, no. 12, pp. 2475–2484, 2007.

[15] F. Pan, J. Kubby, E. Peeters, A. T. Tran, and S. Mukherjee, "Squeeze film damping effect on the dynamic response of a MEMS torsion mirror," *Journal of Micromechanics and Microengineering*, vol. 8, no. 3, pp. 200–208, 1998.

[16] I. Daubechies, *Ten Lectures on Wavelets*, vol. 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, SIAM, Philadelphia, Pa, USA, 1992.

[17] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, Calif, USA, 1998.

[18] M.-Q. Chen, C. Hwang, and Y.-P. Shih, "The computation of wavelet-Galerkin approximation on a bounded interval," *International Journal for Numerical Methods in Engineering*, vol. 39, no. 17, pp. 2921–2944, 1996.

[19] S. L. Ho and S. Y. Yang, "Wavelet-Galerkin method for solving parabolic equations in finite domains," *Finite Elements in Analysis and Design*, vol. 37, no. 12, pp. 1023–1037, 2001.

[20] S. Yang, G. Ni, J. Qiang, and R. Li, "Wavelet-Galerkin method for computations of electromagnetic fields," *IEEE Transactions on Magnetics*, vol. 34, no. 5, part 1, pp. 2493–2496, 1998.

[21] S. Bertoluzza and G. Naldi, "A wavelet collocation method for the numerical solution of partial differential equations," *Applied and Computational Harmonic Analysis*, vol. 3, no. 1, pp. 1–9, 1996.

[22] Y. Liu, I. T. Cameron, and F. Y. Wang, "The wavelet-collocation method for transient problems with steep gradients," *Chemical Engineering Science*, vol. 55, no. 9, pp. 1729–1734, 2000.

[23] Y.-H. Zhou and J. Zhou, "A modified wavelet approximation for multi-resolution AWCM in simulating nonlinear vibration of MDOF systems," *Computer Methods in Applied Mechanics and Engineering*, vol. 197, no. 17-18, pp. 1466–1478, 2008.

[24] W. Yan, X. Shen, and L. Shi, "A wavelet interpolation galerkin method for the numerical solution of PDEs in 2D magnetostatic field," *IEEE Transactions on Magnetics*, vol. 36, no. 4, part 1, pp. 639–643, 2000.

[25] Y. Y. Kim and G.-W. Jang, "Hat interpolation wavelet-based multi-scale Galerkin method for thin-walled box beam analysis," *International Journal for Numerical Methods in Engineering*, vol. 53, no. 7, pp. 1575–1592, 2002.

[26] T. Veijola, H. Kuisma, J. Lahdenperä, and T. Ryhänen, "Equivalent-circuit model of the squeeze gas film in a silicon accelerometer," *Sensors and Actuators A*, vol. 48, pp. 239–248, 1995.

[27] A. Minikes, I. Bucher, and G. Avivi, "Damping of a micro-resonator torsion mirror in rarefied gas ambient," *Journal of Micromechanics and Microengineering*, vol. 15, no. 9, pp. 1762–1769, 2005.

[28] J. J. Blech, "On isothermal squeeze films," *Journal of Lubrication Technology A*, vol. 105, no. 4, pp. 615–620, 1983.

[29] O. V. Vasilyev and C. Bowman, "Second-generation wavelet collocation method for the solution of partial differential equations," *Journal of Computational Physics*, vol. 165, no. 2, pp. 660–693, 2000.

Research Article

Modeling and Fuzzy PDC Control and Its Application to an Oscillatory TLP Structure

Cheng-Wu Chen

Department of Logistics Management, Shu-Te University, 59 Hun Shan Road, Yen Chau,
Kaohsiung County, 82445, Taiwan

Correspondence should be addressed to Cheng-Wu Chen, cwchen@stu.edu.tw

An analytical solution is derived to describe the wave-induced flow field and surge motion of a deformable platform structure controlled with fuzzy controllers in an oceanic environment. In the controller design procedure, a parallel distributed compensation (PDC) scheme is utilized to construct a global fuzzy logic controller by blending all local state feedback controllers. The Lyapunov method is used to carry out stability analysis of a real system structure. The corresponding boundary value problems are then incorporated into scattering and radiation problems. These are analytically solved, based on the separation of variables, to obtain a series of solutions showing the harmonic incident wave motion and surge motion. The dependence of the wave-induced flow field and its resonant frequency on wave characteristics and structural properties including platform width, thickness and mass can thus be drawn with a parametric approach. The wave-induced displacement of the surge motion is determined from these mathematical models. The vibration of the floating structure and mechanical motion caused by the wave force are also discussed analytically based on fuzzy logic theory and the mathematical framework to find the decay in amplitude of the surge motion in the tension leg platform (TLP) system. The expected effects of the damping in amplitude of the surge motion due to the control force on the structural response are obvious.

1. Introduction

The supply and demand for the products of natural energy have been rapidly increasing in recent years. This has motivated oil/gas producers to explore even the ocean depths to obtain these resources. This has necessitated the in-depth study and analysis of deep water structures, such as the tension leg platform (TLP), which are particularly suited for water depths greater than 300 meters [1–4]. A TLP is a vertically moored floating structure often used for oil exploration and ocean thermal energy conversion (OTEC) [5]. This type of structure is particularly suited for water depths greater than 300 meters. They have also been used in oceanic environments as floating breakwaters [6]. The platform is permanently moored by means of "tethers" or "tendons" grouped at each of the structure's corners.

Each group of tethers forms a "tension leg." A typical tension leg platform usually includes a floating body which is anchored to the seafloor by the pretensioned legs. This semi-submerged floating body may be comprised of different materials with a working platform on top. Previous studies have introduced simplified models considering the surge motion of an impermeable structure with linearly elastic, pretensioned legs [7–10]. In studies of actual engineering applications the motion of the tethers is simplified as on-line rigid-body motion proportional to the top platform for estimating the pretension force without the drag effect. The resultant drag effect on the TLP system gives rise to an important issue related to solution types for the TLP system. Lee and Wang present a complete analytical solution for the dynamic behavior of both the platform and tethers in the TLP system [11]. In their study the TLP is subjected to wave-induced surge motion and the flow-induced drag motion, where the drag surge motion of the TLP is considered. Some research results show that the tether drag significantly reduces the response of the TLPs [11, 12]. In those studies, however, only surge motion was considered for the interaction between the wave and the platform structure.

The reduction of the vibration of structures by passive methods such as mass, stiffness, and damping has long been a subject of study in the fields of oceanic, mechanical, and civil engineering. Interest in the usage of tuned mass dampers and tuned liquid column dampers for reducing vibration in oceanic structures has also increased (see [12, 13] and the references therein). There have been numerous applications of passive damping in structural systems leading to a significant improvement in stability. Passive damping is regarded as a simple, low-cost, and easily implemented method for reducing structural vibrations. Another advantage of passive damping is that it ensures that the control system can remain stable when uncertainties or disturbances exist in the structural system. However, TLP systems have infinite numbers of resonant frequencies (or structural modes); so to meet all exigencies one would need an equal number of absorbers to minimize each mode. This is clearly impractical. Therefore, this study develops an innovative approach for achieving an analytical methodology of fuzzy control for the stabilization of TLP systems. In the remainder of this article, we examine a methodology for designing fuzzy controllers which uses the Lyapunov criterion to ensure the stability of the controlled systems.

Much work has been carried out in both industry and academic fields since Zadeh [14] first proposed using a linguistic approach to simulate the thought processes and judgment of human beings in fuzzy logic (see [15–17] and the references therein). However, there are few available mathematical theories or systematic designs for any of these approaches. Takagi and Sugeno [18] proposed a new fuzzy inference system, called the Takagi-Sugeno (T-S) fuzzy model, which combines the flexibility of fuzzy logic theory and rigorous mathematical analysis tools into a unified framework. Linear models are employed for the consequent parts which makes it simple to use conventional linear system theory for the analysis. The advantage of T-S fuzzy modeling is its greater accuracy, dimensionality, and also the simplification of the structure of nonlinear systems. It provides an effective representation of complex nonlinear systems (see [19–24] and the references therein).

In recent years, many control methods have been proposed and there are increasing research activities in the field of structural systems. These methods include optimal control, fuzzy control, pole placement, and sliding mode control (see [25–33] and the references therein). Hsiao et al. proposed a model-based fuzzy control method for dealing with the structure responses and stabilization analysis of passive tuned mass damper systems [33]. The control of TLPs, however, is more complicated, involving boundary value problems and structure resonant response. The analysis of the stability and stabilization problem of TLPs
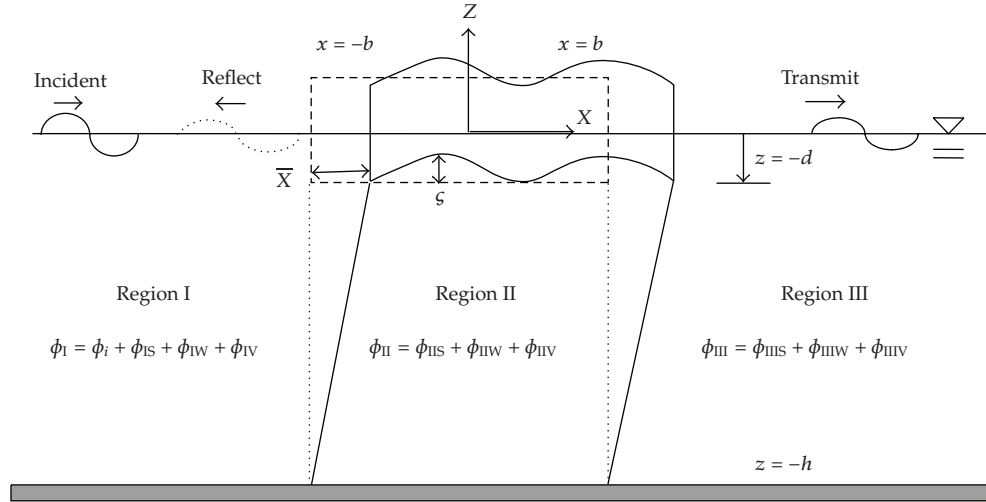
Figure 1: Sketch of a deformable tension leg platform subjected to the wave force.

has not often been discussed. In this study, we apply a fuzzy control technique as well as a T-S fuzzy model to deal with TLP stability problems. A suitable mathematical model of the TLP system and the interaction between a deformable floating structure and surface wave motion are discussed incorporating a partial differential equation as well as fuzzy logic theory. The effects on the dragged surge vibration of a TLP system and the interaction between wave and structure of a set of PDC fuzzy controllers on a platform are considered. It is the purpose of this study to carry out stability analysis and examine the influence of the fuzzy controller on platform motion when the true behavior of the deformable TLP is included in the solution. Along with the solution which is presented as a comparison of the responses between the platform with and without fuzzy controllers, the influences of parameters on the floating structure, the water wave, and the fuzzy controller system are also studied.

2. Boundary Value Problems

The flow field is first divided into three columns with two artificial boundaries at $x = -b$ and $x = b$, as seen in Figure 1, in which b denotes platform width (thus the platform width is $2b$) and d is the platform draft. In Region I, $-\infty < x < -b$, the total velocity potential ϕ_I consists of incident waves ϕ_i, scattered waves ϕ_IS, motion radiation waves ϕ_IW, and vibration radiation waves ϕ_IV. In Region II, $-b < x < b$, and in Region III, $b < x < \infty$, the total velocity potentials ϕ_II and ϕ_III consist of both scattered, ϕ_IIS and ϕ_IIIS, and radiated waves, ϕ_IIW, ϕ_IIIW, ϕ_IIV, and ϕ_IIIV, as seen in Figure 1, where the subscript "s" denotes the scattering problem, "w" the wave-maker (i.e., primitive radiation) problem induced by surge motion, and "v" the vibration radiation problem induced by the deformable platform. The corresponding boundary value problems for the scattering and radiation problems are given in Figure 2, where the displacement of the surge motion with unknown amplitude S is given by $\overline{X} = Se^{-i\sigma t}$, and the deformation of the platform on z-axis is defined as ς, which is function of x and t.

For a discussion of the kinematic boundary condition, dynamic boundary condition and radiation condition please refer to the stability analysis in [1].
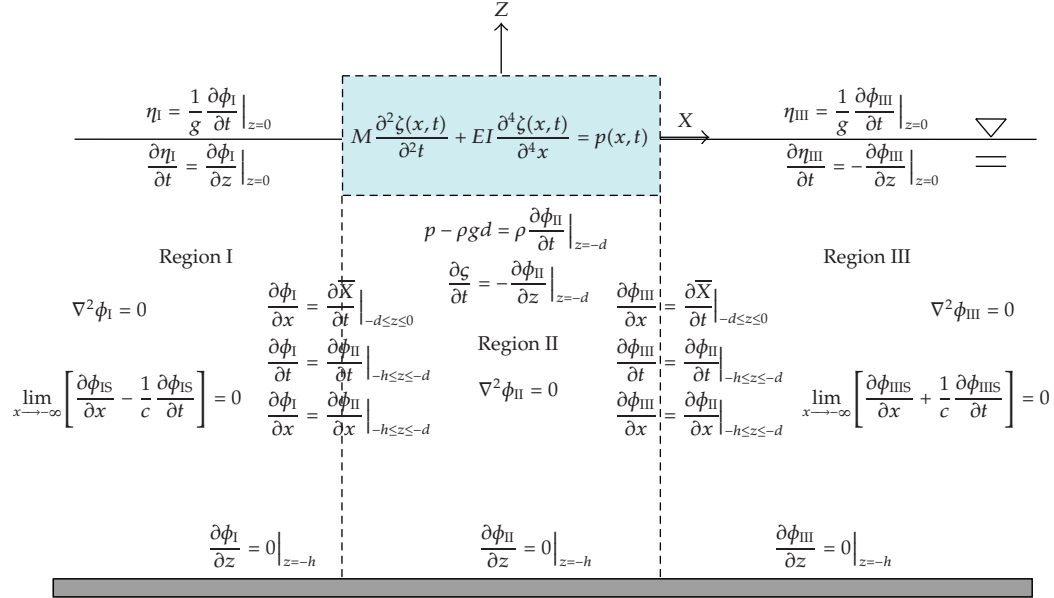
Figure 2: Boundary value problems for scattering and radiation problems.

3. Analytical Solutions

Analytical solutions of the scattering problem, motion radiation problem, and vibration radiation problem can also be derived using a similar approach [1].

3.1. Fuzzy Modeling of a Structural System

Recently, fuzzy-rule-based modeling has become an active research field because of its unique merits for solving complex nonlinear system identification and control problems. Unlike traditional modeling, fuzzy rule-based modeling is essentially a multimodel approach in which individual rules are combined to describe the global behavior of the system [19, 34]. Therefore, fuzzy modeling is employed to represent a structural system in order to simplify the controller design problem.

To ensure the stability of the structural system, Takagi-Sugeno (T-S) fuzzy models and the stability analysis are recalled. There are two approaches for constructing fuzzy models mentioned in literature [35]:

(1) identification (fuzzy modeling) using input-output data,

(2) derivation from given nonlinear interconnected system equations.

There has been extensive work done on fuzzy modeling using input-output data following Takagi's, Sugeno's, and Kang's outstanding work (see [35], and the references therein). The procedure is mainly composed of two parts: structure identification and parameter identification. The identification approach is appropriate for plants that are unable or too difficult to be represented by analytical models. In some cases, because nonlinear dynamic models including algebraic equations for structure systems can be readily obtained, the second approach to derive a fuzzy model is more suitable. This study focuses on using the

second approach to constructing fuzzy models. The linear feedback control techniques can be utilized for feedback stabilization in the T-S type fuzzy model. The procedure is as follows: first, the plant is represented by a Takagi-Sugeno-type fuzzy model. In this type of fuzzy model, local dynamics in different state-space regions are represented by linear models. The overall model of the system is achieved by fuzzy "blending" of these linear models through nonlinear fuzzy membership functions.

3.2. Stability of Fuzzy Control

The momentum equation can be obtained from the motion of the floating structure, extensively derived from Newton's second law. Assume that the momentum equation of a TLP system controlled by actuators can be characterized by the following differential equation:

$$M\ddot{\overline{X}}(t) = BU(t) - M\overline{r}\phi(t), \tag{3.1}$$

where $\overline{X}(t) = [\overline{x}_1(t), \overline{x}_2(t) \cdots \overline{x}_n(t)] \in R^n$ is an n-vector; $\ddot{\overline{X}}(t)$, $\dot{\overline{X}}(t)$, and $\overline{X}(t)$ are acceleration, velocity and displacement vectors, respectively; $\overline{B}$ is an $(n \times m)$ matrix denoting the locations of m control forces; $U(t)$ corresponds to the actuator forces (generated via the active tendon system or an active mass damper). This is only a static model. $M\overline{r}\phi(t)$ is a wave-induced external force which can be expressed by

$$M\overline{r}\phi(t) = F_{wx} - F_{Tx}, \tag{3.2}$$

where F_{wx} is the horizontal wave force acting on the both sides of the structure, and F_{Tx} is the horizontal component of the static (or the pretensioned) tension applied by the tension legs. The static tension is given by $F_{Tx} = f\xi$, in which f is the stiffness coefficient defined by $f = 2\rho g b(d_2 - d_0)/(h - d_2)$. The structure parameters d_0 and d_2 are the drafts of the structure before and after applying pretension, respectively. According to the formula, d_0 increases as the structure's mass increases, and the pretension stiffness f is inversely proportional to the mass for the specific structural geometry.

For the controller design, the standard first-order state equation corresponding to (3.2) is obtained by [36]

$$\dot{X}(t) = AX(t) + BU(t) + E\phi(t), \tag{3.3}$$

where

$$X(t) = \begin{bmatrix} \overline{x}(t) \\ \dot{\overline{x}}(t) \end{bmatrix}, \qquad A = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ M^{-1}\overline{B} \end{bmatrix}, \qquad E = \begin{bmatrix} 0 \\ -\overline{r} \end{bmatrix}. \tag{3.4}$$

To ensure the stability of the TLP system, T-S fuzzy models and some stability analysis are utilized. To design fuzzy controllers, the structural systems are represented by Takagi-Sugeno fuzzy models. The concept of PDC is employed to determine the structures of the fuzzy controllers from the T-S fuzzy models in this section. Some detailed steps for designing

PDC fuzzy controllers are described in literature [35, 37]. First, the ith rule for the T-S fuzzy model, representing the structural system (3.3), is as follows.

Rule i: IF $x_1(t)$ is M_{i1} and $\cdots$ and $x_p(t)$ is M_{ip},

$$\text{THEN} \quad \dot{X}(t) = A_i X(t) + B_i U(t) + E_i \phi(t), \tag{3.5}$$

where $i = 1, 2, \ldots, r$ and r is the rule number; $X(t)$ is the state vector; M_{ip} $(p = 1, 2, \ldots, g)$ are the fuzzy sets; $x_1(t) \sim x_p(t)$ are the premise variables. By using the fuzzy inference method with a singleton fuzzifier, product inference, and center average defuzzifier, the dynamic fuzzy model (3.5) can be expressed as follows:

$$\dot{X}(t) = \frac{\sum_{i=1}^{r} w_i(t) \left[A_i X(t) + B_i U(t) + E_i \phi(t) \right]}{\sum_{i=1}^{r} w_i(t)} = \sum_{i=1}^{r} h_i(t) (A_i X(t) + B_i U(t)) + E_i \phi(t) \tag{3.6}$$

with

$$w_i(t) = \prod_{p=1}^{g} M_{ip}(x_p(t)), \qquad h_i(t) = \frac{w_i(t)}{\sum_{i=1}^{r} w_i(t)}. \tag{3.7}$$

$M_{ip}(x_p(t))$ is the grade of membership of $x_p(t)$ in M_{ip}. It is assumed that

$$w_i(t) \geq 0, \quad i = 1, 2, \ldots, r; \quad \sum_{i=1}^{r} w_i(t) > 0 \tag{3.8}$$

for all t. Therefore, $h_i(t) \geq 0$ and $\sum_{i=1}^{r} h_i(t) = 1$ for all t.

The PDC is adopted to design a global controller for the T-S fuzzy model (3.5). Using the same premise as (3.5) , the ith rule of the FLC can be obtained as follows.

Controller Rule i: IF $x_1(t)$ is M_{i1} and $\cdots$ and $x_g(t)$ is M_{ig},

$$\text{THEN} \quad U(t) = -K_i X(t) \quad (i = 1, 2, \ldots, r), \tag{3.9}$$

where K_i is the local feedback gain vector in the ith subspace. The final model-based fuzzy controller is analytically represented by

$$U(t) = -\frac{\sum_{i=1}^{r} w_i(t) K_i X(t)}{\sum_{i=1}^{r} w_i(t)} = -\sum_{i=1}^{r} h_i(t) K_i X(t). \tag{3.10}$$

The overall closed-loop controlled system obtained by combining (3.6) and (3.10) is

$$\dot{X}(t) = \sum_{i=1}^{r} \sum_{l=1}^{r} h_i(t) h_l(t) \left[(A_i - B_i K_l) X(t) \right] + E_i \phi(t). \tag{3.11}$$

A typical stability condition for a fuzzy system (3.11) is proposed here.

Table 1: Input data for case study of surge motion calculation.

Initial wave conditions	
Wave frequency σ	Variable
Wave amplitude A_i	0.5 m
Environmental conditions	
Water depth h	30 m
Gravitational force g	9.8 m/s^2
Structural parameters	
Mass M	10, 100, and 1000 kg
Draft d	2 m
Width b	10 m
Young's modulus E	1×10^7 N/m^2
Mass moment of inertia	10^{-2} m^4

Theorem 3.1. *The equilibrium point of the fuzzy control system* (3.11) *is stable in the large if there exist a common positive definite matrix P and feedback gains K such that the following two inequalities are satisfied:*

$$(A_i - B_i K_i)^T P + P(A_i - B_i K_i) + \frac{1}{\eta^2} P E_i E_i^T P < 0,$$

$$\left(\frac{(A_i - B_i K_l) + (A_l - B_l K_i)}{2} \right)^T P + P \left(\frac{(A_i - B_i K_l) + (A_l - B_l K_i)}{2} \right) + \frac{1}{\eta^2} P E_i E_i^T P < 0 \quad (3.12)$$

with $P = P^T > 0$, for $i < l \leq r$ and $i = 1, 2, \ldots, r$.

The proof is lengthy and can be derived by the similar approach by Hsiao et al. [33, 37–39] and Yeh et al. [31]. Therefore, the proof is removed in this paper.

4. Numerical Results

In the physical world, engineers measure and observe the effects induced by embedded structures. They might conduct an experiment or carry out real scale numerical simulations. We do not especially consider engineering oriented issues. Instead, we utilize sensitive concepts to explain the effects of physical parameters on a natural transformation. The dynamic response of a platform in a TLP system is dependent on a large set of parameters that include the characteristics of the waves and the material properties of the structure. Several parameters are particularly important for systematic consideration in TLP stability and stabilization problems. These include the amplitude and frequency of the waves (wave characteristics) as well as the structural properties of the platform (including draft, width, thickness mass, etc.). Only platform mass is considered in the following discussion. The parameters are given in Table 1. A plot of the reflection coefficient versus dimensionless wave frequency is shown in Figure 3. It can be seen from this plot that the reflection coefficient increases with the platform mass. On the other hand, as can be seen in Figure 4, the transmission coefficient decreases as the platform mass increases. Obviously, this is in good agreement with the opposing relationship between these two coefficients, which can be defined as the ratio of the amplitude
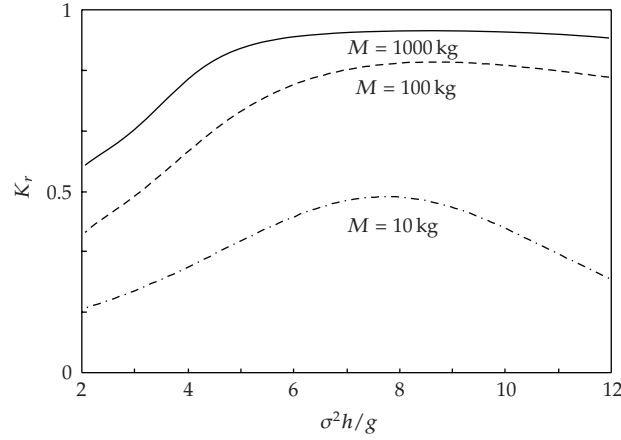
Figure 3: Comparison of the coefficients of reflection and dimensionless frequency for various structure masses.
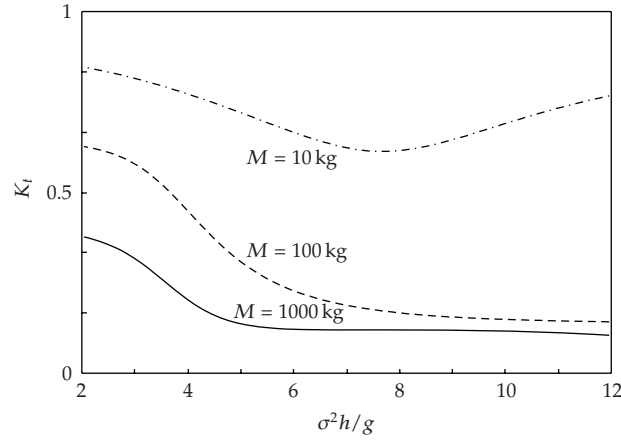

Figure 4: Comparison of the coefficients of transmission and dimensionless frequency for various structure masses.

of the reflected/transmitted wave and the amplitude of the incident wave. Figure 5 shows profiles of surge motion for structures of varied mass but fixed 0.5 m amplitude wave fluctuation. Generally speaking, structural displacement increases with decreasing mass. It is found that the greater the dimensionless wave frequency is, the greater surge motion will be. The calculated maximum displacement of surge motion for a 1000 kg platform mass is estimated to be approximately 60 times the wave amplitude. The dimensionless horizontal platform displacement becomes the largest when in the range of $\sigma^2 h/g$ = 0.4 to 1.0, with values of 58, 52, and 47. The resonance frequency also shifts to a lower value when the mass of the structure increases, which corresponds to a decreasing stiffness coefficient. It is commonly known that an increase in the structural mass reduces the equivalent stiffness caused by increasing draft. Therefore, platform response is enhanced in cases where the mass of the structure increases. This occurs because of the weaker pretension stiffness. In the TLP system, frequencies become smaller as the mass of the system increases. In addition, the control force acts to significantly mitigate the amplitude of the surge motion when the wave-structure interaction is taken
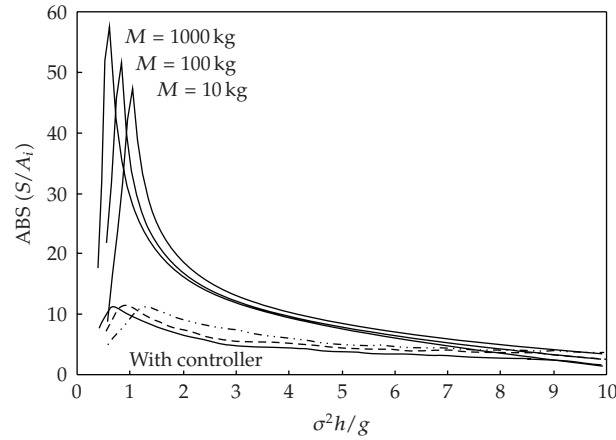
Figure 5: Comparison of the dimensionless amplitude of the surge motion between the platforms with and without controller interactions with respect to the variation of the structure mass.
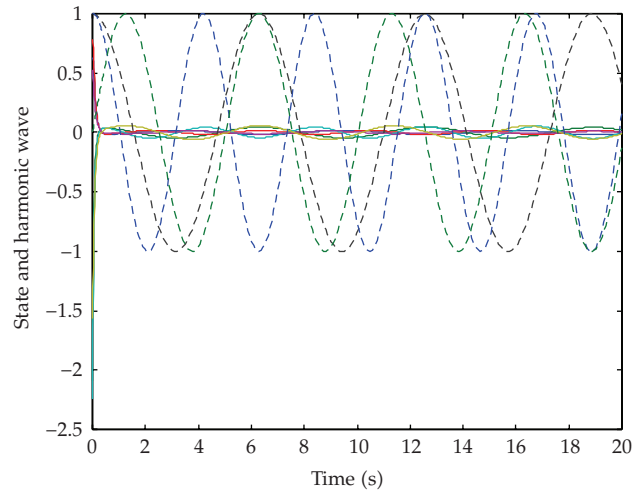

Figure 6: State response of the states and harmonic waves for three conditions.

into account. Table 2 shows the periodic wave-induced structure responses with respect to the several wave conditions shown in Figures 6 and 7. It can be seen by looking at Figure 6 that the state of displacement and velocity of TLP systems will be stabilized under external waves. It can be seen in Figure 7 that the control forces employed by PDC fuzzy schemes can stabilize systems with harmonic inputs. A look at these graphs shows that the TLP system is stable. Note that the trajectories for the three different conditions starting from the initial nonzero states all approach close to the origin under arbitrary harmonic excitation. This demonstrates that the effects of the control force have acted on the TLP system to stabilize the oceanic structure. A comparison of the results for these three figures shows the periodic wave fluctuation. The implication is that the control force enhances the decay of the structural displacement and motion velocity over time. The results of analysis show that despite the fact that much progress has been made in successfully applying a passive control force to TLP system, many basic issues remain that need to be further addressed.

Table 2: Input data for three parametric cases.

Subsystem 1	
Wave period T	10 s
Structure mass M	100 kg
Structure width	4 m
Subsystem 2	
Wave period T	12 s
Structure mass M	200 kg
Structure width	6 m
Subsystem 3	
Wave period T	8 s
Structure mass M	150 kg
Structure width	2 m

Others are as follows.
Wave amplitude $A_i = 0.5$ m; Water depth $h = 20$ m; Young's modulus $E = 10^7$ N/m^2; Mass moment of inertia $= 10^{-2}$ m^4; Draft $d = 2$ m.
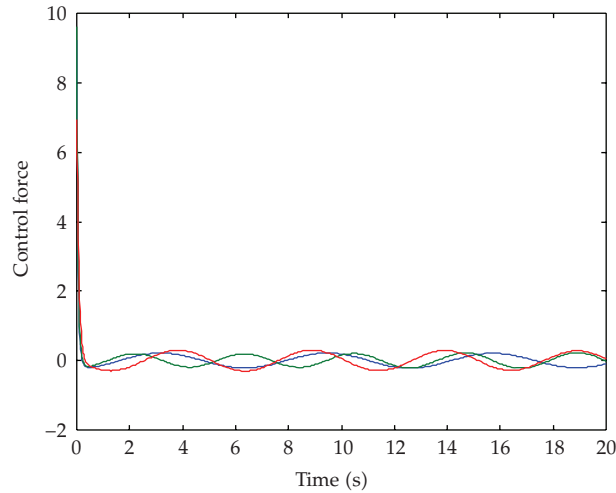
Figure 7: Control forces for three conditions.

Firstly, traditional structural design depends on structural strength and the capability to dissipate energy arising from dynamic forces such as machine loading, wind forces, and earthquakes. The control effects can be improved by the active control forces. Secondly, stability analysis and systematic design are certainly among the most important issues for active control systems. The proposed T-S fuzzy model and PDC fuzzy control (based on Lyapunov functions) can easily ensure stabilization of TLP structures. The simulation results demonstrate the TLP system to be stable. The trajectory of the initial conditions, starting from nonzero initial states, is kept within certain bounds under wave excitation.

5. Conclusion

In order to exploit rich new sources of oil and gas lying beneath the ocean, we must be ready to delve into very deep waters. Many promising fields are situated under water

that is in the 4000–7000 ft depth range. A favored platform design in that environment is the TLP system, where the hull is connected to the seabed by strings of tendon pipes. Since tendon length as well as water pressure increases in proportion to water depth, it is essential to produce a stable structure that the tendon pipes be highly resistant to collapse, especially in deep water environments. We present here a new control force concept for the stabilization of TLP systems, an alternative to the tether drag effects used previously. The results of the present study show that this new controller improves the limitations of steel performance for the maximum water depth attainable with the old TLP system. The dependence of wave reflection, transmission, as well as structure surge motion on incident wave conditions and structure properties has been demonstrated. The results show that the response of the floating structure is influenced by the its mass (as demonstrated by the different surge motion profiles). In particular, the response of a floating structure reaches the largest level of displacement during resonance. The resonant phenomenon takes place when the dimensionless wave frequency moves to lower values, approximately in the 0.5–1.0 range. The response to the wave-structure interaction results in platform vibration, but this can be slowed down and stabilized by means of the control force.

Acknowledgment

References

[1] C.-W. Chen, "The stability of an oceanic structure with T-S fuzzy models," *Mathematics and Computers in Simulation*, vol. 80, no. 2, pp. 402–426, 2009.

[2] S. Chandrasekaran, N. R. Chandak, and G. Anupam, "Stability analysis of TLP tethers," *Ocean Engineering*, vol. 33, no. 3-4, pp. 471–482, 2006.

[3] TLP, http://www.globalsecurity.org/military/systems/ship/platform-tension-leg.htm.

[4] M. M. Gadagi and H. Benaroya, "Dynamic response of an axially loaded tendon of a tension leg platform," *Journal of Sound and Vibration*, vol. 293, no. 1-2, pp. 38–58, 2006.

[5] IOA NEWSLETTER, "British Engineer Proposed A New CWP Design," International OTEC/DOWA Association, vol. 1, no. 3, 1990.

[6] V. W. Harms, "Design criteria for floating tire breakwaters," *Journal of Waterway, Port, Coastal, and Ocean Engineering*, vol. 105, no. 2, pp. 149–170, 1979.

[7] C. P. Lee and J. F. Lee, "Interaction between waves and tension leg platform," in *Proceedings of the Engineering Mechanics Conference on Mechanics Computing in 1990's and Beyond*, Columbus, Ohio, USA, May 1991.

[8] T. Yamamoto, A. Yoshida, and T. Ijima, "Dynamics of elastically moored floating objects," in *Dynamic Analysis for Offshore Structure*, C. L. Kirk, Ed., pp. 106–113, CML, Southampton, UK, 1982.

[9] C. C. Mei, "Numerical methods in water-wave diffraction and radiation," *Annual Review of Fluid Mechanics*, vol. 10, pp. 393–416, 1978.

[10] C.-P. Lee and J.-F. Lee, "Wave-induced surge motion of a tension leg structure," *Ocean Engineering*, vol. 20, no. 2, pp. 171–186, 1993.

[11] H. H. Lee and W.-S. Wang, "Analytical solution on the dragged surge vibration of tension leg platforms (TLPS) with wave large body and small body multi-interactions," *Journal of Sound and Vibration*, vol. 248, no. 3, pp. 533–556, 2001.

[12] H. H. Lee and W.-S. Wang, "On the dragged surge vibration of twin TLP systems with multi-interactions of wave and structures," *Journal of Sound and Vibration*, vol. 263, no. 4, pp. 743–774, 2003.

[13] S. D. Xue, J. M. Ko, and Y. L. Xu, "Optimal performance of the TLCD in structural pitching vibration control," *Journal of Vibration and Control*, vol. 8, no. 5, pp. 619–642, 2002.

[14] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

[15] S. S. L. Chang and L. A. Zadeh, "On fuzzy mapping and control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 2, pp. 30–34, 1972.

[16] W. J. M. Kickert and E. H. Mamdani, "Analysis of a fuzzy logic controller," *Fuzzy Sets and Systems*, vol. 1, no. 1, pp. 29–44, 1978.

[17] M. Braae and D. A. Rutherford, "Theoretical and linguistic aspects of the fuzzy logic controller," *Automatica*, vol. 15, no. 5, pp. 553–577, 1979.

[18] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 15, no. 1, pp. 116–132, 1985.

[19] C. W. Chen, W. L. Chiang, and F. H. Hsiao, "Stability analysis of T-S fuzzy models for nonlinear multiple time-delay interconnected systems," *Mathematics and Computers in Simulation*, vol. 66, no. 6, pp. 523–537, 2004.

[20] C.-W. Chen, "Modeling and control for nonlinear structural systems via a NN-based approach," *Expert Systems with Applications*, vol. 36, no. 3, part 1, pp. 4765–4772, 2009.

[21] C.-W. Chen, M. H. L. Wang, and J.-W. Lin, "Managing target the cash balance in construction firms using a fuzzy regression approach," *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*, vol. 17, no. 5, pp. 667–684, 2009.

[22] T.-Y. Hsieh, M. H. L. Wang, C.-W. Chen, et al., "A new viewpoint of s-curve regression model and its application to construction management," *International Journal on Artificial Intelligence Tools*, vol. 15, no. 2, pp. 131–142, 2006.

[23] M.-L. Lin, C.-W. Chen, Q.-B. Wang, et al., "Fuzzy model-based assessment and monitoring of desertification using MODIS satellite imagery," *Engineering Computations*, vol. 26, no. 7, pp. 745–760, 2009.

[24] C.-H. Tsai, C.-W. Chen, W.-L. Chiang, and M.-L. Lin, "Application of geographic information system to the allocation of disaster shelters via fuzzy models," *Engineering Computations*, vol. 25, no. 1, pp. 86–100, 2008.

[25] C.-Y. Chen, C.-W. Shen, C.-W. Chen, K. F.-R. Liu, and M.-J. Cheng, "A stability criterion for time-delay tension leg platform systems subjected to external force," *China Ocean Engineering*, vol. 23, no. 1, pp. 49–57, 2009.

[26] C. Y. Chen, J. W. Lin, W. I. Lee, and C. W. Chen, "Fuzzy control for an oceanic structure: A case study in time-delay TLP system," *Journal of Vibration and Control*, vol. 16, no. 1, pp. 147–160, 2010.

[27] P. C. Chen, C. W. Chen, and W. L. Chiang, "GA-based fuzzy sliding mode controller for nonlinear systems," *Mathematical Problems in Engineering*, vol. 2008, Article ID 325859, 16 pages, 2008.

[28] C.-W. Chen, K. Yeh, and K. F.-R. Liu, "Adaptive fuzzy sliding mode control for seismically excited bridges with lead rubber bearing isolation," *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*, vol. 17, no. 5, pp. 705–727, 2009.

[29] P. C. Chen, C.-W. Chen, and W. L. Chiang, "GA-based modified adaptive fuzzy sliding mode controller for nonlinear systems," *Expert Systems with Applications*, vol. 36, no. 3, part 2, pp. 5872–5879, 2009.

[30] P. C. Chen, C. W. Chen, W. L. Chiang, and K. Yeh, "A novel stability condition and its application to GA-based fuzzy control for nonlinear systems with uncertainty," *Journal of Marine Science and Technology*, vol. 17, no. 4, pp. 293–299, 2009.

[31] K. Yeh, C.-Y. Chen, and C.-W. Chen, "Robustness design of time-delay fuzzy systems using fuzzy Lyapunov method," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 568–577, 2008.

[32] L.-T. Lu, W.-L. Chiang, J.-P. Tang, M.-Y. Liu, and C.-W. Chen, "Active control for a benchmark building under wind excitations," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 91, no. 4, pp. 469–493, 2003.

[33] F.-H. Hsiao, C.-W. Chen, Y.-W. Liang, S.-D. Xu, and W.-L. Chiang, "T-S fuzzy controllers for nonlinear interconnected systems with multiple time delays," *IEEE Transactions on Circuits and Systems I*, vol. 52, no. 9, pp. 1883–1893, 2005.

[34] C.-W. Chen, "Stability conditions of fuzzy systems and its application to structural and mechanical systems," *Advances in Engineering Software*, vol. 37, no. 9, pp. 624–629, 2006.

[35] K. Tanaka and H. O. Wang, *Fuzzy Control Systems Design and Analysis*, John Wiley & Sons, New York, NY, USA, 2001.

[36] C.-W. Chen, K. Yeh, W.-L. Chiang, C.-Y. Chen, and D.-J. Wu, "Modeling, H^∞ control and stability analysis for structural systems using Takagi-Sugeno fuzzy model," *Journal of Vibration and Control*, vol. 13, no. 11, pp. 1519–1534, 2007.

[37] F.-H. Hsiao, C.-W. Chen, Y.-H. Wu, and W.-L. Chiang, "Fuzzy controllers for nonlinear interconnected TMD systems with external force," *Journal of the Chinese Institute of Engineers*, vol. 28, no. 1, pp. 175–181, 2005.

[38] F.-H. Hsiao, W.-L. Chiang, C.-W. Chen, S.-D. Xu, and S.-L. Wu, "Application and robustness design of fuzzy controller for resonant and chaotic systems with external disturbance," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 13, no. 3, pp. 281–295, 2005.

[39] F.-H. Hsiao, J.-D. Hwang, C.-W. Chen, and Z.-R. Tsai, "Robust stabilization of nonlinear multiple time-delay large-scale systems via decentralized fuzzy control," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 1, pp. 152–163, 2005.

Research Article

Adaptive Fuzzy Tracking Control for a Permanent Magnet Synchronous Motor via Backstepping Approach

Jinpeng Yu,[1,2] Junwei Gao,[1,3] Yumei Ma,[1] and Haisheng Yu[1,2]

[1] *Institute of Complexity Science, Qingdao University, Qingdao 266071, China*
[2] *Shan Dong Province Key Laboratory of Industrial Control Technique, Qingdao University, Qingdao 266071, China*
[3] *State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China*

Correspondence should be addressed to Jinpeng Yu, yjp1109@hotmail.com

The speed tracking control problem of permanent magnet synchronous motors with parameter uncertainties and load torque disturbance is addressed. Fuzzy logic systems are used to approximate nonlinearities, and an adaptive backstepping technique is employed to construct controllers. The proposed controller guarantees the tracking error convergence to a small neighborhood of the origin and achieves the good tracking performance. Simulation results clearly show that the proposed control scheme can track the position reference signal generated by a reference model successfully under parameter uncertainties and load torque disturbance without singularity and overparameterization.

1. Introduction

Permanent magnet synchronous motors (PMSMs) are of great interest for industrial applications requiring dynamic performance due to their high speed, high efficiency, high power density, and large torque to inertia ratio. Then it is still a challenging problem to control PMSM to get the perfect dynamic performance because the motor dynamic model of PMSM is nonlinear and multivariable, the model parameters such as the stator resistance and the friction coefficient are also not be exactly known. The control of PMSM drivers has recently received wide attention and become an active research area. Some advanced control techniques, such as sliding mode control [1, 2], feedback linearization control [3], adaptive control [2, 4], backstepping principles [5–7], and Fuzzy logic control [8–10], are used to the problems of speed or position control of PMSMs.

Table 1: The denotation of the PMSM parameters.

Parameter	Denotation	Parameter	Denotation
ω	The rotor angular velocity	R_s	The stator resistance
i_d	The d axis current	i_q	The q axis current
L_d	The d axis stator inductor	L_q	The q axis stator inductor
u_d	The d axis voltage	u_q	The q axis voltage
n_p	The pole pair	J	The rotor moment of inertia
T	The electromagnetism torque	T_L	The load torque
B	The viscous friction coefficient	Φ	The magnet flux linkage of inertia

Backstepping is a newly developed technique to control the nonlinear systems with parameter uncertainty, particularly those systems in which the uncertainty does not satisfy matching conditions. Though backstepping is successfully applied to the control of PMSM drivers recently, it usually makes the designed controllers' structure to be very complex.

Fuzzy logic control (FLC) has been found one of the most popular and conventional tools in functional approximations. An FLC [11, 12] has strong ability of handling uncertain information and can be easily used in the control of systems which is ill-defined or too complex to have a mathematical model. It provides an effective way to design control system that is one of important applications in the area of control engineering.

In this paper, an adaptive fuzzy control approach is proposed for speed tracking control of PMSM drive system via the backstepping technique. During the controller design process, fuzzy logic systems are employed to approximate the nonlinearities, the adaptive fuzzy controllers are constructed via backstepping. The designed fuzzy controller can track the reference signal quite well even the existence of the parameter uncertainties and load torque disturbance. Compared with the existing controller design schemes via backstepping, the proposed method is very simple and the proposed controller has a simple structure.

2. Mathematical Model of the PMSM Drive System and Preliminaries

In this section, some preparatory knowledge of a PMSM will be introduced. The following assumptions are made in the derivation of the mathematical model of a PMSM [13].

Assumption 2.1. Saturation and iron losses are neglected although it can be taken into account by parameter changes.

Assumption 2.2. The back emf is sinusoidal.

The model of a PMSM can be described in the well known $(d - q)$ frame through the Park transformation as follows. The stator d, q equations in the rotor frame are expressed as follows [14]:

$$J\frac{d\omega}{dt} = T - T_L - B\omega = \frac{3}{2}n_p\left[(L_d - L_q)i_d i_q + \Phi i_q\right] - B\omega - T_L,$$

$$L_d\frac{di_d}{dt} = -R_s i_d + n_p\omega L_q i_q + u_d, \tag{2.1}$$

$$L_q\frac{di_q}{dt} = -R_s i_q - n_p\omega L_d i_d - n_p\omega\Phi + u_q.$$

The denotation of the PMSM parameters is shown in Table 1.

To simplify the previous method mode, the following notations are introduced:

$$x_1 = \omega, \qquad x_2 = i_q, \qquad x_3 = i_d,$$

$$a_1 = \frac{3n_p\Phi}{2}, \qquad a_2 = \frac{3n_p(L_d - L_q)}{2},$$

$$b_1 = -\frac{R_s}{L_q}, \qquad b_2 = -\frac{n_p L_d}{L_q}, \qquad b_3 = -\frac{n_p\Phi}{L_q}, \qquad b_4 = \frac{1}{L_q},$$

$$c_1 = -\frac{R_s}{L_d}, \qquad c_2 = \frac{n_p L_q}{L_d}, \qquad c_3 = \frac{1}{L_d}.$$

(2.2)

By using these notations, the dynamic model of a PMSM motor can be described by the following differential equations:

$$\dot{x}_1 = \frac{a_1}{J}x_2 + \frac{a_2}{J}x_2 x_3 - \frac{B}{J}x_1 - \frac{T_L}{J},$$

$$\dot{x}_2 = b_1 x_2 + b_2 x_1 x_3 + b_3 x_1 + b_4 u_q,$$

$$\dot{x}_3 = c_1 x_3 + c_2 x_1 x_2 + c_3 u_d.$$

(2.3)

The control objective is to design an adaptive fuzzy controller such that the state variable x_1 tracks the given reference signal x_d and all signals of the resulting closed-loop system are uniformly ultimately bounded. In this paper, we adopt the singleton fuzzifier, product inference, and the center-defuzzifier to deduce the following fuzzy rules:

$$R_i : \text{IF } x_1 \text{ is } F_1^i \text{ and} \ldots \text{and } x_n \text{ is } F_n^i \text{ THEN } y \text{ is } B^i \quad (i = 1, 2, \ldots, N),$$

(2.4)

where $x = [x_1, \ldots, x_n]^T \in R^n$, and $y \in R$ are the input and output of the fuzzy system, respectively, F_i^j and B^i are fuzzy sets in R. The fuzzy inference engine performs a mapping from fuzzy sets in R^n to fuzzy set in R based on the IF-THEN rules in the fuzzy rule base and the compositional rule of inference. The fuzzifier maps a crisp point $x = [x_1, \ldots, x_n]^T \in R^n$ into a fuzzy set A_x in R. The defuzzifier maps a fuzzy set in R to a crisp point in R. Since the strategy of singleton fuzzification, center-average defuzzification, and product inference is used, the output of the fuzzy system can be formulated as

$$y(x) = \frac{\sum_{j=1}^{N} W_j \prod_{i=1}^{n} \mu_{F_i^j}(x_i)}{\sum_{j=1}^{N} \left[\prod_{i=1}^{n} \mu_{F_i^j}(x_i) \right]},$$

(2.5)

where W_j is the point at which fuzzy membership function $\mu_{B^j}(W_j)$ achieves its maximum value, and it is assumed that $\mu_{B^j}(W_j) = 1$. Let $p_j(x) = \prod_{i=1}^{n} \mu_{F_i^j}(x_i) / \sum_{j=1}^{N}[\prod_{i=1}^{n} \mu_{F_i^j}(x_i)]$, $S(x) = [p_1(x), p_2(x), \ldots, p_N(x)]^T$ and $W = [W_1, \ldots, W_N]^T$, then the fuzzy logic system above can be

rewritten as

$$y(x) = W^T S(x). \tag{2.6}$$

If all memberships are taken as Gussian functions, then the following lemma holds.

Lemma 2.3 (see [15]). *Let $f(x)$ be a continuous function defined on a compact set Ω. Then for any scalar $\varepsilon > 0$, there exists a fuzzy logic system in the form (2.6) such that*

$$\sup_{x \in \Omega} |f(x) - y(x)| \leq \varepsilon. \tag{2.7}$$

3. Adaptive Fuzzy Controller with the Backstepping Technique

For the system (2.3), the backstepping design procedure contains 3 steps. At each design step, a virtual control function α_i ($i = 1, 2$) will be constructed by using an appropriate Lyapunov function V. At the last step, a real controller is constructed to control the system. In the following, we will give the procedure of the backstepping design.

Step 1. For the reference signal x_d, define the tracking error variable as $z_1 = x_1 - x_d$. From the first subsystem of (2.3), the error dynamic system is given by $\dot{z}_1 = (a_1/J)x_2 + (a_2/J)x_2x_3 - (B/J)x_1 - T_L/J - \dot{x}_d$.

Choose Lyapunov function candidate as $V_1 = (J/2)z_1^2$, then the time derivative of V_1 is given by

$$\dot{V}_1 = Jz_1\dot{z}_1 = z_1(a_1x_2 + a_2x_2x_3 - Bx_1 - T_L - J\dot{x}_d). \tag{3.1}$$

As the parameters B, T_L, and J are unknown, they cannot be used to construct the control signal. Thus, let $\widehat{B}, \widehat{T}_L$, and $\widehat{J}$ be their estimations of B, T_L, and J, respectively. The corresponding adaptation laws will be determined later. Now, construct the virtual control law α_1 as

$$\alpha_1(Z_1) = \frac{1}{a_1}\left(-k_1z_1 + \widehat{B}x_1 + \widehat{T}_L + \widehat{J}\dot{x}_d\right), \tag{3.2}$$

where $k_1 > 0$ is a design parameter and $Z_1 = [x_1, x_d, \dot{x}_d, \widehat{B}, \widehat{T}_L, \widehat{J}]^T$. Defining $z_2 = x_2 - \alpha_1$ and substituting (3.2) into (3.1) yield

$$\dot{V}_1 = -k_1z_1^2 + a_1z_1z_2 + a_2z_1x_2x_3 + z_1\left(\widehat{B} - B\right)x_1 + z_1\left(\widehat{T}_L - T_L\right) + z_1\left(\widehat{J} - J\right)\dot{x}_d. \tag{3.3}$$

Step 2. Differentiating z_2 and using the second subsystem of (2.3) give

$$\dot{z}_2 = \dot{x}_2 - \dot{\alpha}_1 = b_1x_2 + b_2x_1x_3 + b_3x_1 + b_4u_q - \dot{\alpha}_1. \tag{3.4}$$

Now, choose the Lyapunov function candidate as $V_2 = V_1 + (1/2)z_2^2$. Obviously, the time derivative of V_2 is given by

$$
\begin{aligned}
\dot{V}_2 &= \dot{V}_1 + z_2 \dot{z}_2 \\
&= -k_1 z_1^2 + a_2 z_1 x_2 x_3 + z_1 (\widehat{B} - B) x_1 + z_1 (\widehat{T}_L - T_L) + z_1 (\widehat{J} - J) \dot{x}_d + z_2 (f_2 + b_4 u_q),
\end{aligned}
\tag{3.5}
$$

where

$$
\begin{aligned}
\dot{\alpha}_1 &= \frac{\partial \alpha_1}{\partial x_1} \dot{x}_1 + \sum_{i=0}^{1} \frac{\partial \alpha_1}{\partial x_d^{(i)}} x_d^{(i+1)} + \frac{\partial \alpha_1}{\partial \widehat{B}} \dot{\widehat{B}} + \frac{\partial \alpha_1}{\partial \widehat{T}_L} \dot{\widehat{T}}_L + \frac{\partial \alpha_1}{\partial \widehat{J}} \dot{\widehat{J}} \\
&= \frac{\partial \alpha_2}{\partial x_1} \left(\frac{a_1}{J} x_2 + \frac{a_2}{J} x_2 x_3 - \frac{B}{J} x_1 - \frac{T_L}{J} \right) + \sum_{i=0}^{1} \frac{\partial \alpha_1}{\partial x_d^{(i)}} x_d^{(i+1)} + \frac{\partial \alpha_1}{\partial \widehat{B}} \dot{\widehat{B}} + \frac{\partial \alpha_1}{\partial \widehat{T}_L} \dot{\widehat{T}}_L + \frac{\partial \alpha_1}{\partial \widehat{J}} \dot{\widehat{J}},
\end{aligned}
\tag{3.6}
$$

$$
f_2(Z_2) = a_1 z_1 + b_1 x_2 + b_2 x_1 x_3 + b_3 x_1 - \dot{\alpha}_1,
$$

$$
Z_2 = \left[x_1, x_2, x_3, x_d, \dot{x}_d, \ddot{x}_d, \widehat{B}, \widehat{T}_L, \widehat{J} \right]^T.
$$

Apparently, there are two nonlinear terms in (3.5), that is, $a_2 z_1 x_2 x_3$ and f_2, therewithal, f_2 contains the derivative of $\dot{\alpha}_1$. This will make the classical adaptive backstepping design become very complex and troubled, and the designed control law u_q will have the complex structure. To avoid this trouble in design procedure and simplify the control signal structure, we will employ the fuzzy logic system to approximate the nonlinear function f_2. As shown later, the design procedure of u_q becomes simple and u_q has the simple structure. According to Lemma 2.3, for any given $\varepsilon_2 > 0$, there exists a fuzzy logic system $W_2^T S_2(Z_2)$ such that

$$
f_2(Z_2) = W_2^T S_2(Z_2) + \delta_2(Z_2),
\tag{3.7}
$$

with $\delta_2(Z)$ being the approximation error and satisfying $|\delta_2| \leq \varepsilon_2$. Consequently, a simple method computing produces the following inequality:

$$
z_2 f_2 = z_2 \left(W_2^T S_2 + \delta_2 \right) \leq \frac{1}{2 l_2^2} z_2^2 \|W_2\|^2 S_2^2 + \frac{1}{2} l_2^2 + \frac{1}{2} z_2^2 + \frac{1}{2} \varepsilon_2^2.
\tag{3.8}
$$

It follows immediately from substituting (3.8) into (3.5) that

$$
\begin{aligned}
\dot{V}_2 &\leq -\sum_{i=1}^{2} k_i z_i^2 + a_2 z_1 x_2 x_3 + z_1 (\widehat{B} - B) x_1 + z_1 (\widehat{T}_L - T_L) \\
&\quad + z_1 (\widehat{J} - J) \dot{x}_d + \frac{1}{2 l_2^2} z_2^2 \|W_2\|^2 S_2^2 + \frac{1}{2} l_2^2 + \frac{1}{2} z_2^2 + \frac{1}{2} \varepsilon_2^2 + z_2 b_4 u_q.
\end{aligned}
\tag{3.9}
$$

The control input u_q is designed as

$$u_q = \frac{1}{b_4}\left(-k_2 z_2 - \frac{1}{2}z_2 - \frac{1}{2l_2^2}z_2\widehat{\theta}S_2^2\right), \tag{3.10}$$

where $\widehat{\theta}$ is the estimation of the unknown constant θ which will be specified later. Using equality (3.10), the derivative of V_2 becomes as

$$\dot{V}_2 \leq -\sum_{i=1}^{2}k_i z_i^2 + a_2 z_1 x_2 x_3 + z_1\left(\widehat{B} - B\right)x_1 + z_1\left(\widehat{T}_L - T_L\right)$$
$$+ z_1\left(\widehat{J} - J\right)\dot{x}_d + \frac{1}{2l_2^2}z_2^2\left(\|W_2\|^2 - \widehat{\theta}\right)S_2^2 + \frac{1}{2}l_2^2 + \frac{1}{2}\varepsilon_2^2. \tag{3.11}$$

Step 3. At this step, we will construct the control law u_d. To this end, define $z_3 = x_3$ and choose the following Lyapunov function candidate as $V_3 = V_2 + (1/2)z_3^2$. Then the derivative of V_3 is given by

$$\dot{V}_3 = \dot{V}_2 + z_3\dot{z}_3 \leq -\sum_{i=1}^{2}k_i z_i^2 + a_2 z_1 x_2 x_3 + z_1\left(\widehat{B} - B\right)x_1 + z_1\left(\widehat{T}_L - T_L\right)$$
$$+ z_1\left(\widehat{J} - J\right)\dot{x}_d + \frac{1}{2l_2^2}z_2^2\left(\|W_2\|^2 - \widehat{\theta}\right)S_2^2 + \frac{1}{2}l_2^2 + \frac{1}{2}\varepsilon_2^2 + z_3(f_3(Z_3) + c_3 u_d), \tag{3.12}$$

where $f_3(Z_3) = a_2 z_1 x_2 + c_1 x_3 + c_2 x_1 x_2$ and $Z_3 = [x_1, x_2, x_3, x_d]^T$. Similarly, by Lemma 2.3 the fuzzy logic system $W_3^T S_3(Z_3)$ is utilized to approximate the nonlinear function f_3 such that for given $\varepsilon_3 > 0$,

$$z_3 f_3 \leq \frac{1}{2l_3^2}z_3^2\|W_3\|^2 S_3^2 + \frac{1}{2}l_3^2 + \frac{1}{2}z_3^2 + \frac{1}{2}\varepsilon_3^2. \tag{3.13}$$

Substituting (3.13) into (3.12) gives

$$\dot{V}_3 = \dot{V}_2 + z_3\dot{z}_3 \leq -\sum_{i=1}^{2}k_i z_i^2 + \frac{1}{2l_2^2}z_2^2\left(\|W_2\|^2 - \widehat{\theta}\right)S_2^2 + \sum_{i=2}^{3}\frac{1}{2}\left(l_i^2 + \varepsilon_i^2\right)$$
$$+ z_1\left(\widehat{B} - B\right)x_1 + z_1\left(\widehat{T}_L - T_L\right) + z_1\left(\widehat{J} - J\right)\dot{x}_d + \frac{1}{2l_3^2}z_3^2\|W_3\|^2 S_3^2 + \frac{1}{2}z_3^2 + c_3 z_3 u_d. \tag{3.14}$$

Now design u_d as

$$u_d = \frac{-1}{c_3}\left(k_3 z_3 + \frac{1}{2}z_3 + \frac{1}{2l_3^2}z_3\widehat{\theta}S_3^2\right). \tag{3.15}$$

Then, define $\theta = \max\{\|W_2\|^2, \|W_3\|^2\}$. Then, combining (3.14) with (3.15) results in

$$
\begin{aligned}
\dot{V}_4 &\leq -\sum_{i=1}^{3} k_i z_i^2 + \sum_{i=2}^{3} \frac{1}{2}\left(l_i^2 + \varepsilon_i^2\right) + z_1\left(\widehat{B} - B\right)x_1 + z_1\left(\widehat{T}_L - T_L\right) \\
&\quad + z_1\left(\widehat{J} - J\right)\dot{x}_d + \sum_{i=2}^{3} \frac{1}{2l_i^2}z_i^2\left(\|W_i\|^2 - \widehat{\theta}\right)S_i^T(Z_i)S_i(Z_i) \\
&\leq -\sum_{i=1}^{3} k_i z_i^2 + \sum_{i=2}^{3} \frac{1}{2}\left(l_i^2 + \varepsilon_i^2\right) + z_1\left(\widehat{B} - B\right)x_1 + z_1\left(\widehat{T}_L - T_L\right) \\
&\quad + z_1\left(\widehat{J} - J\right)\dot{x}_d + \sum_{i=2}^{3} \frac{1}{2l_i^2}z_i^2 S_i^T(Z_i)S_i(Z_i)\left(\theta - \widehat{\theta}\right).
\end{aligned}
\tag{3.16}
$$

At the present stage, to estimate the unknown constants B, T_L, J, and θ, define the adaptive variables as follows:

$$
\begin{aligned}
\widetilde{T}_L &= \widehat{T}_L - T_L, \\
\widetilde{B} &= \widehat{B} - B, \\
\widetilde{J} &= \widehat{J} - J, \\
\widetilde{\theta} &= \widehat{\theta} - \theta.
\end{aligned}
\tag{3.17}
$$

In order to determine the corresponding adaptation laws, choose the following Lyapunov function candidate:

$$
V = V_3 + \frac{1}{2r_1}\widetilde{T}_L^2 + \frac{1}{2r_2}\widetilde{B}^2 + \frac{1}{2r_3}\widetilde{J}^2 + \frac{1}{2r_4}\widetilde{\theta}^2,
\tag{3.18}
$$

where r_i, $i = 1, 2, 3, 4$ are positive constant. By differentiating V and taking (3.16)–(3.18) into account, one has

$$
\begin{aligned}
\dot{V} &\leq -\sum_{i=1}^{3} k_i z_i^2 + \sum_{i=2}^{3} \frac{1}{2}\left(l_i^2 + \varepsilon_i^2\right) + z_1\widetilde{B}x_1 + z_1\widetilde{T}_L \\
&\quad + z_1\widetilde{J}\dot{x}_d - \sum_{i=2}^{3} \frac{1}{2l_i^2}z_i^2\widetilde{\theta}S_i^T(Z_i)S_i(Z_i) + \frac{1}{r_1}\widetilde{T}_L\dot{\widehat{T}}_L + \frac{1}{r_2}\widetilde{B}\dot{\widehat{B}} + \frac{1}{r_3}\widetilde{J}\dot{\widehat{J}} + \frac{1}{r_4}\widetilde{\theta}\dot{\widehat{\theta}} \\
&= -\sum_{i=1}^{3} k_i z_i^2 + \sum_{i=2}^{3} \frac{1}{2}\left(l_i^2 + \varepsilon_i^2\right) + \frac{1}{r_1}\widetilde{T}_L\left(r_1 z_1 + \dot{\widehat{T}}_L\right) \\
&\quad + \frac{1}{r_2}\widetilde{B}\left(r_2 z_1 x_1 + \dot{\widehat{B}}\right) + \frac{1}{r_3}\widetilde{J}\left(r_3 z_1 \dot{x}_d + \dot{\widehat{J}}\right) + \frac{1}{r_4}\widetilde{\theta}\left[-\sum_{i=2}^{3} \frac{r_4}{2l_i^2}z_i^2 S_i^T(Z_i)S_i(Z_i) + \dot{\widehat{\theta}}\right].
\end{aligned}
\tag{3.19}
$$

According to (3.19), the corresponding adaptive laws are chosen as follows:

$$\dot{\hat{T}}_L = -r_1 z_1 - m_1 \tilde{T}_L,$$

$$\dot{\hat{B}} = -r_2 z_1 x_1 - m_2 \hat{B},$$

$$\dot{\hat{J}} = -r_3 z_1 \dot{x}_d - m_3 \hat{J},$$

$$\dot{\hat{\theta}} = \sum_{i=2}^{3} \frac{r_4}{2 l_i^2} z_i^2 S_i^T (Z_i) S_i(Z_i) - m_4 \hat{\theta},$$

(3.20)

where m_i for $i = 1, 2, 3, 4$ and l_i for $i = 2, 3$ are positive constant.

4. Stability Analysis

In this section, the stability analysis of the resulting closed-loop system will be addressed. Substituting (3.20) into (3.19) yields

$$\dot{V} \le -\sum_{i=1}^{3} k_i z_i^2 + \sum_{i=2}^{3} \frac{1}{2} \left(l_i^2 + \varepsilon_i^2 \right) - \frac{m_1}{r_1} \tilde{T}_L \hat{T} - \frac{m_2}{r_2} \tilde{B} \hat{B} - \frac{m_3}{r_3} \tilde{J} \hat{J} - \frac{m_4}{r_4} \tilde{\theta} \hat{\theta}.$$

(4.1)

For the term $-\tilde{T}_L \hat{T}$, one has

$$-\tilde{T}_L \hat{T}_L = -\tilde{T}_L \left(\tilde{T}_L + T_L \right) \le -\frac{1}{2} \tilde{T}_L^2 + \frac{1}{2} T_L^2.$$

(4.2)

Similarly, we have

$$-\tilde{B} \hat{B} \le -\frac{1}{2} \tilde{B}^2 + \frac{1}{2} B^2,$$

$$-\tilde{J} \hat{J} \le -\frac{1}{2} \tilde{J}^2 + \frac{1}{2} J^2,$$

$$-\tilde{\theta} \hat{\theta} \le -\frac{1}{2} \tilde{\theta}^2 + \frac{1}{2} \theta^2.$$

(4.3)

Consequently, by using these inequalities, (4.1) can be rewritten in the following form:

$$\dot{V} \le -\sum_{i=1}^{3} k_i z_i^2 - \frac{m_1}{2 r_1} \tilde{T}_L^2 - \frac{m_2}{2 r_2} \tilde{B}^2 - \frac{m_3}{2 r_3} \tilde{J}^2 - \frac{m_4}{2 r_4} \tilde{\theta}^2$$

$$+ \sum_{i=2}^{3} \frac{1}{2} \left(l_i^2 + \varepsilon_i^2 \right) + \frac{m_1}{2 r_1} T_L^2 + \frac{m_2}{2 r_2} B^2 + \frac{m_3}{2 r_3} J^2 + \frac{m_4}{2 r_4} \theta^2$$

$$\le -a_0 V + b_0,$$

(4.4)

where $a_0 = \min\{2k_1/J, 2k_2, 2k_3, m_1, m_2, m_3, m_4\}$ and $b_0 = \sum_{i=2}^{3}(1/2)(l_i^2 + \varepsilon_i^2) + (m_1/2r_1)T_L^2 + (m_2/2r_2)B^2 + (m_3/2r_3)J^2 + (m_4/2r_4)\theta^2$. Furthermore, (4.4) implies that

$$V(t) \leq \left(V(t_0) - \frac{b_0}{a_0}\right)e^{-a_0(t-t_0)} + \frac{b_0}{a_0} \leq V(t_0) + \frac{b_0}{a_0}, \quad \forall t \geq t_0. \tag{4.5}$$

As a result, all $z_i (i = 1, 2, 3)$, $\tilde{T}_L, \tilde{B}, \tilde{J}$ and $\tilde{\theta}$ belong to the compact set

$$\Omega = \left\{\left(z_i, \tilde{T}_L, \tilde{B}, \tilde{J}, \tilde{\theta}\right) \mid V \leq V(t_0) + \frac{b_0}{a_0}, \ \forall t \geq t_0\right\}. \tag{4.6}$$

Namely, all the signals in the closed-loop system are bounded. Especially, from (4.5) we have

$$\lim_{t \to \infty} z_1^2 \leq \frac{2b_0}{a_0}. \tag{4.7}$$

From the definitions of a_0 and b_0, it is clear that to get a small tracking error we can take r_i large and l_i and ε_i small enough after giving the parameters k_i and m_i.

5. Simulation

To illustrate the effectiveness of the proposed results, the simulation will be done for the PMSM motor with the parameters:

$$J = 0.00379\,\text{Kgm}^2, \qquad R_s = 0.68\,\Omega, \qquad L_d = 0.00315\,\text{H},$$
$$L_q = 0.00285\,\text{H}, \qquad B = 0.001158\,\text{Nm}/(\text{rad}/\text{s}), \qquad \Phi = 0.1245\,\text{Wb}, \qquad n_p = 3. \tag{5.1}$$

Then, the proposed adaptive fuzzy controllers are used to control this PMSM motor. Given the reference signal is $x_d = 30$ and the control parameters are chosen as follows:

$$k_1 = 2.5, \qquad k_2 = k_3 = 50, \qquad r_1 = r_2 = r_3 = r_4 = 2.5,$$
$$m_1 = m_2 = m_3 = m_4 = 0.0005, \qquad l_2 = l_3 = 5. \tag{5.2}$$
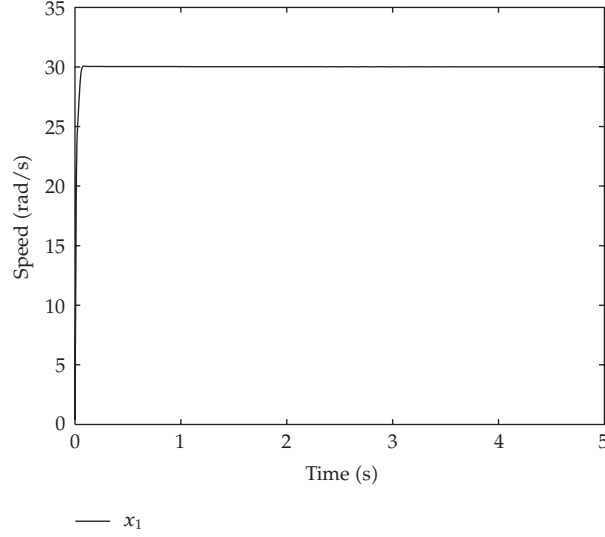
Figure 1: The curve of the rotor speed x_1.

The fuzzy membership functions are chosen as

$$\mu_{F_i^1} = \exp\left[\frac{-(x+5)^2}{2}\right], \qquad \mu_{F_i^2} = \exp\left[\frac{-(x+4)^2}{2}\right],$$

$$\mu_{F_i^3} = \exp\left[\frac{-(x+3)^2}{2}\right], \qquad \mu_{F_i^4} = \exp\left[\frac{-(x+2)^2}{2}\right],$$

$$\mu_{F_i^5} = \exp\left[\frac{-(x+1)^2}{2}\right], \qquad \mu_{F_i^6} = \exp\left[\frac{-(x-0)^2}{2}\right],$$

$$\mu_{F_i^7} = \exp\left[\frac{-(x-1)^2}{2}\right], \qquad \mu_{F_i^8} = \exp\left[\frac{-(x-2)^2}{2}\right], \qquad (5.3)$$

$$\mu_{F_i^9} = \exp\left[\frac{-(x-3)^2}{2}\right], \qquad \mu_{F_i^{10}} = \exp\left[\frac{-(x-4)^2}{2}\right],$$

$$\mu_{F_i^{11}} = \exp\left[\frac{-(x-5)^2}{2}\right].$$

The simulation is carried out under the zero initial condition for two cases. In the first case, $T_L = 1.5$ and in the second case,

$$T_L = \begin{cases} 1.5, & 0 \le t \le 1, \\ 3, & t \ge 1. \end{cases} \qquad (5.4)$$
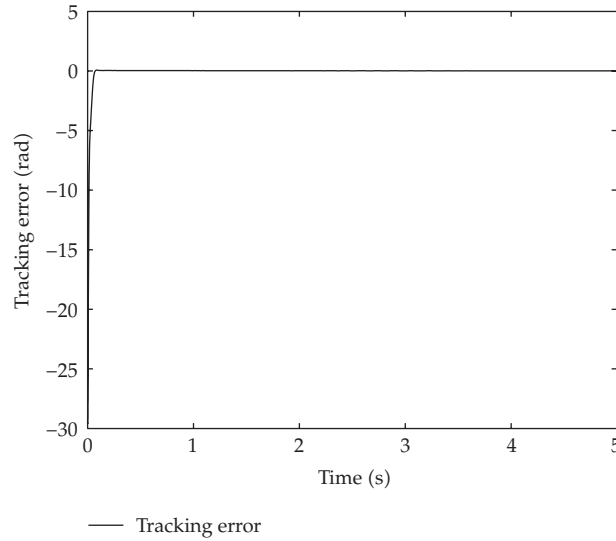
Figure 2: The curve of the speed tracking error.
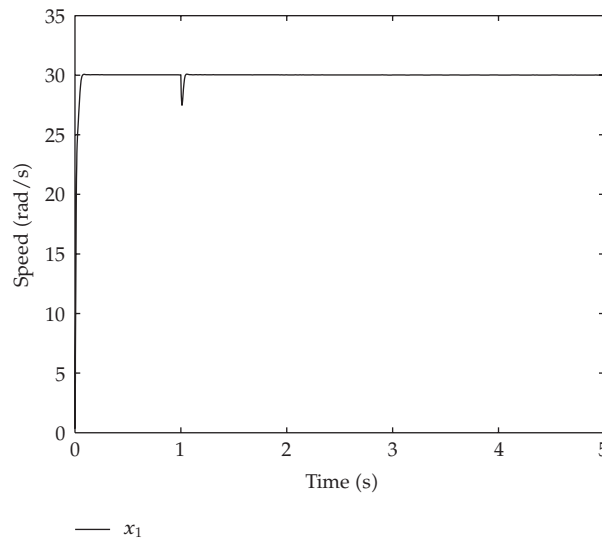
Figure 3: The curve of the rotor position x_1.

Figures 1 and 2 show the simulation results for case 1, and Figures 3 and 4 show the simulation results for the second case. From these figures, it is seen clearly that the tracking performance has been achieved very well. This means that the proposed controller can track the reference signal satisfactorily even under parameter uncertainties and load torque disturbance.

6. Conclusion

Based on adaptive fuzzy control approach and backstepping technique, an adaptive fuzzy control scheme is proposed to control a permanent magnet synchronous motor. The proposed
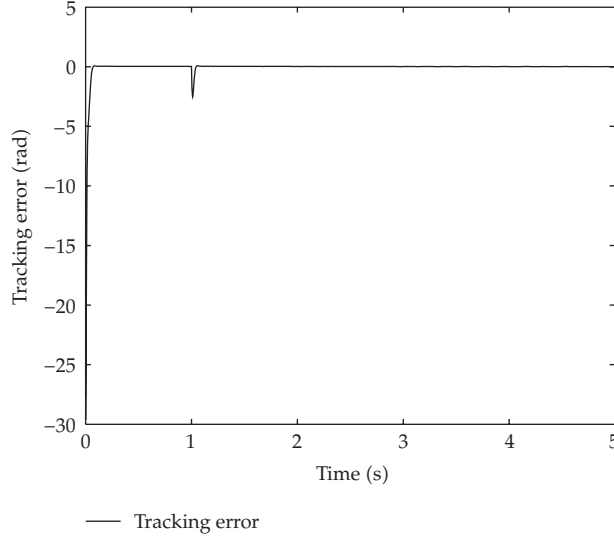
Figure 4: The curve of the speed tracking error.

controllers guarantee that the tracking error converges to a small neighborhood of the origin and all the closed-loop signals are bounded. Simulation results illustrate the effectiveness of the presented method.

Appendix

The Controller of PMSM Based on Conventional Backstepping

The controller of PMSM based on conventional backstepping technique is briefly reviewed here. The detailed design procedure is in [7]. The controllers based on conventional backstepping are shown as follows:

$$
\begin{aligned}
u_q = \frac{1}{b_4}\Bigg(& -k_2 z_2 + b_4\Bigg(\frac{\dot{\widehat{J}}}{a_1}\Big(-k_1 z_1 + \widehat{F} x_1 + \widehat{\Gamma} + \dot{x}_d\Big) + \frac{\widehat{J}}{a_1}\Big(\dot{\widehat{F}} x_1 + \dot{\widehat{\Gamma}} + \ddot{x}_d + k_1 \dot{x}_d\Big) \\
& + \Big(\widehat{F} - k_1\Big)\Big(x_2 + \frac{a_2}{a_1} x_2 x_3\Big) - \frac{\widehat{J}}{a_1}\Big(\widehat{F} - k_1\Big)\Big(\widehat{F} x_1 + \widehat{\Gamma}\Big) - (b_1 x_2 + b_2 x_1 x_3 + b_3 x_1)\Bigg)\Bigg), \\
u_d = \frac{1}{c_3}\Bigg(& -k_3 z_3 - c_1 x_3 - c_2 x_1 x_2 - \frac{a_2}{\widehat{J}} c_3 z_1 x_2\Bigg).
\end{aligned}
$$

$$(A.1)$$

Comparing the presented method with the above controller designed via conventional back-stepping, it can be seen that the proposed controllers based on adaptive fuzzy backstepping have a very simple structure. And the simulation results illustrate the effectiveness of the presented method in this paper.

Acknowledgments

References

[1] F.-J. Lin and S.-L. Chiu, "Adaptive fuzzy sliding-mode control for PM synchronous servo motor drives," *IEE Proceedings: Control Theory and Applications*, vol. 145, no. 1, pp. 63–72, 1998.

[2] S. C. Tong and H. H. Li, "Fuzzy adaptive sliding model control for mimo nonlinear systems," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 3, pp. 354–360, 2003.

[3] A. Isidori, *Nonlinear Control Systems*, Communications and Control Engineering Series, Springer, Berlin, Germany, 3rd edition, 1995.

[4] H. Lee and M. Tomizuka, "Robust adaptive control using a universal approximator for SISO nonlinear systems," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 1, pp. 95–106, 2000.

[5] M. Krstic, I. Kanellakopoulos, and P. Kokotovic, *Nonlinear and Adaptive Control Design*, John Wiley & Sons, New York, NY, USA, 1995.

[6] X. Liu, G. Gu, and K. Zhou, "Robust stabilization of MIMO nonlinear systems by backstepping," *Automatica*, vol. 35, no. 5, pp. 987–992, 1999.

[7] J.-H. Hu and J.-B. Zou, "Adaptive backstepping control of permanent magnet synchronous motors with parameter uncertainties," *Control and Decision*, vol. 21, no. 11, pp. 1264–1269, 2006.

[8] L. A. Zadeh, "Fuzzy sets," *Information and Computation*, vol. 8, pp. 338–353, 1965.

[9] C. Elmas, O. Ustun, and H. H. Sayan, "A neuro-fuzzy controller for speed control of a permanent magnet synchronous motor drive," *Expert Systems with Applications*, vol. 34, no. 1, pp. 657–664, 2008.

[10] S. Tong and H.-X. Li, "Direct adaptive fuzzy output tracking control of nonlinear systems," *Fuzzy Sets and Systems*, vol. 128, no. 1, pp. 107–115, 2002.

[11] C. C. Lee, "Fuzzy logic in control systems: fuzzy logic controller—I," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 20, no. 2, pp. 404–418, 1990.

[12] C. C. Lee, "Fuzzy logic in control systems: fuzzy logic controller—II," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 20, no. 2, pp. 419–435, 1990.

[13] P. Pillay and R. Krishnan, "Modeling of permanent magnet motor drives," *IEEE Transactions on Industrial Electronics*, vol. 35, no. 4, pp. 537–541, 1988.

[14] W. Leonhard, *Control of Electrical Drives*, Springer, Berlin Germany, 1985.

[15] L.-X. Wang and J. M. Mendel, "Fuzzy basis functions, universal approximation, and orthogonal least-squares learning," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 807–814, 1992.

Research Article

Design for Motor Controller in Hybrid Electric Vehicle Based on Vector Frequency Conversion Technology

Jing Lian, Yafu Zhou, Teng Ma, and Wei Wang

School of Automotive Engineering, Faculty of Vehicle Engineering and Mechanics, State Key Laboratory of Structural Analysis for Industrial Equipment, Dalian University of Technology, Dalian 116024, China

Correspondence should be addressed to Yafu Zhou, dlzyf62@126.com

Motor and its control technology are one of the main components of Hybrid Electric Vehicle (HEV). To meet HEV's fast torque response, vector control algorithm based on rotor flux-oriented and simulation model is concerned and modular designs for controller's hardware and software are presented in the paper in order to build a platform to achieve the vector control of asynchronous induction motor. Analyze the controller's electromagnetic compatibility, introduce the corresponding antijamming measures to assure the normal operation of the electromagnetic sensitive devices such as CAN bus; experiment proves that the measure is practical and feasible. On the basis of the control logic correct, such as improving CAN bus communication reliability, assuring power-on sequence and fault treatment, carry on the motor bench experiment, test its static properties, and adjust the controller parameters. The experimental results show that the designed driving system has the performance of low speed and high torque, a wide range of variable speed and high comprehensive efficiency.

1. Introduction

Hybrid Electric Vehicle (HEV) is the vehicle with two or more power supplies (one of these is electrical energy) [1, 2], such as unit of internal combustion engine and secondary cell or unit of fuel cell and secondary cell. At present, developing HEV research is one of the most meaningful ways to solve the problems of pollution and energy.

Drive motor is one of the HEV's core components. In accordance with different vehicles' operating environment of level road, ramp, acceleration-deceleration, start-stop, and so on, the motor operating mode switches frequently among fractional load, heavy load, and excess load, hence improving the integrated efficiency of the electrical motor is a key

issue for the development of motor and its controller, which directly affects vehicle's power performance, fuel economy, and emission. The basic performance requirements of HEV's motor drive system [3, 4] are high-performance and low-loss, high power density, low speed and high torque, a wide range of variable speed, strong overload capacity, good reliability, and so on. Based on the analysis of HEV's mainstream motor drive systems in the present market, this paper chooses the AC asynchronous induction motor which has low cost and high efficiency as a drive motor selection.

The AC asynchronous induction motor is the nonlinear time-varying system [5] of high-order, multivariable, and strong coupling, hence the control method based on motor's static mathematical model has been unable to meet the requirements of its dynamic performance. Currently, the control methods of AC asynchronous induction motor in HEV are mainly two kinds of direct torque control and vector control [6]. In the low speed, direct torque control is easy to produce torque fluctuation and no closed-loop current and is easy to produce over-current, which cannot meet the requirements of high stability torque when the vehicle is in the low-speed climbing state, at the same time there are the problems of motor temperature rise and high noise. Vector control method is able to achieve the AC asynchronous induction motor's decoupling control of magnetic flux and torque, which has good torque control characteristics, being analogous to DC motor, and can improve the efficiency of drive system and achieve maximum efficiency control. In addition, vector control can achieve a fixed switching frequency PWM modulation and reduce the harmonic content of the motor current, which, to a certain extent, reduce the motor temperature rise and noise.

Therefore, this paper introduces the vector control algorithm based on the rotor flux linkage oriented and conducts research to design the motor controller of Hybrid Electric Vehicle with high performance, high efficiency, and high reliability. First, introduce the basic principles of algorithm, build the modularized simulation platform, and then design the controller's software and hardware; meanwhile, analyze controller's electromagnetic compatibility and take appropriate antijamming measures to ensure the normal operation of the electromagnetic sensitive equipments such as CAN bus. Finally, carry out the motor bench experiment on the basis of the control logic correct, such as improving the CAN bus communication reliability, ensuring the power-on sequence, and fault treatment, and complete the performance experiment of the motor drive system.

2. Vector Control Based on Rotor Flux Linkage Oriented

In terms of AC asynchronous motor, the current, voltage, magnetic flux and electromagnetic torque are interrelated. The basic idea of vector control [7, 8] uses the mathematical coordinate transformation method to transform current i_A, i_B, i_C of AC three-phrase winding A, B, C to i_α, i_β of two-phrase static winding α, β, once again by mathematical coordinate transformation transform i_α, i_β to direct current i_M, i_T of two-phrase rotating winding M, T. In essence, through mathematical coordinate transformation transform stator current of AC asynchronous motor into two components, one is excitation component i_M which is used to generate rotating magnetic potential, and another is torque component i_T which is used to generate electromagnetic torque, as shown in Figure 1. Adjusting i_M can adjust the strength of the magnetic field, and adjusting i_T can adjust the size of torque in the constant magnetic field.
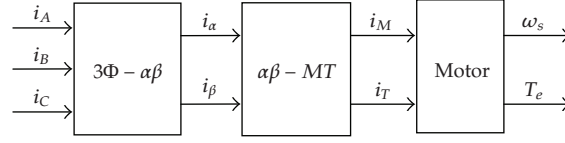
Figure 1: Basic idea of vector control in asynchronous motor.

2.1. Coordinate Transformation

(1) *Transform static three-phrase coordinate system A, B, C into static two-phrase coordinate system α-β.*

Take α axis to be coincident with A axis, the effective turn of each phrase in the three-phrase winding is N_3, the effective turn of each phrase in the two-phrase winding is N_2, $N_3/N_2 = \sqrt{2/3}$. In order to facilitate the inverse transform, add an imaginary zero axis current i_0, and $i_0 = K(i_A + i_B + i_C) = 0$, K is an undetermined constant. Take the principle that the total magnetic potential and total power are permanent before and after transformation, there is

$$
\begin{bmatrix} i_\alpha \\ i_\beta \\ i_0 \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} 1 & -\dfrac{1}{2} & -\dfrac{1}{2} \\ 0 & \sqrt{\dfrac{3}{4}} & -\sqrt{\dfrac{3}{4}} \\ \sqrt{\dfrac{1}{2}} & \sqrt{\dfrac{1}{2}} & \sqrt{\dfrac{1}{2}} \end{bmatrix} \begin{bmatrix} i_A \\ i_B \\ i_C \end{bmatrix}.
\tag{2.1}
$$

(2) *Transform rotating two-phrase coordinate system into static two-phrase coordinate system.*

Rotating two-phrase coordinate system are represented by d axis and q axis which are perpendicular to each other, the angle between d axis and α axis is θ, the rotating speed of d-q axis is ω, $\omega = d\theta/dt$, then there is

$$
\begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} i_d \\ i_q \end{bmatrix}.
\tag{2.2}
$$

2.2. The Motor Equations in Different Coordinate System

Define stator current i_A, i_B, i_C, rotor current i_a, i_b, i_c, stator flux linkage Ψ_A, Ψ_B, Ψ_C, rotor flux linkage Ψ_a, Ψ_b, Ψ_c. The motor equations in different coordinate system are shown as follows

(1) *Three-Phrase Static Coordinate System*

We have the equation of flux linkage

$$
\begin{bmatrix} \Psi_A \\ \Psi_B \\ \Psi_C \\ \Psi_a \\ \Psi_b \\ \Psi_c \end{bmatrix} = \begin{bmatrix} L_{AA} & L_{AB} & L_{AC} & L_{Aa} & L_{Ab} & L_{Ac} \\ L_{BA} & L_{BB} & L_{BC} & L_{Ba} & L_{Bb} & L_{Bc} \\ L_{CA} & L_{CB} & L_{CC} & L_{Ca} & L_{Cb} & L_{Cc} \\ L_{aA} & L_{aB} & L_{aC} & L_{aa} & L_{ab} & L_{ac} \\ L_{bA} & L_{bB} & L_{bC} & L_{ba} & L_{bb} & L_{bc} \\ L_{cA} & L_{cB} & L_{cC} & L_{ca} & L_{cb} & L_{cc} \end{bmatrix} \begin{bmatrix} i_A \\ i_B \\ i_C \\ i_a \\ i_b \\ i_c \end{bmatrix},
\tag{2.3}
$$

where L_{AA}, L_{BB}, L_{CC}, respectively, denote the self-inductance of stator's each phrase and $L_{AA} = L_{BB} = L_{CC} = L_{m1} + L_{e1}$, and where L_{aa}, L_{bb}, L_{cc}, respectively, denote the self-inductance of rotor's each phrase and $L_{aa} = L_{bb} = L_{cc} = L_{m2} + L_{e2}$. Because the stator turns are equivalent to rotor turns after conversion, $L_{m1} = L_{m2}$. We have mutual inductance:

$$
L_{AB} = L_{BA} = \cdots = L_{CB} = L_{m1} \cos 120^\circ = -0.5 L_{m1},
$$

$$
L_{ab} = L_{ba} = \cdots = L_{cb} = L_{m1} \cos 120^\circ = -0.5 L_{m1},
$$

$$
L_{Aa} = L_{aA} = \cdots = L_{cC} = L_{m1} \cos \theta,
\tag{2.4}
$$

$$
L_{Ab} = L_{bA} = \cdots = L_{aC} = L_{m1} \cos(\theta + 120^\circ),
$$

$$
L_{Ac} = L_{cA} = \cdots = L_{bC} = L_{m1} \cos(\theta - 120^\circ),
$$

where L_{e1} denotes stator's leakage inductance and L_{e2} denotes rotor's leakage inductance L_{m1} denotes mutual inductance between each stator and L_{m2} denotes mutual inductance between each rotor.

There is equation of voltage

$$
\begin{bmatrix} u_A \\ u_B \\ u_C \\ u_a \\ u_b \\ u_c \end{bmatrix} = \begin{bmatrix} R_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & R_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & R_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & R_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & R_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & R_2 \end{bmatrix} \begin{bmatrix} i_A \\ i_B \\ i_C \\ i_a \\ i_b \\ i_c \end{bmatrix} + p \begin{bmatrix} \Psi_A \\ \Psi_B \\ \Psi_C \\ \Psi_a \\ \Psi_b \\ \Psi_c \end{bmatrix},
\tag{2.5}
$$

where u denotes voltage, p denotes d/dt, R_1 and R_2, respectively, denote the resistance of stator winding and rotor winding.

There is equation of torque. On the condition that the current is constant and only mechanical displacement changes, electromagnetic torque is equivalent to the partial derivative of the magnetic field energy to mechanical angular displacement.

$$T_e = \frac{1}{2}n_p i^T \frac{\partial L}{\partial \theta} i = \frac{1}{2}n_p \left[i_s^T \frac{\partial L_{rs}}{\partial \theta} i_s + i_s^T \frac{\partial L_{sr}}{\partial \theta} i_r \right]$$

$$= -n_p L_m [(i_A i_a + i_B i_b + i_C i_c) \sin\theta + (i_A i_b + i_B i_c + i_C i_a) \sin(\theta + 120°)$$

$$+ (i_A i_c + i_B i_a + i_C i_b) \sin(\theta - 120°)], \tag{2.6}$$

where T_e denotes electromagnetic torque, θ denotes rotor space angular displacement which is shown by electrical angle, n_p denotes the number of pole-pairs, and L denotes inductance matrix of formula (2.3).

Equation of motion is

$$T_e - T_L = \frac{J}{n_p} \frac{d\omega}{\partial t}, \tag{2.7}$$

where T_L denotes load torque and J denotes moment of inertia of units.

(2) d-q Rotating Coordinate System

Given the mathematical model of the two-phrase motor, stator winding are d_1 and q_1 which is perpendicular to each other, and rotor winding is d_2 and q_2 which are also perpendicular to each other. In order to calculate conveniently, let d_1 and q_1 be, respectively, coincident with d_2 and q_2 which are, respectively, known as d and q axes. Let ω_1 be the synchronous angular velocity of the stator frequency, ω_{11} rotational speed of d axis relative to A axis, and ω_{12} rotational speed of d axis relative to a axis. The equation of flux linkage is

$$\begin{bmatrix} \Psi_{d1} \\ \Psi_{q1} \\ \Psi_{d2} \\ \Psi_{q2} \end{bmatrix} = \begin{bmatrix} L_s & 0 & L_m & 0 \\ 0 & L_s & 0 & L_m \\ L_m & 0 & L_r & 0 \\ 0 & L_m & 0 & L_r \end{bmatrix} \begin{bmatrix} i_{d1} \\ i_{q1} \\ i_{d2} \\ i_{q2} \end{bmatrix}, \tag{2.8}$$

where L_m denotes mutual inductance of stator winding and rotor winding and $L_m = (3/2)L_{m1}$, L_s denotes self-inductance of stator winding and $L_s = L_{e1} + (3/2)L_{m1}$, and L_r denotes self-inductance of rotor winding and $L_r = L_{e2} + (3/2)L_{m1}$.

Equation of voltage is

$$\begin{bmatrix} u_{d1} \\ u_{q1} \\ u_{d2} \\ u_{q2} \end{bmatrix} = \begin{bmatrix} R_1 + L_s p & -\omega_{11}L_s & L_m p & -\omega_{11}L_m \\ \omega_{11}L_s & R_1 + L_s p & \omega_{11}L_m & L_m p \\ L_m p & -\omega_{12}L_m & R_2 + L_r p & -\omega_{12}L_r \\ \omega_{12}L_m & L_m p & \omega_{12}L_r & R_2 + L_r p \end{bmatrix} \begin{bmatrix} i_{d1} \\ i_{q1} \\ i_{d2} \\ i_{q2} \end{bmatrix}. \tag{2.9}$$

Equation of torque is

$$T_e = n_p L_m (i_{q1} i_{d2} - i_{d1} i_{q2}).$$
(2.10)

Equation of motion is same with formula (2.7).

(3) *Equation of Asynchronous Motor in M-T Coordinate System*

In order to make the system decoupling, define that d axis is coincident with the direction of rotor flux linkage vector Ψ_2 which is entitled M axis. The axis being perpendicular to Ψ_2 is q axis which is entitled T axis. Assuming an equal number of turns of the winding, we can eliminate the number of turns in the magnetic potential and the magnetic potential F can be marked as vector i_1, that is, stator current and stator magnetic potential F_1 are only same with the direction of magnetic potential F.

Transform d, q in formula (2.8) into m, t. Due to M axis on Ψ_2 axis, $\Psi_{m2} = \Psi_2$, $\Psi_{t2} = 0$, the new equation of flux linkage is

$$\begin{bmatrix} \Psi_{m1} \\ \Psi_{t1} \\ \Psi_{m2} \\ \Psi_{t2} \end{bmatrix} = \begin{bmatrix} L_s & 0 & L_m & 0 \\ 0 & L_s & 0 & L_m \\ L_m & 0 & L_r & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} i_{m1} \\ i_{t1} \\ i_{m2} \\ i_{t2} \end{bmatrix}.$$
(2.11)

Given the characteristics that the terminal voltage of the cage asynchronous motor is zero, formula (2.9) can be simplified

$$\begin{bmatrix} u_{m1} \\ u_{t1} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} R_1 + L_s p & -\omega_1 L_s & L_m p & -\omega_1 L_m \\ \omega_1 L_s & R_1 + L_s p & \omega_1 L_m & L_m p \\ L_m p & 0 & R_2 + L_r p & 0 \\ \omega_s L_m & 0 & \omega_s L_r & R_2 + L_r p \end{bmatrix} \begin{bmatrix} i_{m1} \\ i_{t1} \\ i_{m2} \\ i_{t2} \end{bmatrix}.$$
(2.12)

The equation of torque is same with formula (2.10) and the equation of motion is same with formula (2.7).

2.3. Vector Control Algorithm and Simulation

By formula (2.11) and (2.12), there are

$$i_{m2} = \frac{-p(L_m i_{m1} + L_r i_{m2})}{R_2} = -\frac{p\Psi_2}{R_2},$$

$$\Psi_2 = L_m i_{m1} + L_r \left(\frac{-p\Psi_2}{R_2} \right).$$
(2.13)

Thus, there is

$$\Psi_2 = \frac{L_m}{1 + T_2 p} i_{m1},$$

(2.14)

where $T_2 = L_r/R_2$ denotes time-constant of rotor excitation, as we can see, rotor flux Ψ_2 is only produced by i_{m1} and has nothing to do with i_{t1}, therefore, i_{m1} is known as the component of the stator excitation.

Similarly, there is

$$\omega_s = \frac{L_m}{\psi_2 T_2} i_{t1}.$$

(2.15)

Regarding the equation of torque, the solution of simultaneous formula (2.10) and (2.11) is

$$T_e = n_p \frac{L_m}{L_r} \Psi_2 i_{t1}.$$

(2.16)

From formula (2.16) we can see that because rotor flux Ψ_2 is only produced by i_{m1} and has nothing to do with i_{t1}, when Ψ_2 is permanent, torque T_e only has something to do with the component of torque of stator current i_{t1}, amounting to the armature current of DC motor, we achieve a good solution of decoupling.

How to get the direction of rotor flux linkage vector Ψ_2 is one of the technical difficulties of vector control algorithm of asynchronous motor. This paper introduces current hall sensor to measure the instantaneous value of motor's three-phrase stator current i_A, i_B, i_C and introduces photoelectric encoder to measure the signal of rotational speed ω. Then according to the coordinate conversion of formulas (2.1) and (2.2) and the equation of motor, design the magnetic flux viewer, through the indirect method calculate to obtain the size of Ψ_2 and the direction of ϕ.

According to the decoupling mathematical model of induction motor, make use of Matlab/Simulink to build simulation model of vector control system of slip type of induction motor that mainly the contain following modules.

(1) The module of rotational speed controller: achieve closed-loop control of motor rotational speed, introduce traditional PI controller, the module's input is rotational speed demand signal ω^* and rotational speed return signal ω. Take the error between ω^* and ω as the input of PI controller, through the set proportion parameter K_P, integral parameter K_I and torque amplitude limiter calculate the torque given value T_e^*.

(2) The calculated module through the component of stator current magnetic flux and torque: based on vector control system of rotor flux-oriented design, so flux linkage introduces open-loop control, in the steady state, there is

$$i_{sM}^* = \frac{\Psi_r^*}{L_{md}}.$$

(2.17)

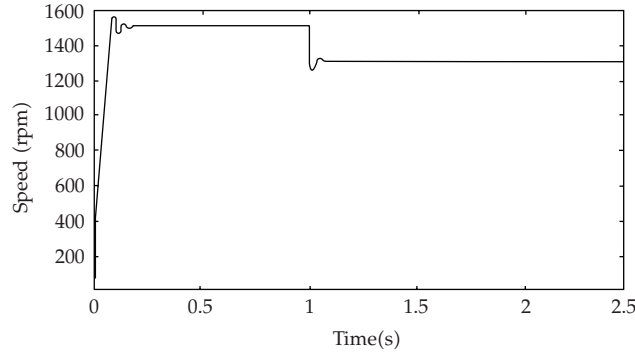Figure 2: The simulation waveform of motor no-load startup on the first second adding load.

Thereby, according to the set flux linkage Ψ_r^* calculate the magnetic flux component of stator current. Then, according to the given torque of the module of rotational speed controller and the given flux linkage, by the equation of torque control of formula (2.16) calculate the torque direction of stator current.

(3) The control module of current hysteresis loop: draw a comparison between the given value of stator three-phrase current and actual value and obtain a deviation, the deviation passes the high-gain amplifier with hysteresis loop character, that is, the hysteresis loop comparator DHC, and compares with the set maximum current deviation to control the make-and-break of two of the upper and lower bridge arm of each phrase of inverter. Make the actual current continually track the waveform of the given current and fluctuate only in the range of deviation.

In addition, the simulation structure also includes the calculation module of the magnetic field orientation angle, the module of 2/3 converter, the module of inverter, the module of induction motor, and so on. Figure 2 is no-load startup and waveform of rotor palstance when time is the first second adding 50 $N \cdot m$ load. We can see when the motor load changes the speed responses very quickly, and the system has good dynamic following and reaches the stable state quickly, which proves the feasibility of oriented vector control based on rotor flux linkage.

3. Controller Design

3.1. Controller Hardware Design

Based on the technical requirements of Hybrid Electric Vehicle motor controller, introduce the special purpose chip of motor control of Microchip Corporation's, dspic33FJ128MC706, to be the controller core and according to the pulse width modulation signal of voltage of asynchronous induction motor generated by the manner of space vector modulation carry on the overall design of the controller. Asynchronous induction motor parameters: rated power $P_N = 55$ kw, rated current $I_N = 300$ A, rated speed $n_N = 2500$ rpm, rated torque $T_e = 200$ $N \cdot m$.

The hardware structure of motor controller is shown in Figure 3, mainly including the following.
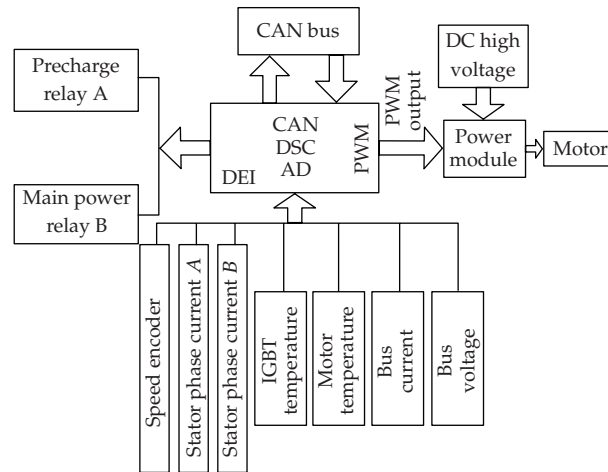
**Figure 3:** The block diagram of motor controller hardware.

(1) Main circuit: the main circuit of controller introduces intelligent power module (IPM) of IGBT to constitute inverter circuit of three-phrase full-bridge and combines of control algorithms for motor control to achieve the following basic functions.

  (i) Load matching: the controller's output power matches the selected motor, not only meet the requirements of electrical rating values, but also ensure that the motor can run in short-term overload.

  (ii) Four-quadrant operation: when Hybrid Electric Vehicle is in the run-time, the forward and reverse drive motor as well as the back coupling of the regeneration energy is the normal running status, requiring that the motor can operate in four-quadrant.

  (iii) High reliability and stability.

(2) Power supply module: provide 15 V, 5 V, 3.3 V power supply for controller and its peripheral devices and complete the power supply monitoring function.

(3) The A/D module: it is responsible to convert multichannel analog quantity into digital quantity, such as the temperature of motor stator, the radiator temperature, the current of motor stator, DC bus, and voltage, which is provided for DSC to calculate and real-time monitor the state of controller and motor as well as providing the necessary parameters i_A and i_B for vector control. The isolation of analog signal uses the precise capacitive isolation amplifier ISO124.

(4) The module of input-output isolation: use optoelectronic isolator TLP521-4/2 to isolate with DSC.

(5) The DEI module: process the signal from the speed sensor and calculate the motor speed, providing the necessary parameter ω for vector control.

(6) The PWM module: export the PWM signal through optocoupler isolation to IGBT drive module, and control the breakover of IGBT's up and down bridge arms to achieve the control of motor.

(7) The module of CAN controller: define a node of CAN bus on the network to achieve the communication among motor controller, power train controller, and battery management system. In order to increase the communication distance, improve the system's instantaneous anti-interference ability, protect bus, reduce radio frequency interference (RFI), and achieve thermal protection; this paper increases CAN transceiver chip PCA82C250 between the module of CAN controller and CAN bus. Considering the car's bad environments, to further enhance the anti-interference measures, between the two CAN notes also adds to the isolation circuit being consisted of high-speed optocoupler 6N137.

(8) Protection circuit: the controller controls the motor's operation, needing real-time monitor motor and its state, including the temperature of the motor stator, the temperature of the inverter radiator, and DC busbar voltage and current. After being converted by the A/D the analog signals compare with the comparator, when there is abnormal situation, for example, the motor load is too high, the back coupling of the energy is too high and illegal operation when debugging, the protection signals after reversed-phrase enter the breakdown input pin of DSC. The software turns off IGBT and coordinates with the IPM's hardware protection to achieve controller's protection function.

(9) The electromagnetic interference suppression: on the basis of analyzing a variety of electromagnetic interference of HEV, introduce different means to enhance the anti-interference ability of electromagnetic sensitive devices such as CAN bus; the detail will be shown in Section 4.

(10) Design for controller vibration isolation: in order to reduce the effects of vibration by the frame to controller, according to GB HG/T3080-1988 (the shockproof rubber material uses rubber material), select the D11-type rubber which has vibration attenuation property to design rubber vibration isolator, and carry on the multi-degree-of-freedom vibration model checking to ensure that in the motivation of a simulation road the amplitude and acceleration of controller meet the requirements of vibration.

3.2. Design for Controller Software

Controller software mainly uses C language and assembly language to develop under the MPLABC30 environment, of which the main program, interrupt handling, CAN communication program use C language, because C language makes the whole process frame structure clear, easy debugging, and maintenance, in line with the concept of structural process design. The related core algorithm like vector control algorithm completely uses the assembly language, which can optimize the code structure, significantly reduce the process running time, improve operating efficiency, and ensure real-time control.

In order to make the software architecture clear, easy debugging, and maintenance, introduce modular design concept, according to the function divide into the module of the main program, vector control, CAN communication, A/D sampling, I/O input-output. The flow of the main function is shown in Figure 4. The controller is electrified to carry out the main function. Initialize clock, peripheral, interface and be connect to 24 V power supply, precharge power capacitors as well as the main power supply relay turns on, then enter the major cycle. Monitor IGBT radiator, the motor temperature, DC bus bar current, and voltage

Figure 4: The main function flow chart.
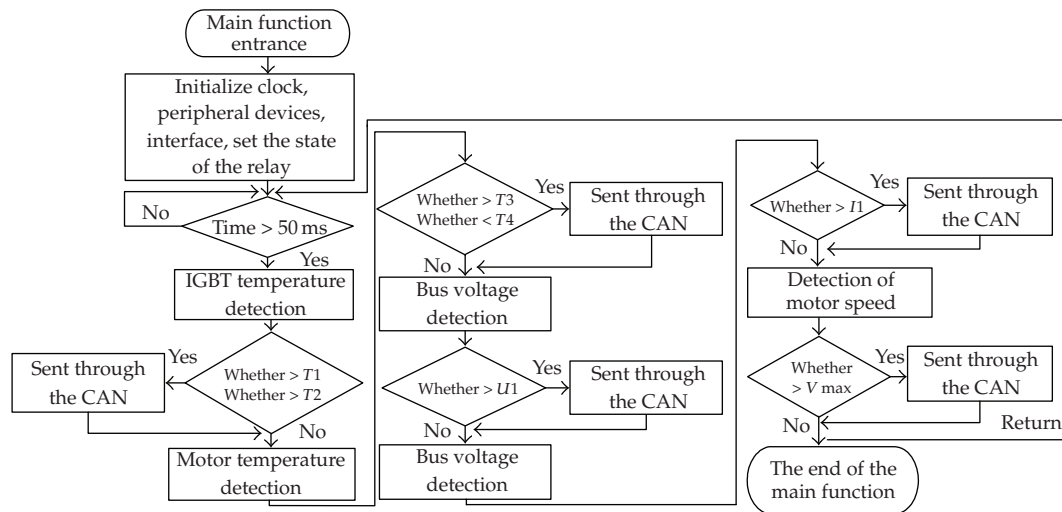
value. At the same time, momentarily respond the interrupt, complete the update of vector control parameters and the communication of CAN with power train controller.

4. Research of Motor Controller's Electromagnetic Compatibility

Compared with the traditional vehicle, HEV has the following characteristics in the control aspect. The control objects are more decentralized, data exchange between the various units is more frequent, and the requirements of reliability and real-time of data exchange are higher. Hence, it is necessary to introduce the simple, efficient, and reliable communication in the HEV.

CAN bus, a serial bus, have the merits of simple structure, fault-tolerant capability, technical maturity, and so forth, which is widely used in the field of automotive electronic control. HEV needs more equipments and higher coordination control, therefore, the use of CAN bus achieving in-vehicle network system is a good solution. All of the Prius of Toyota, the Precept of GM, and the Prodigy of Ford use the CAN bus.

The HEV drive motor of this project is AC asynchronous motor, being driven by the inverter. The power module, (insulated gate bipolar transistor) IGBT, is widely used in the inverter, which is a dual current-carrying device and whose switching characteristics are controlled by the gain of PNP transistor. Therefore, in essence, the current peak of IGBT will be higher than that of (metal-oxide semiconductor field effect transistor) MOSFET. Meanwhile, the carrier frequency of the inverter is up to 20 kHz and bus bar voltage is over a hundred volts, so when there is (pulse-width modulation) PWM wave, there will be a very high voltage peak value, which will inevitably lead to serious interference noise of conduction and electromagnetic radiation [9–11]. In addition, the number of automotive electronic equipment increases day by day, which make the electromagnetic environment of the HEV complex and poor. So, it is of great significance to analyze the system's electromagnetic interference source and its propagation path, take the corresponding measures to make electromagnetic sensitive devices, such as air bag and CAN bus network, work regularly
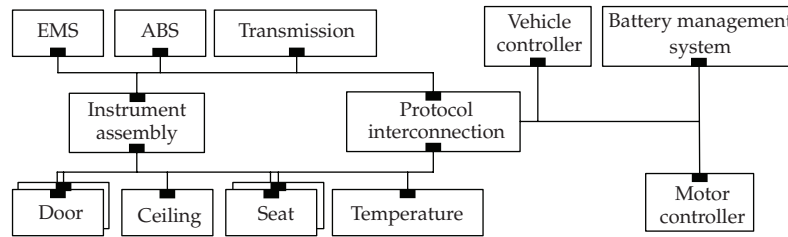
Figure 5: Frame of HEV CAN bus.

under strong interference condition. The experimental results indicate that the antijamming measures are practical and the system operates well.

4.1. Hybrid Electric Vehicle CAN Bus System

The in-vehicle network system of HEV of this project is composed of two major parts, which are connected by the protocol interconnection component and consist of unitive CAN bus network. The basic structure is shown in Figure 5.

The first part includes three node controllers, that is, motor controller, vehicle controller, and battery management system. Each node controller can independently accomplish the function of command decomposition and transfer bottom layer instruction. In this way, for an action unit, the node controller will act as "command host."

The second part includes body measurement and control points, such as car door, ceiling, seat, and the temperature control point, meanwhile include the communication among engine, ABS, and transmission and transfer the data to automobile instrument assembly.

The protocol interconnection component: this part is responsible for the connection of the above two parts. Although these two major parts use the CAN bus network and their physical layer and partial data link layer are established on the CAN2.0B, partial data link layer and application layer protocol due to controlled device use different communication protocols. Therefore, introduce the protocol interconnection component to connect the bus protocol of the two segments so as to consist of unitive bus control network.

Obviously, the CAN bus network system is very important, in relation to vehicle safe information, such as EMS, ABS, and other related information, must be timely, reliable transmission. Due to covering the entire body, its performance and reliability directly involve the normal work of the entire automotive.

4.2. Analysis of System Electromagnetic Interference

Analysis of the source of electromagnetic interference and its propagation path is very important for solving the problem of system electromagnetic compatibility. Use the asynchronous motor, power up to 30 KW, for debugging experiments. At first, carry on the point-to-point communication debugging, the power supply voltage of bus bar is 48 V. After using the conventional antijamming measures, CAN bus system can be normal communication. However, when we do the bench experiment, the power supply is 336 V battery pack, there will always present the phenomena of CAN bus inaccessibility and drop
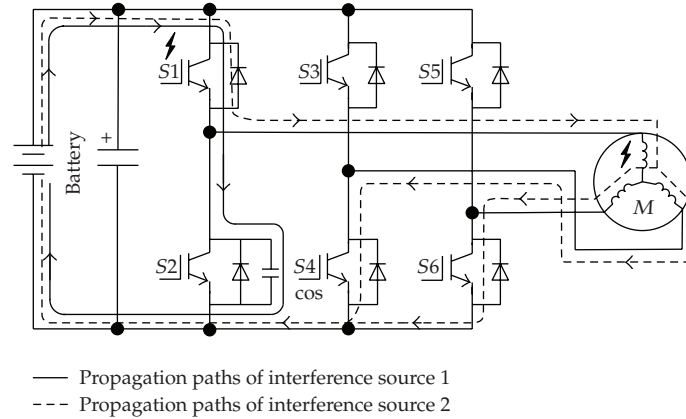
———— Propagation paths of interference source 1
- - - - Propagation paths of interference source 2

Figure 6: Differential mode interference sources and their propagation paths.

frames during data transmission. This indicates that when the inverter operates under the rated conditions, there will be a very strong electromagnetic interference which seriously affects the normal communication of the CAN bus system by conduction and radiation. The experiment shows that the main sources of electromagnetic interference include the following aspects.

(1) *Differential Mode Interference*

Make the bus bar as the interference of shuttle route, which forms the loop through bus bar, battery, and inverter. By the high frequency equivalent circuit of the three-phrase inverter, when IGBT switch acts form a kind of differential mode interference source [12], which returns to battery through bus bar. Besides, when the motor is high frequency voltage, the stator coil will produce voltage peak, which forms another differential mode voltage through bus bar [13]. Both measure battery terminal voltage and motor phase voltage and carry on Fourier transformation and compare or use the linear impedance stabilization network series in the bus bar can analyze the two differential mode interferences. Its propagation path is shown in Figure 6.

(2) *Common Mode Interference*

The interference forms through the parasitic capacitance of phase and GND and forms the loop again through GND. When the IGBT switch acts, there will be not only differential mode interference voltage but also common mode interference voltage [12]. The common mode interference voltage, through the parasitic capacitance of IGBT base relative to the radiator, forms the common mode interference current flowing to the chassis. Besides, form the voltage peak in the stator coil, at the same time induce the shaft bearing voltage, which produces the shaft bearing current flowing to the chassis through the motor shaft, thus form another common mode interference current [13]. Measure the current and carry on Fourier transformation, obtain its propagation path as shown in Figure 7.
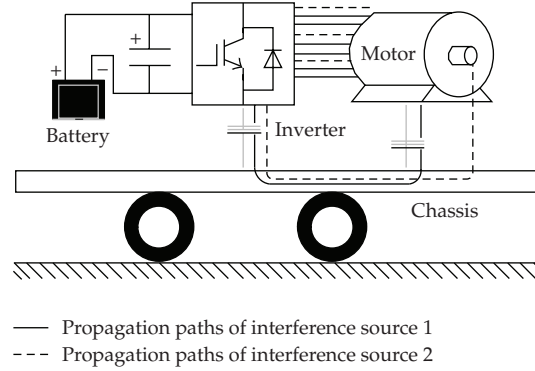
──── Propagation paths of interference source 1
- - - Propagation paths of interference source 2

Figure 7: Common mode interference sources and their propagation paths.

(3) *radiated Interference*

The interference source takes the energy to radiate to the surrounding space in the form of electromagnetic wave. The experiment shows that around the inverter and the entire chassis the radiation intensity is very high, even when two vehicles approach, the electromagnetic radiation level will intensify. Because the differential mode interference formed in the inverter radiate energy to the surrounding space, and because the common mode current flowing to the chassis forms the current loop on its surface due to the conductor skin effect, there will be very strong electromagnetic interference. Obviously, the best way to suppress the radiation interference is to weaken the intensity of differential mode and common mode interference.

In addition, when the generator commutation or load current change suddenly, there will be electromagnetic wave. The switching frequency, the multiples of switching frequency, and high frequency harmonic of the PWM signal will produce radiation interference.

4.3. Solutions for the Problem of System Electromagnetic Interference

Practice shows that the occurrence of any electromagnetic interference must have three elements of interference source, propagation path, and sensitive devices. Any element is weakened or lost, the problem of electromagnetic interference will be improved or solved. Based on the above analysis, we take the following antijamming measures.

(1) *Attenuation of Differential Mode Interference and Radiation Interference*

Introduce the multilayer printed power circuit board to suppress the differential mode interference in the inverter and the corresponding electromagnetic radiation. The results of the experiment indicate that the frequency of the differential mode interference is 10–20 M. By formula (4.1), δ, the skin depth of interference on copper, is 0.01-0.02 mm. All of the interference due to conductor skin effect flow from the conductor surface

$$\delta = \sqrt{\frac{1}{\pi f \mu \sigma}}, \tag{4.1}$$

where f denotes the interference frequency, μ denotes the permeability of copper, $\mu = 4\pi \times 10^{-7}$ H/m, σ denotes the electrical conductivity of copper, $\sigma = 5.8 \times 10^7$ S/m.

Thus, the effect of traditional filter with lumped parameter circuit model is not good. The multilayer power circuit board is a sandwich structure. The second and third layers are used to take the bus bar current, the symmetrical structure, appropriate wire width, and dielectric substance thickness form the reasonable coupling capacitance, which is used to absorb the differential mode interference entering the bus bar to reduce the corresponding radiation energy.

(2) *Attenuation of Common Mode Interference and Radiation Interference*

Introduce the appropriate ground impedance to attenuate the common mode interference current flowing to the ground and the corresponding electromagnetic radiation. Usually, only the tires of the vehicle contact with the ground, so the whole body is insulated to the ground. Being known as the previous analysis, the common mode interference current flows to the chassis through the inverter radiator base or motor shaft, its voltage is likely to endanger the safety of the passengers. Therefore, the radiator base and motor enclosure must be grounded. The grounding impedance must be appropriate to furthest attenuate the common mode interference current. The value of R, C, L can be calculated by the following two formulas:

$$i' = \frac{E}{\sqrt{L/C}} \varepsilon^{-1/T^t},\tag{4.2}$$

$$L = \sqrt{\frac{\sqrt{L/C}^2}{4\pi^2 f_c^2 + (R/2L)^2}},\tag{4.3}$$

where i' denotes the peak common mode interference current flowing to the chassis, E denotes common mode voltage, T denotes the time constant of peak current, the size is $R/2L$, f_c denotes the frequency of peak current i'.

Measure the common mode interference current flowing to the chassis and carry on Fourier transformation to find out peak current i', its amplitude is $E/\sqrt{L/C}$. E is common mode voltage, which usually takes one-third of bus bar voltage, thus obtain $\sqrt{L/C}$. $(R/2L)$ can be estimated through time constant of peak current, calculate L by formula (4.3), then figure out R, C. Meanwhile, consider the upper limit of safe voltage. The experiment indicates that the increase of grounding impedance is helpful to reduce common mode current, but simultaneously enhance the potential of motor enclosure.

(3) *π-Type Filter Circuit Improves the CAN Antijamming Ability*

This filter circuit ingeniously uses the parallel resonant characteristic of inductance coil and capacitance, appropriate to adjust its value to make frequency characteristic curve steep and peak high. Because the coil impedance is very small, the signal waveform will not distort when suppress the noise. Insert loss to indicate the filter's attenuation to the interference, use

IL to express. It is related to the parameters of inductance, capacitance, noise frequency and so on, the relationship is shown as follows:

$$IL = 20\lg\left[\frac{\omega^2 LC + \left(LC^2\omega^3 + 2\omega C\right)R_L R_S}{R_S + R_L}\right], \tag{4.4}$$

where R_L denotes load impedance, R_S denotes source impedance, ω denotes noise angular frequency, L denotes inductance value of coil, C denotes the size of capacitance. In view of the corresponding interference frequency and the requirements of *IL,* calculate the parameter value of π-type filter. Considering the inaccurate estimates of load impedance R_L and source impedance R_S and the impact of parasitic capacitance on inductance coil when it is high frequency, the calculated value needs to adjust according to the actual debugging effect.

(4) *The Matching Resistance Improves the CAN Antijamming Ability*

Appropriate choice of matching resistance can greatly improve the antijamming ability of CAN bus system. The signal at the end of the transmission line suddenly encounters very small cable impedance, then it will cause reflection in this place and create the waveform distortion. In theory, as long as connecting the terminal resistance matching cable characteristic impedance at the end of the transmission cable, there will not present the phenomenon of signal reflection. However, in fact, the terminal resistance of the whole communication system is also affected by the number of node, the impedance of each node, capacitance and bus capacitance, and it is often different from the recommendation value $120\,\Omega$ of the twisted-pair terminal matching impedance. When debugging let the sending end send data 55 H in use of the actual baud rate, at the receiving end use the oscilloscope to observe the output square wave, adjust the matching impedance to obtain the best waveform whose amplitude voltage is no less than 200 mV, this value is the best matching impedance.

5. Experimental Results

5.1. CAN Communication Debugging

In the HEV complex electromagnetic environment, the real-time, reliability, and accuracy of the CAN communication must be first assured. In the third part of the electromagnetic compatibility of motor controller, analyze the source of electromagnetic interference and its propagation path in detail. In order to verify whether the used antijamming measures are reliable and effective, the following experiments have been done.

(1) *The Waveform Debugging of CAN Bus Communication*

To have a direct-viewing understanding of the antijamming effect of CAN bus communication, use digital oscilloscope to observe the communication waveform. As shown in Figure 8, after adding the antijamming measures, the waveform quality is improved obviously.
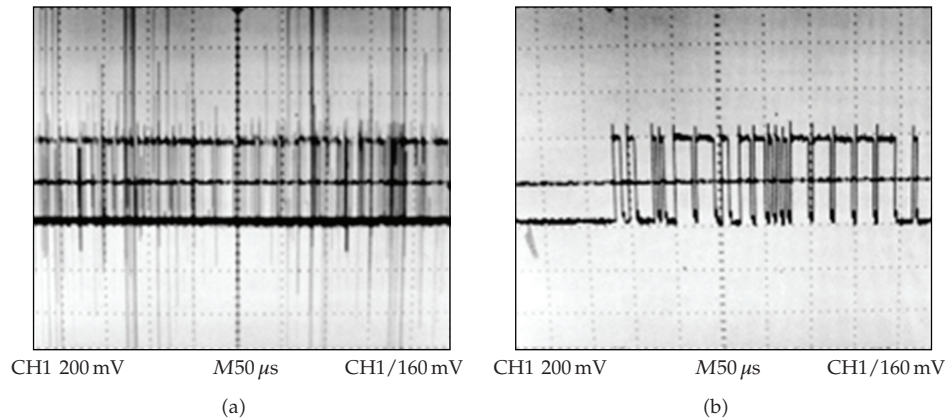
CH1 200 mV $M50\,\mu s$ CH1/160 mV   CH1 200 mV $M50\,\mu s$ CH1/160 mV

(a)   (b)

Figure 8: The CAN wave before and after taking antijamming measures.

Table 1: The experiment results of error rate.

	Speed/(kbps)	Transmission distance/(m)	Bit error rate
Take no measures	128	20	$<10^{-6}$
Take no measures	512	20	$<10^{-4}$
Take measures	128	20	$<10^{-8}$
Take measures	512	20	$<10^{-6}$

(2) *The Experiment of Error Rate of CAN Bus Communication*

In order to verify the stability of CAN bus system after using the above antijamming measures, design and carry on the following experiment. Take two nodes from the bus to carry on the data transmission experiment, the nodes connect with the upper monitor through CAN and USB converter interface. The node 1 sends quantitative bits data to the node 2 in accordance with predetermined data format. On the node 2 the software of upper monitor compares the transmitter and receive data and analyzes the wrong bits. Table 1 is the experiment result. After using the corresponding antijamming measures, the error rate of the communication decreases two orders of magnitude.

5.2. The Experiment of the Control Logic

During the process of controller controlling motor, in order to assure the safety of the system, the control logic, such as power-on sequence and fault treatment, is particularly important. The power-on sequence is the controller of power train sends self-checking instruction to each controller, when the battery management system receives the self-checking instruction, it will carry on self-checking, then connect the relays of all of the battery packs, the battery packs precharge the capacitance through current-limiting resistance; the motor controller is power-on, self-checking, after completion sends the ready instruction to the controller of power train, the controller of power train sends the enabled instruction, when the motor controller receive the enabled instruction, connect the capacitance precharge relay and short current-limiting resistance, and connect the main relay to electrify the controller, at this point the controller will be able to start the motor running. The logic of fault treatment is that controller realtime
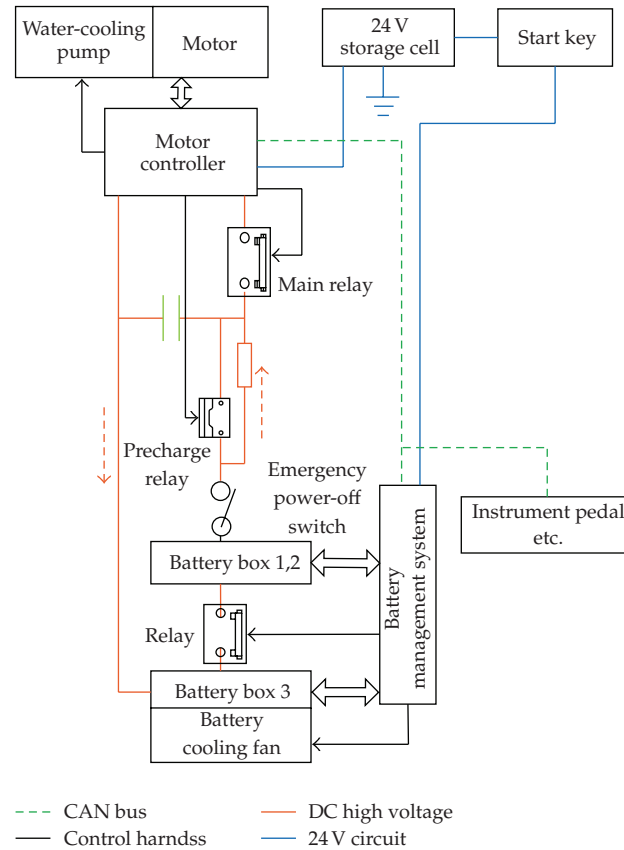
Figure 9: The principle diagram of test-bench.

monitors the voltage and current of DC bus bar, the temperature of rotor and controller, when overruns, immediately cut the output of PWM and close IGBT and the main relay to ensure the safety of the whole system.

5.3. The Motor Bench Experiment

The motor drive system of HEV need has the performance of low speed and high torque, a wide range of variable speed, strong overload capacity, fast torque output respond, high efficiency of multipoint operating mode, and so on. Obviously, the performance of motor and controller and their matching are very important. Therefore, after completing the hardware debugging of motor controller and the main software writing, carry on the bench experiment by controller driving motor.

The bench experiment can effectively detect whether the performance of the motor and its controller meet the design requirements, and it is the most effective measure to assure motor and its controller work safe and reliable in the vehicle and to reduce the unnecessary trouble of debugging on board. The bench of this paper is constituted of motor, controller, cooling system, electric eddy current dynamometer, upper monitor, power battery pack, and its management system, as shown in Figure 9. The dynamometer is connected with motor, as

Table 2: The motor performance parameters.

Type	YPQ160M-4 (55 kW)	
Performance parameters	Design value	Trial value
Peak power (kW)	110	110
Peak power efficiency (%)	89	89.2
Peak power slippage (%)	4.5	4.45
Overload power (kW)	83	83
Overload power efficiency (%)	91	91.15
Overload power slippage (%)	3.5	3.4
Rated power (kW)	55	55
Rated power efficiency (%)	92.5	92.8
Rated power slippage (%)	2.5	2.43
Half-power (kW)	27.5	27.5
Half-power efficiency (%)	92.4	92.55
Half-power slippage (%)	1.5	1.4

upper monitor software can realize the real-time measurement and record of motor torque and rotational speed. The controller controls the motor operation and communicates with power train controller with CAN to get the request torque information, at the same time manages the motor and controller's cooling system. Power battery pack is divided into three groups and are in series by relays, the battery management system monitors and manages its state.

When reverse the key to electrify, the battery management system connects the battery's internal relay, and the battery precharges the capacitor through current-limiting resistance; then controller makes precharge relay connect, shorts charging resistance and connects main relay to electrify the controller, at this time controller can start motor running. In order to prevent accidents from happening, during the experiment add an emergency power-off switch, which can be artificially cut off the battery power supply, to fully guarantee the security of the entire system.

According to "Electric vehicle motor and its controller Part 2: test methods" (GB/T 18488.2-2006), carry on the experiment of torque-speed characteristic, power, efficiency, slip. The experimental data is shown in Table 2, the torque-speed curve is shown in Figure 10.

Note. In Figure 9, in order to conveniently illustrate, draw capacitor, main relay, precharge capacitor, and current-limiting resistance outside of controller which originally locate in controller.

On the analysis of the experimental data and motor characteristic curve, the performance parameters of the motor rated power, overload power, peak power and slip can reach or approach the design requirements, and the system efficiency of the selected experimental points are all above 89.2%; that is, the comprehensive efficiency is high, this is very important for HEV sometimes light load, sometimes heavy load and sometimes overload. On the condition of overload motor can constantly output at torque $T = 300\,N \cdot m$ and speed $n = 0$–2500 rpm, on the condition of rated torque $T_e = 200\,N \cdot m$ motor can even remain the speed $n = 3500$ rpm, which is higher than the rated speed $n_e = 2500$ rpm. What we can see is that the drive system not only can output high torque $T = 300\,N \cdot m$ in low speed, but also can govern speed in a wide range $n = 0$–3500 rpm in the rated torque.
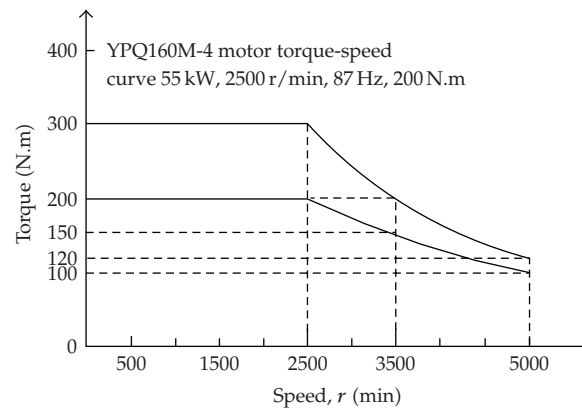
Figure 10: Motor torque-speed characteristic curve.

6. Conclusion

Nowadays, research of Hybrid Electric Vehicle is one of the most meaningful ways to solve the problems of pollution and energy. Motor and its control technology are one of main components of Hybrid Electric Vehicle, and largely affect vehicle's power performance, fuel economy, and emission. This paper carries on the research of motor control technology of application HEV, designs HEV motor controller based on vector frequency conversion technology, and completes the controller's software and hardware design. In accordance with the characteristics of the HEV electromagnetic interference source, propagation path, and CAN bus communication, this paper proposes the corresponding antijamming measures, which greatly improves the antijamming ability of the CAN bus system and controller working stability. Finally, the motor bench experiment verifies that the designed driving system has the performance of low speed and high torque, a wide range of variable speed and high comprehensive efficiency, which create advantages for system dynamic performance experiment, durability experiment and vehicle road experiment.

Acknowledgment

References

[1] Q. Q. Chen, F. C. Sun, and J. G. Zhu, *Modern EV Technology*, Beijing Institute of Technology Press, Beijing, China, 2002.

[2] H. Hu and H. Song, *Electric Vehicle*, China Communications Press, Beijing, China, 2003.

[3] R. Saeks, C. J. Cox, J. Neidhoefer, P. R. Mays, and J. J. Murray, "Adaptive control of a hybrid electric vehicle," *IEEE Transactions on Intelligent Transportation Systems*, vol. 3, no. 4, pp. 213–233, 2002.

[4] W. Liu, G. W. Cheng, and C. N. Deng, "Analysis of electric vehicle's driving motor," *Motor Technology*, pp. 26–28, 2006.

[5] H. D. Li, *The Control System of AC Speed-Governing*, Publishing House of Electronics Industry, 2003.

[6] K. Zhou and D. Wang, "Relationship between space-vector modulation and three-phase carrier-based PWM: a comprehensive analysis," *IEEE Transactions on Industrial Electronics*, vol. 49, no. 1, pp. 186–196, 2002.

[7] S. Shinnaka, S. Takeuchi, A. Kitajima, F. Eguchi, and H. Haruki, "Frequency-hybrid vector control for sensorless induction motor and its application to electric vehicle drive," in *Proceedings of the IEEE Applied Power Electronics Conference and Exposition (APEC '01)*, vol. 1, pp. 32–39, 2001.

[8] Z. X. Fu, J. Xiang, W. C. Reynolds, and B. Nefcy, "Vector control of an IPM synchronous machine capable of full range operations for hybrid electric vehicle application," in *Proceedings of the IEEE Industry Applications Society Conference (IAS '03)*, vol. 3, pp. 1443–1450, 2003.

[9] D. A. Rendusara and P. N. Enjeti, "An improved inverter output filter configuration reduces common and differential modes dv/dt at the motor terminals in PWM drive systems," *IEEE Transactions on Power Electronics*, vol. 13, no. 6, pp. 1135–1143, 1998.

[10] N. Mutoh, J. Nakashima, and M. Kanesaki, "Multilayer power printed structures suitable for controlling EMI noises generated in power converters," *IEEE Transactions on Industrial Electronics*, vol. 50, no. 6, pp. 1085–1094, 2003.

[11] Y. Onozawa, M. Otsuki, and Y. Seki, "Investigation of carrier streaming effect for the low spike fast IGBT turn-off," in *Proceedings of the 18th International Symposium on Power Semiconductor Devices & IC's*, pp. 173–176, University of Naples, Naples, Italy, 2006.

[12] C. C. Chen and X. Y. Xu, "Modeling the conducted EMI emission of an electric vehicle (EV) traction drive," in *Proceedings of IEEE International Symposium on Electromagnetic Compatibility*, vol. 2, pp. 796–801, Denver, Colo, USA, 1998.

[13] N. Mutoh, M. Nakanishi, M. Kanesaki, and J. Nakashima, "Control methods for EMI noises appearing in electric vehicle drive systems," in *Proceedings of the IEEE Applied Power Electronics Conference and Exposition (APEC '05)*, vol. 2, pp. 1022–1028, Austin, Tex, USA, 2005.

Research Article

Biologically Inspired Robotic Arm Control Using an Artificial Neural Oscillator

Woosung Yang,[1] Jaesung Kwon,[1] Nak Young Chong,[2] and Yonghwan Oh[1]

[1] *Center for Cognitive Robotics Research, Korea Institute of Science and Technology, P.O. Box 131, Cheongryang, Seoul 130-650, South Korea*
[2] *School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*

Correspondence should be addressed to Woosung Yang, dreamrize@gmail.com

We address a neural-oscillator-based control scheme to achieve biologically inspired motion generation. In general, it is known that humans or animals exhibit novel adaptive behaviors regardless of their kinematic configurations against unexpected disturbances or environment changes. This is caused by the entrainment property of the neural oscillator which plays a key role to adapt their nervous system to the natural frequency of the interacted environments. Thus we focus on a self-adapting robot arm control to attain natural adaptive motions as a controller employing neural oscillators. To demonstrate the excellence of entrainment, we implement the proposed control scheme to a single pendulum coupled with the neural oscillator in simulation and experiment. Then this work shows the performance of the robot arm coupled to neural oscillators through various tasks that the arm traces a trajectory. With these, the real-time closed-loop system allowing sensory feedback of the neural oscillator for the entrainment property is proposed. In particular, we verify an impressive capability of biologically inspired self-adaptation behaviors that enables the robot arm to make adaptive motions corresponding to an unexpected environmental variety.

1. Introduction

Recently biologically inspired systems and control methods have been studied widely, in particular in robotics field. Thus, a number of virtual human or animal-like robots and control approaches have been yielded for the last decade. Owing that such approaches enable robots to embody autonomous dynamic adaptation motion against unknown environmental changes, its attraction has become generally gained and issued. This is

because the musculoskeletal system is activated like a mechanical spring by means of the central pattern generators (CPGs) and their entrainment property [1–3]. The CPGs consist in the neural oscillator network and produce a stable rhythmic signal. Entrainment of the neural oscillator plays a key role to adapt the nervous system to the natural frequency of the interacted environments incorporating a sensory feedback. Hence, the neural oscillator in the nervous system offers a potential controller, since it is known to be robust and have an entrainment characteristic as a local controller in humans or animals.

Relating these previous works, the mathematical description of a neural oscillator was presented in Matsuoka's works [1]. He proved that neurons generate the rhythmic patterned output and analyzed the conditions necessary for the steady state oscillations. He also investigated the mutual inhibition networks to control the frequency and pattern [2], but did not include the effect of the feedback on the neural oscillator performance. Employing Matsuoka's neural oscillator model, Taga et al. investigated the sensory signal from the joint angles of a biped robot as feedback signals [3, 4], showing that neural oscillators made the robot robust to the perturbation through entrainment. This approach was applied later to various locomotion systems [5–7]. In addition to the studies on robotic locomotion [8], more efforts have been made to implement the neural oscillator to a real robot for various applications. Williamson showed the system that had biologically inspired postural primitives [9]. He also proposed the neuromechanical system that was coupled with the neural oscillator for controlling its arm [10]. Arsenio [11] suggested the multiple-input describing function technique to evaluate and design nonlinear systems connected to the neural oscillator.

As above, existing works in field of biologically inspired system based on neural oscillators have yielded notable results in many cases. However approaches for a proper behavior generation and complex task were not clearly described due to the difficulty in application considering a real robotic manipulator coupled with the neural oscillator. Yang et al. has presented simulation and experimental results in controlling a robot arm and humanoid robot incorporating neural oscillators [12–15]. Apart from such the proposed parameter optimization method, we newly address an intuitive and efficient approach for a desired task of the neural-oscillator-based control. In addition, this work addresses how to control a real system coupled with the neural oscillator for a desired task. For this, real-time feedback is implemented to exploit the entrainment feature of the neural oscillator against unpredictable disturbances.

In the following section, a neural oscillator is briefly explained and its entrainment property is described and verified. Details of the dynamic stability of the developed methodology are discussed in Section 3. The experimental results are presented in Sections 4 and 5. Finally, conclusion is drawn in Section 6.

2. Rhythmic Movement Using an Artificial Neural Oscillator

2.1. Matsuoka's Neural Oscillator

Matsuoka's neural oscillator consists of two simulated neurons arranged in mutual inhibition as shown in Figure 1 [1, 2]. If gains are properly tuned, the system exhibits limit cycle behaviors. The trajectory of a stable limit cycle can be derived analytically, describing the firing rate of a neuron with self-inhibition. The neural oscillator is represented by a set of
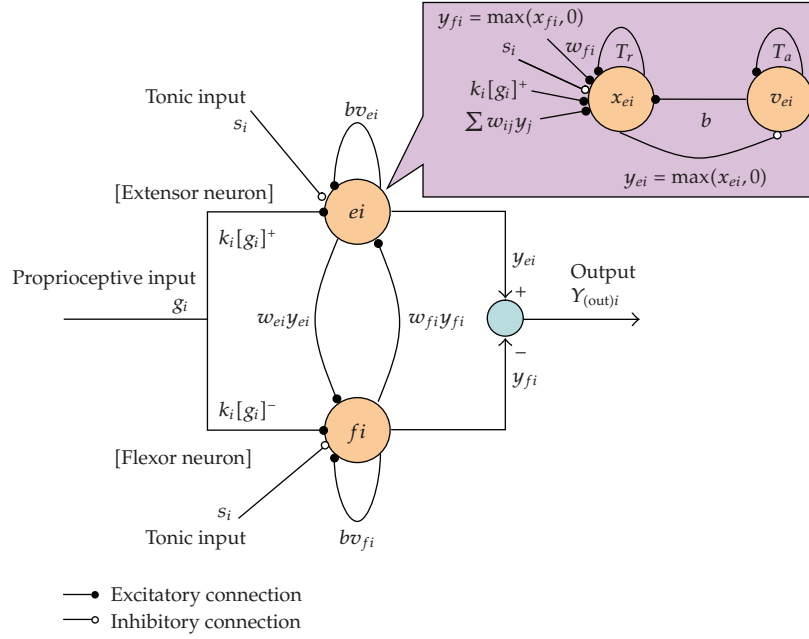
Figure 1: Schematic diagram of Matsuoka's neural oscillator.

nonlinear coupled differential equations given as

$$T_r \dot{x}_{ei} + x_{ei} = -w_{fi} y_{fi} - \sum_{j=1}^{n} w_{ij} y_j - b v_{ei} - \sum k_i [g_i]^+ + s_i,$$

$$T_a \dot{v}_{ei} + v_{ei} = y_{ei},$$

$$y_{ei} = [x_{ei}]^+ = \max(x_{ei}, 0),$$

$$T_r \dot{x}_{fi} + x_{fi} = -w_{ei} y_{ei} - \sum_{j=1}^{n} w_{ij} y_j - b v_{fi} - \sum k_i [g_i]^- + s_i,$$

$$T_a \dot{v}_{fi} + v_{fi} = y_{fi},$$

$$y_{fi} = [x_{fi}]^+ = \max(x_{fi}, 0), \quad (i = 1, 2, \ldots, n),$$

(2.1)

where $x_{e(f)i}$ is the inner state of the ith neuron which represents the firing rate; $v_{e(f)i}$ represents the degree of the adaptation, modulated by the adaptation constant b, or self-inhibition effect of the ith neuron; the output of each neuron $y_{e(f)i}$ is taken as the positive part of x_i, and the output of the whole oscillator as $Y_{(\text{out})}$; w_{ij} (0 for $i \neq j$ and 1 for $i = j$) is the weight of inhibitory synaptic connection from the jth neuron to the ith neuron, and w_{ei}, w_{fi} are also weights from the extensor neuron to the flexor neuron, respectively; $w_{ij} y_i$ represents the total input from the neurons inside the network; the input is arranged to excite one neuron and inhibit the other, by applying the positive part to one neuron and the negative part to the other; T_r and T_a are time constants of the inner state and the adaptation effect of
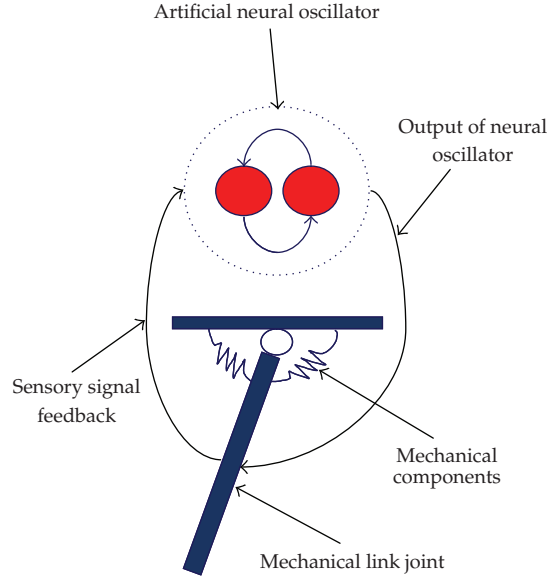
Figure 2: Mechanical system coupled to the neural oscillator.

the ith neuron, respectively; s_i is the external input, and g_i indicates the sensory input from the coupled system which is scaled by the gain k_i.

Figure 2 shows a simple mechanical system connected to the neural oscillator. The coupled method enables a robot to adapt to changing conditions. The desired torque signal to the ith joint can be given by

$$\tau_i = -k_{oi}(q_i - q_{\text{odi}}) - b_i\dot{q}_i, \tag{2.2}$$

where k_{oi} is the stiffness of the joint, b_i is the damping coefficient, q_i is the joint angle, and q_{doi} is the output of the neural oscillator that produces rhythmic commands of the ith joint. The neural oscillator follows the sensory signal from the joints; thus the output of the neural oscillator may change corresponding to the sensory input. This is what is called "entrainment" that can be considered as the tracking of sensory feedback signals so that the mechanical system can exhibit adaptive behavior interacting with the environment.

2.2. Entrainment Property of Neural Oscillator

Generally, according to Matsuoka's work [1, 2, 16], the entrainment can be realized under stable oscillation conditions of the neural oscillator. For stable oscillations, if tonic input exists, then T_r/T_a should be in the range $0.1 \sim 0.5$, for which the natural frequency of the oscillator is proportional to $1/T_r$. The magnitude of the output signal also increases as the tonic input increases. T_r and T_a have an effect on control of the delay time and the adaptation time of the entrained signal, respectively. Thus, as these parameters decrease, the input signal is well entrained. And the minimum gain k_i of the input signal enlarges the entrainment capability, because the minimum input signal is needed to be entrained appropriately in the range of the natural frequency of an input signal. In this case, regardless of the generated natural
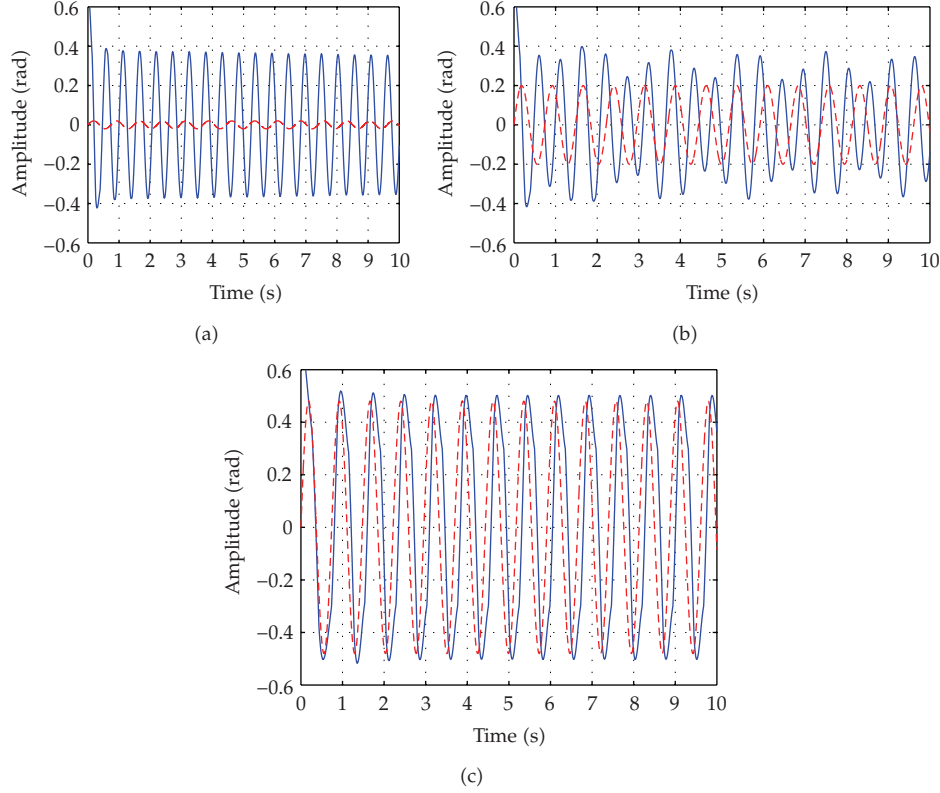
(a)

(b)

(c)

Figure 3: Simulation results on the entrainment property of the neural oscillator in cases that $k = 0.02$ (a), $k = 0.2$ (b), and $k = 0.53$ (c), respectively. The solid line is the output of the neural oscillator and the dashed line indicates the sensory signal input.

frequency of the neural oscillator and the natural frequency of an input signal, the output signal of the neural oscillator locks onto an input signal well in a wide range.

Figure 3 illustrates the entrainment procedure of the neural oscillator. If we properly tune the parameters of the neural oscillator, then the oscillator exhibits the stable limit cycle behaviors. In Figure 1, the gain k of the sensory feedback was sequentially set as 0.02, 0.2, and 0.53 such as in Figures 3(a), 3(b), and 3(c). When k is 0.02, the output of the neural oscillator cannot entrain the sensory signal input as shown in Figure 3(a). The result of Figure 3(b) indicates the signal partially entrained. If the gain k is properly set as 0.53, the neural oscillator produces the fully entrained signal as illustrated in Figure 3(c) in contrast to the result of Figure 3(b).

2.3. Verification of Entrainment Property through Experiment

In this subsection, we experimentally verify the entrainment capability of the neural oscillator and its validation addressed in above Section 2.2. Figure 4 shows the experiment setup to implement entrainment to real robotic systems. As illustrated in Figure 2, the single pendulum is tightly coupled with the neural oscillator. This means that the neural oscillator observes and entrains the encoder value of the motor in terms of the sensory feedback and
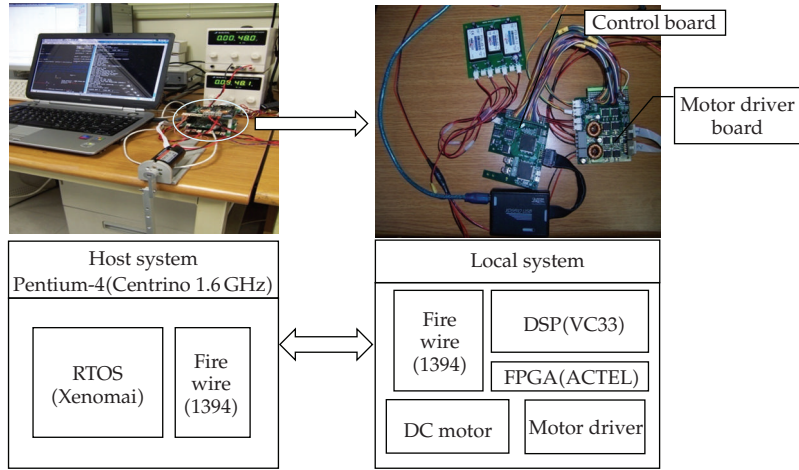
Figure 4: Experimental setup for driving the single pendulum coupled with the neural oscillator. This operating system runs at 200 Hz in real time.
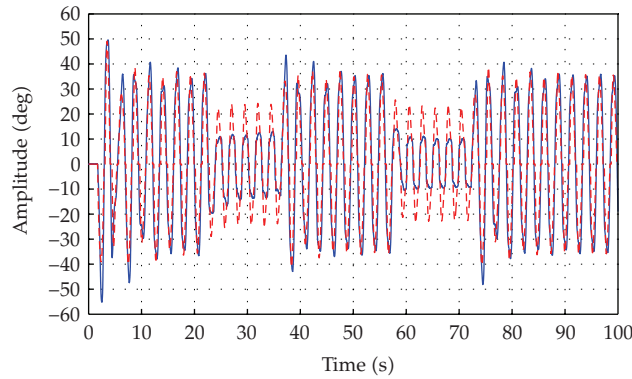

Figure 5: Experiment result on self-adapting motions of the coupled oscillator-pendulum. The red line is the encoder value and the dashed line indicates the output of the neural oscillator.
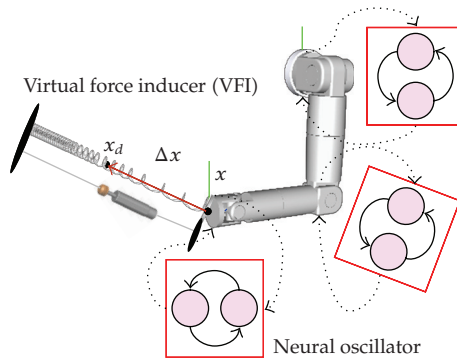

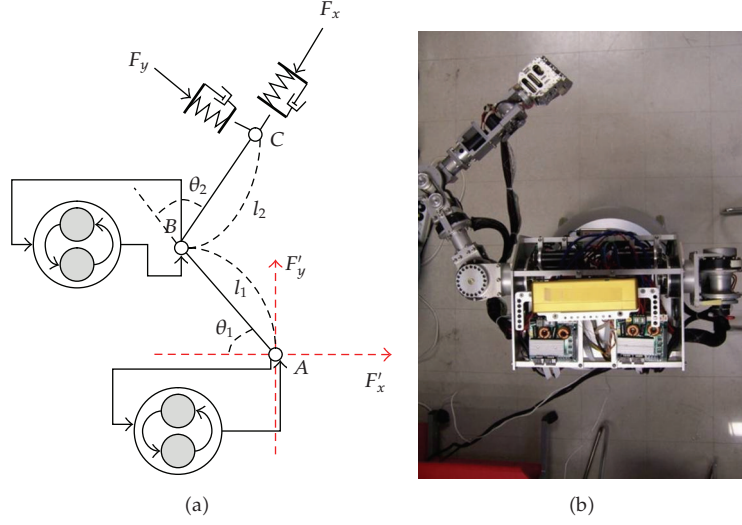**Figure 6:** Schematic robot arm control model coupled with neural oscillators.

Figure 7: (a) Schematic robot arm model and (b) real robot arm coupled with the neural oscillator for experimental test.

the output of the neural oscillator drives the motor directly. Hence, the pendulum is excited periodically by the output generated in the neural oscillator. And also the coupled oscillator-pendulum exhibits natural adaptive motions even though we swing the pendulum arbitrary 22 s to 38 s and 56 s to 74 s sequentially as shown in Figure 5. It can be confirmed from the experimental result that entrainment of the neural oscillator enables the coupled system to show naturally self-adapting motions against unpredictable disturbances.

3. Control Scheme Based on Neural Oscillator

The neural oscillator is a nonlinear system; thus it is generally difficult to analyze the dynamic system when the oscillator is connected to it. Therefore we studied the existence of singular points and their stability of the neural oscillator in time domain analysis investigating equilibrium states [16]. Also analysis of a simple mechanical system coupled with the neural oscillator through a graphical approach known as the describing function analysis has been proposed earlier [17]. The main idea is to plot the system response in the complex plane and find the intersection points between two Nyquist plots of the dynamic system and the neural oscillator. The intersection points indicate limit cycle solutions. However, even if a rhythmic motion of the dynamic system is generated by the neural oscillator, it is usually difficult to obtain the desired motion required by the task. This is because many oscillator parameters need to be tuned, and different responses occur according to the interoscillator network. Hence, we propose the control method that enables a robot system to perform a desired motion without precisely tuning parameters of the neural oscillator within the range of its well-known stable condition.

Figure 6 illustrates a schematic model of a robot arm whose each joint is coupled to the neural oscillators. And a virtual force leads the coupled robot arm to a given motion. The virtual force inducer (VFI) such as springs and dampers which is supposed to exist virtually
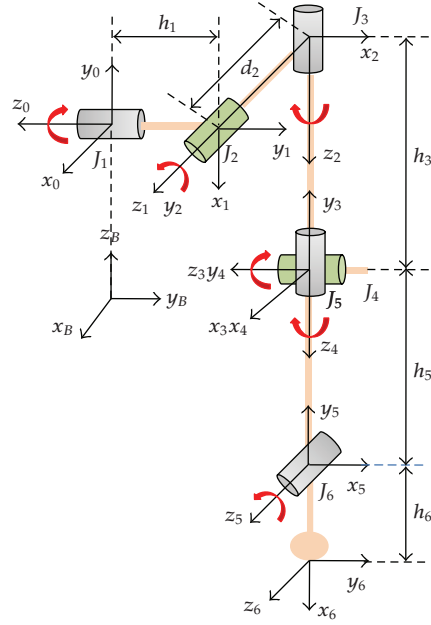
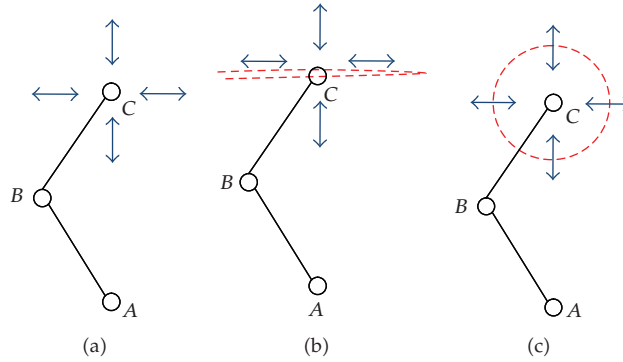**Figure 8:** Schematic figure for D-H parameter of the robot arm.

Figure 9: The given tasks of the two-link robot arm. The dashed lines are the desired motions. The arrows indicate the direction in which an unknown external force is applied to the end-effector of the robot arm.

at the end-effector of a manipulator can be transformed into equivalent torques. This causes the end-effector of a robot arm to draw according to the desired trajectory calculating position error. Also, it is shown that ill-posedness of inverse kinematics can be resolved in a natural way without using any artificial optimization criterion [12–14]. However, even in such a method, kinematic configurations including redundant joints may not be guaranteed, even though the posture of a robot arm could be set only within a certain boundary.

From this point of view, it would be advantageous if neural oscillators are hardly coupled to each joint of a robot arm. When the oscillators are implemented to a robotic arm,
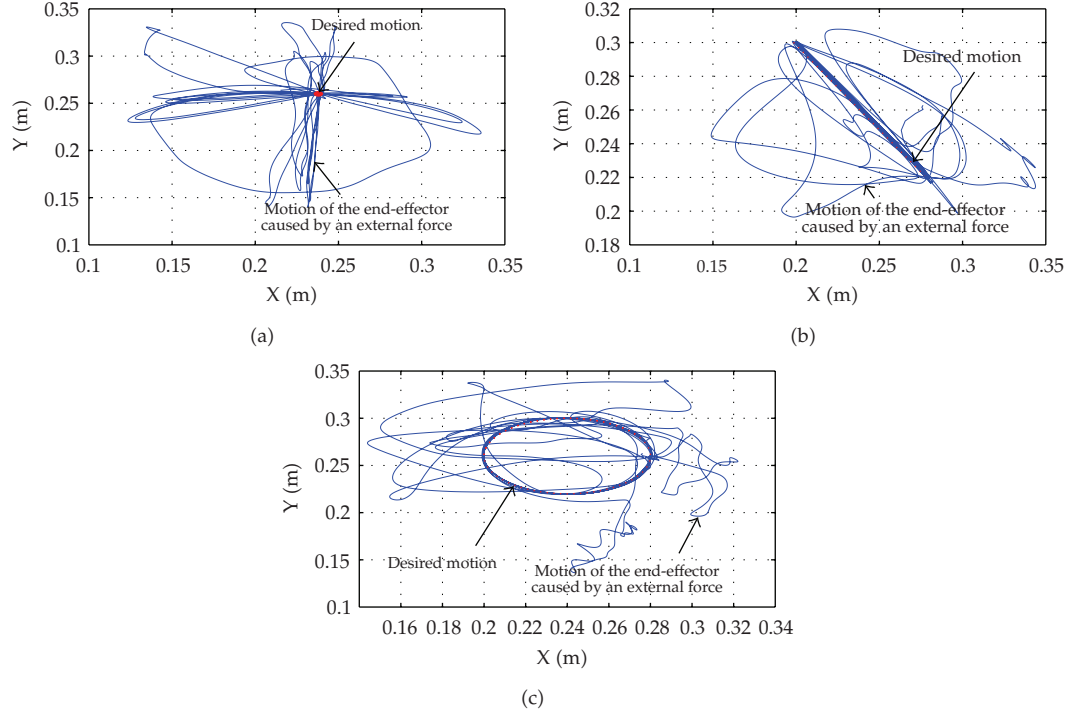
(a)

(b)

(c)

Figure 10: The trajectories drawn by the end-effector of the real robot arm.

they provide a proper motor command considering the movements of the joints with sensory signals. Since biologically inspired motions of each joint as described in Section 2 are attained by entrainment of the neural oscillator, the coupled joint can respond intuitively with the entrainment property according to environmental changes or unknown disturbance inputs performing an objective motion. In addition, each neural oscillator can be tuned in order to give the criterion for limitation of motion within a driving range to the joints considering the amplitude of the sensory feedback signal.

In general, dynamics of a robot system with n DOFs could be expressed as

$$H(q)\ddot{q} + \left\{ \frac{1}{2} \dot{H}(q) + S(q,\dot{q}) \right\} \dot{q} + g(q) = u, \tag{3.1}$$

where H denotes the $n \times n$ inertia matrix of a robot, the second term in the left-hand side of (3.1) stands for coriolis and centrifugal force, and the third term is the gravity effect. Then a control input for a rhythmic motion of the dynamic system shown in (3.1) is introduced as follows:

$$u = -C_0\dot{q} - J^T\left(k\Delta x + \varsigma\sqrt{k}\dot{x}\right) - k_o\Delta q + g(q), \tag{3.2}$$
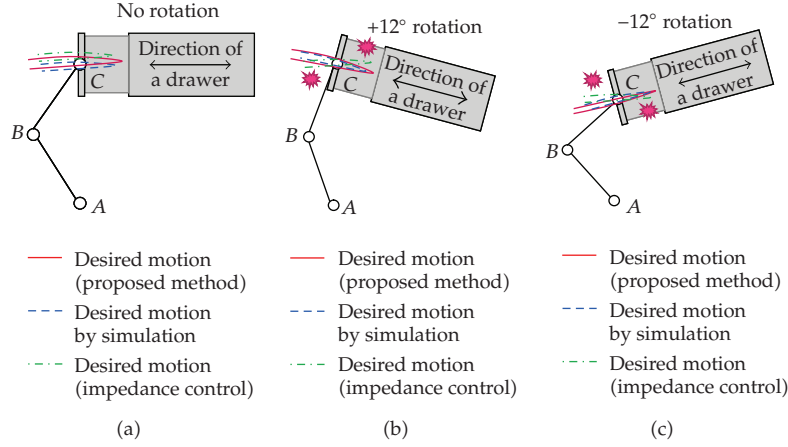
Figure 11: Schematic figure on the experiments that robot arm opens and closes a drawer repeatedly. (a) Fix the drawer in accordance with the robot arm motion, (b) rotate the drawer clockwise about 12°, and (c) rotate the drawer counter-clockwise about 12°.

where

$$C_0 = \text{diag}(c_1, c_2, \ldots, c_n),$$

$$c_i = \varsigma_0 \sqrt{k} \sqrt{\sum_{j=1}^{n} |H_{ij}|}, \quad (i = 1, 2, \ldots, n),$$

$$\Delta x = x - x_d,$$

$$\Delta q = q - q_{\text{od}},$$

(3.3)

where k and ς_0 are the spring stiffness and damping coefficient, respectively, for the virtual components. C_0 is the joint damping. k_o and q_{odi} are the stiffness gain and the output of the neural oscillator that produces rhythmic commands, respectively.

The control inputs as seen in (3.2) consist of two control schemes. One is based on Virtual spring-damper Hypothesis [18, 19] and the other is determined in terms of the output of the neural oscillator as illustrated in (2.2). In the control input of (3.2), the first term describes a joint damping for restraining a certain self-motion which could occurr in a robot system with redundancy, and the second term means PD control in task space by using of Jacobian transpose, and also a spring and a damper in the sense of physics. Appropriate selection of the parameters such as joint damping factors C_0, stiffness k, and damping coefficient ς renders the closed-loop system dynamics convergent, that is, x is converged into x_d and both of $\dot{x}$ and $\dot{q}$ become 0 as time elapses. In general, the neural oscillators coupled to the joints perform the given motion successively interacting with a virtual constraint owing to the entrainment property if gains of the neural oscillator are properly tuned [13, 16]. In the proposed control method, the VFI is considered as a virtual constraint. Also, the coupled model enables a robotic system to naturally exhibit a biologically inspired motion employing sensory signals obtained from each joint under an unpredictable environment change.
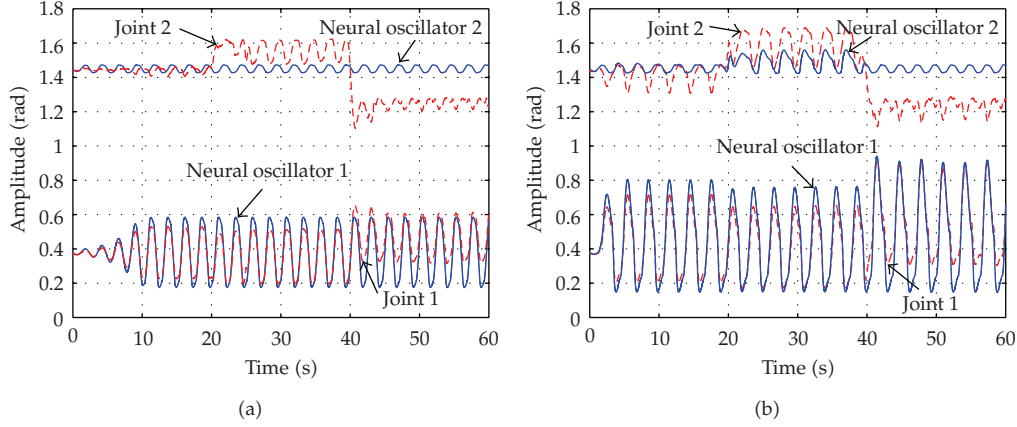
(a)   (b)

Figure 12: The outputs of each joint and neural oscillator as the sensory feedbacks of the neural oscillators are turned off (a) and turned on (b).

4. Experimental Verifications

For considering the possibility of the proposed control scheme described in Section 3, a real robot arm with 6 degrees of freedom (see Figure 7(b)) is employed and a real-time control system is constructed. This arm controller runs at 200 Hz and is connected via IEEE1394 for data transmission at 4 KHz. ATI industrial automation's Mini40 sensor was fitted to the wrist joint of the arm to detect external disturbances. The appropriate parameters in Table 1 were used for the neural oscillator. Also Table 1 illustrates the parameters on the arm dynamics of the real robot. The initial parameters of the neural oscillators were basically selected under the stable oscillation condition of the neural oscillator by investigating the dynamic response of the neural oscillator. The stable oscillation condition of the population of the connected unit neuron is well described in the Matsuoka's works [1, 2]. And then, the amplitude and period of the neural oscillator's output for generating the desired input are determined by controlling the parameter s (tonic input) and T_r/T_a (time constants), respectively. Figure 8 shows the arm kinematics of the real robot arm. Since the desired motions are generated in the horizontal plane, q_1 and q_3 are set to $90°$. The initial values of q_5 and q_6 are set to $0°$, respectively. q_2 and q_4, corresponding to θ_1 and θ_2 in Figure 7(a), respectively, are controlled by the neural oscillators.

Various tasks in cases 1 through 3 of Figures 9(a), 9(b), and 9(c) are verified with respect to adaptive motion of the arm against arbitrary forces. As drawn in Figure 9, the end-effector of the two-link robot arm follows the desired trajectories (dashed lines). Then while the robot arm shows the desired motion, arbitrary external forces are applied to the end-effector of the robotic arm according to the direction of the arrows indicated in Figures 9(a), 9(b), and 9(c). Figures 10(a) to 10(c) indicate experimental results on case 1 to case 3 of Figure 9, respectively. Through these cases, we examine whether various desired motions such as motionless status as well as linear and circular motions can be attained or not. Basically kinds of these motions were verified from the results of Figure 10. In Figure 10, the dotted lines in the center part of the figure show the desired motions and overlapping lines illustrate the motion trajectories that are drawn in terms of the end-effector of the real robot arm. In addition, we pushed and pulled the end-effector along the positive and negative x direction as shown in Figure 10(a). And such conditions were also applied to the robot

Table 1: Parameters of the neural oscillator and robot arm model.

Initial parameters			
Neural oscillator (1)		Neural oscillator (2)	
Inhibitory weight (w_1)	1.7	Inhibitory weight (w_2)	1.7
Time constant (T_{r1})	0.68	Time constant (T_{r2})	0.7
Time constant (T_{a1})	1.36	Time constant (T_{a2})	1.4
Sensory gain (k_1)	3.1	Sensory gain (k_2)	15.6
Tonic input (s_1)	1.0	Tonic input (s_2)	1.0
Robot arm model			
Mass 1 (m_1)	2.347 kg	Mass 2 (m_2)	0.834 kg
Inertia 1 (I_1)	0.0098 kgm^2	Inertia 2 (I_2)	0.0035 kgm^2
Length 1 (l_1)	0.224 m	Length 2 (I_2)	0.225 m

arm along the y direction in order to evaluate an adaptive feature of the proposed control method under additive external disturbances. It can be verified from the experimental result of Figure 10(a) that the robot arm is moved well according to the direction of the applied force (about 10 N and below). If an arbitrary force exists, then it follows that the end-effector of the robot arm shows a compliant motion even in the linear motion and circular motion of the robot arm as seen in Figures 10(b) and 10(c).

The force and torque (F/T) sensor value in the x and y direction are measured and calculated into each joint value. By this, the joint angles are changed according to the direction of the impact of the force induced by the collision, which makes the neural oscillators entrain the joint angles for biologically inspired motion. Hence a change in the output produced intuitively from the neural oscillator causes a change in the joint torque. Finally the joint angles are modified adequately. Thus, it can be confirmed that the proposed neural-oscillator-based robot arm control approach successfully dealt with unexpected collisions sustaining desired motions.

5. Case Study: Opening and Closing a Drawer

5.1. Experimental System

Figure 11 conceptually illustrates the objective tasks with experimental setup for the validation of the proposed control scheme. We evaluate the entrainment capability of the neural oscillator that enables a manipulator to implement and sustain the given task under various environmental changes. Hence, in order to verify the possibility of such adaptation performance, we apply various circumstances to the coupled oscillator-robot arm with the tasks with respect to opening and closing a drawer as seen in Figure 11. We tightly joined the end-effector of the robot arm to the drawer. The end-effector's direction of the robot arm is designed in accordance with the direction to open or close the drawer under the condition that the drawer is not rotated but fixed. In Figures 11(b) and 11(c), the drawer was rotated clockwise and counter-clockwise about 12° for considering unknown environmental changes. Then, the end-effector of the robot arm brings about various collision problems with the drawer due to a different direction between the end-effector of the robot arm and the drawer. Now, we will examine what happens in the arm motion on performing the objective task if additive external disturbances exist.
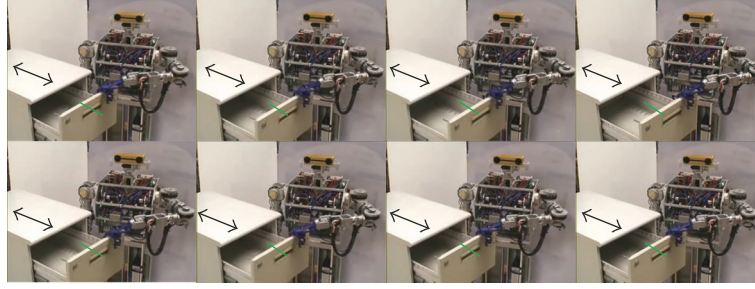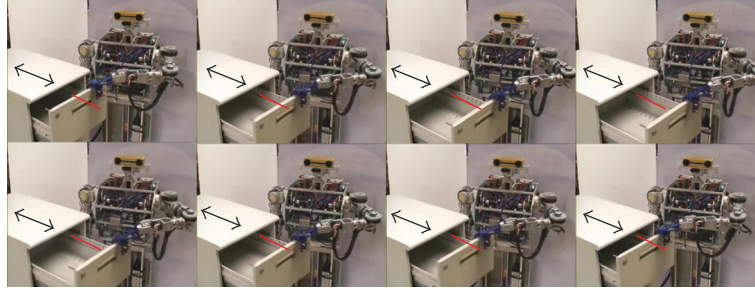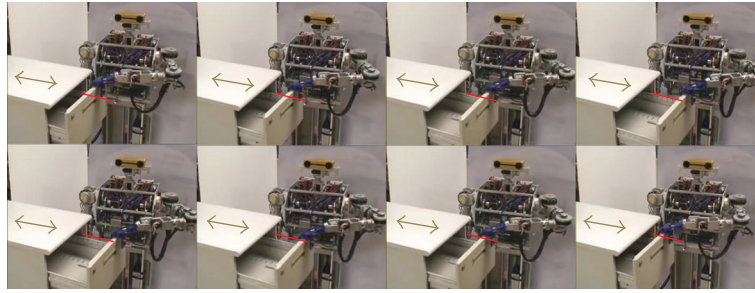
(a)
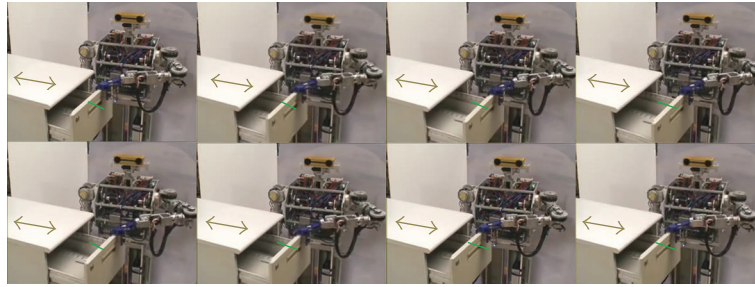
(b)
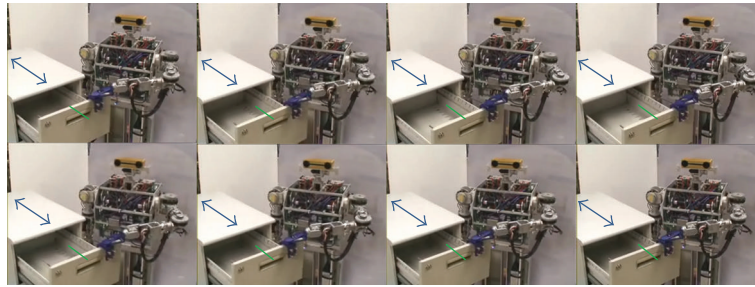
(c)

Figure 13: Snap shots of the robot arm motions when sensory information is not fed again in cases of $0°$ (a), $-12°$ (b) and $12°$ (c) rotation of the drawer.

5.2. Experimental Results

Figures 12(a) and 12(b) illustrate the experimental results on each joint output of the robot arm as the sensory feedback of the neural oscillator is turned off and on, respectively. In the first case of Figure 11(a), the desired motion of robot arm is not changed owing that the drawer is immovable during 0 s to 20 s. The first joint (q_2) and the second one (q_4) are actuated to move to the distance corresponding to an external force as explained in above Section 3. Hence, if the drawer rotates about $+12°$ and $-12°$ as illustrated in Figures 11(b) and 11(c) during 20 s to 40 s and 40 s to 60 s, then the robot arm's motion is autonomously altered. In Figures 12(a) and 12(b), the blue lines indicate the desired trajectories produced by means of the neural oscillators for the joints 1 and 2. The red dotted lines are the output of the joints

(a)

(b)
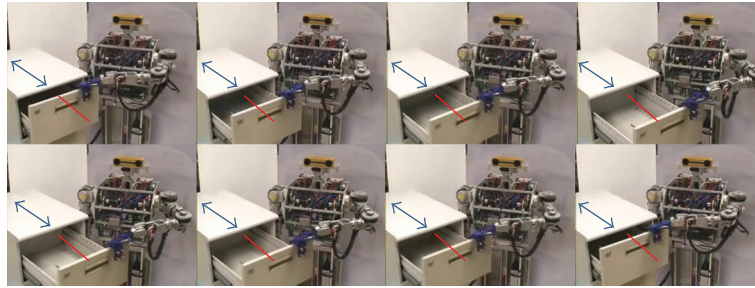
(c)

Figure 14: Snapshots of the robot arm motions when sensory information is fed again in cases of 0° (a), −12° (b), and 12° (c) rotations of the drawer.

1 and 2 that is changed in terms of forces applied when the drawer is rotated. Comparing the result of Figure 12(a) with the neural-oscillator-based control (see Figure 12(b)) if the sensory information is fed again, it can be observed that the outputs of each joint and neural oscillator are changed whenever unknown disturbances are induced into the robot arm. Such effect could be accomplished owing that the oscillator-based control reproduces the desired joint input entraining the joint motion coupled with the neural oscillator through sensory feedback. In Figure 12(a), the output of the neural oscillator sustains a certain oscillation because the sensory feedbacks were turned off and the joint desired input is slightly changed in terms of the torque sensor feedback in comparison with Figure 12(b). Figure 13 shows the snap shots of the robot arm controlled when sensory information is not fed again. The snap shots in Figure 14 show the motion of the robot arm implementing the proposed control
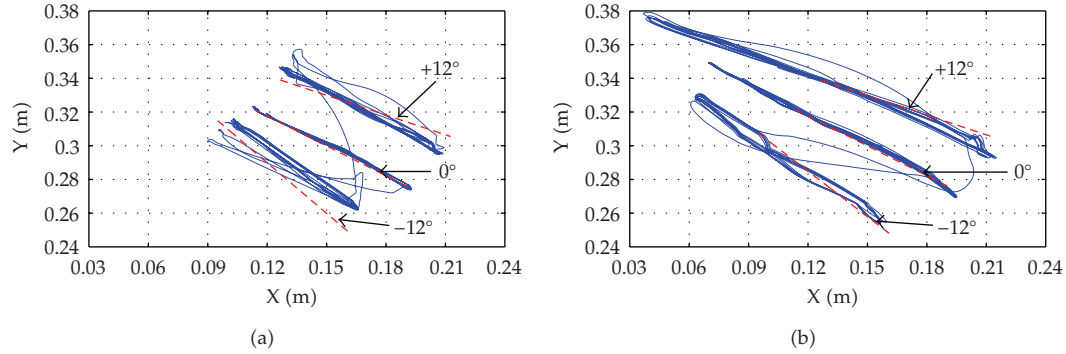
Figure 15: Trajectories of the end-effector of the robot arm in case that the sensory feedbacks are turned off (a) and turned on (b).

approach based on the neural oscillator, where we can observe that the end-effector traces the rotated drawer direction.

As shown in Figures 15(a) and 15(b), the end-effector of the robot arm draws the trajectories corresponding to the desired motion for opening and closing the drawer. The straight dotted lines indicate the desired trajectories of the robot arm generated by simulation. The blue lines show the trajectories measured at the end-effector of the robot arm in experiments related with Figures 13 and 14. In Figure 15(a), movements of the robot arm are identical with the expected performance although there are inefficient motions due to unknown disturbances. This is because the desired input of each joint is modified adequately by the impedance model measuring external forces with the F/T sensor even though sensory information of the neural oscillators is not fed again. In comparison with this, the individual trajectories drawn by the robot arm in Figure 15(b) are completely consistent with the rotated direction of the drawer. Thus, the robot arm coupled with the neural oscillator exhibits the superior potential in adaptive motion exploiting the sensory feedback of the neural oscillator for the capability of entrainment. From experiment results of Figures 12(b), 14, and 15(b), the measured trajectories and movements of real robot arm imply that the neural oscillator enables the robot arm to exhibit the self-adapting motion to enhance adaptive motion sustaining the objective task and motion stability.

6. Conclusion

We have presented a control scheme for technically achieving a biologically inspired self-adapting robotic motion. In contrast to existing works that were only capable of rhythmic pattern generation for simple tasks, our approach allowed the robot arm to precisely trace a trajectory correctly through entrainment. With this, the proposed method is verified through more complex behaviors of the real robot arm under unknown environmental changes. Also our approach causes appropriate desired motions irrespective of precisely modelling with respect to external disturbances. For such reason, it was observed from experimental results that the novel adaptive motions corresponding to an external force appear clearly. This approach will be extended to a more complex task toward the realization of biologically inspired robot control architectures.

References

[1] K. Matsuoka, "Sustained oscillations generated by mutually inhibiting neurons with adaptation," *Biological Cybernetics*, vol. 52, no. 6, pp. 367–376, 1985.

[2] K. Matsuoka, "Mechanisms of frequency and pattern control in the neural rhythm generators," *Biological Cybernetics*, vol. 56, no. 5-6, pp. 345–353, 1987.

[3] G. Taga, Y. Yamaguchi, and H. Shimizu, "Self-organized control of bipedal locomotion by neural oscillators in unpredictable environment," *Biological Cybernetics*, vol. 65, no. 3, pp. 147–159, 1991.

[4] G. Taga, "A model of the neuro-musculo-skeletal system for human locomotion. I. Emergence of basic gait," *Biological Cybernetics*, vol. 73, no. 2, pp. 97–111, 1995.

[5] G. Taga, "A model of the neuro-musculo-skeletal system for human locomotion. II. Real-time adaptability under various constraints," *Biological Cybernetics*, vol. 73, no. 2, pp. 113–121, 1995.

[6] S. Miyakoshi, G. Taga, Y. Kuniyoshi, and A. Nagakubo, "Three-dimensional bipedal stepping motion using neural oscillators-towards humanoid motion in the real world," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 1, pp. 84–89, October 1998.

[7] Y. Fukuoka, H. Kimura, and A. H. Cohen, "Adaptive dynamic walking of a quadruped robot on irregular terrain based on biological concepts," *International Journal of Robotics Research*, vol. 22, no. 3-4, pp. 187–202, 2003.

[8] G. Endo, J. Nakanishi, J. Morimoto, and G. Cheng, "Experimental studies of a neural oscillator for biped locomotion with QRIO," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 596–602, August 2005.

[9] M. M. Williamson, "Postural primitives: interactive behavior for a humanoid robot arm," in *Proceedings of the 4th International Conference on Simulation of Adaptive Behavior*, pp. 124–131, MIT Press, 1996.

[10] M. M. Williamson, "Rhythmic robot arm control using oscillators," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, vol. 1, pp. 77–83, October 1998.

[11] A. M. Arsenio, "Tuning of neural oscillators for the design of rhythmic motions," in *Proceedings of the IEEE International Conference on Robotics and Automation*, vol. 2, pp. 1888–1893, April 2000.

[12] W. Yang, N. Y. Chong, C. Kim, and B. J. You, "Optimizing neural oscillators for rhythmic movement control," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, pp. 807–814, August 2007.

[13] W. Yang, N. Y. Chong, J. Kwon, and B. J. You, "Self-sustaining rhythmic arm motions using neural oscillators," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '08)*, pp. 3585–3590, Nice, France, September 2008.

[14] W. Yang, N. Y. Chong, C. Kim, and B. J. You, "Entrainment-enhanced neural oscillator for rhythmic motion control," *Intelligent Service Robotics*, vol. 1, no. 4, pp. 303–311, 2008.

[15] W. Yang, N. Y. Chong, S. Ra, C. H. Kim, and B. J. You, "Self-stabilizing bipedal locomotion employing neural oscillators," in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, pp. 8–15, December 2008.

[16] W. Yang, J. Kwon, J.-H. Bae, N. Y. Chong, and B. J. You, "Biologically inspired self-adapting motion control employing neural oscillator," in *Proceeding of the IEEE International Symposium on Industrial Electronics*, pp. 161–168, July 2009.

[17] J.-J. E. Slotine and W. Li, *Applied Nonlinear Control*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1991.

[18] S. Arimoto, M. Sekimoto, H. Hashiguchi, and R. Ozawa, "Natural resolution of ill-posedness of inverse kinematics for redundant robots: a challenge to bernstein's degrees-of freedom problem," *Advanced Robotics*, vol. 19, no. 4, pp. 401–434, 2005.

[19] S. Arimoto, M. Sekimoto, J.-H. Bae, and H. Hashiguchi, "Three-dimensional multi-joint reaching under redundancy of DOFs," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1898–1904, 2005.

Research Article

Shape Sensitivity Analysis in Flow Models Using a Finite-Difference Approach

Imran Akhtar,[1] Jeff Borggaard,[1] and Alexander Hay[1, 2]

[1] *Interdisciplinary Center for Applied Mathematics, Virginia Tech, Blacksburg, VA 24061, USA*
[2] *Industrial Materials Institute, National Research Council, Boucherville, Qc, Canada J4B6Y4*

Correspondence should be addressed to Imran Akhtar, akhtar@vt.edu

Reduced-order models have a number of practical engineering applications for unsteady flows that require either low-dimensional approximations for analysis and control or repeated simulation over a range of parameter values. The standard method for building reduced-order models uses the proper orthogonal decomposition (POD) and Galerkin projection. However, this standard method may be inaccurate when used "off-design" (at parameter values not used to generate the POD). This phenomena is exaggerated when parameter values describe the shape of the flow domain since slight changes in shape can have a significant influence on the flow field. In this paper, we investigate the use of POD sensitivity vectors to improve the accuracy and dynamical system properties of the reduced-order models to problems with shape parameters. To carry out this study, we consider flows past an elliptic cylinder with varying thickness ratios. Shape sensitivities (derivatives of flow variables with respect to thickness ratio) computed by finite-difference approximations are used to compute the POD sensitivity vectors. Numerical studies test the accuracy of the new bases to represent flow solutions over a range of parameter values.

1. Introduction

Reduced-order modeling of incompressible flows plays an important role in academic and industrial research. In order to reduce the complexity of the governing equation, reduced-order models are often developed to represent the dynamical system with a few degrees of freedom. These models can provide analytical insight into the physical phenomena and enable application of dynamical systems theory and control methods.

Since Roshko [1] measured the vortex shedding period behind a bluff body, many researchers have investigated this phenomenon experimentally and numerically for a wide range of Reynolds numbers. The most frequently investigated bluff geometry is the circular

cylinder. The flow behind a circular cylinder has become the canonical problem for studying such external separated flows [2–7]. Bishop and Hassan [2] were among the first to suggest a *phenomenological* model reduction by proposing Rayleigh/van der Pol self-excited oscillators to represent the forces over a cylinder due to vortex shedding. Several other analytical models have also been proposed along similar lines for stationary or moving cylinders [8–13].

On the other hand, engineering applications often involve flows over complex bodies, such as wings, submarines, missiles, and rotor blades, which can hardly be modeled as a circular cylinder. It would however be interesting to broaden the application of reduced-order models to other complex geometries. A systematic approach would be to start by studying a configuration that is more general than a circular cylinder and can characterize typical engineering flow configurations. Elliptic cylinders seem to provide such a configuration and changes in the eccentricity allow for shapes ranging from a circular cylinder to a flat plate. By tuning the damping coefficients, self-excited oscillator models [14] can be extended to elliptic cylinders.

In some cases, the reduced-order models represent the space variation with a limited number of basis functions, while still capturing the physics and dominant features of the flow. Most widely used reduced-order model techniques in fluid dynamics are derived from the proper orthogonal decomposition- (POD-) Galerkin projection approach [15, 16]. The POD provides a tool to formulate an optimal basis (or modes) to represent given trajectories of a dynamical system. The most common approach is to compute a "representative" collection of states of the system (commonly referred to as *snapshots*) through numerical simulations or experiments, then use the POD to find a low-dimensional set of basis functions [16, 17]. The POD processes time snapshot data to find the *best* basis functions on which to represent this data. Later, a low-dimensional dynamical system is typically obtained by Galerkin projection. These models give insight to the flow physics, often reproduce the data, and may be used for control purposes. POD-based reduced-order models have been successfully applied to model vortex shedding past a cylinder [18–23]. In fact, the flow past a cylinder has become a benchmark problem for reduced-order modeling.

Despite the accurate reproduction of the data from which it is originated, the POD-based reduced-order model lacks robustness away from the *reference* simulations. This is a serious limitation of the POD-based reduced-order models since most applications use reduced-order models in a *predictive* setting. In [24], the authors noted that *the accuracy of the model predictions rapidly deteriorates as we move away from the decomposition value*. Thus, the POD basis generated from an experiment or a numerical simulation is only useful within a narrow bandwidth of parameters close to that *reference*. In addition to the accuracy of the model away from the *reference*, numerical stability of these models for long-time integration is also an issue [23, 25–30].

Various methods have been proposed to extend the applicability and increase the accuracy of POD-based reduced-order models away from the *reference* data. Ma and Karniadakis [19] investigated the stability and dynamics of three-dimensional limit-cycle states near the Mode A [6] instability for flow past a cylinder using a low-dimensional modeling. They employed two procedures to construct this *hybrid* system, either by concatenating the two sets of snapshots at Re = 182 and Re = 185 or by extracting the POD modes at each state and orthonormalizing the entire set. They concluded that reference data must contain sufficient information about the flow field to predict any qualitative change within that flow regime. Noack et al. [20] suggested a generalization for POD-based Galerkin models to include the transient behavior. In their Galerkin approximation, they included an additional vector and termed it a *shift-mode* (also known as a centering trajectory

of a mean-field correction). They obtained the steady solution with a Newton iteration, employing the discretized steady Navier-Stokes equations. The shift mode represents the shift of the short term averaged flow away from the POD space and is effectively a normalized mean-field correction. Within the control setting of POD-based reduced-order models, a similar strategy is required in cylinder wake problems to make the control effective.

In the application of reduced-order models to optimal control, one desires a model that is accurate over a wide range of actuation data. In a recent study, Graham et al. [21] used reduced-order models to develop optimal control strategies for flow past a circular cylinder at Re = 100 through cylinder rotation. When the POD modes were generated from the snapshots at a given oscillation frequency, the active control model only worked near that particular frequency. Therefore, they incorporated multiple frequency information in their snapshot data set using a varying frequency sinusoid, or *chirp*. With this approach, one may require a large parametric data set as the reference data to broaden the applicability of POD-based reduced-order models over a wide range of parameter values.

In our previous work [31], we proposed the inclusion of parametric derivatives of the POD basis functions (computed using sensitivity analysis). The study considered two approaches for using the POD derivatives to define a reduced basis. One approach used linear extrapolation of the standard POD basis to the parameter value of interest (*extrapolated basis*). A second approach combined with the POD and POD derivatives to construct the basis (*expanded basis*). Both approaches provided promising results in the study of flow past a square cylinder where the viscosity (equivalently, the Reynolds number) was varied from the reference value.

A follow-up study [32] considered the development of reduced-order models for flows where the parameter described the orientation of the square cylinder relative to the free-stream flow direction. We demonstrated our shape sensitivity analysis approach by rotating the square cylinder. This problem also featured a "hidden" parametric dependence on the Reynolds number as the characteristic length scale changes with the orientation of the square cylinder to the mean-stream flow.

In this study, we use a similar basis improvement methodology. However, we consider modeling the flow as the obstacle is changed from a circular cylinder to an elliptic cylinder. The shape parameter is chosen as the thickness ratio (τ) between the minor axis to the major axis. It is important to note that the Reynolds number is defined using the projected area of the cylinder, as seen by the flow, and is kept constant in our analysis. Moreover, we use finite differences to compute the shape sensitivity as the ellipse parameter τ is varied. We employ an O-grid for the cylinders such that the outer domain is circular in all cases. Since the outer domains of both the circular and elliptic cylinders are the same, the shape sensitivity approaches to zero as it moves away from the body. In this regard, the finite-difference-based methodology provides a quick and easier way to compute the shape sensitivity and extend the application of POD basis to actual shape deformation. This study can find its application in the aerospace industry where the wing cross-section changes its actual shape, such as a morphing wing. This study also provides a validation to the basis selection approach used by Ma and Karniadakis [19]. In their study, they built a POD basis using two simulations with similar parameters. Their approach produces a similar basis as we would obtain with our expanded approach.

The paper is organized as follows. Section 2 explains the numerical methodology to solve the Navier-Stokes equations and simulate the flow past circular and elliptic cylinders. In Section 3, we present the POD-based reduced-order model procedure after collecting
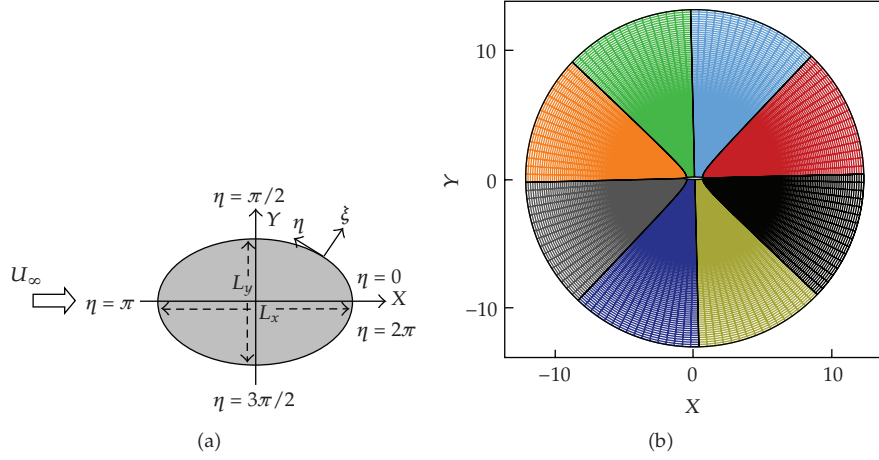
Figure 1: (a) Geometry of an elliptic cylinder and (b) "O" grid distributed among 8 processors in the η direction for $\tau = 2.0$.

the *snapshots* of the flow field. In Section 4 we present the sensitivity analysis procedure and numerical results are discussed in Section 5.

2. Numerical Methodology

The Navier-Stokes and continuity equations are the governing equations for the present problem. For incompressible flow, they can be represented as follows:

$$\frac{\partial u_j}{\partial x_j} = 0, \tag{2.1}$$

$$\frac{\partial u_i}{\partial t} + \frac{\partial}{\partial x_j}(u_j u_i) = -\frac{1}{\rho}\frac{\partial p}{\partial x_i} + \mu \frac{\partial^2 u_i}{\partial x_j \partial x_j}, \tag{2.2}$$

where $i, j = 1, 2$ for two-dimensional flows; the u_i represent the Cartesian velocity components (u, v); p is the pressure; ρ is the fluid density; μ is the fluid viscosity.

We employ curvilinear coordinates in an Eulerian reference frame; a planar view is shown in Figure 1(a). Here, X and Y indicate the Cartesian coordinates, ξ and η represent a curvilinear coordinate system, and L_x and L_y are the major and minor axes, respectively. The thickness ratio of the ellipse is given by $\tau = L_x/L_y$ and is the parameter used in our parametric modeling study.

The Navier-Stokes equations are written in curvilinear coordinates and strong conservation form as follows:

$$\frac{\partial U_m}{\partial \xi_m} = 0, \tag{2.3}$$

$$\frac{\partial (J^{-1} u_i)}{\partial t} + \frac{\partial F_{im}}{\partial \xi_m} = 0, \tag{2.4}$$

where the flux is defined as

$$F_{im} = U_m u_i + J^{-1} \frac{\partial \xi_m}{\partial x_i} p - \frac{1}{\text{Re}} G^{mn} \frac{\partial u_i}{\partial \xi_n}, \tag{2.5}$$

where $J^{-1} = \det(\partial x_i / \partial \xi_j)$ is the inverse of the Jacobian or the volume of the cell, $U_m = J^{-1}(\partial \xi_m / \partial x_j) u_j$ is the volume flux (contravariant velocity multiplied by J^{-1}) normal to the surface of constant ξ_m, and $G^{mn} = J^{-1}(\partial \xi_m / \partial x_j)(\partial \xi_n / \partial x_j)$ is the "mesh skewness tensor." Equations (2.3) and (2.4) are nondimensionalized using the projected area L_y of the ellipse as the length scale and the freestream velocity (U_∞) as the velocity scale. Thus, the Reynolds number is given by $\text{Re} = L_y U_\infty / \nu$, where $\nu = \mu / \rho$. All of these parameters remain fixed in our computations.

In this study, a body conformal O-type grid is employed to simulate the flow over a body as shown in Figure 1(b). The grid is decomposed into different processors to achieve parallel performance. A non-staggered-grid layout is employed to solve the transformed Navier-Stokes equations. The Cartesian velocity components (u, v) and pressure (p) are defined at the center of the control volume in the computational space, and the volume fluxes (U, V) are defined at the mid points of control volume. All of the spatial derivatives are approximated with second-order accurate central differences except for the convective terms. Using the same central differencing for the convection terms may lead to spurious oscillations in the coarser regions of the grid, thereby leading to erroneous results. In the present formulation, we discretize the convective terms using a variation of QUICK [33]; that is, we calculate the face values of the velocity variables (u_i) from the nodal values using quadratic upwinding interpolation. The upwinding of QUICK is carried out by computing the positive and negative volume fluxes $((U_m + |U_m|)/2)$ and $((U_m - |U_m|)/2)$, respectively, and using the generic stencil.

A semi-implicit scheme is employed to advance the solution in time. The diagonal viscous terms are advanced implicitly using the second-order accurate Crank-Nicolson method; whereas all of the other terms are advanced using the second-order accurate Adams-Bashforth method. In the present formulation, we apply a fractional-step method to advance the solution in time. The fractional-step method splits the momentum equation into

(a) an advection-diffusion equation—the momentum equation solved without the pressure term,

(b) a pressure Poisson equation—constructed by implicit coupling between the continuity equation and the pressure in the momentum equation, thus satisfying the constraint of mass conservation.

The governing equations are solved using a methodology similar to that employed by Zang et al. [34]. However, the algorithm is extended to parallel computing platforms and the 1D domain decomposition technique (for two-dimensional flows) is employed to distribute the problem among different processors. Details of discretization, parallel implementation, validation, and verification can be found in [35].

The CFD results for the numerical methodology have been validated for three-dimensional [35] and two-dimensional [23] flows by comparing them with published experimental and numerical studies. We have also performed grid and domain dependence studies to verify the numerical results for circular and elliptic cylinders [14].

3. POD-Galerkin Reduced-Order Model

3.1. The Proper Orthogonal Decomposition

Mathematically, we compute Φ for which the following quantity is maximum:

$$\frac{\left\langle |(\mathbf{u}, \Phi)|^2 \right\rangle}{\|\Phi\|^2},$$ (3.1)

where $\langle \cdot \rangle$ denotes either an ensemble or a time average. Applying variational calculus, one can show that (3.1) is equivalent to a Fredholm integral eigenvalue problem represented as

$$\int_\Omega R_{ij}(\mathbf{x}, \mathbf{x}')\Phi^j(\mathbf{x}')\mathrm{d}\mathbf{x}' = \lambda\Phi^j(\mathbf{x}),$$ (3.2)

where i, j are the number of velocity components and $R_{ij}(\mathbf{x}, \mathbf{x}')$ is the two-point spatial autocorrelation tensor.

In the classical POD or direct method, originally introduced by Bakewell and Lumley [36], R_{ij} is a two-point spatial-correlation tensor and the eigenfunctions are the POD modes. In this approach, the averaging operator is performed in time. On the other hand, if the averaging operator is evaluated as a spatial average over the domain of interest, the method is known as the method of *snapshots* [17]. In this approach, we formulate a temporal-correlation function from the snapshots and transform it into an eigenvalue problem as follows:

$$C_{ij} = \left\langle \left(\mathbf{u}^i, \mathbf{u}^j\right) \right\rangle,$$ (3.3)

where $(a, b) = \int_\Omega a \cdot b \, \mathrm{d}\Omega$ represents the inner product between a and b. The POD modes are then computed by solving the eigenvalue problem

$$\mathbf{CQ} = \mathbf{Q}\lambda,$$ (3.4)

where $\mathbf{Q}$ and λ are the eigenvectors and eigenvalues, respectively. Since $\mathbf{C}$ is nonnegative Hermitian, $\mathbf{Q}$ is orthogonal by definition. The POD modes are computed as follows:

$$\Phi_i = \frac{1}{\sqrt{\lambda_i}}\mathcal{W}Q_i.$$ (3.5)

An important characteristic of these modes is orthogonality; that is, $\Phi_i \cdot \Phi_j = \delta_{ij}$, where δ_{ij} is the kronecker delta. The optimality of the POD modes lies in capturing the greatest possible fraction of the total kinetic energy for a projection onto the given set of modes.

The flow data or *snapshots* of the steady-state velocity field are sampled with a constant time interval (ΔT_s). The velocity field data (u, v) are assembled in a matrix $\mathcal{W}_{2N \times S}$ as follows:

$$\mathcal{W} = \begin{bmatrix} u_1^{(1)} & u_1^{(2)} & \cdots & u_1^{(S)} \\ \vdots & \vdots & & \vdots \\ u_N^{(1)} & u_N^{(2)} & \cdots & u_N^{(S)} \\ v_1^{(1)} & v_1^{(2)} & \cdots & v_1^{(S)} \\ \vdots & \vdots & & \vdots \\ v_N^{(1)} & v_N^{(2)} & \cdots & v_N^{(S)} \end{bmatrix}. \tag{3.6}$$

Each column represents one time instant or a *snapshot* and S is the total number of snapshots for N grid points in the domain. The vorticity field can also be used for POD; however, in the case of the velocity field, the eigenvalues of $\mathcal{W}$ are a direct measure of the kinetic energy in each mode. Deane et al. [24] observed that 20 snapshots are sufficient for the construction of the first eight eigenfunctions at Re = 100–200. In general, numerical studies [37] suggest that the first M POD modes, where M is even, resolve the first $M/2$ temporal harmonics and require $2M$ number of snapshots for convergence.

We write the velocity field as the sum of the mean flow $(\overline{\mathbf{u}})$ and the velocity fluctuations $(\mathbf{u}')$. The mean flow $\overline{\mathbf{u}} = \langle \mathbf{u} \rangle$, where $\langle \cdot \rangle$ is the time average of the assembled data, is subtracted from $\mathcal{W}$. Then, the fluctuations are expanded in terms of the POD eigenfunctions (Φ_i) as follows:

$$\mathbf{u}(\mathbf{x}, t) \approx \overline{\mathbf{u}}(\mathbf{x}) + \sum_{i=1}^{M} q_i(t) \Phi_i(\mathbf{x}), \tag{3.7}$$

where M is the number of POD modes used in the projection. The singular value decomposition (SVD) of this matrix provides the divergence-free velocity POD modes (Φ_i).

3.2. Galerkin Projection

Substituting (3.7) in (2.4) and projecting this equation along the elements of the M-dimensional POD basis give

$$\dot{q}_k(t) = \mathcal{A}_k + \sum_{m=1}^{M} \mathcal{B}_{km} q_m(t) + \sum_{m=1}^{M} \sum_{n=1}^{M} \mathcal{C}_{kmn} q_n(t) q_m(t), \tag{3.8}$$

where

$$\mathscr{A}_k = \frac{1}{\mathrm{Re}_{L_y}}\left(\Phi_k, \nabla^2\overline{\mathbf{u}}\right) - (\Phi_k, \overline{\mathbf{u}} \cdot \nabla\overline{\mathbf{u}}),$$

$$\mathscr{B}_{km} = -(\Phi_k, \overline{\mathbf{u}} \cdot \nabla\Phi_m) - (\Phi_k, \Phi_m \cdot \nabla\overline{\mathbf{u}}) + \frac{1}{\mathrm{Re}_{L_y}}\left(\Phi_k, \nabla^2\Phi_m\right), \qquad (3.9)$$

$$\mathscr{C}_{kmn} = -(\Phi_k, \Phi_m \cdot \nabla\Phi_n).$$

In (3.8), $\mathscr{A}$ is an $M \times 1$ vector resulting from the average flow field, $\mathscr{B}$ is an $M \times M$ matrix corresponding to the linear part, and $\mathscr{C}$ is a tensor that represents the quadratic nonlinearity of the dynamical system. From linear stability analysis, we can compute the eigenvalues of $\mathscr{B}$ and hence ascertain the stability of the trivial solution. Details of the projection for each term are provided in the appendix.

3.3. Numerical Simulations and Reduced-Order Model

We perform numerical simulations of the flows past elliptic cylinders with thickness ratios ranging from $\tau = 0.5$ to $\tau = 1.5$. The flow is simulated over a 192×256 ($\xi \times \eta$) grid with an outer domain size of $30L_y$. In the present case, the grid is divided into eight domains/processors, such that each processor has a load of 192×32 grid points, as shown in Figure 1(b). All simulations are performed at Re = 200. It is important to note that the Reynolds number is defined with the characteristic length of $L_y = 1$ and freestream velocity $U_\infty = 1$. The projected area L_y of the ellipse, as seen by the flow, is fixed for all values of τ. Thus, the nondimensionalization is fixed for all shape variations.

We develop the reduced-order model for the flow past a circular cylinder (as an example) and use a similar procedure for the model reduction of the flow past elliptic cylinders with different thickness ratios. In the current study, we took 64 snapshots of the flow field over one vortex shedding cycle of the circular cylinder. The flow data or *snapshots* of the velocity and pressure fields are sampled in the steady-state of the dynamical system with a constant time interval (ΔT_s). The velocity field (u, v) is stored in $\mathcal{W}$ after subtracting the mean flow. We compute the POD modes from respective snapshot data. In Figure 2, we plot the first six POD modes of u-velocity for the circular cylinder ($\tau = 1$). Projecting the Navier-Stokes equations onto these modes, we develop a reduced-order model and compute $\mathscr{A}, \mathscr{B}$, and $\mathscr{C}$ as given in (3.8). We integrate (3.8) to compute the $q_i(t)$ and plot them in Figure 4. The results we obtained from the reduced-order model match well with the CFD numerical results found in [23].

4. Sensitivity Analysis

In this section, we derive the first-order total derivatives of the POD modes with respect to a generic shape parameter α. We refer to these as the Lagrangian sensitivities of the POD vectors. In our previous study [31], we presented POD sensitivities with respect to a flow parameter, such as the fluid viscosity and the geometry of the structure was fixed. In other words, the sensitivities were computed using the same domain and no grid deformation was required. In this study, the flow domain changes and this leads to a corresponding change in
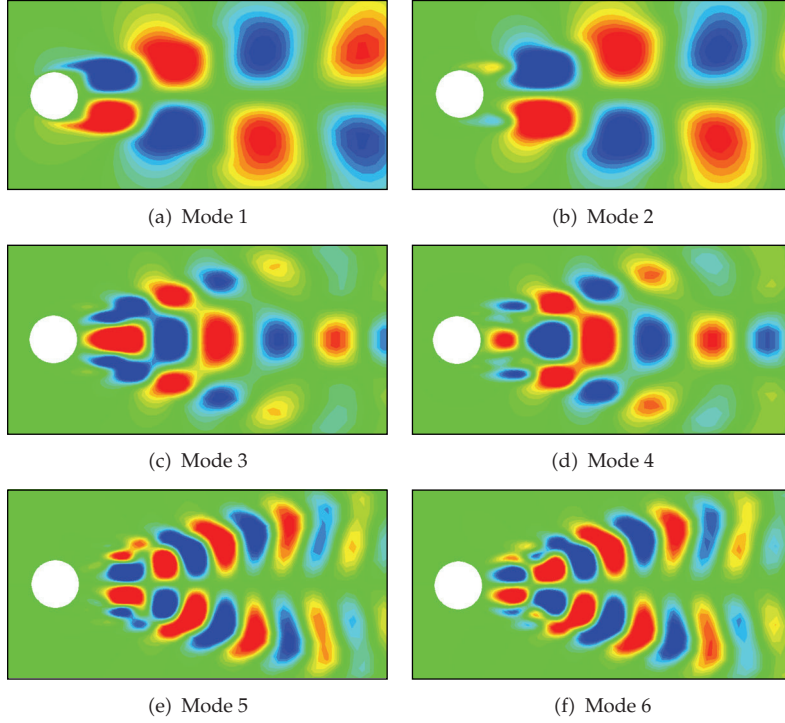
(a) Mode 1

(b) Mode 2

(c) Mode 3

(d) Mode 4

(e) Mode 5

(f) Mode 6

Figure 2: Streamwise POD modes for $\tau = 1$.

the grid with the elliptic parameter τ. It is important to note that the outer domain of for all the cases is fixed with radius $30L_y$.

The sensitivities with respect to the *shape* parameter α (for our case) are computed by a second-order centered finite-difference (FD) approximation

$$\left.\frac{D\Phi_i}{D\alpha}(\mathbf{x}(\alpha_0);\alpha_0)\right|_{\text{FD}} = \frac{\Phi_i(\mathbf{x}(\alpha_0 + \Delta\alpha);\alpha_0 + \Delta\alpha) - \Phi_i(\mathbf{x}(\alpha_0 - \Delta\alpha);\alpha_0 - \Delta\alpha)}{2\Delta\alpha}, \tag{4.1}$$

where α_0 is the parameter value at which the sensitivities are computed and $\Delta\alpha$ is the step size in the finite-difference scheme. The notation $D \cdot /D\alpha$ represents the total derivative with respect to α. The parameter increment $\Delta\alpha$ is chosen sufficiently small for the FD computation to be accurate and sufficiently large for the difference between the two nearby POD vectors to be at least one order of magnitude larger than the discretization error. It is also important to note that since the outer domain is the same for all elliptic cylinders and the flow is nearly freestream, these sensitivities become smaller as we move away from the cylinder and approach zero close to the outer domain of the cylinder.

The traditional approach in reduced-order modeling is to build the POD basis for one particular value of the parameter of the system. This will be referred to as the baseline approach and denoted by α_0, which refers to the baseline state, solution, and POD basis. We aim at producing reduced-order solutions at perturbed states for $\alpha = \alpha_0 + \Delta\alpha$. As in our previous study [31], we define various basis elements used in our analysis.
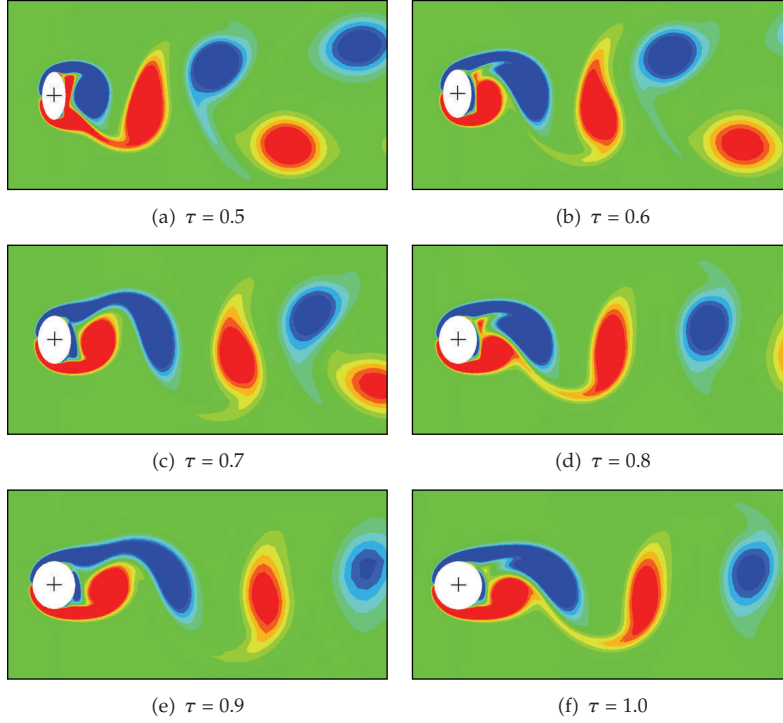
(a) $\tau = 0.5$

(b) $\tau = 0.6$

(c) $\tau = 0.7$

(d) $\tau = 0.8$

(e) $\tau = 0.9$

(f) $\tau = 1.0$

Figure 3: Instantaneous spanwise vorticity contours.

(i) *Baseline POD Basis* (BL). This is the classical approach where the POD basis, built from the data at the baseline value α_0, is used in (3.7) to subsequently produce a reduced-order model at the perturbed value α. These spatial modes are only available on the baseline geometry but they can easily be mapped onto the perturbed geometry.

(ii) *Perturbed POD Basis* (PR). The reduced-order model is constructed using the POD modes extracted from the solution data obtained by a full-order simulation at the perturbed value α. This is a costly approach since each new reduced-order simulation requires a full-order data at the perturbed parameter value. Thus, it has little interest in practice but will be used in the remainder of this study as the benchmark low-dimensional solution.

Following the previous studies for flow parameters in [31] and shape parameters in [32], we examine two different ideas for constructing improved reduced-order bases using the POD modes as well as the Lagrangian sensitivity of the POD modes at the baseline value α_0.

(i) *Extrapolated Basis* (ET). We treat each POD mode as a function of both space and parameter α: $\Phi_i = \Phi_i(\mathbf{x}; \alpha)$. A change $\Delta\alpha$ in the parameter from its baseline value α_0 is reflected in the modes through a first-order expansion in the parameter space:

$$\Phi_i(\mathbf{x}; \alpha) = \Phi_i(\mathbf{x}(\alpha_0); \alpha_0) + \Delta\alpha \frac{D\Phi_i}{D\alpha}(\mathbf{x}(\alpha_0); \alpha_0) + O\left(\Delta\alpha^2\right). \tag{4.2}$$

(a) q_1 (solid) and q_2 (dashed)

(b) q_3 (solid) and q_4 (dashed)

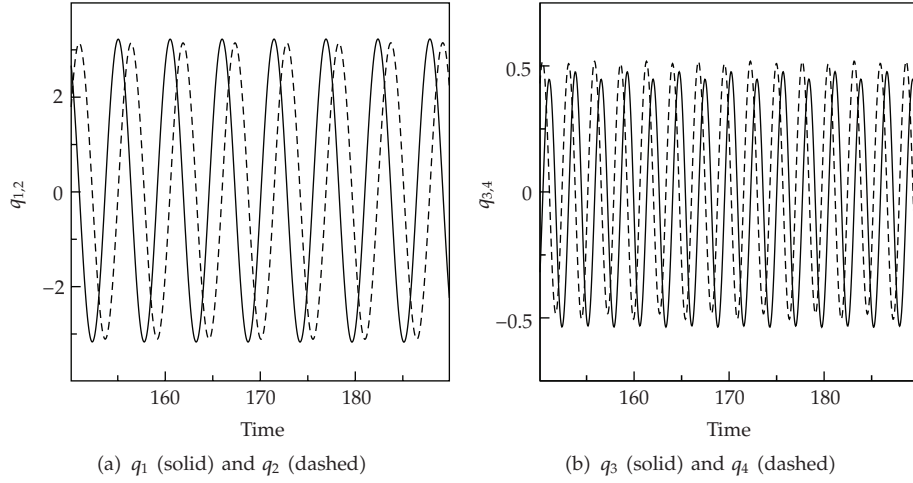**Figure 4:** The velocity coefficients q_i $(i = 1, 2, 3, 4)$ at Re $= 200$.

The effectiveness of this approach clearly depends on whether or not the POD modes exhibit a (nearly) linear dependence with respect to the parameter α. However, the dimension of the reduced basis is preserved and the reduced approximation of the flow variables still uses (3.7). Once again, the spatial functions Φ and $D\Phi_i/D\alpha$ are only available on the baseline geometry but they can be mapped on the perturbed geometry in a straight forward manner.

(ii) *Expanded Basis* (EP). The sensitivities of the modes can be shown to span a different subspace than the POD modes (see, e.g., [38]). Thus, it is natural to expect that if the approximated solution is selected in the union of the two subspaces generated by the POD modes and their sensitivities, a broader class of solutions can be represented. The expanded basis consists of the M first eigenfunctions with their M sensitivities: $[\phi_1; \ldots; \phi_M; D\Phi_1/D\alpha; \ldots; D\Phi_M/D\alpha]$. The underlying assumption behind this approach is that the subspace spanned by the mode sensitivities is well suited to address the change in the solutions induced by a change in the parameter. However, the dimension of the reduced basis has doubled and the reduced approximation of the flow variables is now expressed as

$$\mathbf{u}(\mathbf{x}, t) \approx \overline{\mathbf{u}}(\mathbf{x}) + \frac{D\overline{\mathbf{u}}(\mathbf{x})}{D\alpha} + \sum_{i=1}^{M} q_i(t)\Phi_i(\mathbf{x}) + \sum_{i=M+1}^{2M} q_i(t)\frac{D\Phi_i}{D\alpha}(\mathbf{x}). \tag{4.3}$$

5. Numerical Results and Discussion

In the current study, we numerically simulate the flow past elliptic cylinders with varying thickness ratios ranging from $\tau = 0.5$ to 1.5. In terms of the flow field, the projected area of the cylinder, as seen by the flow, is kept constant (i.e., $L_y = 1$). In other words, the width of the wake is relatively constant for all the elliptic cylinders. In Figure 3, we plot instantaneous spanwise vorticity contours for various elliptic cylinders. It is interesting to note that as the eccentricity decreases ($\tau < 1$), the vortex shedding pattern is no longer a von Karman vortex street as observed behind a circular cylinder. We observe different phenomena as the

vortex shedding transitions to steady and unsteady secondary shedding. This pattern is the manifestation of bifurcations in the flow as geometry of the cylinder is varied. However, as τ increases ($\tau > 1$), the elliptic cylinder is transformed towards an elliptic foil. We still observe von Karman vortex street in the wake; however, the shedding period changes with τ [14]. The periodic nature of these von Karman vortex flows leads to a better approximation by reduced-order models (as the basis only needs to accurately predict the first period). Based on this observation, we can predict that the sensitivity analysis would tend to show better results for increasing τ rather than decreasing τ.

From a reduced-order modeling point of view, variation in any geometric parameter would require a new set of flow data or snapshots, computation of the POD modes, and the projection onto these modes to develop a new model. This procedure corresponds to using the *perturbed* POD basis and is computationally expensive. This undermines the motivation of building a reduced-order model. However, it provides a mean for error calculations and will be used to determine how well the other approaches work. A simple solution could be the *baseline* approach, where we use the POD basis computed at the baseline, that is, for a circular cylinder ($\tau = 1.0$), map it onto the elliptic cylinder, and compute the modified flow field using the transformed baseline modes. Clearly, this classical approach where the solution at any parameter value is sought in the subspace spanned by the POD vectors generated at another parameter value performs poorly. This is because the baseline POD vectors are best suited to represent flow solutions at the parameter value for which they have been built.

Using the sensitivity analysis, we can modify the POD basis to include the effect of parameter variation in the flow field. To do so, we consider the thickness ratio of the ellipse ($\alpha \equiv \tau = L_x/L_y$) defining the cylinder shape as a parameter. Initially, we vary τ by 1% (i.e., $\tau = 0.99$) and compute the sensitivity from the finite-difference approach. The accuracy of the method is limited by the step size Δ_τ. Using the sensitivity data, we compute the *extrapolated* and *expanded* bases and project the snapshot data onto these bases to obtain the time-coefficients (q_i). We plot the two-dimensional projection of the phase portrait for one period in the (q_1, q_2), (q_1, q_3), and (q_1, q_4) planes and compare the results obtained with different basis functions in Figure 5. We observe that the data projected onto all of the bases compare well to the projection onto the perturbed POD basis. Using the coefficient data for each case, we compare the solutions to the full-order solution obtained by DNS at the perturbed state. The relative error in the velocity field u_{mod} on the modified geometry Ω_α for any basis is measured as

$$\text{error} = \int_0^T \sqrt{\int_{\Omega_\alpha} (u_{\text{mod}} - u_{\text{DNS}})^2 \, d\Omega_\alpha} \, dt. \tag{5.1}$$

In Figure 6, we plot the relative error obtained for each set of basis vectors. Obviously, the perturbed approach has the least error while the baseline approach corresponds to the maximum error. On the other hand, we observe that the extrapolated and expanded bases perform better than the baseline approach. Comparing the two new bases, the expanded bases seem to work better than the extrapolated bases. The primary reason could be the additional subspace provided by the sensitivity vectors; however, it also doubles the degrees of freedom in this case. Thus, modification of the POD basis from geometric sensitivity broadens the general applicability of the reduced-order models.
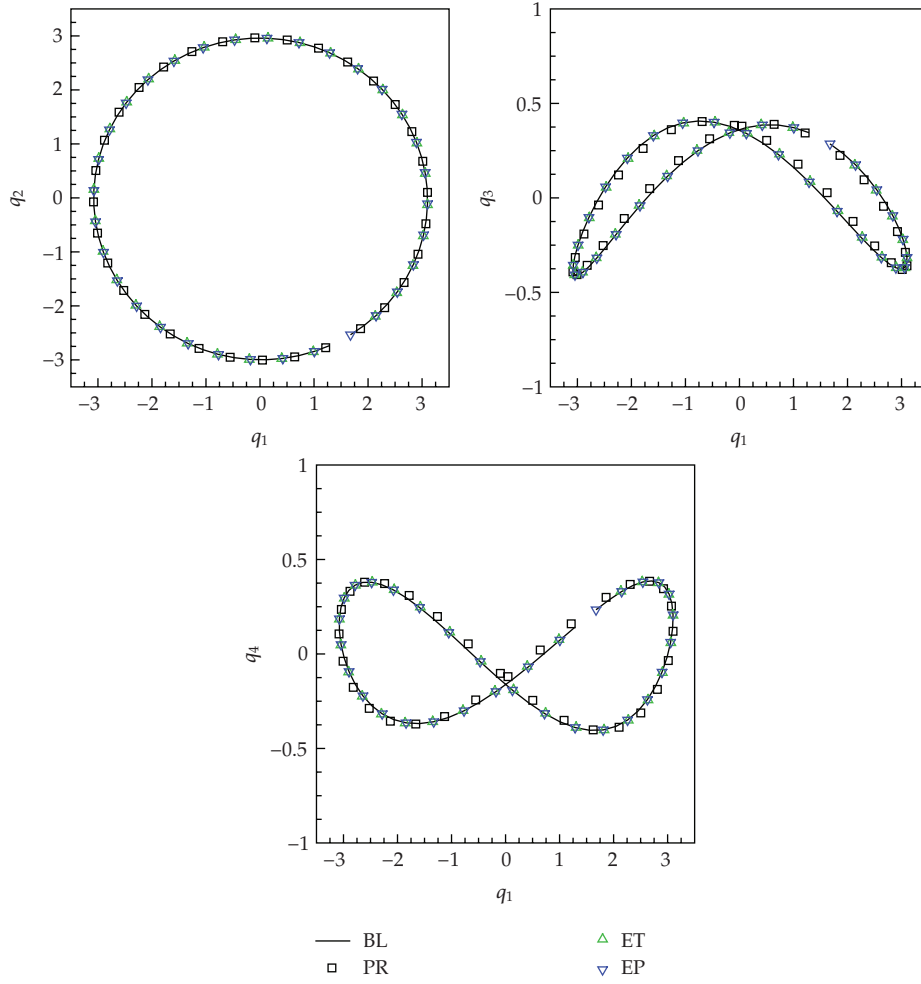
Figure 5: Two-dimensional projections of the phase portraits for different bases.

We vary τ in both directions (i.e., $\tau = 0.5–1.5$) and compute the corresponding shape sensitivities. We then use the projections and the reduced-order model to quantify the error with respect to perturbed basis (DNS data). We use 6, 12, and 24 POD modes for all cases. We compute the error for the baseline, extrapolated, and expanded basis in terms of velocity data obtained from projections (continuous curves) and reduced-order model (discrete points) as shown in Figure 7. Our first observation is that the error is small in the vicinity of $\tau \approx 1$. In fact, for a 10% perturbation in the geometric parameter, we gain nearly an order of magnitude improvement in the error over the baseline model. However, as we move away from unity, the error in all models increases and is weakly dependent on the number of modes used in the expansion. It is interesting to note that the error increases gradually as τ is increased above the value 1 while the error deteriorates quickly for τ less than 1. This fact can be explained from the flow field observed in the wake of elliptic cylinders. Lower values of τ, while keeping $L_y = 1$, means more *bluffness* in the structure and the flow tends to separate due to high
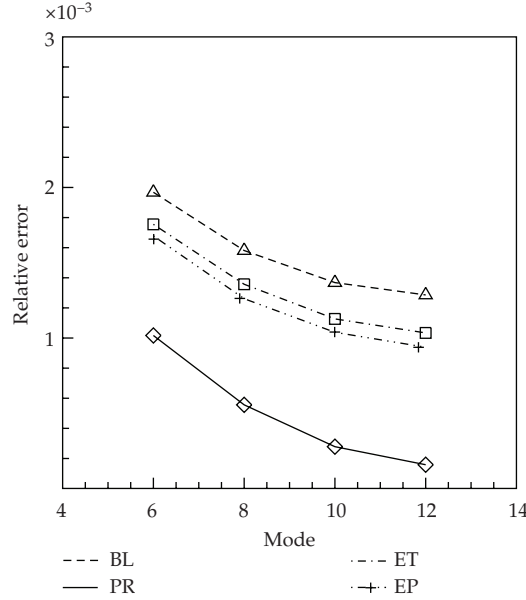
Figure 6: Relative error in reduced-order models at $\tau = 0.99$.

inverse pressure gradient. Earlier separation of the shear layer leads to transition from von Karman vortex shedding and the shedding pattern is altered in the wake. This corresponds to a bifurcation in the flow field and mere sensitivity analysis cannot capture the effect. On the other hand, increasing τ, while keeping $L_y = 1$, *smooths* the flow and the separation is delayed. Vortex shedding frequency is changed while the shedding pattern remains the same. Thus, the flow structures are similar when $\tau > 1$ and can be represented by first-order extrapolations from $\tau = 1$.

In general, the expanded basis tends to perform better than extrapolated basis. Since we assume linear dependency of the POD modes on the geometric parameter, finite-difference approach makes the extrapolated approach close to the baseline methodology. In general, the model based on expanded basis shows relatively better results close to τ.

6. Conclusion

We investigated the possibility of using the POD sensitivity vectors corresponding to a change in shape to improve the accuracy and dynamical system properties of the reduced-order models. As a part of the ongoing research in this area, we modified a circular cylinder to an elliptic cylinder by changing its thickness ratio and computed the sensitivity in the POD modes with respect to this thickness ratio. We defined different POD bases functions, with and without sensitivity, and used them to approximate the velocity field. We then compared the performance of these bases and found that the inclusion of shape sensitivity information in the POD bases performs better than the baseline approach. However,
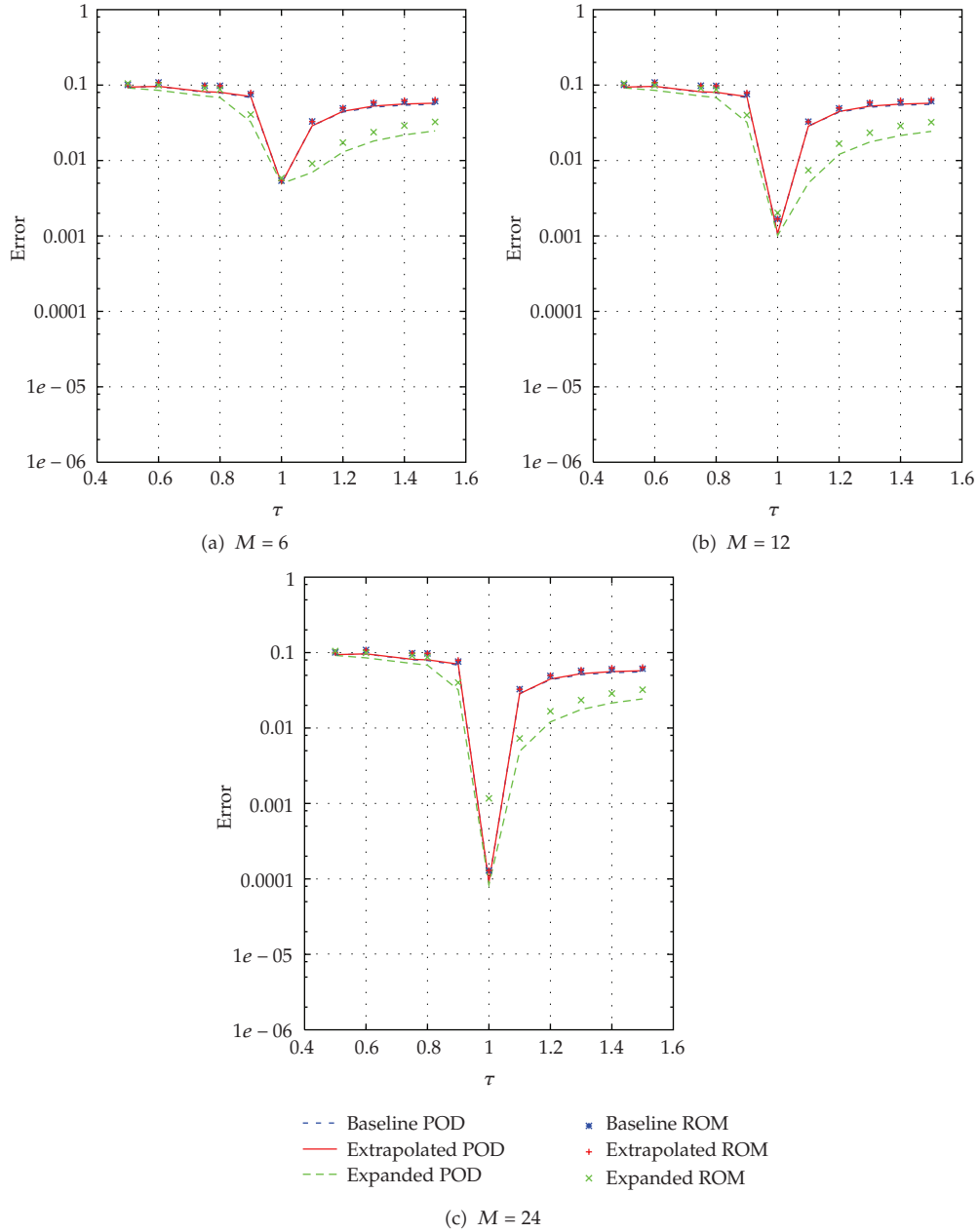
(a) $M = 6$

(b) $M = 12$

(c) $M = 24$

Figure 7: Relative error in reduced-order approximations.

the Sensitivity Analysis for shape parameters is more challenging than for flow parameters since the extrapolation explicitly requires the mapping from one domain to another. The results from the shape sensitivity are encouraging and require further investigation in this field, especially when the parameter changes lead to bifurcations which would require higher order sensitivities (and not merely the first order derivatives).

Appendix

The POD eigenfunctions are used as a basis in a Galerkin projection of the incompressible Navier-Stokes equations. The projection is performed by the inner product of the POD modes with (2.2) as

$$
\left(\Phi_k, \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p - \frac{1}{\mathrm{Re}} \nabla^2 \mathbf{u} \right) = 0,
\tag{A.1}
$$

where $(a, b) = \int_\Omega a \cdot b \, d\Omega$ represents the inner product between a and b, and $k = 1, 2, \ldots, M$. We substitute (3.7) into (A.1) and perform the inner product of each term. This leads to a reduced-order model comprising a set of M ordinary-differential equations. The inner product reduces the two (or three) momentum equations. To elaborate various terms in the reduced-order model, we expand each term individually.

Time-Derivative Term

From the Galerkin expansion in (3.7), we see that the first term is

$$
\begin{aligned}
\left(\Phi_k, \frac{\partial \mathbf{u}}{\partial t} \right) &= \left(\Phi_k, \frac{\partial \overline{\mathbf{u}}(\mathbf{x})}{\partial t} + \frac{\partial \mathbf{u}'(\mathbf{x}, t)}{\partial t} \right) \\
&= \left(\Phi_k, 0 + \frac{\partial \sum_{m=1}^{M} q_m(t) \Phi_m(\mathbf{x})}{\partial t} \right) \\
&= \left(\Phi_k, \Phi_m \sum_{m=1}^{M} \frac{dq_m}{dt} \right).
\end{aligned}
\tag{A.2}
$$

From the definition of the POD modes, the eigenfunctions $\Phi_i(\phi_i^u, \phi_i^v, \phi_i^w)$ are orthogonal by construction; that is,

$$
\left(\Phi_k, \sum_{m=1}^{M} \Phi_m \right) = (\Phi_k, \Phi_k) = (\phi_k^u, \phi_k^u) + (\phi_k^v, \phi_k^v) + (\phi_k^v, \phi_k^v) = \sigma_k.
\tag{A.3}
$$

Substituting (A.3) into (A.2), we obtain

$$
\left(\Phi_k, \frac{\partial \mathbf{u}}{\partial t} \right) = \sigma_k \frac{dq_k}{dt}.
\tag{A.4}
$$

Convection Term

We substitute (3.7) into the second term of (A.1) and separate the terms,

$$
\begin{aligned}
(\Phi_k, \mathbf{u} \cdot \nabla \mathbf{u}) &= \left(\Phi_k, (\overline{\mathbf{u}} + \mathbf{u}') \cdot \nabla (\overline{\mathbf{u}} + \mathbf{u}')\right) \\
&= \left(\Phi_k, \overline{\mathbf{u}} \cdot \nabla \overline{\mathbf{u}} + \overline{\mathbf{u}} \cdot \nabla \mathbf{u}' + \mathbf{u}' \cdot \nabla \overline{\mathbf{u}} + \mathbf{u}' \cdot \nabla \mathbf{u}'\right) \\
&= \left(\Phi_k, \overline{\mathbf{u}} \cdot \nabla \overline{\mathbf{u}} + \overline{\mathbf{u}} \cdot \nabla \sum_{m=1}^{M} q_m \Phi_m + \sum_{m=1}^{M} q_m \Phi_m \cdot \nabla \overline{\mathbf{u}} + \sum_{m=1}^{M} q_m \Phi_m \cdot \nabla \sum_{n=1}^{M} q_n \Phi_n\right) \\
&= \left(\Phi_k, \overline{\mathbf{u}} \cdot \nabla \overline{\mathbf{u}} + \overline{\mathbf{u}} \cdot \sum_{m=1}^{M} \nabla \Phi_m q_m + \sum_{m=1}^{M} \Phi_m \cdot \nabla \overline{\mathbf{u}} q_m + \sum_{m=1}^{M}\sum_{n=1}^{M} \Phi_m \cdot \nabla \Phi_n q_m q_n\right)
\end{aligned}
$$
(A.5)

We expand the inner product of each term on the right-hand side of (A.5). The first term is

$$
\begin{aligned}
(\Phi_k, \overline{\mathbf{u}} \cdot \nabla \overline{\mathbf{u}}) &= (\phi_k^u, \overline{\mathbf{u}} \cdot \nabla u) + (\phi_k^v, \overline{\mathbf{u}} \cdot \nabla v) + (\phi_k^w, \overline{\mathbf{u}} \cdot \nabla w) \\
&= \phi_k^u \left(\overline{u}\frac{\partial \overline{u}}{\partial x} + \overline{v}\frac{\partial \overline{u}}{\partial y} + \overline{w}\frac{\partial \overline{u}}{\partial z}\right) \\
&\quad + \phi_k^v \left(\overline{u}\frac{\partial \overline{v}}{\partial x} + \overline{v}\frac{\partial \overline{v}}{\partial y} + \overline{w}\frac{\partial \overline{v}}{\partial z}\right) \\
&\quad + \phi_k^w \left(\overline{u}\frac{\partial \overline{w}}{\partial x} + \overline{v}\frac{\partial \overline{w}}{\partial y} + \overline{w}\frac{\partial \overline{w}}{\partial z}\right)
\end{aligned}
$$
(A.6)

The second term is

$$
\begin{aligned}
\left(\Phi_k, \overline{\mathbf{u}} \cdot \sum_{m=1}^{M} \nabla \Phi_m q_m\right) &= \sum_{m=1}^{M} \phi_k^u \left(\overline{u}\frac{\partial \phi_m^u}{\partial x} + \overline{v}\frac{\partial \phi_m^u}{\partial y} + \overline{w}\frac{\partial \phi_m^u}{\partial z}\right) q_m \\
&\quad + \sum_{m=1}^{M} \phi_k^v \left(\overline{u}\frac{\partial \phi_m^v}{\partial x} + \overline{v}\frac{\partial \phi_m^v}{\partial y} + \overline{w}\frac{\partial \phi_m^v}{\partial z}\right) q_m \\
&\quad + \sum_{m=1}^{M} \phi_k^w \left(\overline{u}\frac{\partial \phi_m^w}{\partial x} + \overline{v}\frac{\partial \phi_m^w}{\partial y} + \overline{w}\frac{\partial \phi_m^w}{\partial z}\right) q_m \\
&= \left[\phi_k^u \left(\overline{u}\frac{\partial \phi_m^u}{\partial x} + \overline{v}\frac{\partial \phi_m^u}{\partial y} + \overline{w}\frac{\partial \phi_m^u}{\partial z}\right)\right. \\
&\quad + \phi_k^v \left(\overline{u}\frac{\partial \phi_m^v}{\partial x} + \overline{v}\frac{\partial \phi_m^v}{\partial y} + \overline{w}\frac{\partial \phi_m^v}{\partial z}\right) \\
&\quad \left. + \phi_k^w \left(\overline{u}\frac{\partial \phi_m^w}{\partial x} + \overline{v}\frac{\partial \phi_m^w}{\partial y} + \overline{w}\frac{\partial \phi_m^w}{\partial z}\right)\right] * \mathbf{q}.
\end{aligned}
$$
(A.7)

The third term is

$$
\begin{aligned}
\left(\Phi_k, \sum_{m=1}^{M} q_m \Phi_m \cdot \nabla \overline{\mathbf{u}} \right) &= \sum_{m=1}^{M} \phi_k^u \left(\phi_m^u \frac{\partial \overline{u}}{\partial x} + \phi_m^v \frac{\partial \overline{u}}{\partial y} + \phi_m^w \frac{\partial \overline{u}}{\partial z} \right) q_m \\
&+ \sum_{m=1}^{M} \phi_k^v \left(\phi_m^u \frac{\partial \overline{v}}{\partial x} + \phi_m^v \frac{\partial \overline{v}}{\partial y} + \phi_m^w \frac{\partial \overline{v}}{\partial z} \right) q_m \\
&+ \sum_{m=1}^{M} \phi_k^w \left(\phi_m^u \frac{\partial \overline{w}}{\partial x} + \phi_m^v \frac{\partial \overline{w}}{\partial y} + \phi_m^w \frac{\partial \overline{w}}{\partial z} \right) q_m \\
&= \left[\phi_k^u \left(\phi_m^u \frac{\partial \overline{u}}{\partial x} + \phi_m^v \frac{\partial \overline{u}}{\partial y} + \phi_m^w \frac{\partial \overline{u}}{\partial z} \right) \right. \\
&+ \phi_k^v \left(\phi_m^u \frac{\partial \overline{v}}{\partial x} + \phi_m^v \frac{\partial \overline{v}}{\partial y} + \phi_m^w \frac{\partial \overline{v}}{\partial z} \right) \\
&+ \left. \phi_k^w \left(\phi_m^u \frac{\partial \overline{w}}{\partial x} + \phi_m^v \frac{\partial \overline{w}}{\partial y} + \phi_m^w \frac{\partial \overline{w}}{\partial z} \right) \right] * \mathbf{q}.
\end{aligned} \tag{A.8}
$$

The fourth term is

$$
\begin{aligned}
\left(\Phi_k, \sum_{m=1}^{M} q_m \Phi_m \cdot \nabla \sum_{n=1}^{M} q_n \Phi_n \right) &= \left(\phi_k^u, \sum_{m=1}^{M} \sum_{n=1}^{M} q_m q_n \Phi_n \cdot \nabla \phi_m^u \right) \\
&+ \left(\phi_k^v, \sum_{m=1}^{M} \sum_{n=1}^{M} q_m q_n \Phi_n \cdot \nabla \phi_m^v \right) \\
&+ \left(\phi_k^w, \sum_{m=1}^{M} \sum_{n=1}^{M} q_m q_n \Phi_n \cdot \nabla \phi_m^w \right).
\end{aligned} \tag{A.9}
$$

Equation (A.9) is further split into subterms for convenience as follows:

$$
\begin{aligned}
\left(\phi_k^u, \sum_{m=1}^{M} \sum_{n=1}^{M} q_m q_n \Phi_n \cdot \nabla \phi_m^u \right) &= \left(\phi_k^u, \sum_{m=1}^{M} \sum_{n=1}^{M} q_m q_n \left(\phi_n^u * \frac{\partial \phi_m^u}{\partial x} + \phi_n^v * \frac{\partial \phi_m^u}{\partial y} + \phi_n^w * \frac{\partial \phi_m^u}{\partial z} \right) \right) \\
&= \mathbf{q}' * \left[\phi_k^u \left(\phi_n^u * \frac{\partial \phi_m^u}{\partial x} + \phi_n^v * \frac{\partial \phi_m^u}{\partial y} + \phi_n^w * \frac{\partial \phi_m^u}{\partial z} \right) \right] * \mathbf{q}.
\end{aligned} \tag{A.10}
$$

Similarly, the other subterms can be expanded as

$$
\begin{aligned}
\left(\phi_k^v, \sum_{m=1}^{M} \sum_{n=1}^{M} q_m q_n \Phi_n \cdot \nabla \phi_m^v \right) &= \left(\phi_k^u, \sum_{m=1}^{M} \sum_{n=1}^{M} q_m q_n \left(\phi_n^u * \frac{\partial \phi_m^v}{\partial x} + \phi_n^v * \frac{\partial \phi_m^v}{\partial y} + \phi_n^w * \frac{\partial \phi_m^v}{\partial z} \right) \right) \\
&= \mathbf{q}' * \left[\phi_k^v \left(\phi_n^u * \frac{\partial \phi_m^v}{\partial x} + \phi_n^v * \frac{\partial \phi_m^v}{\partial y} + \phi_n^w * \frac{\partial \phi_m^v}{\partial z} \right) \right] * \mathbf{q}, \\
\left(\phi_k^w, \sum_{m=1}^{M} \sum_{n=1}^{M} q_m q_n \Phi_n \cdot \nabla \phi_m^w \right) &= \left(\phi_k^u, \sum_{m=1}^{M} \sum_{n=1}^{M} q_m q_n \left(\phi_n^u * \frac{\partial \phi_m^w}{\partial x} + \phi_n^v * \frac{\partial \phi_m^w}{\partial y} + \phi_n^w * \frac{\partial \phi_m^w}{\partial z} \right) \right) \\
&= \mathbf{q}' * \left[\phi_k^w \left(\phi_n^u * \frac{\partial \phi_m^w}{\partial x} + \phi_n^v * \frac{\partial \phi_m^w}{\partial y} + \phi_n^w * \frac{\partial \phi_m^w}{\partial z} \right) \right] * \mathbf{q}.
\end{aligned}
\tag{A.11}
$$

Combining (A.10) and (A.11), we obtain

$$
\begin{aligned}
\left(\Phi_k, \sum_{m=1}^{M} q_m \Phi_m \cdot \nabla \sum_{n=1}^{M} q_n \Phi_n \right) = \mathbf{q}' * \Bigg[&\phi_k^u \left(\phi_n^u * \frac{\partial \phi_m^u}{\partial x} + \phi_n^v * \frac{\partial \phi_m^u}{\partial y} + \phi_n^w * \frac{\partial \phi_m^u}{\partial z} \right) \\
&+ \phi_k^v \left(\phi_n^u * \frac{\partial \phi_m^v}{\partial x} + \phi_n^v * \frac{\partial \phi_m^v}{\partial y} + \phi_n^w * \frac{\partial \phi_m^v}{\partial z} \right) \\
&+ \phi_k^w \left(\phi_n^u * \frac{\partial \phi_m^w}{\partial x} + \phi_n^v * \frac{\partial \phi_m^w}{\partial y} + \phi_n^w * \frac{\partial \phi_m^w}{\partial z} \right) \Bigg] * \mathbf{q}.
\end{aligned}
\tag{A.12}
$$

Equation (A.12) represents the quadratic term in the model.

Pressure Term

The pressure term in the model is also projected onto the POD modes as follows:

$$
(\Phi_k, \nabla p) = \int \Phi_k \cdot \nabla p = -\int (\nabla \cdot \Phi_k) p + \int_{\Omega_s} p (\mathbf{n} \cdot \Phi_k).
\tag{A.13}
$$

We note that using Green's theorem and the divergence-free property, the pressure term drops out from (3.8) for the case of $p = 0$ on the outerflow boundary Ω_{so} [19]. The POD eigenfunctions are identically zero on the inflow boundary because the average flow is subtracted from the total flow. However, in case of Neumann boundary conditions on Ω_{so}, the contribution of the pressure term is not exactly zero for the cylinder wake. The outer domain is intentionally kept at $25D$ from the cylinder to minimize the pressure effects. Hence, the pressure is neglected on the outflow boundary so the pressure term vanishes in the reduced-order model [37]. Thus,

$$
(\Phi_k, \nabla p) = 0.
\tag{A.14}
$$

Diffusion Term

We substitute (3.7) into the diffusion term to obtain

$$
\left(\Phi_k, \frac{1}{\mathrm{Re}_D}\nabla^2\mathbf{u}\right) = -\frac{1}{\mathrm{Re}_D}\left(\Phi_k, \nabla^2(\overline{\mathbf{u}}+\mathbf{u}')\right) = -\frac{1}{\mathrm{Re}_D}\left(\Phi_k, \nabla^2\overline{\mathbf{u}} + \sum_{m=1}^{M}q_m\nabla^2\Phi_m\right). \qquad (A.15)
$$

We expand each term on the right-hand side of (A.15) into its components. The term containing the mean velocity term is

$$
\begin{aligned}
\left(\Phi_k, \nabla^2\overline{\mathbf{u}}\right) &= \phi_k^u\left(\frac{\partial^2\overline{u}}{\partial x^2} + \frac{\partial^2\overline{u}}{\partial y^2} + \frac{\partial^2\overline{u}}{\partial z^2}\right) \\
&\quad + \phi_k^v\left(\frac{\partial^2\overline{v}}{\partial x^2} + \frac{\partial^2\overline{v}}{\partial y^2} + \frac{\partial^2\overline{v}}{\partial z^2}\right) \\
&\quad + \phi_k^w\left(\frac{\partial^2\overline{w}}{\partial x^2} + \frac{\partial^2\overline{w}}{\partial y^2} + \frac{\partial^2\overline{w}}{\partial z^2}\right).
\end{aligned} \qquad (A.16)
$$

Likewise, the term containing the POD mode is expanded as follows:

$$
\begin{aligned}
\left(\Phi_k, \sum_{m=1}^{M}q_m\nabla^2\Phi_m\right) &= \sum_{m=1}^{M}q_m\phi_k^u\left(\frac{\partial^2\phi_m^u}{\partial x^2} + \frac{\partial^2\phi_m^u}{\partial y^2} + \frac{\partial^2\phi_m^u}{\partial z^2}\right) \\
&\quad + \sum_{m=1}^{M}q_m\phi_k^v\left(\frac{\partial^2\phi_m^v}{\partial x^2} + \frac{\partial^2\phi_m^v}{\partial y^2} + \frac{\partial^2\phi_m^v}{\partial z^2}\right) \\
&\quad + \sum_{m=1}^{M}q_m\phi_k^w\left(\frac{\partial^2\phi_m^w}{\partial x^2} + \frac{\partial^2\phi_m^w}{\partial y^2} + \frac{\partial^2\phi_m^w}{\partial z^2}\right) \\
&= \left[\phi_k^u\left(\frac{\partial^2\phi_m^u}{\partial x^2} + \frac{\partial^2\phi_m^u}{\partial y^2} + \frac{\partial^2\phi_m^u}{\partial z^2}\right)\right. \\
&\quad + \phi_k^v\left(\frac{\partial^2\phi_m^v}{\partial x^2} + \frac{\partial^2\phi_m^v}{\partial y^2} + \frac{\partial^2\phi_m^v}{\partial z^2}\right) \\
&\quad \left. + \phi_k^w\left(\frac{\partial^2\phi_m^w}{\partial x^2} + \frac{\partial^2\phi_m^w}{\partial y^2} + \frac{\partial^2\phi_m^w}{\partial z^2}\right)\right] * \mathbf{q}.
\end{aligned} \qquad (A.17)
$$

Acknowledgments

References

[1] A. Roshko, "On the wake and drag of bluff bodies," *Journal of Aerospace Science and Technologies*, vol. 22, no. 2, pp. 124–132, 1955.

[2] R. E. D. Bishop and A. Y. Hassan, "The lift and drag forces on a circular cylinder in flowing fluid," *Proceedings of the Royal Society A*, vol. 277, pp. 32–50, 1963.

[3] G. E. Karniadakis and G. S. Triantafyllou, "Three-dimensional dynamics and transition to turbulence in the wake of bluff objects," *Journal of Fluid Mechanics*, vol. 238, pp. 1–30, 1992.

[4] A. Roshko, "On the development of turbulent wakes from vortex streets," Tech. Rep. 1191, NACA, 1954.

[5] A. G. Tomboulides, M. Israeli, and G. E. Karniadakis, "Efficient removal of boundary-divergence errors in time-splitting methods," *Journal of Scientific Computing*, vol. 4, no. 3, pp. 291–308, 1989.

[6] C. H. K. Williamson, "Vortex dynamics in the cylinder wake," in *Annual Review of Fluid Mechanics*, vol. 28, pp. 477–539, Annual Reviews, Palo Alto, Calif, USA, 1996.

[7] J. Wu, J. Sheridan, J. Soria, and M. C. Welsh, "An experimental investigation of streamwise vortices in the wake of a bluff body," *Journal of Fluids and Structures*, vol. 8, no. 7, pp. 621–635, 1994.

[8] R. T. Hartlen and I. G. Currie, "Lift-oscillator model of vortex-induced vibration," *ASCE Journal of Engineering Mechanics*, vol. 96, pp. 577–591, 1970.

[9] W. D. Iwan and R. D. Blevins, "A model for vortex induced oscillation of structures," *Journal of Applied Mechanics*, vol. 41, no. 3, pp. 581–586, 1974.

[10] R. Landl, "A mathematical model for vortex-excited vibration of cable suspensions," *Journal of Sound and Vibration*, vol. 42, no. 2, pp. 219–234, 1975.

[11] R. A. Skop and O. M. Griffin, "On a theory for the vortex-excited oscillations of flexible cylindrical structures," *Journal of Sound and Vibration*, vol. 41, no. 3, pp. 263–274, 1975.

[12] S. Krenk and S. R. K. Nielsen, "Energy balanced double oscillator model for vortex-induced vibrations," *Journal of Engineering Mechanics*, vol. 125, no. 3, pp. 263–271, 1999.

[13] O. Marzouk, A. H. Nayfeh, I. Akhtar, and H. N. Arafat, "Modeling steady-state and transient forces on a cylinder," *Journal of Vibration and Control*, vol. 13, no. 7, pp. 1065–1091, 2007.

[14] I. Akhtar, O. A. Marzouk, and A. H. Nayfeh, "A van der Pol-Duffing oscillator model of hydrodynamic forces on canonical structures," *Journal of Computational and Nonlinear Dynamics*, vol. 4, no. 4, Article ID 041006, 2009.

[15] G. Berkooz, P. Holmes, and J. L. Lumley, "The proper orthogonal decomposition in the analysis of turbulent flows," *Annual Review of Fluid Mechanics*, vol. 53, pp. 321–575, 1993.

[16] P. Holmes, J. L. Lumley, and G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge Monographs on Mechanics, Cambridge University Press, Cambridge, UK, 1996.

[17] L. Sirovich, "Turbulence and the dynamics of coherent structures," *Quarterly of Applied Mathematics*, vol. 45, pp. 561–590, 1987.

[18] A. E. Deane and C. Mavriplis, "Low-dimensional description of the dynamics in separated flow past thick airfoils," *AIAA Journal*, vol. 32, no. 6, pp. 1222–1227, 1994.

[19] X. Ma and G. E. Karniadakis, "A low-dimensional model for simulating three-dimensional cylinder flow," *Journal of Fluid Mechanics*, vol. 458, pp. 181–190, 2002.

[20] B. R. Noack, K. Afanasiev, M. Morzynski, G. Tadmor, and F. Thiele, "A hierarchy of low-dimensional models for the transient and post-transient cylinder wake," *Journal of Fluid Mechanics*, vol. 497, pp. 335–363, 2003.

[21] W. R. Graham, J. Peraire, and K. Y. Tang, "Optimal control of vortex shedding using low-order models—part I: open-loop model development," *International Journal for Numerical Methods in Engineering*, vol. 44, no. 7, pp. 945–972, 1999.

[22] M. Bergmann, L. Cordier, and J.-P. Brancher, "Optimal rotary control of the cylinder wake using proper orthogonal decomposition reduced-order model," *Physics of Fluids*, vol. 17, no. 9, Article ID 097101, 21 pages, 2005.

[23] I. Akhtar, A. H. Nayfeh, and C. J. Ribbens, "On the stability and extension of reduced-order galerkin models in incompressible flows: a numerical study of vortex shedding," *Theoretical and Computational Fluid Dynamics*, vol. 23, no. 3, pp. 213–237, 2009.

[24] A. E. Deane, I. G. Kevrekidis, G. E. Karniadakis, and S. A. Orszag, "Low-dimensional models for complex geometry flows: application to grooved channels and circular cylinders," *Physics of Fluids A*, vol. 3, no. 10, pp. 2337–2354, 1991.

[25] E. Tadmor, "Convergence of spectral methods for nonlinear conservation laws," *SIAM Journal on Numerical Analysis*, vol. 26, no. 1, pp. 30–44, 1989.

[26] C. Foias, M. S. Jolly, I. G. Kevrekidis, and E. S. Titi, "Dissipativity of numerical schemes," *Nonlinearity*, vol. 4, no. 3, pp. 591–613, 1991.

[27] N. Aubry, W. Y. Lian, and E. S. Titi, "Preserving symmetries in the proper orthogonal decomposition," *SIAM Journal on Scientific Computing*, vol. 14, no. 2, pp. 483–505, 1993.

[28] M. Couplet, P. Sagaut, and C. Basdevant, "Intermodal energy transfers in a proper orthogonal decomposition—Galerkin representation of a turbulent separated flow," *Journal of Fluid Mechanics*, vol. 491, pp. 275–284, 2003.

[29] S. Sirisup and G. E. Karniadakis, "A spectral viscosity method for correcting the long-term behavior of POD models," *Journal of Computational Physics*, vol. 194, no. 1, pp. 92–116, 2004.

[30] M. Bergmann, C.-H. Bruneau, and A. Iollo, "Enablers for robust POD models," *Journal of Computational Physics*, vol. 228, no. 2, pp. 516–538, 2009.

[31] A. Hay, J. T. Borggaard, and D. Pelletier, "Local improvements to reduced-order models using sensitivity analysis of the proper orthogonal decomposition," *Journal of Fluid Mechanics*, vol. 629, pp. 41–72, 2009.

[32] A. Hay, J. Borggaard, and D. Pelletier, "Reduced-order models for parameter dependent geometries based on shape sensitivity analysis of the POD," in *Proceedings of the 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Victoria, Canada, September 2008, AIAA-2008-5962.

[33] B. P. Leonard, "A stable and accurate convective modelling procedure based on quadratic upstream interpolation," *Computer Methods in Applied Mechanics and Engineering*, vol. 19, no. 1, pp. 59–98, 1979.

[34] Y. Zang, R. L. Street, and J. R. Koseff, "A non-staggered grid, fractional step method for time-dependent incompressible Navier-Stokes equations in curvilinear coordinates," *Journal of Computational Physics*, vol. 114, no. 1, pp. 18–33, 1994.

[35] I. Akhtar, *Parallel simulations, reduced-order modeling, and feedback control of vortex shedding using fluidic actuators*, Ph.D. thesis, Virginia Tech, Blacksburg, Va, USA, 2008.

[36] H. P. Bakewell Jr. and J. L. Lumley, "Viscous sublayer and adjacent wall region in turbulent pipe flow," *Physics of Fluids*, vol. 10, no. 9, pp. 1880–1889, 1967.

[37] B. R. Noack, P. Papas, and P. A. Monkewitz, "The need for a pressure-term representation in empirical Galerkin models of incompressible shear flows," *Journal of Fluid Mechanics*, vol. 523, pp. 339–365, 2005.

[38] A. Hay, J. Borggaard, and D. Pelletier, "On the use of sensitivity analysis to improve reduced-order models," in *Proceedings of the 4th AIAA Flow Control Conference*, Seattle, Wash, USA, June 2008, AIAA-2008-4192.

Research Article

Investigations on Actuator Dynamics through Theoretical and Finite Element Approach

Somashekhar S. Hiremath and M. Singaperumal

Precision Engineering and Instrumentation Laboratory, Department of Mechanical Engineering, Indian Institute of Technology Madras, Chennai 6000 36, India

Correspondence should be addressed to M. Singaperumal, msingam@iitm.ac.in

This paper gives a new approach for modeling the fluid-structure interaction of servovalve component-actuator. The analyzed valve is a precision flow control valve-jet pipe electrohydraulic servovalve. The positioning of an actuator depends upon the flow rate from control ports, in turn depends on the spool position. Theoretical investigation is made for No-load condition and Load condition for an actuator. These are used in finite element modeling of an actuator. The fluid-structure-interaction (FSI) is established between the piston and the fluid cavities at the piston end. The fluid cavities were modeled with special purpose hydrostatic fluid elements while the piston is modeled with brick elements. The finite element method is used to simulate the variation of cavity pressure, cavity volume, mass flow rate, and the actuator velocity. The finite element analysis is extended to study the system's linearized response to harmonic excitation using direct solution steady-state dynamics. It was observed from the analysis that the natural frequency of the actuator depends upon the position of the piston in the cylinder. This is a close match with theoretical and simulation results. The effect of bulk modulus is also presented in the paper.

1. Introduction

The electrohydraulic servovalves (EHSVs), probably the youngest of the standard hydraulic components, made it appearance in the latter 1940s to satisfy the aerospace needs for a fast response servo-controlled system. After 1940s these EHSVs were not significantly faster than electrical servos because they were lacking by an element which could rapidly translate electrical signals into a hydraulic flows. These early servovalves were actuated by small electric servomotors with a large time constants and hence they are the slowest element in the control loop and have a limited system performance. In the early 1950s permanent magnet torque motors having fast response gained favor as a method of stroking valves and the EHSV took its present form. EHSV connects the electronic and hydromechanical portions of a system and hence it is treated as a mechatronic component.

EHSVs are faster responding directional, pressure, and flow control valves, which are frequently used in closed-loop control arrangements to produce the highly sophisticated performance, in terms of high-frequency response, required by modern machines. They serve as an interface between electrical devices and hydraulic systems. They are capable of converting low-power electrical input into movement of spool to precisely control large power, low-speed hydraulic actuators. Thus, they are extensively used in such applications as computer numerical controlled machine tools, aircraft, robotic manipulators, motion simulators, injection molding, remotely controlled mechanisms, and many off-highway and mobile applications.

Most present day EHSVs have three major components:

(1) torque motor,

(2) first-stage hydraulic amplifier,

(3) second-stage sliding spool valve.

The torque motor consists of an armature mounted on a torsion pin pivot and suspended in the air gap of a magnetic field. Two pole pieces, one permanently polarized north and the other south by the permanent magnet, form the framework around the armature and provide paths for the magnetic flux. When current is made to flow through the armature coils, the armature ends become polarized and each end is attracted to one pole piece and repelled by the other. A torque thus produced rotates the armature proportional to the input current. The servovalve to be positioned may be mechanically attached at either armature end. The static and dynamic characteristic of the torque motor plays a vital role in EHSV operation and is well elaborated in the books [1, 2]. Maskrey and Thayer [3] elaborated the historical development, market growth, and wide application areas of EHSV. The selection and performance criteria for electrohydraulic servodrives are given Moog technical bulletin 122 [4] and the performance estimation for electrohydraulic control systems are given Moog technical bulletin 126 [5].

First-stage hydraulic amplifier substantially multiplies the force output of the torque motor to a level sufficient to overcome considerable flow forces, stiction forces, and forces resulting from the acceleration or vibration. Flapper, jet pipe, and spool valves may be used as a hydraulic amplifier in EHSV. Much of the work has been done to study dynamics of flapper/nozzle EHSV. The design aspects and various configurations of EHSV particularly on flapper/nozzle are available in many text books [1, 2, 6–9].

Second-stage sliding spool valves are classified based on many parameters.

(i) *Number of ways "flow" can enter and leave the valve.* Generally all servovalves require at least one supply port, one return port and one line to the load and hence they are classified as a three-way EHSV or a four-way EHSV.

(ii) *Number of spool lands*. The number of lands on a spool may vary from one in a primitive valve to the usual three or four, and special valves may have as many as six lands.

(iii) *Type of valve center when the spool is in neutral position.* If the width of the spool land is smaller than the control port width in the valve sleeve, the valve is said to have an open center or to be an under lapped. If the width of the spool land is identical to the control port width in the valve sleeve, the valve is said to have a critical center or a zero lapped. If the width of the spool land is greater than the control port width in the valve sleeve, the valve is said to have a closed center or to be an overlapped.

The flow forces on the spool dynamics of flapper/nozzle EHSV and their compensation methods are elaborated by Lee and Blackburn [10]. Taft and Twill [11] developed a flow model of a three-way underlapped hydraulic servovalve and derived a mathematical description of the flow momentum forces acting on the spool valve. The effect of these forces on valve performance is investigated by examining both the linearized system of differential equations and digital computer solutions of the system nonlinear differential equations, and by experimental measurements. Lee et al. [12] developed a new linearized equation for modeling servovalve as a modified form of the conventional linearized equation. The conventional linearized equation described by the first-order terms of Taylor's series is effective just near the operating point. However, pressure and flow rate in actual hydraulic systems are usually not confined near an operating point. From the evaluations of time responses and frequency responses obtained from simulations for a hydraulic control system, the effectiveness of the new linearized equation and the methods to determine effective operating points is confirmed.

A detailed transfer function of the flapper/nozzle EHSV has been developed and simulated by Nikiforuk et al. [13]. LeQuoc et al. [14] proposed a novel electrohydraulic configuration, in which the drain line is connected to the tank through a direction control valve, a metering valve, and a relief valve which allow external adjustment of the drain, orifice and back pressure. Servosystems with the conventional servovalve and the new servovalve configuration are modeled and simulated for step input for various values of system parameters. The simulation results indicated that the servosystem with this new configuration would offer a higher steady-state velocity, a lower percent overshoot, and a lesser settling time, if the drain line orifice opening and the back pressure were properly tuned. Also they carried out experiments to validate the simulation results and it has been demonstrated that the mathematical model is relatively adequate to predict the performance of the two servosystems.

Das et al. [15] modeled the electrohydraulic actuation system used for launch vehicle attitude control. The unified nonlinear model has been developed taking into account the associated nonlinearities. The developed model has been simulated using the Matlab and the simulation results are compared with experimental values.

However, studies on jet pipe EHSV are limited. Jet pipe EHSV finds main applications in jet engine, flight control, and turbofan engine control system. Fredlake and Adams [16], Bruce [17], and Harschburger [18] worked on the development of flight control system for fighter aircrafts and hydromechanical control. Ogata [19] explained the operation of a hydraulic jet pipe controller and how the jet pipe controller governs the position of the butterfly valve.

The basic components and principle of operation of jet pipe EHSV are available in text books [1, 2, 6–9]. The major parameters affecting the first-stage pressure recovery are presented by Allen [1]. Experimentally these pressure recoveries in the receiver holes have been measured and studied the valve dynamics using the frequency response method as detailed by Dushkes and Cahn [20]. The pressure recoveries in the receiver holes have been simulated using a bond graph method—a dynamic model tool, and the spool dynamics is studied by Thoma [21]. An analytical and experimental investigation of a jet pipe—controlled electropneumatic servoactuator designed for use in the Utah/MIT dextrous hand has been performed by Henri et al. [22].

The modification in the jet pipe EHSV results in a new design called the deflector jet EHSV. The design aspects and operating principle of the deflector jet EHSV are elaborated in Moog technical bulletin 121 [23]. The jet pipe EHSV is not used as much as the flapper/nozzle
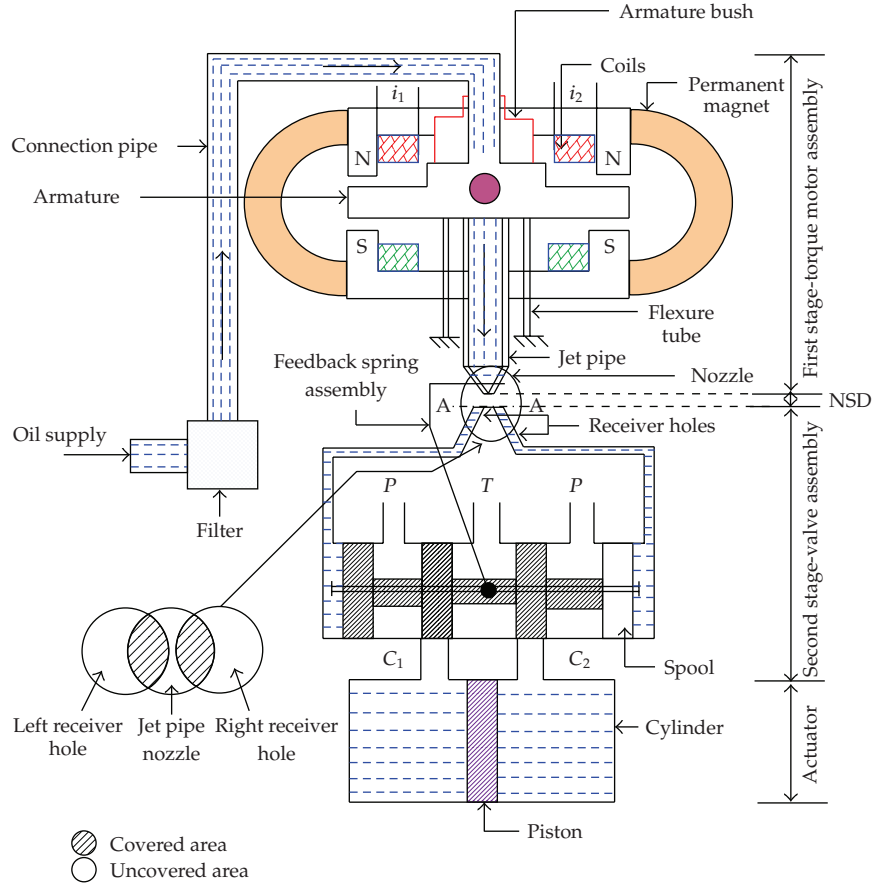
Figure 1: Schematic diagram of a jet pipe electrohydraulic servovalve.

EHSV because of large null flow, slower response, and rather unpredictable characteristics. Its main advantage lies in its insensitivity to dirty fluids.

1.1. Jet Pipe Electrohydraulic Servovalve

The schematic diagram of the analyzed jet pipe EHSV along with actuator is shown in Figure 1. It consists of two main assemblies-first-stage torque motor assembly and second-stage valve assembly. Between the first and the second stages, there is a mechanical feedback spring assembly to stabilize the valve operation. The analyzed jet pipe EHSV is of miniature type with the specifications shown in Table 1.

The jet pipe EHSV operation is as follows.

(i) At null position (no current input to the torque motor coils), the jet is directed exactly between the two receivers, making the pressures on both sides of the spool end equal. The force balance created by equal pressures in both end chambers holds the spool in a stationary position.

(ii) When a current passes through a torque motor coils, it generates an electromagnetic force (emf) on the armature ends and hence a torque on the armature and jet pipe.

Table 1: Specification of a jet pipe electrohydraulic.

No	Parameters	Value
1	System pressure	1.7 MPa
2.	Tank pressure	0.1 MPa
3.	Max. jet pipe deflection	0.2 mm
4.	Max. spool displacement	0.85 mm
5.	No load flow rate	17 L/min. for 1.7 MPa pressure drop
6	Spool diameter and length	11 mm and 40 mm
7.	Jet and receiver diameters	0.3 mm
8.	Piston diameter	54 mm
9.	Piston stroke length	100 mm
10.	Working fluid	Aviation kerosene

(iii) The torque rotates the jet pipe assembly around a pivot point and hence the fluid jet is directed to one of the two receiver holes in the receiver plate, creating a higher pressure in the spool end chamber connected to that receiver. The differential pressure pushes the spool in opposite direction to the jet pipe deflection.

(iv) As the spool starts moving, it pushes the feedback spring assembly, creating the restoring torque on the jet pipe to bring it back to the null position. When the restoring torque due to spool movement equals the applied torque on the armature, the spool stops at a particular position, until the value of the applied current is changed or the polarity of the applied current reversed. The resulting spool position opens a specified flow passage at the control ports of the second-stage to actuator.

The main investigations carried out on the jet pipe EHSV are no-load flow and load flow analysis. The obtained results are used in finite element analysis.

2. Theoretical Flow Modeling

The main purpose of the servovalve is to regulate precisely the required amount of fluid flow to the actuator to control the load position. The actuator load positioning depends on how fast the servovalve responds to the input signal. Two important recommended tests to study the valve performance are no-load and load flow characteristics. Both these are obtained through static analysis.

2.1. No-Load Flow Analysis

The standard procedure to ascertain the no-load flow for the servovalve is to connect the control ports C_1 and C_2 shown in Figure 2. As the spool starts moving, the control ports are opened and fluid flows from the supply port (P) to tank port (T). The amount of fluid flow depends on the opened port area and the differential pressure across the ports. Assuming a symmetric valve configuration, the fluid flow to and from the ports is calculated using the orifice theory:

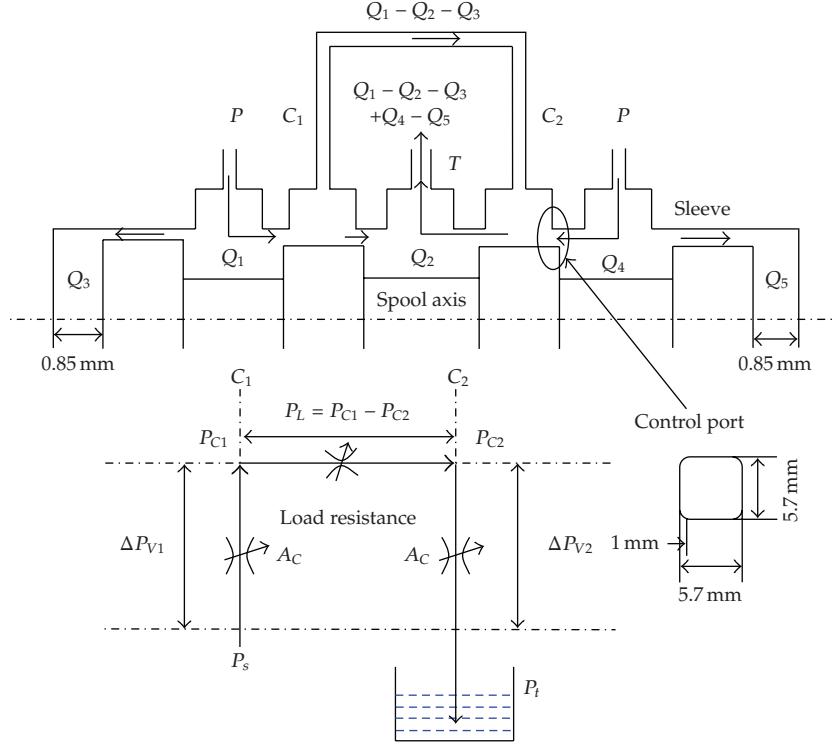$$Q_{ac} = K_2\sqrt{\Delta P_{V1}} = K_2\sqrt{\Delta P_{V2}}, \tag{2.1}$$

Figure 2: Port configuration for no-load flow analysis.

where $K_2 = C_d A_c \sqrt{2/\rho_o}$,

$$\begin{aligned}
\Delta P_{V1} &= P_s - P_{C1}, \\
\Delta P_{V2} &= P_{C2} - P_t.
\end{aligned} \tag{2.2}$$

Equating the above pressure drops for a symmetric valves,

$$P_s - P_t = P_{C1} - P_{C2}. \tag{2.3}$$

If a load resistance exists between C_1 and C_2, then the load pressure drop is given by

$$P_L = P_{C1} - P_{C2}. \tag{2.4}$$

By solving the (2.3) and (2.4), the control port pressures are given by

$$\begin{aligned}
P_{C1} &= \frac{P_s + P_L + P_t}{2}, \\
P_{C2} &= \frac{P_s - P_L + P_t}{2}.
\end{aligned} \tag{2.5}$$
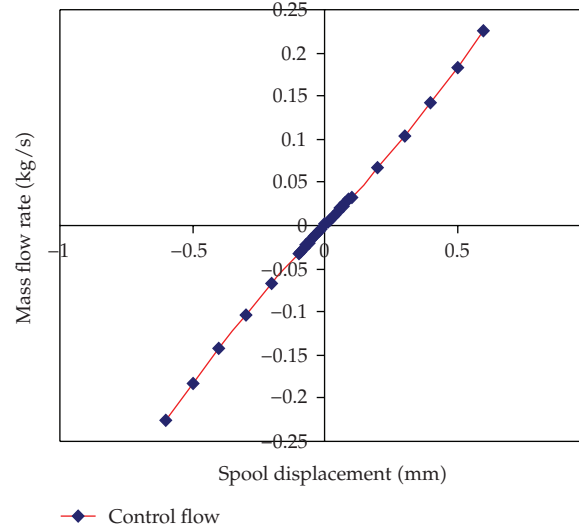
Figure 3: Mass flow rate in the no-load flow analysis.

Hence the valve pressure drops across the ports is given by

$$\Delta P_{V1} = \Delta P_{V2} = \frac{P_s - P_L - P_t}{2}. \tag{2.6}$$

The fluid flow to the actuator is given by

$$Q_{ac} = K_2 \sqrt{\frac{P_s - P_L - P_t}{2}}. \tag{2.7}$$

For the no-load flow analysis, the load pressure is $P_L = 0$, and so the control port pressures P_{C1} and P_{C2} are calculated from (2.5) and (7) as 0.9 MPa assuming the supply and tank pressures as 1.7 MPa and 0.1 MPa, respectively. The valve pressure drops, ΔP_{V1} and ΔP_{V2}, are obtained from (2.6) as 0.8 MPa. The flow through the control ports is calculated using (2.7). The variation of the control port area A_c with respect to the spool displacement is obtained from the solid model. Figure 3 shows the no-load mass flow rate variation with the spool displacement.

2.2. Load Flow Analysis

The flow analysis was carried out for different load conditions and flow is calculated using (2.7). The differential pressure across the valve is calculated using (2.6). Figure 4 shows the ideal normalized pressure-flow curves, of a jet pipe EHSV.

Corresponding to the input current to the torque motor there is a definite spool displacement. For the various spool displacements, there are different flow curves and they are all parabolas which pass through point 1 in either direction. To maximize the power
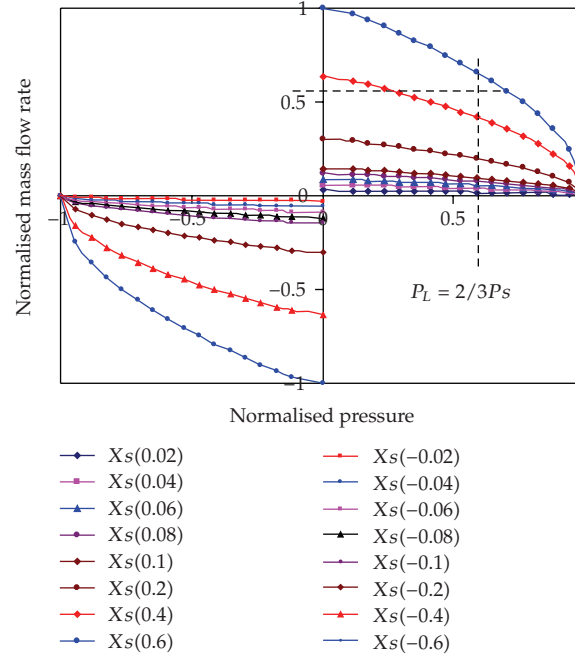
Figure 4: Mass flow rate in the load flow analysis.

output (H), the load pressure is 2/3rd of the supply pressure which provides sufficient flow to move the piston at the desired velocity [2].

3. Coupled Problems

Currently, solving coupled problems using FEM plays a major role in engineering field particularly in aerospace, mechanical and biomedical. In coupled problems, two or more physical systems interact with each other and independent solution of any one system being impossible without simultaneous solution of the others.

The coupled problems are classified as Class I and Class II depending on type of coupling that occur in the domains. In Class I type of problems, the coupling occurs in domain interfaces via the boundary conditions imposed at the interface, while in Class II problems domains overlap (totally or partially) and the coupling occurs through the governing differential equations [24].

The present paper aims to solve the Class I type problem that basically includes the Fluid-Structure-Interaction (FSI) where physically different domains (fluid and structure interaction or structure and structure interaction) interact with each other. The interfaces simply divide arbitrarily chosen regions in which different numerical discretization are used.

The examples of Class II problem are metal extrusion and soil dynamics. In the metal extrusion, the plastic flow is strongly coupled with the temperature field while at the same time the latter is influenced by the heat generated in the plastic flow. In earthquake response of a dam (Soil dynamics), the seepage flow and pressure interact with the dynamic behavior of the soil skeleton [24].

Modeling and analysis studies using numerical approach like FEM have not been carried out so far for the prediction of servovalve characteristics. The literature on applications of FEM to model the hydraulic components is limited. Some of the literature findings are as follows.

Tanaka and Sato [25] studied the affect of eddy current on the switching speed on/off and proportional solenoids by measuring the transient characteristics of the electric current and magnetic flux-flow, and next by calculating transient flux-flow on the process of permeating into the armature with the aid of newly programmed FEM.

Rosu and Vasiliu [26] applied the FE technique for the analysis of a swash plate axial piston pump components using the commercial FE code ANSYS. The analysis includes both static and dynamic study. The static analysis involved the study of static stresses and displacements of the yoke due to the hydrostatic pressure applied on the swash plate. The dynamic analysis includes the modal and transient analysis. The modal analysis revealed the natural frequencies and mode shapes of the components. The transient analysis revealed the response of the components under the action of a time-dependent pressure.

Meikandan et al. [27] and Sadashivappa [28] solved the 2D Reynold's equation, to determine the pressure distribution in the clearance between the piston and cylinder using the FEM. The film thickness equation for the clearance between the cylinder bore and the piston with form errors is introduced into the FEM formulation along with the operating parameters like piston velocity, supply pressure, viscosity of oil, and piston geometry. The program computes the pressure distribution around the circumference of the piston and the net centering force acting on the piston. Depending on the lateral load acting on the piston and the centering force developed, the eccentricity of piston inside the bore is determined. Here Galerkin's weighted residual method is used for obtaining the element equation.

4. Hydrostatic Fluid Model

In the present application of the actuator, it will be necessary to predict the mechanical response of a fluid-filled structure. The response of the structure depends not only on the external loads but also on the pressure exerted by the contained fluid. The major difficulty in these applications is the coupling between the structure and the fluid-filled cavity. In the present work, the mechanical structures are modeled with general purpose finite elements-beams, shells, and solids while the fluid cavities are modeled with special purpose hydrostatic fluid elements.

The general-purpose finite element formulation and their selection for different applications are widely elaborated by Cook [29]. The formulation of special-purpose hydrostatic fluid elements is detailed by Hibbitt et al. [30].

Hydrostatic fluid elements are surface elements that cover the boundary of the fluid cavity. Each cavity has a cavity reference node and has a single degree of freedom representing the pressure inside the fluid cavity. The cavity reference node is also used in calculating the cavity volume. The hydrostatic fluid elements appear as surface elements that cover the cavity boundary, but they are actually volume elements when the cavity reference node is accounted for. The fluid properties like bulk modulus, density, expansion coefficient, and so forth are defined properly for each cavity. The fluid inside the cavity can be compressible (pneumatic) or incompressible (hydraulic). The fluid volume of the cavity is a function of fluid pressure, fluid temperature, and fluid mass.

The volume $\overline{V}$ derived from the fluid pressure and temperature equals to the actual volume V of the cavity. In ABAQUS/Standard, this is achieved by augmenting the virtual work expression for the structure with constraint equation as

$$\overline{V} = f(p, \theta, m), \qquad V - \overline{V} = 0. \tag{4.1}$$

In addition, the virtual work contribution due to the cavity pressure is given by

$$\delta \Pi^* = \delta \Pi - p \delta V - \delta p \left(V - \overline{V} \right), \tag{4.2}$$

where $\delta \Pi^*$ is the augmented virtual work expression and $\delta \Pi$ is the virtual work expression for the structure without the fluid cavity. The negative signs imply that an increase in the cavity volume releases energy from the fluid. This represents a mixed formulation in which the structural displacements and fluid pressure are primary variables. The rate of the augmented virtual work expression is obtained as

$$
\begin{aligned}
d\delta \Pi^* \quad &= d\delta \Pi - p d\delta V - dp \delta V - \left(dV - d\overline{V} \right) \delta p, \\
d\delta \Pi^* \quad &= d\delta \Pi - p d\delta V - p d\delta V - dV \delta p + \frac{d\overline{V}}{dp} dp \delta p, \\
\delta \Pi^* \quad &= \delta \Pi - p \sum_e \delta V^e - \delta p \left[\sum_e V^e - \sum_e \overline{V}^e \right],
\end{aligned}
\tag{4.3}
$$

where $p d\delta V$ represents the pressure load stiffness, and $d\overline{V}/dp$ is the volume-pressure compliance of the fluid. Since the pressure is the same for all elements in the cavity, the augmented virtual work expression can be written as sum of the expressions for individual elements as

$$\delta \Pi^* = \sum_e \left[\delta \Pi^e - p \delta V^e - \delta p \left(V^e - \overline{V}^e \right) \right]. \tag{4.4}$$

Since the temperature is same for all elements in the cavity, the fluid volume can be calculated for each element individually as

$$\overline{V}^e = \overline{V}(p, \theta, m^e), \tag{4.5}$$

where m^e is the element mass. In the solution, the actual volume of the element may be different from the element volume as

$$V^e - \overline{V}^e \neq 0. \tag{4.6}$$

However, the total fluid volume will match the volume of the cavity. The fluid flow from one cavity to another cavity is modeled using two-node hydrostatic fluid link (FLINK), which is a function of the pressure difference, the average temperature, and the average pressure:

$$q = q\left(\Delta P, \overline{p}, \overline{\theta} \right). \tag{4.7}$$
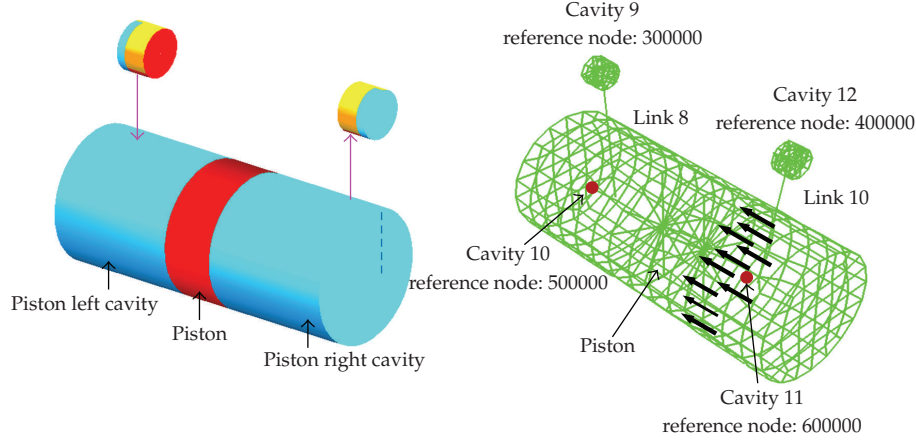
Figure 5: Solid and FE model of an actuator.

The flow rate needs to be integrated over a finite increment. The dependence on the average pressure is assumed to be weak. Hence, a semi-implicit method is used wherein Δp at the end of the increment and $\overline{p}^o$ at the start of the increment are used. For the temperature, $\overline{\theta}^m$ to be the average of $\overline{\theta}$ at the start and end of the increment is chosen, because this is likely to be the most accurate. Hence, the mass flow through the link is obtained as

$$\Delta m = q\left(\Delta P, \overline{p}^o, \overline{\theta}^m\right)\Delta t. \tag{4.8}$$

This mass change needs to be converted to a volume change in each cavity. It is assumed that the fluid in both cavities is the same, but the pressures (and possibly the temperatures) may be different. With the use of the pressure and the temperature-dependent density $\rho_i(p, \theta)$ for each cavity i, the relations become

$$\Delta \overline{V_1} = -\frac{\Delta m}{\rho_1}, \qquad \Delta \overline{V_2} = \frac{\Delta m}{\rho_2}. \tag{4.9}$$

5. Finite Element Modeling of Actuator

The finite element analysis carried out here to study the actuator under different external load conditions includes only the actuator and not the complete servovalve in order to study only the actuator dynamics. The fluid flow to the actuator for different load conditions is calculated and explained in Section 2. Figure 5 shows the solid and FE model of the actuator. Totally, there are four fluid cavities and two fluid links that are created. Cavity 9 and Cavity 12 correspond to the control port connections C_1 and C_2 which are in turn connected to either supply port or tank port depending on the servovalve spool displacement. Cavity 10 and Cavity 11 are piston end cavities. The piston is modeled with 3-dimensional 8-noded solid elements. A common boundary is created between the Cavity 10 and piston left face, similarly between Cavity 11 and piston right face. The FE model comprises of 402 elements, 582 nodes, and 3,356 variables.
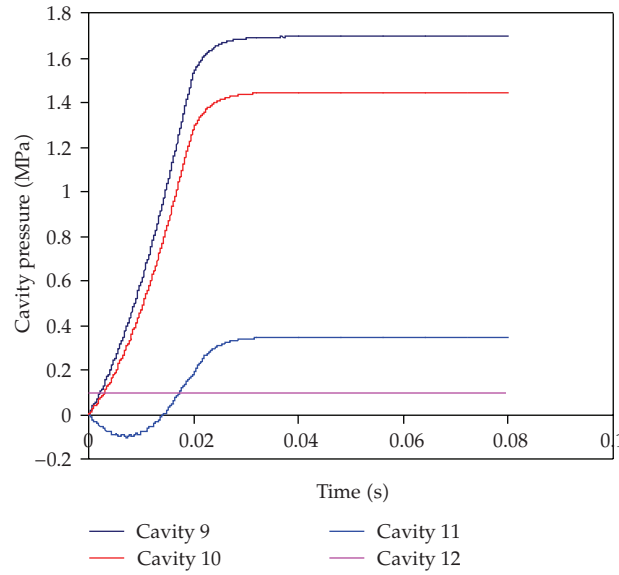
Figure 6: Variation of cavity pressures with time.

The FE analysis detailed here is for one load condition, which is selected as 2/3rd of the supply pressure of 1.7 MPa. Corresponding to this a distributed load (Dload) of 1.1 MPa is applied to the piston right face. In order to simulate the flow to the control ports C_1 or C_2, as the case may be, due to the displacement of the spool, a flux input (fluid) is added to the appropriate cavity. In the present case, the fluid flow responsible for the piston displacement is applied at Cavity 9 as a ramp input opposite to the applied Dload. The properties of the Links 8 and 10 are assumed to be same, about 0.2 MPa when full flow occurs through the link. As the flux is added to the Cavity 9, due to the fluid flow into the Cavity 10, the pressure in the Cavity 10 gradually increases. The variation of cavity pressures is shown in Figure 6.

Due to the application of the load on the right side face of the piston, initially the piston would have moved towards left, increasing the volume of Cavity 11 and at the same time reducing the volume of Cavity 10. Since no flow is present in any of the links, initially the pressure in the Cavity 11 shows a slight under pressure. As the flow into Cavity 10 starts, the pressure in this cavity starts increasing. At a value corresponding to the pressure required to overcome the load, the piston starts moving. From this point onwards, the pressure in the Cavity 11 also starts increasing due to a fluid flow through the Link 10 to the tank port (Cavity 12). The tank port (Cavity 12) pressure is maintained constant. When the steady state is reached, all the cavity pressures reach a constant value. The pressure difference between the pressures in the Cavity 9 and Cavity 10 as well as that between the Cavity 11 and Cavity 12 is equal and corresponds to the maximum pressure drop due to the flow input through Link 9 and that going out through Link 10. The two flows are same due to a symmetric actuator. The difference in pressure at the Cavities 10 and 11 corresponds to the load pressure drop.

The mass flow rate through the Links 8 and 10 is shown in Figure 7. As the fluid flows through the Link 8, it builds up pressure in the piston left side Cavity 10. After overcoming the applied load only, the piston starts displacing to right side and hence flows through Link 10 start after this delay.
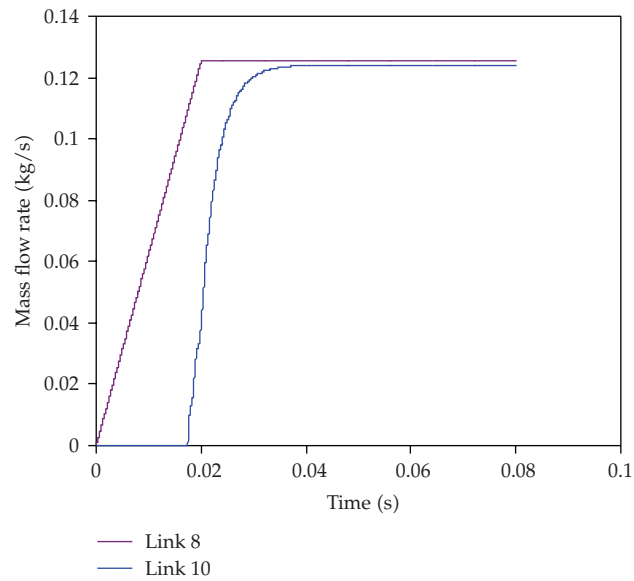
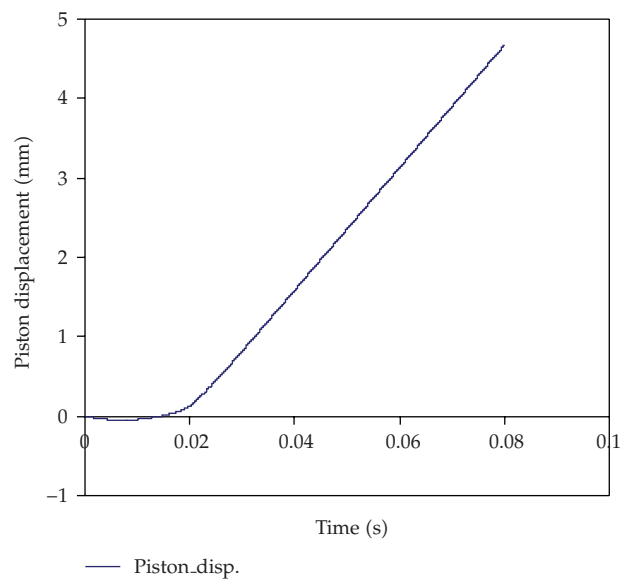**Figure 7:** Variation of mass flow rate with time.

Figure 8: Variation of piston displacement with time.

Figure 8 shows the piston displacement and even here the above effect can be seen. The piston starts moving only when the pressure in the Cavity 10 overcomes the applied load. Due to the bulk modulus of the fluid, the piston is displaced slightly in the opposite direction and then it moves to the right side.

The piston velocity obtained from the FE simulation is compared with that obtained by dividing the flow rate by the area of cross section of the piston. This is shown in Figure 9.
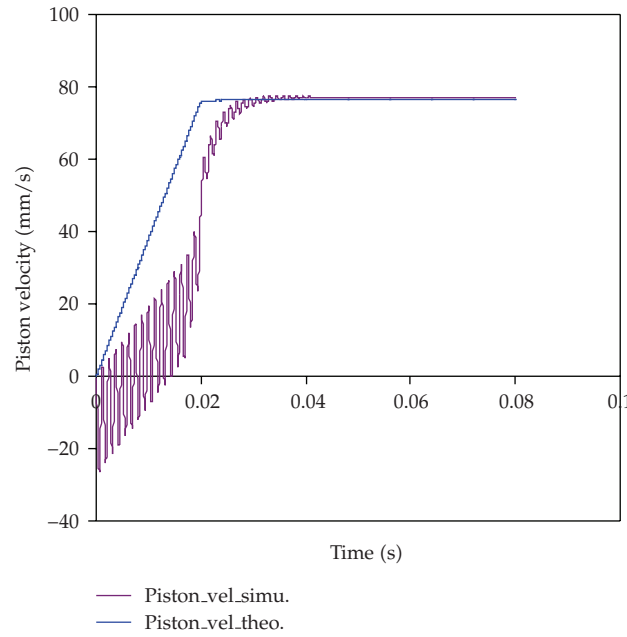
Figure 9: Comparison of piston velocity with time.

Even here, the piston velocity obtained through the FE simulation matches well with that calculated based on the fluid flow into the piston cavity and the piston area.

Figure 10 shows the piston end cavity volume variation. As stated earlier, as the piston starts moving, the cavity volume on the left side (Cavity 10) starts increasing and at the same time the cavity volume on the right side (Cavity 11) starts decreasing due to the piston displacement.

The bulk modulus of fluid affects the actuator dynamics. The variation in the bulk modulus is due to the air entrapment in the fluid flow. Hence, the analysis was extended to study the effect of change in the value of fluid bulk modulus.

Variation of piston displacement with time for different bulk modulus is shown in Figure 11. It is evident that as the bulk modulus reduces, the fluid compressibility increases and hence the delay in the start of the piston displacement increases.

6. Direct-Solution Steady-State Dynamics

Direct-solution steady-state dynamics is adopted to study the system's linearized response to harmonic excitation. The steady-state amplitude and phase of the system over the frequency range of excitation are obtained.

The main features of the direct-solution steady-state dynamic analysis [30] areas follows

(i) It is a linear perturbation procedure.

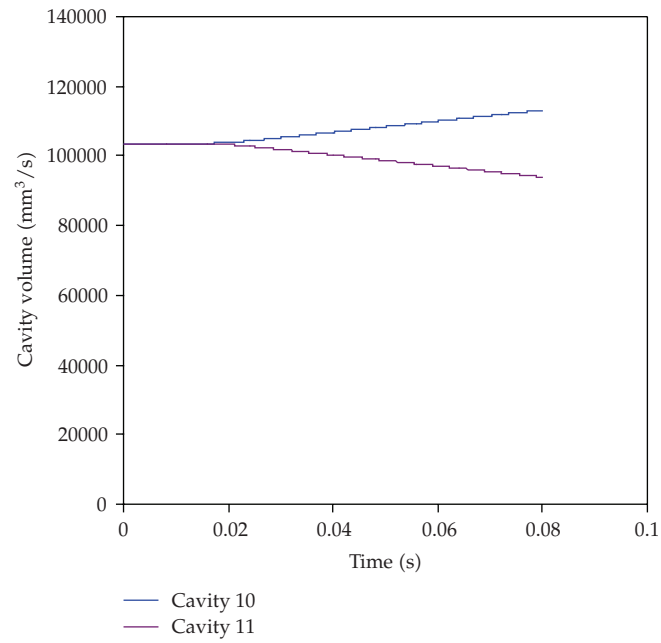(ii) It calculates the response directly in terms of the physical degrees of freedom of the model.

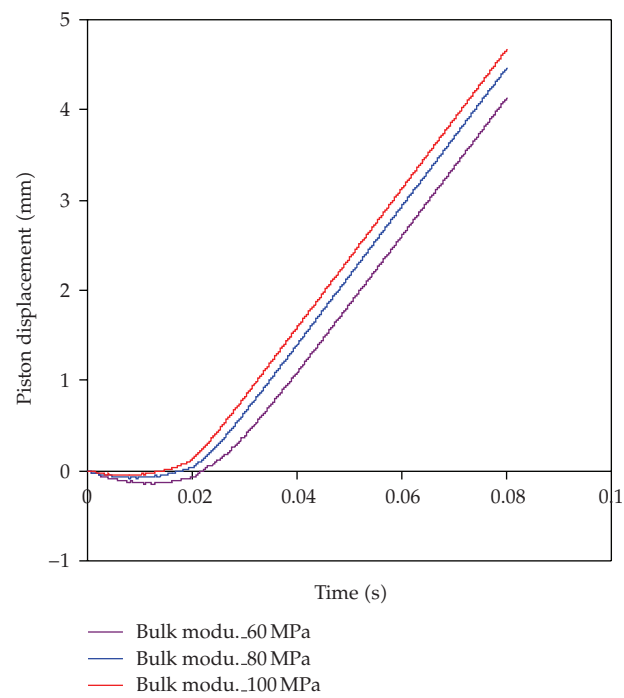**Figure 10:** Variation of piston cavity volume with time.

Figure 11: Variation of piston displacement for different bulk modulus with time.
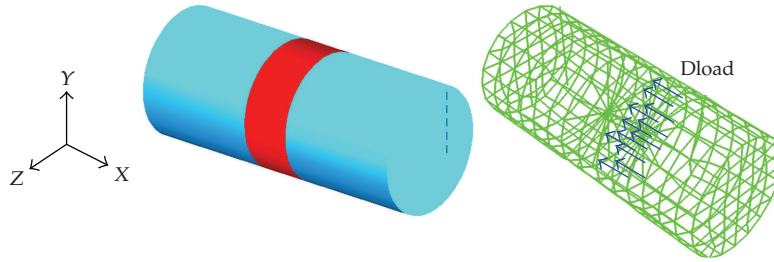
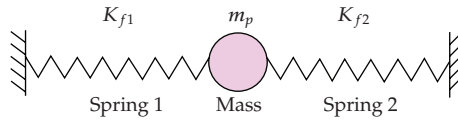**Figure 12:** Solid and FE model of actuator.

Figure 13: Simplified model of actuator: mass with two springs.

(iii) It is an alternative to mode-based steady-state dynamic analysis, in which the response of the system is calculated on the basis of the eigenmodes.

(iv) It is more expensive computationally than mode-based or subpace-based steady-state dynamics.

(v) It is more accurate than mode-based or subspace-based steady-state dynamics, in particular if significant frequency-dependent material damping or viscoelastic material behavior is present in the structure.

(vi) It is able to bias the excitation frequencies toward the approximate values that generate a response peak.

This procedure can be used in the following cases, for which the eigenvalues cannot be extracted using the mode-based analysis:

(i) for nonsymmetric stiffness

(ii) when any form of damping other than modal damping must be included,

(iii) when viscoelastic material properties must be taken into account.

7. Finite Element Modeling of an Actuator for Harmonic Excitation

Figure 12 shows the solid and FE model of the actuator. Figure 13 shows the simplified model of actuator. It consists of a mass with two springs. These two springs correspond to piston end fluid columns. The actuator considered here has an effective piston diameter of 51.84 mm, piston mass of 0.32 kg, and a stroke length of 50 mm on either side of the null position. The piston end cavities are modeled with 3-dimensional hydrostatic fluid elements and the piston with 3-dimensional 8-noded brick elements. The FSI is achieved using the common boundary between the piston end faces and fluid cavities. This is achieved while creating the fluid element by using the nodes of the piston end faces. The FE model comprises of 402 elements, 582 nodes, and 3356 variables.
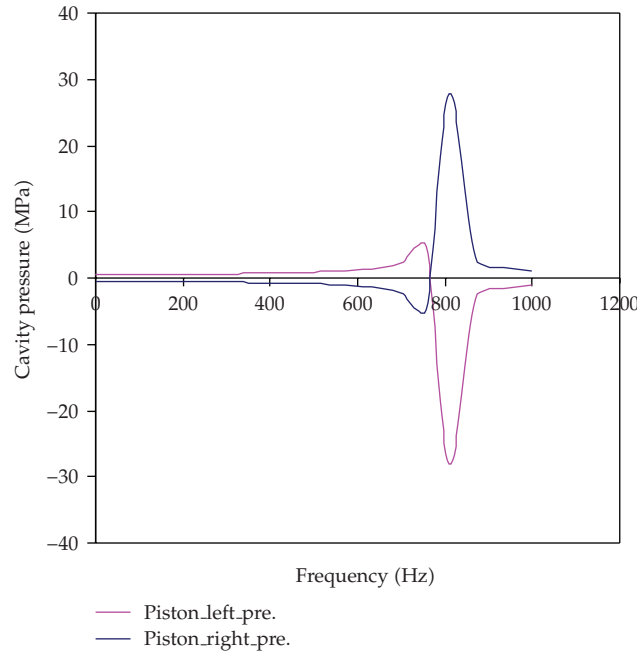
Figure 14: Variation of cavity pressure with frequency.

A distributed force of 1.1 MPa is applied on the right piston face. The frequency is swept from 1 to 1000 Hz. The variation of cavity volume and cavity pressure as well as the piston response is plotted. Figure 14 shows the variation of cavity pressure with frequency. The maximum pressure of 28 MPa in both cavities occurs at a frequency of 811 Hz. For particular combination of load on the piston and fluid bulk modulus, the appearance of the first peak is observed in Figure 14. This analysis is carried out for 100 MPa fluid bulk modulus. Figure 15 shows the variation of cavity pressure with frequency. This analysis is carried out for 80 MPa bulk modulus of fluid.

Figure 16 shows the variation of cavity volume on either side of the piston. It may be observed that maximum cavity volume of 29000 mm^3 occurs at a frequency of 811 Hz.

Figure 17 shows the piston response and the peak amplitude of 14 mm occurs at the same frequency of 811 Hz as before. It is an established fact that the natural frequency of the actuator depends on the piston positions and bulk modulus of the fluid [6, 7]. This was simulated and compared with the theoretical results. The FE simulation has been carried out for different piston position and bulk modulus and the results are shown in Figure 18.

Theoretical equation for calculating the natural frequency of a hydraulic cylinder is given by

$$\overline{\omega}_n = \sqrt{\frac{\beta}{m_p}\left[\frac{A_{AL}^2}{V_{AL}} + \frac{A_{AR}^2}{V_{AR}}\right]}. \tag{7.1}$$

The natural frequency of the actuator changes as the piston moves since the left and right cavity volume of the piston changes. The natural frequency of the actuator is
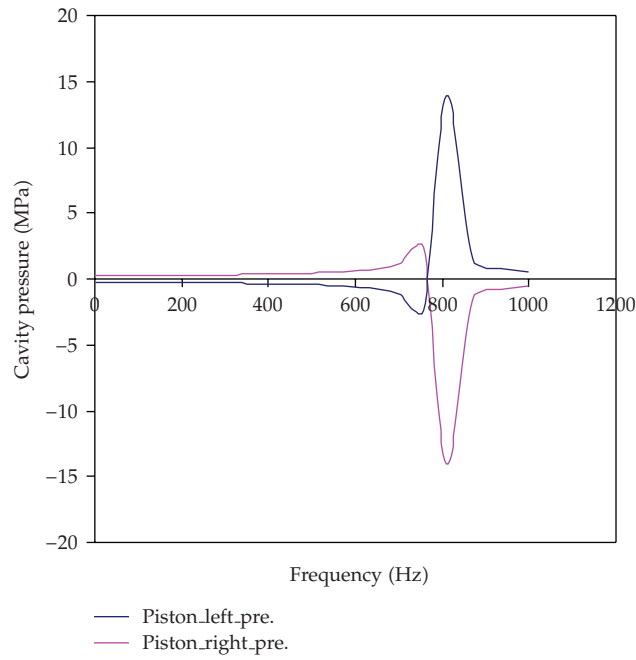
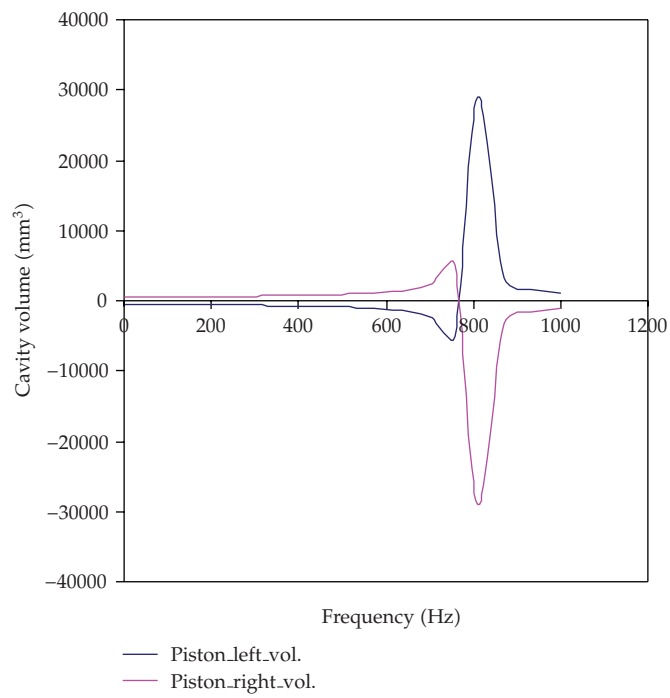**Figure 15:** Variation of cavity pressure with frequency.

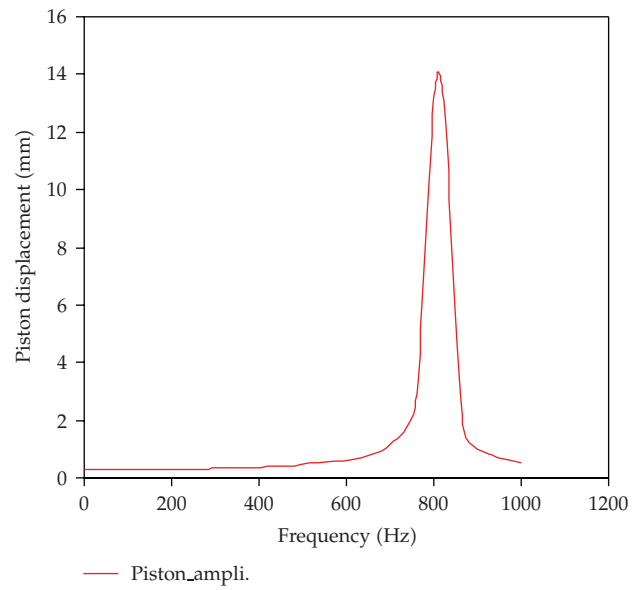**Figure 16:** Variation of cavity volume with frequency.
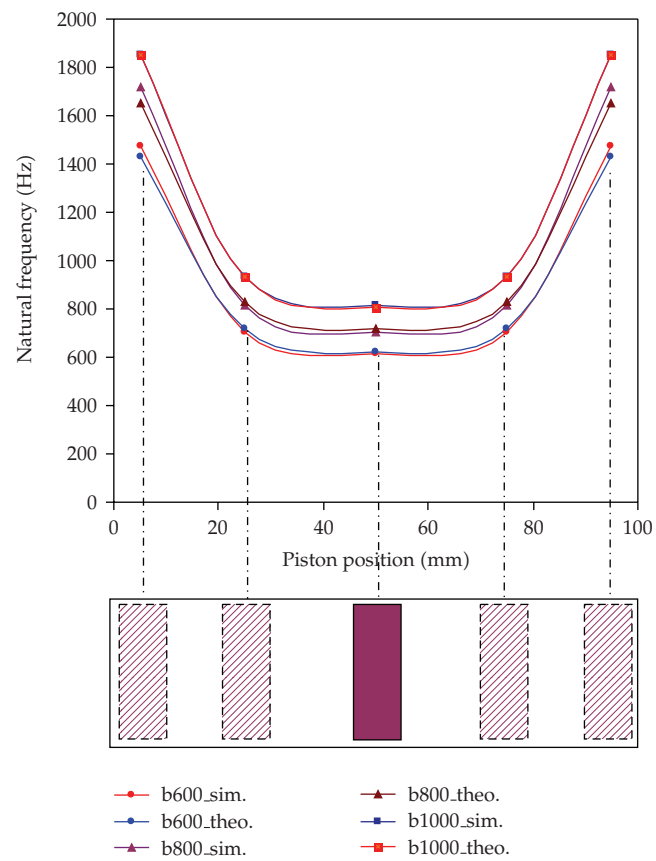
Figure 17: Frequency response of piston.

Figure 18: Variation of natural frequency of piston with piston position.

a maximum when the piston is fully retracted and decreases to a minimum at some point in the stroke and then increases to maximum as the piston approaches the end of the stroke. For a symmetric actuator the minimum natural frequency occurs when the piston is in mid position. The closeness of the theoretical and simulation results is observed (Figure 18).

8. Conclusions

(i) Finite element approach was adopted to model the fluid-structure interaction of servovalve component-actuator.

(ii) The dynamics of the jet pipe servovalve actuator was presented as the variation of cavity pressure, cavity volume, mass flow rates, and the actuator velocity with time.

(iii) The effect of bulk modulus was observed—as bulk modulus reduces, the fluid compressibility increases and hence the delay in the start of the piston displacement increases.

(iv) It was observed that the natural frequency of the actuator changes as the piston moves since the left and right cavity volume of the piston changes.

(v) The natural frequency of the actuator is a maximum when the piston is fully retracted and decreases to a minimum at some point in the stroke and then increases to maximum as the piston approaches the end of the stroke.

(vi) The natural frequency as obtained through FE modeling is well matched with theoretical modeling.

Notation

A_C: Control port area, mm^2
A_{AL}, A_{AR}: Actuator left side and right side cavity area, mm^2,
C_1, C_2:   Control ports
Q_{ac}:   Mass flow rate through the actuator, kg/s
Q_1 to Q_5:   Fluid flow at the second-stage spool valve, l/min.
P: Pressure port
K_{f1}, K_{f2}:   Fluid column spring stiffness, N/mm
m_a: Actuator mass, kg
m_p: Piston mass, kg
P_{C1}, P_{C2}: Control port pressure, MPa
P_L: Load pressure, MPa
P_s: Supply pressure, MPa
P_t: Tank pressure, MPa
T: Tank port
V_{AL}, V_{AR}: Actuator left side and right side cavity volume, mm^3
ΔP_{V1}:   Valve pressure drop between the supply port and control port C_1, MPa
ΔP_{V2}:   Valve pressure drop between the control port C_2 and tank port, MPa
ω_n: Natural frequency of an actuator, Hz
β: Bulk modulus of oil, MPa
ρ_o: Density of oil, kg/mm^3.

References

[1] C. M. Allen, *Electrohydraulic Servomechanisms*, McGraw-Hill, London, UK, 1963.

[2] H. E. Merrit, *Hydraulic Control Systems*, John Wiley & Sons, Chichester, UK, 1967.

[3] R. H. Maskrey and W. J. Thayer, "A brief history of electrohydraulic servomechanisms," *Journal of Dynamic Systems, Measurement and Control*, vol. 100, no. 2, pp. 110–116, 1978.

[4] D. C. Clark, *Selection and Performance Criteria for Electrohydraulic Servodrives*, Moog Technical Bulletin no. 122, Moog, East Aurora, NY, USA, 1969.

[5] T. P. Neal, *Performance Estimation for Electrohydraulic Control System*, Moog Technical Bulletin no. 126, Moog, East Aurora, NY, USA, 1974.

[6] F. D. Norvelle, *Electrohydraulic Control Systems*, Prentice-Hall, Upper Saddle River, NJ, USA, 2000.

[7] T. J. Viersma, *Analysis, Synthesis and Design of Hydraulic Servosystems and Pipelines*, Elsevier, Amsterdam, The Netherlands, 1980.

[8] J. Watton, *Fluid Power Systems: Modeling, Simulation, Analog and Microcomputer Control*, Prentice-Hall, London, UK, 1989.

[9] R. Walters, *Hydraulic and Electro-Hydraulic Servo Systems*, Iliffe Books, London, UK, 1967.

[10] S. Lee and J. F. Blackburn, "Contribution to hydraulic control-1. Steady-state axial forces on control-valve pistons," *Transactions of the ASME*, vol. 74, no. 6, pp. 1005–1011, 1952.

[11] C. K. Taft and J. P. Twill, "An analysis of the three-way underlapped hydraulic spool servovalve," *Journal of Dynamic Systems, Measurement, and Control*, vol. 100, pp. 117–123, 1978.

[12] I. Lee, T. Kim, and A. Kitagawa, "A new linearized equation for modelling a servovalve in hydraulic control systems," in *Proceedings of the JFPS International Symposium on Fluid Power*, vol. 3, pp. 899–904, Nara, Japan, 2002.

[13] P. N. Nikiforuk, P. R. Ukrainetz, and S. C. Tsai, "Detailed analysis of a two-stage four-way electrohydraulic flow-control valve," *Journal of Mechanical Engineering Science*, vol. 11, pp. 168–174, 1969.

[14] S. LeQuoc, R. M. H. Cheng, and A. Limaye, "Investigation of an electrogydraulic servovalve with tuneable return pressure and drain orifice," *Transactions of the ASME*, vol. 276, no. 109, pp. 276–285, 1987.

[15] B. B. Das, S. K. Zachariah, and B. Sebastian, "Nonlinear modeling for launch vehicle electrohydraulic servo actuation system," in *Proceedings of the National Conference on Emerging Challenges in Fluid Power Technology*, pp. 1–9, 2003.

[16] J. J. Fredlake and M. R. Adams, "Electronic fuel controls for executive jet aircraft," Tech. Rep. 831478, SAE, 1983.

[17] S. L. Bruce, "Development of control surface actuation systems on various configurations of the F-16," Tech. Rep. 831483, SAE, 1983.

[18] H. E. Harschburger, "Development of redundant flight control actuation systems for the F/A-18 strike fighter," Tech. Rep. 831484, SAE, 1983.

[19] K. Ogata, *Modern Control Engineering*, Prentice-Hall, New Delhi, India, 4th edition, 2003.

[20] S. Z. Dushkes and S. L. Cahn, "Analysis of some hydraulic components used in regulators and servo mechanisms," *Transactions of the ASME*, pp. 595–601, 1952.

[21] J. U. Thoma, "Fluid mechanics, bond graphs and jet pipe servo valves," *Modeling and Simulation of Systems*, pp. 77–81, 1989.

[22] P. D. Henri, J. M. Hollerbach, and A. Nahvi, "An analytical and experimental investigation of a jet pipecontrolled electropneumatic actuator," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 4, pp. 601–611, 1998.

[23] Moog Inc., *The Deflector Jet Servovalve*, Moog Technical Bulletin no. 121, Moog, East Aurora, NY, USA.

[24] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method-I*, Butterworth–Heinemann, London, UK, 2000.

[25] H. Tanaka and Y. Sato, "FEM analysis and measurement on the eddy current in magnetic circuit of PWM on/off and proportional solenoid valves," in *Proceedings of the 3rd JHPS International Symposium*, pp. 385–390, 1996.

[26] C. Rosu and N. Vasiliu, "Researches on the main components of a positive displacement pump by FEM," in *Proceedings of the 2nd International FPNI—PhD Symposium*, pp. 1–6, Modena, Italy, July 2002.

[27] N. Meikandan, R. Raman, M. Singaperumal, and K. N. Seetharamu, "Theoretical analysis of tapered pistons in high speed hydraulic actuators," *Wear*, vol. 137, no. 2, pp. 299–321, 1990.

[28] K. Sadashivappa, *The effect of form errors on the performance of hydraulic actuators*, Ph.D. thesis, Indian Institute of Technology Madras, Chennai, India, 1996.

[29] R. D. Cook, *Concepts and Applications of Finite Element Analysis*, John Wiley & Sons, New York, NY, USA, 1974.

[30] *ABAQUS/Standard Finite Element Simulation Package (ver.6.4), Vols. I-III*, Hibbitt, Karlsson & Sorensen, Pawtucket, RI, USA, 2002.